INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team ASAP

As Scalable As Possible: Foundations of large-scale dynamic systems

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

Theme : Distributed Systems and Services

Activity Report

2009

Table of contents

ASAP is a bi-localized project-team at INRIA *Rennes - Bretagne Atlantique (*IRISA*) and* INRIA *Saclay - Ile de France sud. ASAP has been officially created on July 1st, 2007.*

1. Team

Research Scientist

Anne-Marie Kermarrec [Team Leader, Research Director, HdR]
Gérard Le Lann [Research Director,, HdR]
Fabrice Le Fessant [Research Scientist, INRIA SACLAY - ILE DE FRANCE SUD]
Aline Carneiro Viana [Research Scientist, INRIA SACLAY - ILE DE FRANCE SUD (On sabbatical since November 2009)]

Faculty Member

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

Antoine Boutet [Ingénieur-Expert]
Davide Frey [Ingénieur-Expert (Since November 2009)]
Erwan Le Merrer [Ingénieur-Expert (Until October 2009)]
Guang Tan [Ingénieur-Expert (Since November 2009)]
Massimo Vecchio [Ingénieur-Expert INRIA SACLAY - ILE DE FRANCE SUD (Since March 2009)]
Paulo Roberto Walenga Junior [Ingénieur-Expert INRIA SACLAY - ILE DE FRANCE SUD (June-July 2009)]

PhD Student

Xiao Bai [INSA- UTC - Chinese Council,]
François Bonnet [MENRT Grant,]
Kévin Huguenin [MENRT Grant,]
Damien Imbs [MENRT Grant,]
Kostantinos Kloudas [INRIA Grant, (since October 2009)]
Nicolas Le Scouarnec [Cifre Thomson Grant,]
Vincent Leroy [MENRT Grant,]
Afshin Moin [INRIA Grant,]
Gilles Trédan [MENRT Grant, (until November 2009)]
Mubashir Husain Rehmani [MENRT Grant, INRIA SACLAY - ILE DE FRANCE SUD]

Post-Doctoral Fellow

Davide Frey [Post-Doc (Until October 2009)]
Guang Tan [Post-Doc (Until October 2009)]
Christopher Thraves [Post-Doc (Since March 2009)]
Golnaz Karbaschi [Post-Doc INRIA SACLAY - ILE DE FRANCE SUD]
Silvija Kokalj-Filipovic [Post-Doc (since October 2009)]

Visiting Scientist

Ajoy Datta [University of Nevada, USA, June 2009]
Anwittaman Datta [University of Singapore, June 2009]
Anelise Munaretto [Technological and Federal University of Parana, Brazil, June-July 2009]
Ymir VigFusson [IBM Haifa, October 2009]

Administrative Assistant

Christine Biard [INRIA SACLAY - ILE DE FRANCE SUD (Since October 2009)]
Cécile Bouton [INRIA RENNES - BRETAGNE ATLANTIQUE]
Marie-Jeanne Gaffard [INRIA SACLAY - ILE DE FRANCE SUD (Until October 2009)]

Other

Jacques-Henri Jourdan [L3 student ENS Paris (June-July 2009)]
Greg Bigwood [University of St. Andrews, UK (March-August 2009)]
Massimo Vecchio [IMP Institute for Advanced Studies, Lucca, Italy (until February 2009)]

2. Overall Objectives

2.1. Highlights of the year

1. **Best paper award** at the International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009) "Looking for the Weakest Failure Detector for k-Set Agreement in Message-passing Systems: Is k the End of the Road", François Bonnet and Michel Raynal.
2. **Best student paper award** (Kevin Huguenin is a PhD student in ASAP) at the International Conference on Principles of Distributed Systems (OPODIS 2009) "Decentralized Polling With Respectable Participants", Rachid Guerraoui, Kevin Huguenin, Anne-Marie Kermarrec and Maxime Monod.

2.2. 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.

2.2.1. 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*" [101] (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 in the past 5 years, 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, such as process mobility, low computing capacity, network dynamicity, scale, etc. This creates a new deal that makes distributed computing more diverse and more challenging.

2.2.2. 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 and dynamicity 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 the targeted application and/or functionality. Among the core functionalities that we are targeting here are efficient search, notification and content dissemination.

2.3. 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.

2.3.1. 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).

2.3.2. 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.

2.4. Resource management in peer-to-peer overlays

Managing resources on a large scale, be them computing resources, data, events, bandwidth, requires a fully decentralized solution. Our research in this area focuses on building the relevant overlay networks to provide core functionalities of resource management and discovery. This includes broadcast, anycast, search, notification. Overlay networks organize peers in a logical network on top of an existing networking infrastructure. The system automatically and dynamically adapts to frequent peer arrivals and departures. Two main classes have been designed. (1) Structured overlay networks, such as distributed hash tables (DHT), rely on a logical structure and map object keys to overlay nodes. While structured networks were initially more popular among scientists, their exact-match interface limits their flexibility and use for various applications, notably for non-exact information retrieval. (2) Unstructured overlay networks connect peers randomly (or pseudo-randomly). This class of networks is dominated by broadcast-based searching techniques, where the goal has become to enforce restrictions on broadcasting so that efficiency can be guaranteed.

In this area, our approach is original for three reasons. First, we are focusing on creating overlays taking into account application characteristics. This translates into either connecting applications objects themselves as peers (which obviously are eventually hosted on a physical computing entity), or influencing the overlay links so that the structure of the application itself can be leveraged for a better performance. We strongly believe that it is not possible to rely on a generic framework applicable to all potential large-scale platforms (as the Internet, grids, or wireless autonomic networks), because of the heterogeneity and limitation of resources among participating entities.

Second, we strongly believe in weakly-structured networks and most of our projects rely 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 [100].

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.

2.5. Peer-to-peer computing and wireless autonomic networks

In this area, we are investigating the use of peer-to-peer algorithms in wireless autonomic networked systems. At the moment, this research area is essentially studied by the network community, although many of these issues are common to the ones encountered in distributed computing such as information propagation, resource discovery, etc.

Wireless autonomic networks and peer-to-peer networks exhibit many similarities that are worth leveraging. Scale and dynamicity are among the most striking ones. The need for scalability prevents the use of any form of centralization. The dynamic nature of such networks imposes to design self-organizing solutions to be able to support churn, disconnection, mobility *etc.*. However, wireless network specificities imply major adaptations of Internet-based peer-to-peer algorithms. More specifically, the fact that the neighbourhood of a node is entirely fixed by the physical network topology, the necessity to take into account the energy consumption, and the broadcast property of the radio communication of wireless nodes have a strong impact on the algorithm design. This recent research area allows us to leverage our peer-to-peer expertise in another application domain with specific applications.

2.6. 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. 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. This is one of the hottest topics of distributed computing today.

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, resource management in IP-based systems, and data management in wireless autonomic networks. 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 [104], [95], [96]. 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 [105], [106], [107] following a specific topology or an unstructured network [99], [108] 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 [98], [97], 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 proven efficient to build and maintain large-scale distributed systems in the context of many applications such as broadcasting [97], 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. Resource management in Internet-based applications

Internet-based applications comprise a large number of applications deployed over the Internet. Such applications however share some common characteristics. First of all, a basic assumption is that participating entities are potentially able to communicate with every other entity using IP. This has a large impact on the possible structure of an overlay network. However, the characteristics of the underlying network in terms of delay and bandwidth might have to be taken into account. This model may serve as a basis to formalize overlay connectivity in such contexts where memory or power consumptions are not an issue, but latency matters.

The applications that we are targeting in this area are related to resource management in large-scale distributed systems. Resource might be related to data, computing power or bandwidth. Among the numerous applications fitting in this denomination, we are especially interested in search, collaborative storage systems, resource discovery and allocation and large-scale content distribution and indexing. Core functionalities of such applications are search, notification and dissemination.

4.3. Sensor-based applications

The advances in hardware development have made possible the miniaturization of micro-electro-mechanical systems and consequently, the development of wireless sensor networks. The combination of inexpensive, autonomous, low-power sensing, and compact devices has improved the viability of deploying large and dense wireless sensor networks able to sense the physical world. By essence, such networks require fully decentralized solutions in which the load is evenly balanced in the system, merely because participating entities have limited in power, storage and communication capabilities.

As opposed to Internet-based applications, entities, here sensors, communicate through radio links and have therefore a limited communication range. This imposes hard constraints on the structure of the resulting topology. More specifically, the overlay structure is highly dependent on the physical topology. Also, sensors, if embedded in human body for example, might be mobile. They might also fail, having some limiting physical capabilities. These properties make such systems highly dynamic.

In this context, we are targeting two main applications: data monitoring and *physical databases*. In the latter applications, as opposed to software databases virtualizing the real objects, sensors embedded on objects themselves can communicate to provide similar functionalities.

5. Software

5.1. Peerple

Participants: Anne-Marie Kermarrec, Fabrice Le Fessant, Silvija Kokalj-Filipovic.

Contact:	Fabrice Le Fessant
Licence:	GPL
Presentation:	Friend-to-Friend Private Peer-to-Peer Client
Status:	under development

Peerple is a peer-to-peer client to share personal documents with friends in a secure and reliable way, and to backup these documents on these “friends” Peerple clients. This work is done in tight collaboration with Laurent Viennot (GANG project-team INRIA Paris - Rocquencourt). We have developed the client, with the following functionalities: a web interface allows friends to connect, authentify on a client and access photo albums using an AJAX interface. A server offers a DNS service for the peer-to-peer clients, and a Mail service to notify friends about the presence of new shared files. Finally, the client is able to backup incrementally files on a local hard disk. The prototype has been released at the beginning of 2007 as an open-source project for external contributions. We are working on adding distributed backups and code modularization.

5.2. Move&Play and MNPlight

Participant: Fabrice Le Fessant.

Contact:	Fabrice Le Fessant
Licence:	Proprietary
Presentation:	Scalable Online File-Sharing System
Status:	under development

MoveNPlay is a web-based system enabling its users to easily access and share their content online. It provides a Web 2.0 interface extended with a powerful Java applet, runs on Google App Engine and uses online storage from the Amazon S3 service. MoveNPlay is developed in the context of the creation of a startup, who received a ANVAR/OSEO grant in May 2008. A release is expected before the end of 2009.

5.3. 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.

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

5.4. HEAP: Heterogeneity-aware gossip protocol.

Participants: Davide Frey, Anne-Marie Kermarrec.

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Java Application
Status:	release expected in the first half of 2010

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.

