SPREADS
Safe P2p-based REliable Architecture for Data Storage

Soumission de Projet ANR
Agence Nationale de la Recherche
Programme “Télécommunications”
Appel à projets de recherche 2007

Ubiquitous-Storage SA,
LACL (Paris XII),
INRIA MASCOTTE project-team,
INRIA REGAL project-team,
EURECOM NS Team

Contents

Introduction

Nowadays, information is essentially digital. The main questions about digital data are: "where to store them ?" and "for how long will they be available ?". Consider for instance the photograph case: today each family snapshot is taken through a digital camera and photographs are stored on hard disks or CDs. Will you be able to retrieve your family snapshots in 1, 10, or 20 years? For companies data storage is also an important responsibility (losing data can quickly bring a company to bankruptcy) for which currently existing data storage solutions are often complicated or constraining (that's why many small and medium companies don't seriously backup their data).

We believe that the future of data storage is on the network through easy-to-use on-line storage systems. New services must be provided to store/retrieve data to/from network in a transparent way.

Network storage can be easily achieved using data-centers on which all data are concentrated, it requires large centralized infrastructures (to ensure the persistence of data it must be replicated in at least two data centers) which means high investment costs, network bottlenecks and sensitivity to attacks.

The peer-to-peer (P2P) paradigm proposes to decentralize the storage among large number of peers distributed in the network to avoid these disadvantages. The counterpart is to design distributive schemes which are more complicated in term of network communications, security and studies as their models are more complicated than for the well-known client/server scheme.

This proposal presents a collaborative research effort to study and design highly dynamic secure P2P storage systems on large scale networks like the Internet, dealing with most of the domains described by the 3^{rd} research theme ("Logiciels pour les télécommunications et réseaux") of this call for proposal. Compared to other popular peer-to-peer data sharing systems, a reliable storage system has a different set of functional constraints. First it must ensure reliability, which means that the system must be self-organized to guarantee that the data will never get lost (in peer-to-peer data sharing systems, the perenniality of the data in the system is due to their popularity among other peers). Second, it must ensure confidentiality, which means that nobody except the data owner (and possibly some trusted third-party authority) should be able to retrieve(exploit) the data.

Partners are the UbiStorage SA company, the LACL (Univ. Paris XII) laboratory, the REGAL INRIA/LIP6 project-team, the MASCOTTE Sophia INRIA project-team and the EURECOM NS Team. UbiStorage is the leading partner which is currently developing and commercializing an on-line data storage system based on the P2P paradigm. The expected industrial outcome of this project is to be able to provide a secure on-line storage system which can deal with several million of Internet users without centralized infrastructure and with low administrative costs. The aim of this proposal is to harness the skills of one industrial and four complementary academic partners from distinct communities to investigate this issue.

The scientific program covered by this proposal is mainly the design of new mathematical safety, security and performance models, secure patterns, simulation to evaluate the quality of service of a peer-to-peer storage system in the context of a dynamic large scale network. These models and simulations will eventually be corroborated by experimentation on the Grid 5000 and Grid eXplorer platforms. Thanks to results provided by simulation and experimentation, improvements to current P2P storage systems will be designed.

UbiStorage is already commercializing a peer-to-peer private storage solution, but its solution does not yet deal with more than thousand of peers, and as it is difficult to test on large scale in a real context, models and simulations have to be done to optimize trade-offs between used resources and guarantees

of the system efficiency. Strategies for perennial maintenance also have to be designed and evaluated. In the current version of the product, security was not the priority (main work was done on the system architecture design and market penetration to validate that this kind of solution can take a part in the data storage market). Now all security aspects have to be taken into account to convince customers and insurance companies that the peer-to-peer scheme allows a high level of security (even if the peer-to-peer paradigm suffers of a bad reputation due to its large piracy usage, its distributed nature makes its robustness, in the way the Internet was originally designed).

With this project UbiStorage wants to improve and validate its solution to be able to market a very large scale storage system with scientific guarantee of efficiency and robustness. Even if some distributed solutions exist on the market, UbiStorage is one of the first to commercialize such a system where storage peers are collocated with customers and on a such large scale while retaining reliability. Theses studies will provide an important body of knowledge strongly needed by the community which will design the distributed systems of tomorrow. Most of the studies which will be done in this project should be usable in other peer-to-peer approaches like collaborative data solutions, distributive decision processes.

Chapter 1

Context and State-of-the-Art

1.1 Scientific State of the art

Persistent storage

The main problem of data storage is to ensure quality of service with transparent high security level: the data will not be lost or stolen and would be easily retrieved, user should also be in confidence with it. A resilient data storage system should have the following characteristics:

- integrity and availability of data. The data should only be lost or altered with a low probability defined by quality of service parameters (redundancy and integrity checking policies).
- confidentiality and security. Access to a private data should be under the control of the owner (cryptography, certification protocols and access control).
- ability to give maximum feedback to the final user. The final user should be able to evaluate on his own the quality of the service such that he can be in confidence with it. That corresponds to the item (“qualité de service fournie/perçue”) in the 3rd thematic of this call.
- ubiquity and access policy. Most people want to access ot their data from anywhere, share them with other users in read or write mode.
- low cost. Infrastructure, management, supervision resources should be low to provide services to most users.
- scalability. Solutions should be able to support five users as well as thousands of users without significant performance issues.

In a seminal work, Luby introduced the idea of using Erasure Correcting Codes [1] in order to achieve high reliability at much lower cost. Those ideas are daily used in modern RAID disk arrays, and their impact on distributed storage has been studied recently [10]. UbiStorage hold exclusivity on a patent on such a system.

To guarantee data survival, it is necessary to introduce a self repair mechanism which detects and rebuilds lost data. The reconstruction uses the redundancy mechanism previously described: dispersed redundancy fragments are collected to recover lost fragments on other peers. In previous works, we showed that P2P storage system has to face a continuous large number of reconstructions to insure data durability [45]. A consequence is that the network is flooded by a continuous data stream necessary to maintain data integrity. New communication and distribution strategy must be designed [37]. Such systems can be hybrid and controlled (some servers are watching the system and faithful users execute a

[1] those codes can retrieve missing bits of a data, but unlike Error Correcting codes they are unable to correct an erroneous bit

rigid protocol) but ideally it should be completely decentralised. This involves issues generally studied in the context of file Peer-to-Peer networks.

A dispersal of data, using such redundancy mechanism was first proposed by Rabin in [35].

Most proposals have addressed security with three objectives: confidentiality (enforced through plain symmetric encryption), integrity (and which may be addressed through digital signature or cryptographic hash functions) and availability (enforced through the verification of storage durability on peers involved). It should be noted that the interaction of protection schemes with redundancy schemes has seldom been taken into account: for instance, some integrity protection is provided by data coding itself, which some proposals simply do not assume to be present; confidentiality protection also may hamper the proper working of replication when required.

Ensuring storage durability, the most original protection scheme in data storage security, first involves the development of appropriate protocols to check that the data stored are still owned by one peer. Deterministic protocols like [5, 16, 19] to name a few, in which a fresh "signature" can be extracted out of the data stored, are the most frequent type of such protocols. Probabilistic protocols [30, 34], in which verifications consist in randomly checking a slice of the data, make it easier to limit the verification overhead. At the same time, the latter provide a more powerful and realistic approach to verifying the durability of data storage in a distributed architecture, whereas the former is more attractive for securing data backup, in which case the original data are still available to the verifier for comparison.

Verifying storage durability makes it possible to assess the cooperation of a node towards achieving storage and to react accordingly. Two behaviors should be prevented among the peers participating to the system: selfishness is represented by nodes that do not want to offer storage and by nodes that destroy part of the data they had previously stored when they are out of space; maliciousness is represented by nodes that perform deliberate attacks on other peers or that systematically destroy the data they receive, without trying to obtain any storage advantage out of it. Most often, reacting to such attacks has essentially consisted in creating new replicas of the data, and avoiding peers perceived as selfish or malicious, but more systematic studies for designing adapted cooperation schemes are still at an early stage of investigation [33] compared with the work achieved in cooperation for mobile ad hoc routing and forwarding.

Peer-to-Peer networks

Peer-to-Peer data sharing have first emerged with centralized control, the idea was simple: reduce the server load (bandwidth, CPU) by limiting the server duty to manage the system while the actual data will be transmitted using cross-links between the peers. Despite the huge success of those networks (Napster grew up to 80 millions of user before being bared down by the Recording Industry Association of America), their shortcoming were clear: such hybrids are very vulnerable (disabling one or a few servers destroys the system), they are not scalable, and difficult to support economically. The next step was to study the following issue: removing any centralized control. First some naive answers were provided, the network was self-organized but unstructured. To query a data a peer would literally either flood the network (gnutella, edonkey), or look randomly in the hope of finding it [22]. The efficiency of such systems is low, and scalability is not achieved since one peer may need to query all the others to retrieve data [2].

The major answer came with a so called Distributed Routing Protocol (DRP). A distributed routing protocol is a routing scheme enabling one peer to contact another one just by knowing its ID (one usually assumes that IDs are binary strings of $\Theta(\log n)$, equiv. numbers from $[1, n]$). This must be achieved with limited memory and time (scalability) in a dynamic network. In the non dynamic case, networks like butterflies, shuffle-exchange hypercubes or hypercube-like networks in which x is adjacent to $x + 2i$ achieve this goal. The dynamic analogue of hypercube is the chord network [43, 29]. In Chord like networks, nodes are organized on a ring of size n, each nodes knows about $\log n$ nodes at distance $2i$, $i = 1, 2, \ldots, \log n$. In such networks, searching a node takes order of $\log n$ hops (ie TCP&IP connections), a node can come and go and the probability to be unable to contact an existing node is close to zero. Chord is one practical aspect of the small world phenomena [23], and it can also be

considered as similar to Milgram experiment, if one considers an ID as a field, and if a node queries about a particular field, each node knows another node (which is significantly more expert than it) in that field. Several scalable DRP, all quite similar to Chord, have been proposed (CAN [39], PAST [40], tapestry [49]).

A DRP can be considered as the core of the current and future peer-to-peer networks. The routing is there established over the Internet in a virtual network on which routing is easy to perform. From a DRP one can provide a Distributed Hash Table (DHT). For this it is enough to use a good hash function that maps keys on peers IDs. The DHT can be made resilient by using a suitable amount of replicas. File systems can then be implemented above this layer like in [12].

Peer to Peer file systems

Although many peer-to-peer file systems have been proposed by different research groups during the last few years [26, 7, 18, 1], only a handful are designed to scale to hundreds of thousands of nodes and to offer read-write access to a large community of users. Moreover, very few prototypes of these large-scale multi-writer systems exist to this date, and the available experimental data are still very limited.

One of the reasons is that, as the system grows to a very large scale, allowing updates to be made anywhere at anytime while maintaining consistency, ensuring security, and achieving good performance is not an easy task. Multi-writer designs must face a number of issues not found in read-only systems, such as maintaining consistency between replicas, enforcing access control, guaranteeing that update requests are authenticated and correctly processed, and dealing with conflicting updates.

The Ivy system [32], for instance, stores all file system data in a set of logs using the DHash distributed hash table. In Ivy each update is stored by appending a record to a log. Since records are never removed from the logs, every client has access to the whole file system history, which greatly simplifies conflict detection and resolution. Furthermore, each Ivy user has its own log to which he appends his own updates. This has two advantages: first, writes are fast since there is no central serialization point (like Oceanstore's primary tier), and second, data cannot be overwritten by a malicious user since only the log owner can append data to it. However, as the number of users sharing a given file system increases, the number of logs that need to be traversed to satisfy a read operation also becomes larger, thus increasing network traffic. Although the number of DHash servers can grow to hundreds of thousands, the number of Ivy users sharing a given file does not scale. Another problem in Ivy is that applications have little control over the consistency of data. Although Ivy uses a consistency model similar to close-to-open consistency, applications cannot fully decide when written data are propagated to the network (this due to the lack of a CLOSE RPC in the NFS v3 protocol specification).

Oceanstore [25] uses a completely different approach for updates handling by introducing some degree of centralization. A primary tier of nodes uses a Byzantine-fault tolerant (BFT) [6] algorithm to serialize all file system updates coming from secondary tier nodes. Since BFT is quite expensive, primary tier nodes must be highly resilient nodes located in high-bandwidth areas of the network. Oceanstore's designers assume that these nodes will be set up and maintained by a commercial service provider. Thus, Oceanstore may not be suitable for a community of cooperative users wishing to use a system which does not depend on a centralized authority. Oceanstore also takes into account network locality to optimize replica location. An introspection layer provides information about the network conditions, allowing the system to dynamically adapt itself to the current environment. Locality management is absent in Ivy.

Pangaea [42] differs from Ivy and Oceanstore because it does not rely on a key-based routing layer. Instead, object location is achieved by maintaining a graph of live replicas through which updates are propagated by flooding a special message called harbinger. Moreover, a replica of a file or directory is created on each client that accesses the file. Although this reduces read latency, it can generate an important amount of traffic when updates are propagated.

