INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MESCAL

Middleware Efficiently SCALable

Grenoble - Rhône-Alpes

THEME NUM

Activity Report

2007

Table of contents

1. Team

The MESCAL project-team is a common project-team supported by CNRS, INPG, UJF and INRIA located in the LIG laboratory (UMR 5217).

Head of project-team

Bruno Gaujal [Research Director (DR) INRIA, HdR]

Administrative staff

Marion Ponsot [Secretary (SAR) INRIA; half time]

INRIA Staff

Corinne Touati [Research Associate (CR) INRIA]
Derrick Kondo [Research Associate (CR) INRIA]

CNRS Staff

Arnaud Legrand [Research Associate (CR) CNRS]

INPG Staff

Yves Denneulin [Associate Professor, HdR]
Brigitte Plateau [Associate Professor, HdR]

UJF Staff

Vania Martin [Associate Professor]
Jean-François Méhaut [Associate Professor, HdR]
Florence Perronnin [Associate Professor]
Olivier Richard [Associate Professor, INRIA Delegation]
Jean-Marc Vincent [Associate Professor]

Project-Team technical staff

Saïd Oulahal [Engineer Assistant]

Invited Scientist

Jean-Michel Fourneau [Associate Professor, HdR]

PostDoc

Ana Bušić [ANR SMS, October 2007]
Sébastien Lagrange [ARC Coinc, November 2006]

PhD students

Carlos Barrios [2005, EGIDE, co-tutelle]
Rémi Bertin [2007, ANR DOCCA]
Léonardo Brenner [2004, Brazilian CAPES scholarship]
Ricardo Czekster [2007, Brazilian CAPES scholarship]
Nicolas Gast [2007, AC]
Yiannis Georgiou [2006, CIFRE BULL scholarship]
Ahmed Harbaoui [2006, CIFRE France Télécom R&D scholarship]
Hussein Joumma [2006, MNRT scholarship]
Pedro Antonio Madeira [2006, Brazilian CAPES scholarship]
Duc Nguyen [2005, INRIA scholarship]
Lucas Nussbaum [2005, BDI-CNRS MNRT scholarship]
Carlos Rojas [2007, CIFRE STMicroelectronics]
Afonso Sales [2005, Brazilian CAPES scholarship]
Nazha Touati [2004, Rhône-Alpes scholarship]
Olivier Valentin [2003, MNRT scholarship]
Jérome Vienne [2006, CIFRE BULL scholarship]
Brice Videau [2005, MNRT scholarship]
Blaise Yenké [2004, Ngaundere University scholarship]
Thais Webber [2006, CAPES-COFECUB scholarship, cotutelle]

Former PhD students

Maxime Martinasso [CIFRE BULL scholarship. Currently at Shell Inc.]
Ihab Sbeity [MRNT scholarship. Currently professor at Lebanese International University]

2. Overall Objectives

2.1. Objectives

The recent evolutions in computer networks technology, as well as their diversification, goes with a tremendous change in the use of these networks: applications and systems can now be designed at a much larger scale than before. This scaling evolution concerns at the same time the amount of data, the number of computers, the number of users, and the geographical diversity of these users.

This race towards *large scale* computing questions many assumptions underlying parallel and distributed algorithms and operating middleware. Today, most software tools developed for average size systems cannot be run on large scale systems without a significant degradation of their performances.

The goal of the MESCAL project-team is to design and validate efficient exploitation mechanisms (middleware and system services) for large distributed infrastructures.

MESCAL's target applications are intensive scientific computations such as cellular micro-physiology, protein conformations, particle detection, combinatorial optimization, Monte Carlo simulations, and others. Such applications are constituted of a large set of independent, equal-sized tasks and therefore may benefit from large-scale computing platforms. Initially executed on large dedicated clusters (CRAY, IBM, COMPAQ), they have been recently deployed on collections of homogeneous clusters aggregating a large number of commodity components. The experience showed that such clusters offer a huge computing power at a very reasonable price. MESCAL's target infrastructures are aggregations of commodity components and/or commodity clusters at metropolitan, national or international scale. Examples of target infrastructures are grids obtained through sharing of available resources inside autonomous computing services, lightweight grids (such as the local CIMENT Grid) which are limited to trusted autonomous systems, clusters of intranet resources (Condor) or aggregation of Internet resources (SETI@home, Xtremweb).

