INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team ASAP

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

Rennes - Futurs

THEME COM

Table of contents

1. Team **1**
2. Overall Objectives **1**
2.1. General objectives 1
2.1.1. A challenging new setting 2
2.1.2. Mastering uncertainty in distributed computing 2
2.2. Models and abstractions for large-scale distributed computing 3
2.2.1. Distributed computability. 3
2.2.2. Distributed computing abstractions. 3
2.3. Resource management in peer to peer overlays 4
2.4. Peer to peer computing and sensor networks 4
3. Scientific Foundations **5**
3.1. Introduction 5
3.2. Models and abstractions of large-scale dynamic systems 5
3.3. Peer to peer overlay networks 5
3.4. Epidemic protocols 6
4. Application Domains **6**
4.1. Panorama 6
4.2. Resource management in Internet-based applications 6
4.3. Sensor-based applications 7
5. Software **7**
5.1. Palabre 7
5.2. Peer to peer systems simulators 7
6. New Results **8**
6.1. Panorama 8
6.2. Models and abstractions: dealing with dynamics 8
6.2.1. Leader Election 8
6.2.2. Set Agreement in synchronous systems 9
6.2.3. Agreement: what is possible and what is not possible 10
6.2.4. Decentralized estimation size algorithms 10
6.2.5. Core Persistence in peer to peer Systems: Relating size to lifetime 11
6.3. Resource management in peer to peer systems 11
6.3.1. Querying peer to peer systems 11
6.3.2. Decentralized Content-based Publish and Subscribe 12
6.3.3. Palabre: a Peer to-peer back-up system 13
6.3.4. Distributed Ordered Slicing in Large-Scale Systems 13
6.3.5. Adaptive replica placement strategies 14
6.3.6. Combining structured and unstructured peer to peer networks 14
6.4. Peer to peer sensor networks 14
7. Contracts and Grants with Industry **15**
7.1. France Telecom 15
7.2. Advestigo 15
8. Other Grants and Activities **16**
8.1. National grants 16
8.1.1. ANR MD ALPAGE 16
8.1.2. RNRT project SVP 16
8.1.3. ARC Inria Recall. 16
8.1.4. Academic collaborations 16
8.2. International grants 16
8.2.1. The ReSIST european project 16

8.2.2. Epi-Net associated team with Vrije Universiteit, Amsterdam, NL 17
8.2.3. Collaboration with Madrid, Spain 17
8.2.4. PAI Mexique 18
8.2.5. Collaboration with Univeristy of Illinois, Urbana Champain 18
8.2.6. PAI Hong Kong 18
8.3. Visits (2006-2007) 19
9. Dissemination **19**
9.1. Community animation 19
9.1.1. Leaderships, Steering Committees and community service 19
9.1.2. Editorial boards, steering and program committees 19
9.1.3. Evaluation committees, consulting 21
9.2. Academic teaching 21
9.3. Conferences, seminars, and invitations 22
9.4. Administrative responsibilities 22
10. Bibliography **22**

1. Team

ASAP is a bi-localized project-team, in the process of creation, at INRIA-*Rennes (*IRISA*) and* INRIA-*Futurs (Saclay). The project proposal has been presented at and approved by the board committee of* INRIA-*Rennes in June 2006 and* INRIA-*Futurs in September 2006.*

Head of project-team

Anne-Marie Kermarrec [Research Director INRIA (DR), HdR]

Administrative assistants

Fabienne Cuyollaa [Technical Assistant (TR) INRIA (from April 2006)]
Angelina Lemaître [Secretary (SAR) INRIA (until October 2006)]
Caroline Ollivier [Secretary (SAR) INRIA (since October 2006)]

Staff members Inria

Fabrice Le Fessant [Research Associate (CR) INRIA]
Aline Carneiro Viana [Research Associate (CR) INRIA (since October 2006)]

Staff members University Rennes 1

Achour Mostefaoui [Associate Professor (MdC), HdR]
Michel Raynal [Professor (Pr), HdR]

Staff members Insa de Rennes

Marin Bertier [Assistant Professor (MdC)]

PhD students

François Bonnet [ENS CACHAN Grant (since October 2006)]
Yann Busnel [MENRT/ENS CACHAN Cachan Grant]
Vincent Gramoli [MENRT Grant]
Erwan Le Merrer [Cifre France Telecom industrial Grant]
Étienne Rivière [MENRT Grant]
Corentin Travers [MENRT Grant]
Gilles Trédan [MENRT Grant (since October 2006)]

Post-doctoral fellow

Aline Carneiro Viana [Post-Doc INRIA (until September 2006)]

Student intern

Yoong Kim [Master student UNIVERSITY RENNES 1, February-July 2006]
Balasubramaneyan Maniymaran [Internship INRIA, Ph.D student at McGill University, Canada, May-July 2006]
Gilles Trédan [Master student UNIVERSITY RENNES 1 February-July 2006]

Visitors

Antonio Fernandez [Juan Carlos University, Madrid, Spain, from June 1st till August 15]
Mark Jelasity [UNIVERSITY RENNES 1 Grant/University of Bologna, Italy, from March 1st till April 15]
Oliver Theel [University of Oldenburg, Germany, May-June 2006]

2. Overall Objectives

2.1. General objectives

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

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

2.1.1. A challenging new setting

Distributed computing was born in the late seventies when people started taking into account the intrinsic characteristics of physically distributed systems. The field then emerged as a specialized research area distinct from networks, operating systems and parallelism. Its birth certificate is usually considered as the publication in 1978 of Lamport's most celebrated paper "*Time, clocks and the ordering of events in a distributed system*" [57] (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 domain 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 operated 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 to acknowledge 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 intiative 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, sensors, ...). 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.

2.1.2. Mastering uncertainty in distributed computing

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

- **Fully decentralized model**: a fully decentralized system does not rely on any central entity to control the system. Participating entities may act both as clients and servers. The number of potential servers thus increases linearly with the size of the system, avoiding the performance bottleneck imposed by the presence of servers in traditional distributed systems. Such systems are therefore naturally protected from failures since there is no single point of failure and many services are naturally replicated.
- **Self-organizing capabilities**: participating entities are by essence highly dynamic as they might be disconnected, mobile or faulty. The system should be able to handle such dynamic behaviors and get automatically reorganized to face entity arrival and departure.

- **Local system knowledge**: individual entities behavior is based on a restricted knowledge of the system. 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 algorithms 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**: to deal with large-scale and dynamic behavior, 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. Based on this, 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.

2.2. Models and abstractions for large-scale distributed computing

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

2.2.1. Distributed computability.

Among the fundamental theoretical results of distributed computing, there is a list of problems (e.g. consensus or non-blocking atomic commit) that have been proved to have no deterministic solution in asynchronous distributed computing systems prone to failures. In order such a problem becomes 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 parsed).

2.2.2. Distributed computing abstractions.

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

2.3. Resource management in peer to peer overlays

Managing resources on a large-scale whether they are computing resources, data, events, bandwidth, requires some fully decentralized solutions. 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 an overlay network, defining a logical network on top of an existing networking infrastructure. The system automatically and dynamically adapts to frequent peer arrival and departure. In practice, two main classes have been designed: structured overlay networks rely on a name structure and map object keys to overlay nodes. They provide a distributed hash table functionality (DHT). While structured peer to peer systems initially dominated the academic research, their exact-match interface limits their flexibility and use for various applications, notably when it comes to non-exact information retrieval. At the other end of the spectrum, 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 the area of overlay networks, our approach is original for the following reasons.

- First of all, we step away from the traditional approaches consisting in creating overlay networks based solely on randomization. Instead we are focusing on creating overlays taking into account applications 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 if possible for a better performance. In order to purchase this goal, we strongly believe that it is not possible to rely on a generic framework applicable to all potential large-scale platforms (as Internet, grids or sensor networks) applications. Instead, a large-scale system is an environment where constraints are imposed by the (potentially limited) resources of the participating entities. However, this does not prevent a service specifically designed for an application domain to be applied in another context. Therefore, we tightly couple the design of distributed systems to application environments.
- Second, we strongly believe in weakly structured peer to peer systems, and most of our projects rely on epidemic-based unstructured overlay networks. Epidemic communication models have recently started to be explored as a general paradigm to build and maintain unstructured overlay networks. More specifically, they have shown to provide a scalable way of implementing and maintaining highly dynamic unstructured overlays in which nodes can frequently join and leave. The basic property of these epidemic protocols is that periodically, each peer exchanges information with some other peers selected from a local list of neighbors. One important observation from [56] is that epidemic algorithms cope very well with high dynamicity.
- 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 developing our own platform in the context of a peer to peer collaborative backup platform we are currently building. Similarly, we leverage the models and abstractions defined in the first theme of ASAP to provide guarantees and analysis of the protocols we develop in this area.