Computer Simulations

Computer simulations of large scale distributed applications is still an open issue. Some specialized simulators, such as the SimGrid project[2], appeared in the last few years and start to be functional. However, the largest configurations which can actually be simulated in a reasonable computation time hardly exceed a few thousand nodes. With P2P this limit needs to be extended to as much as several hundreds thousands nodes and even up to millions nodes! In order to cope with this extremely challenging scalability issue, the complexity of the model must be lowered everywhere where it makes sense. However, here, the complexity has not the usual meaning of complexity theory. Several definitions of the notion of complexity in simulation have been proposed, such as the ones given in [48] or [8]. Despite the fact that the vocabulary used is slightly different, the idea is the same: the complexity in simulation is a quantity that depends both on the size of the model and on the detail level of the model (also called the model resolution). In our case, the size of the model corresponds to the number of nodes in the P2P system, and the resolution represents how accurately the behavior of each peer is reproduced. In order to reduce the complexity of the model, the key issue to address is to decide which part(s) of a (large) model have a significant impact on the results produced by the simulation, and which parts have a low or null impact on the results. In [48], Zeigler et al. define the complexity as a product of the size and the resolution of the model. Following this definition, when the size of the model is fixed to a very large value, the only way to reduce complexity is to dramatically reduce the resolution. Unfortunately, a simplistic application of this principle will inexorably lead to wrong or useless results. However, several "enabling" techniques have been proposed to implement such reduction strategies[48]:

- ignore some parts of the model that do not significantly impact the result of the simulation
- group and simulate together, as a single entity, several elements of the model instead of simulating independently each of these elements
- replace a deterministic model/variable by a stochastic one or conversely, depending on which is the most complex to compute
- change the modeling formalism to the most efficient one depending on the problem
- model discrete parts of the system as a linear system

Very few of these techniques have been used so far for the modelling of vary large P2P systems. Some pioneering developments, such as the PeerSim project, for example[3], that combines two simulation techniques, have started to explore such directions. On the other hand, another (orthogonal) approach for dealing with this complexity consists in increasing the computation power of the computer used for running the simulation. The classical approach for this is to use parallel and distributed techniques, such has the ones presented by Fujimoto in [20]. Indeed, several parallelisation approaches may be envisaged in the case of simulation. Fujimoto classifies these approaches in two categories: the conservative ones, in which the processing order of the events in the simulator is strictly enforced, even between the entities that execute in parallel, and the optimistic ones, in which the order is relaxed between the parallel processing units, but at the expense of a roll-back operation when an inconsistent simulation state is reached.

Specification and verification of communication protocols

Verification by model-checking [9, 3] is a well-established, industrially applied methodology that can be integrated into the development circuit, and, from its early starts, has been applied in protocol development. Various tools exist, like the popular Spin model checker [4] that targets software verification by

[2] http://simgrid.gforge.inria.fr/

[3] http://peersim.sourceforge.net/

[4] http://www.spinroot.com/

means of Linear Temporal Logic specifications and Buchi automata models of systems, or the CADP toolset[5] that supports several specification formalisms (temporal logics, mu-calculi, bisimulation-based, automata-based) and verification techniques, based on various data structures (BDDs, enumerative) or various techniques for speeding-up the verification process (partial orders, compositional or distributed model checking).

Verification of infinite-state systems is tackled by various techniques, of which abstraction techniques [11] play an important role; an alternative to abstractions is the so-called "regular model checking" approach. Recently, the two approaches have been combined into a so-called "abstract regular model checking" formalism [4].

Recently, there has been a growing interest in extending the model-checking approach to larger classes properties, like the epistemic properties related to the "knowledge" of agents in a multi-agent system. Some tools exist for decidable subclasses of the temporal logic of knowledge, like the MCMAS tool [6] or the MCK tool [7]. Research is more at its beginning on the aspects related to handling infinite-state systems by abstractions [17]; some authors [44, 46] propose coding knowledge formulas into existing temporal frameworks, where this can be done. The distributed model checking problem does not seem to have been addressed yet.

Specification-based testing of security protocols

The domain of security protocols verification has seen an important number of improvements in the last decade, and a number of tools are already available for protocol analysis. The AVISPA tool[8] is a joint project that aims at “developing a push-button, industrial-strength technology for the analysis of large-scale Internet security-sensitive protocols”. We may also mention the ISpi tool[9], which is based on equational reasoning.

Testing of security protocols has been addressed in [47], where tests are generated using so-called "mutations", which are similar to fault injections, but into the specification. The CADP toolset also provides a test generation from Lotos specifications, but it does not seem to have been applied to testing security protocols.

1.2 Economic Context

We address the on-line backup sector. Currently, all competitors provide their services through a client server approach based on data-center. Peer to peer aproach is a good alternative to this client/server model because this technology avoid the needs for heavy infrastructure.

Today, Peer-to-Peer systems (P2P) are widely used mechanisms to share resources on Internet at a very large scale. Very popular systems were designed to share CPU (Seti@home, XtremWeb, Entropia) or to publish files (Napster, Gnutella, Kazaa). These systems are able to carry on several thousand of peers, in spite of the high dynamicity and heterogeneity of the network. In the same time, some systems were designed to share disk space (OceanStore [26], Intermemory [7], PAST [18], Farsite [1]). The primary goal of such systems is to provide a transparent distributed storage service over Internet.

These systems share common issues with CPU or files sharing systems: resource discovery, localization mechanisms, dynamic point to point overlay network infrastructure... But for sharing disk systems, data lifetime is the primary concern. P2P CPU or file publishing systems can deal with node failures: the computation can be restarted anywhere or the published files resubmitted to the system. For disk sharing systems, node failure is a critical event: the stored data are definitively lost. So data redundancy and data recovery mechanisms are crucial for such systems. To provide such a distributed data storage service,

[5] http://www.inrialpes.fr/vasy/cadp/

[6] http://www.cs.ucl.ac.uk/staff/f.raimondi/MCMAS/

[7] http://www.cse.unsw.edu.au/ mck/

[8] http://www.avispa-project.org/

[9] preliminary version, http://www.lsv.ens-cachan.fr/ goubault/ISpi/

more constraint is put on the overlay network. Today, none of these systems was able to demonstrate its feasibility and effectiveness for a number of peer greater than 10^2!

At this time, some companies are proposing data backup systems based on P2P technology:

- Vembu (StoreGrid)
- Pensamos (Magic Mirror Backup)
- Hispread
- 312inc (BitVault)

But their products are restricted to small scale networks, typically for intranets.

To our best knowledge, we are the first to address the definition of on-line backup P2P storage system for highly dynamic large scale network.

1.3 Market analysis

On-line data backup is a fast growing market[10] (greater than 50%/year). The 2005 annual sales turnover in France was close to 50M€.

The professional market may be segmented according to company size. In 2004, the annual sales turnover for on-line backup in France for SME/SOHO (resp. Large Enterprises) was equal to 21M€ (resp. 1M€).

Currently, the on-line data backup sector is dispersed. In France, there are several dozens of actors. The leader is Adhersis (15M€ of sales turnover in 2004), with 15 000 customers over the EU. Challengers are OODRIVE, Backup Avenue, and AGS Backup.

The sector is dispersed because:

- the market is new and attractive: competitor seize the advisability of taking significant parts of market from the very start. Moreover customers are in general captive with regard to data backup solution (it is hard to change) which guaranteed to you recurrent income;
- the technological barriers are weak: to propose an on-line data backup solution to your customers, what you need is a data center and an on-line backup software.
- the Infrastructure are costly: to have a big part of the market, you need a large data-center, the break-even point is high.

We observe the arrival of new actors on the market:

- Software data-backup editors: some of them make partnership with data-center providers to propose optional on-line backup with their software (eg. Neobe Backup with RedBUS Interhouse).
- Web hosting providers: they propose private backup space for their customers.
- ISP and telecom operators: Xdrive, the first on-line data backup company was acquired by AOL. KPN, the main telecom operator in Netherland, proposes an on-line data backup service. FT provides such a service with the *securitoo* package.

Peer-to-peer backup system is clearly a disruptive technology in the market of data backup: it allows to take a large part of the market without heavy investments.

[10] Source from ACTEMIS Consulting.

Chapter 2

Partnership

2.1 Ubiquitous Storage SA

UbiStorage SA is a french company issued of several years of academic research. The company was created in the framework of the french innovative laws (Loi française sur l'innovation, juillet 1999). Its founder was involved in two French national ACI GRID grant devoted to distributed P2P applications and storage (CGP2P and DataGraal). Previous academic work focused on data durability and data distribution in peer to peer system [37, 45, 36, 38]. The team was also inventor of a patent on distributed P2P storage system. The project of company creation was two times laureate of the grant for innovative company creation project ("concours d'aide à la création d'entreprises de technologies innovantes" category "emergence" and "création développement"). The project was also supported by the HEC-Challenge+[1] help management program for innovative project, and by a regional "company incubator" (Incubateur Régional de Picardie) until August 2006. The incubator provided a grant to finance a post-doc position in 2005, who developed the main part of the system currently sold. The UbiStorage company was created in early 2006 and is currently addressing the online backup market for small and medium companies.
Know-how and competences: Research, Development and Commercialisation of P2P storage system for small business companies.
UbiStorage is not involved in any other project.

2.2 LACL - Paris XII

LACL is the computer science laboratory of Paris-XII University. Its "communicating systems" team works on tools for the safety and security analysis of communicating systems. This research has attracted several national and international grants and is ongoing within the ACI-SATIN national project (between LORIA, LIFO, LACL, France-Telecom R&D). LACL researchers are also specialising in parallel programming, and this researche has also attracted various fundings and an EADS PhD prize. Our techniques for high-performance declarative programming are now being applied to the parallelisation of modelchecking and to the audit of safeness-secureness properties.

Know-how and competences: specification, modeling and verification of protocol security properties. High-performance computing and its application to verification.
LACL research in those areas has been funded by national basic-research grants (ACI), has received one national EADS thesis prize, appears in numerous international publications and has been supported by several international grants (bi-lateral projects).
Involvment in other projects:

- [2002-2004] Coordination of ACI-GRID CARAML "CoordinAtion et Répartition des Applications Multiprocesseurs en objective camL" a national basic-research project ended in 2004. High-

[1] http://www.hec.fr/hec/eng/instituts/startup.html

performance declarative programming as developed by CARAML at LACL is directly applicable to the extensive verification problems faced by the UbiStorage partner.

- [2004-2007] Participation in ACI-Sécurité SATIN "Security Analysis for Trusted Infrastructures and Network protocols" a national basic-research project ending in July 2007. The main issue addressed by SATIN is the symbolic and automatic verification of security properties.

2.3 INRIA - Projet MASCOTTE

Mascotte is a joint project-team of CNRS, INRIA and the University of Nice Sophia-Antipolis. It involves 25 members among them 14 permanent researchers (from CNRS, INRIA and University) and it is strongly associated with FranceTelecom R&D.
Mascotte's main objective is to develop methods and tools for efficient use of telecommunication networks. This involves high level research in the fields of simulation, algorithms, and discrete mathematics. Mascotte's work in the telecommunications field relies on a strong theoretical background (graph theory, combinatorial optimization, discrete probabilities), with a large number of publications in the best international conferences and journals in fields like distributed computing, efficient telecommunications, graph theory or radio networks. Experience in simulation includes the specification and development of several software platforms (Prosit, ASIMUT, OSA) and the modeling expertise has been applied both to computer networks and road traffic simulations.
Mascotte has developed industrial collaborations with various partners such as France Telecom, Alcatel, and CNES, for the design and optimization of telecommunication networks (for example with RNRT PORTO). Mascotte was involved in various projects funded by the EC, in particular recently in the FET CRESCCO projects and RTN ARACNE project, and is currently involved in the Aeolus FET project whose main goal is to study overlay networks.
It has also many bilateral cooperations with European countries and also with Canada, Brazil, Israel.
Know-how and competences: modeling, simulation, analysis and optimization of large network architectures.
Involvment in other projects:

- [2005-2009] IST-FET "AEOLUS" project (Overlay computers)
- [2005-2008] ANR "Jeunes Chercheurs" OSERA (Meshed Radio Networks in Urban Areas)

2.4 EURECOM - NS Team

Eurecom Institute is a graduate school of engineering and research institute in telecommunications located in Sophia Antipolis, France. It is organized as a consortium of industrial and academic members including EPFL, Télécom Paris, Telecom INT, ETH Zürich, ENST Bretagne, Swisscom, Hasler Stiftung, Thalès, SFR, France Telecom, HITACHI Europe, Texas Instruments, ST Microelectronics, Bouygues Telecom, SHARP, Cisco Systems, BMW Group Research & Technology, Politecnico di Torino, and Helsinki University of Technology.

With more than 70 researchers in three research departments (computer communications, multimedia and mobile communications), Eurécom has been actively involved in the ACTS, Telematics, TEN TELECOM, and IST programmes (BETEUS, NICE, SUZIE, WEB4GROUPS, WAND, WITNESS, MOBY-DICK, SPATION, PRIME, NEWCOM, ...), as well as in the FET programme (MOBILEMAN) and in the Safer Internet Action Plan (3W3S). Eurecom has also been involved in a number of national research projects supported by the French RNRT program (SEVA, ICARE, METROPOLIS, VTHD++, COBASCA, PLATON, ANTIPODE, @IRS++, ...) or ACI programs (SPLASH, MOSAIC, ...).
Know-how and competences: The Computer Communications Department is actively investigating

self-organized systems, like peer-to-peer architectures, in terms of protocols and their performance analysis as well as in terms of their security.
Involvment in other projects:

- [2004-2007] ACI SI MOSAIC - Collaborative data backup for mobile systems
- [2006-2009] IST IP R4egov - Secure e-Administration
- [2006-2009] IST NoE ReSIST - Resilience in IST
- [2006-2009] IST FET Cascadas - Autonomic services
- [2006-2010] IST FET Haggle - Secure opportunistic networking

2.5 INRIA - Projet REGAL

Regal is a joint project-team with LIP6 (CNRS and Univ. Paris 6). The Regal project-team aims to manage resources in large scale networks. Regal investigates solutions to deploy applications (with code and data) in highly distributed environments. The project targets large scale configurations (in terms of the number of nodes and distance between them), highly dynamic (with failure, disconnection and partitioning). Regal is focused on replication techniques to tolerate failure, to increase the availability, and to provide efficient access to distributed services. Research themes include resource management in distributed systems, monitoring and failure detection, fault tolerance, reactive replication and dynamic adaptation of virtual machines. Regal was involved in the following international and industrial relations: RNRT Cyrano, GEMPLUS, IST COACH, AS CNRS Operating Systems, ACI GRID2, and ACI DATAGRAAL.
Know-how and competences: distributed algorithms, peer-to-peer technology, fault-tolerance, experimentations in grid environments
Involvment in other projects:

- [2003-2006] ACI MD Grid Data Service
  Members: IRISA (Paris Team), ENS-Lyon (LIP - Remap Team), Regal
- [2003-2006] ACI Data Grid eXplorer
  Members:IMAG-ID, Laria, LRI, LAAS, LORIA, LIP Ens-Lyon, LIFL, INRIA Sophia Antipolis, LIP6, IBCP, CEA, IRISA INRIA Rocquencourt
- [2004-2007] ACI Gedeon
  Members: IMAG-ID, IMAG-LSR, IBCP, Regal
- [2005-2008] ANR (ARA MDSA) Respire
  Members: LIP6, Atlas (IRISA), Paris (IRISA), Regal

Chapter 3

Project Organisation and resources

This project, planned for 36 months, is made of four Work Packages (WP). Each WP will end by a workshop presenting all works done during the WP. The WPs will be composed of sub phases of research, development, simulation, experimentation.

