INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team asap

As Scalable As Possible: foundations of large scale dynamic distributed systems

Rennes - Bretagne-Atlantique, Saclay - Île-de-France

Theme : Distributed Systems and Services

Table of contents

1. Team

Research Scientists

Anne-Marie Kermarrec [Team Leader, Research Director, HdR]
Davide Frey [Research Scientist, INRIA SACLAY - ILE DE FRANCE SUD (since October 2010)]
Fabrice Le Fessant [Research Scientist, INRIA SACLAY - ILE DE FRANCE SUD (until July 2010)]
Aline Carneiro Viana [Research Scientist, INRIA SACLAY - ILE DE FRANCE SUD (until July 2010 and on sabbatical since 2009)]

Faculty Members

Achour Mostefaoui [Associate Professor (MdC), University Rennes 1, HdR]
Michel Raynal [Professor (Pr), University Rennes 1, HdR]
Marin Bertier [Assistant Professor (MdC), INSA Rennes]

Technical Staff

Nicolas Destor [Ingénieur-Expert (until July 2010)]
Davide Frey [Ingénieur-Expert (until September 2010)]
Guang Tan [Ingénieur-Expert (until October 2010)]

PhD Students

Xiao Bai [INSA UT - China Scholarship Council (until December 2010)]
François Bonnet [MENRT Grant (until July 2010)]
Antoine Boutet [INRIA Grant]
Kévin Huguenin [MENRT Grant]
Damien Imbs [MENRT Grant]
Konstantinos Kloudas [INRIA Grant]
Alexandre Van Kempen [Cifre Technicolor Grant (since February 2010)]
Nicolas Le Scouarnec [Cifre Technicolor Grant (until November 2010)]
Vincent Leroy [MENRT Grant (until September 2010)]
Afshin Moin [INRIA Grant]
Tyler Crain [Marie-Curie European Grant (since April 2010)]
Arnaud Jegou [INRIA Grant (since October 2010)]
Mohammad Nabil Al-Aggan [MENRT Grant (since October 2010)]

Post-Doctoral Fellows

Christopher Thraves-Caro [Post-Doc (Since March 2009)]
Silvija Kokalj-Filipovic [Post-Doc (until August 2010)]

Administrative Assistants

Christine Biard [INRIA SACLAY - ILE DE FRANCE SUD (until July 2010)]
Cécile Bouton [INRIA RENNES - BRETAGNE ATLANTIQUE]

Others

Olivier Baldellon [(until June 2010)]
Arnaud Jegou [(until June 2010)]
Mohammad Nabil Al-Aggan [(until July 2010)]

2. Overall Objectives

2.1. General objectives

Recent evolutions in distributed computing significantly increased the degree of uncertainty inherent to any distributed system and led to a scale shift that traditional approaches can no longer accommodate. The key to scalability in this context lies into fully decentralized and self-organizing solutions. The objective of the ASAP project team is to provide a set of abstractions and algorithms to build serverless, large-scale, distributed applications involving a large set of volatile, geographically distant, potentially mobile and/or resource-limited computing entities.

The ASAP Project-Team is engaged in research along three main themes: *Distributed computing models and abstractions*, *Peer-to-peer distributed systems and applications* and *Data management in wireless autonomic networks*. These research activities encompass both basic research, seeking conceptual advances, and applied research, to validate the proposed concepts against real applications.

2.1.1. A challenging new setting

Distributed computing was born in the late seventies when people started taking into account the intrinsic characteristics of physically distributed systems. The field then emerged as a specialized research area distinct from networks, operating systems and parallelism. Its birth certificate is usually considered as the publication in 1978 of Lamport's most celebrated paper "*Time, clocks and the ordering of events in a distributed system*" [105] (that paper was awarded the Dijkstra Prize in 2000). Since then, several high-level journals and (mainly ACM and IEEE) conferences are devoted to distributed computing. This distributed system area has continuously been evolving, following the progresses in all the abovementioned areas such as networks, computing architecture, operating systems. We believe that the changes that occurred in the past decade involve a paradigm shift that can be much more than a "simple generalization" of previous works. Several conferences such as NSDI and IEEE P2P were created, acknowledging this evolution. The NSDI conference is an attempt to reassemble the networking and system communities while the IEEE P2P conference was created to be a forum specialized in peer-to-peer systems. At the same time, the EuroSys conference has been created as an initiative of the European Chapter of the ACM SIGOPS to gather the system community in Europe.

The past decade has been dominated by a major shift in scalability requirements of distributed systems and applications mainly due to the exponential growth of network technologies (Internet, wireless technology, sensor devices, etc.). Where distributed systems used to be composed of up to a hundred of machines, they now involve thousand to millions of computing entities scattered all over the world and dealing with a huge amount of data. In addition, participating entities are highly dynamic, volatile or mobile. Conventional distributed algorithms designed in the context of local area networks do not scale to such extreme configurations. Therefore, they have to be revisited to fit into this new challenging setting. Precisely, *scalability* is one of the main focus of the ASAP project-team. Our ambitious goal is to provide the algorithmic foundations of large-scale dynamic distributed systems, ranging from abstractions to real deployment.

More specifically, distributed computing as such is characterized by how a set of distributed entities, whether they are called processes, agents, sensors, peers, processors or nodes, having only a partial knowledge of many parameters involved in the system, communicate and collaborate to solve a specific problem. While parallelism and real-time deal respectively with efficiency and on-time computing, distributed computing can be characterized by the word *uncertainty*. Uncertainty used to be created by the effect of asynchrony and failures in traditional distributed systems, it is now the result of many other factors. These include process mobility, low computing capacity, network dynamics, scale, and more recently the strong dependence on personalization which characterizes user-centric Web 2.0 applications. This creates new challenges such as the need to manage large quantities of personal data in a scalable manner while guaranteeing the privacy of users.

2.1.2. Mastering uncertainty in distributed computing

The peer-to-peer communication paradigm emerged in the early 2000s and is now one of the prevalent models to cope with the requirements of large-scale dynamic distributed systems. In order to successfully manage the increasing level of uncertainty, distributed systems should now rely on the following properties:

Fully decentralized model: A fully decentralized system does not rely on any central entity to control the system. Participating entities may act both as clients and servers. The number of potential servers thus increases linearly with the size of the system, avoiding the performance bottleneck imposed by the presence of servers in traditional distributed systems. Such systems are therefore naturally protected from failures since there is no single point of failure and many services are naturally replicated.

Self-organizing capabilities: Participating entities are by essence highly dynamic as they might be disconnected, mobile or faulty. The system should be able to handle such dynamic behaviors and get

automatically reorganized to face entities arrival and departure.

Local system knowledge: Individual entities behavior is based on a restricted knowledge of the system and yet the system should converge toward global properties.

The objective of the ASAP project-team is to cope efficiently with the intrinsic uncertainty of distributed systems and provide the foundations for a new family of distributed systems for which scalability, dynamics, and privacy are first class concerns, and to provide the basis for the design and the implementation of distributed algorithms suited to this new challenging setting. More specifically, our objectives are to work on the following complementary axes:

Distributed computing models and abstractions: While many protocols have been proposed dealing with dynamic large-scale systems, there is still a lack of formal definitions with respect to the underlying computing model. In this area, our objectives are to investigate distributed computing problem solvability, and define a realistic model for dynamic systems along with the related abstractions.

Customizable overlay networks for scalability: Many peer-to-peer overlay networks, organizing nodes in a logical network on top of a physical network, have been proposed in the past five years in order to deal with large-scale and dynamic behavior. Following this trend, we intend to step away from general-purpose overlay networks that have been proposed so far and build domain-specific overlays customized for a given application and/or functionality. In particular, our core target has recently been the design and evaluation of large-scale user-centric applications with a focus on personalization and privacy. Among these applications are for example personalized search, notification and content dissemination.

2.2. Models and abstractions for large-scale distributed computing

A very relevant challenge (maybe a Holy Grail) lies in the definition of a computation model appropriate to dynamic systems. This is a fundamental question. As an example there are a lot of peer-to-peer protocols but none of them is formally defined with respect to an underlying computing model. Similarly to the work of Lamport on "static" systems, a model has to be defined for dynamic systems. This theoretical research is a necessary condition if one wants to understand the behavior of these systems. As the aim of a theory is to codify knowledge in order it can be transmitted, the definition of a realistic model for dynamic systems is inescapable whatever the aim we have in mind, be it teaching, research or engineering.

2.2.1. Distributed computability

Among the fundamental theoretical results of distributed computing, there is a list of problems (e.g., consensus or non-blocking atomic commit) that have been proved to have no deterministic solution in asynchronous distributed computing systems prone to failures. In order such a problem to become solvable in an asynchronous distributed system, that system has to be enriched with an appropriate oracle (also called failure detector). We have been deeply involved in this research and designed optimal consensus algorithms suited to different kind of oracles. This line of research paves the way to rank the distributed computing problems according to the "power" of the additional oracle they required (think of "additional oracle" as "additional assumptions"). The ultimate goal would be the statement of a distributed computing hierarchy, according to the minimal assumptions needed to solve distributed computing problems (similarly to the Chomsky's hierarchy that ranks problems/languages according to the type of automaton they need to be solved).

2.2.2. Distributed computing abstractions

Major advances in sequential computing came from machine-independent data abstractions such as sets, records, etc., control abstractions such as while, if, etc., and modular constructs such as functions and procedures. Today, we can no longer envisage not to use these abstractions. In the "static" distributed computing field, some abstractions have been promoted and proved to be useful. Reliable broadcast, consensus, interactive consistency are some examples of such abstractions. These abstractions have well-defined specifications. There are both a lot of theoretical results on them (mainly decidability and lower bounds), and numerous implementations. There is no such equivalent for dynamic distributed systems.

2.3. User-centric peer-to-peer architectures

Recent research has shown that effective management of resources on a large scale, be them computing resources, data, events, bandwidth, requires fully decentralized solutions. This need is even more important when applications aim to be user-centric with significant personalization features. The need to manage huge amounts of personal information on a large scale therefore impacts all aspects of the design of distributed applications such as the management of the relevant overlay networks, resource discovery and information dissemination.

In GOSSPLE, and the other projects within ASAP, we are tackling these challenges through a number of contributions. First we are combining our expertise in the domain of large scale peer-to-peer overlays with techniques from the data-mining community. Applications objects become themselves peers (although obviously hosted on a physical computing entity), while their data directly influences the overlay links. This allows us to define overlays that directly take into account application characteristics. For example, users sharing similar interests can be aggregated to improve the performance of search, or recommendation applications [32].

Second, we strongly believe in weakly-structured networks and most of our work relies on epidemic-based unstructured overlays. Epidemic communication models have recently started to be explored as a general paradigm to build and maintain unstructured overlay networks. The basic principle of such epidemic protocols is that periodically, each peer exchanges information with some other peers selected from a local list of neighbors. Such protocols have shown to be extremely resilient to network dynamics [104].

Finally, we are convinced that we can greatly benefit from the experience gathered from both existing systems and theoretical models. We spend a significant amount of energy to find, gather and analyze workloads of real systems as well as to develop our own platform in the context of our peer-to-peer collaborative backup platform.

2.4. Malicious behaviors in large scale networks

A failure model is always considered and clearly stated when designing fault-tolerant applications. The most benign faults consist of processes that execute their protocol correctly before silently stopping execution. However, processes may exhibit malicious (or arbitrary) behaviors (commonly called Byzantine processes), voluntarily or not. A Byzantine process can send spurious information, send multiple information to processes, etc. Such a behavior could be due to an external attack or even to an unscrupulous person with administrative access. More generally, Byzantine processes can also cooperate to maximize the damage caused to the system. We refer to the notion of "adversary". When defining the system failure model, it is necessary to explicit the assumed adversary. For example, can the adversary delay messages exchanged among correct processes? Can the adversary delay a correct process (by jamming the system)? Can the Byzantine processes cooperate? Is the computational power of Byzantine processes "unbounded"? In such a case, the use of cryptography is useless.

Considering malicious behaviors is therefore related to fault-tolerance but it is also at the core of security. System security encompasses a family of mechanisms and techniques that help protect the system from internal and external attacks. These mechanisms control different aspects of the system (cryptography, secured links, controlled access, etc.). Protecting a distributed system, partially under the control of an adversary is an extremely challenging task, particularly when the system itself is managing potentially sensitive personal information. Dealing with process crashes is far from being trivial, many problems are known to be impossible in pure asynchronous systems. Assuming Byzantine processes complicates the problem even further.

Finally, managing malicious behaviors is not only one of the hottest topics in today's distributed computing, but it is also a stringent requirement for all user-centric applications, and thus one of the constituting pillars of GOSSPLE. The complexity of human behavior introduces new nuances to failure models applicable to social distributed systems. Users may not be completely byzantine, but they may, for example, act rationally in an attempt to maximize their benefit with the minimal amount of work. This opens new opportunities for models and protocols that can be more efficiently applied in the context of real system.