2.4. Peer to peer computing and sensor networks

In this area, recently initiated, we are investigating the use of peer to peer algorithms in sensor network systems. At the moment, this research area is essentially studied by the network community although many of these problematics are similar to distributed computing ones, like information propagation, resource discovery, *etc.* Sensors and peer to peer networks present many similarities that we plan to leverage. Scale and dynamicity are among the most striking similarities between the two types of networks. 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.*. We identified peer to peer solutions as being

also relevant for sensor networks, providing us with another application domain. However, sensor network specificities imply major adaptations of Internet-based peer to peer algorithms. More specifically, the fact that the neighbourhood of a node is entirely determined by the physical network topology, and the necessity to take into account the energy consumption, have strong consequences on the algorithm design. The broadcast property of the radio communication of the sensor has also a major impact on algorithm behaviour. This new research area allows us to widen our application domain with specific applications, but above all to vary some major properties of our target system and therefore generalize our work on fully decentralized algorithms.

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 sensor networks. We have chosen to 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, and a lot of research in that domain is focused on fundamental questions: what can be computed in such models, 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 (e.g.,what can be computed when failure detectors are combined with conditions on process input values). Another part is devoted to model equivalence (e.g.what can be computed with a given class of failure detectors, and to which synchronization primitives is equivalent a given failure class). 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, researchers in that domain consider that nodes can enter and leave the system, but do not provide an abstract characterization (based on properties of computation instead of description of possible behaviours) [58], [50], [51]. This shows that finding dynamics distributed computing models is today a "holy grail", whose discovery would allow a better understanding of the deep 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 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 peer arrivals and departures. Finally, individual peer behaviour is based on a local knowledge of the system. Yet the system converges toward global properties.

A peer to peer overlay network logically connect peers on top of a network infrastrcuture to form an overlay network. Two main classes of such overlays dominate, structured and unstructured. The differences relate to the choice of the neighbors in the overlay, and to 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 whether it is IP, a wireless or a sensor network. Such an overlay might form a structured overlay network [59], [60], [61] following a given topology or an unstructured network [55], [54], [62] 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 [53], [52], are also consistently used in our research. In the context of distributed systems, epidemic protocols are mainly used to create overlay networks and 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 throughout the system. Epidemic algorithms have proven efficient to build and maintain scalable distributed systems in the context of many applications such as broadcasting [52], monitoring, resource management, search, and more generally in building unstructured peer to peer networks.

4. Application Domains

4.1. Panorama

Keywords: *Scientific computing*, *Sensor networks*, *cooperative applications*, *large-scale computing*, *voice on IP*.

The results of the research targeted in ASAP span over a wide range of application areas ranging from Internet-based applications, Grid computing, and sensor networks. 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 plan to 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 able to potentially 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 actual 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 falling in this denomination, we are especially interested in collaborative storage systems, resource discovery and allocation in grid-like environments, and large-scale content distribution and indexing. Core functionalities of such applications are search, notification and dissemination. We will discuss in more details in the next sections the particular case of a peer to peer backup system we are currently developing.

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 systems, simply because participating entities are 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 on human 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: health medical 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. Palabre

Keywords: *Peer to peer backup system*, *distributed storage systems*, *open-source software*.

Participants: Fabrice Le Fessant, Anne-Marie Kermarrec.

Contact: Fabrice Le Fessant

Licence: GPL

Status: under development

Palabre 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 clients. This work is done in tight collaboration with Laurent Viennot and Anh-Tuan Gai (GYROWEB INRIA Rocquencourt). We are currently developing 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 already able to incrementally backup files on its local hard disk. We are now working on inter-clients authentification to allow distributed backups. The prototype is expected to be released early 2007 as an open-source project for external contributions.

5.2. Peer to peer systems simulators

Keywords: *Peer to peer*, *simulation*, *unstructured overlays*.

Participants: Anne-Marie Kermarrec, Erwan Le Merrer, Etienne Rivière.

Contact: Anne-Marie Kermarrec

Licence: Not defined yet

Status: Under development

Several simulators were developed in Java to evaluate the proposed peer to peer systems. The SizeWalker simulator provides a generic framework to simulate unstructured peer to peer overlays. The simulator enables to set the way the unstructured peer to peer overlay is built, as well as the associated counting algorithm. This simulator has been used to evaluate the SizeWalker algorithm as well as two competitors. Second, we developed a large scale simulator for VoroNet, that can handle up to millions of nodes. This simulator is implementing both the protocol, and a set of tools to examine the behaviour of the system under various workloads or nodes behaviours.

The Sub-2-Sub workload generator software has been developed in collaboration with Marteen van Steen, Vrije Universiteit in Amsterdam, and provides a workload generator for comprehensive publish-and-subscribe systems evaluation and comparison. This workload generator is currently used to evaluate Sub-2-Sub system [47], and will be used to evaluate several existing peer to peer approaches.

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 sensor networks. We focus our research on two main areas: resource management and dissemination. We believe 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. It is organized along the following themes:

1. Models and abstractions in large-scale systems,
2. Resource management in large-scale dynamic systems,
3. Data collection and propagation in sensor networks systems.

For each of these themes, we detail the results we obtained in 2006.

6.2. Models and abstractions: dealing with dynamics

Keywords: *Leader election*, *asynchronous message-passing systems*, *decentralized system size estimation*, *distributed shared memory systems*, *failure detectors*, *failure resilience*, *persistence*, *random walks*, *set-agreement*, *synchronous systems*.

6.2.1. Leader Election

Participants: Achour Mostefaoui, Michel Raynal, Corentin Travers.

This work considers the eventual leader election problem in asynchronous message-passing systems where an arbitrary number t of processes can crash (t<n, where n is the total number of processes). It considers weak assumptions both on the initial knowledge of the processes and on the network behavior. More precisely, initially, a process knows only its identity and the fact that the process identities are different and totally ordered (it knows neither n nor t). Two eventual leader election protocol and a lower bound are presented.

The first protocol assumes that a process also knows the lower bound α on the number of processes that do not crash. This protocol requires the following behavioral properties from the underlying network: the graph made up of the correct processes and fair lossy links is strongly connected, and there is a correct process connected to $t - f$ other correct processes (where f is the actual number of crashes in the considered run) through eventually timely paths (paths made up of correct processes and eventually timely links). This protocol is not communication-efficient in the sense that each correct process has to send messages forever.

The second protocol is communication-efficient: after some time, only the final common leader has to send messages forever. This protocol does not require the processes to know alpha, but requires stronger properties from the underlying network: each pair of correct processes has to be connected by fair lossy links (one in each direction), and there is a correct process whose output links to the rest of correct processes have to be eventually timely. The lower bound result shows that this a necessary requirement. This protocol enjoys also the property that each message is made up of several fields, each of which taking values from a finite domain.

On the other side we tackeled the problem of electing a leader in a shared memory distributed system. While protocols that elect an eventual common leader in asynchonous message-passing systems have been proposed, no such protocol has been proposed for the shared memory communication model to our knowledge. This work presents a leader election protocol suited to the shared memory model. In addition to its design simplicity, the proposed protocol has two noteworthy properties. It does not use timers, and it is optimal with respect to the number of processes that have to write forever the shared memory: a single process has to do it (namely, the leader that is eventually elected). Among the many possible uses of such a leader protocol, one is Lamport's Paxos protocol. Paxos is an asynchronous consensus algorithm that relies on an underlying eventual leader abstraction. As recently, several versions of Paxos have been designed for asynchronous shared memory systems (the shared memory being an abstraction of a physically shared memory or a set of commodity disks that can be read and written by the processes), the proposed leader protocol makes Paxos effective in these systems.

6.2.2. Set Agreement in synchronous systems

Participants: Achour Mostefaoui, Michel Raynal, Corentin Travers.

The k-set agreement problem is a generalization of the consensus problem: considering a system made up of n processes where each process proposes a value, each non-faulty process has to decide a value such that a decided value is a proposed value, and no more than k different values are decided. It has recently be shown that, in the crash failure model, min(f/k +2,t/k +1) is a lower bound on the number of rounds for the non-faulty processes to decide (where t is an upper bound on the number of process crashes, and f (0 <= f<= t) is he actual number of crashes). This work considers the k-set agreement problem in synchronous systems where up to t<n/2 processes can experience general omission failures (i.e., a process can crash or omit sending or receiving messages). It first introduces a new property, called "strong termination". This property is concerned with the processes deciding. It is satisfied if, not only every non-faulty process, but any process that neither crashes nor commits receive omission failures decides. We presented a k-set agreement protocol that enjoys the following features. First, it is strongly terminating (to our knowledge, it is the first agreement protocol to satisfy this property, whatever the failure model considered). Then, it is "early deciding and stopping" in the sense that a process that either is non-faulty or commits only send omission failures decides and halts by round $\min(f/k + 2, t/k + 1)$. To our knowledge, this is the first early deciding k-set agreement protocol for the general omission failure model. Moreover, the protocol provides also the following additional "early stopping" properties.

