INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Project-Team PARIS*

# *Programming Parallel and Distributed Systems for Large Scale Numerical Simulation Applications*

*Rennes - Bretagne-Atlantique*

THEME NUM

Activity Report

2008

# Table of contents

**1. Team** . . . . . . . . **1**

**2. Overall Objectives** . . . . . . . . **2**

2.1. General objectives 2

2.1.1. Parallel processing to go faster 2

2.1.2. Distributed processing to go larger 2

2.1.2.1. Large-scale numerical simulation. 2

2.1.2.2. Resource aggregation. 3

2.1.3. Scientific challenges of the Paris Project-Team 3

2.2. Highlights of the year 3

**3. Scientific Foundations** . . . . . . . . **4**

3.1. Introduction 4

3.2. Data consistency 4

3.3. High availability 4

3.4. Data management in Grids 5

3.5. Component model 5

3.6. Adaptability 6

3.7. Chemical programming 6

**4. Application Domains** . . . . . . . . **7**

**5. Software** . . . . . . . . **7**

5.1. Kerrighed 7

5.2. PaCO++ 8

5.3. Adage 8

5.4. JuxMem 9

5.5. CoRDAGe 9

5.6. BlobSeer 10

5.7. Dynaco 10

5.8. Vigne 11

5.9. XtreemOS 11

5.10. Kargo and VMdeploy 11

**6. New Results** . . . . . . . . **12**

6.1. Introduction 12

6.2. Operating system and runtime for clusters and grids 12

6.2.1. Cluster operating systems 12

6.2.2. Grid operating systems 13

6.2.2.1. Vigne, a system for large-scale, dynamic Grids 13

6.2.2.2. XtreemOS Grid operating system 13

6.3. Middleware systems for computational grids 14

6.3.1. Parallel CORBA objects and components 14

6.3.2. Spatio-temporal skeleton software component models 15

6.3.3. Application deployment on computational Grids 15

6.3.4. Adaptation for data management 16

6.3.5. Adaptation for fault tolerance 16

6.4. Large-scale data management for grids 16

6.4.1. Using transparent sharing of distributed data for databases 16

6.4.2. Hierarchical Grid Storage based on the JuxMem Grid Data-Sharing Service and on the Gfarm Global File System 17

6.4.3. Toward transparent management of interactions between applications and resources 17

6.4.4. Toward efficent versioning for large objects under heavy access concurrency 17

6.5. Advanced programming models for the Grid 18

6.6. Experimental Grid infrastructures 19

**7. Contracts and Grants with Industry . . . 20**

7.1. EDF Contract 2 (2006-2008) 20

7.2. Sun Microsystems (2005-2008) 20

**8. Other Grants and Activities . . . 20**

8.1. Regional grants 20

8.1.1.1. PhD grants 20

8.1.1.2. 5000NET Project 20

8.1.1.3. Support to XtreemOS Project Management 20

8.2. National grants 21

8.2.1. ANR WP: ANR White Program 21

8.2.2. ANR CI: ANR Program on High-Performance Computing and Simulation 21

8.2.2.1. ANR CI DISC Project 21

8.2.2.2. ANR CI LEGO Project 21

8.2.2.3. ANR CI NUMASIS Project 21

8.2.3. ANR MD: ANR Program on Data Masses and Ambient Knowledge 21

8.2.4. ANR SI: ANR Program on Security and Informatics 21

8.2.5. INRIA ADT 22

8.2.6. ANR TLOG: ANR Program on Software Technologies 22

8.3. European grants 22

8.3.1. CoreGRID NoE Project 22

8.3.2. EchoGRID Specific Support Action 23

8.3.3. XtreemOS IP Project 23

8.3.4. S-Cube NoE Project 24

**9. Dissemination . . . 24**

9.1. Community animation 24

9.1.1. Leaderships, Steering Committees and community service 24

9.1.2. Editorial boards, direction of program committees 24

9.1.3. Program Committees 24

9.1.4. Evaluation committees, consulting 25

9.2. Academic teaching 25

9.3. Conferences, seminars, and invitations 26

9.4. Administrative responsibilities 26

9.5. Miscellaneous 27

**10. Bibliography . . . 27**

# 1. Team

**Research Scientist**

Thierry Priol [ Research Director (DR), team leader, HdR ]
Gabriel Antoniu [ Research Associate (CR) ]
Yvon Jégou [ Research Associate (CR) ]
Christine Morin [ Research Director (DR), HdR ]
Christian Pérez [ Research Associate (CR), until September 30, 2008, HdR ]

**Faculty Member**

Françoise André [ Professor, HdR ]
Jean-Pierre Banâtre [ Professor, HdR ]
Jean-Louis Pazat [ Professor, HdR ]
Luc Bougé [ Professor, ENS CACHAN Brittany Campus, HdR ]

**Technical Staff**

David Margery [ Research Engineer (IR, part-time 50%) ]
Pascal Morillon [ Engineer (IE) ]
Landry Breuil [ INRIA Associate Engineer (IA), ANR MD Project LEGO ]
Matthieu Fertré [ INRIA Associate Engineer (IA), XTREEMOS IP Project, till November 2008 ]
Serge Guelton [ INRIA Associate Engineer (IA), ANR SI Project Safescale ]
Mathieu Kermarrec [ INRIA Associate Engineer (IA), ANR CI Project NUMASIS ]
Surbhi Chitre [ INRIA Associate Engineer (IA), XTREEMOS IP Project, since December 2008 ]
Raúl López Lozano [ INRIA Associate Engineer (IA), ANR CI Project DISC ]
Oscar Sanchez [ INRIA, XTREEMOS IP Project ]

**PhD Student**

Julien Bigot [ MENRT Grant ]
Hinde Lilia Bouziane [ INRIA Grant ]
Loïc Cudennec [ INRIA and Brittany Regional Council Grant ]
Boris Daix [ CIFRE EDF Industrial Grant ]
Jérôme Gallard [ INRIA Grant ]
Sylvain Jeuland [ INRIA Grant ]
Bogdan Nicolae [ MENRT Grant ]
Thomas Ropars [ MENRT Grant ]
Pierre Riteau [ MENRT Grant, since October 2008 ]
Mohamed Zouari [ INRIA and Brittany Regional Council Grant ]
Alexandra Carpen-Amarie [ INRIA CORDI Grant, since October 2008 ]
Diana Moise [ INRIA and Brittany Regional Council Grant, since October 2008 ]
André Lage [ INRIA Grant, since october 2008 ]
Guillaume Gauvrit [ MENRT Grant, since october 2008 ]

**Post-Doctoral Fellow**

Xingwu Liu [ ECHOGRID Post-Doc, till November 2008 ]
Adrien Lèbre [ INRIA Post-Doc, XTREEMOS IP Project, till August 2008 ]
Manuel Caeiro Rodríguez [ COREGRID Post-Doc, till March 2008 ]

**Administrative Assistant**

Maryse Auffray [ Secretary (TR) INRIA ]
Sandrine L'Hermitte [ XTREEMOS Scientific Coordinator Assistant INRIA ]
Olivia Vasselin [ CoreGRID Scientific Coordinator Assistant INRIA, till August 2008 ]

# 2. Overall Objectives

## 2.1. General objectives

The PARIS Project-Team aims at contributing to the programming of parallel and distributed infrastructures for large-scale numerical simulation applications. Its goal is to design operating systems and middleware to ease the use of such computing infrastructures for the targeted applications. Such applications enable the speed-up of the design of complex manufactured products, such as cars or aircrafts, thanks to numerical simulation techniques.

As computer performance rapidly increases, it is possible to foresee in the near future comprehensive simulations of these designs that encompass multi-disciplinary aspects (structural mechanics, computational fluid dynamics, electromagnetism, noise analysis, etc.). Numerical simulations of these different aspects are not carried out by a single computer due to the lack of computing and memory resources. Instead, several clusters of inexpensive PCs, and probably federations of clusters (aka. *Grids*), have to be simultaneously used to keep simulation times within reasonable bounds. Moreover, simulation have to be performed by different research teams, each of them contributing its own simulation code. These teams may all belong to a single company, or to different companies possessing appropriate skills and computing resources, thus adding geographical constraints. By their very nature, such applications will require the use of a computing infrastructure that is *both* parallel and distributed.

The PARIS Project-Team is engaged in research along five topics: *Operating System and Runtime for Clusters and Grids*, *Middleware Systems for Computational Grids*, *Large-Scale Data Management for Grids*, *Advanced Programming Models for the Grid* and *Experimental Grid Infrastructures*.

The research activities of the PARIS Project-Team encompass both basic research, seeking conceptual advances, and applied research, to validate the proposed concepts against *real* applications. The project-team is also heavily involved in managing a national grid computing infrastructure (GRID'5000) enabling large-scale experiments.

### *2.1.1. Parallel processing to go faster*

Given the significant increase of the performance of microprocessors, computer architectures and networks, clusters of standard personal computers now provide the level of performance to make numerical simulation a handy tool. This tool should not be used by researchers only, but also by a large number of engineers, designing complex physical systems. Simulation of mechanical structures, fluid dynamics or wave propagation can nowadays be carried out in a couple of hours. This is made possible by exploiting multi-level parallelism, simultaneously at a fine grain within a microprocessor, at a medium grain within a single multi-processor PC, and/or at a coarse grain within a cluster of such PCs. This unprecedented level of performance definitely makes numerical simulation available for a larger number of users such as SMEs. It also generates new needs and demands for more accurate numerical simulation. Parallel processing alone cannot meet this demand.

### *2.1.2. Distributed processing to go larger*

These new needs and demands, mixing high-performance and collaborative multidisciplinary works, are motivated by the constraints imposed by a worldwide economy: making things faster, better and cheaper.

*2.1.2.1. Large-scale numerical simulation.*

Large scale numerical simulation will without a doubt become one of the key technologies to meet such constraints. In traditional numerical simulation, only one simulation code is executed. In contrast, it is now required to *couple* several such codes together in a single simulation.

A large-scale numerical simulation application is typically composed of several codes, not only to simulate one physics, but to perform multi-physics simulation. One can imagine that the simulation times will be in the order of weeks and sometimes months depending on the number of physics involved in the simulation, and depending on the available computing resources.

Parallel processing extends the number of computing resources locally: it cannot significantly reduce simulation times, since the simulation codes will not be localized in a single geographical location. This is particularly true with the global economy, where complex products (such as cars, aircrafts, etc.) are not designed by a single company, but by several of them, through the use of subcontractors. Each of these companies brings its own expertise and tools such as numerical simulation codes, and even its private computing resources. Moreover, they are reluctant to give access to their tools as they may at the same time compete for some other projects. It is thus clear that distributed processing cannot be avoided to manage large-scale numerical applications

*2.1.2.2. Resource aggregation.*

More generally, the development of large scale distributed systems and applications now rely on resource sharing and aggregation. Distributed resources, whether related to computing, storage or bandwidth, are aggregated and made available to the whole system. Not only this aggregation greatly improves the performance as the system size increases, but many applications would simply not have been possible without such a model (peer-to-peer file sharing, ad-hoc networks, application-level multicast, publish-subscribe applications, etc.).

### *2.1.3. Scientific challenges of the Paris Project-Team*

The design of large-scale simulation applications raises technical and scientific challenges, both in applied mathematics and computer science. The PARIS Project-Team mainly focuses its effort on Computer Science. It investigates new approaches to build software mechanisms that hide the complexity of programming computing infrastructures that are *both* parallel and distributed. Our contribution to the field can thus be summarized as follows:

*combining parallel and distributed processing whilst preserving performance and transparency.*

This contribution is developed along five directions.

Operating system and runtime for clusters and grids. The challenge is to design and build an operating system for clusters and grids hiding to the programmers and the users, the fact that resources (processors, memories, disks) are distributed.

Middleware systems for computational grids. The challenge is to design a middleware implementing a component-based approach for grids. Large-scale numerical applications will be designed by combining together a set of components encapsulating simulation codes. The challenge is to seamlessly mix both parallel and distributed processing.

Large-scale data management for grids. One of the key challenges in programming grid computing infrastructures for real, is data management. It has to be carried out at an unprecedented scale, and to cope with the native dynamicity and heterogeneity of the underlying grids.

Advanced programming models for the Grid. This topic aims at contributing to study unconventional approaches for the programming of grids based on the *chemical metaphors*. The challenge is to exploit such metaphors to make the use, including the programming, of grids more intuitive and simpler.

Experimental Grid Infrastructures. The challenge here is to be able to design and to build an *instrument* (in the sense of a large scientific instrument, like a telescope) for computer scientists involved in grid research. Such an instrument has to be highly reconfigurable and scalable to several thousand of resources.

## 2.2. Highlights of the year

In 2008, the project-team has the following highlights:

- The CoreGRID Network of Excellence funded by the European Commission has been successfully completed after four years of existence. It has been positively evaluated by the experts appointed by the European Commission. This NoE was scientifically coordinated by the project-team (T. Priol).
- The XtreemOS European project, coordinated by the project-team (C. Morin) announced the first public version of the XtreemOS Grid Operating System.

# 3. Scientific Foundations

## 3.1. Introduction

Research activity within the PARIS Project-Team encompasses several areas: operating systems, middleware and programming models. We have chosen to provide a brief presentation of some of the scientific foundations associated with them.

## 3.2. Data consistency

A shared virtual memory system provides a global address space for a system where each processor has only physical access to its local memory. Implementing of such a concept relies on the use of complex cache coherence protocols to enforce data consistency. To allow the correct execution of a parallel program, it is required that a read access performed by one processor returns the value of the last write operation previously performed by any other processor. Within a distributed or parallel a system, the notion of the *last* memory access is sometimes only partially defined, since there is no global clock to provide a total order of the memory operation.

It has always been a challenge to design a shared virtual memory system for parallel or distributed computers with distributed physical memories, capable of providing comparable performance with other communication models such as message-passing. *Sequential Consistency* [92] is an example of a memory model for which all memory operations are consistent with a total order. Sequential Consistency requires that a parallel system having a global address space appears to be a multiprogramming uniprocessor system to any program running on it. Such a strict definition impacts on the performance of shared virtual memory systems due to the large number of messages that are required (page access, invalidation, control, etc.). Moreover Sequential Consistency is not necessarily required to correctly run parallel programs, in which memory operations to the global address space are guarded by synchronization primitives.