2.5. Highlights of the year

1. **Best paper award** ICDCS 2010: Champel Mary-Luc; Huguenin Kévin ; Kermarrec Anne-Marie; Le Scouarnec Nicolas. LT Network Codes. In 30th International Conference on Distributed Computing Systems (ICDCS) [44].
2. **Best student paper award** DISC 2010: Bonnet François; Raynal Michel. Anonymous Asynchronous Systems: the Case of Failure Detectors. In Symposium on Distributed Computing (DISC'09) [73]
3. **Distinguished paper** EUROPAR 2010: Imbs Damian; Raynal Michel. The x-Wait-freedom Progress Condition. In European Parallel Computing Conference (EUROPAR'10) [53]
4. **Talks at College de France** by Anne-Marie Kermarrec and Michel Raynal in January 2010.
5. **Talk at the British Royal Society** by Anne-Marie Kermarrec "Web science: a new Frontieer" for the 350 anniversary of the BRS.
6. **Nomination** of Michel Raynal at the Institut Universitaire de France

3. Scientific Foundations

3.1. Introduction

Research activities within the ASAP Project-Team encompass several areas in the context of large-scale dynamic systems: models and abstraction, user-centric distributed architectures, and high-bandwidth data dissemination. We provide a brief presentation of some of the scientific foundations associated with them.

3.2. Models and abstractions of large-scale dynamic systems

Finding models for distributed computations prone to asynchrony and failures has received a lot of attention. A lot of research in that domain focuses on what can be computed in such models, and, when a problem can be solved, what are its best solutions in terms of relevant cost criteria. An important part of that research is focused on distributed computability: what can be computed when failure detectors are combined with conditions on process input values for example. Another part is devoted to model equivalence. What can be computed with a given class of failure detectors? Which synchronization primitives is a given failure class equivalent to?). Those are among the main topics addressed in the leading distributed computing community. A second fundamental issue related to distributed models, is the definition of appropriate models suited to dynamic systems. Up to now, the researchers in that area consider that nodes can enter and leave the system, but do not provide a simple characterization, based on properties of computation instead of description of possible behaviors [107], [98], [99]. This shows that finding dynamics distributed computing models is today a "Holy Grail", whose discovery would allow a better understanding of the essential nature of dynamics systems.

3.3. Peer-to-peer overlay networks

As mentioned before, the past decade has been dominated by a major shift in scalability requirements of distributed systems and applications mainly due to the exponential growth of the Internet. A standard distributed system today is related to thousand or even millions of computing entities scattered all over the world and dealing with a huge amount of data. In this context, the peer-to-peer communication paradigm imposed itself as the prevalent model to cope with the requirements of large scale distributed systems. Peer-to-peer systems rely on a symmetric communication model where peers are potentially both client and servers. They are fully decentralized, thus avoiding the bottleneck imposed by the presence of servers in traditional systems. They are highly resilient to peers arrivals and departures. Finally, individual peer behavior is based on a local knowledge of the system and yet the system converges toward global properties.

A peer-to-peer overlay network logically connect peers on top of IP. Two main classes of such overlays dominate, structured and unstructured. The differences relate to the choice of the neighbors in the overlay, and the presence of an underlying naming structure. Overlay networks represent the main approach to build large-scale distributed systems that we retained. An overlay network forms a logical structure connecting participating entities on top of the physical network, be it IP or a wireless network. Such an overlay might form a structured overlay network [108], [109], [110] following a specific topology or an unstructured network [103], [111] where participating entities are connected in a random or pseudo-random fashion. In between, lie weakly structured peer-to-peer overlays where nodes are linked depending on a proximity measure providing more flexibility than structured overlays and better performance than fully unstructured ones. Proximity-aware overlays connect participating entities so that they are connected to close neighbors according to a given proximity metric reflecting some degree of affinity (computation, interest, etc.) between peers. We extensively use this approach to provide algorithmic foundations of large-scale dynamic systems.

3.4. Epidemic protocols

Epidemic algorithms, also called gossip-based algorithms [101], [100], are consistently used in our research. In the context of distributed systems, epidemic protocols are mainly used to create overlay networks and to ensure a reliable information dissemination in a large-scale distributed system. The principle underlying the technique, in analogy with the spread of a rumor among humans via gossiping, is that participating entities continuously exchange information about the system in order to spread it gradually and reliably. Epidemic algorithms have proved efficient to build and maintain large-scale distributed systems in the context of many applications such as broadcasting [100], monitoring, resource management, search, and more generally in building unstructured peer-to-peer networks.

3.5. Malicious process behaviors

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

4. Application Domains

4.1. Panorama

The results of the research targeted in ASAP span over a wide range of application areas ranging from Internet-based applications, Grid computing, and wireless autonomic networked systems. Most applications are nowadays distributed and we believe that many new potential applications are yet to be discovered.

To tackle our challenging goals, we focus on a few sets of applications, which we believe are representative of large-scale distributed applications. More specifically, the constraints imposed by those applications are representative of those we deal with in ASAP.

4.2. User-centric decentralized web

The Web 2.0 has radically changed the way people interact with the Internet turning it from a read-only infrastructure to a collaborative read-write platform with active players. Users can now classify Web content based on their interests, and share it with their friends or even unknown users. The content is no longer generated only by experts but pretty much by everyone. Web 2.0 applications, such as LastFM, Flickr, Delicious, Twitter, or Digg, contain a goldmine of information. Yet, matching a specific query in such a mine might rapidly turn out to be like looking for a needle in a haystack. Content is continually changing and is not indexed with a controlled vocabulary, e.g, *ontology*. Rather, millions of users can choose arbitrary keywords to *tag* billions of items, forming a complex *folksonomy* (folk + taxonomy).

The freedom left to the users to express their interests underlies the success of Web 2.0 but is also an impediment to navigation. Within GOSSPLE, we have been proposing novel mechanisms to tackle these challenge. Rather than relying on centralized architecture as most Web 2.0 applications are doing, we focus achieving high levels of scalability and privacy through decentralization. This means obtaining effective ways to manage and exploit personalized information in the presence of limited resources such as computing power and bandwidth.

To maximize and at the same time evaluate the impact of our research, we are operating directly on a number of applications. As an example we have been working on personalized search, collaborative storage systems, resource discovery and large-scale personalized content distribution and indexing. This is evident in our production comprising both scientific results and software.

4.3. Streaming

One key application in the context of the Internet of the future is the dissemination of high-bandwidth content. Be it video or audio, multimedia content is constantly being streamed to millions of devices around the world. This introduces significant technical challenges for researchers and application designers. The significant need for bandwidth that characterizes these applications calls for decentralized protocols that can optimize the resource utilization. Such protocols should be able to operate in highly heterogeneous environments by adapting the quality of the streamed content and their dissemination mechanisms according to available resources.

Our research in this context focuses on the application of epidemic protocols to the problem of content dissemination. This covers aspects that include the design of scalable and adaptable dissemination protocols, their integration with network coding techniques, and the ability to operate in the presence of malicious users by identifying and punishing freeriding behaviors.

5. Software

5.1. GossipLib: library for effective development of gossip-based applications

Participants: Davide Frey, Anne-Marie Kermarrec.

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Library for Gossip protocols
Status:	released version 0.7alpha

GossipLib is a library consisting of a set of JAVA classes aimed to facilitate the development of gossip-based application in a large-scale setting. It provides developers with a set of support classes that constitute a solid starting point for building any gossip-based application. GossipLib is designed to facilitate code reuse and testing of distributed application and as thus also provides the implementation of a number of standard gossip protocols that may be used out of the box or extended to build more complex protocols and applications. These include for example the peer-sampling protocols for overlay management.

GossipLib also provides facility for the configuration and deployment of applications as final-product but also as research prototype in environments like PlanetLab, clusters, network emulators, and even as event-based simulation. The code developed with GossipLib can be run both as a real application and in simulation simply by changing one line in a configuration file.

5.2. YALPS

Participants: Davide Frey, Anne-Marie Kermarrec.

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Library for Gossip protocols
Status:	released version 0.3alpha

YALPS is an open-source Java library designed to facilitate the development, deployment, and testing of distributed applications. Applications written using YALPS can be run both in simulation and in real-world mode without changing a line of code or even recompiling the sources. A simple change in a configuration file will load the application in the proper environment. A number of feature make YALPS useful both for the design and evaluation of research prototypes and for the development of applications to be released to the public. Specifically, YALPS makes it possible to run the same application as a simulation or in a real deployment without a single change in the code. Applications communicate by means of application-defined messages which are then routed either through UDP/TCP or through YALPS's simulation infrastructure. In both cases, YALPS's communication layer offers features for testing and evaluating distributed protocols and applications. Communication channels can be tuned to incorporate message losses or to constrain their outgoing bandwidth.

The work has been done in collaboration with Maxime Monod (EPFL).

5.3. HEAP: Heterogeneity-aware gossip protocol.

Participants: Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec.

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Java Application
Status:	release & ongoing development

This work has been done in collaboration with Vivien Quéma (CNRS Grenoble), Maxime Monod and Rachid Guerraoui (EPFL), and has lead to the development of a video streaming platform based on HEAP, *HEterogeneity-Aware gossip Protocol*. The platform is particularly suited for environment characterized by heterogeneous bandwidth capabilities such as those comprising ADSL edge nodes. HEAP is, in fact, able to dynamically leverage the most capable nodes and increase their contribution to the protocol, while decreasing by the same proportion that of less capable nodes. During the last few months, we have integrated HEAP with the ability to dynamically measure the available bandwidth of nodes, thereby making it independent of the input of the user.

5.4. WhatsUp: A Distributed News Recommender

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

Contact:	Antoine Boutet
Licence:	Open Source
Presentation:	A Distributed News Recommender
Status:	Beta version

WhatsUp is a distributed recommendation system aimed to distribute instant news in a large scale dynamic system. WhatsUp has two parts, an embedded application server in order to exchange with others peers in the system and a fully dynamic web interface for displaying news and collecting opinions about what the user reads. Underlying this web-based application lies Beep, a biased epidemic dissemination protocol that delivers news to interested users in a fast manner while limiting spam. Beep is parameterized on the fly to manage the orientation and the amplification of news dissemination. Every user forwards the news of interest to a randomly selected set of users with a preference towards those that have similar interests (orientation). The notion of interest does not rely on any explicit social network or subscription scheme, but rather on an implicit and dynamic overlay capturing the commonalities between users with respect to they are interested in. The size of the set of users to which a news is forwarded depends on the interest of the news (amplification).

5.5. AskBuddies

Participants: Davide Frey, Anne-Marie Kermarrec, Guang Tan.

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Web Application
Status:	Deployed

AskBuddies is a Web-based personalized Q&A service developed on the opensocial platform. Serving as a third-party application embedded in current online social services such as Orkut, MySpace, etc, it allows a question asker to find appropriate answerers from the network of her direct or indirect acquaintances. We have implemented AskBuddies as a web application and have deployed it on Orkut to start collecting user data.

6. New Results

6.1. Panorama

Our research activities range from theoretical bounds to practical protocols and implementations for large-scale distributed dynamic systems. The target applications range from Internet-based applications to wireless autonomic networks. We focus our research on two main areas: resource management and dissemination. We believe that such services are basic building blocks of many distributed applications. We also examine these services in two networking contexts: Internet and wireless sensors. These two classes of applications, although exhibiting very different behaviors and constraints, clearly require scalable solutions.

To achieve this ambitious goal, we tackle the issues both along the theoretical and practical sides of scalable distributed computing and ASAP is organized along the following themes:

1. Models and abstractions: dealing with dynamics;
2. User-centric web, the main project in this area is the GOSSPLE project.
3. Streaming

6.2. Models and abstractions: dealing with dynamics

6.2.1. Signature-Free Broadcast-Based Intrusion Tolerance

Participants: Achour Mostefaoui, Michel Raynal.

Providing application processes with strong agreement guarantees despite failures is a fundamental problem of fault-tolerant distributed computing. Correct processes must not to be “polluted” by the erroneous behavior of faulty processes. This paper considers the consensus agreement problem in a setting where some processes can behave arbitrarily (Byzantine behavior). In such a context it is possible that Byzantine processes collude to direct the correct processes to decide on a “bad” value (a value proposed only by faulty processes).

We proposed in [61] several contributions. We present a family of consensus algorithms in which no bad value is ever decided by correct processes. These processes always decide a value they have proposed (and this is always the case when they all propose the same value) or a default value $\perp$. These algorithms are called *intrusion-free* consensus algorithms. To that end, each consensus algorithm is based on an appropriate underlying broadcast algorithm. One of these abstractions, called *validated broadcast* is new and allows the design of a resilience-optimal consensus algorithm (i.e., it copes with up to $t < n/3$ faulty processes where n is the total number of processes). All proposed consensus algorithms assume the underlying system is enriched with additional computational power provided by a binary Byzantine consensus algorithm. We present also a resilience-optimal randomized binary consensus algorithm based on the validated broadcast abstraction. An important feature of all these algorithms lies in the fact that they are signature-free (and hence particularly efficient).

6.2.2. Time-Free Authenticated Byzantine Consensus

Participant: Achour Mostefaoui.

The work [62], done in collaboration with Hamouma Moumen from University of Béjaia, proposes a simple protocol that solves the authenticated Byzantine Consensus problem in asynchronous distributed systems. To circumvent the FLP impossibility result in a deterministic way, synchrony assumptions should be added. In the context of Byzantine failures for systems where at most t processes may exhibit a Byzantine behavior and where not all the system is assumed eventually synchronous, Moumen et al. provide the main result. They assume at least one correct process, called $2t$-bisource, connected with $2t$ privileged neighbors with eventually timely outgoing and incoming links. We show that a deterministic solution for the authenticated byzantine consensus problem is possible if the system model satisfies an additional assumption that does not rely on physical time but on the pattern of messages that are exchanged. The basic message exchange between processes is the query-response mechanism. To solve the Consensus problem, we assume a correct process p, called $\diamond 2t$-winning process, and a set Q of $2t$ processes such that, eventually, for each query issued by p, any process q of Q receives a response from p among the $(n - t)$ first responses to that query. The processes in the set Q can exhibit a Byzantine behavior and this set may change over time. Whereas many time-free solutions have been designed for the consensus problem in the crash model, this is, to our knowledge, the first time-free deterministic solution to the Byzantine consensus problem.