On the other side, we investigated the use of additional synchronization messages in round-based message-passing synchronous systems. It first presents a synchronous computation model allowing a process to send such messages. The difference with respect to the traditional round-based synchronous model lies in the sending phase, where a process can first send a data message to each other process, and then, without a break, a synchronization message (their sendings can be pipelined). This model is suited to the class of local area networks where communication channels are reliable. (It is not for networks where unreliable communication requires message retransmission.) To illustrate the model, the study presents a uniform consensus algorithm suited to this model. This algorithm, based on the rotating coordinator paradigm, allows the processes to decide in at most f+1 rounds where f is the actual number of processes that crash in the corresponding run. (This improves the f+2 lower bound of the traditional synchronous model.) In addition to its efficiency, the algorithm enjoys another first class property, namely, design simplicity. The work focuses also on lower bound results, and shows that any uniform consensus algorithm designed for the proposed model, requires at least f+1 rounds in the worst case. The proposed algorithm is consequently optimal. In that sense the approach has to be

seen as an investigation of both the power and the limit of adding synchronization messages to synchronous systems built on top of local networks with reliable communication.

6.2.3. Agreement: what is possible and what is not possible

Participants: Achour Mostefaoui, Michel Raynal, Corentin Travers.

We addressed the problem of solving a task $T = (T_1, ...T_m)$, in which a processor returns in an arbitrary one of m-simultaneous consensus subtasks $T_1, ...T_m$. Processor p_i submits to T an input vector of proposals $(prop - i1, ..., prop - im)$, one entry per subtask, and outputs form just one subtask $T_x l$, a pair $(l, prop - j)$ for some j. All processors that output at the same output the same proposal. Let d be a bound on the number of distinct input vectors that may be submitted to T. A wait-free algorithm regardless of the number of processors solves T provided m<= d is presented. What is the power of $T = (T_1, ..., T_m)$ when given as a subroutine, to be used by any number of processors with any number of input vectors? Obviously, T solves m-set consensus since each processor p_i can submit the vector $(id_i, id_i, ...id_i)$, but can m-set consensus solve T? We show it does, and thus simultaneous consensus is a new characterization of set-consensus. Finally, what if each T_j is just a binary-consensus rather than consensus? Then we get the novel problem that was recently introduced of the Committee-Coordination. We show that a task that returns one of m simultaneous binary-consensus subtasks when used by n processors is equivalent to (n, m)-set consensus. Thus, while set-consensus unlike consensus, has no binary version, now that we characterize m-set consensus through simultaneous consensus $T = (T_1, ..., T_m)$, the notion of binary-set-consensus is well defined. We have then showed that binary-set-consenusus is equivalent to set consensus as it was with consensus.

In a second work, we looked for the weakest failure detector for wait-free set agreement. Asynchronous failure detector-based set agreement algorithms proposed so far assume that all the processes participate in the algorithm. This means that (at least) the processes that do not crash propose a value and consequently execute the algorithm. It follows that these algorithms can block forever (preventing the correct processes from terminating) when there are correct processes that do not participate in the algorithm. This work investigates the wait-free set agreement problem, i.e., the case where the correct participating processes have to decide a value whatever the behaviour of the other processes (i.e., the processes that crash and the processes that are correct but do not participate in the algorithm). The approach presents a wait-free set agreement algorithm. This algorithm is based on a leader failure detector class that takes into account the notion of participating processes. Interestingly, this algorithm enjoys a first class property, namely, design simplicity.

Finally, we introduce a new problem, the (b, n)-Committee Decision Problem (CD) - a generalization of the consensus problem. While set agreement generalizes consensus in terms of the number of decisions allowed, the CD problem generalizes consensus in the sense of considering many instances of consensus and requiring a processor to decide in at least one instance. In more detail, in the CD problem each one of a set of n processes has a (possibly distinct) value to propose to each one of a set of b consensus problems, which we call "committee". Yet a process has to choose only one of the b committee, and decide a value for at least one of these committees, such that all processes deciding for the same committee decide the same value. We study the CD problem in the context of a wait-free distributed system and analyze it using a combination of distributed algorithmic and topological techniques, introducing a novel reduction technique.

We use the reduction technique to obtain the following results. We show that the (2,3)-CD problem is equivalent to the musical benches problem and both are equivalent to (2,3)-set agreement, closing an open question left there. Thus, all three problems are wait-free unsolvable in a read/write shared memory system, and they are all solvable if the system is enriched with objects capable of solving (2,3)-set agreement. While the previous proof of the impossibility of musical benches was based on the Borsuk-Ulam (BU) Theorem, it now relies on Sperner's Lemma, opening intriguing questions about the relation between BU and distributed computing tasks.

6.2.4. Decentralized estimation size algorithms

Participants: Anne-Marie Kermarrec, Erwan Le Merrer.

Peer to peer systems are characterized by the fact that peers only have a limited knowledge of the system. Therefore, no peer is aware of the global membership and able to compute the system size. We propose two algorithms, *the random tour* and the *sample and collide* algorithms to estimate the system size. These algorithms are fully decentralized, and based on random walks. The basic idea of the *Random tour* algorithm is that a peer initiates a random walk in the system. The message associated to this random walk, carries a density information (number of neighbor) across the network and provides an accurate estimation when it returns to the process initiator. In the *sample and collide* approach, an initiator iterates on a sampling approach. the estimation of the system size is based on the redundancy observed in the samples. In collaboration with Laurent Massoulié (Microsoft Research, UK), we provided a theoretical proof of the properties of those algorithms [38]. We also compared a this approach to two others recent approaches (Aggregation and probabilistic polling approaches) [36].

6.2.5. Core Persistence in peer to peer Systems: Relating size to lifetime

Participants: Vincent Gramoli, Anne-Marie Kermarrec, Achour Mostefaoui, Michel Raynal.

While organizing peers in an overlay network has generated a lot of interest leading to a large number of solutions, maintaining critical data in such a network remains an open issue. In this work, we were interested in defining the portion of nodes and the probing, given the churn observed in the system, in order to achieve a given probability of maintaining the persistence of some critical data. More specifically, we provided a clear result relating the size, the frequency of the probing, along with its proof, as well as an analysis of the way of leveraging such an information for parameter setting in large-scale dynamic systems [30].

6.3. Resource management in peer to peer systems

Keywords: *Peer to peer content searching*, *RSS feeds*, *backup systems*, *gossip-based overlay construction*, *objects networks*, *publish and ubscribe*, *random walk*, *structured overlay*, *system size estimation*.

6.3.1. Querying peer to peer systems

Participants: Anne-Marie Kermarrec, Fabrice Le Fessant, Étienne Rivière.

Managing resources on a large scale, whether they are computing resources, data, events, bandwidth, requires some fully decentralized solutions. Our research in this area focuses on building the relevant overlay networks to provide core functionalities of resource management and discovery.

Efficient search algorithms are crucial for a wide range of distributed applications. We worked in this area along two main directions: content-based publish-subscribe systems and generic query mechanisms. In publish-subscribe systems, subscribers register their interest in an event or a pattern of events in order to be asynchronously notified of any event published matching their subscription. On the contrary, query mechanisms are symmetric: items are stored permanently and queries are the events looking for matching items. While existing P2P generic infrastructures provide a scalable support for topic-based publish-subscribe systems, they are not well adapted to content-based ones (in which events are filtered according to their content). In this area, we pursued the following objectives of (i) raising expressiveness provided by search mechanisms in peer to peer data storage systems and publish and suscribe middleware; (ii) investing a new way of organizing data in such overlays, by linking application objects rather than physical nodes; and finally (iii) to propose self-organizing distributed algorithms achieving these goals while providing scalability, resilience to dynamicity and failures, and communication efficiency. These objectives are tackled for two different distributed communication paradigms: query-retrieval and publish-subscribe systems, and for content distribution networks in the context of RSS feeds dissemination. We strongly believe that dedicated overlays present a sound basis for most applications, rather than generic lookup mechanisms such as distributed hash tables, and our approachs are based on this primary design goal.

Two projects have been conducted in this area: GosSkip and Voronet. They share a common design decision, as they link application objects themselves rather than computing entities. This permits more expressive queries, and more design possibilities to achieve distributed data structures. This is a relatevily new approach to structured peer to peer overlays, in that most overlays in the litterature rely on an organization of computing entities, mostly using a distributed hash table structure, and therefore often exhibit low expressivness and flexibility.

In the GosSkip project, we built a gossip-based structured overlay using gossip-based construction. This work has been done in collaboration with Sidath Handurukande, Rachid Guerraoui (EPFL, Switzerland) and Kévin Huguenin (student from ENS Cachan, during an internship in Summer 2005). GosSkip is an efficient attribute-based publish-subscribe system. GosSkip is a structured overlay the structure of which reflects the actual structure of the underlying application properties. GosSkip relies on gossip messages to construct a structure eventually similar to a perfect Skip list, preserving the semantic locality of the items stored in the overlay. In GosSkip, events are delivered to matching subscriptions in $O(logN)$ routing hops, N being the total number of subscriptions. Experimental results based on a real peer to peer trace demonstrate the scalability, the failure resilience, the efficiency and the fairness of the approach both in static and dynamic scenarios. In 2006, we extended this approach along two directions. First, a mechanism to leverage the presence of multiple data objects on a single physical node has been proposed. Second an emulation of GosSkip on the Grid 5000 testbed has been performed, using traces of the KaZaa, a peer to peer file sharing system. This work was published in the fifth P2P conference [31].