The chart of Figure 3.1 represents the human resources repartition on the four aspects of this project.

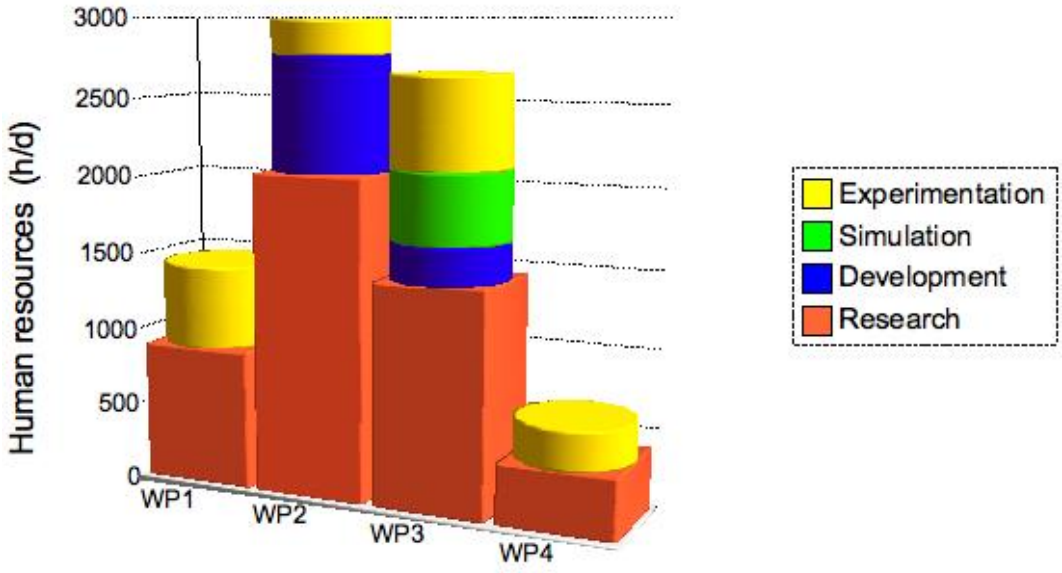


Figure 3.1: Human resources repartion

3.1 Studied problems

3.1.1 Specification and formal verification of the communication and distribution protocols used

This part of the project is concerned with the specification of the intended functionalities of the communication protocols and the server distribution, and the verification of the functional parts of the system w.r.t. the specification. During the very early stages of the project, we will identify appropriate formalisms for specification, in the framework of modal and temporal logics. The verification will be done

using abstraction techniques developed at LACL.

Two variants of the system will be model-checked: a single server one and a distributed server one. For the second variant, we intend to identify appropriate parametric model checking techniques, by considering that the number of servers is a parameter of the deployment, that may even be dynamically modified.

One of the functionalities that will be specified and verified is the following: the system will output, in "real time", a "view" of its state to the end-user, in order for the user to gain confidence in the system, as he/she will achieve some essential information about the system state. Specifying and verifying this type of properties implies the use of modal logics of knowledge, specific in the study of multi-agent systems. We intend to use existing techniques for embedding such properties into the temporal logic framework.

We intend to utilize existing open-source model checkers, that will be used "in parallel" to check a given specification formula, using various techniques for parallelization.

Recent work [31, 21, 41] has also addressed security issues in collaborative multiprocessing environments, as well as tools for the analysis of performance and information flow in protocols and distributed systems.

3.1.2 Optimisation of data placement

In the context of peer-to-peer storage, a file is broken into pieces by the originator of the data that need to be stored. The main goals of such a system could be summarized as follows:

- Goal 1: replicate (with redundancy) the file onto multiple destination.
- Goal 2: restore the original file at the source.

Goal 1 could be seen as multi-message gossiping (or multicast): a (possibly large) amount of data is sent from one node to multiple destinations. Goal 2 could be seen as a reverse gossiping: multiple sources need to send (possibly redundant) data to a single destination. In this work, we set aside the challenging task of designing efficient techniques for content distribution and focus on data object placement strategies, as driven by user requests. The following sketch represents our system model.

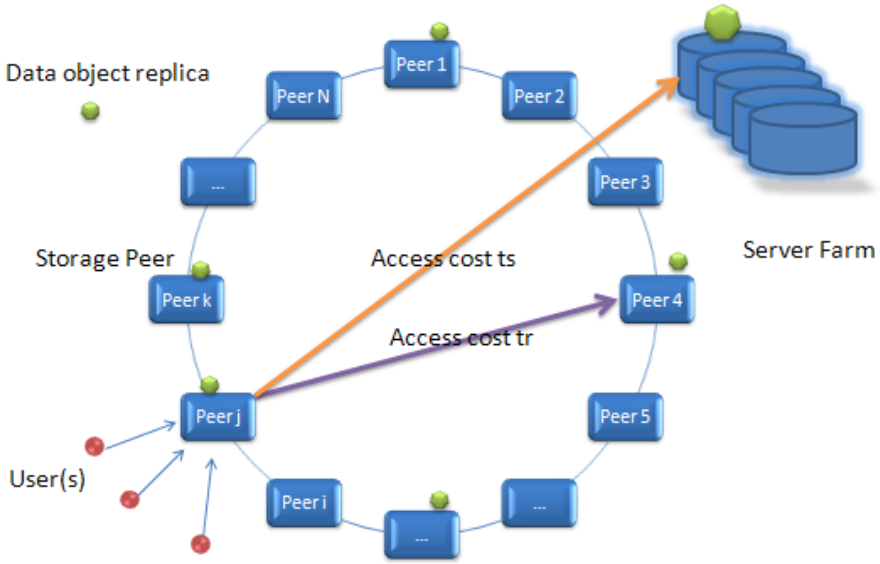


Figure 3.2: A sketch of the p2p storage system model.

As shown in Figure 3.2, we assume storage peers (i.e. the storage devices) to be organized in an virtual network, that we call the p2p overlay. In the Figure the overlay network takes the form of a

structured overlay (e.g., a DHT) but our system model is suitable for an unstructured approach, wherein nodes can form an overlay with arbitrary topology. We assume the presence of a set of centralized entities (a Server, in the terminology used for this proposal) that we call a server farm. Note that our model supports also the degenerate case in which the server farm is composed of one server only. With the aim of extending the original system architecture of the proposal, we assume the server farm to carry out, besides the role of a centralized data dispatcher, the role of a centralized storage facility. In other terms, users' data may potentially be stored on the server farm. Lastly, in Figure 3.2 it is possible to identify the system users. Our model extends the typical assumptions of a dedicated storage peer per user to take into account the possibility that a single storage peer might be used by multiple users. Users are characterized by a request rate to access to the data objects stored in the P2P system, and the cost for accessing an object varies depending on the location the object is stored into. We assume access costs to be ts when data is stored at the server farm, tr when data is stored in a remote storage peer, and tl when the data is available locally. Note that $ts > tr > tl$.

In such a system, we suggest to study the situation in which cooperation among storage peers is the key issue: while no peer has a strong incentive to cooperate, each peer needs that most of (or all) the others cooperate for the service to be of some value. Intuitively, in a P2P storage system, data should be uniformly spread over the peers taking part to the storage overlay network to ensure a certain level of system fairness, which would mean that:

- The situation in which a minority of peers stores the majority of data should be avoided;
- Data (or fractions of it) should be retrieved or stored in almost the same time;
- The impact of peer joins and leaves (i.e. peer churn) should be minimized: system dynamics (for example in terms of user requests, and peer churn rate) should have similar costs for every peer.

Random distribution may appear as a good strategy, but all peers may not exhibit the same behavior or be in the same context. As a simple example, in order to ensure that data will not be lost after an earthquake, all fragments of the same data should not be stored on peers from the same geographical area. The literature is rich of efficient schemes [37, 36, 38] making use of a structured approach in the data distribution phase, wherein combinatorial structures are used to improve redundancy. On the contrary, in this task we will investigate a new set of constraints that derive from the observation that peers may not abide a centralized set of rules (or algorithms) and deviate from the "socially optimal" object allocation rule if this is beneficial. Specifically, we introduce the notion of "soft-security" by relaxing the Boolean assumption made in traditional security schemes wherein agents are assumed to be malicious attackers or legitimate entities. We allow agents to belong a new class of illegitimate parties whose goal is not to subvert the system, although their actions could jeopardize its functioning. Under this perspective, every entity of the (distributed) system calibrate their object allocation rules to maximize their benefits, or conversely, to minimize their costs. We call this abstract model, which finds its roots in game theoretic modeling, the distributed selfish storage problem. In our model we define the set of players to be the storage peers: the set of strategies available to each player correspond to data object placement. We then model the set of utilities for each individual player, under an object placement, the capacity of a peer and the request patter by users to access data, and study equilibrium object placement strategies (if any exists) that derive from the uncoordinated action of all system players. Based on previous work on distributed selfish replication and caching [28, 27] we will first examine the case in which access time to the objects stored in the system is minimized. In this scenario, we will delve into techniques to actually find equilibrium object placement (in game theory, a Nash equilibrium is a descriptive concept, as opposed to a prescriptive one). We will then extend the model to take into account other objectives a peer may have, such as maximizing the reliability and/or availability of the stored data. A key concept that we will further examine is the "price of anarchy" [24], which informally states the performance loss of letting system participants acting in a selfish way as opposed to a "socially maximizing" centralized solution. Our work can be placed in between a purely local approach (for example a greedy strategy to place objects on a set of storage peers) which does not require global information, and a centralized approach (the optimal one) which requires full knowledge of the system state.

3.1.3 Securisation of protocols

We will specify the security properties - confidentiality and various types of authenticity - that are meant to be ensured by the security protocols. We will also specify safety properties of the security protocols related to the protocol construction, like strong typing and timestamp management.

On the other hand, we will model attacker capabilities against which the system is supposed to be secure - these may include the capacity to inject, from different sources, fake packets into the network, or the existence of malicious participants that want to attack others. Attacker modeling will consider only perfect cryptography, hence focusing explicitely on the conception and implementation of communication protocols and not on cryptographic strength.

From (negations of formulas giving) protocol specifications and from the attacker models, we will generate attack scenarios, in the form of conformance testing. The attack scenarios will then be experienced on the system.

3.1.4 Handling protected data

To meet several classical security requirements ranging from confidentiality and privacy to content protection through source identification and conditional access control the storage systems have to protect the data segments within the storage or during the transfer operations using various data protection mechanisms such as encryption, integrity protection and digital signatures. Most data handling operations performed as part of data storage or communications are not compatible with these protection mechanisms since either the data handling operations require clear-text access to the entire data segments or to parts thereof or data handling operations disrupt the effect of data integrity mechanisms in case of clear-text data protected with integrity mechanisms only. For instance, data forwarding mechanisms would be in conflict with encryption since the former require access to some data header in clear-text in order to be able to perform basic forwarding operations. Similarly, a lookup request cannot be served by a simple storage system if the data kept in storage is encrypted. Conversely, integrity mechanisms such as Message Authentication Codes (MAC) would be affected if integrity-protected data segments have to be merged by some intermediate nodes that do not have access to the secret MAC function. The conflict between data handling and security transforms is a fundamental research problem that calls for new security solutions. We plan to investigate possible solutions based on homomorphic security transforms for encryption, integrity and authentication using cryptographic primitives such as encrypted keyword search, searchable encryption, private information retrieval, and computing with encrypted data.

3.1.5 Enforcement and verification of self-organized data storage functions

A mandatory requirement in self-organized systems is for cooperation enforcement among the peers. In self-organized systems the overall function (data storage, communication, file replication, etc.) is carried out through the contribution of several peers. In an unmanaged setting whereby peers do not have any incentive to cooperate, the overall operation can be jeopardized if the individual rate of peer cooperation falls below some threshold. Past work on ad hoc networks and peer-to-peer file replication intensively addressed the problem of cooperation enforcement through various methods based on reputation, rewarding and threshold cryptography. The cooperation enforcement problem in self-organized data storage is inherently harder than with packet forwarding in ad hoc networks or with peer-to-peer file replication. Cooperation enforcement mechanisms rely on a basic operation through which a peer's elementary cooperation with respect to the overall operation can be verified. While in case of ad hoc networks and peer-to-peer file replication, this verification can consist of immediate monitoring of a response from the peer party (the peer forwards a packet; the peer provides a file share upon request), in the case of storage systems, the elementary cooperation verification is not amenable to an atomic response verification since the storage operations take place over a long duration. For instance, in order to check if a peer properly cooperates with respect to the storage of some data segment, the basic verification mechanism should perform several tests over a significant amount of time. Cooperation enforcement

for self-organized data storage is made even more difficult when taking into account the fact that data segments can be protected with some encoding transform such as erasure correcting codes or with some security transforms such as encryption, integrity, or digital signature. We plan to investigate possible solutions for cooperation enforcement in self-organized data storage along two directions: cryptographic primitives for the verification of encrypted stored data and general cooperation mechanisms based on these primitives. The design of such cryptographic primitives is a fundamental research problem with a potential impact broader than the scope of self-organized storage area. Regarding data storage, obtaining assurances about security while enjoying correct performance with such primitives is an especially important area of research that will require resorting to theoretical tools like game theoretical analyses as well as probabilistic and stochastic models.