6.2.3. A Necessary and Sufficient Condition for Byzantine Consensus

Participants: Olivier Baldellon, Achour Mostefaoui, Michel Raynal.

Solving the consensus problem requires in one way or another that the underlying system satisfies synchrony assumptions. Considering a system of n processes where up to $t < n/3$ may commit Byzantine failures. We have investigated the synchrony assumptions that are required to solve consensus and presented a corresponding necessary and sufficient condition in [28].

Such a condition is formulated with the notions of a symmetric synchrony property and property ambiguity. A symmetric synchrony property is a set of graphs, where each graph corresponds to a set of bi-directional eventually synchronous links among correct processes. Intuitively, a property is ambiguous if it contains a graph whose connected components are such that it is impossible to distinguish a connected component that contains correct processes only from a connected component that contains faulty processes only. We have then connected the notion of a symmetric synchrony property with the notion of eventual bi-source, and shown that the existence of a virtual $\diamond[t + 1]$bi-source is a necessary and sufficient condition to solve consensus in presence of up to t Byzantine processes in systems with bi-directional links and message authentication. Finding necessary and sufficient synchrony conditions when links are timely in one direction only, or when processes cannot sign messages, still remains open (and very challenging) problems.

6.2.4. Anonymous Asynchronous Systems: the Case of Failure Detectors

Participants: François Bonnet, Michel Raynal.

Trivially, agreement problems such as consensus, that cannot be solved in non-anonymous asynchronous systems prone to process failures, cannot be solved either if the system is anonymous. This work investigated failure detectors that allow processes to circumvent this impossibility. It has several contributions. It first presents four failure detectors (denoted AP, $\overline{AP}$, $A\Omega$ and $A\Sigma$) and show that they are the "identity-free" counterparts of the two perfect failure detectors, eventual leader failure detectors and quorum failure detectors, respectively. $A\Sigma$ is new and showing that $A\Sigma$ and Σ have the same computability power in a non-anonymous system is not trivial. We also showed that the notion of failure detector reduction is related to the computation model. We proved correct a uniform anonymous consensus algorithm based on the failure detector pair $(A\Omega, A\Sigma)$ ("uniform" means that not only processes have no identity, but no process is aware of the total number of processes). This new algorithm is not a "straightforward extension" of an algorithm designed for non-anonymous systems. To benefit from $A\Sigma$, it uses a novel message exchange pattern where each phase of every round is made up of sub-rounds in which appropriate control information is exchanged. Finally, we examined the notions of failure detector hierarchy, weakest failure detector for anonymous consensus, and the implementation of identity-free failure detectors in anonymous systems. The results of this work were published at DISC 2010 [35].

6.2.5. *The multiplicative power of consensus numbers*

Participants: Damien Imbs, Michel Raynal.

The Borowsky-Gafni (BG) simulation algorithm is a powerful reduction algorithm that shows that t-resilience of decision tasks can be fully characterized in terms of wait-freedom. Said in another way, the BG simulation shows that the crucial parameter is not the number n of processes but the upper bound t on the number of processes that are allowed to crash. The BG algorithm considers colorless decision tasks in the base read/write shared memory model. (Colorless means that if, a process decides a value, any other process is allowed to decide the very same value.)

This work considers system models made up of n processes prone to up to t crashes, and where the processes communicate by accessing read/write atomic registers (as assumed by the BG) and (differently from the BG) objects with consensus number x, accessible by at most x processes (with $x \leq t < n$). Let $ASM(n,t,x)$ denote such a system model. While the BG simulation has shown that the models $ASM(n,t,1)$ and $ASM(t+1,t,1)$ are equivalent, this paper focuses the pair (t,x) of parameters of a system model. Its main result is the following: the system models $ASM(n_1,t_1,x_1)$ and $ASM(n_2,t_2,x_2)$ have the same computational power for colorless decision tasks if and only if $\lfloor \frac{t_1}{x_1} \rfloor = \lfloor \frac{t_2}{x_2} \rfloor$. As can be seen, this contribution complements and extends the BG simulation. It shows that consensus numbers have a multiplicative power with respect to failures, namely the system models $ASM(n,t',x)$ and $ASM(n,t,1)$ are equivalent for colorless decision tasks iff $(t \times x) \leq t' \leq (t \times x) + (x-1)$. This work has been published in PODC 2010 [52].

6.2.6. *Asymmetric progress conditions*

Participants: Damien Imbs, Michel Raynal.

Wait-freedom and obstruction-freedom have received a lot of attention in the literature. These are symmetric progress conditions in the sense that they consider all processes as being "equal". Wait-freedom has allowed to rank the synchronization power of objects in presence of process failures, while obstruction-freedom (that is a much weaker progress condition) allows for simpler and more efficient object implementations.

This work introduces the notion of asymmetric progress conditions. Given an object O in a read/write system of n processes, such a condition assumes that O can be accessed by a subset of $y \leq n$ processes only (i.e., O has y ports), and requires that O guarantees wait-freedom for x processes and obstruction-freedom for the remaining $y-x$ processes. The paper investigates the power of such a progress condition, called (y,x)-liveness ((n,n)-liveness is wait-freedom while $(n,0)$-liveness is obstruction-freedom). Our main contributions are the following. (1) An impossibility result showing that $(n,1)$-liveness cannot be obtained from $(n-1,n-1)$-live objects (i.e., from any number of wait-free objects with $n-1$ ports). (2) An (n,x)-liveness hierarchy for $0 \leq x \leq n$. (3) An impossibility result showing that there is no consensus algorithm that

is obstruction-free with respect to all processes and fault-free with respect to one process only even if one can use underlying $(n-1, n-1)$-live consensus objects (fault-freedom means here that at least one process has to always terminate when there are no crashes). (4) An algorithm based on (x, x)-live objects that constructs a consensus object with an asymmetric group-based progress condition. This work has been published in PODC 2010 [54].

6.2.7. Software transactional memories

Participants: Tyler Crain, Damien Imbs, Michel Raynal.

The aim of a Software Transactional Memory (STM) is to discharge the programmers from the management of synchronization in multiprocess programs that access concurrent objects. To that end, a STM system provides the programmer with the concept of a transaction. The job of the programmer is to design each process the application is made up of as a sequence of transactions. A transaction is a piece of code that accesses concurrent objects, but contains no explicit synchronization statement. It is the job of the underlying STM system to provide the illusion that each transaction appears as being executed atomically. Of course, for efficiency, a STM system has to allow transactions to execute concurrently. Consequently, due to the underlying STM concurrency management, a transaction commits or aborts.

This work studies the relation between two STM properties (read invisibility and permissiveness) and two consistency conditions for STM systems, namely, opacity and virtual world consistency. Both conditions ensures that any transaction (be it a committed or an aborted transaction) reads values from a consistent global state, a noteworthy property if one wants to prevents abnormal behavior from concurrent transactions that behave correctly when executed alone. A read operation issued by a transaction is invisible if it does not entail shared memory modifications. This is an important property that favors efficiency and privacy. An STM system is permissive with respect to a consistency condition if it accepts every history that satisfies the condition. This is a crucial property as a permissive STM system never aborts a transaction "for free". We show that read invisibility, permissiveness and opacity are incompatible, which means that there is no permissive STM system that implements opacity while ensuring read invisibility. We also show that invisibility, permissiveness and opacity are compatible. To that end we describe a new STM protocol called IR_VWC_P. This protocol presents additional noteworthy features: it uses only base read/write objects and locks which are used only at commit time; it satisfies the disjoint access parallelism property; and, in favorable circumstances, the cost of a read operation is $O(1)$.

6.3. GOSSPLE : A radically new approach to navigating the digital information universe

GOSSPLE is the topic of the ERC Starting Grant led by Anne-Marie Kermarrec.

While the Internet has fully moved into homes, creating tremendous opportunities to exploit the huge amount of resources at the edge of the network, the Web has changed dramatically over the past years. There has been an exponential growth of user-generated content (Flickr, Youtube, Delicious, ...) and a spectacular development of social networks (Twitter, FaceBook, etc.). This represents a fantastic potential in leveraging such kinds of information about the users: their circles of friends, their interests, their activities, the content they generate. This also reveals striking evidence that navigating the Internet goes beyond traditional search engines. New and powerful tools that could empower individuals in ways that the Internet search will never be able to do are required.

The objective of GOSSPLE is to provide an innovative and fully decentralized approach to navigating the digital information universe by placing *users affinities and preferences* at the heart of the navigation process. Building on the peer to peer communication paradigm and harnessing the power of gossip-based algorithms, GOSSPLE aims at personalizing Web navigation, by means of a fully decentralized solution, for the sake of scalability and privacy.

6.3.1. The Gossple anonymous social network

Participants: Marin Bertier, Davide Frey, Anne-Marie Kermarrec, Vincent Leroy.

This work [32] was carried out in collaboration with Prof. Rachid Guerraoui (EPFL). GOSSPLE exploits the social dimension of the Internet to get "related" users indirectly connected and refine each other's filtering procedures through implicit preferences. The network is organized around such preferences and affinities between users. Such a network of affinities is at the heart of GOSSPLE. The GOSSPLE anonymous network provides each user with a personalized view of the network through a *thrifty decentralized* protocol that *automatically* infer *personalized* connections in Internet-scale systems. GOSSPLE nodes continuously gossip digests of their corresponding interest profiles and locally compute a personalized view of the network which is then leveraged to improve their Web navigation. The view covers *multiple interests* without any explicit support (such as explicit social links or ontology) and without violating *anonymity*: the association between users and profiles is hidden.

Basically, every GOSSPLE node has a proxy, chosen randomly, gossiping its profile digest *on its behalf*; the node transmits its profile to its proxy in an encrypted manner through an intermediary, which cannot decrypt the profile. To reduce bandwidth consumption, the gossip exchange procedure is *thrifty*: nodes do not exchange profiles but only Bloom filters of those until time reveals that the two nodes might indeed benefit from the exchange. To limit the number of profiles maintained by each node, while encompassing the various interests of the user associated with the node, we introduce a new *set cosine similarity*, as a generalization of the classical *cosine similarity* metric and an effective heuristic to compute it. The work was presented at the Middleware conference in November 2010.

6.3.2. Gossip-based top-k processing in Gossple

Participants: Xiao Bai, Marin Bertier, Anne-Marie Kermarrec, Vincent Leroy.

This work [27] has been done in collaboration with Prof. Rachid Guerraoui (EPFL). A fine grained personalization to process top-k queries requires to maintain inverted lists on a user basis, relying on the information held by users that share interests. This is almost impossible to achieve in a centralized approach as the storage required is prohibitively large and the maintenance of millions of users' inverted lists would overwhelm a central server. Alternatively, each user may be in charge of storing her entire personal inverted lists. This requires each user to store the information stored by all *related* users. This number potentially grows linearly with the number of users and would be also ultimately be prohibitively large.

Instead, we propose the first fully decentralized personalized top-k algorithm. The inverted lists are not precomputed, but computed on the fly based on information collected in a fully decentralized manner in the network. Each user identifies its personal network by gossiping user profiles and measuring similarity between users. Yet, each user stores a very small number of full profiles (say 20) along with the ID of the users of her personal network. Each top-k query is then gossiped in the network, harvesting at each hop the relevant information. Partial results are remotely computed and sent back to the requester who sees her request refined by the minute. While the network is maintained at a low frequency to avoid overloading the network, top-k queries speeds up that frequency, refreshing the part of the network involved in the query and generating a wave of updates in the personalization process. Results obtained on a 10,000 Delicious trace show that with each node storing 20 profiles, top-k queries are satisfied in less than 10 cycles.

6.3.3. On-the-fly personalized top-k processing

Participants: Xiao Bai, Anne-Marie Kermarrec.

This work has been done in collaboration with Prof. Rachid Guerraoui (EPFL). The rapidly increasing amount of user-generated content in collaborative tagging systems provides a huge source of information. Yet, performing effective search in these systems is very challenging, especially when we seek the most appropriate items that match a potentially ambiguous query. Personalization is appealing in this context as it limits the search for the items within a small network of participants with similar preferences. Off-line personalization, which consists in maintaining, for every user, a network of similar participants based on their tagging behaviors, is effective for queries that are close to the seeker's tagging profile but performs poorly when the queries, depicting emerging interests, have little correlation with the seeker's profile.

We present the first algorithms, DT^2 (Do Top-k Twice) and DT^{2}P^2 (Peer-to-Peer Do Top-k Twice), which perform the personalization on-line in centralized and peer-to-peer systems respectively. Both algorithms rely on the hybrid interest of the seeker, taking into account both her profile and her current query, to determine the network of similar participants for processing her query. In DT^2, at query time, we first execute a top-k processing on users to associate such a network to the seeker, then we process the query within this network to find the k most relevant items. In DT^{2}P^2, such a network, as well as the top-k results within this network, are gradually refined with the partial knowledge of the users in the system. We evaluate DT^2 and DT^{2}P^2 on CiteULike and Delicious traces involving up to $50,000$ users and highlight the advantages of on-line personalization. The experimental results also convey the scalability of DT^2 and DT^{2}P^2 in terms of storage and processing time.