The Voronet project aims at building a fully distributed overlay network for data storage system with strictly proven algorithmic costs. This work is done in collaboration with Loris Marchal (ENS Lyon) and Olivier Beaumont (LABRI Bordeaux). The idea of Voronet is to generalize the Kleinberg model where each peer in an overlay is connected to its neighbors on a grid as well as to a remote node picked at random with a probability proportional to its distance. The fact that remote nodes are chosen according to an harmonic distribution of distances to original nodes, enables the network to exhibit two keys properties of small-world systems: existence of short paths and navigability. Each node is linked to a small number of other nodes: these neighbors are those who share vertexes in the Voronoi tesselation of the euclidean space. This links are eventually forming the Delaunay complex of the set of elements. Every data item is specified by a set of values over k attributes, positioning the item in an k-euclidean space. Each data item represents a logical peer in the Voronet overlay and is connected to a set of neighbor, sharing vertexes in the Voronoi tessellation of the Euclidean space. These links eventually form the Delaunay complex of the set of elements. Each peer also knows a remote peer to provide efficient polylogarithmic routing between any two peers in the overlay, independently of the distribution of peers in the space. VoroNet has been implemented in a simulation tool chain, and extendend simulation results involving up to millions of nodes demonstrate its good properties in terms of routing efficiency and scalability [49], [22]. This work opens several research perspectives: First we plan to develop and evaluate complex query mechanisms for VoroNet, second we are in the process of investigating a gossip-based construction of the overlay, to provide more fault-tolerance and self-organization.

6.3.2. Decentralized Content-based Publish and Subscribe

Participants: Anne-Marie Kermarrec, Étienne Rivière.

The work on the evaluation of publish-subscribe systems has been done in collaboration with Spyros Voulgaris and Prof. Marteen Van Steen, researchers at the Vrije Universiteit in Amsterdam, the Netherlands. Even if research on efficient and decentralized implementations of the publish and subscribe communication paradigm has generated a lot of interest in the past few years, most proposed systems were evaluated using ad-hoc workloads, and comparing systems is most of the time not possible. To be able to accurately measure performances of the systems we develop, we proposed two workload generation techniques, that can be reused to compare different PubSub systems with the same synthetic or real-world settings. We designed a workload generator and simulation tools which goals are to highlight problematic behaviors of such a system or potential target applications. Second, we obtained traces from the usage of a centralized publish and subscribe system, with more than 100,000 users, publishers and subsribers. These traces have been used to further evaluate Sub-2-Sub, which is described in the next paragraph.

In 2006, we carried on the Sub-2-Sub project initiated in 2005 in the context of the Epi-Net associated team. This work has been done in collaboration with Spyros Voulgaris and Prof. Marteen Van Steen from the Vrije Universiteit in Amsterdam. The Sub-2-Sub project [47] aims at providing a self-organizing overlay network for content-based publish-subscribe systems. PubSub systems are loosely coupled distributed systems that enable communication between information producers (publishers) and information consummers (subscribers) based on registered consummers interests in information (subscriptions). Sub-2-Sub builds an unstructured

peer to peer overlay linking subscriptions using epidemic algorithms so that subscribers having similar interests are automatically clustered. Events are then routed towards clusters and efficiently disseminate within such clusters. Sub-2-Sub has been implemented and tested using the open source simulator PeerSim. We also used real traces of a centralized Publish Subscribe system to evaluate Sub-2-Sub performance with a realistic workload. We also extended the protocol to handle undefined attributes values for publications attributes.

A third project started in June 2006 a new project, called Rappel: a self-organizing dedicated overlay for dissemination of RSS feeds update. This work is done in collaboration with Jay A. Patel and Prof. Indranil Gupta, respectively PhD student and assistant professor at the University of Illinois at Urbana Champaign (UIUC), United States. RSS Feeds are XML data associated to a website or any content provider. They include a set of entries, called update, along with their timestamp. RSS Feeds can be very popular (Google news, healine newspaper) or unpopular (as most personnal blogs). A user can subscribe to a set of feeds using a feed reader client. This feed reader client polls periodically (i.e., every 30 minutes) the feed server to check for new updates. This may involve a high stress on a server if it hosts popular feeds, and this also involves lots of unecessary requests if the period of update publications is higher than the polling period from the client applications. Rappel's key design choices are: (i) a peer to peer distributed system that is convenient both for rare and popular feeds; (ii) a design taking into account network proximity as a mean to reduce the stress on the network that other peer to peer systems may exhibit; (iii) fault tolerance and self organization as a primary design goal; and (iv) a more efficient rate of update discovery than the one that is achieved through direct polling, by rapidly disseminating updates through redundant dissemination trees.

Rappel organizes peers (subscribers and publishers) in a peer to peer resilient dissemination network. This network is optimized using inherent correlation existing in human preferences graph, and using network proximity, that helps reducing the stress on the infrastructure. The latter is achieved by using network coordinates. Rappel will be evaluated using real traces with more than 300,000 subscribers and publishers and 1 million updates, from a popular personnal blog and online aggreation platform, using a event-driven simulator that simulates the underlying infrastructure using a transit-stub topology model.

6.3.3. Palabre: a Peer to-peer back-up system

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 telecommuting. Backuping all this data has become a new challenge of 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 Palabre open-source project, uses both a structured overlay, to localize stored data during restauration, and an unstructured overlay, to query for storage availability among neighbours. 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. A novel methodology has been also developed to allow automatic encoding and decoding of message and file formats while keeping full backward compatibility with almost no extra programming cost. This work is done in collaboration with Laurent Viennot and Anh-Tuan Gai from GYROWEB, INRIA-Rocquencourt.

6.3.4. Distributed Ordered Slicing in Large-Scale Systems

Participants: Anne-Marie Kermarrec, Vincent Gramoli, Michel Raynal.

Recently, there has been an increasing interest to harness the potential of peer to peer technology to design and build rich environments where services are provided and multiple applications can be supported in a flexible and dynamic manner. In such a context, resource assignment to services and applications is crucial. Current approaches require significant "manual-mode" operations and/or rely on centralized servers to maintain resource availability. Such approaches are neither scalable nor robust enough.

During his visit in Spring 2006, Mark Jelasity, together with Anne-Marie Kermarrec designed a solution to this problem. They proposed and evaluated a gossip-based protocol to automatically partition the available nodes into "slices", also taking into account specific attributes of the nodes. These slices can be assigned to run services or applications in a fully self-organizing but controlled manner. The main advantages of the proposed protocol are extreme scalability and robustness. The ordered nature of the slicing means that specific attributes can be taken into account to partition the network: the partitioning is done along a fixed attribute of the nodes. For example, a service might require a slice composed of the top 20% of the nodes providing the largest bandwidth. Besides, we need to provide this top 20% constantly, even if the nodes in the top 20% constantly change due to churn or changing node properties. Many metrics may be used to sort the nodes such as available resources (memory, bandwidth, computing power) or some specific behavioral pattern such as up-time. Note that slicing the network at random, and focusing only on the size of the slices is a special case of our ordered slicing protocol. We also note that the slice sizes are expressed as a percentage of the network, that is, if the network grows, slices grow accordingly.

We base our approach on a class of gossip-based protocols, that have been proven to be able to maintain connectivity in large-scale dynamic systems in the presence of high churn and other extreme failure scenarios [56]. This is achieved through maintaining a dynamic, pseudo-random overlay network. Due to their low cost, simplicity, scalability and robustness, these networks represent an ideal foundation to build our protocol upon. We present approximative theoretical models and extensive empirical analysis of the proposed protocol.

Building upon this experience, in collaboration with Antonio Fernandez and Ernesto Jimenez from Madrid university, Spain, and Mark Jelasity, now in Hungary at the University of Szeged, we are currently investigating other approaches, able to deal with correlated failures.

6.3.5. *Adaptive replica placement strategies*

Participants: Vincent Gramoli, Anne-Marie Kermarrec, Erwan Le Merrer.

This work takes place in the context of a peer to peer system where a set of services is evenly replicated in the network in such a way that the service can be reached in a small number of routing steps. Organizing such a service in a fully decentralized way so that the right replication degree is reached is a real challenge. In this context, we are working on an algorithm, called Sonde [37], able to automatically ensure a given level of service replication in a large-scale peer to peer network. Not only, such algorithms should deal with churn (high rate of nodes arrival and departures) but also adapt the replication so that the load is evenly replicated, matching the potential constraints of participating entities.

6.3.6. *Combining structured and unstructured peer to peer networks*

Participants: Balasubramaneyan Maniymaran, Marin Bertier, Anne-Marie Kermarrec.

As many different peer to peer overlay networks providing various functionalities have been proposed, it is likely that multiple overlays may be deployed over a set of nodes. Therefore a physical peer may hosts several instances of logical peers belonging to various overlay networks. In this work, we show that the co-existance of a structured peer to peer overlay and an unstructured one may be leveraged so that by building one, the other is automatically constructed as well. More specifically, we show that the randomness provided by an unstructured gossip-based overlay may be used to build the routing tables of a structured peer to peer overlay and the other way around. Simulation results, comparing our approach with both a Pastry-like system and a gossip-based unstructured overlay, show that we significantly reduce the overhead while providing similar functionalities.

6.4. Peer to peer sensor networks

Keywords: *Peer to peer overlays*, *coverage*, *gossip-based algorithms*, *power comsumption*, *sensor networks*.