Several other memory models have thus been proposed to relax the requirements imposed by sequential consistency. Among them, *Release Consistency* [86] has been thoroughly studied since it is well adapted to programming parallel scientific applications. The principle behind Release Consistency is that memory accesses are (should?) always be guarded by synchronization operations (locks, barriers, etc.), so that the shared memory system only needs to ensure consistency at synchronization points. Release Consistency requires the use of two new operations: *acquire* and *release*. The aim of these two operations is to specify when to propagate the modifications made to the shared memory systems. Several implementations of Release Consistency have been proposed [90]: an *eager* one, for which modifications are propagated at the time of a release operation; and a *lazy* one, for which modifications are propagated at the time of an acquire operation. These alternative implementations differ in the number of messages that needs to be sent/received, and in the complexity of their implementation [91].

Implementations of Release Consistency rely on the use of a logical clock such as a vector clock [94]. One of the drawback of such a logical clock is its lack of scalability when the number of processors increases, since the vector carries one entry per processor. In the context of computing systems that are both parallel and distributed, such as a grid infrastructure, the use of a vector clock is impossible in practice. It is thus necessary to find new approaches based on logical clocks that do not depend on the number of processors accessing the shared memory system. Moreover, these infrastructures are natively *hierarchical*, so that the consistency model should better take advantage of it.

## 3.3. High availability

> “A distributed system is one that stops you getting any work done when a machine you’ve never even heard about crashes.” (Leslie Lamport)

The *availability* [87] of a system measures the ratio of service accomplishment conforming to its specifications, with respect to elapsed time. A system *fails* when it does not behave in a manner consistent with its specifications. An error is the consequence of a *fault* when the faulty part of the system is activated. It may lead to the system *failure*. In order to provide highly-available systems, *fault tolerance techniques* [93] based on redundancy can be implemented. Abstractions like *group membership*, *atomic multicast*, *consensus*, etc. have been defined for fault-tolerant distributed systems.

*Error detection* is the first step in any fault tolerance strategy. *Error treatment* aims at avoiding that the error leads to the system failure.

*Fault treatment* consists in avoiding that the fault be activated again. Two classes of techniques can be used for fault treatment: *reparation* which consists in eliminating or replacing the faulty module; and *reconfiguration* which consists in transferring the load of the faulty element to valid components.

Error treatment can be of two forms: *error masking* or *error recovery*. Error masking is based on hardware or software redundancy in order to allow the system to deliver its service despite the error. Error recovery consists in restoring a correct system state from an erroneous state. In *forward error recovery* techniques, the erroneous state is transformed into a safe state. *Backward error recovery* consists in periodically saving the system state, called a *checkpoint*, and rolling back to the last saved state if an error is detected.

A *stable storage* guarantees three properties in presence of failures: (1) *integrity*, data stored in stable storage is not altered by failures; (2) *accessibility*, data stored in stable storage remains accessible despite failures; (3) *atomicity*, updating data stored in stable storage is an all or nothing operation. In the event of a failure during the update of a group of data stored in stable storage, either all data remain in their initial state or they all take their new value.

## 3.4. Data management in Grids

Past research on distributed data management led to three main approaches. Currently, the most widely-used approach to data management for distributed grid computation relies on *explicit data transfers* between clients and computing servers. As an example, the *Globus* [76] platform provides data access mechanisms (like data catalogs) based on the *GridFTP* protocol. Other explicit approaches (e.g., *IBP*) provide a large-scale data storage system, consisting of a set of buffers distributed over Internet. The user can “rent” these storage areas for efficient data transfers.

In contrast, *Distributed Shared Memory* (DSM) systems provide *transparent* data sharing, via a virtual, unique address space accessible to physically distributed machines. It is the responsibility of the DSM system to localize, transfer, replicate data, and guarantee their consistency according to some semantics. Within this context, a variety of consistency models and protocols have been defined. Nevertheless, existing DSM systems have generally shown satisfactory efficiency only on small-scale configurations, up to a few tens of nodes.

Recently, *peer-to-peer* (P2P) has proven to be an efficient approach for large-scale resource (data or computing resources) sharing [95]. The peer-to-peer communication model relies on a symmetric relationship between peers which may act both as clients and servers. Such systems have proven able to manage very large and dynamic configurations (millions of peers). However, several challenges remain. More specifically, as far as data sharing is concerned, most P2P systems focus on sharing *read-only* data, that do not require data consistency management. Some approaches, like *OceanStore* and *Ivy*, deal with *mutable* data in a P2P with restricted use. Today, one major challenge in the context of large-scale, distributed data management is to define appropriate models and protocols allowing to guarantee both *consistency* of replicated data and *fault tolerance*, in *large-scale, dynamic environments*.

## 3.5. Component model

Software component technology [101] has been emerging for some years, even though its underlying intuition is not very recent. Building an application based on components emphasizes programming by *assembly*, that is, *manufacturing*, rather than by *development*. The goals are to focus expertise on domain fields, to improve software quality, and to decrease the time-to-market thanks to reuse of existing codes.

The CORBA Component Model (CCM), which is part of the latest CORBA [97] specifications (Version 3), appears to be the most complete specification for components. It allows the deployment of a set of components into a distributed environment. Moreover, it supports heterogeneity of programming languages, operating systems, processors, and it also guarantees interoperability between different implementations. However, CCM does not provide any support for parallel components.

The CORBA Component Architecture (CCA) Forum [79] aims at developing a standard which specifically addresses the needs of the HPC community. Its objective is to define a minimal set of standard interfaces that any high-performance component framework should provide to components, and may expect from them, in order to allow disparate components to be composed together into a running application. CCA aims at supporting *both* parallel and distributed applications.

## 3.6. Adaptability

Due to the dynamic nature of large-scale distributed systems in general, and the Grid in particular, it is very hard to design an application that fits well in any configuration. Moreover, constraints such as the number of available processors, their respective load, the available memory and network bandwidth are not static. For these reasons, it is highly desirable that an application could take into account this dynamic context in order to get as much performance as possible from the computing environment.

Dynamic adaptation of a program is the modification of its behavior according to changes of the environment. This adaptivity can be achieved in many different ways, ranging from a simple modification of some parameters, to the total replacement of the running code. In order to achieve adaptivity, a program needs to be able to get information about the environment state, to make a decision according to some optimization rules, and to modify or replace some parts of its code.

Adaptivity has been implemented by designing ad hoc applications that take into account the specificities of the target environment. For example, this was done for the Web applications access protocol on mobile networks by defining the WAP protocol [78]. A more general way is to provide mechanisms enabling dynamic self-adaptivity by changing the program's behavior. In most cases, this has been achieved by embedding the adaptation mechanism within the application code. For example, the AdOC compression algorithm [89] includes such a mechanism to dynamically change the compression level according to the available resources.

However, it is desirable to separate the adaptation engine from the application code, in order to make the code easier to maintain, and to easily change or improve the adaptation policy. This was done for wireless and mobile environments by implementing a framework [85] that provides generic mechanisms for the adaptation process, and for the definition of the adaptation rules.

## 3.7. Chemical programming

The chemical reaction metaphor has been discussed in various occasions in the literature. This metaphor describes computation in terms of a chemical solution in which molecules (representing data) interact freely according to reaction rules. Chemical models use the multiset as their basic data structure. Computation proceeds by rewritings of the multiset which consume elements according to reaction conditions and produce new elements according to specific transformation rules.

To the best of our knowledge, the GAMMA formalism was the first "chemical model of computation" proposed as early as in 1986 [82] and extended later [83].

A GAMMA program is a collection of reaction rules acting on a multiset of basic elements. A reaction rule is made of a condition and an action. Execution proceeds by replacing elements satisfying the reaction condition by the elements specified by the action. The result of a GAMMA program is obtained when a stable state is reached that is to say when no more reactions can take place. Here is an example illustrating the GAMMA style of programming:

$$primes = \mathbf{replace}\ x, y\ \mathbf{by}\ y\ \mathbf{if}\ multiple(x, y)$$

The reaction $primes$ computes the prime numbers lower or equal to a given number $N$ when applied to the multiset of all numbers between 2 and $N$ ($multiple(x, y)$ is true if and only if $x$ is a multiple of $y$). Let us emphasize the conciseness and elegance of these programs. Nothing had to be said about the order of evaluation of the reactions. If several disjoint pairs of elements satisfy the condition, the reactions can be performed in parallel.

GAMMA makes it possible to express programs without artificial sequentiality. By artificial, we mean sequentiality only imposed by the computation model and unrelated to the logic of the program. This allows the programmer to describe programs in a very abstract way. In some sense, one can say that GAMMA programs express the very idea of an algorithm without any unnecessary linguistic idiosyncrasies. The interested reader may find in [83] a long series of examples (string processing problems, graph problems, geometry problems, etc.) illustrating the GAMMA style of programming and in [81] a review of contributions related to the chemical reaction model. Later, the idea was developed further into the CHAM [84], the P-systems [98], etc. Although built on the same basic paradigm, these proposals have different properties and different expressive powers.

The $\gamma$-calculus [80] is an attempt to identify the basic principles behind chemical models. It exhibit a minimal chemical calculus, from which all other “chemical models” can be obtained by addition of well-chosen features. Essentially, this minimal calculus incorporates the $\gamma$-reduction which expresses the very essence of the chemical reaction, and the associativity and commutativity rules which express the basic properties of chemical solutions.

# 4. Application Domains

## 4.1. Application Domains

**Keywords:** *Scientific computing*, *co-operative applications*, *large-scale computing*.

The project-team research activities address scientific computing and specifically numerical applications that require the execution of several codes simultaneously (code-coupling). This kind of applications requires both the use of parallel and distributed systems. Parallel processing is required to address performance issues. Distributed processing is needed to fulfill the constraints imposed by the localization and the availability of resources, or for confidentiality reasons. Such applications are being experimented within contracts with the industry or through our participation to application-oriented research grants.

# 5. Software

## 5.1. Kerrighed

**Keywords:** *Cluster operating system*, *checkpointing*, *distributed file system*, *single system image (SSI)*.

**Participants:** Matthieu Fertré, Jérôme Gallard, Adrien Lèbre, Christine Morin, Pierre Riteau.

Contact: Christine Morin, `Christine.Morin@irisa.fr`

URL: http://www.kerrighed.org/ and http://ssi-oscar.gforge.inria.fr/

Status: Registered at APP, under Reference `IDDN.FR.001.480003.006.S.A.2000.000.10600`.

License: GNU General Public License (GPL) version 2. KERRIGHED is a registered trademark.

Presentation: KERRIGHED is a *Single System Image* (SSI) operating system for high-performance computing on clusters. It provides the user with the illusion that a cluster is a virtual SMP machine. KERRIGHED is based on Linux which is slightly patched and extended with a kernel module. It is Posix compliant. Legacy sequential or parallel applications running on Linux can be executed without modification on top of KERRIGHED. KERRIGHED (version V2.3.0) includes 40,000 lines of code (mostly in C). It involved more than 250 persons-months. Professional support is provided by Kerlabs http://www.kerlabs.com, a spin-off from PARIS project-team created in 2006. In 2008, kDFS distributed file system [30] has been improved from a stability point of view and optimized. It has been extended to provide an efficient support to file checkpointing [70]. We have also implemented full support for providing cluster-wide IPC. Finally, we have improved the checkpoint/recovery mechanisms to support multithreaded applications and applications using IPC [73]. Kerrighed is used in the cluster flavour of XTREEMOS Grid operating.KERRIGHED has been demonstrated at *Supercomputing 2008 Conference* (Austin, Texas, November 2008, P. Riteau, Ch. Morin).

## 5.2. PaCO++

**Keywords:** *CORBA*, *Grid*, *data parallelism*, *middleware system*.

**Participants:** Raúl López Lozano, Christian Pérez, Thierry Priol.

Contact: Christian Pérez, `Christian.Perez@inria.fr`

URL: http://paco.gforge.inria.fr/

Status: Registered at APP, under Reference `IDDN.FR.001.450014.000.S.P.2004.000.10400`.

License: GNU General Public License (GPL) version 2 and GNU Lesser General Public License (LGPL) version 2.1.

Presentation: The PACO++ objective is to allow a simple and efficient embedding of a SPMD code into a parallel CORBA object, and to allow parallel communication flows and data redistribution during an operation invocation on such a parallel CORBA object. PACO++ provides an implementation of the concept of parallel object applied to CORBA. PACO++ extends CORBA, but does not modify the underlying model. The parallelism of an object is in fact considered to be an implementation feature of this object, and the OMG IDL is not dependent on it. The development of PACO++ started at the end of 2002. It involved 65 persons-months. The current version (0.3) has been released in October 2008. The version 0.3 of PACO++ includes 63,000 lines of Java, 7,700 lines of Python, 15,000 lines of C++ and 2,000 lines of `shell`, `make` and `configure` scripts. It is currently used within two French ANR CI projects: DISC and NUMASIS. PACO++ is co-developed with the EDF R&D company which is including it into Salome, an open source integration platform for numerical simulation.

## 5.3. Adage

**Keywords:** *Grid*, *deployment*, *middleware system*.

**Participants:** Landry Breuil, Loïc Cudennec, Christian Pérez, Thierry Priol.

Contact: Christian Pérez, `Christian.Perez@inria.fr`

URL: http://www.irisa.fr/paris/ADAGE/

Status: Registered at APP, under Reference `IDDN.FR.001.270020.000.S.P.2007.000.10000`.

License: GNU General Public License (GPL) version 2.

Presentation: ADAGE (*Automatic Deployment of Applications in a Grid Environment*) is a research prototype that aims at studying the deployment issues related to multi-middleware applications. Its original contribution is to use a *generic* application description model (*GADe*) to transparently handle various middleware systems. With respect to application submission, ADAGE requires an application description, which is specific to a programming model, a reference to a resource information service (MDS2, or an XML file), and a control parameter file. The support of multi-middleware applications is based on a plug-in mechanism. ADAGE currently deploys static applications and also pseudo-dynamic applications by re-deploying new application elements. It supports standard programming models like MPI (*MPICH1*, *MPICH2* and *OpenMPI*), CCM, DIET, JXTA, P2P, and Gfarm. The version 0.4 of ADAGE includes a core of 10,000 lines of C++, 6,000 lines of planners and 10,500 lines for the 9 application plugins. It has been registered at APP in June 2007 and the public release has been delivered in Mars 2008.

## 5.4. JuxMem

**Keywords:** *JXTA*, *Peer-to-peer*, *data grids*, *large-scale data management*.

**Participants:** Gabriel Antoniu, Luc Bougé, Landry Breuil, Loïc Cudennec.

Contact: Gabriel Antoniu, `Gabriel.Antoniu@irisa.fr`

URL: http://juxmem.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 2.1.