6.3.4. Decentralized recommender systems

Participants: Afshin Moin, Vincent Leroy, Anne-Marie Kermarrec, Christopher Thraves-Caro.

The need for efficient decentralized recommender systems has been appreciated for some time, both for the intrinsic advantages of decentralization and the necessity of integrating recommender systems into P2P applications. On the other hand, the accuracy of recommender systems is often hurt by data sparsity. In this work we first focused on comparing different decentralized user-based and item-based Collaborative Filtering (CF) algorithms with each other. Then we proposed a new user-based random walk approach customized for decentralized systems, specifically designed to handle sparse data. We showed how the application of random walks to decentralized environments is different from the centralized version. We examine the performance of our random walk approach in different settings by varying the sparsity, the similarity measure and the neighborhood size. In addition, we introduce the *popularizing* disadvantage of the significance weighting term traditionally used to increase the precision of similarity measures, and elaborate how it can affect the performance of the random walk algorithm. Simulations on MovieLens 10,000,000 ratings dataset demonstrate that over a wide range of sparsity, our algorithm outperforms other decentralized CF schemes. Moreover, our results show decentralized user-based approaches perform better than their item-based counterparts in P2P recommender applications. The results of this work were published at Opodis 2010 [55].

6.3.5. Converging Quickly to Independent Uniform Random Topologies

Participants: Anne-Marie Kermarrec, Vincent Leroy, Christopher Thraves-Caro.

The peer sampling service is a core building block for gossip protocols in peer-to-peer networks. Ideally, a peer sampling service continuously provides each peer with a sample of peers picked uniformly at random in the network. While empirical studies have shown that uniformity was achieved, analysis proposed so far assume strong restrictions on the topology of the overlay network it continuously generates. In this work, we analyze a *Generic Random Peer Sampling Service* (GRPS) that satisfies the desirable properties for any peer sampling service –*small views*, *uniform sample*, *load balancing*, and *independence*– and relieve strong degree connections in the nodes assumed in previous works. The main result we prove is: starting from any simple (without loops and parallel edges) directed graph with out-degree equal to c for all nodes, and recursively applying GRPS, eventually results in a *random* simple directed graph with out-degree equal to c for all nodes. We test empirically convergence time and independence time for GRPS. Finally, We use this empirical evaluation to show that GRPS performs better than previously presented peer sampling services. This work has been accepted for publication at PDP 2011.

6.3.6. Peer-to-peer polling without cryptography

Participants: Kévin Huguenin, Anne-Marie Kermarrec.

This work has been done in collaboration with Rachid Guerraoui, Andrei Giurgiu, and Maxime Monod (EPFL, Switzerland). The emergence of social networks provides a framework for polling a community easily by the mean of peer-to-peer techniques. Polling is not as critical as voting as the accuracy on the tally is less important. Yet, it must provide similar properties to electronic voting, such as voter privacy, fairness, and probabilistic accuracy. The core idea of the work is to build a decentralized protocol without cryptography ensuring this properties with high probability by the mean of peer-to-peer deterrent power : *every action in the protocol may*

be subject to verification by peers. Ensuring that any malicious action is detected with probability one or at least close to one, we increase the accuracy of the tally by limiting the proportion of peers misbehaving. Privacy is ensured probabilistically by using peers as proxies for emitting ballots making vote recovery impossible for reasonable proportion of malicious nodes. We developed an extension to operate with aggregation functions, which was published at SSS 2010 [48].

6.3.7. Social-aware navigable peer-to-peer storage systems

Participants: Kévin Huguenin, Anne-Marie Kermarrec, Konstantinos Kloudas.

In this work we are addressing the problem of scalability and navigability in distributed storage systems for socially-interconnected content. The main focus of our research is to see how the "social" aspect of the content in these sites, can be leveraged to build scalable, distributed storage systems with properties that enable the easy navigation of users through the content graph.

The main idea behind the proposed architecture is that by letting the content graph guide the replica placement strategy, one can obtain a scalable storage infrastructure that translates content proximity in the graph into storage proximity in the overlay. The latter, combined with the fact that users usually navigate through the content graph by picking the next content fetched among the neighbors of the current one, is very important as it enables efficient content localization. In our system, each item is stored along with a subset of its related items on the same node and we make sure that nodes storing related items are close in the overlay topology.

The above architecture is evaluated through extensive trace-driven simulations using data from the largest and most popular on-line video-sharing website, i.e., Youtube, where the social interconnections between contents are captured by the so-called "related video" feature.

6.3.8. Availability-aware storage systems

Participants: Anne-Marie Kermarrec, Alexandre Van Kempen.

In this work [97] - done in collaboration with Erwan Le Merrer from Technicolor, Rennes, France - we focused on leveraging hosts availability patterns in order to increase distributed storage system performances. We show that by using the availability history of hosts, the performances of two important faces of distributed storage can be significantly improved namely (i) replica placement is achieved based on complementary nodes with respect to nodes availability, improving the overall data availability, and (ii) repairs can be scheduled according to node availability, so as to decrease the number of repairs while achieving comparable availability; this is achieved by an adaptive per-node timeout, instead of relying on a system-level timeout. We propose practical heuristics for those two issues. We evaluate our approach through extensive simulations based on real and well-known availability traces. Results clearly show the benefits of our approach with regards to the critical tradeoff between availability, load-balancing and bandwidth consumption.

6.3.9. Distributed social graph embedding

Participants: Anne-Marie Kermarrec, Vincent Leroy.

In collaboration with Gilles Trdan, Deutsch telecom, we proposed SoCS (Social Coordinate Systems), a fully distributed algorithm that embeds any social graph in an Euclidean space. Inspired by recent works on non-isomorphic embeddings, the SoCS embedding preserves the community structure of the original graph, while relying on gossip protocols which provide a scalable and churn-resilient behavior. Nodes thus get assigned coordinates that reflect their social position. We illustrate the use of these social coordinates through two applications, namely link prediction and long link detection. Link prediction consists in predicting new links in a social graph. Thus, SoCS can be used to predict new friends in a P2P social network. Long link detection was formalized by Kleinberg in 2006 and refers to the problem of identifying long links in a small-world graph. SoCS provides a novel approach to solve both problems in distributed systems, where they only received little attention.

Experiments on various real and synthetic data sets have showed that SoCS provide accurate and cheap answers to these both problems. SoCS even outperforms state of the art centralized link prediction algorithms.

6.3.10. Private similarity computation in distributed systems: from cryptography to differential privacy

Participants: Mohammad Nabil Al-Aggan, Anne-Marie Kermarrec.

In collaboration with Sbastien Gambs (ADEPT), we address the problem of computing the similarity between two users (according to their profiles) while preserving their privacy in a fully decentralized system and for the passive adversary model. First, we introduce a novel two-party protocol for privately computing a threshold version of the similarity and apply it to well-known similarity measures such as the scalar product and the cosine similarity. The output of this protocol is only one bit of information telling whether or not two users are similar beyond a predetermined threshold.

Afterwards, we explore the computation of the exact and threshold similarity within the context of differential privacy. Differential privacy is a recent notion developed within the field of private data analysis guaranteeing that an adversary that observes the output of the differentially private mechanism, will only gain a negligible advantage (up to a privacy parameter) from the presence (or absence) of a particular item in the profile of a user. This provides a strong privacy guarantee that holds independently of the auxiliary knowledge that the adversary might have. More specifically, we design several differentially private variants of the exact and threshold protocols that rely on the addition of random noise tailored to the sensitivity of the considered similarity measure. We also analyze their complexity as well as their impact on the utility of the resulting similarity measure. Finally, we provide experimental results validating the effectiveness of the proposed approach on real workloads.

6.3.11. AskBuddy: Search over social networks

Participants: Anne-Marie Kermarrec, Guang Tan.

AskBuddies is a Web-based personalized Q&A service developed on the opensocial platform. Serving as a third-party application embedded in current online social services such as Orkut, MySpace, etc, it allows a question asker to find appropriate answerers from the network of her direct or indirect acquaintances.

AskBuddies considers personalization along two dimensions: content and social distance. The former is of the traditional type, reflecting users' interest match, while the latter is a new type. The social distance based personalization is intended to capture the following observation: that given constrained human effort (in terms of energy, time, etc) devoted to the service, and for equally interesting questions, an answerer tends to offer help first to askers that come close to her in social relations. We propose decentralized methods to calculate this social distance, in its simplest form measured by link hop count. The 2D personalization is embodied by a ranking algorithm that provides, for each user, a local list of questions in order of her servicing preference, resembling the ranking effort by a standard web search engine, except with an additional sociological consideration.

We have deployed this service on Orkut and started collecting user data. We hope our practice will shed light on the understanding and design of a broader range of networked systems where human instead of machines act as servers, and where sociological behaviors play a role in resource/load distribution.

6.3.12. Cold start link prediction in social network

Participant: Vincent Leroy.

This work [59] has been done in collaboration with B. Barla Cambazoglu and F. Bonchi from Yahoo Research, Spain. In the traditional link prediction problem, a snapshot of a social network is used as a starting point to predict, by means of graph-theoretic measures, the links that are likely to appear in the future. In this paper, we introduce *Cold Start Link Prediction* as the problem of predicting the structure of a social network when the network itself is totally missing while some other information regarding the nodes is available. We propose a two-phase method based on the *Bootstrap Probabilistic Graph*. The first phase generates an implicit social network under the form of a probabilistic graph. The second phase applies probabilistic graph-based measures to produce the final prediction. We assess our method empirically over a large data collection obtained from Flickr, using interest groups as the initial information. The experiments confirm the effectiveness of our approach.

6.3.13. WhatsUp: P2P news recommender

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

WhatsUp is a recommendation system aimed to distribute instant news in a large scale dynamic system with no central bottleneck, single point of failure or censorship authority. Users express their opinions about the news they see by pushing a button (like-dislike) or simply ignoring them. The users expect to subsequently receive more interesting news, following a collaborative filtering scheme, based on what news have interested users with similar taste without explicitely subscribe to topics of interest. Underlying WhatsUp are two distributed protocols. The first, Wup, dynamically clusters active users with similar tastes using a multi-interest cosine similarity metric accounting for emerging interests. The second, Beep, is used to disseminate news from a user to a randomly selected set of users with a bias towards those with similar interests. The size of that set itself depends on the interest expressed by the user about the news it read. WhatsUp has been implemented and extensively tested using traces involving thousands users from system Digg. We report on its efficiency to deliver relevant news to interested users while limiting spam. This work has been carried out in collaboration with Rachid Guerraoui from EPFL and was demonstrated at P2P 2010 [41].

6.4. Streaming and dynamic systems

6.4.1. Heterogeneous gossip

Participants: Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec.

This work has been done in collaboration with Vivien Quéma (CNRS Grenoble), Maxime Monod and Rachid Guerraoui (EPFL). Gossip-based information dissemination protocols are considered easy to deploy, scalable and resilient to network dynamics. Load-balancing is inherent in these protocols as the dissemination work is evenly spread among all nodes. Yet, large-scale distributed systems are usually heterogeneous with respect to network capabilities such as bandwidth. In practice, a blind load-balancing strategy might significantly hamper the performance of the gossip dissemination.

These observations have led to the development of HEAP, *HEterogeneity-Aware gossip Protocol*, where nodes dynamically adapt their contribution to the gossip dissemination according to their bandwidth capabilities. Using a continuous, itself gossip-based, approximation of relative bandwidth capabilities, HEAP dynamically leverages the most capable nodes by increasing their fanout, while decreasing by the same proportion those of less capable nodes. HEAP preserves the simplicity and proactive (churn adaptation) nature of gossip, while significantly improving its effectiveness. We extensively evaluated HEAP in the context of a video streaming application on a 270 PlanetLab node testbed. Our results show that HEAP significantly improves the perceived quality of the streaming over standard gossip protocols, especially when the stream rate is close to the average available bandwidth. During the last year, we have integrated HEAP with novel retransmission and FEC encoding mechanisms [47] as well as introduced the ability to measure bandwidth variations dynamically.

6.4.2. LIFTING

Participants: Kévin Huguenin, Anne-Marie Kermarrec.

This work has been done in collaboration with Maxime Monod and Rachid Guerraoui (EPFL). LiFTinG is the first protocol to detect freeriders, including colluding ones, in gossip-based content dissemination systems with asymmetric data exchanges. LiFTinG relies on nodes tracking abnormal behaviors by cross-checking the history of their previous interactions, and exploits the fact that nodes pick neighbors at random to prevent colluding nodes from covering each other up. We present extensive analytical evaluations of LiFTinG, backed up by simulations and PlanetLab experiments. In a 300-node system, where a stream of 674 kbps is broadcast, LiFTinG incurs a maximum overhead of only 8%.With 10% of freeriders decreasing their contribution by 30%, LiFTinG detects 86% of the freeriders after only 30 seconds and wrongfully expels only a few honest nodes. This work was published at Middleware 2010 [49].

6.4.3. Codes in Distributed Systems

Participants: Kévin Huguenin, Anne-Marie Kermarrec, Nicolas Le Scouarnec.

This project has been carried out in collaboration with Mary-Luc Champel (Technicolor) and Gilles Straub (Technicolor). Coding has been widely applied in distributed systems for it provides improvement in dissemination performance and storage resilience. We explore ways to make such codes more appealing by reducing side costs that may limit their uses in distributed systems.