Participants: Marin Bertier, Yann Busnel, Anne-Marie Kermarrec, Erwan Le Merrer, Aline Viana.

In this area, we are investigating the use of peer to peer algorithms in sensor network systems. We observe many similarities between this two domains that we plan to leverage. Scale and dynamicity are among the most striking similarities between the two types of networks. However, sensor network specificities imply major adaptations of Internet-based peer to peer algorithms. Thus, this new research area allows us to widen our application domain with specific applications, but above all to vary some major properties of our target system and therefore generalize our work on fully decentralized algorithms. At the following, we give a brief description of our current activities.

We conducted a first activity in the area of information dissemination in sensor network in the context of monitoring children activities to detect obesity pathologies. This application brings us a challenging setting with respect to dynamics as sensors are embedded on human beings and are by essence mobile. In this context, we are investigating methods for energy-efficient route discovery and for the reliable relaying of data from the sensor to the sink (monitoring station), so that the network lifetime can be maximized. While many approaches in this area rely on precise location information such as GPS, we want to avoid such techniques for cost, size but also performance reason, as GPS information is not available indoors. We are investigating the use of epidemic algorithms to trace sink location and efficiently transport data to a sink without flooding the network.

We also proposed a gossip-based algorithm for software updates in Sensor networks. this work has been done in collaboration with Eric Fleury, ARES project-team, Inria Rhone Alpes.

Finally, we initiated a project in the area of power management in sensor networks. Energy consumption is the most important factor that determines sensor node lifetime. The optimization of wireless sensor network lifetime targets not only the reduction of energy consumption of a single sensor node, but also the extension of the entire network lifetime. The sensor network lifetime is defined as the period during which the routing fidelity and the sensing fidelity of the network are guaranteed. Our goal is then to leverage node redundancy in wireless sensor networks (WSNs) to reduce and distribute the computational and communication energy consumption of the network between sensors. We consider that the cooperative nature of sensors offers significant opportunities to manage energy consumption. We then propose a simple and adaptive energy-conserving topology management scheme, called SAND (Self-Organizing Active Node Density). SAND is fully decentralized and relies on a distributed probing approach and on the redundancy resolution of sensors for energy optimizations, while preserving the data forwarding and sensing capabilities of the network. The SAND proposal was recently published at the ACM MobiShare workshop, the 1st International Workshop on Decentralized Resource Sharing in Mobile Computing and Networking [35].

7. Contracts and Grants with Industry

7.1. France Telecom

Participant: Anne-Marie Kermarrec.

Since October 2004, we have a collaboration with France Télécom R&D, Lannion on applying peer to peer techniques to telecom operator frameworks. More specifically, in this area, we are working on timely dissemination of voice over IP and a reliable and distributed telecom infrastructure. In this context, Anne-Marie Kermarrec acts as the Ph.D advisor of Erwan le Merrer.

7.2. Advestigo

Participants: Marin Bertier, Anne-Marie Kermarrec, Fabrice Le Fessant.

We have a consulting contract with Advestigo, which just received the 2006 European IST Prize Awards, a small company working in the content protection area. In this context Marin Bertier, Anne-Marie Kermarrec and Fabrice Le Fessant are providing expertise in the area of monitoring peer to peer systems.

8. Other Grants and Activities

8.1. National grants

8.1.1. ANR MD ALPAGE

Participants: Marin Bertier, Anne-Marie Kermarrec, Fabrice Le Fessant, Étienne Rivière.

Alpage, is an ANR MD project starting in January 2006 focusing on algorithms for large-scale platforms. The project gathers several teams with complementary expertise ranging from algorithms design and scheduling techniques, to macro-communications primitives and routing protocols and to peer to peer architectures and distributed systems. In this project, we aim at designing algorithms for large-scale dynamic platforms and will concentrate our efforts on the following complementary axes:

- Large-scale distributed platform modeling
- Overlay network topologies
- Scheduling for regular parallel applications
- Scheduling for file-sharing applications

The partners includes the LABRI (contact: Olivier Beaumont), ENS Lyon (contact: Yves Robert), LRI (Contact: Pierre Fraigniaud) and IRISA (contact: Anne-Marie Kermarrec). In this context, the ASAP project-team is mostly involved in the overlay network topologies theme and in this context we are actively collaborating with Olivier Beaumont (LABRI).

8.1.2. RNRT project SVP

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

The SVP project addresses the understanding, the conception, and the implementation of an integrated ambient architecture that would ease the optimization in the deployment of monitoring and prevention services in various types of dynamic networks. The main objective is to develop an environment which is able to accommodate a high number of dynamic entities completely dedicated to a specific service. The different 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 sensor networks for health monitoring applications takes place in this context.

8.1.3. ARC Inria Recall.

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

Anne-Marie Kermarrec, Achour Mostefaoui and Michel Raynal are involved in the ARC *Recall* on optimistic replication for collaborative editing in peer to peer networks. The INRIA project-teams CASSIS INRIA LORRAINE and REGAL INRIA ROCQUENCOURT as well as the LIRMM and EPFL (Rachid Guerraoui) are involved in this project as well.

8.1.4. Academic collaborations

Francois Baccelli (TREC, INRIA ROCQUENCOURT/ENS) and Anne-Marie Kermarrec have been collaborating on virtual coordinates network in the context of the master thesis of Bruno Kauffmann. Dominique Lavenier (SYMBIOSE, INRIA RENNES) and Anne-Marie Kermarrec are collaborating on the use of peer to peer infrastructures to index and search DNA sequences libraries. Fabrice Le Fessant and Anne-Marie Kermarrec are working with Laurent Viennot (GYROWEB, INRIA-ROCQUENCOURT) on collaborative back-up systems.

8.2. International grants

8.2.1. The ReSIST european project

Participants: Marin Bertier, Achour Mostefaoui, Michel Raynal, Corentin Travers.

ReSIST is an NoE (Network of Excellence) that addresses the strategic objective "Towards a global dependability and security framework" of the European Union's FP6 Work Programme for IST (Information Society Technologies), and responds to the stated "need for resilience, self-healing, dynamic content and volatile environments". The contract supporting the ReSIST activities extends on 3 years, starting on the 1st of January 2006.

ReSIST integrates leading researchers active in the multidisciplinary domains of Dependability, Security, and Human Factors, in order that Europe will have a well-focused coherent set of research activities aimed at ensuring that future "ubiquitous computing systems", the immense systems of ever-evolving networks of computers and mobile devices which are needed to support and provide Ambient Intelligence (AmI), have the necessary resilience and survivability, despite any residual development and physical faults, interaction mistakes, or malicious attacks and disruptions.

ReSIST's partners are: Budapest UTE (HG), London City U. (UK), TU Darmstadt (DE), Deep Blue Srl (IT), France Telecom R&D (FR), IBM Research GmbH (CH), Institut Eurecom (FR), IRISA (FR), IRIT (FR), LAAS-CNRS (FR), Lisbon U. (PT), Newcastle upon Tyne U. (UK), Pisa U. (IT), Qinetiq (UK), Roma U. La Sapienza (IT), Ulm U. (DE), Southampton U. (UK), Vytautas Magnus U. (LT).

The current state-of-knowledge and state-of-the-art reasonably enables the construction and operation of critical systems, be they safety-critical (e.g., avionics, nuclear control) or availability-critical (e.g., back-end servers for transaction processing). The situation drastically worsens when considering large, networked, evolving, systems either fixed or mobile, with demanding requirements driven by their domain of application. There is statistical evidence that these emerging systems suffer from a significant drop in dependability and security in comparison with the former systems. There is thus a dependability and security gap opening in front of us. Filling the gap clearly needs dependability and security technologies to scale up, in order to counteract the two main drivers of the creation and widening of the gap: complexity and cost pressure.

8.2.2. Epi-Net associated team with Vrije Universiteit, Amsterdam, NL

Participants: Marin Bertier, Yann Busnel, Anne-Marie Kermarrec, Etienne Rivière.

Following a PAI Van Gogh action in 2005, Epi-Net is an associated team from January 1st, 2006. Epi-Net addresses several applications using epidemic-based unstructured networks. Gossip-based communication models have recently started to be explored as a general paradigm to build and maintain unstructured overlay networks. More specifically, they have shown to provide a scalable way of implementing and maintaining highly dynamic unstructured overlays in which nodes can frequently join and leave. Many variants of such protocols exist and they mainly differ in deciding which neighbor to communicate with, deciding on exactly which neighbors to exchange information on, and, in the end, deciding on which peers to keep in the list to prevent it from growing unboundedly.

In collaboration with Rachid Guerraoui (EPFL) and Mark Jelasity (University of Bologna), Maarten van Steen (VU) and Anne-Marie Kermarrec have been the first to systematically study the effects of the aforementioned decisions, providing a deep insight in the tradeoffs that need to be made. One important observation from this empirical study is that epidemic algorithms cope very well with high dynamics. The goal of Epi-Nets is to start from these results on generic unstructured networks, to exploit other forms of epidemic algorithms in specific and complex settings, namely (i) large-scale publish-subscribe applications (ii) ad-hoc networks and (iii) file sharing applications.