5.5. Papeer: Peer-to-peer social-centric paper management platform.

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

Contact:	Davide Frey
Licence:	Open Source
Presentation:	Java Application
Status:	first release expected in December 2009

Papeer is a novel peer-to-peer bibliography management system. Papeer's user-centric gossip-based paper-indexing platform enables users to search for scientific papers and share them with their collaborators. Its recommendation facilities allows papeer to locate not only interesting papers but also new people to collaborate with on new research topics. Different from web-based tagging platforms, Papeer's decentralized architecture makes locally stored tags and content available to other users with similar research interests, through a personalized interest-based social network. This social network constitutes the basis for Papeer's ability search for papers, collaborators and to identify which coworkers are most likely to be interested in a given paper or topic.

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. Resource management in large-scale dynamic systems, the main project in this area is the GOSSPLE project;
3. Peer-to-peer wireless autonomic networked systems.

6.2. Models and abstractions: dealing with dynamics

6.2.1. Reducing the cost of Byzantine consensus in synchronous systems

Participant: Achour Mostefaoui.

In a system composed of n processes where at most t can exhibit a Byzantine behavior, it is known since early eighties that t need to be smaller then a third of the total number of process to make the Byzantine consensus problem decidable. Moreover, it has been proved that the minimum number of communication steps needed is $t+1$ in the worst case. Yet, this protocol is extremely costly in terms of the size of messages and local computation (this protocol is called EIG for Exponential Information Gathering). So far, lowering this cost has led to consider smaller values of t (such as $t < n/4$) and an increased latency ($2t+2$ steps). One protocol exist with a reasonable (polynomial) cost with the initial number of steps but is extremely complex. The goal of this work is to design a simple algorithm that is as simple and resilient as the EIG protocol but with a quadratic number of steps.

6.2.2. Peer-to-peer polling without cryptography

Participants: Kévin Huguenin, Anne-Marie Kermarrec.

This work has been done in collaboration with Rachid Guerraoui 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. This work has been published in 2009 in the international conference OPODIS [45]. We are currently working on extension to n-ary polling.

6.2.3. Bridging the Gap between population and gossip-based protocols

Participants: Marin Bertier, Anne-Marie Kermarrec.

In this work, we establish a correlation between population and gossip-based protocols. Studying the equivalence between them, we propose a classification of gossip-based protocols, based on the nature of the underlying peer-sampling service. First, we show that the class of gossip protocols, where each node relies on an arbitrary sample, is equivalent to population protocols. Second, we show that gossip-based protocols, relying on a more powerful peer sampling providing peers with a clearly identified set of other peers, are equivalent to community protocols, a variant of population protocols. Leveraging the resemblances between these areas enables to provide a theoretical framework for distributed systems where global behaviours emerge from a set of local interactions, both in wired and wireless settings. This work has been presented at SIRROCCO 2009 [31]. We also study the impact of the agents mobility model on the convergence speed of population protocols. We perform our study by considering several mobility models traditionally used in the ad-hoc network community. We propose an augmented population protocol model where each edge of the interaction graph is weighted, representing the probability of two agents to interact.

6.2.4. STAR

Participants: Erwan Le Merrer, Anne-Marie Kermarrec.

STAR is a fully decentralized self-stabilizing randomized membership protocol building a strongly connected overlay graph with sub-logarithmic diameter and almost homogeneous logarithmic degree. STAR is the first protocol to simultaneously maintain the following properties on the resulting graph G: (i) The graph maintains the *Eulerian* property, i.e., that the in-degree and out-degree of each node are the same, and G is strongly connected. (ii) The out-degree of each node automatically converges to an average of $2\ln(n) + O(1)$ without any node knowing the exact size n of the network. (iii) The diameter of the overlay graph is $\frac{\ln n}{\ln \ln n + \ln 2} + O(1)$ with high probability. (iv) STAR is self-stabilizing. Starting from an arbitrary graph topology, or after disruptive error, STAR causes the overlay graph to converge to the desired properties. This work has been done in collaboration with Prof. Ajoy Datta, University of Nevada, Las Vegas.

6.2.5. Agreement in anonymous systems

Participants: François Bonnet, Michel Raynal.

This work addresses the consensus problem in asynchronous systems prone to process crashes, where additionally the processes are anonymous (they cannot be distinguished one from the other: they have no name and execute the same code). To circumvent the three computational adversaries (asynchrony, failures and anonymity) each process is provided with a failure detector of a class denoted ψ, that gives it an upper bound on the number of processes that are currently alive (in a non-anonymous system, the classes ψ and $\mathcal{P}$ -the class of perfect failure detectors- are equivalent).

After having designed a simple ψ-based consensus algorithm where the processes decide in $2t+1$ asynchronous rounds (where t is an upper bound on the number of faulty processes), we have shown that $2t+1$ is a lower bound for consensus in the anonymous systems equipped with ψ. Then addressing early-decision, we have designed and proved correct an early-deciding algorithm where the processes decide in $\min(2f+2, 2t+1)$ asynchronous rounds (where f is the actual number of process failures). This leads to think that anonymity doubles the cost (wrt. synchronous systems) and it is conjectured that $\min(2f+2, 2t+1)$ is the corresponding lower bound. These results have then been extended to k-set agreement problem, for which we have introduced a family of failure detector classes $\{\psi_\ell\}_{0\leq\ell<k}$ that generalizes the class $\psi(=\psi_0)$ and designed an algorithm that solves the k-set agreement in $R_{t,\ell} = 2\left\lfloor \frac{t}{k-\ell} \right\rfloor + 1$ asynchronous rounds. This last formula relates the cost ($R_{t,\ell}$), the coordination degree of the problem (k), the maximum number of failures (t) and the the strength (ℓ) of the underlying failure detector.

6.2.6. Looking for weakest failure detectors

Participants: François Bonnet, Michel Raynal.

In the k-set agreement problem, each process (in a set of n processes) proposes a value and has to decide a proposed value in such a way that at most k different values are decided. While this problem can easily be solved in asynchronous systems prone to t process crashes when $k > t$, it cannot be solved when $k \leq t$. Since several years, the failure detector-based approach has been investigated to circumvent this impossibility. While the weakest failure detector class to solve the k-set agreement problem in read/write shared-memory systems has recently been discovered (PODC 2009), the situation is different in message-passing systems where the weakest failure detector classes are known only for the extreme cases $k = 1$ (consensus) and $k = n-1$ (set agreement). We have introduced introduces a candidate for the general case. It presents a new failure detector class, denoted Π_k, and shows $\Pi_1 = \Sigma \times \Omega$ (the weakest class for $k = 1$), and $\Pi_{n-1} = \mathcal{L}$ (the weakest class for $k = n-1$). Then, we have investigated the structure of Π_k and shows it is the combination of two failures detector classes denoted Σ_k and Ω_k (that generalize the previous “quorums” and “eventual leaders” failure detectors classes). Finally, we have proved that Σ_k is a necessary requirement (as far as information on failure is concerned) to solve the k-set agreement problem in message-passing systems. This work has obtained the Best Paper Award at SSS 2009.

6.2.7. Implementing a register in a dynamic distributed system

Participants: François Bonnet, Michel Raynal, Anne-Marie Kermarrec.

This work was carried out in collaboration with S. Bonomi and R. Baldoni, Universita’ di Roma, La Sapienza, Roma, Italy. Providing distributed processes with concurrent objects is a fundamental service that has to be offered by any distributed system. The classical shared read/write register is one of the most basic ones. Several protocols have been proposed that build an atomic register on top of an asynchronous message-passing system prone to process crashes. In the same spirit, we have addressed the implementation of a regular register (a weakened form of an atomic register) in an asynchronous dynamic message-passing system. The aim is here to cope with the net effect of the adversaries that are asynchrony and dynamicity (the fact that processes can enter and leave the system). Our work focuses on the class of dynamic systems the churn rate c of which is constant. It presents two protocols, one applicable to synchronous dynamic message passing systems, the other one to eventually synchronous dynamic systems. Both protocols rely on an appropriate broadcast communication service (similar to a reliable broadcast). Each requires a specific constraint on the churn rate c. Both protocols are first presented in an as intuitive as possible way, and are then proved correct (ICDCS 2009, SIROCCO 2009).

6.2.8. *The extended Boroswky-Gafni simulation*

Participants: Damien Imbs, Michel Raynal.

The *Borowsky-Gafni* (BG) *simulation* algorithm is a powerful tool that allows a set of $t+1$ asynchronous sequential processes to wait-free simulate (i.e., despite the crash of up to t of them) a large number n of processes under the assumption that at most t of these processes fail (i.e., the simulated algorithm is assumed to be t-resilient). The BG simulation has been used to prove solvability and unsolvability results for crash-prone asynchronous shared memory systems. In its initial form, the BG simulation applies only to colorless decision tasks, i.e., tasks in which nothing prevents processes to decide the same value (e.g., consensus or k-set agreement tasks). Said in another way, it does not apply to decision problems such as renaming where no two processes are allowed to decide the same new name. Very recently (STOC 2009), Eli Gafni has presented an *extended* BG *simulation* algorithm (GeBG) that generalizes the basic BG algorithm by extending it to "colored" decision tasks such as renaming. His algorithm is based on a sequence of sub-protocols where a sub-protocol is either the base agreement protocol that is at the core of BG simulation, or a commit-adopt protocol.

We have designed a core algorithm for the extended BG simulation algorithm that is particularly simple. This algorithm is based on two underlying objects:the base agreement object used in the BG simulation (as does GeBG), and (differently from GeBG) a new simple object that we call *arbiter*. As in GeBG, while each of the n simulated processes is simulated by each simulator, each of the first $t+1$ simulated processes is associated with a predetermined simulator that we called its "owner". The arbiter object is used to ensure that the permanent blocking (crash) of any of these $t+1$ simulated processes can only be due to the crash of its owner simulator (SSS 2009).

6.2.9. *Software transactional systems*

Participants: 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 decompose each sequential process the application is made up of into 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. For efficiency, a STM system allows transactions to execute concurrently. Consequently, due to the underlying STM concurrency management, a transaction commits or aborts. Our work on STM has several facets. We have designed a new STM consistency condition, called *virtual world* consistency. This condition states that no transaction reads object values from an inconsistent global state. It is similar to opacity for the committed transactions but weaker for the aborted transactions. More precisely, it states that (1) the committed transactions can be totally ordered, and (2) the values read by each aborted transaction are consistent with respect to its causal past only. Hence, virtual world consistency is weaker than opacity while keeping its spirit. Then, assuming the objects shared by the processes are atomic read/write objects, the paper presents a STM protocol that ensures virtual world consistency (while guaranteeing the invisibility of the read operations). From an operational point of view, this protocol is based on a vector-clock mechanism. Finally, the paper considers the case where the shared objects are regular read/write objects. It also shows how the protocol can be weakened to satisfy the *causal consistency* condition (that is weaker than virtual world consistency) (SIROCCO 2009, PaCT 2009).