In the context of dissemination, where network codes have been shown to provide optimal throughput, their current forms suffer from a high decoding complexity. This is an issue when applied to systems composed of nodes with low processing capabilities. We propose a novel network coding approach based on LT codes [106], initially introduced in the context of erasure coding. Our coding scheme, called LTNC [44], fully benefits from the low complexity of belief propagation decoding. Yet, such decoding schemes are extremely sensitive to statistical properties of the code. Maintaining such properties in a fully decentralized way with only a subset of encoded data is challenging. This is precisely what the recoding algorithms of LTNC achieve. We evaluate LTNC against random linear network codes in an epidemic content-dissemination application. Results show that LTNC slightly increases communication overhead (20%) and convergence time (30%) but greatly reduces the decoding complexity (99%) when compared to random linear network codes. This work received the best paper award at ICDCS 2010 [44].

In the context of storage, the use of erasure correcting codes greatly increases the repair cost (i.e., the amount of data that must be downloaded to reconstruct the redundancy lost during a crash). It has been shown that the use of network codes-based techniques allows reducing the repair cost [102]. Yet, the existing works are limited to single failures and static systems. We extend these works by defining Coordinated Regenerating Codes and Adaptive Regenerating Codes [85] that allow optimal repairs in presence of multiple failures and that allow the system to self-adapt to the condition so as to always perform repairs optimally. We show that the repair method we propose is optimal and permanently outperforms existing methods.

6.4.4. *Incentive-compatible peer-to-peer Video-on-Demand*

Participants: Kévin Huguenin, Anne-Marie Kermarrec.

This work has been done in collaboration with Maarten Van Steen and Vivek Rai (Vrije Universiteit Amsterdam) in the context of the EPINET associate team. Video-on-demand (VoD) is increasingly attractive Internet application allowing a client to begins to play a movie stored on a server almost instantaneously. The missing pieces of the video are downloaded in a sequential order while playing the downloaded pieces. Peer-to-peer swarming has proved to be a very efficient paradigm for file distribution even with a very low proportion of seeds provided that all the peers actually contribute to the system by uploading pieces to others. This problem is critical in VoD applications since peers must achieve their deadlines, and thus rely on the other peers, to provide a smooth playback to the user. The incentives used by BitTorrent, referred as Tit for tat are not usable in VoD downloads due to the sequential nature of VoD downloads and since any two peers may not have mutual interest. Our work consist in building an efficient Incentive-compatible peer-to-peer Video-on-Demand system by imposing a loose structure on the peer sets : grouping nodes whose positions in the video are close in groups and build a linked list of groups to ensure feeding groups by more advanced ones in exchange of most advanced pieces injected in less advanced groups. Using an appropriate pieces seeding/feeding strategy the full swarm comes down to several very efficient independent swarms fed by each others, the seed being appealed only by the most advanced group. The structured VoD swarming protocol outperforms existing protocols based on random graphs in terms of throughput, goodput (i.e., playback rate) and memory usage (by keeping a limited memory of past pieces instead of storing the entire video). This work was published in the proceedings of NOSDAV 2010 [50].

6.4.5. *Localization and efficient routing in large scale sensor networks*

Participants: Guang Tan, Anne-Marie Kermarrec.

We studied the localization problem for large-scale sensor network in a very general setting: no means for physical distance measurement, no anchor nodes with known positions, 2D/3D, no knowledge of network boundaries, possible holes in the network [66]. The outcome is a scheme named CATL that is able to assign each node a set of (relative) coordinates which jointly form a network layout resembling the original. This will

help nodes recognize their relative positions to each other which benefits many sensor network applications. We also investigated efficient geographic routing in sensor networks by a systematically study of the conditions of greedy forwarding and how to decompose networks to facilitate routing [57]. Theoretical properties of the the problem are explored and efficient algorithms are designed and validated with simulation.

6.4.6. Deterministic Recurrent Communication and Synchronization in Restricted Sensor Networks

Participant: Christopher Thraves-Caro.

This work has been done in collaboration with Antonio Fernández Anta (Institute IMDEA Networks, Madrid) and Miguel A. Mosteiro (Department of Computer Science Rutgers, The State University of New Jersey).

Monitoring physical phenomena in Sensor Networks requires guaranteeing permanent communication between nodes. Moreover, in an effective implementation of such infrastructure, the delay between any two consecutive communications should be minimized. The problem is challenging because, in a restricted Sensor Network, the communication is carried out through a single and shared radio channel without collision detection. Dealing with collisions is crucial to ensure effective communication between nodes. Additionally, minimizing them yields energy consumption minimization, given that sensing and computational costs in terms of energy are negligible with respect to radio communication. In this work, the authors present a deterministic recurrent-communication protocol for Sensor Networks. After an initial negotiation phase of the access pattern to the channel, each node running this protocol reaches a steady state, which is asymptotically optimal in terms of energy and time efficiency. As a by-product, a protocol for the synchronization of a Sensor Network is also proposed. Furthermore, the protocols are resilient to an arbitrary node power-up schedule and a general node failure model. This work was published at ALGOSENSOR 2010.

6.4.7. Online Task Allocation in Client-Server Large Scale Heterogeneous Platforms

Participant: Christopher Thraves-Caro.

This work has been done in collaboration with Olivier Beaumont, Lionel Eyraud-Dubois and Hejer Rejeb from INRIA Bordeaux – Sud-Ouest, University of Bordeaux, LaBRI.

In this research, the authors consider the problem of the online allocation of a very large number of identical tasks on a master-slave platform. Initially, several masters hold or generate tasks that are transfered and processed by slave nodes. The goal is to maximize the overall throughput achieved using this platform, i.e., the (fractional) number of tasks that can be processed within one time unit. The authors model the communications using the so-called bounded degree multi-port model, in which several communications can be handled by a master node simultaneously, provided that bandwidths limitation are not exceeded and that a given server is not involved in more simultaneous communications than its maximal degree. Under this model, it has been proved that maximizing the throughput (MTBD problem) is NP-Complete in the strong sense but that a small additive resource augmentation (of 1) on the servers degrees is enough to find in polynomial time a solution that achieves at least the optimal throughput. In this research, the authors consider the reasonable setting where the set of slave processors is not known in advance but rather join and leave the system at any time, i.e., the online version of MTBD. The authors prove that no fully online algorithm (where only one change is allowed for each event) can achieve a constant approximation ratio, whatever the resource augmentation on servers degrees. Then, The authors prove that it is possible to maintain the optimal solution at the cost of at most four changes per server each time a new node joins or leaves the system. At last, the authors propose several other greedy heuristics to solve the online problem and compare the performance (in terms of throughput) and the cost (in terms of disconnections and reconnections) of proposed algorithms through a set of extensive simulation results. Results from this work were published at PDP 2010 [29].

6.4.8. Statistically Anonymous Sources in Wireless Sensor Networks

Participants: Silvija Kokalj-Filipovic, Fabrice Le Fessant.

In this work, we proposed a scheme for generating fake network traffic in order to disguise the real source in the presence of a global eavesdropper which is especially relevant for large WSNs with single data collector, and for delay-intolerant monitoring applications. The scheme provides statistical source anonymity in an efficient way. Under the dummy-traffic framework of the source anonymity protection, we also propose a metric for its goodness that includes the work of the adversary in terms of the number of samples required for statistical hypothesis tests, used to discover the source location with given probability. By including the adversary's work, we aim to better model a global eavesdropper, and to present the quality of the location protection strategy as relative to the adversary's strength. In addition, the goodness metric includes the statistically guaranteed anonymity level, the work spent to generate the fake traffic, and the latency guarantees by the proposed algorithm. This is a work in progress with the goal to formalize the source anonymity as an optimization problem that can offer different solutions for different efficiency criteria.

6.4.9. Building secured links in sensor networks

Participants: Marin Bertier, Achour Mostefaoui.

This work [34], done in collaboration with Gilles Trédan from TU Berlin, deals with malicious behaviors in the context of sensor networks. Such a behavior can be due to an adversary that has some sensors under control or more generally to a problem of the sensor itself. Effectively, as sensors are small devices that are industrially built, many of them may be defective. Moreover, it is known that when a sensor is running out of energy, it can enter a state where it behaves abnormally. Malicious behaviors in sensor networks are less hard to handle as the power of the adversary is lower. Indeed a sensor has a limited energy. The more it is active the less it will survive and thus even its computation power is bounded. In the case of a sensor network with static sensors, we try to build secured links between sensors. The objective is to avoid the case of an adversary that collects the whole information exchanged among the sensors.

7. Contracts and Grants with Industry

7.1. Technicolor

Participants: Anne-Marie Kermarrec, Nicolas Le Scouarnec, Alexandre Van Kempen.

Since November 2007, we have collaborated with Technicolor R&D France to study coding for resource optimization, as well as storage in large-scale distributed systems. In this context, Anne-Marie Kermarrec was the PhD adviser of Nicolas Le Scouarnec since November 2007 and of Alexandre van Kempen since 2010.

7.2. National grants

7.2.1. ANR VERSO project Shaman

Participants: Marin Bertier, Achour Mostefaoui, Anne-Marie Kermarrec, Michel Raynal, Christopher Thraves-Caro.

The Shaman project started in 2009, gathering several members of the team working on distributed systems and distributed algorithms. The aim of this project is to propose new theoretical models for distributed algorithm inspired from real platform characteristics. From these models, we elaborate new algorithms and try to evaluate their theoretical power.

7.2.2. ANR ARPÈGE project Streams

Participants: Achour Mostefaoui, Marin Bertier, Michel Raynal.

The Streams project started in November 2010. Beside the ASAP group, it includes teams from INRIA Nancy and PARIS. Its aim it to design a real-time collaborative platform based on a peer-to-peer network. For this it is necessary to design a support architecture that offers garantees on the propagation, security and consistency of the operations and the updates proposed by the differents collaborating sites.

7.2.3. Rnrt Project SensLab

Participants: Marin Bertier, Nicolas Destor, Antoine Boutet, Anne-Marie Kermarrec.

SensLab is an RNRT project started in 2008 focusing on the deployment of a very large-scale open wireless sensor network platform to be used as an efficient scientific tool for designing, tuning, and experimenting real sensor-based applications. A SensLAB platform composed of 1024 nodes is deployed among 4 sites. This infrastructure represents the unique scientific tool for the research on wireless sensor networks.

7.2.4. ADT Project SensTools

Participants: Marin Bertier, Antoine Boutet, Anne-Marie Kermarrec.

SensTools is an ADT project supported by INRIA. SensTools provides a set of hardware and software tools for the WSN430 platform used within the Senslab project. Some basic drivers and several OSes are provided (FreeRTOS, TinyOS, Contiki).

8. Other Grants and Activities

8.1. International grants

8.1.1. GOSSPLE ERC Starting Grant

Participants: Mohammad Nabil Al-Aggan, Xiao Bai, Marin Bertier, Antoine Boutet, Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec, Konstantinos Kloudas, Vincent Leroy, Afshin Moin, Guang Tan.

Anne-Marie Kermarrec is the principal investigator of the GOSSPLE ERC starting Grant (Sept. 2008 - Sept. 2013). GOSSPLE aims at providing a radically new approach to navigating the digital information universe. This project has been granted a 1.250.000 euros budget for 5 years.

GOSSPLE aims at radically changing the navigation on the Internet by placing users affinities and preferences at the heart of the search process. Complementing traditional search engines, GOSSPLE will turn search requests into live data to seek the information where it ultimately is: at the user. GOSSPLE precisely aims at providing a fully decentralized system, auto-organizing, able to discover, capture and leverage the affinities between users and data.

8.1.2. Transform Marie Curie Initial Training Network

Participants: Tyler Crain, Anne-Marie Kermarrec, Achour Mostefaoui, Michel Raynal.

Transform is a Marie Curie Initial Training Networks European project devoted to the Theoretical Foundations of Transactional Memory (Grant agreement no.: 238639 Date of approval of Annex I by Commission: May 26, 2009). It involves the following universities : Foundation for Research and Technology Hellas ICS FORTH Greece, University of Rennes 1 UR1 France, Ecole Polytechnique Federale de Lausanne EPFL Switzerland, Technische Universitaet Berlin TUB Germany, and Israel Institute of Technology Technion.

Major chip manufacturers have shifted their focus from trying to speed up individual processors into putting several processors on the same chip. They are now talking about potentially doubling efficiency on a 2x core, quadrupling on a 4x core and so forth. Yet multi-core is useless without concurrent programming. The constructors are now calling for a new software revolution: the concurrency revolution. This might look at first glance surprising for concurrency is almost as old as computing and tons of concurrent programming models and languages were invented. In fact, what the revolution is about is way more than concurrency alone: it is about concurrency for the masses. The current parallel programming approach of employing locks is widely considered to be too difficult for any but a few experts. Therefore, a new paradigm of concurrent programming is needed to take advantage of the new regime of multicore computers. Transactional Memory (TM) is a new programming paradigm which is considered by most researchers as the future of parallel programming. Not surprisingly, a lot of work is being devoted to the implementation of TM systems, in hardware or solely in software. What might be surprising is the little effort devoted so far to devising a sound theoretical framework

to reason about the TM abstraction. To understand properly TM systems, as well as be able to assess them and improve them, a rigorous theoretical study of the approach, its challenges and its benefits is badly needed. This is the challenging research goal undertaken by this MC-ITN. Our goal through this project is to gather leading researchers in the field of concurrent computing over Europe, and combine our efforts in order to define what might become the modern theory of concurrent computing. We aim at training a set of Early Stage Researchers (ESRs) in this direction and hope that, in turn, these ESRs will help Europe become a leader in concurrent computing. Its keywords are Transactional Memory, Parallelization Mechanisms, Parallel Programming Abstractions, Theory, Algorithms, Technological Sciences

8.1.3. Demdyn: INRIA/CNPq Collaboration

Participants: Achour Mostefaoui, Marin Bertier, Michel Raynal.