Apart from short visits of both team members on eiter side, Etienne Rivière visited Vrije Universiteit in June and July 2006 and Stevens Leblond, from Vrije Universiteit visited the ASAP team for three months (May-July 2006).

In this context Maarten van Steen and Anne-Marie Kermarrec are organizing in December 2006, a workshop on the gossip-based communication networks, gathering all the world-wide experts in the area.

8.2.3. Collaboration with Madrid, Spain

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

This project financed by the Urban community of Madrid includes researchers from the university of Madrid (mainly Professor Antonio Fernandez) and Michel Raynal, Achour Mostefaoui and Anne-Marie Kermarrec.

The Distributed shared memory (DSM) is a well-known mechanism for inter-process communication in a distributed environment. With this mechanism, application processes exchange information by reading and writing shared objects. These objects are the basic elements of a shared memory abstraction which is in fact supported by the distributed memory of the distributed system and implemented by a distributed DSM algorithm (or algorithm for short).

In this project we plan to study and hopefully solve the problem of implementing DSM systems with useful consistency criteria in the presence of failures. For that, we need first to define the different consistency criteria in the presence of failures. Then, we need to study the use of basic fault-tolerant services, like consensus or atomic broadcasts, to derive reliable distributed shared memory systems that implement these consistencies.

We are also working in this context on the distributed 'slicing" algorithms in large-scale systems.

In the context of this collaboration Antonio Fernandez and Ernesto Jimenez visited the ASAP team over the summer 2006.

8.2.4. PAI Mexique

Participants: Achour Mostefaoui, Michel Raynal, Corentin Travers.

The goal of our LAFMI 2003-2006 project (Franco-Mexican research laboratory in computer science) is to study in depth the *condition-based* approach to solve consensus and other distributed problems, mainly on synchronous distributed systems. The partners are Professor Sergio Rajsbaum from the Autonomous university of Mexico and professors Michel Raynal and Achour Mostefaoui. Several visits occurred during these last years in both directions.

The approach, which we introduced in 2001, studies *conditions* which identify restrictions on the set of possible inputs to a problem, which allow to solve it faster, or allow to solve it under model assumptions where it was unsolvable. We have studied the approach in asynchronous systems, mainly for consensus and set-agreement. The first main result we obtained was defining *legal* conditions and show that these are exactly the conditions for which a consensus protocol exists in an asynchronous system where at most t processes can crash. We presented a generic consensus protocol for any legal condition. In sequel publications we identified conditions which allow consensus to be solved faster, and we studied set-agreement and other problems. We discovered relations to coding theory, topology, and other approaches in distributed computing like failure-detectors. We started exploring the condition-based approach in synchronous systems and found out that it yields a rich and important research area that we have been exploring in this project.

8.2.5. Collaboration with Univeristy of Illinois, Urbana Champain

Participants: Marin Bertier, Anne-Marie Kermarrec, Etienne Rivière.

In 2005 we started a 2-year collaboration with Indranil Gupta's team from the University of Illinois at Urbana Champaign (UIUC). Indranil Gupta visited INRIA in June 2006 and he initiated, in collaboration with Anne-Marie Kermarrec and Étienne Rivière, the Rappel project. Étienne Rivière visited UIUC in September for two weeks to carry on the collaboration.

8.2.6. PAI Hong Kong

Participant: Michel Raynal.

The aim of this research project is to explore the world of agreement problems in the context of MANET (Mobile Adhoc NETwork) and in the context of mobile agents. The main partners of this project are Professor Cao from the university of HK Poytechnic and Michel Raynal and Corentin Travers. During these two last years, three visits took place.

The agreement problems have been extensively studied in classical distributed systems, namely, systems where the processes communicate through a shared memory or a fixed network. Here we want to solve in the context of MANET (Mobile Adhoc NETwork) and in the context of mobile agents. What makes the problem difficult to solve is the combination of failures and asynchrony that prevents the entities to know which of them can actually actively participate in the consensus algorithm. These new contexts add new difficulties.

8.3. Visits (2006-2007)

Amr El Abbadi, UCSB, May 2006.

Antonio Fernàndez, University "Rey Juan-Carlos", Madrid, Spain, (June-september 2006).

Indranil Gupta, University of Illinois, Urbana Champaign, USA (June 2006).

Mark Jelasity, University of Bologna is visiting researcher in March-April 2006. He worked with Anne-Marie Kermarrec on resource allocation in large-scale dynamics networks. This work heavily relies on gossip-based algorithms.

Stevens Leblond, Master student at the Vrije Universiteit in Amsterdam (NL) spent 8 weeks in the context of the associated team Epi-Net. He worked in collaboration with A.-M. Kermarrec in the area of collaborative downloading in May-July 2006.

Oliver Theel, University of Oldenburg, Germany (May-June 2006).

Weigang Wu, Hong-Kong Polytechnic university (February-March 2006).

In 2006, we received a number of short-term visits among them Pascal Felber, University of Neuchale, Switzerland; Roy Friedman, Technion, Haifa, Israel; Sara Tucci, University of Roma, Italy.

9. Dissemination

9.1. Community animation

9.1.1. Leaderships, Steering Committees and community service

Lorentz workshop: A.-M. Kermarrec, with M. van Steen, organized in December 2006, a workshop on "Gossip-based networking" in Leiden, The Netherlands, gathering the international experts in the area (50 attendees,).

ANR Masses de données et connaissances ambiantes (MDCA): A.-M. Kermarrec is a member of the *Selection Committee* of the evaluation committee of the *ANR Masses de données et connaissances ambiantes*.

SPECIF award: A.-M Kermarrec is a member of the selection committee of the *Prix de thèse SPECIF*.

9.1.2. Editorial boards, steering and program committees

M. Bertier served in the Program Committees for the following conferences:

- MSN'06: *Second International Conference on Mobile Ad-hoc and Sensor Networks*, Hong-Kong, China, December 2006.
- CFIP 2006: *Colloque Francophone sur l'Ingénierie des Protocoles*, Tozeur, Tunisia, November 2006.
- Algotel 2007: *The International Conference on Dependable Systems and Networks* to be held in Ile d' Oléron, France, Mai 2007.

He also serves as the Web responsible in the Euro-Par 2007 organisation committee.

A.-M. Kermarrec is the General Chair (PC chair and Organization) of Euro-Par 2007, to be held in Rennes in August 2007. She also served as Publicity Chair for the *EuroSys 2006*, held in Leuven, Belgium in April 2006.

She serves in the *Steering Committee* of the *Euro-Par* annual conference on parallel computing (250 attendees, http://www.europar.org/)

She served in the Program Committees for the following conferences:

IPTPS '06: *International workshop on peer-to-peer Computing*, Santa-Barbara, CA, USA, February 2006.

IEEE Global Internet Symposium 2006. In conjunction with Infocom in Barcelona, Spain, April 2006

DSN 2006: *The International Conference on Dependable Systems and Networks*, Philadelphia, PA, US, June 2006.

ICDCS '06: *International Conference on Distributed Computing Systems*, Operating system and Middleware track, Lisboa, Portugal, July 2006.

IWDDS '06: *The first International Workshop on Dynamic Distributed Systems* in conjunction with ICDCS, July 2006.

Algotel 2006: *Rencontres Francophones sur les aspects algorithmiques des télécommunications*, Trégastel, France, June, 2006.

Euro-Par 2006: *Peer to peer and Web Computing* Topic of Euro-Par 2006, Dresden, Germany, August 2006.

IWSOS '06: *New Trends in Network Architectures and Services: International Workshop on Self-Organizing Systems* Passau, Germany, September 2006.

Global Internet Symposium 2006: *IEEE Global Internet Symposium 2006*, Barcelona, Spain, April 2006.

Middleware 2006: *ACM/IFIP/USENIX 7th International Middleware Conference* Melbourne, Australia, 2006.

CFSE 2006: *Conférence Française en système d'exploitation*, Perpignan, France 2006.

Co-Next'06: *2nd Conference on Future Networking Technologies*, Lisboa, Portugal, December 2006.

SSC'06: *Eighth International Symposium on Stabilization, Safety, and Security of Distributed Systems*, Dallas, USA, November 2006.

ICDCS'07: *International Conference on Distributed Computing Systems*, Data Management track, to be held in Toronto, Canada in June 2007.

Infocom 2007: *IEEE Conference on Computer Communications and Networking*, to be held in Anchorage, Alaska, USA, May 2007.

NSDI'07: *USENIX Symposium on Networked Systems Design & Implementation* to be held in Cambridge, MA, Boston, April 2007.

SASO 2007: *First IEEE International Conference on Self-Adaptive and Self-Organizing Systems* to be held in Boston, USA, July, 2007

HotDep'07: *the Third Workshop on Hot Topics in System Dependability*, to be held in conjunction with DSN'2007 in Edinburgh, UK.

ICDE'2008: *IEEE 24th International Conference on Data Engineering (ICDE 2008)*, Distributed, Parallel, and Peer to Peer Databases Track, to be held at Cancun Mexico, 2008.

Fabrice Le Fessant served in the Program Committees for the following conferences:

Euro-Par 2007: *Peer to peer and Web Computing* Topic of Euro-Par 2006, to be held in Rennes, France, August 2007.