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. The GOSSPLE challenges have been published in SSS 2009 [55].

6.3.1. The Gossple anonymous social network

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

This work [25] has been done 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.

6.3.2. Query expansion in Gossple

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

This work [32] has been done in collaboration with Prof. Rachid Guerraoui (EPFL). A query expansion system seeks to extend a set of keywords (depicting a query) with additional ones to improve the query results. We consider here a collaborative tagging system, such as Delicious or CiteUlike, where every node is associated with a tagging profile. We describe how to use the GOSSPLE network to expand queries in a personalized way and significantly improve the completeness (recall) and accuracy (precision) of the results, with respect to the state of the art personalized centralized approach, namely Social Ranking. Note GOSSPLE's personalized approach also automatically handles cases like homonym ambiguity. For instance, the terms "computer" or "fruit" will be added to the "apple" query depending on the node's profile.

This is applied in a collaborative tagging system such as delicious where users tag items with tags. The way we achieve the query expansion is as follows. We use GOSSPLE personalized network on each node (called the *GNet*) to compute a data structure we call *TagMap*, a personalized view of the relations between the tags in the node's profile and those in its *GNet*. The *TagMap* is the only source of information that is used for the query expansion and it is updated periodically to reflect the changes in the *GNet*. Only the scores associated with the tags affected by the modifications need to be updated. A query per se is then expanded using the *TagMap* through a *centrality* algorithm we call *TagRank*, which we derived from the celebrated *PageRank* algorithm. While PageRank computes the relative importance of Web pages (eigenvector centrality), *TagRank* computes the relative importance of tags on a given node. Our *TagRank* algorithm estimates the relevance of each tag in the *TagMap* with respect to the query and assigns a score to the tag. Our results obtained on real traces crawled from Delicous show that the *Gossple* query expansion mechanism improves both recall and precision.

6.3.3. Gossip-based top-k processing in Gossple

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

This work [25] has been done in collaboration with Prof. Rachid Guerraoui (EPFL). A fine grained personalisation 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.

6.3.4. Recommendation in Gossple

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

We are working on a decentralized recommender system to be used in peer to peer systems. This recommender system relies on epidemic protocols to form a personal network for each peer, using a correlation measure suitable for decentralized systems. A random walk graph model is then used to detect further the *confidence* between users. This confidence finally serves as the weight for the ratings of the neighbors in the personal network in order to predict the unknown ratings of the central user. At the end, a diffusion phase is added to retreive a higher number of recommendations.

Papeer is a novel user-centric gossip-based paper indexing platform, through which users can search for scientific papers, share them with their collaborators as well as find new people to collaborate with on new research topics. Different from web-based tagging platforms, Papeer's decentralized architecture makes locally stored tags and content available to other users with similar research interests, through a personalized interest-based social network. The same social network constitutes the basis for Papeer's ability to recommend the coworkers who are most interested in a given paper, or who are most likely to be interested in a given paper or topic. We validated Papeer's recommendation approach by means of an experimental evaluation on a data trace with 13000 users.

6.3.5. Gossip protocols in practice: NAT-resilient protocols

Participant: Anne-Marie Kermarrec.

Gossip peer sampling protocols now represent a solid basis to build and maintain peer to peer (p2p) overlay networks. They typically provide peers with a random sample of the network and maintain connectivity in highly dynamic settings. They rely on the assumption that, at any time, each peer is able to establish a communication with any of the peers of the sample provided by the protocol. Yet, this ignores the fact that there is a significant proportion of peers that now sit behind NAT devices (70% is a fair ratio in the current

Internet), preventing direct communication without specific mechanisms. This has been largely ignored so far in the community. Our experiments demonstrate that the presence of NATs, introducing some restrictions on the communication between peers, significantly hurts both the randomness of the provided samples and the connectivity of the p2p overlay network, in particular in the presence of high rate of peers arrivals, departures and failures (aka churn). We proposed a NAT resilient gossip peer sampling protocol, called *Nylon* that accounts for the presence of NATs. *Nylon* is fully decentralized and spreads evenly the extra load caused by the presence of NATs, between peers. *Nylon* ensures that a peer can always establish a communication, and therefore initiates a gossip, with any peer in its sample. This is achieved through a simple, yet efficient mechanism, establishing a path of relays between peers. Our results show that the randomness of the generated samples is preserved, that the connectivity is not impacted even in the presence of high churn and a high ratio of peers sitting behind NAT devices. This work has been published in ICDCS 2009 [60].

6.3.6. Search over social networks

Participants: Anne-Marie Kermarrec, Guang Tan.

This project targets advanced information needs that can hardly be satisfied by generic search engines even in the predictable future, and thus require human's explicit effort. Specificly we study a popular network application: the online question/answer services. Different from existing solutions provided by centralized sites (e.g., Yahoo! Answers), our sysem, called AskBuddies, takes a decentralized architecture over social networks such as MySpace, Friendster, MSN and Yahoo Messenger. The system aims to find for a query the top-k best answerers, considering both knowledge match and social distance which we believe plays an important role in providing incentive. Efficient algorithms of identifying short chains of acquaintance, or referral chains, and of distributed top-k processing will be devised and implemented. Through deploying and running a real system we will examine user behaviors in such an application, and how social networks may benefit distributed information retrieval in more general contexts.

6.3.7. Cold start link prediction in social network

Participant: Vincent Leroy.

This work 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.

6.4. Resource management in large-scale dynamic systems

6.4.1. Heterogeneous gossip

Participants: Davide Frey, 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. This work has been published at DSN 2009 [43] and Middleware [42].

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.

6.4.3. LTNC

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

This work has been done in collaboration with Mary-Luc Champel (Thomson Labs). Network coding has been successfully applied in large-scale content dissemination systems. While network codes provide optimal throughput, its 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 [103], initially introduced in the context of erasure coding. Our coding scheme, called LTNC, 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. In addition, LTNC consistently outperforms dissemination protocols without codes, thus preserving the benefit of coding.

6.4.4. Phosphite

Participants: Anne-Marie Kermarrec, Nicolas Le Scouarnec.

Phosphite [38] is a mechanism to preserve out of order download in peer to peer video-on-demand applications, in the presence of selfish peers. In such applications, peers have a natural trend to download blocks in order to start watching videos as soon as possible. Without specific mechanism to enforce a fair amount of out of order download, the last blocks of the video tend to be lost due to peers leaving soon after having downloaded the last blocks thus forcing peers to rely on the central server for re-introducing those lost blocks. This issue can be solved if peers dedicate a portion of their bandwidth for out of order downloads. Yet, this heavily relies on the goodwill of peers to collaborate. Phosphite is a simple yet efficient approach ensuring that all peers dedicate a part of their bandwidth to out of order download. Phosphite relies on a computational challenge where peers are provided with a combination of the requested blocks and other blocks. This forces peers to download out of order blocks to be able to decode the requested blocks. We evaluate Phosphite and show that it successfully prevents the system from losing blocks, even in the presence of selfish peers, thus offering an appealing alternative to state of the art approaches. With Phosphite, the last blocks remain available (with a probability higher than 0.98), while such result cannot be guaranteed (with a probability lower than 0.5) without enforcement mechanism. Phosphite ensures that a peer to peer download is almost always possible, even in the presence of selfish peers.

This work has been done in collaboration with Mary-Luc Champel. A preliminary paper has been accepted at the French conference Algotel 2009 [37] and the final version of the paper has been accepted at the international conference P2P 2009 [38].

6.4.5. Distributed channel switching in P2P-IPTV

Participants: Erwan Le Merrer, Anne-Marie Kermarrec.

It is now common for IPTV systems attracting millions of users to be based on a peer-to-peer architecture. In such systems, each channel is typically associated with one P2P overlay network connecting the users. Yet, the joining process resulting in a peer to be integrated in channel overlay usually requires a significant amount of time. As a consequence, switching from one channel to another is far to be as fast as in IPTV solutions provided by telco operators. In this work, we tackle the issue of efficient channel switching in P2P IPTV system. We formulate the switching problem, and propose a simple distributed algorithm, as an illustration of the concept, which aims at leveraging the presence of peers in the network to fasten the switch process. This work has been published at Euro-Par 2009 [94].

6.4.6. Combining gossip and fountain codes for fast content dissemination

Participants: Anne-Marie Kermarrec, Nicolas Le Scouarnec.

Gossip-based dissemination reaches all nodes with high probability. However, some peers may not receive all the data. In order to enhance the performance of gossip-based dissemination protocol, fountain codes can be used to recover the missing data. We propose a protocol that combine a gossip-based dissemination protocol and fountain codes. The gossip protocol builds a structured split-graph overlay to split the peers between encoders and forwarders. Forwarders become encoders as soon as they have received the whole content so they can start encoding. In order to benefit even further from encoders, we deliberately bias the dissemination process through the gossip protocol so that forwarders become earlier potential encoders. We implemented this protocol on the PlanetLab testbed. By simulation in PeerSim, we assess that this protocol outperforms a simple protocol that does not introduce a bias or that does not update the overlay to match the peers'role.

6.4.7. Peer-to-peer collaborative back-up

Participants: Anne-Marie Kermarrec, Fabrice Le Fessant.

The storage capacity of computers has increased a lot in the past years: in the meantime, final users have started using this storage for important personal data, with the democratization of digital cameras, and professional data with the rise of telecomputing. Backuping all this data has become a new challenge for peer-to-peer systems, since these users are connected most of the time, often with large unused storage capacity on their disks, and unfortunately seldom take the time to properly save these important data.

Anne-Marie Kermarrec and Fabrice Le Fessant are currently designing a platform for a collaborative backup system, and this problem tackles a large set of problems: making the backup resilient to the large number of failures characterizing peer-to-peer networks, choosing where to backup the data, designing the protocols to place and retrieve the data from the network, while ensuring secrecy/privacy of the data. The prototype, currently developed by Fabrice Le Fessant within the Peerple open-source project, uses both a structured overlay, to localize stored data during restoration, and an unstructured overlay, to query for storage availability among neighbors. Contrary to most peer-to-peer backup systems, files are not stored separately on the overlay network, but gathered in volumes, encrypted using strong cryptography for privacy, and replicated using Reed-Solomon coding, to ensure availability even in the presence of high failure rates at a minimal extra storage cost. This work is done in collaboration with Laurent Viennot from the GANG project-team, INRIA Paris - Rocquencourt.