The aim of this project is to exploit dependable aspects of dynamic distributed sys- tems such as VANETs, WiMax, Airborn Networks, DoD Global Information Grid, P2P, etc. Applications that run on these kind of networks have a common point: they are extremely dynamic both in terms of the nodes that take part of them and available resources at a given time. Such dynamics results in instability and uncer- tainty of the environment which provide great challenges for the implementation of dependable mechanisms that ensure the correct work of the system.

This requires applications to be adaptive, for instance, to less network bandwidth or degraded Quality-of-Service (QoS). Ideally, in these highly dy- namic scenarios, adaptiveness characteristics of applications should be self-managing or autonomic. Therefore, being able to detect the occurrence of partitions and automatically adapting the applications for such scenarios is an important dependable requirement for such new dynamic environments.

8.2. Visits (2010)

Rachid Guerraoui, EPFL Lausanne, Switzerland, Several Visits in 2010 (Rennes). **Yann Gripay**, LIRIS Lyon, April 2010 (Rennes). **Roberto Cascella**, INRIA Sophia-Antipolis, April 2010 (Rennes). **Eric Ruppert**, York University, Canada, April 2010 (Rennes). **Hamouma Moumen**, Universtié de Béjaia, May 2010 (Rennes). **Eli Gafni**, UCLA, USA, June and December 2010 (Rennes). **Marc Shapiro**, LIP6, France, December 2010 (Rennes). **Ajoy Datta**, University of Las Vegas, USA, December 2010 (Rennes). **Robert Elsässer**, University of Paderborn, Germany, December 2010 (Rennes). **Carole Delporte**, Université Paris Diderot-Paris 7, France, April, and December 2010 (Rennes). **Hugues Fauconnier**, **Carole Delporte**, Université Paris Diderot-Paris 7, France, April, and December 2010 (Rennes).

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Leaderships and community service

A.-M. Kermarrec is a nominated member of the ACM Software System Award Committee since October 2009.

9.1.2. Editorial boards, steering and program committees

M. Bertier served in the program committees of the following conferences:

DCOSS 2010 *The 6th IEEE International Conference on Distributed Computing in Sensor Systems*, June 2010, Santa Babara, US

DEBS 2010 *The 4th ACM International Conference on Distributed Event-Based Systems*, July 2010, Cambridge, United Kingdom

Algotel 2010 *12mes Rencontres Francophones sur les Aspects Algorithmiques de Tlcommunications*, Mai 2010 Belle Dune - Cote d'Opale, France

A. Boutet served in the program committee of TridentCom 2010.

D. Frey served in the program committees for the following conferences:

EDCC'10: Eighth European Dependable Computing Conference. Valencia, Spain. April 28-30, 2010.

SSS'10: 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems, New York, USA, September 2010.

Davide Frey and **A.-M. Kermarrec** organized the Workshop on Social Networks and Distributed Systems (SNDS 2010), colocated with PODC 2010, Zurich Switzerland, July 2010.

A.-M. Kermarrec also organized the first GOSSPLE workshop on Social Networks.

A.-M. Kermarrec served in the steering committee of the Eurosys Social Network Systems (SNS), and is a member of the steering committee of the Winter School Hot topics in distributed computing.

A.-M. Kermarrec served in the program committees for the following conferences:

EuroSys 2010: *European Conference on Computer Systems*, Paris, France, April 2010.

Middleware 2010: *ACM/IFIP/USENIX International Middleware Conference*, Bangalore, India, December 2010.

Dial-POMC 2010: *International Workshop on FOUNDATIONS OF MOBILE COMPUTING*, Cambridge, Massachusetts, Usa, September 2010.

P2P 2010: *IEEE Conference on Peer-to-Peer systems*, Delft, Netherlands, August 2010.

DISC 2011: *International Symposium on Distributed Computing*, Rome, Italy, September 2011.

Eurosys SNS 2011: *ACM Workshop on Social Network Systems*, Salzburg, Austria, April 2011.

EuroSys Doctoral Workshop 2011: at the *European Conference on Computer Systems*, Salzburg, Austria, April 2011.

EDBT 2011 (Demo): International Conference on Extending Database Technology (Demo track), Uppsala, Sweden, March 2011

A. Mostefaoui served in the program committees for the following conferences:

SSS'10: 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems, New York, USA, September 2010.

IWSN/DCOSS 2011: International Workshop on Interconnections of Wireless Sensor Networks, in conjunction with DCOSS'11, Barcelona, Spain, June 2011.

M. Raynal was conference chair of the 11th International Conference on Distributed Computing and Networking (ICDCN'10), january 2010.

He was worskshop co-Chair of the 29th IEEE International conference on Distributed Computing Systems (ICDCS'10), june 2010.

He served in the program committees for the following conferences.

- 12th IEEE International Conference on High Performance Computing and Communications (HPCC'10), Melbourne, September 2010.
- 3rd International Workshop on Reliability, Availability, and Security (WRAS'10), in conjunction with PODC, July 2010 Zurich, Switzerland.
- 24th International Symposium on Distributed Computing (DISC'10), MIT, Boston, September 2010.
- 14th International Conference on Principles of Distributed Systems (OPODIS'10), 2010.

He served also in the editorial board of the following journals:

JPDC Journal of Distributed and Parallel Computing (since 2005)

IEEE TPDS IEEE Transactions on Parallela nd Distriubted Systems (since 2006)

FCDS Foundations of Computing and Decision Sciences (since 1995)

IJCSSE International Journal of Computer Systems Science and Engineering (since 1998)

He served also in the steering committees of the following conferences:

ICDCS "International Colloquium on Structural Information and Communication Complexity"

ICDCN "International Conference on Distributed Computing and Networking"

9.2. Administrative responsibilities

C. Bouton is an elected member of the "comité de centre".

A.-M. Kermarrec is an elected member of the INRIA Evaluation Committee since September 2005.

She was head of the 2010 INRIA Selection Committees for for Junior Researcher permanent positions (CR2) (INRIA Lille)

She has been a member of the "bureau du CP" since November 2009.

She took part in a public debate "regards croisés, les champs libres: Internet et Democratie", December 2010.

Antoine Boutet has been a member of the "Commission de formation" of INRIA Rennes since 2010.

Fabrice Le Fessant is a member of the Project Committee for Free Software of the Pôle de Compétitivité System@tic, Paris.

9.3. Academic teaching

There is a strong teaching activity in the ASAP project team as three of the permanent members are Professor or Assistant Professor.

M. Bertier is responsible of the 5th year of the Engineer school INSA Rennes and responsible of a Master's course entitled "Operating System"(INSA)

A.-M. Kermarrec and **M. Raynal** are each responsible of a Master's courses (University of Rennes 1 and ENS Cachan, Brittany extension) entitled respectively "peer-to-peer systems and applications (PAP)" and "Foundations of Distributed Systems". The teaching in the PAP module is shared with Gabriel Antoniu from the KERDATA project-team.

A.-M. Kermarrec gave a 10 hour lecture at University of Madrid on gossip protocols in March 2009.

F. Le Fessant is a half-time associate professor at Ecole Polytechnique.

A. Mostefaoui ensures part of one of the three basic courses of the Master in Computer Science. He is responsible of a Master's course (University of Bougie, Algeria) entitled "Distributed Algorithms and Systems".

In addition, **A. Boutet** and **N. Le Scouarnec** are teaching respectively at University of Rennes 1 and INSA, while **F. Bonnet, K. Huguenin, V. Leroy,** and **D. Imbs** were Teaching Assistants (*moniteurs*) (INSA, University of Rennes 1, ENS Cachan).

9.4. Conferences, seminars, and invitations

9.4.1. Invited Talks

A.-M. Kermarrec was a keynote speaker at P2P 2010.

A.-M. Kermarrec was an invited speaker at the Royal Society in September 2010 (celebrating 350 years), in the “Web science: A new frontier” seminar.

A.-M. Kermarrec and **M. Raynal** were invited speakers at the Collège de France in 2010.

D. Frey was an invited speaker at the Summer School, Masses de donnés distribués, Les Houches (France), May 2010.

9.4.2. Seminars

A.-M. Kermarrec was invited to a debate on “Internet and democracy” at Les champs librés Rennes, December 2010.

A.-M. Kermarrec was invited to give a talk at the University of Neuchatel, Switzerland, December 2010.

A.-M. Kermarrec was invited to give a talk at LINA (University of Nantes) in June 2010.

Marin Bertier was invited to give a talk at "Journée PucesCom", Rennes, September 2010.

Vincent Leroy, **Kévin Huguenin**, and **Nicolas Le Scouarnec** gave invited talks at Technicolor R&D, Rennes, France in June 2010.

Vincent Leroy was invited to give a talk at Yahoo! New York, USA, in July 2010.

Vincent Leroy and **Kévin Huguenin** gave invited talks at Yahoo! India, in December 2010.

Kévin Huguenin gave an invited talk at McGill University, Montreal, Canada, in September 2010.

M. Raynal was invited to give a talk at Singapore University (march 2010).

M. Raynal was invited to give a talk at La Sapienza (Roma) May 2010.

M. Raynal was invited to give a talk at UNAM (MX), August 2010.

M. Raynal was invited to give a talk at the Polytechnic University of Hong-Kong (march 2010).

9.4.3. Visits

X. Bai spent three months (June-Sept 2010) at Yahoo! Research, Barcelona, Spain.

10. Bibliography

Major publications by the team in recent years

[1] J. CAO, M. RAYNAL, X. YANG, W. WU. *Design and Performance Evaluation of Efficient Consensus Protocols for Mobile Ad Hoc Networks*, in "IEEE Transactions on Computers", 2007, vol. 56, n^{o} 8, p. 1055–1070.

[2] A. CARNEIRO VIANA, S. MAAG, F. ZAIDI. *One step forward: Linking Wireless Self-Organising Networks Validation Techniques with Formal Testing approaches*, in "ACM Computing Surveys", 2009, http://hal.inria.fr/inria-00429444/en/.

[3] D. FREY, R. GUERRAOUI, A.-M. KERMARREC, M. MONOD, K. BORIS, M. MARTIN, V. QUÉMA. *Heterogeneous Gossip*, in "Middleware 2009", Urbana-Champaign, IL, USA, 2009, http://hal.inria.fr/inria-00436125/en/.

[4] R. FRIEDMAN, A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL. *Distributed agreement problems and their connection with error-correcting codes*, in "IEEE Transactions on Computers", 2007, vol. 56, n^{o} 7, p. 865–875.

[5] A. J. GANESH, A.-M. KERMARREC, E. LE MERRER, L. MASSOULIÉ. *Peer counting and sampling in overlay networks based on random walks*, in "Distributed Computing", 2007, vol. 20, n° 4, p. 267-278.

[6] M. JELASITY, S. VOULGARIS, R. GUERRAOUI, A.-M. KERMARREC, M. VAN STEEN. *Gossip-Based Peer Sampling.*, in "ACM Transactions on Computer Systems", August 2007, vol. 41, n° 5.

[7] B. MANIYMARAN, M. BERTIER, A.-M. KERMARREC. *Build One, Get One Free: Leveraging the Coexistence of Multiple P2P Overlay Networks.*, in "Proceedings of ICDCS 2007", Toronto, Canada, June 2007.

[8] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *From Diamond W to Omega: a simple bounded quiescent reliable broadcast-based transformation*, in "Journal of Parallel and Distributed Computing", 2007, vol. 61, n° 1, p. 125–129.

[9] J. PATEL, É. RIVIÈRE, I. GUPTA, A.-M. KERMARREC. *Rappel: Exploiting interest and network locality to improve fairness in publish-subscribe systems.*, in "Computer Networks", 2009, vol. 53, n° 13, http://hal.inria.fr/inria-00436057/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[10] X. BAI. *Traitement personnalisé de requête top-k: des systèmes centralisés aux systèmes décentralisés*, INSA de Rennes, December 2010, http://hal.inria.fr/tel-00545642/en.

[11] F. BONNET. *Cohérence de calculs répartis face aux défaillances, à l'anonymat et au facteur d'échelle*, Université Rennes 1, July 2010, http://hal.inria.fr/tel-00549364/en.

[12] K. HUGUENIN. *Méthodes dissuasives contres les utilisateurs malhonnêtes dans les systèmes répartis*, Université Rennes 1, December 2010, http://hal.inria.fr/tel-00545663/en.

[13] N. LE SCOUARNEC. *Codage pour l'optimisation de ressources dans les systèmes distribués*, INSA de Rennes, December 2010, http://hal.inria.fr/tel-00545641/en.

[14] V. LEROY. *Distributing Social Applications*, INSA de Rennes, December 2010, http://hal.inria.fr/tel-00545639/en.

Articles in International Peer-Reviewed Journal

[15] Y. AFEK, E. GAFNI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *The k-simultaneous consensus problem*, in "Distributed Computing", 2010, vol. 22, n° 3, p. 185-196, http://hal.inria.fr/inria-00543100/en.