3.1.6 Multiple writer system

With the aim of finding a solution to the shortcomings of multi-writers systems the REGAL team has designed Pastis, a highly-scalable, completely decentralized multi-writer peer-to-peer file system. For every file or directory Pastis keeps an inode object in which the file's metadata are stored. As in the Unix File System, inodes also contain a list of pointers to the data blocks in which the file or directory contents are stored. All blocks are stored using the Past distributed hash table, thus benefiting from the locality properties of both Past and Pastry. The system is completely decentralized. Security is achieved by signing inodes before inserting them into the Past network. Each inode is stored in a special block called User Certificate Block, or UCB. Data blocks are stored in immutable blocks, called Content-Hash Blocks, the integrity of which can easily be verified. All blocks are replicated in order to ensure fault tolerance and to reduce the impact of network latency and throughput limits. However, persistence can only be achieved if nodes are highly available, that is, if they stay most of the time connected to the overlay. Churn (i.e., nodes connecting and disconnecting from the overlay) in peer-to-peer networks is mainly due to the fact that users have total control on theirs computers, and thus may not see any benefit in keeping its peer-to-peer client running all the time. This is very common in existing peer-to-peer file sharing networks, as many users connect to the overlay to download a particular file, and disconnect soon after the download has finished. Although intermittent connections are not particularly harmful in file sharing networks, this kind of unstable user behavior is undesirable on DHTs. Contrary to file sharing systems, DHTs are designed to guarantee data persistence. This is achieved by replicating data blocks on geographically dispersed nodes, which minimizes the probability of correlated failures, and by regenerating replicas as soon as they leave the network so that the replication factor is kept constant. This reduces the risk of data becoming unavailable if all replicas leave the network, but it also means that as nodes join and leave the network the DHT maintenance algorithm needs to transfer a large number of replicas from one node to another, consuming a lot of bandwidth. Furthermore, DHTs clients lack any flexibility to choose where their data is stored in the overlay. In the project we propose to study the effect of churn on multi-writer DHT. We propose two main tasks:

- Studying incentives-based mechanism to increase the availability of DHT nodes, thereby providing better data persistence for DHT users. High availability increases a node's reputation, which translates into access to more DHT resources and a better Quality-of-Service. The mechanism required for tracking a node's reputation is completely decentralized, and is based on certificates reporting a node's availability which are generated and signed by the node's neighbors. An audit mechanism deters collusive neighbors from generating fake certificates to take advantage of the system.

- Modelling of replica loss and replica repair in P2P systems by a simple Markov chain model, and derive an expression for the lifetime of the replicated state. Then applying this model to study the impact of practical considerations like storage and bandwidth limits on the system, and describe methods to optimally choose system parameters so as to maximize lifetime.

3.1.7 Erasure Correcting Codes DHT

In DHTs, the durability of data is obtained by a duplication mechanism. Each data has a key and is sent to some peers whose ids are closest from this key. If a peer disappears, then a mechanism on these backup peers duplicates the data on another close peer. So, if each data is duplicated on n peers, data can be lost only if more than n peers disappear in almost the same time, and the total memory space needed is n times the size of the data.

Using Erasure Correction Codes (ECC) allow to use less memory space for the same fault-tolerance constraint. For example in the case of Reed-Solomon codes, data are split up in s fragments and r fragments of redundancy are generated for the s ones, then the $s + r$ fragments are sent to some peers in the network. The key point of the usage of these codes is to retrieve a data initially consisting of s fragments, only any s fragments from the $s + r$ sent need to be retrieved. To compare with replication: if $r = n$, still n peers can disappear without lost of data but the total needed memory space is only $1 + n/s$ times the size of the data. The counterpart of using ECC is that data maintenance needs more communications (In [45] it has been proved that ECC are the best in case of highly available peers).

In order to have a robust storage system, not only the data must be spread redundantly on several peers, the management of the data (which mainly means where data are, how many fragments of each data still remain in the system, let reconstruct some fragments, ...) must also be done in a distributed way. Even if one uses ECC instead of DHTs to not replicate data, all the indexes (which can be named as meta-data), may be stored in a DHT because this meta-data does not require much space. Then we aim to define a new kind of generic DHT which store data using ECC and meta-data with replicas and then ensure the maintenance and integrity of the data linked to this meta-data. When using ECC, the following parameters have to be take in account:

1. type of ECC used (Reed-Solomon, Tornado like, ...)
2. redundancy factor: the ratio r/s between the number of fragments generated from the initial data (r) and the number of fragments in which the data is split up (s).
3. Minimum Time Before Failure (MTBF) of the peers, when a peer definitively disappears, some fragments of the data stored also disappear. Following, some data will see their redundancy level decrease and it is necessary to have a mechanism for maintaining the redundancy to a safe level.
4. connectivity of the peers in the network,
5. bandwidths of the peers,
6. memory space of the peers,
7. communication structure in case of overlays.

In that part we want to study the interactions and trade-offs between these parameters. Some parameters are under control (1,2,7), some are not or depends of the peer (3,4,5,6). Therefore the controlled parameters must be tuned to achieve the required quality of service under the constraint of the uncontrolled parameters. For example, if the redundancy is high, the perennial should be better as a lot of fragments of the data will be in the system, but at the expense of using more memory space and more communications to maintain the redundancy level. Depending on the required service level (less space usage, less bandwidth usage, ...), one must optimize these parameters. As the uncontrolled parameters may evolve dynamically, the tuning of the controllable parameters also must be done dynamically.

3.1.8 Simulation of very large distributed systems

Computer simulations of large scale distributed applications is still an open issue. Because of the very large scale of P2P systems, the simulation complexity of the system reaches unprecedented levels. In

order to cope with such a high complexity, we propose to combine several approaches of the following ones, which to our knowledge has never been achieved yet.

First we will explore several approaches for reducing the complexity:

- combine discrete event models with fluid models. This kind of approach has been successfully used in LAAS by J.-M. Garcia and his team for network protocol simulations and lead to the creation of the QosDesign start-up company.
- explore several ways of using aggregation techniques. For example group peers in clusters or split peer clusters according to various situations or conditions: occurrence of a rare event or according to some criteria reflecting a similar state (eg. peers with $0, 1, 2, \ldots, k$ blocks of data pending for repair), etc.
- combine several formalisms and techniques: discrete-events, fluid models, Markovian models, etc. A few simulators, such as Ptolemy, developed in Berkeley, already provide means of exploring such an approach. However, it is worth stressing that the key issue here is not (only) to build a multi-paradigms simulator, but rather (also) to decide in which relevant situations it is worth using one instead of another.

Then (concurrently), we will explore the various parallel and distributed execution approaches. Several well-known parallelisation techniques proposed for simulation (especially the ones described by Fujimoto in [20]) should be considered. However, given the complexity of the biggest P2P systems we want to study, we will also try to apply these techniques on extremely massively parallel configurations. In particular, we plan to experiment grid computing solutions at a very large scale, on the french Grid5000 network first, and then, if possible on larger (meta-)clusters. We are already working on establishing international collaborations in the simulation and grid communities in order to connect multiple grid computing facilities (currently these discussions involve teams in France, Canada, USA and south America).

Last, but not least, all these experiments around the simulation methodologies and techniques will require a lot of software developments as well as a very flexible and versatile simulation architecture. Regarding the simulation architecture, we already started working on a new OpenSource Simulation Platform called OSA (Open Simulation Architecture) [13, 14]. This new simulation software is also carefully designed to allow a collaborative model of development, fully inspired of the successful Eclipse model[15].

3.2 WP1: A common base for work (7 months)

The goal of the first phase is to merge the knowledge of the partners who come from distinct communities. In large-scale distributed storage systems, both security and performance are key issues which should be studied jointly as they influence each another. This phase will be the starting point of a collaboration between research teams on these topics. The output of this WP will be common models which will serve as a basis for studies in next WPs.

3.2.1 System Description (60 days), Partners: All

First, UbiStorage will describe its P2P storage system to partners based on its experience with it. A description of the architecture of the system and software will be produced (30d). During the analysis described below, UbiStorage will also describe all its known bottlenecks and planned improvements (30d). During this phase, partners will give feedback to UbiStorage on the understanding of their description.

During this phase, Eurecom will cooperate with UbiStorage and possibly other partners of the project to fully grasp the architecture of the system and security issues regarding storage with the goal of preparing a solid grounding for the abstraction effort carried out during the analysis below. The tools that will

be used to model the system and its dynamics should allow studying more general architectures than the original system, hence the need for abstract and general models of cooperation and performance.

Resources needed

- UbiStorage: S. Choplin (50%), S. Drapeau (50%), G. Utard (15%), **Engineer R&D #1 (50%)**
- LACL: G. Hains (10%), C. Dima (10%), F. Pommereau (10%)
- REGAL: P. Sens (10%), L. Arantes (10%)
- MASCOTTE: PhD-M (100%)
- EURECOM: R. Molva (8%), P. Michiardi (10%), Y. Roudier (20%), **Postdoctoral researcher - engineer (100%)**

3.2.2 Analysis and First Experimentations (180 days), Partners: All

Each partner will then describe a state of the art in its area of knowledge about P2P storage issues that will compare the UbiStorage solution to other existing solutions or research projects (180d).

This analysis will constitute a common base of understanding so that all partners will describe the issues they tackle during the project using a shared terminology and model.

For their own experimentations, each partner will set up a network of ten dedicated computers to experiment with existing solutions. These computers will also be connected together to run larger experimentations at later stages of the project. We will later on also use the tools developed in the second phase.

We will also experiment existing solutions on a large scale (including the UbiStorage one). The experimentations will be done on several complementary platforms (the project's network of dedicated nodes, Grid 5000, PlanetLab, GridExplorer). The REGAL team has already experimented their Pastis system on Grid 5000 and EURECOM is already working with PlanetLab, so these partners will be of great value for this task. This task will begin as soon as possible, UbiStorage PhD #1 and Engineer #1 will be in charge of these experimentations as a practical base of knowledge of existing solutions. PhD #1 will meet both partners spending 1 month in Sophia-Antipolis and 1 month in Paris to make experimentations fit as much as possible with the problems studied. An important part of this work will be to established benchmarks in order to be able to compare the different approaches and solutions. After each prototype phase of development, the prototypes will also be experimented on such platforms. The final prototype will be challenging the platform with constraints which should reflect the Internet situation as much as possible.

On the specification and verification side, this period will also serve to identify the appropriate formalisms - logics, abstractions, system models - needed for the specification of the functionalities of the communication protocols.

Resources needed (Human)

- UbiStorage: S. Choplin (30%) (supervision of experimentations, WP1 report), G. Carpentier (20%) (setting up and adminstration of the experimentation platform), CIFRE PhD #1 (100%) (including 50% on experimentations), **Engineer R&D #1 (100%)** (experimentations),
- LACL: C. Dima (30%) (choice of formalisms - logics, abstractions, system models),G. Hains (40%) (modelisation, algorithm design for verification), **PhD #2 (100%)** (same research topic as C. Dima)
- EURECOM: R. Molva (8%), P. Michiardi (10%), Y. Roudier (20%), **Postdoctoral researcher - engineer (100%)**, **PhD-E (100%)**

- MASCOTTE: O. Dalle (60%) (simulation modeling, study of protocols and algorithms, scenarios specifications, performance metrics definition), P. Mussi (30%) (simulation modeling, performance metrics definition), S. Pérennes (40%) (analytical modeling, performance metrics definition), M. Syska (40%) (study of protocols and algorithms, analytical modeling, performance metrics definition), **Ing-M (20%)** (engineering support, test case specification for validation of future simulations), **PhD-M (100%)** (study of protocols and algorithms, simulation and analytical modeling),
- REGAL: P. Sens (30%) (supervision of experimentations on Grid'5000 and Grid eXplorer), L. Arantes (30%) (study of communication protocols and algorithms), **Internship #1 (100%)** (experimentations on Grid'5000)

Resources needed (Material)

- 50 computers (10 for each partner) for experimentations, future tests and simulations, 5 x 15K€

Resources needed (Travels)

- UbiStorage: CIFRE PhD #1: 1 month in Sophia-Antipolis (MASCOTTE and EURECOM), 1 month in Paris (LACL and REGAL), G. Carpentier: 2 days on each partner site to set up the experimentation platform, S. Choplin: 2x2 days in Sophia-Antipolis (MASCOTTE and EURECOM), 2x2 days in Paris (LACL and REGAL)
- LACL: PhD #2 (LACL): 2 weeks stay with UbiStorage for getting acquainted with protocol requirements, 3 persons x 1 travel (2 days) to Amiens for all LACL members.
- REGAL: P. Sens: 1 travel (2 days) to Amiens, Internship #1: 1 travel (2 days) to Amiens
- MASCOTTE: O. Dalle, S. Pérennes, M. Syska, P. Mussi, PhD-M: 1 travel (2 days) to Amiens, O. Dalle, M. Syska, S. Pérennes: 2 travels (2 Days) to Paris (REGAL/LACL), PhD-M: 1 week in Amiens PhD-M: 1 week in Paris (REGAL/LACL)
- EURECOM: Y. Roudier, P. Michiardi, PhD-E, Postdoctoral researcher/engineer: project meetings, conference(s) registration and travel expenses.

3.2.3 Meeting (3 days), Partners: All

Then partners will meet to present their results, define the main direction of the future research, and specify ways and tools to prove, experiment and simulate the future proposed solutions. Problem to study will be scheduled in WP2 and WP3 such that WP2 will end with most solutions as possible, then WP3 will contain some real experimentations of these solutions.

Resources needed

- Ubi: 4 persons x 3 days + travels: 2K€
- LACL: 4 persons x 3 days + travels: 2K€
- REGAL: 4 persons x 3 days + travels: 2K€
- MASCOTTE: 6 persons x 3 days + travels: 4K€
- EURECOM: 4 persons x 3 days + travels: 2K€

3.3 WP2: Find solutions for the bottlenecks (12 months)

The goal of this WP is to find solutions on some of the identified bottlenecks which should be ready to experiment in the WP3.