First results have been published in DAMAP'09[30].

6.4.8. Measuring availability in peer-to-peer systems

Participants: Anne-Marie Kermarrec, Fabrice Le Fessant.

Tracking peer availability in a peer-to-peer network is of utmost importance for many collaborative applications. For instance, such information is invaluable for identifying the most stable peers or group of peers with similar uptime characteristics. However, as many applications tend to reward the most stable peers, there is a clear incentive for peers to try to appear more available than their real availability. We developed a scalable and lightweight protocol that enables nodes to measure the peer availability in the presence of such selfish peers. In our protocol, which is called Pacemaker each peer is in charge of maintaining proofs of its own availability over time by collecting pulses disseminated by a trusted entity using asymmetric cryptographic signatures. Essentially, using these pulses, peers gain the ability to challenge other peers and verify that their real uptime matches the advertised one. Simulation results show that our protocol provides accurate availability measures even in the presence of selfish peers. Furthermore, our results are verified by experiments in Planetlab, which also illustrates the deployability of Pacemaker in real networks.

6.4.9. Finding good partners in availability-aware p2p networks

Participants: Fabrice Le Fessant, Erwan Le Merrer.

We studied the problem of finding peers matching a given availability pattern in a peer-to-peer (P2P) system. Motivated by practical examples, we specified two formal problems of availability matching that arise in real applications: disconnection matching, where peers look for partners expected to disconnect at the same time, and presence matching, where peers look for partners expected to be online simultaneously in the future. As a scalable and inexpensive solution, we proposed to use epidemic protocols for topology management; we provided corresponding metrics for both matching problems. We evaluated this solution by simulating two P2P applications, task scheduling and file storage, over a new trace of the eDonkey network, the largest available with availability information. We first proved the existence of regularity patterns in the sessions of 14M peers over 27 days. We also showed that, using only 7 days of history, a simple predictor could select predictable peers and successfully predicted their online periods for the next week. Finally, simulations showed that our simple solution provided good partners fast enough to match the needs of both applications, and that consequently, these applications performed as efficiently at a much lower cost. We believe that this work will be useful for many P2P applications for which it has been shown that choosing good partners, based on their availability, drastically improves their performance and stability. This work has been done in collaboration with Stevens Leblond from INRIA Sophia-Antipolis.

Our results have been published in SSS'09 [61], CFSE'09 [70], RR-6795 [88].

6.5. Peer-to-peer wireless autonomic networked systems

6.5.1. Low stretch and small state geographic routing in sensor networks

Participants: Marin Bertier, Anne-Marie Kermarrec, Guang Tan.

This work addresses a central problem in ad hoc sensor (and more generally wireless) networks: that of finding (asymptotically) shortest paths using minimum possible per-node state. Assuming the availability of location information (hence geographic) to each node, and that a source node knows the location of a destination node, we explore this problem in two approaches. In the first approach, we abstract the network into a highly compact global structure, known as visibility graph in computational geometry, that produces constant-stretch paths while getting rid of the notorious reliance on $O(n)$ per-node state. To our knowledge this is the first geographic routing protocol that offers provably optimal performance with per-node state independent of network size. In the context of a more general location-free (or virtual coordinate) environment, our second approach uses the idea of divide-and-conquer, whereby the network is partitioned to convex pieces that are amenable to greedy routing. Based on new insights into the performance pitfalls of previous work, the notion of convexity results in a significantly better tradeoff between stretch and state size than existing solutions.

6.5.2. Adaptive data forwarding in delay-tolerant networks

Participants: Marin Bertier, Anne-Marie Kermarrec, Aline Carneiro Viana.

We conducted studies in information dissemination in mobile networks where a contemporaneous path may never exist between two nodes (a source and a destination) in the network, known as delay-tolerant networks (DTNs). These networks are further challenged by strict resource constraints (e.g., memory, CPU and energy limitations). Our goal in this domain was to support intelligent and adaptive forwarding, which allows a good trade-off between reliability and resource-efficiency. We then design a new protocol, called Seeker, which empowers nodes with the ability to estimate *favorable contact opportunities* by taking advantage of any information they can locally infer. Hence, nodes are able to adapt and self-organize in dynamic environments with minimal control overhead. This work has been performed in collaboration with researcher Roy Friedman (from Technion - Israel Institute of Technology). The general idea was published in the ExtremeCom workshop [67], which, although at its first edition, gave us the opportunity to meet other active researchers and practitioners in areas related to delay tolerant networks. The complete description of the protocol can also be found in [92].

6.5.3. Reliable data dissemination in cognitive radio networks

Participants: Mubashir Husain Rehmani, Aline Carneiro Viana.

Recent advances in cognitive radio technology have enabled opportunistic wireless cognitive radio (CR) nodes to efficiently locate and exploit spectrum under-utilized by licensed primary radio (PR) nodes. Nevertheless, CR transmissions should not degrade the reception quality of PR nodes and should be immediately interrupted whenever a neighboring PR activity is detected. The main problem we tackle here is *the channel assortment in multi-hop cognitive radio networks and concerns how to increase the transmission coverage of CR nodes by selecting good qualified channels for communicating*. Our *channel assortment strategy*, named 'SURF', provides to CR nodes a strategy to assort channels according to their availabilities, giving to nodes the possibility of selecting the best classified one for transmission and/or overhearing. Finally, our strategy improves network reliability – since less PR-occupied channels will be selected – and receivers coverage – since an acceptable number of CR neighbors in terms of space sharing will compete for the selected channel.

This work has been done in collaboration with Hicham Khalife from LaBRI/Université Sciences et Technologies - Bordeaux I.

6.5.4. A content-based network coding to match social interest similarities in delay tolerant networks

Participants: Golnaz Karbaschi, Aline Carneiro Viana.

Considering users' social characteristics leads to better use of network coding in delay tolerant networks. We present a content-based network coding that aims to match the social interests similarities (i.e., same profession, hobbies, interests, etc.) of people in a community. We aim to provide a more adaptive-to-social-network coding and lower decoding delay for the users that are interested in different *contents*. This work was published in the ExtremeCom 2009 Workshop [53]. Additionally, adaptive-to-social-network coding requires to deal with the unfairness issues of network coding in multi-hop wireless networks. The reason for this unfairness is that by mixing different flows, packets destined to one destination in order to be decoded need to wait for the reception of the whole mixed set of encoded packets that may be totally independent in terms of final destination. This may lead to highly unfair delay for small block data. To mitigate this unfairness, relay nodes may mix only packets going to the same destination. We call this strategy *FairMix* and it was published in the IEEE PIMRC 2009 conference [54]. This work was performed in collaboration with Khaldoun Al Agha and Steven Martin from LRI/University of Paris-Sud. More details about the idea can be also found in [85].

6.5.5. Cryptographic protocols to fight sinkhole attacks on tree-based routing in wireless sensor networks

Participants: Aline Carneiro Viana, Fabrice Le Fessant.

Wireless Sensor Networks (WSN) are penetrating more and more in our daily life. As a consequence, security has become an important matter for these networks. We introduce two new cryptographic protocols of different complexity and strength in limiting network degradation caused by sinkhole attacks on tree-based routing topologies in Wireless Sensor Networks (WSNs). The main goal of both protocols is to provide continuous operation by improving resilience against, rather than detection of, these attacks. The main benefit of providing resilience is that it allows operating (or graceful degradation) in the presence of attacks. Furthermore, while resilience mechanisms do not dismiss detection mechanisms, detection mechanisms often introduce more complexity and so, more weaknesses to the system, which might not justify their benefits. More specifically our two *RESIlient and Simple Topology-based reconfiguration protocols* are: RESIST-1 and RESIST-0. RESIST-1 prevents a malicious node from modifying its advertised distance to the sink by more than one hop, while RESIST-0 does not allow such lying at the cost of additional complexity. This work was firstly published in the NPSec Workshop [102] held in conjunction with ICNP 2009.

6.5.6. Self-management: Virtual coordinates for autonomous networked system

Participants: Anne-Marie Kermarrec, Achour Mostefaoui, Michel Raynal, Gilles Trédan, Aline Carneiro Viana.

The motivation behind this research work comes from the lack in the literature, of an autonomous system able (1) to permanently evolve and self-organize under dynamic changing conditions (due either to the environment or technological issues), and (2) to provide various networking functionalities over the same underlying support system. Our answer to those demands was an autonomous system able not only to be adaptable to environment changing conditions, but also, that provides variety to network functionalities. Thus, we have designed an autonomous and lightweight self-organizing networked system that, by imposing a bounded overhead to wireless devices, constructs a base network structure for supporting network functionalities, commonly required in WSNs. We have only exploited local connectivity information and per-neighbor communication. This work has been performed in collaboration with researcher Roy Friedman (from Technion - Israel Institute of Technology) and is currently in submission. The research related to this subject has been published at the ICDCN conference [58].

6.5.7. Data organization in wireless sensor networks

Participants: Massimo Vecchio, Aline Carneiro Viana.

We have been working on how to make the sensed monitored data available to the mobile sink in a robust, adaptive, and efficient way. In this context, we have proposed, in collaboration with researchers Artur Ziviani (from the National Laboratory for Scientific Computing (LNCC), Brazil) and Roy Friedman (from the Technion - Israel Institute of Technology) an efficient data dissemination approach (in terms of overhead and representativeness) to allow a mobile sink to gather a representative view of the monitored region covered by n sensor nodes by visiting **any** m nodes, where $m \ll n$. This work was firstly published in the journal IEEE Communication Letter [14]. Then, the Post-Doc Massimo Vecchio started working on other probabilistic techniques in order to decrease the overhead incurred by this previous approach. This work is currently in submission to a journal. We have received the first reviewers' comments and resubmit the paper again.

Additionally, with researchers from the LRI/University of Paris-Sud (Fatiha Zaidi, Thomas Herault, Thomas Largillier, and Sylvain Peyronnet) the protocol Supple was also proposed for proactive data dissemination in wireless sensor networks. In Supple, sensor nodes use a simple tree-based structure, constructed during the neighborhood discovery phase, which allows weight distribution among nodes. Weights are based on predefined criterion of selection as well as distribution law, and are used by sensors at the data dissemination phase. At this phase, sensors then make on the fly forwarding and data storing decisions based on their own weights and the weights of their neighbors. Supple takes thus advantage of the bias among different sensors' weights for good data dissemination. The resulting work is currently in submission at the IEEE Percom 2010 Conference.