Status: Registered at APP, under Reference `IDDN.FR.001.180015.000.S.P.2005.000.10000`.

Presentation: JUXMEM is a supportive platform for a data-sharing service for grid computing. This service addresses the problem of managing mutable data on dynamic, large-scale configurations. It can be seen as a hybrid system combining the benefits of *Distributed Shared Memory* (DSM) systems (transparent access to data, consistency protocols) and *Peer-to-Peer* (P2P) systems (high scalability, support for resource volatility). JUXMEM's architecture decouples fault-tolerance management from consistency management. Multiple consistency protocols can be built using fault-tolerant building blocks such as *consensus*, *atomic multicast*, *group membership*. Currently, a hierarchical protocol implementing the entry consistency model is available. A more relaxed consistency protocol adapted to visualization is also available. Up to version 0.4 (included), JuxMem is based on the *JXTA* generic platform for P2P services (Sun Microsystems, http://www.jxta.org/). This version includes 16,700 lines of Java code and 16,000 lines of C code. Implementation started in February 2003. In 2008, a lighter version of JuxMem (0.5), non-dependent on JXTA was released. It includes 4600 lines of C++ code. JUXMEM is currently used for transparent data sharing within the following running projects: ANR CI LEGO project, and ANR MD RESPIRE project. An industrial collaboration with Sun Microsystems has been funded between August 2005 for 3 years (Loïc Cudennec's Ph.D. thesis). JUXMEM is currently used within an international collaboration with the University of Tsukuba. Other past users: University of Illinois at Urbana Champaign, University of Pisa, University of Calabria.

## 5.5. CoRDAGe

**Keywords:** *CoRDAGe*, *Computing grid*, *autonomic*, *co-deployment service*, *distributed application*, *dynamicity*, *re-deployment*.

**Participants:** Gabriel Antoniu, Luc Bougé, Loïc Cudennec.

Contact: Loïc Cudennec, `Loic.Cudennec@irisa.fr`

URL: http://cordage.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 3.

Status: Registration at APP in progress.

Presentation: CoRDAGe is a generic co-deployment and re-deployment service for grid computing applications. It addresses the deployment of applications in a dynamic way: it allows redeployment and reconfiguration during the execution, as well as coordinated deployment of multiple, heterogeneous applications. The service interfaces applications with grid reservation and deployment tools, thus making all interactions transparent for the applications and the final user. It proposes two principal functionalities. CoRDAGe has been developed since January 2008 within an INRIA Gforge project. The current implementation features near all functionalities that come with the model. It includes more than 6,700 lines of C++, C and Perl code. It relies on the Adage deployment tool and the *OAR* resource scheduler. CoRDAGe can handle applications based on the *JXTA* peer-to-peer system, the JuxMem data-sharing service and the *Gfarm* distributed file system. It has been tested within the Grid'5000 experimental platform, using up to 386 nodes and 6 sites in a single experiment. The CoRDAGe prototype has been used within the ANR CI LEGO project, and within the ANR MD RESPIRE project.

## 5.6. BlobSeer

**Keywords:** *blob*, *fine grain access*, *heavy access concurrency*, *huge storage object*, *versioning*.

**Participants:** Gabriel Antoniu, Luc Bougé, Bogdan Nicolae.

Contact: Bogdan Nicolae, `Bogdan.Nicolae@irisa.fr`

URL: http://blobseer.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 3.

Status: This software is available on INRIA's forge. Registration with AP P is in progress.

Presentation: BlobSeer is a huge blob (binary large object) management service to be used as a specialized storage backend for large scale distributed computing applications that abstract data input as huge sequences of bytes (such as MapReduce applications). It exports a simple, yet versatile blob manipulation interface, that allows reading, writing and appending to huge blobs while providing full transparency with regard to data allocation, replication and fault tolerance. BlobSeer has been developed since January 2008 within the JuxMem INRIA Gforge project, and has moved in its separate project since September 2008. The current implementation features 23,000 lines of C++/Perl. The service itself relies on the Boost collection of C++ libraries, libconfig and OpenSSL. Perl code is used to handle deployment on Grid'5000, which is done through the *OAR* resource scheduler. Preliminary tests have proven correctness and performance with up to 400 nodes from 3 different sites. Ongoing development targets integration with Hadoop and PostgreSQL

## 5.7. Dynaco

**Keywords:** *Grid*, *components*, *framework*, *objects*.

**Participants:** Françoise André, Jean-Louis Pazat.

Contact: Jean-Louis Pazat, `Jean-Louis.Pazat@irisa.fr`

URL: http://dynaco.gforge.inria.fr/

Status: Version 0.2 is available.

License: GNU Lesser General Public License (LGPL) version 2.1.

Presentation: Dynaco (*Dynamic Adaptation for Components*) is a framework that helps in designing and implementing dynamically adaptable components. This framework is developed by the Paris Project-Team. The implementation of Dynaco is based on the *Fractal Component Model* and its formalism.

## 5.8. Vigne

**Participants:** Christine Morin, Thomas Ropars.

Contact: Christine Morin, `Christine.Morin@irisa.fr`

URL: http://www.irisa.fr/paris/web/GridOS.html

Status: Version 1.0 soon available

License: GNU General Public License (GPL).

Presentation: VIGNE is a prototype of a grid-aware operating system for grids, whose goal is to ease the use of computing resources in a grid for executing distributed applications. VIGNE is made up of a set of operating system services based on a peer-to-peer infrastructure. This infrastructure currently implements a structured overlay network inspired from *Pastry* and an unstructured overlay network inspired from *Scamp* for join operations. On top of the structured overlay network, a transparent data-sharing service based on the sequential consistency model has been implemented. It is able to handle an arbitrary number of simultaneous reconfigurations. An application execution management service has also been implemented including resource discovery, resource allocation, and application monitoring services. In 2008, the VIGNE prototype has been extended with active replication mechanisms to make the application execution management service highly available on top of the structured overlay network. The VIGNE prototype has been developed in C and includes 30,000 lines of code. This prototype has been coupled with a discrete-event simulator.

## 5.9. XtreemOS

**Participants:** Surbhi Chitre, Matthieu Fertré, Jérôme Gallard, Yvon Jégou, Sylvain Jeuland, Adrien Lèbre, Christine Morin, Pierre Riteau, Oscar Sanchez.

Contact: Christine Morin, `Christine.Morin@irisa.fr`

URL: http://gforge.inria.fr/projects/xtreemos

Status: Version 1.0

License: GPL-2/BSD depending on software packages composing the system

Presentation: XTREEMOS is a prototype of a Grid Operating system based on Linux with native support for virtual organizations. Three flavours of XTREEMOS are developed for individual PCs, clusters and mobile devices. XTREEMOS has been developed by the XtreemOS consortium. The first public version of XtreemOS (PC and cluster flavours) has been released in December 2008 [75]. XTREEMOS has been demonstrated at Internet of Services 2008, Brussels, Belgium, in September 2008 (Y. Jégou), SC'08, Austin, USA (Y. Jégou, S. Jeuland, C. Morin, O. Sanchez), and ICT 2008, Lyon, France (Y. Jégou, S. Jeuland, O. Sanchez) in November 2008. XTREEMOS software is a set of services developed in Java, C++ and C. XtreemOS cluster version leverages KERRIGHED single system image operating system. A permanent testbed composed of computers provided by several XTREEMOS partners has been setup for experimentation and demonstration purposes.

## 5.10. Kargo and VMdeploy

**Participants:** Jérôme Gallard, Adrien Lèbre, Christine Morin.

Contact: Christine Morin, `Christine.Morin@irisa.fr`

Status: Version V1.0 (experimental)

License: GPL-2

Presentation: **Kargo** is a shell script to deploy Kerrighed on cluster nodes. Based on a diskless approach, the script first deploys the master node that hosts the root image (NFSRoot) of the slaves nodes. After this it boots the slaves nodes in the new environment from the NFS root. **VMdeploy** is a shell script that deploys VMware virtual machines on the Grid'5000 platform, allowing the users to run their scripts into virtual machines. The framework reserves the nodes, deploys a customized environment (with VMware server installed) with the help of Kadeploy, and then instantiates the virtual machines on the reserved nodes. VMdeploy is able to reserve nodes either as soon as possible (interactive) or when resources are idle (best effort). To avoid wasting resources due to the direct killing of best effort jobs by the OAR scheduler, the framework is able to automatically and transparently preempt/restart and migrate best effort jobs [65], [56].

# 6. New Results

## 6.1. Introduction

Research results are presented according to the scientific challenges of the PARIS Project-Team.

## 6.2. Operating system and runtime for clusters and grids

### *6.2.1. Cluster operating systems*

**Participants:** Matthieu Fertré, Jérôme Gallard, Adrien Lèbre, Christine Morin, Pierre Riteau.

KERRIGHED is a Single System Image (SSI) OS providing the illusion that a distributed cluster is a virtual multiprocessor machine. In 2008, we continued to contribute to the design and implementation of KERRIGHED in the framework of the XTREEMOS European IP project. We contributed to several new releases (based on Linux 2.6.20) of KERRIGHED (2.2.1 and 2.3.0) and of its customized version for the cluster flavour of XtreemOS (LinuxSSI 0.9 in June 2008 and 1.0 in November 2008). We contributed to the packaging of KERRIGHED and LinuxSSI for Mandriva Linux distribution and XtreemOS LiveCD.

The implementation of IPC System V semaphores arrays and messages queues in KERRIGHED has been finalized [73]. These mechanisms have been heavily tested and fixed to run on SMP cluster nodes. Patches have been submitted and accepted to the Linux Test Project (LTP - http://ltp.sourceforge.net/) to fix concurrent running of IPC tests in an SMP context.

KERRIGHED checkpoint/restart mechanisms have been significantly improved [73]. First, session identifiers and process group identifiers are restored correctly for all application processes. Secondly, an API has been developed to easily handle the checkpoint/restart of different objects that can be shared by processes of a same tree depending of the flags used for the *clone* system call: file pointers and the *files_struct*, *fs_struct*, *mm_struct*, *sighand_struct*, *signal_struct*, *sysvsem* descriptors. Before this work, when an object was shared by two processes at the time of the checkpoint, the object was dumped twice and when restarting the application, each process had its own copy of the object. Now, the object is checkpointed only once and the sharing is rebuilt. Thanks to this API, checkpointing/restarting multi-threaded applications has been implemented (the prototype has been validated by checkpointing/restarting a Java Virtual Machine).

We also worked on the design and implementation of *kDFS* (kernel/KERRIGHED Distributed File System), a distributed file system exploiting the disks attached to the computing nodes of a cluster [58], [30], [73].

In the context of Pierre Riteau's Master internship [70], we focused on reliable execution of applications that use file systems for data storage in a distributed environment. An efficient and portable file versioning framework was designed and implemented in the distributed file system kDFS. This framework can be used to snapshot file data when a process' volatile state is checkpointed and thereby make it possible to restart a process using files in a coherent way. Our experiments showed that the overhead caused by the file versioning framework was negligible. A replication model synchronized with the checkpoint mechanisms was also proposed. It provides stable storage in a distributed architecture. The synchronization allows to reduce network and disk I/O compared to a synchronous replication mechanism like RAID1 [60].

In the framework of Marco Obrovac's internship [69], we studied new scheduling strategies taking into consideration I/O usage. This work led to a theoretical proposal where the load values of the resources are prioritized, i.e. every resource enters in the load calculation with a specific weight. These weights are not fixed, so every cluster architect can adjust the scheduler policy to suite her needs.

We improved the existing prototype in terms of stability and performance. kDFS can now successfully execute Bonnie++ benchmark and the NTFS3G test suite (http://www.ntfs-3g.org/pjd-fstest.html) [73]. Using the Kargo tool, we deployed kDFS upon 48 nodes in the Grid'5000 platform. This experiment led to a 8.1TB storage space. We built virtual images of KERRIGHED including KDFS for QEMU and VmWare systems and made them available to the community for testing (http://www.kerrighed.org/wiki/index.php/KernelDevelKdFS). Finally, kDFS was ported on the kDDM standalone framework [30]. This demonstrates that kDFS only relies on the kDDM kernel level data sharing mechanisms for inter-node communications and that it can be used without KERRIGHED.

## *6.2.2. Grid operating systems*

**Participants:** Surbhi Chitre, Matthieu Fertré, Jérôme Gallard, Yvon Jégou, Sylvain Jeuland, Adrien Lèbre, Christine Morin, Pierre Riteau, Thomas Ropars, Oscar Sanchez.

### *6.2.2.1. Vigne, a system for large-scale, dynamic Grids*

**Participants:** Christine Morin, Thomas Ropars.

Our research aims at easing the execution of distributed computing applications on computational grids composed of a large number of geographically-distributed computing resources and characterized by a high churn. To ease the use of such dynamic, distributed systems, we propose to build a distributed operating system which provides a Single System Image (SSI), which is self-healing, and which can be tailored to the needs of the users. Such an operating system is composed of a set of distributed services, each of them providing a Single System Image for a specific type of resource, in a fault-tolerant way [29]. We are implementing this system on a research prototype called VIGNE.

In the framework of Rajib Kummar Nath's internship we worked on fault tolerance mechanisms to make critical Vigne services highly available (application execution manager, memory coherence manager) [68]. We have demonstrated that combining active replication and peer to peer techniques is an attractive solution to provide transparent high availability mechanisms for grid services. We have specified how to implement an active replication system based on consensus on top of a structured peer to peer network. The implementation of the proposed mechanisms is on-going.

To provide fault tolerance for large scale message passing applications, we proposed two protocols, O2P [42] and O2P-CF[38]. O2P, targeting clusters, is an optimistic message logging protocol that aims at reducing the amount of data piggybacked on the application messages for the need of the optimistic protocol [42]. Experiments conducted on the Grid'5000 platform have shown that the amount of data piggybacked on messages has a significant impact on application performance. O2P-CF combines O2P with a pessimistic message logging protocol. It targets applications executed on cluster federations. We are implementing these two protocols in the Open-MPI library and experimenting them in the context of Vigne system.

### *6.2.2.2. XtreemOS Grid operating system*

**Participants:** Surbhi Chitre, Matthieu Fertré, Jérôme Gallard, Yvon Jégou, Sylvain Jeuland, Adrien Lèbre, Christine Morin, Pierre Riteau, Thomas Ropars, Oscar Sanchez.

The objective of XTREEMOS project is to design, implement and promote a Linux-based Grid operating system providing a native virtual organization support [54]. The scientific coordination of the XTREEMOS European project is done by Ch. Morin, assisted by O. Sanchez, Technical Manager and Release Manager, and S. L'Hermitte, Project Office Assistant [61], [62].