ICFP 2007: *12th ACM SIGPLAN International Conference on Functional Programming*, to be held in Freiburg, Germany, October 2007.

Achour Mostefaoui served in the Program Committees of the following conferences:

AINA'06: *IEEE conference on Advances Information Networking and Applications*

ICDCS'06: *International Conference on Distributed Computing Systems*, Algorithms and Theory track, Lisboa, Portugal in July 2006.

Euro-Par 2007: *Distributed Systems and algorithms* Topic of Euro-Par 2006, to be held in Rennes, France, August 2007.

DISC'07: 21st Annual Conference on Distributed Computing.

FOFDC'07: ARES 2007 Workshop on Foundations of Fault-tolerant Distributed Computing.

ISPS'07: 8th International Symposium on Progamming and Systems.

Michel Raynal is a member of the editorial board of the following journals:

IEEE TPDS: IEEE Transactions on Parallel and Distributed Systems.

JPAC: Journal of Parallel and Distributed Computing

JCSSE: Journal of Computer Systems Sciences and Engineering

He is a member of the steering committees of the following conferences: ACM PODC, SIROCCO and ICDCN.

He served in the Program Committees of the following conferences:

AINA'06: *IEEE conference on Advances Information Networking and Applications*

DISC'06: PC member and chair of the 20th anniversary of the symposium, *20th Annual Conference on Distributed Computing*

ICDCS'06: International Conference on Distributed Computing Systems, Conference co-chair Lisboa, Portugal in July 2006.

OPODIS'06: 10th International Conference on Principles of Distributed Systems.

PODC 2006: 25th ACM Symposium on Principles of Distributed Computing

SSC'06: Eighth International Symposium on Stabilization, Safety, and Security of Distributed Systems, Dallas, USA, November 2006.

AINA'07: IEEE conference on Advances Information Networking and Applications.

9.1.3. Evaluation committees, consulting

A.-M. Kermarrec served as a reviewer of the EVERGROW IP EU-funded project.

She acted as an evaluator for the FP6 Call 4 S.O 2.4.6 *Networked Audio/Visual Systems and Home platform.*

She acted as a referee for the foreign PhD committees of Leonardo Querzoni, University la Sapienza, Roma, Italy and Spyros Voulgaris from Vrije universiteit, Amsterdam, The Netherlands.

She 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 ASR Grid, peer to peer and parallelism.

Michel Raynal acted as an external examinator for the Ph.D of P. Zielinsky from Cambridge University, UK.

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.

Anne-Marie Kermarrec and Michel 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 PARIS project-team.

Fabrice le Fessant is an associate professor at Ecole Polytechnique.

Achour Mostefaoui is responsible of a Master's course (University of Bougie, Algeria) entitled "Distributed Algorithms".

In addition four Ph.D students, Yann Busnel, Etienne Rivière, Corentin Travers and Gilles Trédan are teaching assistants (*moniteurs*).

9.3. Conferences, seminars, and invitations

Only the events not listed elsewhere are listed below.

MINEMA 2007. A.-M. Kermarrec is invited to give a tutorial at the MINEMA Winter School, Switzerland, February 2007.

9.4. Administrative responsibilities

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

She was a member of the 2006 INRIA Selection Committee for the Junior Researcher permanent positions (CR2) at the INRIA Futurs, Rocquencourt and Rhone-Alpes Research Units.

She was a member of the 2006 INRIA Selection Committee for the Senior Researcher permanent positions (DR2).

She is a member of the working group *Actions incitatives* of the INRIA *Conseil d'Orientation Scientifique et Technologique*.

She is a member of the Selection Committee (*Commission de Spécialistes*, CSE) for computer Science of ENS CACHAN.

10. Bibliography

Major publications by the team in recent years

[1] M. BERTIER, L. ARANTES, P. SENS. *Distributed Mutual Exclusion Algorithms for Grid Applications: a Hierarchical Approach*, in "Journal of parallel and distributed Computing (JPDC)", 2005.

[2] M. BERTIER, O. MARIN, P. SENS. *Performane Analysis of Hierarchical failure Detector*, in "International Conference on dependable Systems and networks (DSN' 03), San Francisco, USA", October 2003.

[3] M. CASTRO, P. DRUSCHEL, A.-M. KERMARREC, A. NANDI, A. ROWSTRON, A. SINGH. *SplitStream: High-Bandwidth Multicast in Cooperative Environments*, in "Symposium on Operating System principles (SOSP 2003), Bolton Landing, NY, USA", October 2003.

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

[5] A. J. GANESH, A.-M. KERMARREC, L. MASSOULIÉ. *Peer-to-Peer membership management for gossip-based protocols*, in "IEEE Transactions on Computers", vol. 52, n^{o} 2, February 2003.

[6] S. HANDURUKANDE, A.-M. KERMARREC, F. LE FESSANT, L. MASSOULIÉ, S. PATARIN. *Peer Sharing Behaviour in the eDonkey Network, and Implications for the Design of Server-less File Sharing Systems*, in "Eurosys, Leuven, Belgium", September 2006.

[7] J.-M. HÉLARY, A. MOSTEFAOUI, R. NETZER, M. RAYNAL. *Communication-based prevention of useless checkpoints in distributed computations*, in "Distributed computing", vol. 13, n^{o} 1, 2000, p. 29-43.

[8] A.-M. KERMARREC, L. MASSOULIÉ, A. J. GANESH. *Probabilistic Reliable Dissemination in Large-Scale Systems*, in "IEEE Transactions on Parallel and Distributed Systems", vol. 14, n^{o} 3, March 2003.

[9] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL. *Conditions on Input Vectors for Consensus Solvability in Asynchronous Distributed Systems*, in "Journal of the ACM", vol. 50, n^{o} 6, 2003, p. 922-954.

[10] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL. *Synchronous Condition-Based Consensus*, in "Distributed Computing", vol. 18, n^{o} 5, 2006, p. 325-343.

Year Publications

Books and Monographs

[11] F. LE FESSANT. *Peer-to-Peer: Comprendre et utiliser*, Eyrolles, 2006.

Articles in refereed journals and book chapters

[12] O. BABAOGLU, M. JELASITY, A.-M. KERMARREC, A. MONTRESOR, M. VAN STEEN. *Managing Clouds: A Case for a Fresh Look at Large Unreliable Dynamic Networks*, in "ACM SIGOPS Operating Systems Review", vol. 40, n^{o} 3, July 2006.

[13] R. FRIEDMAN, A. MOSTEFAOUI, M. RAYNAL. *On the respective power of eventual P P and eventual S to solve one-shot agreement problems*, in "IEEE Transactions on Parallel and Distributed Systems", To appear, 2007.

[14] S. GORENDER, R. MACEDO, M. RAYNAL. *An adaptive programming model for fault-tolerant distributed computing*, in "IEEE Transactions on Dependable and Secure Computing", To appear, 2007.

[15] I. GUPTA, A.-M. KERMARREC, A. GANESH. *Efficient and Adaptive Epidemic-style Protocols for Reliable and Scalable Multicast.*, in "IEEE Transactions on parallel and distributed systems", July 2006.

[16] A. MOSTEFAOUI, E. MOURGAYA, M. RAYNAL, C. TRAVERS. *A time-free assumption to implement eventual leadership*, in "Parallel Processing Letters", vol. 16, n^{o} 2, 2006, p. 189-208.

[17] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL. *Synchronous Condition-Based Consensus*, in "Distributed Computing", vol. 18, n^{o} 5, 2006, p. 325-343.

[18] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *From eventual W to Omega: a simple bounded quiescent reliable broadcast-based transformation*, in "Journal of Parallel and Distributed Computing", To appear, 2007.

[19] A. MOSTEFAOUI, M. RAYNAL, C. TRAVERS. *Time-free and timer-based assumptions can be combined to get eventual leadership*, in "IEEE Transactions on Parallel and Distributed Systems", vol. 17, n^{o} 7, 2006, p. 656-666.

[20] M. RAYNAL, R. GUERRAOUI. *The alpha of indulgent consensus*, in "The Computer Journal", To appear, 2007.

Publications in Conferences and Workshops

[21] Y. AFEK, E. GAFNI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *Simultaneous consensus tastks: a tighter characterization of set-consensus*, in "Proc. 12th Int. Conference on Distributed Computing and Networking (ICDCN-06), Guwahati, India", LNCS, n^{o} 4308, Springer Verlag, 2006, p. 331-341.

[22] O. BEAUMONT, A.-M. KERMARREC, L. MARCHAL, É. RIVIÈRE. *VoroNet, réseau objet-à-objet sur le modèle petit monde*, in "Cinquième Conférence Francophone en Systèmes d'Exploitation (CFSE), Perpignan, France", October 2006.

[23] Y. BUSNEL, A.-M. KERMARREC. *PROXSEM : Mesure de proximité sémantique pour les systèmes de partage de fichiers pair-à-pair*, in "5ème Conférence Française en Systèmes d'Exploitation (CFSE'5), Perpignan, France", vol. 5, ACM-SIGOPS France, October 2006, p. 37–48.