3.3.1 Research

Based on the models described in phase one, research will be done to handle open problems:

Specification and verification, Partners: LACL, UbiStorage

For the specification and verification part, during this period the functionalities of the communication protocol will be specified in the chosen logic, for the single server version of the system.

We will also work on parallelizing model checkers, and experiencing with "toy examples" that mimics some features of the protocol functionalities. This phase involves extensive work with parallelized model checkers.

On the other hand, we will identify the formal framework for the generation of attack scenarios. This implies the choice of a formalism for specifying the security properties, as well as a formalism for specifying attacker capabilities.

Price of anarchy, Partners: EURECOM, MASCOTTE, UbiStorage

During the second phase of the project, Eurecom will focus (among the subjects addressed by the team) on game theoretic analysis of the system model defined during the first phase. Formally, we will focus on the definition of distributed algorithms that implement Equilibrium solutions (exact or approximate) to the game theoretic model of the system. These algorithms will be compared to both greedy local algorithms (that use local knowledge only) and to "socially" optimal algorithms where a complete knowledge of the system is assumed. The ultimate goal is to come up with a measure of the loss of efficiency (called the price of anarchy) when system components act selfishly. We also will tackle the problem of augmenting our model to take into account system dynamics (such as peer churn).

Study of relevant techniques for the simulation of Large P2P Systems, Partners: MASCOTTE, LACL

Based on the specification of the system provided in WP1, we will start our studies and experimentations on simulation techniques for P2P networks. We will explore both directions of reducing the complexity of the simulation models and parallelizing the execution of the simulation engine in order to support the execution simulation models that have higher complexity.

Modeling and analysis of relevant performance metrics for P2P storage stystems, Partners: MASCOTTE, EURECOM, UbiStorage

After the identification of a set of relevant performance metrics conducted in WP1, we will start the performance evaluation of the system and study its sensitiveness to the variations of some key functional parameters of the system. These studies will combine both the results of analytical modeling and some preliminary results of simulations. Analytical models will be based on classical stochastic approaches and Queueing Theory classical models.

We will also study how these models might be used as a way to assess cooperation. This should in particular make it possible for us to enhance our game-theoretical model of a cooperation based secure storage protocol and to study its convergence.

Churn resistance, Partners: REGAL, MASCOTTE

The frequency of connections/disconnections (churn) is a real problem in P2P systems since the integrity of the P2P infrastructures is totally dependent on this not trivial parameter. Based on our experience on Pastis file system, we will propose churn resistant protocols where data replicas are stored in stable nodes.

Algorithmic primitives and data protection

We will study the design of cryptographic primitives for data verification and data management. In particular, schemes based on homomorphic functions appear to be especially promising to both areas. We hope to integrate such functions as checks in our probabilistic data storage verification protocols. We also plan to use such functions to design primitives for manipulating encrypted data while preserving their properties (confidentiality, integrity, ...). Other algorithmic approaches will also be investigated for both types of problems.

3.3.2 Tools development

Tools will be developed so that each partner will be able to test/simulate strategy in its context.

Formal specifications (30d), Partners: UbiStorage

Based on common work during the meeting of the previous WP, formal specifications will be defined and proposed to all partners. After validation by the partners, the tools will be developed.

Simulator, Partners: MASCOTTE, UbiStorage, LACL

The simulator will be mainly developed by MASCOTTE. It will be based on the OSA simulation platform and will be available as OpenSource software (under LGPL). Our main concern in WP2 will be to implement realistic, discrete-event simulation models of the system specifiified in WP1. These realistic models will serve as a basis for:

- preliminary performance studies of small size P2P configurations
- experimentations of new simulation techniques for the reduction of the complexity of large scale P2P systems models: the reduction techniques will be experimented on variants of the basic models (eg. replacement of a collection of realistic models of peers by an aggregated model, adjunction of fluid models of background workload to realistic discrete event models, etc).

Experimentation tools, Partners: UbiStorage, EURECOM

UbiStorage will distribute to partners its system with access to the API of the software to allow partners to implement new strategies for their experimentations. Engineer #1 (UbiStorage) will spend one week in Sophia-Antipolis and 1 week in Paris to teach to partners the API usage.

Experimental tools defined in first phase which should be of interest for experimentations will be mainly developed by UbiStorage in collaboration with other partners.

Resources needed (Human)

- UbiStorage: S. Choplin (50%) (modeling and perf. analysis, experimentations), S. Drapeau (40%) (software API maintenance and adaptation), G. Carpentier (20%) (tools development and experimentation platform administration), **Engineer #1 (100%)** (60% API formation and experimentation support, 40% OSA devs), CIFRE PhD #1 (100%) (including 25% of experimentation)

- LACL: G. Hains (30%) (formalisms for security properties and model checker parallelization, sim techniques), C. Dima (40%) (communication protocol specification and formalism for security properties), F. Pommereau (40%) (communication protocol specification and model checker parallelization, sim techniques and sim dev performances), **PhD #2 (100%)** (communication protocol specification).
- REGAL: P. Sens (30%) (specification of a churn injection tools), L. Arantes (30%) (specification of churn resistant protocols)
- MASCOTTE: O. Dalle (50%) (10% OSA devs, 30% sim techniques, 10% analytical modeling), S. Perennes (50%) (10% price of anarchy, 10% churn resistance 10%, 30% analytical modeling), P. Mussi (20%) (sim techniques), M. Syska (40%) (analytical modeling), **Ing-M (100%)** (OSA devs), **Phd-M (100%)** (33% OSA devs, 33% sim techniques, 34% analytical modeling)
- EURECOM: R. Molva (14%), P. Michiardi (10%), Y. Roudier (25%), **Postdoctoral researcher - engineer (100%)**, **PhD-E (100%)**

Resources needed (Travel)

- UbiStorage: Engineer #1: 1 week in Sophia-Antipolis, 1 week in Paris (API Ubi, formation)
- LACL: PhD #2: 2 weeks in Amiens for cross-checking formal and informal specifications
- EURECOM: Y. Roudier, P. Michiardi, PhD-E, Postdoctoral researcher/engineer: project meetings, conference(s) registration and travel expenses.
- MASCOTTE: participation to relevant workshop and conferences in France (simulation and P2P domains), meeting travels, conference registration and travel

3.3.3 Meeting (3 days), Partners: All

Then partners will meet to

- present their research results that will be experimented in WP3,
- define the main direction of the future research,
- present the developed tools that will be used in WP3.

Resources needed

- Ubi: 4 persons x 3 days + travels: 2K€
- LACL: 4 persons x 3 days + travels: 2K€
- REGAL: 4 persons x 3 days + travels: 2K€
- MASCOTTE: 6 persons x 3 days + travels (meeting travels): 4K€
- EURECOM: 4 persons x 3 days + travels: 2K€

3.4 WP3: Implementation, Simulations, Validation of solutions, Find solutions on remaining problems (12 months)

The first goal of this WP is to implement solutions of WP2 to run experimentation on large scale to validate these solutions. The second one is to handle remained bottlenecks and experiment parts of new proposed solutions.

3.4.1 Research

Both research thematics of WP2 will be continued, additionally the following points will be studied:

Specification of protocols, Partners: LACL, EURECOM, UbiStorage

During this period, the (abstraction of) system functionalities will be model-checked against the formal specification of the functionalities of the communication protocol. Eventual bugs will be checked against the real system, for either refining the system abstraction and/or improving the specifications, or for correcting into code.

A further sub-working package includes the specification of protocols achieving server distribution.

We will also formally specify the security properties and the attacker capabilities together with the teams that will design and implement the security protocols

Analysis of performance bottlenecks and scalability issues, Partners: MASCOTTE, EURECOM, UbiStorage

Reusing the outcome of the research and development on modeling and simulation conducted in WP2, studies of larger system models based on simulation will be able to start. These studies will use the advanced simulation modeling techniques developed in WP2 in order to lower the simulation complexity. these techniques will be applied to the models developed in WP2. New models, corresponding to the improvements and modifications suggested by the partners in WP2 will need to be developed. The output of these simulations will provide the necessary feed-back required by all the partners in order to evaluate the performance of their solution (according to the performance metrics described in WP1). The simulation results will be confronted to the results of theoretical analysis as well empirical results obtained by the partners during the experimentations.

Based upon the modeling efforts as well as the experiments carried out in WP1 and WP2, we will extend the game theoretic investigations for both optimal placement strategies and data verification protocols. Specifically for the former, we will tackle the problem of groups or coalitions of peers/users exhibiting different behaviors when the system evolves to understand which strategy will dominate. Regarding the latter, we plan to study how to optimize the trade-offs that arise between security and performance of data storage or management.

Resources needed

- UbiStorage: CIFRE PhD #1 (50%), S. Choplin (15%)
- LACL: G. Hains (50%), C. Dima (40%), F. Pommereau (20%), **PhD #2 (100%)**
- MASCOTE: **PhD-M (50%)**, O. Dalle (25%), P. Mussi (10%), M. Syska (25%), S. Pérennes (50%)
- EURECOM: R. Movla (4%), P. Michiardi (8%), Y. Roudier (9%), **PhD-E** (80%)

3.4.2 Development and Experimentations, Partners: UbiStorage, REGAL, MASCOTTE

Using research results of the previous WP, prototypes based on the UbiStorage solution will be developed to implements the proposed solutions. Then, experimentations will serve as validation and feedback for these results.

We will also implement a churn injection tool and churn resistant protocols specified in WP2.

The developments in simulation will consist in:

- improving the performance of the simulator using both code optimization techniques and parallelization techniques. In particular we will work on the development of a massively parallel version of the simulator and will evaluate its performances compared to the sequential version on High Performance Grid facilities.

- develop new simulation models and update existing ones, based on the results and improvements proposed by the partners

Resources needed

- UbiStorage: **R&D Engineer** (100%), S. Drapeau (40%), G. Carpentier (20%), S. Choplin (35%), CIFRE PhD #1 (50%)
- MASCOTTE: **PhD-M (50%)** (15% of experimentation, 35% of simulator dev.), O. Dalle (25%), M. Syska (5%), P. Mussi (10%), **Engineer-M (40%)** (simulator dev.),
- REGAL: L. Arantes (30%), P. Sens (30%), **Internship #1 (100%)**, **Postdoc (100%)**
- EURECOM: P. Michiardi (2%), Y. Roudier (9%), **PhD-E (80%)**

3.4.3 Meeting (3 days), Partners: All

Then partners will meet to

- present their research and experiment results,
- define what will be challenged in WP4,
- describe what will be in the popularization documentation,
- define the end-user tool using the simulator and model-checking tools.

Resources needed

- Ubi: 4 persons x 3 days + travels: 2K€
- LACL: 4 persons x 3 days + travels: 2K€
- REGAL: 4 persons x 3 days + travels: 2K€
- MASCOTTE: 6 persons x 3 days + travels (meeting travels): 4K€
- EURECOM: 4 persons x 3 days + travels: 2K€

3.5 WP4: Final experimentations and Results Dissemination (5 months)

The goal of the last WP is to present the results and improvements brought by this project.

3.5.1 Scientific result analysis and publication

Main research results will be published to the scientific community.

3.5.2 Verification, Partners: LACL

During this part of the project, the distributed server version will be model-checked against the general specifications. The abstraction methods developed in WP2 and WP3 will be applied to provide state-space reduction. We will also utilize the best parallelization strategy for model checking of the reduced model.

3.5.3 Test generation for security properties, Partners: LACL, EURECOM

The security specifications and attacker model developed during WP3 will be used for generating tests, that will be implemented and experimented on the system. For instance, malicious behaviors will be assigned to a given fraction of the peers in order to test the resilience of the storage function. Denial of service attacks will also be evaluated on the real infrastructure.

3.5.4 Challenging experimentation, Partners: REGAL, UbiStorage

A large challenging experimentation will be done to show the efficiency of results obtained by this project.

3.5.5 Popularizing documents, Partners: UbiStorage

Documents will be producted to popularize the major contributions of this project to the mass public and to increase the trust of potential users towards distributed P2P storage.

3.5.6 Development of end-user simulation/model checking tool, Partners: MASCOTTE, LACL, UbiStorage

A tool merging the simulator and formal model checking will be developed in which user will be able to run by himself a set of test with his entry parameters.

3.5.7 Final public meeting

A final meeting will be done to present the scientific results and experimentations.

Resources needed

- Prints and conception of popularizing documents 3K€
- Communication of results to mass public 5K€

Resources needed - Human

- LACL: G. Hains (40%) (parallel model checking and test generation, end-user simulator), C. Dima (25%) (test generation), F. Pommereau (20%) (parallel model checking), **PhD #2 (100%)** (parallel model checking)
- MASCOTTE: O. Dalle (25%) (results analysis and publications, OSA software diffusion, end-user simulator), M. Syska (25%) (results analysis and publications), S. Perennes (30%) (results analysis and publications), **PhD-M (100%)** (results analysis and publications)
- EURECOM: Y. Roudier (20%) (test generation, result analysis and publications), P. Michiardi (10%) (results analysis and publications), R. Molva (8%) (results analysis and publications), **PhD-E (100%)** (test generation, results analysis and publications)
- UbiStorage: S. Choplin (40%) (popularizing documents,Final report,challenging experimentation), **R&D Engineer (100%)** (challenging experimentation, end-user simulator), G. Utard (15%) (popularizing documents), G. Carpentier (20%) (challenging experimentation), CIFRE PhD #1 (100%) (results analysis and publications)
- REGAL: P. Sens (30%) (results analysis and publications, challenging experimentation), L. Arantes (results analysis and publications)

Resources needed: travels

- LACL: 4 participants x 2 days travel for the final meeting.
- UbiStorage: 4 participants x 2 days travel for the final meeting.
- REGAL: 4 participants x 2 days travel for the final meeting.
- EURECOM: 4 participants x 2 days travel for the final meeting.
- MASCOTTE: 6 participants x 2 days travel for the final meeting.