[16] F. BONNET, M. RAYNAL. *A simple proof of the necessity of the failure detector Sigma to implement an atomic register in asynchronous message-passing systems*, in "Information Processing Letters", 2010, vol. 110, n° 4, http://hal.inria.fr/inria-00543125/en.

[17] Y. BUSNEL, L. QUERZONI, R. BALDONI, M. BERTIER, A.-M. KERMARREC. *Analysis of Deterministic Tracking of Multiple Objects using a Binary Sensor Network*, in "ACM Transactions on Sensor Networks", 2011, http://hal.inria.fr/inria-00544467/en.

[18] V. CHOLVI, A. FERNÁNDEZ, E. JIMÉNEZ, P. MANZANO, M. RAYNAL. *A Methodological Construction of an Efficient Sequentially Consistent Distributed Shared Memory*, in "The Computer Journal", 2010, p. 1523-1534, http://hal.inria.fr/inria-00543094/en.

[19] A. FERNÁNDEZ, E. JIMÉNEZ, P. MANZANO, M. RAYNAL. *Eventual Leader election with weak assumptions on initial knowledge, communication reliability and synchrony.*, in "Springer-Verlag Journal of Computer Science and Technology (JCST)", 2010, vol. 25, n^{o} 6, p. 1267-1281, http://hal.inria.fr/inria-00543126/en.

[20] A. FERNÁNDEZ, E. JIMÉNEZ, M. RAYNAL, G. TREDAN. *A timing assumption and two t-resilient protocols for implementing an eventual leader service in asynchronous shared-memory systems*, in "Algorithmica", 2010, http://hal.inria.fr/inria-00543053/en.

[21] A. FERNÁNDEZ, M. RAYNAL. *From an Asynchronous Intermittent Rotating Star to an Eventual Leader*, in "IEEE Transactions on Parallel and Distributed Systems", 2010, p. 1290-1303, http://hal.inria.fr/inria-00543096/en.

[22] D. IMBS, M. RAYNAL. *Software Transactional Memories: an Approach for Multicore Programming*, in "Springer Journal of Supercomputing", 2010, http://hal.inria.fr/inria-00543127/en.

[23] A.-M. KERMARREC, E. LE MERRER, B. SERICOLA, G. TREDAN. *Second order centrality: Distributed assessment of nodes criticity in complex networks*, in "Computer Communications", 2010, http://dx.doi.org/10.1016/j.comcom.2010.06.007, http://hal.inria.fr/inria-00506385/en.

[24] A. MOSTEFAOUI, M. RAYNAL, C. TRAVERS. *Narrowing power vs efficiency in synchronous set agreement: Relationship, algorithms and lower bound*, in "Theoretical Computer Science", January 2010, vol. 411, n^{o} 1, p. 58-69, http://hal.inria.fr/hal-00543282/en.

[25] M. RAYNAL, C. TRAVERS, P. RAÏPIN-PARVÉDY. *Strongly Terminating Early-Stopping k-set Agreement in Synchronous Systems with General Omission Failures*, in "Theory of Computing Systems", 2010, http://hal.inria.fr/inria-00543051/en.

[26] M. VECCHIO, A. CARNEIRO VIANA, A. ZIVIANI, R. FRIEDMAN. *Deep: Density-based Proactive Data Dissemination Protocol for Wireless Sensor Networks with Uncontrolled Sink Mobility*, in "Computer Communications", November 2010, http://hal.inria.fr/inria-00455651/en.

International Peer-Reviewed Conference/Proceedings

[27] X. BAI, M. BERTIER, R. GUERRAOUI, A.-M. KERMARREC, V. LEROY. *Gossiping Personalized Queries*, in "13th International Conference on Extending Database Technology", Switzerland Lausanne, March 2010, http://hal.inria.fr/inria-00455643/en.

[28] O. BALDELLON, A. MOSTEFAOUI, M. RAYNAL. *A Symmetric Synchrony Condition for Solving Byzantine Consensus*, in "12th International Conference on Distributed Computing and Networking (ICDCN 2011)", India Bangalore, LNCS, Springer Verlag, January 2011, vol. 6522, p. 215–226, http://hal.inria.fr/inria-00544666/en.

[29] O. BEAUMONT, L. EYRAUD-DUBOIS, H. REJEB, C. THRAVES. *On-line Allocation of Clients to Multiple Servers on Large Scale Heterogeneous Systems*, in "PDP 2010 - The 18th Euromicro International Conference

on Parallel, Distributed and Network-Based Computing", Italy Pisa, February 2010, http://hal.inria.fr/inria-00444584/en.

[30] M. BERTIER, F. BONNET, A.-M. KERMARREC, V. LEROY, S. PERI, M. RAYNAL. *D2HT: the best of both worlds, Integrating RPS and DHT*, in "European Dependable Computing Conference", Spain Valencia, April 2010, http://hal.inria.fr/inria-00459944/en.

[31] M. BERTIER, Y. BUSNEL, A.-M. KERMARREC. *Dynamic Computation of Population Protocols*, in "the 17th IEEE International Conference on Telecommunications - Ad-hoc and Sensor Communications track (ICT 2010)", Qatar Doha, April 2010, p. 100-107, http://hal.inria.fr/hal-00480988/en.

[32] M. BERTIER, D. FREY, R. GUERRAOUI, A.-M. KERMARREC, V. LEROY. *The Gossple Anonymous Social Network*, in "ACM/IFIP/USENIX 11th International Middleware Conference", India Bangalore, November 2010, http://hal.inria.fr/inria-00515693/en.

[33] M. BERTIER, A.-M. KERMARREC, G. TAN. *Message-Efficient Byzantine Fault-Tolerant Broadcast in a Multi-Hop Wireless Sensor Network*, in "The 30th International Conference on Distributed Computing Systems", Italy Genoa, June 2010, http://hal.inria.fr/inria-00457215/en.

[34] M. BERTIER, A. MOSTEFAOUI, G. TREDAN. *Low-Cost Secret-Sharing in Sensor Networks*, in "12th IEEE International High Assurance Systems Engineering Symposium (HASE 2010)", United States San Jose, CA, S. SEDIGH (editor), IEEE, November 2010, p. 1-9, http://hal.inria.fr/inria-00544585/en.

[35] F. BONNET, M. RAYNAL. *Anonymous Asynchronous Systems: the Case of Failure Detectors*, in "Proc. 24th Int'l Symposium on Distributed Computing (DISC'10)", United States Cambridge (MA), Springer-Verlag LNCS, 2010, p. 206-220, Best Student-paper award, http://hal.inria.fr/inria-00543857/en.

[36] F. BONNET, M. RAYNAL. *Consensus in Anonymous Distributed Systems: Is There a Weakest Failure Detector?*, in "24th IEEE International Conference on Advanced Information Networking and Applications (AINA'10)", Australia Perth, IEEE Computer Society Press, 2010, p. 206-213, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5474697, http://hal.inria.fr/inria-00426674_v1.

[37] F. BONNET, M. RAYNAL. *Consensus in Anonymous Distributed Systems: Is There aWeakest Failure Detector?*, in "The 24th International Conference on Advanced Information Networking and Applications (AINA 2010)", Australia Perth, April 2010, http://hal.inria.fr/inria-00473001/en.

[38] F. BONNET, M. RAYNAL. *Early Consensus in Message-passing Systems Enriched with a Perfect Failure Detector and its Application in the Theta Model*, in "The 8th European Dependable Computing Conference (EDCC 2010)", Spain Valence, April 2010, http://hal.inria.fr/inria-00473002/en.

[39] F. BONNET, M. RAYNAL. *Early Consensus in Message-passing Systems Enriched with a Perfect Failure Detector and its Application in the Theta Model*, in "8th European Dependable Computing Conference (EDCC'10)", Spain Valencia, IEEE Computer Society Press, 2010, p. 107-116, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5474189, http://hal.inria.fr/inria-00473002.

[40] S. BONOMI, R. BALDONI, M. RAYNAL. *Value-based Sequential Consistency for Set Objects in Dynamic Distributed Systems*, in "Proc. 16th Int'l European Parallel Computing Conference (EUROPAR'10)", Italy Ischia - Naples, 2010, p. 523-534, http://hal.inria.fr/inria-00543132/en.

[41] A. BOUTET, D. FREY, R. GUERRAOUI, A.-M. KERMARREC. *WhatsUp: news from, for, through everyone*, in "10th IEEE International Conference on Peer-to-Peer Computing (IEEE P2P'10)", Netherlands Delft, August 2010, http://hal.inria.fr/inria-00515420/en.

[42] A. CARNEIRO VIANA, T. HERAULT, T. LARGILLIER, S. PEYRONNET, F. ZAIDI. *Supple: A Flexible Probabilistic Data Dissemination Protocol for Wireless Sensor Networks*, in "ACM MSWIM", Turkey Bodrum, October 2010, http://hal.inria.fr/inria-00544749/en.

[43] A. CARNEIRO VIANA, N. MITTON, L. SCHMIDT, M. VECCHIO. *A k-layer self-organizing structure for product management in stock-based networks*, in "IEEE ICEBE", China Shanghai, November 2010, http://hal.inria.fr/inria-00544752/en.

[44] M.-L. CHAMPEL, K. HUGUENIN, A.-M. KERMARREC, N. LE SCOUARNEC. *LT Network Codes*, in "30th International Conference on Distributed Computing Systems (ICDCS)", Italy Genoa, 2010, http://hal.inria.fr/inria-00455639/en.

[45] C. DELPORTE-GALLET, H. FAUCONNIER, R. GUERRAOUI, A.-M. KERMARREC, E. RUPPERT. *Byzantine Agreement with Homonyms*, in "PODC 2010 - brief announcement", Switzerland Zurich, 2010, http://hal.inria.fr/inria-00545914/en.

[46] A. FERNÁNDEZ ANTA, M. MOSTEIRO, C. THRAVES. *Deterministic Recurrent Communication and Synchronization in Restricted Sensor Networks*, in "6th International Workshop on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile Entities (ALGOSENSOR), 2010.", France Bordeaux, 2010, http://hal.inria.fr/inria-00543258/en.

[47] D. FREY, R. GUERRAOUI, A.-M. KERMARREC, M. MONOD. *Boosting Gossip for Live Streaming*, in "P2P 2010", Netherlands Delft, August 2010, http://hal.inria.fr/inria-00517384/en.

[48] A. GIURGIU, R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC. *Computing in Social Networks*, in "12th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)", United States New York, September 2010, http://hal.inria.fr/inria-00498132/en.

[49] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MONOD. *LiFTinG: Lightweight Freerider-Tracking Protocol in Gossip*, in "ACM/IFIP/USENIX 11th International Middleware Conference (MIDDLEWARE)", India Bangalore, November 2010, http://hal.inria.fr/inria-00505268/en.

[50] K. HUGUENIN, A.-M. KERMARREC, V. RAI, M. VAN STEEN. *Designing a Tit-for-Tat Based Peer-to-Peer Video-on-Demand System*, in "20th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)", Netherlands Amsterdam, 2010, http://hal.inria.fr/inria-00467786/en.

[51] D. IMBS, M. RAYNAL. *On Adaptive Renaming under Eventually Limited Contention*, in "Proc. 12th Int'l Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS'10)", United States New York, Springer-Verlag LNCS, 2010, p. 377-387, http://hal.inria.fr/inria-00543135/en.

[52] D. IMBS, M. RAYNAL. *The Multiplicative Power of Consensus Numbers.*, in "29th ACM Symposium on Principles of Distributed Computing (PODC'10)", Switzerland Zurich, ACM Press, 2010, p. 26-35, http://hal.inria.fr/inria-00543130/en.

[53] D. IMBS, M. RAYNAL. *The x-Wait-freedom Progress Condition*, in "Proc. 16th Int'l European Parallel Computing Conference (EUROPAR'10)", Italy Ischia - Naples, Springer-Verlag LNCS, 2010, p. 584-595, http://hal.inria.fr/inria-00543134/en.

[54] D. IMBS, M. RAYNAL, G. TAUBENFELD. *On Asymmetric Progress Conditions*, in "29th ACM Symposium on Principles of Distributed Computing (PODC'10)", Switzerland Zurich, ACM Press, 2010, p. 55-64, http://hal.inria.fr/inria-00543131/en.

[55] A.-M. KERMARREC, V. LEROY, A. MOIN, C. THRAVES. *Application of Random Walks to Decentralized Recommender Systems*, in "14th International Conference On Principles Of Distributed Systems", Tunisia Tozeur, 2010, http://hal.inria.fr/inria-00520214/en.

[56] A.-M. KERMARREC, V. LEROY, C. THRAVES. *Converging Quickly to Independent Uniform Random Topologies*, in "19th Euromicro International Conference on Parallel, Distributed and Network-Based Computing (PDP), 2011.", Cyprus Ayia Napa, February 2011, http://hal.inria.fr/inria-00543249/en.

[57] A.-M. KERMARREC, G. TAN. *Greedy Geographic Routing in Large-Scale Sensor Networks: A Minimum Network Decomposition Approach*, in "ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)", United States Chicago, September 2010, http://hal.inria.fr/inria-00493387/en.

[58] S. LE BLOND, A. LEGOUT, F. LE FESSANT, W. DABBOUS, M. A. KAAFAR. *Spying the World from your Laptop – Identifying and Profiling Content Providers and Big Downloaders in BitTorrent*, in "3rd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET'10)", United States San Jose, CA, Usenix, April 2010, http://hal.inria.fr/inria-00470324/en.