In 2008, the research activities of the PARIS Project-Team were focused on the design and implementation of virtual organization and security services and of a Grid checkpointing service, on the study of virtualized environments and on the design and implementation of *LinuxSSI*, leveraging KERRIGHED SSI operating system for the cluster flavour of XTREEMOS system. Our work on *LinuxSSI* is described in Section 6.2.1.

A key feature of XTREEMOS is its support for *Virtual Organizations* (VO). We participated in the design of the XTREEMOS approach for VO management in close collaboration with ICT, STFC and TID [15], [41]. We contributed to the integration of the VO and security services with the other XtreemOS services: application execution management (AEM), XtreemFS Grid file system, overlays, LinuxSSI [67], [66]. The current XTREEMOS prototype does not properly address dynamic VO. VO are dynamic in a number of directions: addition and removal of users and resources, creation and deletion of subVOs, addition and removal of users and resources in subVOs, creation and deletion of attributes, addition and removal of user attributes, generation and invalidation of identity and attribute certificates, automatic VO generation when a new project is set up, VO federation. We have started to revisit XTREEMOS VO and security management services to support dynamic VO.

We contributed in close collaboration with the University of Duesseldorf and BSC to the design and implementation of XTREEMOS grid checkpointing service [32], [33], [59]. This service comprising of three layers is in charge of ensuring reliable application execution despite failures. It selects and applies the fault tolerance policy, manages data related to application life cycle, coordinates the fault tolerance actions for distributed applications spanning multiple Grid nodes and interacts with the AEM service which monitors the jobs and takes suspend, restart and migration decisions and with the XtreemFS Grid file system which is used to store checkpoints. We started to study the integration of the O2P-CF checkpointing protocol in the framework of XTREEMOS Grid checkpointing service.

We have also investigated the use of virtualization technologies in the context of XTREEMOS and studied scenarios of use of XTREEMOS in the area of Cloud computing [74]. We have initialized a study aiming at optimizing the performance of the migration of virtual clusters. Continuing the study on virtual clusters initialized in 2007 [28], we have proposed an extension of Goldberg model [55]. Goldberg classifies virtualization techniques in two models (Type-I and Type-II), which does not enable the classification of the latest virtualization technologies such as abstraction, emulation, partitioning and so on. Our extension formally defines these mechanisms by rigorously formalizing the following terms: virtualization, emulation, abstraction, partitioning, and identity. We also demonstrate that a single virtualization solution is generally composed of several layers of virtualization capabilities, depending on the granularity of the analysis.

In the framework of Oana Goga's internship [65], we worked on designing and implementing a framework allowing to (i) to deploy virtual machines upon the Grid'5000 platform, and (ii) to deploy Kerrighed upon a set of resources of Grid'5000 (physical or virtual nodes). Two software tools have been implemented: (i) *VMdeploy* to deploy virtual machines on the top of Grid'5000 and (ii) *Kargo* to deploy Kerrighed on top of physical/virtual nodes [56].

Oscar Sanchez, as release manager, coordinated the production and the testing of the first integrated version of XTREEMOS Grid operating system, publicly released in November 2008 [71], [75]. A permanent geographically distributed testbed made up of several computers provided by different XTREEMOS partners have been set up and used for testing and demonstrating XTREEMOS prototype.

## 6.3. Middleware systems for computational grids

### *6.3.1. Parallel CORBA objects and components*

**Keywords:** *CORBA*, *Grid*, *distributed component*, *distributed object*, *parallelism*.

**Participants:** Mathieu Kermarrec, Raúl López Lozano, Christian Pérez, Thierry Priol.

Distributed parallel object/component appears to be a key technology for programming distributed numerical simulation systems. It extends the well-known object/component-oriented model with a parallel execution model. Previous works such as PACO and GRIDCCM focused on communications between two parallel objects and components.

In 2008, we worked on hierarchical parallel component models as well as on the adaptation of hierarchical component models for Numamachines. With respect to hierarchical component models, we developed DHICO, an implementation of the DISCOGRID API. The originality of this model relies on the hierarchical management of partitioned data so as to let the runtime to optimize the communications (neighborhood as well global communications) while enabling a resource transparency for the user. The preliminary experiments on GRID'5000 showed that DHICO is able to outperform grid-enabled MPIimplementations while easing the developer task for real CEM and CFD applications. In order to evaluate component models on top of Numamachines, we developed FRIM, a multithread implementation of the FRACTAL component model. Experiments showed the inability of plain implementation to fully exploit Numamachines because of thread and memory placement. Hence, we have started studying how transfer placement and workflow information available at the assembly to the thread and memory sub-systems of the operating sytem.

On one hand, we plan to complete the undergoing experiments of DHICO on GRID'5000 to consolidate the results with real applications. It should lead to the actual resolution of large problem size. On the other hand, we will continue to study the adaptation of component models on Numamachines. We target to combine both effort within a common model and implementation.

### 6.3.2. Spatio-temporal skeleton software component models

**Keywords:** *CORBA Component Model (CCM)*, *Grid*, *software component*, *spatial description*, *temporal description*.

**Participants:** Julien Bigot, Hinde Lilia Bouziane, Christian PÃÂ©rez, Thierry Priol.

Software component models have succeeded in handling another level of the software complexity by dealing with system architecture. Moreover, we showed through STCM that they can be enhanced to also support temporal composition such as workflow or data flow.

In 2008, we start an implementation of STCM to show its feasibility and its benefits through real applications, and in particular a climatology application. Moreover, we deal with the next challenge that was to combine both advantages of component models and of skeleton models so as to enable more abstract and and generic compositions. In cooperation with the university of Pisa, we define STKM, an enhancement of STCM with algorithmic skeleton concepts. Programmers are therefore enabled to assembly applications specifying both temporal and spatial relations among components and instantiating predefined skeleton composite components to implement all those application parts that can be easily modeled with the available skeletons. We also explore the feasibility of such a model on top of SCA. Experimental results on kernel applications show the need and benefits of the high level of abstraction offered by STKM.

STKM seems to be a model rich enough to express applications independently of the resources. Hence, next steps are to study how to efficiently implement STKM and how to efficiently execute a STKM application on heterogeneous and dynamic resources.

### 6.3.3. Application deployment on computational Grids

**Participants:** Landry Breuil, Boris Daix, Christine Morin, Christian PÃÂ©rez, Thierry Priol.

The deployment of parallel, component-based applications is a critical issue in using computational Grids. It consists in selecting a number of nodes and in launching the application on them. We proposed a generic deployment model that aims to automatically deploy complex, static applications on Grids and ADAGE, an implementation of the model.

In 2008, we add a simple mechanism for handling dynamicity to ADAGE based on the concept of re-deployment. While it is a pseudo-dynamic mechanism, it turns out to be enough to validate other works like STCM/STKM and CORDAGE. Moreover, we finalize the specification of SAMURAAIE, a generic data model to abstract (dynamic) deployment of users' applications on resources. Not only, it abstracts instances of applications and of resources, but also of actions and events from them. SAMURAAIE systematically considers deployment as containers fitting contents. Therefore, it maintains information about containers, contents, and linkages. A running prototype shows the feasibility of the model and some its advantages. It clearly outperforms its predecessor GADE by being more expressive and generic.

In the future, we will study the integration of SAMURAAIE with SALOME, an open source integration platform for numerical simulation as well as the benefits that can be derived from SAMURAAIE with respect to the scheduling of structure applications on resources.

### *6.3.4. Adaptation for data management*

**Participants:** Françoise André, Mohamed Zouari.

The usage of context-aware data management in mobile environments has been investigated by Françoise André in collaboration with Mayté Segarra and Jean-Marie Gilliot from ENST Bretagne (Brest). A context-aware data replication and consistency system that adapts dynamically to changes in the environment has been proposed, based on the use of the DYNACO framework. This work has been supported by a contract (*ReCoDEM*) between ENST Bretagne and Orange Labs (previously known as France-Télécom R&D)

In the *ReCoDEM* project, the distributed aspects of the adaptation system has not been thoroughly investigated. Therefore, a new subject is launched since October 2007 (with M. Zouari as PhD student) to propose a generic distributed adaptation framework. This work will use data management in Grid and mobile environments as an illustrative application. Mayté Segarra from ENST Bretagne is co-adviser for the PhD thesis of M. Zouari.

### *6.3.5. Adaptation for fault tolerance*

**Participants:** Serge Guelton, Jean-Louis Pazat.

The use of adaptive framework as been studied to build dependable applications for Grids in the context of the *SafeScale* Project. Standard cases of attacks have been simulated and taken into account using the DYNACO framework and the *MPICH-V communication library* developed at LRI. The use of such a framework for a platform for ubiquitous computing has been studied in [100]. We have connected the DYNACO framework to the *Kaapi* environment developed at IMAG/LIG in order to be able to adapt the execution of task graphs to faulty environments. We have been able to demonstrate the control of task steeling and task cloning to generate challenges with this environment.

## 6.4. Large-scale data management for grids

### *6.4.1. Using transparent sharing of distributed data for databases*

**Keywords:** *DSM*, *Databases*, *Grid data-sharing*, *JXTA*, *peer-to-peer*.

**Participants:** Gabriel Antoniu, Luc Bougé, Landry Breuil, Marius Moldovan, Bogdan Nicolae.

Since 2003, we have been working on the concept of *data-sharing service* for Grid computing, that we defined as a compromise between two rather different kinds of data-sharing systems:

- *DSM systems*, which propose consistency models and protocols for efficient transparent management of *mutable data, on static, small-scaled configurations (tens of nodes)*;
- *P2P systems*, which have proven adequate for the management of *immutable data* on *highly dynamic, large-scale configurations (millions of nodes)*.

We illustrated this concept through the JUXMEM software platform, mainly developed by our group within the framework of Mathieu Jan's PhD thesis [88] and Sébastien Monnet's PhD thesis [96]. JUXMEM relies on the JXTA [77] generic peer-to-peer framework, which provides basic building blocks for user-defined, peer-to-peer services. L. Cudennec's PhD thesis is specifically devoted to improving the deployment of JXTA-based programs in the context of large-scale grid platforms such as GRID'5000.

In 2007, we explored the possibility of building a distributed database management system (DBMS) on top of JUXMEM, as a natural extension of previous approaches based on the distributed shared memory paradigm. The proposed approach consisted in providing the DBMS with a transparent, persistent and fault-tolerant access to the stored data, within a unstable, volatile and dynamic environment. The DBMS is thus alleviated from any concern regarding the dynamic behavior of the underlying nodes. This work has been done within the framework of the *RESPIRE* ANR project.

In 2008, this direction continued by exploring ways of integrating the concept of data-sharing service into an existing DBMS. During Marius Moldovan's Master internship, we experimented the interconnection between the BlobSeer prototype (mainly developed by Bogdan Nicolae) and the BerkeleyDB DBMS. In our prototype setting, BlobSeer serves as a block device for BerkeleyDB.

This work is continued within the framework of the PhD thesis of B. Nicolae, started in September 2007, with a focus on efficient storage and access to large data chunks.

### 6.4.2. Hierarchical Grid Storage based on the JuxMem Grid Data-Sharing Service and on the Gfarm Global File System

**Keywords:** *DSM*, *Grid file systems*, *JXTA*, *grid data sharing*.

**Participants:** Gabriel Antoniu, Loïc Cudennec, Landry Breuil, André Lage.

Within the framework of our collaboration with the Gfarm team from the Tsukuba University (Japan), we have defined a hybrid architecture which relies on both the JUXMEM grid-data sharing service and the *Gfarm* Grid file system (http://datafarm.apgrid.org/), and combines their specific benefits. The main idea was to allow applications to use JUXMEM's efficient memory-oriented API, while letting JUXMEM persistently store data on disk files by transparently making calls to *Gfarm* in the background. This work has been validated through a prototype that couples the *Gfarm* file system with the JUXMEM data-sharing service. The results have been published at Euro-Par 2008.

During André Lage's Master internship, we explored an alternative interaction between JuxMem and Gfarm: the goal was to use JuxMem to build a distributed, fault-tolerant metadata server that would replace Gfarm's centralized metadata server. The goal of sharing metadata thanks to JuxMem was only partially reached, for a subset of the metadata.

### 6.4.3. Toward transparent management of interactions between applications and resources

**Keywords:** *CoRDAGe*, *Computing grid*, *autonomic*, *co-deployment service*, *distributed application*, *dynamicity*, *re-deployment*.

**Participants:** Gabriel Antoniu, Luc Bougé, Landry Breuil, Loïc Cudennec, Christian Perez.

As a result of the past collaboration experience between the JUXMEM group with the *JXTA* and *Gfarm* teams, the need of launching complex distributed applications on large-scale testbeds led us to work on the automation of the deployment process. This has been done in close collaboration with the designers of the ADAGE deployment tool (Christian Pérez, Landry Breuil). Software *plugins* for ADAGE were designed to handle the deployment, in a static manner, for *JXTA*, JUXMEM and *Gfarm*-based applications. These *plugins* have been used in most of the experimentations involving such types of application.

Despite these efforts, the deployment of such applications remains quite painful for regular users, mostly because they are still in charge of the interactions with reservation and deployment tools. Therefore, the CORDAGE model has been proposed to significantly facilitate the deployment of applications by introducing transparent, on-the-fly resource reservations in response to possibly variable needs. This model has been presented at the STHEC workshop [26]. A prototype has been implemented and experimentations have been conducted within the GRID'5000 platform. These results are part of Loïc Cudennec thesis to be defended on January 15th, 2009.

### 6.4.4. Toward efficent versioning for large objects under heavy access concurrency

**Keywords:** *blob*, *fine grain access*, *heavy access concurrency*, *huge storage object*, *versioning*.

**Participants:** Gabriel Antoniu, Luc Bougé, Bogdan Nicolae.

Considering the problem of efficiently managing massive data in a large-scale distributed environment, we focus on data strings of sizes reaching the order of TB, shared and accessed at a fine grain by concurrent clients, as is the case with many data processing abstractions such as MapReduce. On each individual access, a segment of the blob, reaching the order of MB, is read or modified. Our goal is to provide the clients with an efficient fine-grain access and versioning interface to the blob that copes with heavy access concurrency, without the need to lock the blob itself. Our approach is illstrated through the BLOBSEER prototype. The overall design enables several features:

1. support massive, distributed data blobs management (in the order of TB);
2. support for a large number of blobs;
3. efficient atomic fine grain access to each blob (e.g., in the order of MB);
4. implicit versioning: updates to each blob add rather than replace data and generate a new virtual global view of the blob;
5. powerful concurrency management : high performance concurrent read/read, read/write, write/write access;
6. little overhead of storage space despite versioning.