3.6 Project Management

3.6.1 Management Structure

This project will be coordinated by UbiStorage (S. Choplin).

Two main aspects of the project can be defined and distributed as follows between partners:

- "Research and Experimentation" which will be mainly handled by LACL, MASCOTTE, REGAL and EURECOM,
- "Engineer and Verification" which will be mainly handled by UbiStorage and MASCOTTE.

S. Choplin will spend two days every six months with both partners to discuss and work on main project progress and objectives.

A public web site presenting the project and its progresses will be maintained by UbiStorage. A restricted to partners collaborative web site will also be provided by UbiStorage to serve as a common space of work.

3.6.2 Measures taken relative to the major identified risks

The major risk is in the scheduling of this project between WP2 and WP3, if not enough research results are provided at the end of WP2, then development for experimentations would be delayed in WP3. To reduce this risk, a intermediate milestone in WP2 will consist in evaluating the research advances and reorient research to simpler problem if needed.

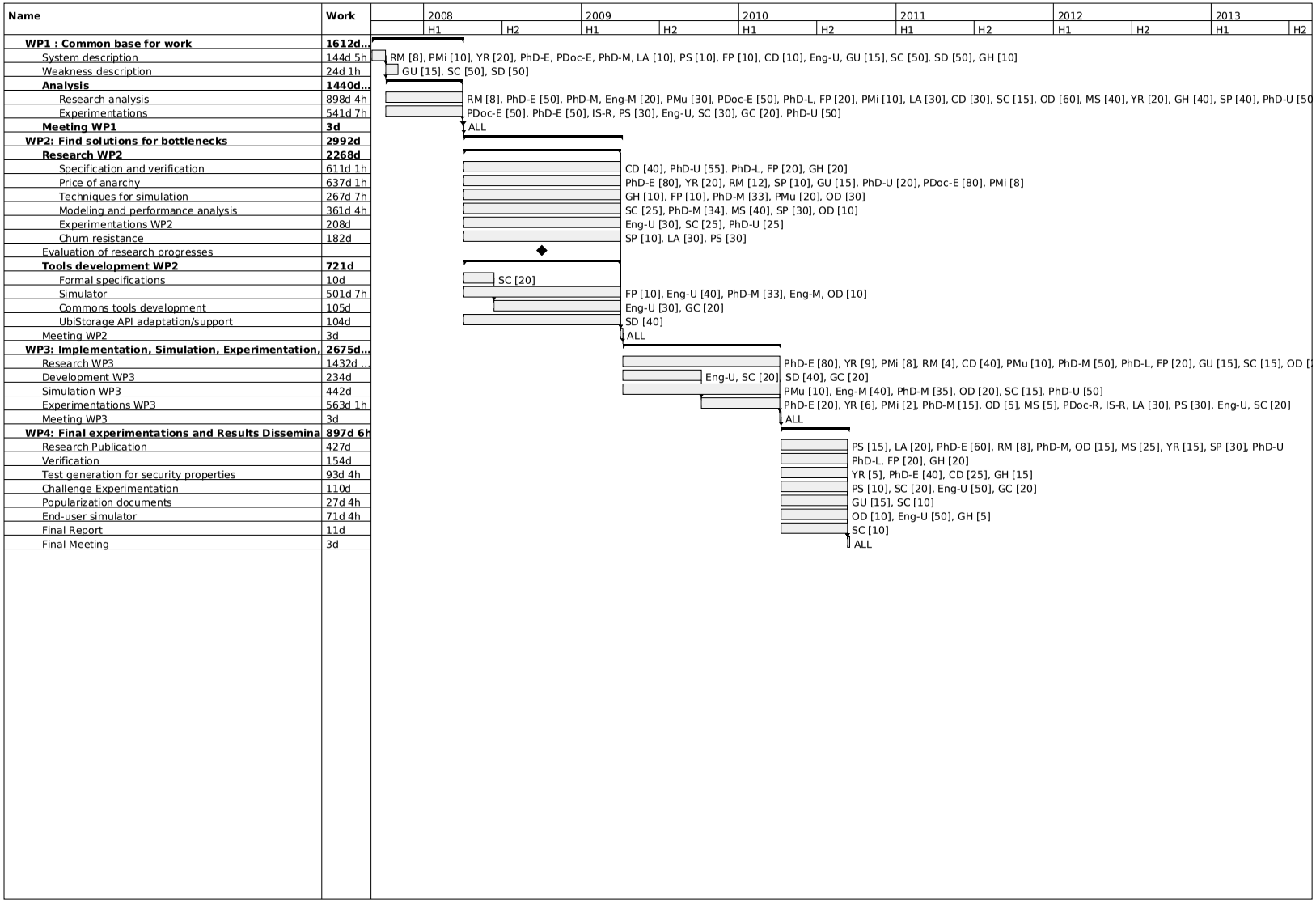
Even if this problem occurs, development resources will be assigned to the simulator developments and when enough results from research will come, the original development resources assigned to the simulator developments will be assigned to development for experimentations.

3.6.3 Main milestones

The main milestones will be at the end of each WP, as each WP is mainly composed of parallel sub-phases. One extra milestone will be in the middle of WP2 to ensure that, at the end of WP2, enough results could be exploited as experimentation in WP3.

3.6.4 Planning

Table 3.1 describe the Gantt chart of the project and table 3.2 its associated human resources.



Name	Work	2008 H1	2008 H2	2009 H1	2009 H2	2010 H1	2010 H2	2011 H1	2011 H2	2012 H1	2012 H2	2013 H1	2013 H2
WP1 : Common base for work	**1612d...**												
System description	144d 5h	RM [8], PMi [10], YR [20], PhD-E, PDoc-E, PhD-M, LA [10], PS [10], FP [10], CD [10], Eng-U, GU [15], SC [50], SD [50], GH [10]											
Weakness description	24d 1h	GU [15], SC [50], SD [50]											
Analysis	**1440d...**												
Research analysis	898d 4h	RM [8], PhD-E [50], PhD-M, Eng-M [20], PMu [30], PDoc-E [50], PhD-L, FP [20], PMi [10], LA [30], CD [30], SC [15], OD [60], MS [40], YR [20], GH [40], SP [40], PhD-U [50											
Experimentations	541d 7h	PDoc-E [50], PhD-E [50], IS-R, PS [30], Eng-U, SC [30], GC [20], PhD-U [50]											
Meeting WP1	**3d**	ALL											
WP2: Find solutions for bottlenecks	**2992d**												
Research WP2	**2268d**												
Specification and verification	611d 1h	CD [40], PhD-U [55], PhD-L, FP [20], GH [20]											
Price of anarchy	637d 1h	PhD-E [80], YR [20], RM [12], SP [10], GU [15], PhD-U [20], PDoc-E [80], PMi [8]											
Techniques for simulation	267d 7h	GH [10], FP [10], PhD-M [33], PMu [20], OD [30]											
Modeling and performance analysis	361d 4h	SC [25], PhD-M [34], MS [40], SP [30], OD [10]											
Experimentations WP2	208d	Eng-U [30], SC [25], PhD-U [25]											
Churn resistance	182d	SP [10], LA [30], PS [30]											
Evaluation of research progresses													
Tools development WP2	**721d**												
Formal specifications	10d	SC [20]											
Simulator	501d 7h	FP [10], Eng-U [40], PhD-M [33], Eng-M, OD [10]											
Commons tools development	105d	Eng-U [30], GC [20]											
UbiStorage API adaptation/support	104d	SD [40]											
Meeting WP2	3d	ALL											
WP3: Implementation, Simulation, Experimentation,	**2675d...**												
Research WP3	1432d ...	PhD-E [80], YR [9], PMi [8], RM [4], CD [40], PMu [10], PhD-M [50], PhD-L, FP [20], GU [15], SC [15], OD [											
Development WP3	234d	Eng-U, SC [20], SD [40], GC [20]											
Simulation WP3	442d	PMu [10], Eng-M [40], PhD-M [35], OD [20], SC [15], PhD-U [50]											
Experimentations WP3	563d 1h	PhD-E [20], YR [6], PMi [2], PhD-M [15], OD [5], MS [5], PDoc-R, IS-R, LA [30], PS [30], Eng-U, SC [20]											
Meeting WP3	3d	ALL											
WP4: Final experimentations and Results Dissemina	**897d 6h**												
Research Publication	427d	PS [15], LA [20], PhD-E [60], RM [8], PhD-M, OD [15], MS [25], YR [15], SP [30], PhD-U											
Verification	154d	PhD-L, FP [20], GH [20]											
Test generation for security properties	93d 4h	YR [5], PhD-E [40], CD [25], GH [15]											
Challenge Experimentation	110d	PS [10], SC [20], Eng-U [50], GC [20]											
Popularization documents	27d 4h	GU [15], SC [10]											
End-user simulator	71d 4h	OD [10], Eng-U [50], GH [5]											
Final Report	11d	SC [10]											
Final Meeting	3d	ALL											

Table 3.1: Gantt chart

Short name	Name	Group
CD	Catalin Dima	LACL
Eng-U	Engineer #1 - U	UbiStorage
Eng-M	Engineer #2 - M	MASCOTTE
FP	Franck Pommereau	LACL
GH	Gaétan Hains	LACL
GU	Gil Utard	UbiStorage
GC	Guillaume Carpentier	UbiStorage
IS-R	Internship R	REGAL
LA	L. Arantes	REGAL
MS	Michel Syska	MASCOTTE
OD	Olivier Dalle	MASCOTTE
PhD-E	PhD-E Eurecom	EURECOM
PhD-L	PhD #2 LACL	LACL
PhD-M	PhD-M Mascotte	MASCOTTE
PhD-U	CIFRE PhD #1 UbiStorage/LACL	UbiStorage
PMu	Philippe Mussi	MASCOTTE
PS	Pierre Sens	REGAL
PMi	Pietro Michiardi	EURECOM
PDoc-E	PostDoc Eurecom	EURECOM
PDoc-R	Postdoc REGAL	REGAL
RM	Refik Molva	EURECOM
SC	Sébastien Choplin	UbiStorage
SD	Stéphane Drapeau	UbiStorage
SP	Stéphane Pérennes	MASCOTTE
YR	Yves Roudier	EURECOM

Table 3.2: Humane resources

3.6.5 PhD Students

CIFRE PhD #1 (UbiStorage)

Subject: design and experimentation of scalable secure P2P storage system.

This PhD will be employed by UbiStorage, supervised by S. Choplin (UbiStorage) and G. Hains (LACL Laboratory). Under a CIFRE convention, the salary of this PhD will not be funded by the ANR.

He (She) will be involved in both theoretical aspects and experimentation to describe a P2P storage system with good properties (according to mentioned constraints) and analyze in real environment the outcomes of this project.

As the main topic of this PhD will be on security, he will work on this project with mainly the LACL laboratory (to which he will be attached) and the Eurecom NS Team.

PhD #2 (LACL)

The PhD Student at Université Paris-12 will be supervised by LACL participants. He will work mainly at LACL, participate in all project correspondence and meetings, and spend a dedicated work-term at UbiStorage to maximize interaction with the CIFRE PhD student working there. The student will specialize in formal and automated verification of performance and safety-security properties. All the models and case studies he develops will be centered on the UbiStorage architecture. The thesis will result in lists of possible security / performance leaks, as well as automated tools/libraries for the verification of global behaviour. Symbolic and parallel processing will be core techniques in attaining these results.

PhD #3 (MASCOTTE)

The PhD. Student at university of Nice Sophia antipolis will be supervised by the Mascotte participants (main advisors: O. Dalle and S. Pérennes, official advisor: JC Bermond). This PhD. will address two kinds of closely related issues: (1) explore new analytical modeling techniques and simulation techniques that enable the study of very large P2P systems and (2) exploit the previous analytical and simulation techniques in order to optimize the performances of the system (or, more generally, its Quality of Service levels). This PhD. thesis will result in (1) a set of analytical and computer simulation models (mostly distributed in Open Source) and (2) a list of performance analysis results and recommendations about the system architecture and underlying protocols. These results, recommendations and improvements will be published in conferences and journals of the domain.

PhD #4 (EURECOM)

The Ph.D. Student at Eurecom Institute will be supervised by the NSTEAM participants (co-advisors: R. Molva and Y. Roudier). The objective of this Ph.D. is to explore algorithmic protection primitives making it possible to extract new types of signatures, hashes, parity checks, etc. out of encrypted or blinded data without breaking their security properties. Homomorphic functions are especially envisioned as building blocks for such primitives. This Ph.D. will also explore how these primitives can be applied to several problems in the data storage area: encrypted data lookup and routing, data reencryption, cooperative data durability enforcement, etc. These results will be published in conferences and journals of the domain.

Furthermore, we envision that the Ph.D. might lead to the development of open source tools (possibly based on PlanetLab) for experimenting and validating the cooperative schemes designed and how algorithmic protection primitives impact the data storage application performance.

Chapter 4

Expected results and perspectives

The results of this project are both scientific for the research community and strategic for the UbiStorage company.

4.1 Scientific outcomes

- This project joins 5 teams with complementary skills, this mix will provide a better understanding of a hot topic involving several scientific domains (security, distributed algorithms, complexity and optimisation, simulation, network design). Publication in international conferences will be published on the scientific understanding of the problems.
- All the models, simulations softwares and scenarios developed within this project, except the ones that are critical for the competitiveness or security of the UbiStorage system, will be distributed as Open Source software, under a free Copyright licence such as GPL or LGPL.

4.2 Evaluation criteria for measuring its success

The following criteria will serve as measuring the success of this project:

- the number of scientific publications
- the number of scientific publications in communities which are not the original community of the authors (to show the success of the association of the complementary teams)
- the ratio (number of experimented solutions)/(number of theoretical proposed solutions)
- the ratio (number of integrated solution in the UbiStorage system)/(number of experimented solutions)
- size of the biggest simulations
- size of the biggest experimentations

The public website will provide daily up-to-date evaluations of these criteria.