[59] V. LEROY, B. B. CAMBAZOGLU, F. BONCHI. *Cold Start Link Prediction*, in "The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining", United States Washington DC, July 2010, 12 p, http://hal.inria.fr/inria-00485619/en.

[60] S. LIN, F. TAÏANI, M. BERTIER, A.-M. KERMARREC. *Transparent Componentisation: High-level (Re)configurable Programming for Evolving Distributed Systems*, in "26th ACM Symposium on Applied Computing", Taiwan, Province Of China Taichung, March 2011, http://hal.inria.fr/inria-00544510/en.

[61] A. MOSTEFAOUI, M. RAYNAL. *Signature-Free Broadcast-Based Intrusion Tolerance*, in "14th International Conference On Principles Of Distributed Systems (OPODIS 2010)", Tunisia Tozeur, C. LU, T. MASUZAWA, M. MOSBAH (editors), LNCS, Springer Verlag, December 2010, vol. 6490, p. 143-159, http://hal.inria.fr/inria-00544650/en.

[62] H. MOUMEN, A. MOSTEFAOUI. *Time-Free Authenticated Byzantine Consensus*, in "10th IEEE International Symposium on Network Computing and Applications (NCA'11)", United States Cambridge, MA, F. CAPELLO, H.-P. SCHWEFEL (editors), IEEE, July 2010, p. 140-146, http://hal.inria.fr/inria-00544518/en.

[63] P. R. WALENGA JR, M. FONSECA, A. MUNARETTO, A. CARNEIRO VIANA, A. ZIVIANI. *ZAP: Um Algoritmo de Atribuicao Distribuida de Canais para Mitigaç ao de Interferencias em Redes com R´adio Cognitivo*, in "SBRC 2010", Brazil Gramado, May 2010, http://www.tkn.tu-berlin.de/publications/papers/64070_correcao.pdf.

[64] T. RAZAFINDRALAMBO, N. MITTON, A. CARNEIRO VIANA, M. DIAS DE AMORIM. *Adaptive Deployment for Pervasive Data Gathering in Connectivity-Challenged Environments*, in "IEEE PERCOM", Germany Mannheim, March 2010, http://hal.archives-ouvertes.fr/hal-00472656_v1/.

[65] T. RAZAFINDRALAMBO, N. MITTON, A. C. VIANA, M. DIAS DE AMORIM, K. OBRACZKA. *Adaptive Deployment for Pervasive Data Gathering in Connectivity-Challenged Environments*, in "Eighth Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM)", France, March 2010, 000, http://hal.inria.fr/hal-00472656/en.

[66] G. TAN, H. JIANG, S. ZHANG, A.-M. KERMARREC. *Connectivity-based and Anchor-Free Localization in Large-Scale 2D/3D Sensor Networks*, in "ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)", United States Chicago, September 2010, http://hal.inria.fr/inria-00493389/en.

National Peer-Reviewed Conference/Proceedings

[67] M. BERTIER, Y. BUSNEL, A.-M. KERMARREC. *Rumeurs, populations et communautés : équivalence uniquement sociologique ? Protocole de population versus protocoles épidemiques*, in "12èmes Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)", France Belle Dune, M. G. POTOP-BUTUCARU, H. RIVANO (editors), 2010, http://hal.inria.fr/inria-00475746/en.

[68] S. CONCHON, J.-C. FILLIÂTRE, F. LE FESSANT, J. ROBERT, G. VON TOKARSKI. *Observation temps-réel de programmes Caml*, in "JFLA (Journées Francophones des Langages Impératifs)", France Vieux-Port La Ciotat, M. MAYERO, S. CONCHON (editors), Studia Informatica Universalis, Hermann Informatique, 2010, p. 195-216, http://hal.inria.fr/inria-00535644/en.

Scientific Books (or Scientific Book chapters)

[69] M. RAYNAL. *Communication and Agreement Abstractions for Fault-Tolerant Asynchronous Distributed Systems*, Morgan & Claypool Publishers, 2010, http://hal.inria.fr/inria-00543036/en.

[70] M. RAYNAL. *Fault-Tolerant Agreement in Synchronous Message-Passing Systems*, Morgan & Claypool Publishers, 2010, http://hal.inria.fr/inria-00543049/en.

Research Reports

[71] O. BALDELLON, A. MOSTEFAOUI, M. RAYNAL. *A Necessary and Sufficient Synchrony Condition for Solving Byzantine Consensus*, INRIA, July 2010, n^{o} PI 1954, http://hal.inria.fr/inria-00521646/en.

[72] G. BIGWOOD, A. CARNEIRO VIANA, M. BOC, A. MARCELO DIAS DE. *Opportunistic data collection through delegation*, INRIA, August 2010, n^{o} RR-7361, http://hal.inria.fr/inria-00508273/en.

[73] F. BONNET, M. RAYNAL. *Anonymous Asynchronous Systems: The Case of Failure Detectors*, INRIA, 2010, n^{o} PI 1945, http://hal.inria.fr/inria-00452799/en.

[74] A. CASTAÑEDA, S. RAJSBAUM, M. RAYNAL. *The Renaming Problem in Shared Memory Systems: an Introduction*, INRIA, November 2010, n^{o} PI-1960, http://hal.inria.fr/inria-00537914/en.

[75] T. CRAIN, D. IMBS, M. RAYNAL. *Read invisibility, virtual world consistency and permissiveness are compatible*, INRIA, November 2010, n^{o} PI-1958, http://hal.inria.fr/inria-00533620/en.

[76] V. ERRAMILLI, K. HUGUENIN, N. LAOUTARIS, I. TRESTIAN. *WYT: Optimized Consistency for Geo-Diverse Online Social Networks*, INRIA, July 2010, n° RR-7343, http://hal.inria.fr/inria-00504913/en.

[77] A. FERNÁNDEZ ANTA, M. MOSTEIRO, C. THRAVES CARO. *Deterministic Recurrent Communication and Synchronization in Restricted Sensor Networks*, INRIA, May 2010, http://hal.inria.fr/inria-00486277/en.

[78] A. FERNÁNDEZ ANTA, M. MOSTEIRO, C. THRAVES CARO. *Deterministic Recurrent Communication in Restricted Sensor Networks*, INRIA, May 2010, http://hal.inria.fr/inria-00486270/en.

[79] A. GIURGIU, R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC. *Computing in Social Networks*, INRIA, May 2010, n° RR-7295, http://hal.inria.fr/inria-00492201/en.

[80] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MAXIME MONOD, S. PRUSTY. *LiFTinG: Lightweight Freerider-Tracking Protocol in Gossip*, INRIA, 2010, n° RR-6913, http://hal.inria.fr/inria-00379408/en.

[81] D. IMBS, M. RAYNAL. *A Simple Snapshot Algorithm for Multicore Systems*, INRIA, July 2010, n° PI 1955, http://hal.inria.fr/inria-00505233/en.

[82] D. IMBS, M. RAYNAL. *The Multiplicative Power of Consensus Numbers*, INRIA, 2010, n° PI 1949, http://hal.inria.fr/inria-00454399/en.

[83] D. IMBS, M. RAYNAL. *The x-Wait-freedom Progress Condition*, INRIA, 2010, n° PI 1944, http://hal.inria.fr/inria-00454386/en.

[84] D. IMBS, M. RAYNAL, G. TAUBENFELD. *On Asymmetric Progress Conditions*, INRIA, May 2010, n° PI-1952, http://hal.inria.fr/inria-00486977/en.

[85] A.-M. KERMARREC, N. LE SCOUARNEC, G. STRAUB. *Beyond Regenerating Codes*, INRIA, September 2010, n° RR-7375, http://hal.inria.fr/inria-00516647/en.

[86] A.-M. KERMARREC, V. LEROY, A. MOIN, C. THRAVES. *Addressing Sparsity in Decentralized Recommender Systems through Random Walks*, INRIA, July 2010, http://hal.inria.fr/inria-00505180/en.

[87] A.-M. KERMARREC, V. LEROY, C. THRAVES. *Ensuring Uniformity in Random Peer Sampling Services*, INRIA, April 2010, http://hal.inria.fr/inria-00477658/en.

[88] A.-M. KERMARREC, V. LEROY, G. TREDAN. *Distributed Social Graph Embedding*, INRIA, June 2010, n° RR-7327, http://hal.inria.fr/inria-00495250/en.

[89] M. LARREA, M. RAYNAL. *Specifying and Implementing an Eventual Leader Service for Dynamic Systems*, INRIA, December 2010, n° PI 1962, http://hal.inria.fr/inria-00543977/en.

[90] S. LE BLOND, A. LEGOUT, F. LE FESSANT, W. DABBOUS. *Angling for Big Fish in BitTorrent*, INRIA, January 2010, http://hal.inria.fr/inria-00451282/en.

[91] A. MOSTEFAOUI, M. RAYNAL. *Signature-Free Broadcast-Based Intrusion Tolerance: Never Decide a Byzantine Value*, INRIA, June 2010, n° PI-1953, http://hal.inria.fr/inria-00495653/en.

[92] M. RAYNAL. *Failure Detectors to Solve Asynchronous k-Set Agreement: a Glimpse of Recent Results*, INRIA, November 2010, n^o PI-1959, http://hal.inria.fr/inria-00536089/en.

[93] M. H. REHMANI, A. C. VIANA, H. KHALIFE, S. FDIDA. *A Cognitive Radio Based Internet Access Framework for Disaster Response Network Deployment*, INRIA, May 2010, n^o RR-7285, http://hal.inria.fr/inria-00482593/en.

[94] A. C. VIANA, N. MITTON, L. SCHMIDT, M. VECCHIO. *SElf-orgaNizing Structures for mAnagemenT In stock Oriented Networks*, INRIA, February 2010, n^o RR-7192, http://hal.inria.fr/inria-00454109/en.

Other Publications

[95] A. BOUTET. *Which aquaintances through distributed social networks?*, July 2010, http://hal.inria.fr/inria-00515421/en.

[96] A. BOUTET, D. FREY, R. GUERRAOUI, A.-M. KERMARREC. *WhatsUp: news from, for, through everyone*, April 2010, http://hal.inria.fr/inria-00468272/en.

[97] A.-M. KERMARREC, E. LE MERRER, G. STRAUB, A. VAN KEMPEN. *Availability-based methods for distributed storage systems*, 2010, in preparation, http://hal.inria.fr/hal-00521034/en.

References in notes

[98] M. AGUILERA. *A Pleasant Stroll Through the Land of Infinitely Many Creatures.*, in "ACM SIGACT News, Distributed Computing Column", 2004, vol. 35, n^o 2.

[99] D. ANGLUIN. *Local and Global Properties in Networks of Processes.*, in "Proc. 12th ACM Symposium on Theory of Computing (STOC'80)", 1980.

[100] K. BIRMAN, M. HAYDEN, O. OZKASAP, Z. XIAO, M. BUDIU, Y. MINSKY. *Bimodal Multicast*, in "ACM Transactions on Computer Systems", May 1999, vol. 17, n^o 2, p. 41-88.

[101] A. DEMERS, D. GREENE, C. HAUSER, W. IRISH, J. LARSON, S. SHENKER, H. STURGIS, D. SWINEHART, D. TERRY. *Epidemic algorithms for replicated database maintenance*, in "Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC'87)", August 1987.

[102] A. G. DIMAKIS, P. B. GODFREY, Y. WU, M. O. WAINWRIGHT, K. RAMCHANDRAN. *Network Coding for Distributed Storage Systems*, in "IEEE Transactions On Information Theory", 2010, vol. 56, p. 4539-4551.

[103] P. EUGSTER, S. HANDURUKANDE, R. GUERRAOUI, A.-M. KERMARREC, P. KOUZNETSOV. *Lightweight Probabilistic Broadcast*, in "ACM Transaction on Computer Systems", November 2003, vol. 21, n^o 4.

[104] M. JELASITY, R. GUERRAOUI, A.-M. KERMARREC, M. VAN STEEN. *The Peer Sampling Service: Experimental Evaluation of Unstructured Gossip-Based Implementations*, in "Middleware", February 2003.

[105] L. LAMPORT. *Time, clocks, and the ordering of events in distributed systems*, in "Communications of the ACM", 1978, vol. 21, n^o 7.

[106] M. LUBY. *LT Codes*, in "FOCS", 2002.

[107] M. MERRITT, G. TAUBENFELD. *Computing Using Infinitely Many Processes.*, in "Proc. 14th Int'l Symposium on Distributed Computing (DISC'00)", 2000.

[108] S. RATNASAMY, P. FRANCIS, M. HANDLEY, R. KARP, S. SHENKER. *A Scalable Content-Addressable Network*, in "Conference of the Special Interest Group on Data Communication (SIGCOMM'01)", 2001.

[109] A. ROWSTRON, P. DRUSCHEL. *Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems*, in "IFIP/ACM Intl. Conf. on Distributed Systems Platforms (Middleware)", 2001.

[110] I. STOICA, R. MORRIS, D. KARGER, F. KAASHOEK, H. BALAKRISHNAN. *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*, in "SIGCOMM'01", 2001.

[111] S. VOULGARIS, D. GAVIDIA, M. VAN STEEN. *CYCLON: Inexpensive Membership Management for Unstructured P2P Overlays*, in "Journal of Network and Systems Management", 2005, vol. 13, n^{o} 2.