BlobSeer, our prototype currently in development serves to contuct experimentation on the GRID'5000 platform. Preliminary results demonstrate good scalability and performance. These results have been published in [35] and [34].

## 6.5. Advanced programming models for the Grid

**Keywords:** *Chemical programming*, *autonomic systems*, *coordination*, *desktop grids*.

**Participants:** Jean-Pierre Banâtre, Manuel Caeiro Rodríguez, Xingwu Liu, Thierry Priol.

In our past work, we developed the $\gamma$-calculus and *HOCL*, a Higher-Order Chemical Language based on the $\gamma$-calculus. HOCL has been used to express workflow enactment and autonomic systems. This was the subject of Yann Radenac's PhD thesis, defended and published in April 2007 [99].

In 2008, we have investigated how to use HOCL to express dynamicity in scientific workflow. This is a joint research activity between SZTAKI (Hungary) and University of Vigo (Spain) that is being carried out within the CoreGRID network of Excellence. Dynamicity is a recurrent topic in traditional workflow systems. The need and feasibility to perform changes in workflow process instances while they are being executed has been a main (and to a long extend yet unsolved) challenge. We first analyzed dynamicity scenarios and requirements in scientific workflows. This work led to the publication of a CoreGRID technical report [53] within which we identified five general scenarios involving different dynamicity needs and illustrated byt concrete examples. These scenarios were used to identify a set of dynamicity requirements for scientific workflows support. After this initial phase, we introduced a proposal for a scientific workflow execution system based on HOCL. This proposal led to several publications [24], [25].

We also studied how to express service orchestration using the Chemical Metaphor. We have shown in [21] that chemical programming can be a good candidate for service programming, such as the composition and coordination of services.

A collaboration has been set up with ICT (China) under the framework of the UNCONV INRIA associated team and the FP6 IST EchoGRID project. We studied how HOCL can give a formal semantics to GSML (Grid Service Markup Language). GSML is a programming language that has been designed at ICT for grid end-users to overcome the programming hurdle and the high learning curve associated with Grid infrastructures that are complex distributed computing systems. This work led to a joint publication [31]. This paper describes the translation of GSML programs into HOCL allowing to give a precise definition of the concepts of GSML, especially sessions. The semantics also bridges the GSML and chemical computing paradigms. Another topics is being investigating to design an autonomic map&reduce framework for the execution of parallel applications over a Desktop Grid.

Finally, we started a collaboration with STFC (UK) under the framework of the CoreGRID Network of Excellence to study security engineering of distributed systems, such as Grids, when following the chemical-programming paradigm. We have analysed in [51], how to model secure systems using HOCL. Emphasis is on modularity, hence we advocate for the use of aspect-oriented techniques, where security is seen as a cross-cutting concern impacting the whole system. We show how HOCL can be used to model Virtual Organisations (VOs), exemplified by a VO system for the generation of digital products. We also develop security patterns for HOCL, including patterns for security properties such as authorisation, integrity and secure logs.

## 6.6. Experimental Grid infrastructures

**Participants:** Yvon Jégou, David Margery, Pascal Morillon, Cyril Rohr.

The deployment of the GRID'5000 site of Rennes was initiated in November 2003. The major steps for the platforms were

| *Date* | *# Nodes* | *# Procs* | *# Cores* | *Processor type* | *Node type* |
|---:|---:|---:|---:|---|---|
| Dec. 2003 | 66 | 132 | 132 | Intel Xeon IA32 | Dell PowerEdge 1750 |
| Oct. 2004 | 33 | 66 | 66 | IBM PowerPC | Apple Xserve G5 |
| Nov. 2004 | 66 | 132 | 132 | AMD Opteron 248 | Sun V20z |
| Dec. 2005 | 102 | 204 | 204 | AMD Opteron 246 | HP DL145 G2 |
| Nov. 2006 | 66 | 132 | 264 | Intel Xeon 5148LV | Dell PowerEdge 1950 |
| Sep. 2007 | 33 | 66 | 132 | Intel Xeon 5148LV | Dell PowerEdge 1950 |
| Dec. 2008 | 65 | 130 | 520 | Intel Xeon L5420 | Digitech Carri Systems CS-5393B |

As of the end of 2008, 196 nodes corresponding to 392 processors and 594 cores are active on our platform. An Additonal 65 nodes, 130 processors, 520 cores are to be installed by the end of the year.

The following interconnection equipments have been acquired since 2003:

| *Date* | *# Ports* | *Throughput* | *Uplink* | *Type* | *Model* |
|---:|---:|---:|---:|---|---|
| Dec. 2003 | 2x48 | 100Mb/s | 1 Gb/s | Ethernet | Foundry EdgeIron |
| Dec. 2004 | 8x24 | 1 Gb/s | 1 Gb/s | Ethernet | Cisco 3750 |
| Dec. 2005 | 66 | 10 Gb/s | | Infiniband | Mellanox/Voltaire |
| Feb 2006 | 320 | 1 Gb/s | 2x10 Gb/s | Ethernet | Cisco 6509 |
| Apr. 2006 | 33 | 10 Gb/s | | Myrinet | Myricom |
| Sep. 2007 | 64 | 10 Gb/s | 2x10 Gb/s | Myrinet | Myricom |

As of the end of 2008, the production network interconnects all nodes at 1 Gb/s using Ethernet technology, and provides connectivity to GRID'5000 sites through a 10 Gb/s optical link. A private Ethernet network, the management network interconnecting all nodes, is used for node management: monitoring, reboot, etc. It is exploited by the management software of the platform (*OAR*, *kadeploy*). Two local high-performance networks are available: an Infiniband network interconnecting 66 nodes at 10 Gb/s and a Myrinet 10G network interconnecting 97 nodes and the main Ethernet interconnect at 10 Gb/s.

The statistics show an average platform usage higher than 75%. The results provided by local users, mainly from the PARIS Project-Team, show that experimentations on our platform are cited in 6 PHD thesis, 6 book chapters or journal articles, 30 communications to international conferences and in 11 communications to national conferences.

2008 is a major milestone for this experimental infrastructure as its further development was planned and accepted as ALADDIN-G5K, starting in June 2008. The work in Rennes has started to design an API to access the plateform and the corresponding service oriented architecture based on the REST architectural style.

# 7. Contracts and Grants with Industry

## 7.1. EDF Contract 2 (2006-2008)

**Participants:** Boris Daix, Christine Morin, Christian Pérez.

The collaboration with EDF R&D aims at improving the dynamic deployment of scientific code-coupling applications on cluster federations, taking into account their execution constraints. In 2008, we finalized and implemented a deployment model for applications and resources that both have properties of parallelism/distribution, heterogeneity, and dynamicity.

## 7.2. Sun Microsystems (2005-2008)

**Participants:** Gabriel Antoniu, Luc Bougé, Loïc Cudennec, Diana Moise.

The work addresses techniques to optimize the use of the JXTA P2P library on Grid infrastructures. The main achievements in 2008 are related to deployment issues. JXTA was our first middleware used as a test-case for the CoRDAGe co-deployment and re-deployment tool developed within Loïc Cudennec's thesis. Hybrid deployment topologies have been experimented by CoRDAGe, where our JXTA-based JuxMem service is coupled and deployed together with the Gfarm grid file system.

# 8. Other Grants and Activities

## 8.1. Regional grants

### *8.1.1. Brittany Council*

*8.1.1.1. PhD grants*

The Brittany Regional Council provides half of the financial support for the PhD theses of the following students: Loïc Cudennec, Mohamed Zouari, Diana Moise. This support amounts to a total of 3 times 14,000 Euros/year, that is, 42,000 Euros/year.

*8.1.1.2. 5000NET Project*

**Participants:** Yvon Jégou, David Margery, Pascal Morillon.

The 5000NET Project is funded by the Brittany Regional Council until July 2007. Its aim was to provide financial support for the integration of high-speed interconnection networking equipments in our Grid'5000 platform.

*8.1.1.3. Support to XtreemOS Project Management*

**Participants:** Sandrine L'Hermitte, Christine Morin.

The Brittany Regional Council provides a financial support for the management of the XtreemOS IP project. This supports amounts to a total of 30,000 Euros. It contributes to funding S. L'Hermitte, who assists the scientific coordinator and ensures the clerical management of the XtreemOS project office and of all XtreemOS management bodies.

## 8.2. National grants

### *8.2.1. ANR WP: ANR White Program*

*8.2.1.1. ANR WP AutoChem Project*

**Participants:** Jean-Pierre Banâtre, Thierry Priol.

This project aims at investigating and exploring an unconventional approach, based on chemical computing, to program complex computing infrastructures, such as Grids and real-time deeply-embedded systems. It is a 3-year project which started in December 2007.

### *8.2.2. ANR CI: ANR Program on High-Performance Computing and Simulation*

*8.2.2.1. ANR CI DISC Project*

**Participants:** Raúl López Lozano, Christian Pérez, Thierry Priol.

It aims at studying and promoting a new paradigm for programming non-embarrassingly parallel scientific computing applications on distributed, heterogeneous, computing platforms. The *DISC* project concentrates its activities on numerical kernels and related issues that are of interest to a large variety of application contexts. It is a 3-year project which started in January 2006.

*8.2.2.2. ANR CI LEGO Project*

**Participants:** Gabriel Antoniu, Landry Breuil, Hinde Lilia Bouziane, Loïc Cudennec, Christian Pérez.

The aim of this project is to provide algorithmic and software solutions for large-scale architectures, focusing on performance issues. The software component approach provides a flexible programming model where resource management issues and performance optimizations are handled by the implementation. The project addresses topics in programming models, communication models, and scheduling. The results are validated on three applications: an ocean-atmosphere numerical simulation, a cosmology simulation, and a sparse-matrix solver. It is a 3-year project which started in January 2006. Project site: http://graal.ens-lyon.fr/LEGO/.

*8.2.2.3. ANR CI NUMASIS Project*

**Participants:** Christian Pérez, Mathieu Kermarrec.

It deals with recent NUMA multiprocessor machines with a deep hierarchy. In order to efficiently exploit it, the project aims at evaluating the features of current systems, at proposing and implementing new mechanisms for process, data and communication management. The target applications come from the seismology field that appear representative of current needs in scientific computing. It is a 3-year project which started in January 2006. Project site: http://numasis.gforge.inria.fr/.

### *8.2.3. ANR MD: ANR Program on Data Masses and Ambient Knowledge*

*8.2.3.1. ANR MD RESPIRE Project*

**Participants:** Gabriel Antoniu, Luc Bougé, Landry Breuil, Loïc Cudennec.

The RESPIRE Project of the ANR MD program aims at providing a peer-to-peer (P2P) environment for advanced data management applications. It gathers research teams from the “databases” area and from the “distributed systems” area, The RESPIRE Project is based on the JXTA infrastructure which provides a complete abstraction from the underlying P2P network organization (DHT, flooding, super-peer). The main actions that will be developed in the project are resource access and sharing, managing logical cluster, handling replication and automated deployment of the environment. The project started in January 2006 for 3 years. Gabriel Antoniu is the local correspondent of RESPIRE for the PARIS Project-Team. Project site: http://respire.lip6.fr/.

### *8.2.4. ANR SI: ANR Program on Security and Informatics*

*8.2.4.1. ANR SI SafeScale*

**Participants:** Jean-Louis Pazat, Françoise André.

The *SafeScale* Project is concerned with security and safety in global ambient computing systems, e.g., computational grids. We have used our adaptive techniques (e.g., DYNACO) to implement application reactions to use-case attacks on an experiment on GRID'5000. We have connected DYNACO to the *Kaapi* task execution environment to study adaptation with work-stealing.

### *8.2.5. INRIA ADT*

#### *8.2.5.1. ALADDIN-G5K*

**Participants:** Yvon Jégou, David Margery, Pascal Morillon, Thierry Priol, Cyril Rohr.

The PARIS project-team coordinated in 2008 a follow-up proposal to the Grid'5000 project of the ACI GRID. This proposal aims at the construction of a scientific instrument for experiments on large-scale parallel and distributed systems, building on the Grid'5000 platform. This proposal was accepted and officially started June 1st. It structures INRIA's leadership role as the institute is present in 8 if the 9 Grid'5000 sites distributed across France.

Thierry Priol is the director of this ADT, Franck Cappello the scientific director, Frédéric Desprez the deputy scientific director and David Margery the technical director. An executive committee, where each of the 10 project-teams supporting Grid'5000 in the 8 research centers is represented, meets every month. It gives recommendations to the directors on scientific animation, access policy to the instrument as well as for the hardware and software development according to the resources devoted to this ADT. Yvon Jegou represents the PARIS project-team in this executive committee.

The technical team is now composed of 12 engineers, of which 3 are from the PARIS project-team (David Margery, Pascal Morillon, Cyril Rohr). This technical team is structured in a sysadmin team, managing the instrument, and a developpment team building the tools to build, execute and analyze experiments.

In 2008, ALADDIN-G5K has enabled the coordinated acquisition of hardware on 2 sites, Rennes and Nancy, therefore reducing unwanted hardware heterogeneity. Moreover, it has enabled the transformation of a team of local system administrators loosely coordinated into a coherent team of engineers competent on all sites composing Grid'5000 with a clear leadership. More than 600 account are open, leading to a great number of experiments and publications on the topic of large scale parallel and distributed systems.

30 project-teams not including in the 10 mentioned above are using or have shown interest in using such an instrument.

### *8.2.6. ANR TLOG: ANR Program on Software Technologies*

#### *8.2.6.1. ANR TLOG NeuroLog Project*

**Participant:** Yvon Jégou.

The *NeuroLog* consortium (*Software technologies for integration of process, data and knowledge in medical imaging*) is targeting software technologies in medical domains for large scale management of data, knowledge and computation: management and access of partly structured data, heterogeneous and distributed in an open environment; access control and protection of private medical data; control of workflows implied in complex computing process on grid infrastructures; extraction and quantification of relevant parameters for different pathologies.

## 8.3. European grants

### *8.3.1. CoreGRID NoE Project*

**Participants:** Françoise André, Gabriel Antoniu, Hinde Lilia Bouziane, Christian Pérez, Thierry Priol, Olivia Vasselin.