MESCAL's methodology in order to ensure **efficiency** and **scalability** of proposed mechanisms is based on systematic modeling and performance evaluation of target architectures, software layers and applications.

3. Scientific Foundations

3.1. Large System Modeling and Analysis

Keywords: *Discrete event dynamic systems*, *Markov chains*, *Performance evaluation*, *Petri nets*, *Queuing networks*, *Simulation*.

Participants: Bruno Gaujal, Derrick Kondo, Arnaud Legrand, Florence Perronnin, Brigitte Plateau, Olivier Richard, Corinne Touati, Jean-Marc Vincent.

Understanding qualitative and quantitative properties of distributed systems and parallel applications is a major issue. The *a posteriori* analysis of the behavior of the system or the design of predictive models are notoriously challenging problems.

Indeed, large distributed systems contain many different features (processes, threads, jobs, messages, packets) with intricate interactions between them (communications, synchronizations). The analysis of the global behavior of the system requires to take into account large data sets.

As for *a priori* models, our current research focuses on capturing the distributed behavior of large dynamic architectures. Actually, both formal models and numerical tools are being used to get predictions on the behavior of large systems.

For large parallel systems, the non-determinism of parallel composition, the unpredictability of execution times and the influence of the outside world are usually expressed in the form of multidimensional stochastic processes which are continuous in time with a discrete state space. The state space is often infinite or very large and several specific techniques have been developed to deal with what is often termed as the "curse of dimensionality".

MESCAL deals with this problem using several complementary tracks:

- Behavior analysis of highly distributed systems,
- Simulation algorithms able to deal with very large systems,
- Fluid limits (used for simulation and analysis),
- Decomposition of the state space,
- Structural and qualitative analysis,
- Game theory methods for resolving resource contention.

3.1.1. Behavior analysis of highly distributed systems

The development of highly distributed architectures running widely spread applications requires to elaborate new methodologies to analyze the behavior of systems. Indeed, runtime systems on such architectures are empirically tuned. Analysis of executions are generally manually performed on post-mortem traces that have been extracted with very specific tools. This tedious methodology is generally motivated by the difficulty to characterize the resources of such systems. For example, big clusters, grids or peer-to-peer (P2P) [1] networks present properties of size, heterogeneity, dynamicity that are usually not taken into account in classical system models. The asynchrony of the architecture also induces perturbations in the behavior of the application leading to significant slow-down that should be avoided. Therefore, when defining the workload of the system, the distributed nature of applications should be taken into account with a specific focus on problems related to synchronizations.

3.1.2. Simulation of distributed systems

Since the advent of distributed computer systems, an active field of research has been the investigation of *scheduling* strategies for parallel applications. The common approach is to employ scheduling heuristics that approximate an optimal schedule. Unfortunately, it is often impossible to obtain analytical results to compare the efficiency of these heuristics. One possibility is to conduct large numbers of back-to-back experiments on real platforms. While this is possible on tightly-coupled platforms, it is infeasible on modern distributed platforms (i.e. Grids or peer-to-peer environments) as it is labor-intensive and does not enable repeatable results. The solution is to resort to *simulations*. Simulations not only enable repeatable results but also make it possible to explore wide ranges of platform and application scenarii.

The SIMGRID framework enables the simulation of distributed applications in distributed computing environments for the specific purpose of developing and evaluating scheduling algorithms. This software is the result of a long-time collaboration with Henri CASANOVA (University of California, San Diego).

3.1.3. Perfect Simulation

Using a constructive representation of a Markovian queuing network based on events (often called GSMPs), we have designed a perfect simulation tool computing samples distributed according to the stationary distribution of the Markov process with no bias. Two softwares have been developed. ψ analyzes a Markov chain using its transition matrix and provides perfect samples of cost functions of the stationary state. ψ^2 samples the stationary measure of Markov processes using directly the queuing network description. Some monotone networks with up to 10^{50} states can be handled within minutes over a regular PC.

[1] Our definition of peer-to-peer is a network (mainly the Internet) over which a large number of autonomous entities contribute to the execution of a single task.