With researchers Fatiha Zaidi from LRI/University of Paris-Sud and Stephane Maag from TELECOM & Management SudParis, the paper "One step forward: Linking Wireless Self-Organising Networks Validation Techniques with Formal Testing approaches" was accepted for publication in one of the main ACM Journals in Computer Science, the ACM Computing Surveys [13].

6.5.8. Energy-efficient route discovery in sensor networks

Participant: Aline Carneiro Viana.

The vast literature on the wireless sensor research community contains many valuable proposals for managing energy consumption, the most important factor that determines sensor lifetime. The goal of this work is to extend the network lifetime. We aim at determining good energy-efficient routes in the network by using the energy level of nodes as a criterion to select good links in the route. In collaboration with Khaldoun Al Agha and PhD student Joseph Rahme (who is co-advised for 30% by Aline) both from LRI/Universite Paris-Sud, our first contribution in this context was the proposal of cost functions to choose energy efficient routes. The originality of the proposed cost functions lies in their completely decentralized and adaptive behavior in considering energy consumption, remaining energy of nodes, and the number of transmissions a node can make before its energy depletion.

6.5.9. Adaptive infrastructure deployment for data gathering in intermittently connected networks

Participant: Aline Carneiro Viana.

A collaboration with Marcelo Dias de Amorim from CNRS/LIP6 focuses on applications of wireless sensor networks that require periodic readings. This means that these readings should be performed following some predefined parameter $f_{\min}$ that denotes the minimum frequency at which the whole target area must be sensed. We consider mobile sensor network and focus on *how* sensors should move, in order to guarantee the coverage of all targets in the network in a timely and efficient way. To answer this issue, we proposed a mobility strategy that consists in making nodes follow a Hilbert space-filling curve and use opportunistic contacts to reduce the data delivery's delay.

Additionally, a joint work with researchers from EPI POPS of INRIA Lille (Tahiry Razafindralambo, Nathalie Mitton), Marcelo Dias de Amorim, and Katia Obraczka from UC Santa Cruz, CA, USA started, where the protocol COVER was proposed. COVER is an adaptive strategy for placement and trajectory control of infrastructure nodes in intermittently connected networks, where data-producing nodes (targets) have to report data on a regular basis. The goal is to cover mobile targets whose mobility patterns are unknown. The main issue is how to *deploy* and *manage* the infrastructure nodes in order to adaptively guarantee network availability, to respect required data gathering latency, to balance load and still to limit deployment cost. The first results are currently in submission at the IEEE Percom 2010 conference.

6.5.10. Building secured links is sensor networks

Participants: Marin Bertier, Achour Mostefaoui, Gilles Trédan.

This work 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. Thomson

Participants: Anne-Marie Kermarrec, Nicolas Le Scouarnec.

From november 2007 to november 2010, we have a collaboration with Thomson R&D France on providing coding and incentives mechanisms for audio/video secure content distribution over Internet using the peer to peer paradigm. In this context, Anne-Marie Kermarrec acts as the PhD advisor of Nicolas Le Scouarnec.

8. Other Grants and Activities

8.1. National grants

8.1.1. ANR VERSO project Shaman

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

The Shaman project started in 2009, grouped together a majority of 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.

8.1.2. Rnrt project SVP

Participants: Marin Bertier, Anne-Marie Kermarrec, Aline Carneiro Viana.

The SVP project addressed the understanding, the conception, and the implementation of an integrated ambient architecture that would ease the optimization in the deployment of surveillance and prevention services in different types of dynamic networks. The main objective was to develop an environment which is able to accommodate a high number of dynamic entities completely dedicated to a specific service. The partners of the project come from various research communities: network, distributed system, sensor architecture and metabolical and mechanical motion control (CEA, ANACT, APHYCARE, INRIA, UPMC/LIP6, LPBEM, Thalès). Our work on adaptive data forwarding in delay tolerant networks took place in this context. This project has officially been concluded in September 2008.

8.1.3. RTRA Digiteo

Participants: Golnaz Karbaschi, Aline Carneiro Viana.

DigiteoLabs is a virtual lab created to gather and promote collaborations between the following research centers: INRIA, University of Paris-Sud, Supelec, Ecole Polytechnique, and CEA. In the context of a call for regional collaborating projects, the ASAP project entitled "Resource management in delay tolerant networks" was selected, which is financing a 1-year Post-Doc fellowship of Golnaz Karbaschir in ASAP Saclay. The project targets the use of network coding to improve the use the network availabilities, like throughput. Khaldoun Al Agha and Steven Martin from LRI/University of Paris-Sud are also part of the project. The work resulting from this project was published in the ExtremeCom 2009 Workshop [53] and in the IEEE PIMRC 2009 conference [54].

Besides, other projects in this context enabled the invitations (1) for 4 months the researcher Aartur Ziviani from the LNCC laboratory, Brazil (2008-2009), and (2) for 2 months the researcher Anelise Munaretto from the Technological and Federal University of Parana, Brazil.

8.1.4. ANR Project USS-SimGrid

Participant: Fabrice Le Fessant.

USS-SimGrid has been accepted in the ANR call “Embedded Systems and Big Infrastructures”. It is composed of seven academic research labs, aiming at improving the SimGrid grid simulator. The main goal is to increase the scalability of the simulator to be able to simulate bigger and more asynchronous systems, such as peer-to-peer systems. The ASAP team is responsible for testing the simulator by using it to simulate collaborative backup and peer-to-peer video-on-demand protocols. The project is a 3-year project starting at the end of 2008.

8.1.5. ICOM Project of the PICOM

Participant: Aline Carneiro Viana.

ICOM (“Infrastructure pour le COMmerce du futur”) is a project of the “Pôles de Compétitivé Industries du commerce (PICOM)”. This project is composed by three INRIA research teams (POPS, ASAM, and ASAP) and by some important industrial and commercial partners (i.e. Declathon, La Redoute, Orange, La Poste, Auchan, GS1, Atos Origin). This project targets the study, the exploitation, and the experimentation of a large-scale ambient platform that will (1) make simple the fast deployment of new ubiquitous applications of infrastructure management and (2) cope with heterogeneous technologies (RFID, NFC, Zigbee, etc). This project has officially started in April 2008. A 2-year research engineer will be selected to work on the project. ASAP has presented his contribution as well as a demo in the meeting of 23rd October 2009 and is preparing its final report for December 2009.

8.1.6. Project Pôle de Competitivité Images & Réseaux - P2Pim@ges

Participants: Anne-Marie Kermarrec, Erwan Le Merrer.

The P2Pim@ges project deals with secure multimedia file distribution in peer-to-peer environments. The role of the ASAP project-team is to provide the peer-to-peer infrastructure to distribute large multimedia files in an efficient way, in particular using epidemic protocols to achieve efficient and relevant clustering. The following partners within the *Pôle de Competitivité Images & Réseaux* are involved : Thomson R&D, Thomson Broadcast & Multimedia, Mitsubishi Electric ITE/TCL, Devoteam, France Télécom, ENST Bretagne, Marsoin, IPdiva, TMG and eOdus. Project has ended in November 2009.

8.1.7. Rnrt project SensLab

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

SensLab is an RNRT project started in 2008 focussing 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. Consequently, a SensLAB platform composed of 1024 nodes is deployed among 4 sites. This infrastructure will represent the unique scientific tool for the research on wireless sensor networks.

8.1.8. 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).

8.2. International grants

8.2.1. GOSSPLE ERC Starting Grant

Participants: Xiao Bai, Marin Bertier, Antoine Boutet, Davide Frey, 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.

8.2.2. Transform Marie Curie Initial Training Network

Participants: 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

8.3. Visits (2009)

Ajoy Datta, University of Nevada, Las Vegas, USA, June 2009 (Rennes).

Anwittaman Datta, University of Singapore, June 2009 (Rennes).

Sarunas Girdzijauskas, SICS, Sweden, November 2009 (Rennes).

Anelise Munaretto, Federal Technological University of the Parana (UTFPR), Brazil, May-July 2009 (Saclay).

Ymir Vigfusson, IBM Haifa, Israel, November 2009 (Rennes).

9. Dissemination

9.1. Community animation

9.1.1. Leaderships and community service

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

A.-M. Kermarrec is a member of a CNRS group of experts on networking (*Comité d'experts réseaux*), a member of the steering committee of RESCOM (*pôle du GDR ASR du CNRS* gathering the French community interested in networking), and a member of the GDR Grid, peer-to-peer and parallelism.

9.1.2. Editorial boards, steering and program committees

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

DEBS'09 *3th ACM International Conference on Distributed Event-Based Systems*. Proceedings published by ACM. 6-9 July, Nashville.

CFIP'09 *Colloque Francophone sur l'ingénierie des Protocoles*. 12-15 october, Strasbourg.

SSS'09 *11th International Symposium on Stabilization, Safety and Security of Distributed Systems*, November 2009, Lyon

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

SSS'09: 11th Int. Symposium on Stabilization, Safety, and Security of Distributed Systems, Lyon, France, November 2009.

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

A.-M. Kermarrec is the Guest Editor with Maarten van Steen of the Computer Networks Elsevier Journal 53(13), Special Issue on Gossip-based networking, Computer Networks, 2009.

She served in the program committees for the following conferences:

EuroSys'10: *European Conference on Computer Systems*, Paris, France, April 2010.

PODC'09: *ACM Symposium on Principles of distributed computing*, Calgary, USA, August 2009.

ICDCS'09: *International Conference on Distributed Computing Systems*, Operating Systems and Middleware track, Montreal, Canada, June 2009.

IPDPS 2009: *IEEE International Parallel & Distributed Processing Symposium*, Miami, Florida, May 2009.

P2P'09 *IEEE Conference on Peer-to-Peer systems*, Seattle, USA, September 2009.

Infocom'09: *IEEE Conference on Computer Communications and Networking*, Rio De Janeiro, Brazil, Apr 2009.

SNS'09: *ACM Workshop on Social Network Systems*, Nuremberg, Germany, March 2009,

F. Le Fessant served in the program committees for the following conferences/workshops :

SSS'09: 11th Int. Symposium on Stabilization, Safety, and Security of Distributed Systems, Lyon, France, November, 2009.

HotP2P'09: *IPDPS Workshop on Hot Topics in P2P Systems*, Roma, May 2009.

DaMaP'09: *EDBT Workshop on Data Management in P2P Systems*, Saint Petersbourg, March 2009

CFSE'09 *Conférence Française en Systèmes d' Exploitation*, Toulouse, September 2009.

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

ISPS'09: 9th International Symposium on Programming and Systems, Alger, Algeria, May 2009.

ICDCS'09: 29th International Conference on Distributed Computing Systems, Montreal, Canada, June 2009.