Thierry Priol is the Scientific Coordinator of a *Network of Excellence* proposal, called COREGRID, in the area of Grid and Peer-to-Peer (P2P). The COREGRID network started on September 1, 2004. As many as 46 partners, mostly from 19 European countries are involved. The COREGRID Network of Excellence aims at building a European-wide research laboratory that will achieve scientific and technological excellence in the domain of large-scale distributed, Grid, and Peer-to-Peer computing. The research programme is structured around six complementary research areas, i.e., work packages that have been selected on the basis of their strategic importance, their research challenges, and the European expertise in these areas to develop next generation Grids: *Knowledge and Data Management*, *Programming Models*, *Architectural Issues: Scalability, Dependability, Adaptability*, *Grid information, Resource and Workflow Monitoring Services*, *Resource Management and Scheduling*, *Grid Systems, Tools and Environments* INRIA is managing the network in collaboration with the ERCIM office. ERCIM is in charge of administrative and financial management. Th. Priol is the *Scientific Coordinator* (SCO), leading the network with respect to the scientific aspects, and looking after its overall management. He is assisted by Olivia Vasselin. The main tasks of the SCO during this fourth year were coordinating and monitoring the activities related to the scientific and technical workpackages, coordinating the COREGRID *Scientific Advisory Board*, performing the first ranking of partners activity, coordinating the preparation of the second *Joint Program of Activities* and providing the first internal assessment of the network. In addition, the SCO participated in dissemination tasks by giving presentations, contributing to the COREGRID Newsletters, etc. The SCO was also responsible to initiate the sustainability of the Network after the completion of the EU contract. COREGRID has ended in August 2008 and went through successfully its last review that was held on November 2008.

Christian Pérez is responsible for the COREGRID contract within INRIA. He is responsible for managing the four INRIA Project-Teams (PARIS, *Grand-Large*, *OASIS* and *SARDES*) with respect to periodic reporting, etc. His main tasks were to represent INRIA in the COREGRID Members General Assembly meetings and votes.

### *8.3.2. EchoGRID Specific Support Action*

**Participants:** Jean-Pierre Banâtre, Thierry Priol.

Th. Priol is the *Scientific Coordinator* (SCO) of the ECHOGRID Specific Support Action that is funded under the FP6 IST Work Programme. This action aims to foster collaboration in Grid research and technologies by defining short-, mid-, and long-term vision in the field. It is a 2-year project which started in February 2007. It involves 10 partners from 4 European countries plus China. Th. Priol participated in 2008 to two workshops, respectivelly in June (Athens) and October (Beijing), and one conference that was held in October at Shenzhen. He was in charge of the organization of a panel session dedicated to Cloud computing and virtualization technologies. In the context of this action, Yann Radennac has spent a 12-month post-doc grant ending in September 2008. It worked at ICT from the Chinese Academia of Science. His research are devoted to advanced programming models for Grids.

### *8.3.3. XtreemOS IP Project*

**Participants:** Surbhi Chitre, Matthieu Fertré, Jérôme Gallard, Yvon Jégou, Sylvain Jeuland, Adrien Lèbre, Sandrine L'Hermitte, Christine Morin, Thierry Priol, Thomas Ropars, Oscar Sanchez.

Ch. Morin is the *Scientific Coordinator* (SCO) of the XTREEMOS Integrated Project (IP) that addresses Strategic Objective 2.5.4 *Advanced Grid Technologies, Systems and Services*, Focus 3 on *Network-centric Grid Operating Systems* as described in the IST 2006 Work Programme. XTREEMOS involves 19 academic and industrial partners from 7 European countries plus China. The XTREEMOS project aims at the design, implementation, evaluation and distribution of an open source Grid operating system with native support for virtual organizations and capable of running on a wide range of underlying platforms, from clusters to mobiles. The approach we propose in this project is to investigate the construction of a new Grid OS, XTREEMOS, based on the existing general-purpose OS Linux [5]. *Y. Jégou* leads the WP4.3 Work-Package, aiming at setting up XTREEMOS testbeds. The GRID'5000 experimental grid platform will be used as a testbed by XTREEMOS partners. Ch. Morin leads WP1.1, Project management, WP2.1, Virtual Organization support in Linux, WP2.2 Federation management and WP5.3, Collaboration with other IST Grid-related projects. We hosted a one-week general technical meeting for the whole XTREEMOS consortium at INRIA Rennes in January 2008.

### *8.3.4. S-Cube NoE Project*

**Participants:** Françoise André, Jean-Louis Pazat, André Lage, Guillaume Gauvrit.

S-Cube is a European Network of Excellence in Software Services and Systems. Its goal is to establish an integrated, multidisciplinary, vibrant research community. This will enable Europe to lead the software-services revolution, thereby helping shape the software-service based Internet which is the backbone of our future interactive society. We are involved in the infrastructure workpackage which aims at defining specification of the infrastructure for building adaptable and self-* service based applications.

# 9. Dissemination

## 9.1. Community animation

### *9.1.1. Leaderships, Steering Committees and community service*

European COREGRID IST-FP6 Network of Excellence. Th. Priol is the *Scientific Coordinator* of the COREGRID Network of Excellence (http://www.coregrid.net/).

European ECHOGRID IST-FP6 Supported Action. Th. Priol is the *Scientific Coordinator* of the ECHOGRID project (http://echogrid.ercim.org/).

European XTREEMOS IST-FP6 Integrated Project. Ch. Morin is the *Scientific Coordinator* of the XTREEMOS Integrated Project (http://www.xtreemos.eu/). Y. Jégou is a member of XTREEMOS Executive Committee. Th. Priol is a member of XTREEMOS Scientific Advisory Committee. J.-P. Banâtre is the INRIA representative in the XTREEMOS Governing Board.

Euro-Par Conference Series. L. Bougé serves as the Vice-Chair of the *Steering Committee* of the *Euro-Par* annual conference series on parallel computing.

CNRS, GDR ASR. J.-L. Pazat is co-director of the GSP working group on Grids, Systems and Parallelism of the CNRS Research Co-operative Federation (*Groupement de recherche*, GDR) ASR (*Architectures, Systems and Networks*). F. André serves as the coordinator of the ADAPT action (*Dynamic Adaptation*) of the GSP working group.

*Agrégation* of Mathematics. L. Bougé serves as one of the Vice-Chairs of the National Selection Committee for High-School Mathematics Teachers.

ALADDIN-G5K Th. Priol is the director of ALADDIN-G5K and D. Margery is the technical director of ALADDIN-G5K.

### *9.1.2. Editorial boards, direction of program committees*

L. Bougé is a member of the *Editorial Advisory Board* of the *Scientific Programming* Journal.

J.-L. Pazat serves as the Chair of the Organizing Committee of the RenPar, CFSE and Sympa federated conference series. He is the chairman of the Steering Committee of RenPar http://www.renpar.org/).

Th. Priol is a member of the Editorial Board of the *Parallel Computing* Journal and of the *International Journal of Web Services Research*. He was Chair of the Program Committee of the *2008 CoreGRID Symposium*, *2008 CCGRID conference*, *2008 ServiceWave conference*.

### *9.1.3. Program Committees*

G. Antoniu served in the Program Committees for the following conferences: MSOP2P 2008, DaMap 2008, HIPS 2008, CCGrid 2008, HPDGrid 2008, Euro-Par 2008, Cluster 2008.

L. Bougé served in the Program Committee for the following conferences: NPC 2008.

Ch. Morin served in the Program Committees of the following conferences: RenPar 18, HPCVirt 2008, ICA3PP 2008, LASCO 2008, SDMAS 2008, SNAPI 2008, HiperIO 2008, ICPADS 2008, ISPA 2008.

J.-L. Pazat served in the Program Committees of the following conferences: GPC 2008, RenPar 18.

Ch. Pérez served in the Program Committees of the following conferences: RenPar 18, ICNS 2008, AINA 2008, CCGrid 2008.

Th. Priol served in the Program Committees of the following conferences: CCGRID 2008, DAPSYS 2008, GCC 2008, HPCVirt 2008,HPDataGrid 2008, HPDC 2008, ICCS 2008, ICSOC 2008, ICWS 2008, SC 2008, ServiceWave 2008, VecPar 2008, WI 2008.

### *9.1.4. Evaluation committees, consulting*

L. Bougé served as a member of the Selection Committee for the *Gilles Kahn PhD Thesis Award 2008*. He served in the Mid-Term Evaluation Committee of the ANR MD, and of the ANR GIGC. He served in the AERES Evaluation Committee for the Research in Mathematics and System Sciences at ENSMP in February 2008 and for the LIFL Laboratory in December 2008.

Ch. Morin acted as a referee for the Foreign PhD Committee of Francesc Guim Bernat (UPC) and of Carlo Bertolli (Pisa University).

Th. Priol was a member of an International Committee appointed by the FCT, Portugal, to evaluate the research units in Electrical Engineering and Computer Science (EECS) in Portugal. He was also an evaluator for RPF in Cyprus. He was expert for the French-Israelo multicomputer initiative. He also acts as an expert to review FP6 and FP7 projects for the European Commission as well as ERC proposals.

## 9.2. Academic teaching

Only the teaching contributions of project-team members on non-teaching positions are mentioned below.

G. Antoniu is teaching part of the *Operating Systems* Module at *IUP 2 MIAGE*, IFSIC. He has given lectures on peer-to-peer systems within the *High Performance Computing on Clusters and Grids* Module and within the *Peer-to-Peer Systems* Module of the Master Program, UNIVERSITY RENNES 1, and within the *Distributed Systems* Module taught for the final year engineering students of INSA Rennes.

B. Daix have given in Spring 2008 a complete course on "GNU/Linux specialized for sight-challanged students in scientific domain" at INSA Toulouse (30 hours).

A. Lèbre taught part of the Operating Systems Module at Master 1 MIAGE, IFSIC. He also gave lectures on high-performance I/O in clusters within the *Distributed Systems: from networks to Grids* Module of the Master Program, UNIVERSITY RENNES 1. Finally, he gave lectures on Computer Architecture and several tutorials on JAVA programming at the Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAE), Rennes.

Ch. Morin is responsible for a graduate teaching Module *Distributed Systems: from networks to Grids* of the Master Program in Computer Science, UNIVERSITY RENNES 1. Within this module, she gave lectures on cluster and Grid computing. She gave a lecture on cluster single system image operating systems within the *Parallelism* Module of the 3rd-year students of INT of Évry.

Ch. Pérez gave lectures to 5th-year students of INSA of Rennes on CORBA and CCM within the course *Objects and components for distributed programming*. He also gave lectures to 5th-year students of Polytech Nantes on CORBA and CCM within the course *Objects and components for distributed programming*.

Th. Priol gave lectures on Distributed Shared Memory and Grid Programming within the *Distributed Systems: from Network to Grids* Module of the Master Program, UNIVERSITY RENNES 1.

T. Ropars taught C programming Language for 3rd-year students of INSA Rennes (practical work). He also taught Parallel Programming for Master students of the University of Rennes 1 within the Operating System Module (practical work).

## 9.3. Conferences, seminars, and invitations

Only the events not listed elsewhere are listed below.

Y. Jégou presented demonstrations of the XTREEMOS system at Internet of Services 2008 collaboration meeting organized by the European Commission (September 2008) and ICT 2008 (November 2008).

B. Daix gave a talk entitle *Enrichir la plate-forme de simulation numérique SALOME par la délégation du déploiement de ses conteneurs de composants à ANGE* at the PhD student day, SINETICS department, EDF R&D, Clmart, France, December 2008.

S. Jeuland gave a talk on XTREEMOS and presented a demonstration of XTREEMOS at LaBri, Bordeaux, France, October, 2008.

A. Lèbre gave a talk entitled *kDFS, Toward an Integrated Cluster File System* at the second Kerrighed Summit, Paris, France, February 2008. He gave a talk on *kDFS et systèmes de fichiers pour grilles* at Ecole des Mines de Nantes (EMN), Nantes, France, April, 2008. He presented a demonstration of the LinuxSSI during the XTREEMOS project formal review, Brussels, Belgium, July 2008.

Ch. Morin gave a talk entitled *XtreemOS plans for the Kerrighed project* at the Second Kerrighed Summit, Paris, February 2008. She gave an invited talk on *System software for peta/exascale machines*, at the SOS12 workshop, Waldhaus, Switzerland, March 2008. She gave a keynote talk on *Beyond Grid middleware: XtreemOS Vision* at the 8th IEEE International Symposium on Cluster Computing and the Grid (CC-GRID 2008) conference, Lyon, France, May 2008. She gave an invited talk on *XtreemOS: a Linux-based Operating System for Large Scale Dynamic Grids*, at the *First Usenix workshop on large-scale computing* (LaSCo'08), Boston, USA, June 2008. She gave a talk on *XtreemOS: a Linux-based Operating System for Large Scale Dynamic Grids*, at the first workshop orghanized in the framework of the UNCONV associated team in Beijing, China in June 2008. She gave a talk on *Overview of XtreemOS*, at the XtreemOS - Grid4All joint meeting, Paris, July 2008. She was invited to gave a talk on *Clouds: A New Playground for XtreemOS Grid Operating System* at the international workshop on Clusters and Computational Grids for Scientific Computing (CCGSC 2008), Flat Rock, USA, September 2008. She gave a talk on *Kerrighed : une étude de cas de développement et valorisation de logiciels en environnement de recherche* and one on *Kerrighed : un système dâexploitation SSI pour grappes* during the school on *dEveloppemeNt et la ValOrisation des Logiciels en environnement de recherche (ENVOL)* organized by CNRS/PLUME, Annecy, France, October 2008. She was invited to participate in the round table on International Collaboration at the Internet of Services 2008 collaboration meeting organized by Unit D3 on Service and Software of the European Commission. She gave a lecture on *XtreemOS Linux-based Grid Operating System* for a MasterClass organized at Vrije Universiteit Amsterdam in the framework of Grid Forum Netherlands, Amsterdam, The Netherlands, October 2008.

Th. Priol gave a keynote presentation entitled *From Conventional to Unconventional Grid Programming* at the 7th International Conference on Grid and Cooperative Computing (GCC 2008) and Second EchoGRID Conference that was held in Shenzhen on October 2008. He was invited to gave a talk on *Unconventional Grid Programming* at the international workshop on Clusters and Computational Grids for Scientific Computing (CCGSC 2008), Flat Rock, USA, September 2008.

## 9.4. Administrative responsibilities

F. André is the vice-chair of the Administrative Committee of IFSIC, the Computer Science Teaching Department of UNIVERSITY RENNES 1.