[24] Y. BUSNEL, A.-M. KERMARREC. *ProxSem: Interest-based Proximity Measure to Improve Search Efficiency in P2P Systems*, in "4th European Conference on Universal Multiservice Networks (ECUMN'2007), IEEE, Toulouse, France", to appear, February 2007.

[25] J. CAO, M. RAYNAL, W. WU, X. WANG. *The power and limit of adding synchronization messages for synchronous agreement*, in "35th Int. Conference on Parallel Processing (ICPP-06), Colombus, Ohio, USA", IEEE Computer Society, August 2006, p. 399-406.

[26] A. FERNANDEZ, E. JIMENEZ, M. RAYNAL. *Leader election with weak assumptions on initial knowledge, communication reliability and synchrony*, in "Int. IEEE Conference on Dependable Systems and Networks (DSN-06), Philadelphia (Pensylvania)", IEEE Computer Society, June 2006, p. 166-175.

[27] E. GAFNI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *The committee decision problem*, in "Latin American Theoretical Informatics Symposium (LATIN-06), Valdivia, Chile", LNCS, n^{o} 3887, Springer Verlag, March 2006, p. 502-514.

[28] V. GRAMOLI, E. ANCEAUME, A. VIRGILLITO. *SQUARE: Scalable Quorum-Based Atomic Memory with Local Reconfiguration*, in "22nd ACM Symposium on Applied Computing (SAC'07), Seoul, Korea", to appear, March 2007.

[29] V. GRAMOLI. *From Fast to Lightweight Atomic Memory in Large-Scale Dynamic Distributed Systems*, in "ReSIST Workshop, Pisa, Italy", September 2006.

[30] V. GRAMOLI, A.-M. KERMARREC, A. MOSTEFAOUI, M. RAYNAL, B. SERICOLA. *Core Persistence in Peer-to-Peer Systems: Relating Size to Lifetime*, in "On The Move International Workshop on Reliability in Decentralized Distributed systems (OTM'06), Montpellier, France", LNCS, n^{o} 4278, Springer Verlag, oct 2006, p. 1470-1479.

[31] R. GUERRAOUI, S. B. HANDURUKANDE, K. HUGUENIN, A.-M. KERMARREC, F. LE FESSANT, E. RIVIÈRE. *GosSkip, an Efficient, Fault-Tolerant and Self Organizing Overlay Using Gossip-based Construction and Skip-Lists Principles*, in "6th International Conference on Peer-to-Peer Computing (P2P), Cambridge, UK", September 2006.

[32] R. GUERRAOUI, M. RAYNAL. *A leader election protocol for eventually synchronous shared memory systems*, in "4th Int. Workshop on Software Technologies for Future Embedded and Ubiquitous Systems (SEUS-06), Gyeongju, Korea", IEEE Computer Society, April 2006, p. 75-80.

[33] S. B. HANDURUKANDE, A.-M. KERMARREC, F. LE FESSANT, L. MASSOULIÉ, S. PATARIN. *Peer Sharing Behaviour in the eDonkey Network, and Implications for the Design of Server-less File Sharing Systems*, in "EuroSys, Leuven, Belgium", 18-21 April 2006.

[34] M. JELASITY, A.-M. KERMARREC. *Ordered Slicing of Very Large-Scale Overlay Networks*, in "The Sixth IEEE Conference on Peer to Peer Computing (P2P), Cambridge, UK", September 2006.

[35] E. LE MERRER, V. GRAMOLI, M. BERTIER, A. VIANA, A.-M. KERMARREC. *Energy Aware Self-organizing Density Management in Wireless Sensor Networks*, in "ACM MobiShare, Los Angeles, CA, USA", September 2006.

[36] E. LE MERRER, A.-M. KERMARREC, L. MASSOULIÉ. *Peer-to-Peer Size Estimation in Large and Dynamic Networks: a Comparative Study*, in "15th International Symposium on High Performance Distributed Computing (HPDC-15), Paris, France", 5-9June 2006.

[37] E. LE MERRER, A.-M. KERMARREC, D. NEVEUX. *SONDe: Contrôle de densité auto-organisante de fonctions réseaux pair à pair*, in "Algotel, Tregastel, France", 2006.

[38] L. MASSOULIÉ, E. LE MERRER, A.-M. KERMARREC, A. GANESH. *Peer Counting and Sampling in overlay networks: random walk methods*, in "25th ACM SIGACT-SIGOPS Int. Symposium on Principles of Distributed Computing (PODC-06), Denver, Colorado, USA", ACM Press, July 2006.

[39] S. MONNET, M. BERTIER. *Using failure injection mechanisms to experiment and evaluate a grid failure detector*, in "Workshop on Computational Grids and Clusters (WCGC 2006), Rio de Janeiro, Brazil", July 2006.

[40] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *From failure detectors with limited scope accuracy to system wide leadership*, in "19th Int. IEEE Conference on Advanced Information Networking and Applications (AINA-06), Vienna, Austria", IEEE, March 2006, p. 81-86.

[41] A. MOSTEFAOUI, S. RAJSBAUM, M. RAYNAL, C. TRAVERS. *Irreductibility and additivity of set agreement-oriented failure detectors*, in "the 25th ACM SIGACT-SIGOPS Int. Symposium on Principles of Distributed Computing (PODC-06), Denver (Colorado)", ACM Press, July 2006.

[42] A. MOSTEFAOUI, M. RAYNAL, C. TRAVERS. *Exploring Gafni's reduction land: from $Omega^k$ to wait-free adaptive (2p-p/k)-renaming via k-set agreement*, in "20th International Symposium on Distributed Computing (DISC-06), Stockholm, Sweeden", LNCS, n$^{\text{o}}$ 4167, Springer Verlag, September 2006, p. 1-16.

[43] A. MOSTEFAOUI, M. RAYNAL, G. TREDAN. *On the fly estimation of the number of the processes that are alive/crashed in an asynchronous message-passing system*, in "12th IEEE Pacific Rim Int. Symposium on Dependable Computing (PRDC-06), Riverside, CA, USA", IEEE Computer Society, December 2006.

[44] P. RAIPIN-PARVEDY, M. RAYNAL, C. TRAVERS. *Strongly terminating early-stopping k-set agreement in synchronous systems with general omission failures*, in "the 13th International Colloquium on Structural Information and Communication Complexity (SIROCCO-06), Chester, UK", LNCS, n° 4056, Springer Verlag, July 2006, p. 182-196.

[45] M. RAYNAL, C. TRAVERS. *In search of the holy grail: looking for the weakest failure detector for wait-free set agreement*, in "10th Int. Conference on Principles of Distributed Systems (OPODIS-06)", LNCS, Invited paper, n° 4305, Springer Verlag, December 2006, p. 1-17.

[46] M. RAYNAL, C. TRAVERS. *Synchronous set agreement: a concise guided tour*, in "12th IEEE Pacific Rim Int. Symposium on Dependable Computing (PRDC-06), Riverside, CA, USA", IEEE Computer Society, December 2006.

[47] S. VOULGARIS, E. RIVIÈRE, A.-M. KERMARREC, M. VAN STEEN. *Sub-2-Sub: Self-Organizing Content-Based Publish and Subscribe for Dynamic and Large Scale Collaborative Networks*, in "IPTPS'06: the fifth International Workshop on Peer-to-Peer Systems, Santa Barbara, USA", FEB 2006.

[48] W. WU, J. CAO, M. RAYNAL. *A hierarchical consensus protocol for mobile adhoc networks*, in "14th Euromicro Int. Conference on Parallel, Distributed and Network-based Processing (PDP-06), Montbéliard, France", February 2006, p. 64-71.

Internal Reports

[49] O. BEAUMONT, A.-M. KERMARREC, L. MARCHAL, É. RIVIÈRE. *VoroNet: A scalable object network based on Voronoi tessellations*, Research Report, n° RR-5833, INRIA, Feb 2006, https://hal.inria.fr/inria-00071210.

References in notes

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

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

[52] 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.

[53] A. DEMERS, D. GREENE, C. HAUSER, W. IRISH, J. LARSON. *Epidemic algorithms for replicated database maintenance*, in "Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing (PODC'87), Vancouver, British Columbia, Canada", August 1987, p. 1-12.

[54] 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.

[55] GNUTELLA. *The Gnutella protocol specification*, 2000.

[56] M. JELASITY, R. GUERRAOUI, A.-M. KERMARREC, M. VAN STEEN. *The Peer Sampling Service: Experimental Evaluation of Unstructured Gossip-Based Implementations*, vol. 52, n^{o} 2, February 2003.

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

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

[59] S. RATNASAMY, P. FRANCIS, M. HANDLEY, R. KARP, S. SHENKER. *A Scalable Content-Addressable Network*, in "Proceedings of ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'01), New York, NY, USA", August 2001, p. 161–172.

[60] 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)", November 2001, p. 329–350.

[61] I. STOICA, R. MORRIS, D. KARGER, M. F. KAASHOEK, H. BALAKRISHNAN. *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*, in "Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'01), San Diego, CA, USA", ACM Press, ACM, August 2001, p. 149–160, http://www.pdos.lcs.mit.edu/.

[62] 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^{o} 2, 2005.