SSS'09: 11th Int. Symposium on Stabilization, Safety, and Security of Distributed Systems, Lyon, France, November 2009.

M. Raynal was the program chair of the 13th Int'l Conference on Principles of Distributed Systems (OPODIS'09), 2009.

He was the International Liaison Co-chair if the 28th IEEE International conference on Distributed Computing Systems (ICDCS'09), 2009. Montréal, Canada.

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

COCOA'09: 3th International Conference on Combinatorial Optimization and Applications, 2009

ICDCS'09: 29th International Conference on Distributed Computing Systems, Montreal, Canada, June 2009.

MFCS'09: 34th International Symposium on Mathematical Foundations of Computer Science

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 Int'l Journal of Computer Systems Science and Engineering (since 1998)

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

PODC "ACM Conference on the Principles of Disitributed Computing"

ICDCS "Int'l Colloquium on Structural Information and Communication Complexity"

ICDCN "Int'l Conference on Distributed Computing and Networking"

Aline C. Viana served in the program committees for the following conferences/workshops :

WICON'10: *5th International Wireless Internet Conference*, Singapore, March 2010.

BWNCP'9: *1st IEEE International Workshop on Broadband Wireless Network Communications Performance*, in conjunction with IEEE IPCCC 2009, Phoenix, Arizona, USA. December 2009.

PIMRC'9: *IEEE Int'l Symposium on Personal, Indoor and Mobile Radio Communications*, Tokyo, Japan, October 2009.

AUTONOMICS'09: Cyprus, September 2009.

Algotel'09: *Algotel*, Carry Le Rouet, June 2009.

9.1.3. Evaluation committees, consulting

A.-M. Kermarrec served as a reviewer of the HAGGLE IP, and NETREFOUND EC-funded projects. She acted as a referee for the PhD committee of Ramses Morales, University of Illinois, Urbana Champain, USA. She served in the recruting committee of the "Full Professor/Associate Professor position in the field of Computer Communication and Networking"at Chalmers University, Sweden in 2009.

F. Le Fessant is a member of the Projet Committee for Free Software of the System@tic parisian cluster.

M. Raynal acted as a referee for the Swedish Research Council, the research agency of Austria and the Natural Sciences and Engineering Research Council of Canada (NSERC).

A. Carneiro Viana served as a reviewer for the PhD committee of Massimo Vecchio, IMT Institute for Advanced Studies, Italy.

9.2. 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 is heading the Master SSI (Security of Computing Systems) of the University of Rennes. He is also responsible of a Master's course (University of Bougie, Algeria) entitled "Distributed Algorithms and Systems".

M. Raynal gave a 10 hour lecture on distributed computing at UNAM, Mexico City in April 2009.

M. Raynal gave a tutorial at the IEEE Conference Dependable Systems and Networks (DSN), Lisbon, 2009.

In addition, **A. Boutet** and **N. Le Scouarnec** are teaching a few hours respectively at University of Rennes 1 and INSA, and **F. Bonnet, K. Huguenin, V. Leroy, G. Trédan** are Teaching Assistants (*moniteurs*) (INSA, University of Rennes 1, ENS Cachan).

9.3. Conferences, seminars, and invitations

9.3.1. Invited Talks

A.-M. Kermarrec and **Michel Raynal** were invited speakers at the METIS Spring School, Morocco, May 2009.

A.-M. Kermarrec was invited speaker at the International Symposium on Stabilization, Safety and Security of Distributed Systems (SSS 2009) in November 2009.

A.-M. Kermarrec was invited speaker at the International Conference on Principles of Distributed Systems (OPODIS 2009) in December 2009.

A.-M. Kermarrec was invited speaker at the Conférence Française en Systèmes d'Exploitation (CFSE 2009) in September 2009.

A.-M. Kermarrec will be 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** will be invited speakers at the Collège de France in 2010.

M. Raynal was invited speaker at the Workshop on Theoretical Aspects of Dynamic Distributed Systems (TADDS'09) in conjunction with DISC 2009. Elche (Spain), 2009.

M. Raynal was invited speaker at the Franco-Brazilian Colloquium on Advances and Challenges in Computer Science (COLIBRI'09), Bento Goncalves (Brazil), July 2009.

D. Frey was an invited speaker at the EDBT Summer School, Presqu'île de Giens, France, August/September 2009.

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

9.3.2. Seminars

D. Frey was invited to give a talk at DIS Universita' di Roma, La Sapienza, Italy, May 2009.

D. Frey was invited to give a talk at ENS-Cachan Antenne de Bretagne, France. November, 2009.

A.-M. Kermarrec was invited to give a talk at the Bell Labs Open days in Antwerp, October 2009.

A.-M. Kermarrec was invited to give a talk at the Bell Labs-INRIA workshop, Paris, October 2009.

A.-M. Kermarrec was invited to give a talk at Deutsch Telecom, Berlin, October 2009.

M. Raynal was invited to give a talk at IBM, Bangalore, January 2009.

M. Raynal was invited to give a talk at Puebla University, Mexico, April 2009.

M. Raynal was invited to give a talk at Fukuoka University, Japan, March 2009.

A. C. Viana was invited to give a talk at the FOKUS laboratory, Germany (March 2009).

A. C. Viana was invited to give a talk at the TKN laboratory of the TU-Berlin, Germany (May 2009).

9.3.3. Visits

K. Huguenin spent three months (May-July 2009) at Telefonica, Barcelona, Spain.

F. Le Fessant spent a week one in Sophia-Antipolis in the PLANETE team (Arnaud Legout).

E. Le Merrer spent one month (October 2009 at the University of Nevada, Las Vegas, USA.

V. Leroy spent three months (May-July 2009) at Yahoo! Research, Barcelona, Spain.

A. C. Viana spent two months (May-July 2009) at Telephonica, Barcelona, Spain.

A. C. Viana spent two weeks (April 2009) at National Laboratory for Scientific Computing (LNCC), Brazil.

A. C. Viana spent one week (April 2009) at Technological and Federal University of Parana, Brazil.

9.4. 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 a member of the 2009 INRIA Selection Committees for Senior Researcher permanent positions (DR2), and for Junior Researcher permanent positions (CR2) (INRIA Lille)

She is a member of the “bureau du CP” since November 2009.

Fabrice Le Fessant is in the Free Software Project Committee of Pôle de Compétitivité System@tic.

Aline C. Viana is a member of the “Commission développement technologique” of the INRIA SACLAY - ILE DE FRANCE SUD since 2007.

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", vol. 56, n^{o} 8, 2007, 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/DEDKCH.

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

[5] A. J. GANESH, A.-M. KERMARREC, E. L. MERRER, L. MASSOULIÉ. *Peer counting and sampling in overlay networks based on random walks*, in "Distributed Computing", vol. 20, n^{o} 4, 2007, 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", vol. 41, n^{o} 5, August 2007.

[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", vol. 61, n^{o} 1, 2007, 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", vol. 53, n^{o} 13, 2009, http://hal.inria.fr/inria-00436057/en/NOUS.

Year Publications

Articles in International Peer-Reviewed Journal

[10] Y. AFEK, E. GAFNI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *The k-simultaneous consensus problem*, in "Distributed Computing", 2009, http://hal.inria.fr/inria-00436496/en/MXUSIL.

[11] F. BONNET, M. RAYNAL. *Conditions for set agreement with an application to synchronous systems*, in "Journal of Computer Science and Technology", 2009, http://hal.inria.fr/inria-00436481/en/.

[12] J. CAO, M. RAYNAL, W. WU. *Eventual Clusterer: a Modular Approach to Designing Hierarchical Consensus Protocols in MANETs*, in "IEEE Transactions on Parallel and Distributed Systems", 2009, http://hal.inria.fr/inria-00436485/en/HK.

[13] 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/.

[14] A. CARNEIRO VIANA, A. ZIVIANI, R. FRIEDMAN. *Decoupling Data Dissemination from Mobile Sink's Trajectory in Wireless Sensor Networks*, in "IEEE Communications Letters", 2009, http://hal.inria.fr/inria-00429445/en/BRIL.

[15] 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", 2009, http://hal.inria.fr/inria-00436493/en/ES.

[16] E. GAFNI, A. MOSTEFAOUI, M. RAYNAL, C. TRAVERS. *From adaptive renaming to set agreement*, in "Theoretical Computer Science", vol. 410, n^{o} 14, 2009, p. 1328-1335, http://hal.archives-ouvertes.fr/hal-00432970/en/US.

[17] V. GRAMOLI, Y. VIGFUSSON, K. BIRMAN, A.-M. KERMARREC, R. VAN RENESSE. *Slicing distributed systems*, in "IEEE Transactions on Computers", vol. 59, n^{o} 11, 2009, http://hal.inria.fr/inria-00435175/en/USCH.

[18] D. IMBS, M. RAYNAL. *A note on atomicity: boosting Test-And-Set to solve Consensus*, in "Information Processing Letters", 2009, http://hal.inria.fr/inria-00436478/en/.

[19] Y. MOSES, M. RAYNAL. *Revisiting Simultaneous Consensus with Crash Failures*, in "Journal of Parallel and Distributed Computing", 2009, http://hal.inria.fr/inria-00436477/en/IL.

[20] A. MOSTEFAOUI, M. RAYNAL, G. TRÉDAN. *On the Fly Estimation of the Processes that Are Alive in an Asynchronous Message-Passing System*, in "IEEE Transactions on Parallel and Distributed Systems", vol. 20, n° 6, 2009, p. 778-787, http://hal.archives-ouvertes.fr/hal-00432964/en/.

[21] A. MOSTEFAOUI, M. RAYNAL, C. TRAVERS. *Narrowing Power vs Efficiency in Synchronous Set Agreement: Relationship, Algorithms and Lower Bound*, in "Theoretical Computer Science", 2009, http://hal.inria.fr/inria-00436490/en/IL.

[22] 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", vol. 53, n° 13, 2009, http://hal.inria.fr/inria-00436057/en/NOUS.

[23] G. TAN, S. JARVIS, A.-M. KERMARREC. *Connectivity-Guaranteed and Obstacle-Adaptive Deployment Schemes for Mobile Sensor Networks*, in "IEEE Transactions on Mobile Computing", 2009, http://hal.inria.fr/inria-00432129/en/GB.

Invited Conferences

[24] A. MOSTEFAOUI. *What Agreement Problems Owe Michel*, in "Distributed Computing, 23rd International Symposium, DISC 2009, Espagne Elche", I. KEIDAR (editor), Springer, 2009, 5, http://hal.archives-ouvertes.fr/hal-00432956/en/.