G. Antoniu serves as Scientific Correspondent for International Relations of INRIA's Research Center of Rennes - Bretagne Atlantique (since September 2008).

L. Bougé chairs the Computer Science and Telecommunication Department (*Département Informatique et Télécommunications, DIT*) of the Brittany Extension of ENS CACHAN. He leads the Master

Program in Computer Science at the Brittany Extension of ENS CACHAN. He serves as the Vice-Chairman of the Selection Committee (*Commission de spécialistes d'Établissement*, CSE) for Computer Science at ENS CACHAN, and as an external deputy-member of the Computer Science CSE at UNIVERSITY RENNES 1.

J.-L. Pazat leads the Master Program of the 5th year of Computer Science at INSA of Rennes. He is responsible for a teaching module on Parallel Processing for engineers at INSA of Rennes.

T. Priol is member of the *Bureau du Comité des projets*.

## 9.5. Miscellaneous

F. André is a member of the Selection Committee (Commission de spécialistes, CSE) of IFSIC (Computer Science department of University of Rennes1), of the Computer Science department of INSA of Rennes and of the Computer Science group of University of Rennes 2.

L. Bougé is a member of the Project-Team Committee of IRISA (*Comité des projets*), standing for the ENS CACHAN partner.

Ch. Morin has served as an external deputy member in the Selection Committee (*Commission de spécialistes*, CSE) for the Computer Science department of INSA Rennes. She served as a member of the INRIA Rennes local committee for attributing INRIA/CORDIS Ph.D. grants in 2008. She has been the patron of the OSCAR d'Ille-et-Vilaine 2008 ceremony awarding local exemplary companies. She also served as a member of the selection committee for the Dream Orange 2008 competition awarding students for their projects on products and services exploiting optical fibers.

J.-L. Pazat is a member of the Computer Science Department committee. He is the local coordinator for the international exchange of students at the computer science department of INSA. He serves as the Chairman of the Selection Committee (*Commission de spécialistes d'Établissement*, CSE) for Computer Science at INSA Rennes

Ch. Pérez is a member of the IRISA Laboratory Committee (*Conseil de laboratoire*).

T. Priol is a member of the Project-Team Committee of IRISA (*Comité des projets*).

# 10. Bibliography

## Major publications by the team in recent years

[1] M. ALDINUCCI, F. ANDRÉ, J. BUISSON, S. CAMPA, M. COPPOLA, M. DANELUTTO, C. ZOCCOLO. *Parallel program/component adaptivity management*, in "ParCo 2005, Málaga, Spain", 13-16 September 2005.

[2] F. ANDRÉ, M. LE FUR, Y. MAHÉO, J.-L. PAZAT. *The Pandore Data Parallel Compiler and its Portable Runtime*, in "High-Performance Computing and Networking (HPCN Europe 1995), Milan, Italy", Lecture Notes in Computer Science, vol. 919, Springer Verlag, May 1995, p. 176–183.

[3] G. ANTONIU, L. BOUGÉ. *DSM-PM2: A portable implementation platform for multithreaded DSM consistency protocols*, in "Proc. 6th International Workshop on High-Level Parallel Programming Models and Supportive Environments (HIPS '01), San Francisco", Lect. Notes in Comp. Science, Available as INRIA Research Report RR-4108, vol. 2026, Springer-Verlag, Held in conjunction with IPDPS 2001. IEEE TCPP, April 2001, p. 55–70, http://hal.inria.fr/inria-00072523.

[4] J.-P. BANÂTRE, D. LE MÉTAYER. *Programming by Multiset Transformation*, in "Communications of the ACM", vol. 36, n$^{o}$ 1, January 1993, p. 98–111.

[5] CONSORTIUM, XTREEMOS. *Annex 1 - Description of Work*, XtreemOS Integrated Project, IST-033576, April 2006, Contract funded by the European Commission.

[6] A. DENIS, C. PÉREZ, T. PRIOL. *PadicoTM: An Open Integration Framework for Communication Middleware and Runtimes*, in "IEEE Intl. Symposium on Cluster Computing and the Grid (CCGrid2002), Berlin, Germany", Available as INRIA Reserach Report RR-4554, IEEE Computer Society, May 2002, p. 144–151, http://hal.inria.fr/inria-00072034.

[7] A.-M. KERMARREC, C. MORIN, M. BANÂTRE. *Design, Implementation and Evaluation of ICARE*, in "Software Practice and Experience", n$^{o}$ 9, 1998, p. 981–1010.

[8] T. KIELMANN, P. HATCHER, L. BOUGÉ, H. BAL. *Enabling Java for High-Performance Computing: Exploiting Distributed Shared Memory and Remote Method Invocation*, in "Communications of the ACM", Special issue on Java for High Performance Computing, vol. 44, n$^{o}$ 10, October 2001, p. 110–117.

[9] Z. LAHJOMRI, T. PRIOL. *KOAN: A Shared Virtual Memory for iPSC/2 Hypercube*, in "Proc. of the 2nd Joint Int'l Conf. on Vector and Parallel Processing (CONPAR'92)", Lecture Notes in Computer Science, vol. 634, Springer Verlag, September 1992, p. 441–452, http://hal.inria.fr/inria-00074927.

[10] B. NICOLAE, G. ANTONIU, L. BOUGÉ. *Distributed Management of Massive Data: an Efficient Fine-Grain Data Access Scheme*, in "International Workshop on High-Performance Data Management in Grid Environments (HPDGrid 2008), Toulouse, France", VECPAR, June 2008, http://hal.inria.fr/inria-00323248/en/.

[11] B. NICOLAE, G. ANTONIU, L. BOUGÉ. *Enabling Lock-Free Concurrent Fine-Grain Access to Massive Distributed Data: Application to Supernovae Detection*, in "Poster Session - IEEE Cluster 2008, Tsukuba, Japan", Cluster/Grid, September 2008, http://hal.inria.fr/inria-00329698/en/.

[12] T. PRIOL. *Efficient support of MPI-based parallel codes within a CORBA-based software infrastructure*, in "Response to the Aggregated Computing RFI from the OMG, Document orbos/99-07-10", July 1999.

[13] Y. RADENAC. *Programmation "chimique" d'ordre supérieur*, Thèse de doctorat, Université de Rennes 1, April 2007.

## Year Publications

### Doctoral Dissertations and Habilitation Theses

[14] H. BOUZIANE. *De l'abstraction des modèles de composants logiciels pour la programmation d'applications scientifiques distribuées*, Thèse de doctorat, Université de Rennes 1, IRISA/INRIA, Rennes, France, February 2008, https://www.irisa.fr/paris/bibadmin/uploads/pdf/These.pdf.

### Articles in International Peer-Reviewed Journal

[15] Y. COPPOLA, C. MORIN, B. MATTHEWS, L. P. PRIETO, O. D. SÁNCHEZ, E. YANG, H. YU. *Virtual Organization Support within a Grid-wide Operating System*, in "IEEE Internet Computing", vol. 12, n$^{o}$ 2, March 2008, p. 20-28.

### Articles in National Peer-Reviewed Journal

[16] A. LEBRE, G. HUARD, Y. DENNEULIN, P. SOWA. *Optimisation des E/S avec QoS dans les environnements multi-applicatif distribués*, in "Technique et Science Informatiques", vol. 27, n$^{o}$ 3-4, 2008, 265, 291.

## International Peer-Reviewed Conference/Proceedings

[17] M. ALDINUCCI, H. BOUZIANE, M. DANELUTTO, C. PÉREZ. *Towards Software Component Assembly Language Enhanced with Workflows and Skeletons*, in "Joint Workshop on Component-Based High Performance Computing and Component-Based Software Engineering and Software Architecture (CBHPC/COMPARCH 2008)", 14-17 October 2008, https://www.irisa.fr/paris/bibadmin/uploads/pdf/0.88095200%201225115375_paper.pdf.

[18] G. ANTONIU, E. CARON, F. DESPREZ, A. FÈVRE, M. JAN. *Towards a Transparent Data Access Model for the GridRPC Paradigm*, in "Proc. of the 13th International Conference on High Performance Computing (HiPC 2007), Goa, India", Lect. Notes in Comp. Science, n$^{o}$ 4873, Springer-Verlag, January 2008.

[19] G. ANTONIU, L. CUDENNEC, M. GHAREEB, O. TATEBE. *Building Hierarchical Grid Storage Using the GFarm Global File System and the JuxMem Grid Data-Sharing Service*, in "Proceedings of the 14th International Euro-Par Conference on Parallel Processing (Euro-Par'08), Las Palmas de Gran Canaria, Spain", Lect. Notes in Comp. Science, vol. 5168, Springer-Verlag, 2008, p. 456-465.

[20] G. ANTONIU, M. JAN, D. NOBLET. *A practical example of convergence of P2P and grid computing: an evaluation of JXTA's communication performance on grid networking infrastructures*, in "Proc. 3rd Int. Workshop on Java for Parallel and Distributed Computing (JavaPDC'08), Miami", Held in conjunction with IPDPS 2008, April 2008, 104.

[21] J.-P. BANÂTRE, T. PRIOL, Y. RADENAC. *Service Orchestration Using the Chemical Metaphor*, in "Software Technologies for Embedded and Ubiquitous Systems", SPRINGER (editor), Lecture Notes in Computer Science, vol. 5287, October 2008, p. 79-89.

[22] J. BIGOT, H. BOUZIANE, C. PÉREZ, T. PRIOL. *On Abstractions of Software Component Models for Scientific Applications*, in "Proc. of the Abstractions for Distributed Systems workshop, Las Palmas, Gran Canaria, Spain", 2008.

[23] H. BOUZIANE, C. PÉREZ, T. PRIOL. *A Software Component Model with Spatial and Temporal Compositions for Grid infrastructures*, in "Proc. 14th Intl. Euro-Par Conference (Euro-Par 08),, Las Palmas de Gran Canaria, Spain", vol. 5168, Springer Berlin / Heidelberg, 26-29 August 2008, p. 698-708, https://www.irisa.fr/paris/bibadmin/uploads/pdf/0.25529500%201220430923_paper.pdf.

[24] M. CAEIRO-RODRIGUEZ, Z. NÉMETH, T. PRIOL. *A Chemical Workflow Engine to Support Scientific Workflows with Dynamicity Support*, in "Proceedings of the 3rd Workshop on Workflows in Support of Large-Scale Science", to appear, IEEE, November 2008.

[25] M. CAEIRO-RODRIGUEZ, T. PRIOL, Z. NÉMETH. *A Proposal to Support the Execution of Scientific Workflows based on a Higher Order Chemical Language*, in "Proceedings of 9th International Workshop on State-of-the-Art in Scientific and Parallel Computing", to appear, May 2008.

[26] L. CUDENNEC, G. ANTONIU, L. BOUGÉ. *CoRDAGe: towards transparent management of interactions between applications and ressources*, in "International Workshop on Scalable Tools for High-End Computing (STHEC 2008), Kos, Grèce", 2008, p. 13-24, http://hal.inria.fr/inria-00288339/en/.

[27] N. CURRLE-LINDE, C. PÉREZ, M. RESCH, M. COPPOLA. *Component Measurable Values and Services: A Technology for the Conclusion of Resource Transactions*, in "Proc. of the CoreGRID Integration Workshop 2008, Heraklion, Crete, Greece", S. GORLATCH, P. FRAGOPOULO, T. PRIOL (editors), 2-4 April 2008, p. 311-322.

[28] J. GALLARD, G. VALLÉE, A. LEBRE, C. MORIN, P. GALLARD, S. L. SCOTT. *Complementarity Between Virtualization and Single System Image Technologies*, in "VHPC 2008, 3rd Workshop on Virtualization in High-Performance Cluster and Grid Computing, Las Palmas de Gran Canaria, Canary Island, Spain", Springer LNCS, Held in conjunction with Euro-par 2008, 2008.

[29] E. JEANVOINE, C. MORIN. *RW-OGS: an Optimized Random Walk Protocol for Resource Discovery in Large Scale Dynamic Grids*, in "Proc. of the 9th IEEE/ACM International Conference on Grid Computing (GRID 2008), Tsukuba, Japan", October 2008.

[30] A. LEBRE, R. LOTTIAUX, E. FOCHT, C. MORIN. *Reducing Kernel Development Complexity in Distributed Environments*, in "Euro-Par", 2008, p. 576-586.

[31] X. LIU, Y. RADENAC, J.-P. BANÂTRE, T. PRIOL, Z. XU. *A Chemical Interpretation of GSML Programs*, in "Proceedings of Seventh International Conference on Grid and Cooperative Computing", IEEE, October 2008, p. 459 – 466.

[32] J. MEHNERT-SPAHN, M. SCHOETTNER, D. MARGERY, C. MORIN. *XtreemOS Grid Checkpointing Architecture*, in "IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2008), Lyon, France", Poster, May 2008.

[33] J. MEHNERT-SPAHN, M. SCHOETTNER, C. MORIN. *Checkpointing Process Groups in a Grid Environment*, in "Proc. of the 9th International Conference on Parallel and Distributed Computing (PDCAT'08)", December 2008.

[34] B. NICOLAE, G. ANTONIU, L. BOUGÉ. *Distributed Management of Massive Data. An Efficient Fine Grain Data Access Scheme*, in "International Workshop on High-Performance Data Management in Grid Environment (HPDGrid 2008), Toulouse", Held in conjunction with VECPAR'08. Electronic proceedings, 2008.

[35] B. NICOLAE, G. ANTONIU, L. BOUGÉ. *Enabling Lock-Free Concurrent Fine-Grain Access to Massive Distributed Data: Application to Supernovae Detection*, in "Poster Session - IEEE Cluster 2008, Tsukuba, Japan", Cluster/Grid, September 2008, http://hal.inria.fr/inria-00329698/en/.

[36] Z. NÉMETH, C. PÉREZ, T. PRIOL. *Chemical coordination: an abstract enactment model for workflows*, in "Proceedings of 9th International Workshop on State-of-the-Art in Scientific and Parallel Computing", to appear, May 2008.

[37] Z. NÉMETH, C. PÉREZ, T. PRIOL. *Towards Dynamic Workflow Enactment by Artificial Chemistry*, in "Proc. of the CoreGRID Integration Workshop 2008, Heraklion, Crete, Greece", S. GORLATCH, P. FRAGOPOULO, T. PRIOL (editors), 2-4 April 2008, p. 407–418.

[38] T. ROPARS, C. MORIN. *Fault Tolerance in Cluster Federations with O2P-CF*, in "Resilience 2008, Workshop on Resiliency in High Performance Computing, Los Alamitos, CA, USA", IEEE Computer Society, Held in conjunction with CCGrid 2008, 2008, p. 807-812.