4.3 Useability and ergonomic aspects

A particular attention will be put on the reusability of the simulation software and models developed within the project. The new simulation models developed by INRIA within the SPREADS project will be implemented in the OSA platform. Indeed, INRIA has recently started the development of a new component-based simulation platform, called OSA. The components of OSA, in turn, are built on top

of the ObjectWeb's Fractal component model, developed by INRIA and France Telecom. Thanks to the Fractal component model, the OSA platform and models will exhibit unique features compared to other exiting simulation platform: multi-programming language support, distributed execution, strong separation of the simulation concerns, etc. The OSA platform also relies on a front-end user interface which is totally integrated in the popular Eclipse Integrated Development Environment. This integration in Eclipse also reflects the Open and Collaborative philosophy of the OSA project. OSA also relies on the Apache foundation Maven building system, which dramatically facilitates the building, maintenance and dependency management of the simulation applications thanks to a dedicated software repository.

To convince the end-user, a tool merging the simulator and formal model checking will be developed in which the user will be able to run by himself a set of test with his entry parameters. It will be publicly available so that any person can have answers to many of Frequently Asked Questions like "what is the number of failures that the system can support ?" (parameters should be input by the user).

4.4 Economic perspectives

Involved in this project, UbiStorage will keep in hot research on its domain: providing safe, distributed data storage services.

The first result of this work will be to permit UbiStorage to address the mass market and then propose a low cost, safe way to people to store their data.

With simulations and models, parameter tuning will be possible such that the guarantee of the perenniality of the data should be insured with very high probability. Also experimentations will be a proof of concept. These points will be important for UbiStorage for several reasons:

- customers have to be in confidence with distributed solutions, then we have to provide them with this kind of proof of concept,
- a partnership with insurance companies can be established (and then we have to estimate all risks of the solution), some of the concurrents using data center already have this strategy.

For long term, the UbiStorage aim is to become the technological leader in the growing market of online data storage. A grant of the ANR program will be a high opportunity to build a strong relationship between UbiStorage and academic partners and to stay at the top of the art.

4.5 Marketing

Using knowledge obtained during this project, UbiStorage will improve its solution after validation of each experimentation and adaptation of its software. Then the first exploitation of these results will start during WP3 and will be in charge of UbiStorage.

Chapter 5

Intellectual property / free or open source software

The partners of this project will contract on the Intellectual Property in the next 6 months on the following points:

- The company Ubiquitous-Storage SA has an exclusive agreement with the "Université de Picardie J.V." to develop its patent on a distributed storage system. This patent can be used for the research purpose of this ANR, if so a contract will be signed between partners and the "Université de Picardie J.V.".
- Patents and Software commonly developed by the partners will be owned by the partners. The modalities of the repartition and the support of the administrative costs to deposit and maintain the patents or licences will be notified in the contract.
- Partners will contract a Non Disclosing Agreement (NDA) for the usage of the UbiStorage software and all knowledge of the UbiStorage system. An agreement will be concludes to allow academic partners to publish their results in respect of the UbiStorage strategic constraints.
- Softwares developed for simulation and experimentations will be published as OpenSource software under LGPL Licence excepted the ones which will be based on the UbiStorage software.

In case of Patent or Software deposite, evaluation will be provided by "Cabinet Breese Derambure Majerowicz"[1] (5K€).

[1] http://www.breese.fr/

Bibliography

[1] A. Adya, W. Bolosky, M. Castro, R. Chaiken, G. Cermak, J. Douceur, J. Howell, J. Lorch, M. Theimer, and R. Wattenhofer. Farsite: Federated, available, and reliable storage for an incompletely trusted environment, 2002.

[2] Fred S. Annexstein, Kenneth A. Berman, and Mijhalo A Jovanovic. Broadcasting in unstructured peer-to-peer overlay networks. *Theoretical Computer Science*, 355, 2006.

[3] B. Berard, M. Bidoit, A. Finkel, F. Laroussinie, A. Petit, L. Petrucci, and P. Schnoebelen. *Systems and Software Verification, Model-Checking Techniques and Tools*. Springer Verlag, 2001.

[4] A. Bouajjani, P. Habermehl, and T. Vojnar. Abstract regular model checking. In *Proc. 16th Intern. Conf. on Computer Aided Verification (CAV'04)*, volume 3114 of *LNCS*, pages 372–386. Springer Verlag, 2004.

[5] G. Caronni and M. Waldvogel. Establishing Trust in Distributed Storage Providers. In *Third IEEE P2P Conference, Linkoping*, 2003.

[6] Castro and Liskov. Practical byzantine fault tolerance. In *OSDI: Symposium on Operating Systems Design and Implementation*. USENIX Association, Co-sponsored by IEEE TCOS and ACM SIGOPS, 1999.

[7] Yuan Chen, Jan Edler, Andrew Goldberg, Allan Gottlieb, Sumeet Sobti, and Peter Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

[8] Leonardo Chwif, Marcos Ribeiro Pereira Barretto, and Ray J. Paul. On simulation model complexity. In J.A. Joines, R.R. Barton, K. Kang, and P.A. Fishwick, editors, *Proceedings of the 2000 Winter Simulation Conference (WSC'00)*, pages 449–455, 2000.

[9] E.M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 2000.

[10] J. Cooley, J. Mineweaser, L. Servi, and E. Tsung. Software-based erasure codes for. scalable distributed storage. In *20th IEEE Mass Storage Systems & Technologies Symposium, MSST2003. Paradise Point Resort, San Diego, California, USA*, 2003.

[11] Patrick Cousot and Radhia Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Conference Record of the Fourth Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 238–252. ACM Press, 1977.

[12] Frank Dabek, M. Frans Kaashoek, David Karger, Robert Morris, and Ion Stoica. Wide-area cooperative storage with CFS. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01)*, Chateau Lake Louise, Banff, Canada, October 2001.

[13] Olivier Dalle. OSA: an Open Component-based Architecture for Discrete-event Simulation. Technical Report RR-5762, INRIA, February 2006. Available from http://www.inria.fr/rrrt/rr-5762.html.

[14] Olivier Dalle. Component-based discrete event simulation using the fractal component model. In *AI, Simulation and Planning in High Autonomy Systems (AIS)-C onceptual Modeling and Simulation (CMS) Joint Conference*, Buenos Aires, AR, Februray 2007.

[15] Jim des Rivières and John Wiegand. Eclipse: A platform for integrating development tools. *IBM Systems Journal*, 43(2):371–383, 2004.

[16] Y. Deswarte, J.-J. Quisquater, and A. Saïdane. Remote Integrity Checking. In *Proceedings of the Sixth Working Conference on Integrity and Internal Control in Information Systems (IICIS)*, 2004.

[17] Catalin Dima. Abstractions of multi-agent systems. in preparation.

[18] P. Druschel and A. Rowstron. PAST: A large-scale, persistent peer-to-peer storage utility. In *Procedings of HOTOS*, pages 75–80, 2001.

[19] D. G. Filho and P. S. L. M. Barreto. Demonstrating Data Possession and Uncheatable Data Transfer, published in IACR Cryptology ePrint Archive, 2006.

[20] Richard M. Fujimoto. *Parallel and distributed simulation systems*. Wiley Series on Parallel and Distributed Computing. J Wiley & Sons, 2000.

[21] G. Hains. Efficient static checking of dynamic access control in shared multiprocessor environments. In *Workshop on Collaboration and Security (COLSEC'07), part of The 2007 International Symposium on Collaborative Technologies and Systems (CTS'07)*, Orlando (FL), USA, May 2007.

[22] Igor Ivkovic. Improving gnutella protocol: Protocol analysis and research proposals. Technical report, University of Waterloo, 2001.

[23] J. Kleinberg. The small-world phenomenon: an algorithmic perspective. In *32nd ACM Symposium on the Theory of Computing*, 2000.

[24] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In *Proc of the 16th Annual Symposium on Theoretical Aspects of Computer Science, STACS*, Trier, Germany, March 2007.

[25] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, D. Eaton, P. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, C. Weimer, W. Wells, and Zhao B. Oceanstore: An architecture for global-scale persistent store. In *Proc. of ASPLOS'2000*, nov 2000.

[26] John Kubiatowicz, Divid Bindel, Yan Chen, Patrick Eaton, Dennis Geels, Ramakrishna Gummadi, Sean Rhea, Hakim Weatherspoon, Westly Weimer, Christopher Wells, and Ben Zhao. Oceanstore: An architecture for global-scale persistent storage. In *Proceedings of ACM ASPLOS*. ACM, November 2000.

[27] Nikolaos Laoutaris, Georgios Smaragdakis, Konstantinos Oikonomou, Ioannis Stavrakakis, and Azer Bestavros. Distributed placement of service facilities in large-scale networks. In *Proc. of the 26th Annual IEEE Conference on Computer Communications, INFOCOM*, Ankorage, Alaska, USA, May 2007.

[28] Nikolaos Laoutaris, Orestis Telelis, Vassilios Zissimopoulos, and Ioannis Stavrakakis. Distributed selfish replication. *IEEE Transactions on Parallel and Distributed Systems*, 17(12):1401–1413, 2006.

[29] D. Liben-Nowell, H. Balakrishnan, and D. Karger. Analysis of the evolution of peer-topeer systems. In *Proceedings of the Twenty-First Annual Symposium on Principles of Distributed Computing*, pages 233–242, 2002.

[30] M. Lillibridge, S. Elnikety, A. Birrell, M. Burrows, and M. Isard. A Cooperative Internet Backup Scheme. In *Proceedings of the 2003 Usenix Annual Technical Conference (General Track), San Antonio, Texas*, pages 29–41, jun 2003.

[31] A. Merlin and G. Hains. A generic cost model for concurrent and data-parallel meta-computing. In *Fourth Workshop on Automated Verification of Critical Systems: (AVOCS'04)*, September 2004. Long preliminary version appears as LIFO RR2004-06.

[32] A. Muthitacharoen, R. Morris, T. M. Gil, and Chen B. Ivy: A read/write peer-to-peer file system. In *Proceedings of 5th Symposium on Operating Systems Design and Implementation (OSDI 2002)*, 2002.

[33] N. Oualha, P. Michiardi, and Y. Roudier. A Game Theoretic Model of a Protocol for Data Possession Verification. In *to appear in Proceedings of the 2007 IEEE International Workshop on Trust, Security, and Privacy for Ubiquitous Computing (TSPUC 2007), Helsinki, Finland*, jun 2007.

[34] N. Oualha and Y. Roudier. Probabilistically Secure Cooperative Distributed Storage, Eurecom Research Report rr-07-188. Technical report, EURECOM, feb 2007.

[35] M. O. Rabin. Efficient dispersal of information for security, load balancing, and fault tolerance. *Journal of ACM*, 38:335–348, 1989.

[36] C. Randriamaro, O. Soyez, G. Utard, and F. Wlazinski. Data distribution for failure correlation management in a P2P storage system. In *Int. Symp. on Parallel and Distributed Computing (ISPDC05)*, 2005.

[37] C. Randriamaro, O. Soyez, G. Utard, and F. Wlazinsky. Data distribution in a peer to peer storage system. *Journal of Grid Computing (JoGC), Special issue on Global and Peer-to-Peer Computing*, 2006.

[38] Cyril Randriamaro, Olivier Soyez, Gil Utard, and Francis Wlazinski. Structured data mapping in a peer to peer storage system., 2006.

[39] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *In Proceedings of ACM SIGCOMM 2001*, 2001.

[40] Antony Rowstron and Peter Druschel. Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*, pages 188–201, New York, NY, USA, 2001. ACM Press.

[41] G. Hains S. Anantharaman, J. Chen. A synchronous process calculus for service costs. In *SEFM'05, The 3rd IEEE Int. Conference on Software Engineering and Formal Methods*, pages 435–444, 2005.

[42] Y. Saito, C. Karamanolis, M. Karlsson, and Mahalingam M. Taming aggressive replication in the pangaea wide-area file system. In *5th Symp. on Op. Sys. Design and Implementation (OSDI 2002)*, dec 2002.

[43] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. *IEEE ACM Transactions on Networking*, 11(1):17–31, 2003.

[44] Kaile Su. Model checking temporal logics of knowledge in distributed systems. In *Proceedings of 19th AAAI*, pages 98–103. AAAI Press / The MIT Press, 2004.

[45] G. Utard and A. Vernois. Data durability in peer-to-peer storage systems. In *4th Workshop on Global and Peer to Peer Computing*, Chicago, April 2004.

[46] W. van der Hoek and M. Wooldridge. Model checking knowledge and time. In *Proceedings of the Ninth International SPIN Workshop on Model Checking of Software*, volume 2318 of *LNCS*, pages 95–111. Springer Verlag, 2002.

[47] Guido Wimmel and Jan Jürjens. Specification-based test generation for security-critical systems using mutations. In *Proceedings of ICFEM 2002*, volume 2495 of *LNCS*, pages 471–482. Springer Verlag, 2002.

[48] Bernard P. Zeigler, Herbert Praehofer, and Tag Gon Kim. *Theory of Modeling and Simulation*. Academic Press, 2nd edition, 2000.

[49] Ben Y. Zhao, Ling Huang, Jeremy Stribling, Sean C. Rhea, Anthony D. Joseph, and John Kubiatowicz. Tapestry: A resilient global-scale overlay for service deployment. *IEEE Journal on Selected Areas in Communications*, 22(1):41–53, 2004.

Appendix A

Team details

A.1 Ubiquitous Storage SA

Sébastien Choplin, 30 years old (CTO)

Sébastien Choplin received his Ph.D. in Computer Science in 2002 from Univ. Nice Sophia-Antipolis. From 2003 to 2006 he was associate professor at the Univ. of Picardie J.Verne. He currently is in delegation as Research and Development Director in Ubiquitous Storage SA, the company he founded with G. Utard and C. Randriamaro. In 2005, he obtained the certification from the HEC-Challenge+ Institute[1] (help management program for innovative project). Research and Development interests: modeling of telecommunication networks, peer-to-peer storage, combinatorial optimization.