International Peer-Reviewed Conference/Proceedings

[25] X. BAI, M. BERTIER, R. GUERRAOUI, A.-M. KERMARREC. *Toward Personalized Peer-to-Peer Top-K Processing*, in "Second ACM Workshop on Social Network Systems (SNS), Allemagne Nuremberg", 2009, http://hal.inria.fr/inria-00430159/en/CH.

[26] R. BALDONI, S. BONOMI, A.-M. KERMARREC, M. RAYNAL. *Implementing a Register in a Dynamic Distributed System.*, in "International Conference on Distributed Computing Systems (ICDCS), Canada Montreal", 2009, http://hal.inria.fr/inria-00437923/en/IT.

[27] R. BALDONI, S. BONOMI, M. RAYNAL. *Joining a Distributed Shared Memory Computation in a Dynamic Distributed System*, in "7th Workshop on Software Technologies for Future Embedded and Ubiquitous Computing Systems (SEUS'09), États-Unis d'Amérique Newport Beach", 2009, http://hal.inria.fr/inria-00440395/en/IT.

[28] R. BALDONI, S. BONOMI, M. RAYNAL. *Regular Register: an Implementation in a Churn Prone Environment*, in "16th Colloquium on Structural Information and Communication Complexity (SIROCCO'09), Slovénie Piran", 2009, http://hal.inria.fr/inria-00440377/en/IT.

[29] O. BEAUMONT, L. EYRAUD-DUBOIS, H. REJEB, C. THRAVES. *Allocation of Clients to Multiple Servers on Large Scale Heterogeneous Platforms*, in "The Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), Chine Shenzhen", 2009, http://hal.inria.fr/inria-00413437/en/.

[30] S. BERNARD, F. LE FESSANT. *Optimizing peer-to-peer backup using lifetime estimations*, in "International Workshop on Data Management in Peer-to-Peer Systems, Russie st Petersburg", 2009, http://hal.inria.fr/inria-00432752/en/.

[31] M. BERTIER, Y. BUSNEL, A.-M. KERMARREC. *On Gossip and populations*, in "16th International Colloquium on Structural Information and Communication Complexity (SIROCCO 2009), Slovénie Piran", 2009, http://hal.inria.fr/inria-00437857/en/.

[32] M. BERTIER, R. GUERRAOUI, A.-M. KERMARREC, V. LEROY. *Toward Personalized Query Expansion*, in "Social Network Systems 2009, Allemagne Nuremberg", 2009, http://hal.inria.fr/inria-00368168/en/CH.

[33] F. BONNET, M. RAYNAL. *Looking for the Weakest Failure Detector for k-Set Agreement in Message-Passing Systems: Is Pkk the End of the Road?*, in "11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009), France Lyon", 2009, http://hal.inria.fr/inria-00436500/en/.

[34] F. BONNET, M. RAYNAL. *The Price of Anonymity: Optimal Consensus Despite Asynchrony, Crash and Anonymity*, in "23rd International Symposium on Distributed Computing., Espagne Elche", 2009, http://hal.inria.fr/inria-00436498/en/.

[35] M.-L. CHAMPEL, K. HUGUENIN, A.-M. KERMARREC, N. LE SCOUARNEC. *LTNC: Low Complexity Network Codes*, in "5th ACM International Conference on emerging Networking EXperiments and Technologies (CoNeXT), Student Workshop, Italie Rome", 2009, http://hal.inria.fr/inria-00429680/en/.

[36] M.-L. CHAMPEL, A.-M. KERMARREC, N. LE SCOUARNEC. *FoG: fighting the Achilles' heel of gossip protocols with fountain codes*, in "11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009), France Lyon", 2009, http://hal.inria.fr/inria-00429998/en/.

[37] M.-L. CHAMPEL, A.-M. KERMARREC, N. LE SCOUARNEC. *Phosphite : Incitation à la collaboration pour la vidéo à la demande en P2P*, in "AlgoTel, France Carry-Le-Rouet", A. CHAINTREAU, C. MAGNIEN (editors), 2009, http://hal.inria.fr/inria-00383172/en/.

[38] M.-L. CHAMPEL, A.-M. KERMARREC, N. LE SCOUARNEC. *Phosphite: Guaranteeing out-of-order download in P2P Video on Demand*, in "IEEE Ninth International Conference on Peer-to-Peer Computing (P2P'09), États-Unis d'Amérique Seattle", 2009, http://hal.inria.fr/inria-00429994/en/.

[39] P. FELBER, A.-M. KERMARREC, L. LEONINI, É. RIVIÈRE, S. VOULGARIS. *PULP: Un protocole epidémique hybride*, in "AlgoTel, France Carry-Le-Rouet", A. CHAINTREAU, C. MAGNIEN (editors), 2009, http://hal.inria.fr/inria-00383063/en/NONLCH.

[40] A. FERNÁNDEZ ANTA, M. MOSTEIRO, C. THRAVES CARO. *An Early-stopping Protocol for Computing Aggregate Functions in Sensor Networks*, in "Pacific Rim International Symposium on Dependable Computing, Chine Shanghai", 2009, http://hal.inria.fr/inria-00432380/en/ES.

[41] E. FLEURY, K. HUGUENIN, A.-M. KERMARREC. *Route in Mobile WSN and Get Self-Deployment for Free*, in "5th IEEE/ACM International Conference on Distributed Computing in Sensor Systems (DCOSS), États-Unis d'Amérique Marina del Rey", 2009, http://hal.inria.fr/inria-00397372/en/.

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

[43] D. FREY, R. GUERRAOUI, A.-M. KERMARREC, M. MONOD, V. QUÉMA. *Stretching Gossip with Live Streaming*, in "DSN 2009, Portugal Estoril", 2009, http://hal.inria.fr/inria-00436130/en/CH.

[44] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MONOD. *Brief Announcement: Towards Secured Distributed Polling in Social Networks*, in "23rd International Symposium on Distributed Computing (DISC), Espagne Elche", 2009, http://hal.inria.fr/inria-00429676/en/CH.

[45] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MONOD. *Decentralized Polling With Respectable Participants*, in "13th International Conference On Principles Of Distributed Systems (OPODIS), France Nimes", 2009, http://hal.inria.fr/inria-00429678/en/CH.

[46] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MONOD. *Short Paper: On Tracking Freeriders in Gossip Protocols*, in "9th IEEE International Conference on Peer to Peer Computing (P2P), États-Unis d'Amérique Seattle", 2009, http://hal.inria.fr/inria-00429682/en/CH.

[47] D. IMBS, M. RAYNAL. *A versatile STM protocol with invisible read operations that satisfies the virtual world consistency condition*, in "16th Colloquium on Structural Information and Communication Complexity (SIROCCO'09), Slovénie Piran", 2009, http://hal.inria.fr/inria-00440375/en/.

[48] D. IMBS, M. RAYNAL. *Help when needed, but no more: Efficient Read/Write Partial Snapshot*, in "23rd International Symposium on Distributed Computing (DISC'09), Espagne Elche", 2009, http://hal.inria.fr/inria-00440392/en/.

[49] D. IMBS, M. RAYNAL. *Provable STM Properties: Leveraging Clock and Locks to Favor Commit and Early Abort*, in "International Conference on Distributed Computing and Networking (ICDCN'09), Inde Hyderabad", 2009, http://hal.inria.fr/inria-00440361/en/.

[50] D. IMBS, M. RAYNAL. *Software Transactional Memories: an Approach for Multicore Programming*, in "10th International Conference on Parallel Computing Technologies (PaCT'09), Russie Novosibirsk", 2009, http://hal.inria.fr/inria-00440382/en/.

[51] D. IMBS, M. RAYNAL. *Visiting Gafni's Reduction Land: From the BG Simulation to the Extended BG Simulation*, in "11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009), France lyon", 2009, http://hal.inria.fr/inria-00440398/en/.

[52] Y. JARMA, G. KARBASCHI, M. DIAS DE AMORIM, F. BENBADIS, G. CHELIUS. *VAPS: Positioning with Spatial Constraints*, in "10th IEEE International Symposium on a "World of Wireless, Mobile and Multimedia networks" (WoWMoM 2009), Grèce Kos", IEEE, 2009, http://hal.inria.fr/inria-00396284/en/.

[53] G. KARBASCHI, A. CARNEIRO VIANA. *A Content-based Network Coding to Match Social Interest Similarities in Delay Tolerant Networks*, in "ExtremeCom Workshop, Suède Laponia", 2009, http://hal.inria.fr/inria-00429439/en/.

[54] G. KARBASCHI, A. CARNEIRO VIANA, S. MARTIN, K. AL AGHA. *On using network coding in multi-hop wireless networks*, in "IEEE PIMRC, Japon Tokyo", 2009, http://hal.inria.fr/inria-00429442/en/.

[55] A.-M. KERMARREC. *Challenges in personalizing and decentralizing the Web: an overview of GOSSPLE*, in "11th International Symposium on Stabilization, Safety and Security (SSS 2009), France Lyon", 2009, http://hal.inria.fr/inria-00435170/en/.

[56] A.-M. KERMARREC, E. LE MERRER, Y. LIU, G. SIMON. *Surfing Peer-to-Peer IPTV: Distributed Channel Switching*, in "Euro-Par, Pays-Bas Delft", LNCS (editor), 2009, http://hal.inria.fr/inria-00437873/en/.

[57] A.-M. KERMARREC, E. LE MERRER, B. SERICOLA, G. TRÉDAN. *Centralite du second ordre : Calcul distribue de l'importance de noeuds*, in "AlgoTel, France Carry-Le-Rouet", A. CHAINTREAU, C. MAGNIEN (editors), 2009, http://hal.inria.fr/inria-00384426/en/.

[58] A.-M. KERMARREC, A. MOSTEFAOUI, M. RAYNAL, G. TRÉDAN, A. CARNEIRO VIANA. *Large-scale networked systems: from anarchy to geometric self-structuring*, in "10th International Conference on Distributed Computing and Networking (ICDCN), Inde Hyderabad", 2009, http://hal.inria.fr/inria-00334355/en/.

[59] A.-M. KERMARREC, A. MOSTEFAOUI, M. RAYNAL, G. TRÉDAN, A. C. VIANA. *Large-Scale Networked Systems: From Anarchy to Geometric Self-structuring*, in "10th International Conference on Distributed Computing and Networking (ICDCN 2009), Inde Hyderabad", K. K. VIJAY K. GARG (editor), Springer, 2009, p. 25-36, http://hal.archives-ouvertes.fr/hal-00432967/en/.