[39] X. SHI, J.-L. PAZAT, E. RODRIGUEZ, H. JIN, H. JIANG. *Dynasa Adapting Grid Applications to Safety using Fault Tolerant Methods*, in "17th ACM/IEEE International Symposium on High Performance Distributed Computing(HPDC 2008)", Poster, 2008.

### National Peer-Reviewed Conference/Proceedings

[40] L. CUDENNEC. *Un service hiérarchique distribué de partage de données pour grille*, in "Rencontres francophones du Parallélisme (RenPar' 18), Fribourg, Suisse", 2008.

[41] S. JEULAND, Y. JÉGOU, O. D. SÁNCHEZ, C. MORIN. *Support dórganisations virtuelles au sein dún système déxploitation pour la grille*, in "Actes de RenPar' 18, Fribourg, Switzerland", February 2008.

[42] T. ROPARS, C. MORIN. *O2P : un protocole à enregistrement de messages extrêmement optimiste*, in "Actes de RenPar' 18", 2008.

### Scientific Books (or Scientific Book chapters)

[43] G. ANTONIU, A. CONGIUSTA, S. MONNET, D. TALIA, P. TRUNFIO. *Peer-to-Peer Metadata Management for Knowledge Discovery Applications in Grids*, in "Grid Middleware and Service Challenges and Solutions", CoreGRID series, Springer Verlag, 2008, p. 219-233.

### Books or Proceedings Editing

[44] L. BOUGÉ, M. FORSELL, J. L. TRÄFF, A. STREIT, W. ZIEGLER, M. ALEXANDER (editors). *Euro-Par 2007 Workshops Parallel Processing*, Lect. Notes in Comp. Science, vol. 4854, Springer-Verlag, Rennes, France, INRIA/IRISA, Rennes, France, 2008.

[45] S. GORLATCH, M. BUBAK, T. PRIOL (editors). *Achievements in European Research on Grid Systems*, CoreGRID Books, Springer, November 2008.

[46] S. GORLATCH, T. PRIOL, P. FRAGOPOULO (editors). *Grid Computing - Achievements and Prospects*, CoreGRID Books, Springer, September 2008.

[47] P. MAHONEN, P. KLAUS, T. PRIOL (editors). *Towards a Service-Based Internet*, Lecture Note in Computer Science, vol. 5377, Springer, November 2008.

[48] T. PRIOL, M. VANNESCHI (editors). *From Grids to Service and Pervasive Computing*, CoreGRID Books, Springer, August 2008.

### Research Reports

[49] M. ALDINUCCI, H. BOUZIANE, M. DANELUTTO, C. PÉREZ. *Towards Software Component Assembly Language Enhanced with Workflows and Skeletons*, Technical report, n$^o$ 0153, CoreGRID - Network of Excellence, Institute of Programming Model, 3 July 2008.

[50] M. ALDINUCCI, M. DANELUTTO, H. BOUZIANE, C. PÉREZ. *Towards a Spatio-Temporal sKeleton Model Implementation on top of SCA*, Technical report, n$^o$ 0171, CoreGRID - Network of Excellence, Institute of Programming Model, 31 August 2008, https://www.irisa.fr/paris/bibadmin/uploads/pdf/0.32734500%201225114949_TR-0171.pdf.

[51] A. Arenas, J.-P. Banâtre, T. Priol. *Developing Secure Chemical Programs with Aspects*, Technical report, n$^{o}$ TR-0166, Institute on Knowledge and Data Management , CoreGRID - Network of Excellence, August 2008, http://www.coregrid.net/mambo/images/stories/TechnicalReports/tr-0166.pdf.

[52] H. Bouziane, C. Pérez, T. Priol. *Combining a Software Component Model and a Workflow Language into a Component Model with Spatial and Temporal Compositions*, Research Report, n$^{o}$ 6421, INRIA, 2008, https://hal.inria.fr/inria-00211158.

[53] M. Caeiro-Rodriguez, T. Priol, Z. Németh. *Dynamicity in Scientific Workflows*, Technical report, n$^{o}$ TR-0162, Institute on Grid Information, Resource and Workflow Monitoring Services , CoreGRID - Network of Excellence, August 2008, http://www.coregrid.net/mambo/images/stories/TechnicalReports/tr-0162.pdf.

[54] T. Cortes, C. Franke, Y. Jégou, T. Kielmann, D. Laforenza, B. Matthews, C. Morin, L. P. Prieto, A. Reinefeld. *XtreemOS: a Vision for a Grid Operating System*, Technical report, n$^{o}$ 4, XtreemOS European Integrated Projet, May 2008, http://www.xtreemos.eu.

[55] J. Gallard, P. Gallard, A. Lebre, C. Morin, S. L. Scott, G. Vallée. *Refinement Proposal of the Goldberg's Theory*, RR-6613, Rapport de recherche, INRIA, 2008, http://hal.inria.fr/inria-00310899/en/.

[56] J. Gallard, O. Goga, A. Lebre, C. Morin. *Using VM Capabilities to Improve Dynamicity and Transparency in Cluster and Grid Usage*, to appear, Technical report, n$^{o}$ RR-XXXX, INRIA, December 2008.

[57] E. Jeanvoine, C. Morin. *RW-OGS: an Optimized Random Walk Protocol for Resource Discovery in Large Scale Dynamic Grids*, to appear, Research report, n$^{o}$ RR-XXXX, INRIA, Rennes, France, 2008.

[58] A. Lebre, R. Lottiaux, E. Focht, C. Morin. *Reducing Kernel Development Complexity In Distributed Environments*, Rapport de recherche, n$^{o}$ RR-6405, INRIA, 2008, http://hal.inria.fr/inria-00201911/en/.

[59] J. Mehnert-Spahn, T. Ropars, M. Schoettner, C. Morin. *XtreemOS grid checkpointing architecture*, to appear, Technical report, n$^{o}$ RR-XXXX, INRIA, December 2008.

[60] P. Riteau, A. Lebre, C. Morin. *Handling Persistent States in Process Checkpoint/Restart Mechanisms for HPC Systems*, to appear, Technical report, n$^{o}$ RR-6765, INRIA, December 2008, http://hal.inria.fr/inria-00346745/fr/.

## Other Publications

[61] Consortium, XtreemOS. , INRIA (editor)*XtreemOS Annex 1 - Description of Work - updated version M25-M42*, XtreemOS Integrated Project, IST-033576, June 2008, Contract funded by the European Commission.

[62] Consortium, XtreemOS. , INRIA (editor)*XtreemOS Periodic Management Report (M13-M24)*, XtreemOS Integrated Project, IST-033576, June 2008, Contract funded by the European Commission.

[63] B. Daix. *Enrichir la plate-forme de simulation numérique SALOME par la délégation du déploiement de ses conteneurs de composants à ANGE*, 2008, Deliverable EDF research contract.

[64] B. Daix. *On-demand Application Deployment for High Performance Computing*, 2008, Deliverable EDF research contract.

[65] O. GOGA. *Combining Virtual Machines with Batch Schedulers for Clusters and Grids*, Masters thesis, Technical University of Cluj-Napoca, Romania, September 2008.

[66] Y. JÉGOU, S. JEULAND, C. MORIN, O. D. SÁNCHEZ. *XtreemOS: A Grid Operating System Providing Native Virtual Organization Support*, September 2008, poster.

[67] C. MORIN, Y. JÉGOU, O. D. SÁNCHEZ. *XtreemOS: A Grid Operating System Providing Native Virtual Organizatio Support*, in "3rd EGEE user forum, Clermont-Ferrand, France", February 2008.

[68] R. K. NATH. *Fault Tolerance of the Application Manager in Vigne*, Internship Report, Masters thesis, University of Tennessee, September 2008.

[69] M. OBROVAC. *Scheduling Strategy in Single System Image Cluster Environments*, Masters thesis, University of Zagreb, Croatia, December 2008.

[70] P. RITEAU. *Coordination du système de fichiers avec les mécanismes de sauvegarde/reprise pour une grappe de calcul haute-performance*, Masters thesis, Université de Rennes 1, June 2008, https://www.irisa.fr/paris/bibadmin/uploads/pdf/0.13990200%201225708489_main.pdf.

[71] O. D. SÁNCHEZ, P. COSTA, G. PIPAN, C. MORIN. *Opportunities and Issues for EU Research Projects in Open Sourc Software: the XtreemOS Approach*, January 2008, Qualipso Conference.

[72] XTREEMOS, CONSORTIUM. , INRIA, TEAM PARIS (editors)*Collaboration report and plan including commitments for contributions to Tasks 1 to 8 (PC5)*, August 2008.

[73] XTREEMOS, CONSORTIUM. , INRIA, TEAM PARIS (editors)*Design and implementation of first advanced version of LinuxSSI*, November 2008.

[74] XTREEMOS, CONSORTIUM. , INRIA, TEAM PARIS (editors)*Evaluation of Linux native isolation mechanisms for XtreemOS flavours*, November 2008.

[75] XTREEMOS, CONSORTIUM. , INRIA, TEAM PARIS (editors)*XtreemOS administrator and user guide*, November 2008.

## References in notes

[76] I. FOSTER, C. KESSELMAN (editors). *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers, 1998.

[77] *Project JXTA: Java programmers guide*, Sun Microsystems, Inc., 2001, http://www.jxta.org/white_papers.html.

[78] *Wireless Application Protocol 2.0: technical white paper*, January 2002, http://www.wapforum.org/what/WAPWhite_Paper1.pdf.

[79] R. ARMSTRONG, D. GANNON, A. GEIST, K. KEAHEY, S. KOHN, L. MCINNES, S. PARKER, B. SMOLINSKI. *Toward a Common Component Architecture for High-Performance Scientific Computing*, in "Proceeding of the 8th IEEE International Symposium on High Performance Distributed Computation", August 1999.

[80] J.-P. BANÂTRE, P. FRADET, Y. RADENAC. *Principles of Chemical Programming*, in "Proceedings of the 5th International Workshop on Rule-Based Programming (RULE 2004)", S. ABDENNADHER, C. RINGEISSEN (editors), ENTCS, vol. 124, n$^{o}$ 1, Elsevier, June 2005, p. 133–147.

[81] J.-P. BANÂTRE, P. FRADET, D. LE MÉTAYER. *Gamma and the Chemical Reaction Model: Fifteen Years After*, in "Multiset Processing", LNCS, vol. 2235, Springer-Verlag, 2001, p. 17–44.

[82] J.-P. BANÂTRE, D. LE MÉTAYER. *A new computational model and its discipline of programming*, Technical report, n$^{o}$ RR0566, INRIA, September 1986, http://hal.inria.fr/inria-00075988.

[83] J.-P. BANÂTRE, D. LE MÉTAYER. *Programming by Multiset Transformation*, in "Communications of the ACM", vol. 36, n$^{o}$ 1, January 1993, p. 98–111.

[84] G. BERRY, G. BOUDOL. *The Chemical Abstract Machine*, in "Theoretical Computer Science", vol. 96, 1992, p. 217–248.

[85] D. CHEFROUR, F. ANDRÉ. *Auto-adaptation de composants ACEEL coopérants*, in "3e Conférence française sur les systèmes d'exploitation (CFSE 3)", 2003.

[86] K. GHARACHORLOO, D. LENOSKI, J. LAUDON, P. GIBBONS, A. GUPTA, J. HENESSY. *Memory Consistency and event ordering in scalable shared memory multiprocessors*, in "17th Annual Intl. Symposium on Computer Architectures (ISCA)", ACM, May 1990, p. 15–26.

[87] J. GRAY, D. SIEWIOREK. *High Availability Computer Systems*, in "IEEE Computer", September 1991.

[88] M. JAN. *JuxMem : un service de partage transparent de données pour grilles de calculs fondé sur une approche pair-à-pair*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes, France, November 2006.

[89] E. JEANNOT, B. KNUTSSON, M. BJORKMANN. *Adaptive Online Data Compression*, in "IEEE High Performance Distributed Computing (HPDC 11)", 2002.

[90] P. KELEHER, A. COX, W. ZWAENEPOEL. *Lazy Release Consistency for Software Distributed Shared Memory*, in "19th Intl. Symposium on Computer Architecture", May 1992, p. 13–21.

[91] P. KELEHER, D. DWARKADAS, A. COX, W. ZWAENEPOEL. *TreadMarks: Distributed Shared Memory on standard workstations and operating systems*, in "Proc. 1994 Winter Usenix Conference", January 1994, p. 115–131.

[92] L. LAMPORT. *How to Make a Multiprocessor Computer That Correctly Executes Multiprocess Programs*, in "IEEE Transactions on Computers", vol. 28, n$^{o}$ 9, September 1979, p. 690–691.

[93] P. LEE, T. ANDERSON. *Fault Tolerance: Principles and Practice*, vol. 3 of Dependable Computing and Fault-Tolerant Systems, Springer Verlag, second revised edition, 1990.

[94] F. MATTERN. *Virtual Time and Global States in Distributed Systems*, in "Proc. Int. Workshop on Parallel and Distributed Algorithms, Gers, France", North-Holland, 1989, p. 215–226.

[95] D. S. MILOJICIC, V. KALOGERAKI, R. LUKOSE, K. NAGARAJA, J. PRUYNE, B. RICHARD, S. ROLLINS, Z. XU. *Peer-to-Peer Computing*, Submitted to Computing Surveys, Research Report, n$^{o}$ HPL-2002-57, HP Labs, March 2002, http://www.hpl.hp.com/techreports/2002/HPL-2002-57R1.pdf.

[96] S. MONNET. *Gestion des données dans les grilles de calcul : support pour la tolérance aux fautes et la cohérence des données*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes, France, November 2006.

[97] OMG. *CORBA Component Model V3.0*, June 2002, OMG Document formal/2002-06-65.

[98] G. PĂUN. *Computing with Membranes*, in "Journal of Computer and System Sciences", vol. 61, n$^{o}$ 1, 2000, p. 108-143.

[99] Y. RADENAC. *Programmation "chimique" d'ordre supérieur*, Thèse de doctorat, Université de Rennes 1, April 2007.

[100] X. SHI, J.-L. PAZAT. *A Novel Adaptive and Safe Framework for Ubicomp*, in "Proceedings of the 2007 International Workshop on Service, Security and its Data management for Ubiquitous Computing (SSDU-07)", May 2007.

[101] C. SZYPERSKI. *Component Software - Beyond Object-Oriented Programming*, Addison-Wesley / ACM Press, 1998.