Gil Utard, 40 years old (CEO)

Gil Utard received his Ph.D. in Computer Science in 1995 from ENS Lyon. He was assistant professor at ENS Lyon and ENS Cachan in 1996 and 1997, then associate professor at Univ. of Picardie J.Verne in 1998. In 2002, he was during two years senior researcher (CR1) at INRIA Rhône Alpes. In 2005, he obtained the certification from the HEC-Challenge+ Institute[1] (help management program for innovative project). Research and Development interests: distributed algorithms, grid computing, memory consistency models.

Stéphane Drapeau, 32 years old (R&D Engineer)

Currently, Stéphane Drapeau is engineer on research and development at the Ubiquitous Storage enterprise. Since 2004 he works on the middleware architecture of the UbiStorage solution. He is architect of software for peer to peer storage systems. He made a PhD thesis (CIFRE funding) at the Polytechnic School of Grenoble. His research was made at two laboratories: the Architecture of Distributed Systems laboratory of France Télécom and the Logiciels, Systèmes et Réseaux laboratory of the IMAG Institute. His PhD research focused on the definition and development of an adaptable framework for replication services. From October 2004 to April 2006, Stéphane Drapeau participated in the Us project where he proposed an open architecture based on middleware services. Stéphane Drapeau has work on several research projects: JORM (Java Object Repository Mapping), PING (Platform for Interactive Networked Games), Continuum (system for virtual worlds). Research and Development interests: middleware, services, separation of concerns, component, adaptability, adaptation, distributed systems aspects: replication, consistency, coherency, fault tolerance, naming, persistency, mobility.

[1] http://www.hec.fr/hec/eng/instituts/startup.html

Guillaume Carpentier, 26 years old (Engineer)

Guillaume Carpentier received is Master Degree in Computer Science (Distributed System Applications speciality) in 2006 from University of Picardie J. Verne. He is member of UbiStorage SA since 2006, in the UbiStorage solution, he developed most of the front-office and part of the back-office. He is in charge of the administration of the test platform used by UbiStorage. Research and Development interests: system administration, software development.

CIFRE PhD to recruit (NOT funded by this project)

Engineer to recruit (funded by this project)

A.2 LACL - Paris XII

Catalin Dima, 38 years old (Maître de conférences)

Catalin Dima is assistant professor at LACL since 2003. He obtained a PhD thesis from the Université Joseph Fourier Grenoble, at the Verimag research lab. His research interests focus on theoretical and algorithmic aspects of specification and verification using automata techniques and/or modal and temporal logics, as well as models of information leak.

Gaétan Hains, 43 years old (Professor and LACL director)

Gaétan Hains is full professor at LACL since 2006. He obtained a DPhil Thesis in 1990 at the University of Oxford, held various positions in Canada, Japan and France since then. He has been director of LIFO in Orléans for 6 years and IST programme officer at ANR in 2005-6. His research deals with parallel processing, formal models of performance and security.

Franck Pommereau, 32 years old (Maître de conférences)

Franck Pommereau is assistant professor at the LACL, where he obtained his PhD thesis in 2002. He has been involved into research on modelling complex systems using a class of structured Petri nets. In particular, his thesis was devoted to the introduction of time and preemption in such a model. His current work is oriented towards the efficient verification of this family of models, in particular, using massive parallelism.

PhD to recruit (funded by this project)

A.3 INRIA - MASCOTTE Project-team

Olivier Dalle (Assistant professor)

Olivier Dalle is assistant professor in the C.S. dept. of Faculty of Sciences at University of Nice-Sophia Antipolis (UNSA). He received his BS from U. of Bordeaux 1 and his M.Sc. and Ph.D. from UNSA. From 1999 to 2000 he was a post-doctoral fellow at the the french space agency center in Toulouse (CNES-CST), where he started working on component- based discrete event simulation of complex telecommunication systems, on the ASIMUT project. He joined the MASCOTTE team in 2000, and since Sept 2006 he has Sabbatical's year for research (délégation) in the MASCOTTE team. In 2005, he started the OSA project, an Open, Component-based Simulation Architecture based on the Fractal Component model and the Eclipse IDE.
His web page can be found here: `http://www.inria.fr/mascotte/Olivier.Dalle/`.

Michel Syska 42 years old (Maître de Conférences UNSA)

Michel Syska received the PhD degree in computer science from the Univ of Nice - Sophia Antipolis in 1992. His main recent research activities were driven by the following projects: RNRT PORTO, Design and Optimisation of WDM Optical Networks (1999 - 2001), IST CRESCCO, Critical Resource Sharing for Cooperation in Complex Systems (2002-2005) and IST AEOLUS, Algorithmic Principles for Building Efficient Overlay Computers(2005 -). He is also leading the Mascopt project-team (Java library for graph and network optimization). Research interests: network design, routing algorithms, distributed algorithms.

Collective duties and responsabilities:

- Member of the technical committee of IST FET AEOLUS
- Member of the I3S laboratory committee
- Menber of the CS 27 - University of Avignon

Selected References

[1] J.-F. Lalande, M. Syska, and Y. Verhoeven. Arrondi aléatoire et protection des réseaux WDM. In Ecole Polytechnique de l'Université de Tours, editor, *ROADEF*, number 6, pages 241–242, Tours, France, 2005.

[2] J-C. Bermond, O. de Rivoyre, S. Pérennes, and M. Syska. Groupage par tubes. In *ALGOTEL*, Banyuls, May 2003.

[3] G. Huiban, S. Pérennes, and M. Syska. Traffic grooming in WDM networks with multi-layer switches. In *IEEE ICC*, New-York, April 2002.

Stéphane Pérennes (CR CNRS)

Stéphane Pérennes is *Chargé de Recherche* CNRS in the I3S Laboratory (UMR 6070 CNRS/Univ Nice Sophia Antipolis). He is member of the Mascotte project-team since 1997. He obtained is MSc. from ENS Lyon in 1992 and his PhD. from Univ of Nice - Sophia Antipolis in 1996. In 1996-97, he was a postdoc fellow in TU-Delft (NL). Stéphane Pérennes has published over 30 papers in international journals and over 40 conference papers, on a wide range of theoretical topics related to networks, discrete event systems and graph theory.

Philippe Mussi (CR INRIA)

Philippe Mussi was born in Bordeaux (France) in 1958. He got an "Agrégation in Mathématiques" (1982) and a Doctorat in Computer Science from University of Paris V (1990). He is now a scientist at INRIA (Institut National de Recherche en Informatique et Automatique), Sophia-Antipolis, France, in charge of public relations, technology transfer, pôles de compétitivité and international partnerships. His current research interests include Parallel and Distributed Simulation of Discrete Event Systems, and Object-Oriented Simulation. He his the author of the Prosit simulation framework.

PhD to recruit (funded by this project)

Engineer to recruit (funded by this project)

A.4 EURECOM - NS Team

Refik Molva (Professor)

Refik Molva is a professor and the head of the Computer Communications Department at Institut Eurécom in Sophia Antipolis, France. His current research interests are in security protocols for self-organizing systems and privacy. He has been responsible for research projects on multicast and mobile

network security, anonymity and intrusion detection. Beside security, he worked on distributed multimedia applications over high speed networks and on network interconnection. Prior to joining Eurécom, he worked as a Research Staff Member in the Zurich Research Laboratory of IBM where he was one of the key designers of the KryptoKnight security system. He also worked as a network security consultant in the IBM Consulting Group in 1997. Refik Molva has a Ph.D. in Computer Science from the Paul Sabatier University in Toulouse (1986) and a B.Sc. in Computer Science (1981) from Joseph Fourier University, Grenoble, France.

Yves Roudier, 36 years old (Assistant Professor)

Yves Roudier is a member of the Computer Communications Department since 1998, after nearly two years of stay in Japan as an STA fellow researcher at the Electrotechnical Laboratory (ETL). His research interests currently include secure cooperative storage and ubiquitous computing security. His other interests include access control, mobile code security, distributed and reactive programming and middleware, as well as reflection based and aspect-oriented approaches to object-oriented language extension. Yves Roudier received a Ph.D. in Computer Science in 1996 and a B.Sc. in Computer Science and Business Management in 1992 from University of Nice Sophia Antipolis.

Pietro Michiardi (Assistant Professor)

Pietro Michiardi received the M.S. in Electrical Engineering from the "Politecnico di Torino" in 2001. In 1999 Pietro was granted an EU scholarship and joined "Institut Eurécom" to obtain a double degree M.S. in Communication Systems. In September 2001 Pietro joined "Ecole Nationale Superieure des Telecommunications" (ENST) as a Ph.D. candidate working on network security for Mobile Ad Hoc Networks. During his Ph.D. Pietro focused on research topics ranging from game theory and computational economic modeling applied to wireless networks, trust and reputation establishment schemes to advanced identity-based cryptographic techniques. He obtained his Ph.D. in December 2004. Since January 2005, Pietro is an assistant professor at Institut Eurécom in the Networking Department.

Postdoctoral researcher (funded by this project)

PhD to recruit (funded by this project)

A.5 INRIA - Projet REGAL

Luciana Arantes, 45 years old (Maître de Conférences U. Pierre et Marie Curie - Paris 6)

Luciana Arantes received her Ph.D. in Computer Science in 2000 from the University of Pierre et Marie Curie (Paris 6). She is also member of INRIA/LIP6 Regal group. Research interests: distributed algorithms, grid computing, memory consistency models.

Pierre Sens, 39 years old (Professeur U. Pierre et Marie Curie - Paris 6)

Pierre Sens received his Ph.D. in Computer Science in 1994, and the "Habilitation à diriger des recherches" in 2000 from the University Pierre et Marie Curie (Paris). Pierre Sens is heading of the Regal group (join research team between LIP6 and INRIA). Research interests: distributed systems, peer-to-peer file systems, fault tolerance and resource management in Grid configurations.

Collective duties and responsibilities: co-director of Laboratory of Computer Science (LIP6), member of several program committees of national and international conferences (CFSE, ISROC, CDUR, AC, NCA, SSS).

2 Internship students (funded by this project)

Postdoctoral researcher - engineer to recruit (funded by this project)

Appendix B

Financial Resources

This summarizes the financial part filled by partners.

B.1 UbiStorage

The following human resources are globally provided by UbiStorage

- S. Choplin (45%), CTO, leading this project
- G. Utard (15%), CEO,
- S. Drapeau (35%), R&D Engineer
- G. Carpentier (20%), Engineer
- CIFRE PhD (100%), to recruit, co-supervised by LACL

The following resources will be founded by this project:

- R&D Engineer, implementation of solution for experimentations
- R&D Equipment: cluster of 10 PC for experimentations (20K€)
- material: 4 personal computer and their softwares, communications, prints, ... (12K€)
- missions: travels for S. Choplin (2 days each six months with partners for project leading), travels for the PhD conferences and 1 months each year with partners (2 sites: Sophia Antipolis and Paris), travels for the Engineer (UbiStorage soft API formation for partners), travels for meeting at then end of each WP for 4 persons (35K€)
- external services: conception and production of popularization documentation (3K€), communication fees for dissemination of results (5K€), Intellectual Property fees (5K€)
- internal service: UbiStorage will provide one noébox(its currently commercialized solution) to each partner, (1K€ for supervision, corresponding to 20% of the customer price)
- compatbility assistance (2K€)

Total funds required: 400 000 €.

B.2 LACL

The following human resources are globally provided by LACL

- G. Hains (40%), Professeur
- C. Dima (35%), Maître de conférences
- F. Pommereau (25%), Maître de conférences

The following resources will be founded by this project:

- CDD Doctorant (100%)
- R&D Equipment: cluster of 10 PC for experimentations (20K€)
- material: 4 personal computer and their softwares, communications, prints, ... (12K€)
- missions: 1 international mission and 8 national missions per year (12K€), travel: PhD 1 travel to UbiStorage (1K€) and 1 international mission (2K€)
- administrative fees: Université Paris-12 (5.3K€ corresponding to 4%).

Total funds required: 136 900 €.

B.3 MASCOTTE

The following human resources are globally provided by MASCOTTE

- O. Dalle (50%), Maître de conférences
- S. Pérennes (45%), CR CNRS
- M. Syska (35%), Maître de conférecnes
- P. Mussi (20%), CR INRIA

The following resources will be founded by this project:

- CDD Doctorant (100%), research and simulator developments
- Engineer (50% = 18 months), simulator developments
- R&D Equipment: cluster of 10 PC for experimentations (20K€)
- material, 1 workstation and 1 laptop (6.4K€)
- missions: project meetings and 2 international conference per year for 2 persons 2 (15K€)
- administrative fees: 7K€ (4%)

Total funds required: 205 570 €.

B.4 REGAL

The following human resources are globally provided by REGAL

- P. Sens (30%), Professeur
- L. Arantes (20%), Maître de conférences
- M. Bouillaguet (10%), Maître de conférences

The following resources will be founded by this project:

- PostDoc (33%)
- Internship (33%)
- R&D Equipment: cluster of 10 PC for experimentations (10K€)
- material: 2 personal computer and their softwares (3.5K€)
- missions: project meeting, international and national conferences (15K€).
- administrative fees: 3K€ (4%)

Total funds required: 80 437 €.

B.5 EURECOM

The following human resources are globally provided by EURECOM

- Y. Roudier (20%), Maître de conférences
- P. Michiardi (10%), Maître de conférences
- R. Molva (8%), Professeur

The following resources will be founded by this project:

- PostDoc (50%)
- PhD (100%)
- R&D Equipment: cluster of 10 PC for experimentations (10K€)
- missions: project meetings for 3 persons (15K€).
- external services: financial audit (1.2K€)
- administrative fees: supervision (20%), personnel costs (40%), equipment (7%)

Total funds required: 273 483 €.