[60] A.-M. KERMARREC, A. PACE, V. QUÉMA, V. SCHIAVONI. *NAT-resilient Gossip Peer Sampling*, in "International Conference on Distributed Computing Systems (ICDCS), Canada Montreal", 2009, http://hal.inria.fr/inria-00435180/en/.

[61] S. LE BLOND, F. LE FESSANT, E. LE MERRER. *Finding Good Partners in Availability-aware P2P Networks*, in "International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS'09), France Lyon", 2009, http://hal.inria.fr/inria-00432741/en/.

[62] R. MACÊDO, M. RAYNAL. *Adding Dynamicity to the Uncertainty that Characterizes Distributed Systems: Challenges Ahead*, in "Franco-Brazilian Colloquium on Advances and Challenges in Computer Science (COLIBRI'09), Brésil Bento Goncalves", 2009, http://hal.inria.fr/inria-00440386/en/BR.

[63] A. PAPADIMITRIOU, F. LE FESSANT, A. CARNEIRO VIANA, C. SENGUL. *Cryptographic Protocols to Fight Sinkhole Attacks on Tree-based Routing in Wireless Sensor Networks*, in "5th Workshop on Secure Network Protocols (NPSec), États-Unis d'Amérique Princeton", 2009, http://hal.inria.fr/inria-00429440/en/.

[64] M. RAYNAL. *Shared Memory Synchronization in Presence of Failures: an Exercise-based Introduction for the Sophomore*, in "International Conference on Complex, Intelligent and Software Intensive Systems (CISIS'09), Japon Fukuoka", 2009, http://hal.inria.fr/inria-00440367/en/.

[65] M. RAYNAL. *Software Transactional Memory: What? Why? How? A New Challenge?*, in "Franco-Brazilian Colloquium on Advances and Challenges in Computer Science (COLIBRI'09), Brésil Bento Goncalves", 2009, http://hal.inria.fr/inria-00440388/en/.

[66] M. H. REHMANI, A. CARNEIRO VIANA, H. KHALIFE, S. FDIDA. *Adaptive and Occupancy-based Channel Selection for unreliable Cognitive Radio Networks*, in "Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL) 2009, France Carry Le Rouet", 2009, http://hal.inria.fr/inria-00441154/en/.

[67] C. SENGUL, A. CARNEIRO VIANA, R. FRIEDMAN, M. BERTIER, A.-M. KERMARREC. *The Importance of Being Adaptive for Forwarding*, in "ExtremeCom Workshop, Suède Laponia", 2009, http://hal.inria.fr/inria-00429438/en/IL.

[68] G. TAN, M. BERTIER, A.-M. KERMARREC. *Convex Partition of Sensor Networks and Its Use in Virtual Coordinate Geographic Routing*, in "INFOCOM 2009, Brésil Rio de Janeiro", 2009, http://hal.inria.fr/inria-00430145/en/.

[69] G. TAN, M. BERTIER, A.-M. KERMARREC. *Visibility-Graph-based Shortest-Path Geographic Routing in Sensor Networks*, in "INFOCOM 2009, Brésil Rio de Janeiro", 2009, http://hal.inria.fr/inria-00430147/en/.

National Peer-Reviewed Conference/Proceedings

[70] S. LE BLOND, F. LE FESSANT, E. LE MERRER. *Choix de partenaires en p2p suivant des critères de disponibilité*, in "conférence francaise sur les systèmes d'exploitation, France Toulouse", 2009, http://hal.inria.fr/inria-00432747/en/.

Workshops without Proceedings

[71] M. RAYNAL. *How to implement a shared memory in a dynamic system? Which are the constraints?*, in "International Workshop on Theoretical Aspects of Dynamic Distributed Systems (TADDS'09), Espagne Elche", 2009, http://hal.inria.fr/inria-00440390/en/.

Research Reports

[72] M. BERTIER, D. FREY, R. GUERRAOUI, A.-M. KERMARREC, V. LEROY. *Personalized Web Search by Gossiping with Unknown Social Acquaintances*, INRIA, 2009, http://hal.inria.fr/inria-00368137/en/, RR-6878, Rapport de rechercheCH.

[73] F. BONNET, M. RAYNAL. *A simple proof of the necessity of the failure detector Σ to implement a register in asynchronous message-passing systems*, INRIA, 2009, http://hal.inria.fr/inria-00392450/en/, PI 1932, Rapport de recherche.

[74] F. BONNET, M. RAYNAL. *Consensus in Anonymous Distributed Systems: Is There a Weakest Failure Detector?*, INRIA, 2009, http://hal.inria.fr/inria-00426674/en/, Rapport de recherche.

[75] F. BONNET, M. RAYNAL. *Early Consensus in Message-passing Systems Enriched with a Perfect Failure Detector and its Application in the Theta Model*, INRIA, 2009, http://hal.inria.fr/inria-00425127/en/, Rapport de recherche.

[76] F. BONNET, M. RAYNAL. *Looking for the Weakest Failure Detector for k-Set Agreement in Message-passing Systems: Is Π_k the End of the Road?*, INRIA, 2009, http://hal.inria.fr/inria-00384993/en/, PI 1929, Rapport de recherche.

[77] M.-L. CHAMPEL, K. HUGUENIN, A.-M. KERMARREC, N. LE SCOUARNEC. *LT Network Codes*, INRIA, 2009, http://hal.inria.fr/inria-00416671/en/, RR-7035, Rapport de recherche.

[78] A. DATTA, A.-M. KERMARREC, L. LARMORE, E. LE MERRER. *Star: A Stable and Robust Membership Protocol*, INRIA, 2009, http://hal.inria.fr/inria-00399546/en/, Rapport de rechercheUS.

[79] E. FLEURY, K. HUGUENIN, A.-M. KERMARREC. *Route in Mobile WSN and Get Self-Deployment for Free*, INRIA, 2009, http://hal.inria.fr/inria-00357240/en/, RR-6819, Rapport de recherche.

[80] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MAXIME MONOD. *Decentralized Polling With Respectable Participants*, INRIA, 2009, http://hal.inria.fr/inria-00371708/en/, RR-6890, Rapport de rechercheCH.

[81] R. GUERRAOUI, K. HUGUENIN, A.-M. KERMARREC, M. MAXIME MONOD. *LiFT: Lightweight Freerider-Tracking Protocol*, INRIA, 2009, http://hal.inria.fr/inria-00379408/en/, RR-6913, Rapport de rechercheCH.

[82] K. HUGUENIN, A.-M. KERMARREC, V. RAI, M. VAN STEEN. *Designing a Tit-for-Tat Based Peer-to-Peer Video-on-Demand System*, INRIA, 2009, http://hal.inria.fr/inria-00416208/en/, RR-7034, Rapport de rechercheNL.

[83] D. IMBS, M. RAYNAL. *A versatile STM protocol with invisible read operations that satisfies the virtual world consistency condition*, INRIA, 2009, http://hal.inria.fr/inria-00362844/en/, PI 1923, Rapport de recherche.

[84] D. IMBS, M. RAYNAL. *Visiting Gafni's Reduction Land: from the BG Simulation to the Extended BG Simulation*, INRIA, 2009, http://hal.inria.fr/inria-00389682/en/, PI 1931, Rapport de recherche.

[85] G. KARBASCHI, A. CARNEIRO VIANA, S. MARTIN, K. AL AGHA. *On Delay Fairness for Multiple Network Coding Transmissions*, INRIA, 2009, http://hal.inria.fr/inria-00399561/en/, RR-6972, Rapport de recherche.

[86] A.-M. KERMARREC, E. LE MERRER, B. SERICOLA, G. TRÉDAN. *Centralité du second ordre : Calcul distribué de l'importance de noeuds dans un réseau complexe*, INRIA, 2009, http://hal.inria.fr/inria-00355667/en/, RR-6809, Rapport de recherche.

[87] A.-M. KERMARREC, C. THRAVES. *Extending the Analysis for Peer Sampling Service: It Chooses the Partial Views Uniformly at Random*, INRIA, 2009, http://hal.inria.fr/inria-00407516/en/, Rapport de recherche.

[88] S. LE BLOND, F. LE FESSANT, E. LE MERRER. *Finding Good Partners in Availability-aware P2P Networks*, INRIA, 2009, http://hal.inria.fr/inria-00352529/en/, RR-6795, Rapport de recherche.

[89] S. MAAG, A. CARNEIRO VIANA, F. ZAIDI. *One step forward: Linking Wireless Self-Organizing Networks Validation Techniques with Formal Testing approaches*, INRIA, 2009, http://hal.inria.fr/inria-00359569/en/, RR-6817, Rapport de recherche.

[90] A. PAPADIMITRIOU, F. LE FESSANT, A. CARNEIRO VIANA, C. SENGUL. *Fighting Sinkhole Attacks in Tree-based Routing Topologies*, INRIA, 2009, http://hal.inria.fr/inria-00355873/en/, RR-6811, Rapport de recherche.

[91] M. H. REHMANI, A. CARNEIRO VIANA, H. KHALIFE, S. FDIDA. *Toward Reliable Contention-aware Data Dissemination in Multi-hop Cognitive Radio Ad Hoc Networks*, INRIA, 2009, http://hal.inria.fr/inria-00441892/en/, RT-375, Rapport Technique.

[92] C. SENGUL, A. CARNEIRO VIANA, R. FRIEDMAN, M. BERTIER, A.-M. KERMARREC. *Adaptive Forwarding to Match Mobility Characteristics in Delay Tolerant Networks*, INRIA, 2009, http://hal.inria.fr/inria-00356601/en/, RR-6816, Rapport de rechercheIL.

[93] M. VECCHIO, A. CARNEIRO VIANA, A. ZIVIANI, R. FRIEDMAN. *Proactive Data Dissemination in Wireless Sensor Networks with Uncontrolled Sink Mobility*, INRIA, 2009, http://hal.inria.fr/inria-00357655/en/, RR-6820, Rapport de rechercheBRILIT.

Other Publications

[94] A.-M. KERMARREC, E. LE MERRER, Y. LIU, G. SIMON. *Surfing Peer-to-Peer IPTV: Distributed Channel Switching*, 2009, http://hal.inria.fr/inria-00363927/en/, RR-6857.

References in notes

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

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

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

[98] 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.

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

[100] 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.

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

[102] F. LE FESSANT, C. SENGUL, A.-M. KERMARREC. *Pace-Maker: Tracking peer availability in large networks*, INRIA, 2008, http://hal.inria.fr/inria-00305620/en/, RR-6594, Rapport de recherche.

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

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

[105] 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.

[106] 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.

[107] 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.

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