3.1.4. Fluid models

When the size of systems grows very large, one may use asymptotic techniques to get a faithful estimate of their behaviors. One such tools is mean field analysis and fluid limits, that are used on a modeling and simulation level. One recent significant application is call centers. Another one is peer to peer systems. Web caches as well as peer-to-peer systems must be able to serve a set of customers which is both large (several tens of thousands) and highly volatile (with short connection times). These features make analysis difficult when classical approaches (like Markovian Models or simulation) are used. We have designed simple fluid models to get rid of one dimension of the problem. This approach has been applied to several systems of web caches (such as Squirrel) and to peer-to-peer systems (such as BitTorrent). This helps to get a better understanding of the behavior of the system and to solve several optimization problems.

3.1.5. Markov Chain Decomposition

The first class of models we will be using is Continuous time Markov chains (CTMC). The usefulness of Markov models is undisputed, as attested by the large number of modeling tools implementing Markov solvers. However their practical applications are limited by the *state-space explosion* problem, which puts excessive demands on memory and execution time when studying large real-life systems. Continuous-time Stochastic Automata Networks describe a system as a set of subsystems that interact. Each subsystem is modeled by a stochastic automaton, and some rules between the states of each automaton describe the interactions between subsystems. The main challenge is to come up with ways to compute the asymptotic (or transient) behavior of the system without ever generating the whole state space. Several techniques have been developed in our group based on bounds, lumpability, symmetry and properties of the Kronecker product. Most of them have been integrated in a software tool (PEPS) which is openly available.

3.1.6. Discrete Event Systems

The interaction of several processes through synchronization, competition or superposition within a distributed system is a big source of difficulties because it induces a state space explosion and a non-linear dynamic behavior. The use of exotic algebra, such as (min,max,plus) can help. Highly synchronous systems become linear in this framework and therefore are amenable to formal solutions. More complicated systems are neither linear in (max,plus) nor in the classical algebra. Several qualitative properties have been established for a large class of such systems called free-choice Petri nets (sub-additivity, monotonicity or convexity properties). Such qualitative properties are sometimes enough to assess the class of routing policies optimizing the global behavior of the system. They are also useful to design efficient numerical tools computing their asymptotic behavior.

3.1.7. Game Theory Methods for Resolving Resource Contention

Resources in large-scale distributed platforms (Grid computing platforms, enterprise networks, peer-to-peer systems) are shared by a number of users having conflicting interests who are thus prone to act selfishly. A natural framework for studying such non-cooperative individual decision-making is game theory. In particular, game theory models the decentralized nature of decision-making.

It is well known that such non-cooperative behaviors can lead to important inefficiencies and unfairness. In other words, individual optimizations often results in global resource waste. In the context of game theory, a situation in which all users selfishly optimize their own utility is known as a *Nash equilibrium* or *Wardrop equilibrium*. In such equilibria, no user has interest in unilaterally deviating from its strategy. Such policies are thus very easy to implement in a fully distributed system and have some stability properties. However, a possible consequence is the *Braess paradox* in which the increase of resource happens at the expense of *every* user. This is why, the study of the occurrence and degree of such inefficiency is of crucial interest. Up until now, little is known about general conditions for optimality or degree of efficiency of these equilibria, in a general setting.

Many techniques have been developed to enforce some form of collaboration and improve these equilibria. In this context, it is generally prohibitive to take joint decisions so that a global optimization cannot be achieved. A possible option relies on the establishment of virtual prices, also called *shadow prices* in congestion networks. These prices ensure a rational use of resources. Equilibria can also be improved by advising policies to mobiles such that any user that does not follow these pieces of advice will necessarily penalize herself (*correlated equilibria*).

3.2. Management of Large Architectures

Keywords: *Administration*, *Clusters*, *Deployment*, *Grids*, *Job scheduler*, *Peer-to-peer*.

Participants: Derrick Kondo, Arnaud Legrand, Olivier Richard, Corinne Touati, Vania Marangozova.

Most distributed systems deployed nowadays are characterized by a high dynamism of their entities (participants can join and leave at will), a potential instability of the large scale networks (on which concurrent applications are running), and the increasing probability of failure. Therefore, as the size of the system increases, it becomes necessary that it adapts automatically to the changes of its components, requiring a self-organization of the system with respect to the arrival and departure of participants, data, or resources.

As a consequence, it becomes crucial to understand and model the behavior of large scale systems, to efficiently exploit these infrastructures. In particular it is essential to design dedicated algorithms and infrastructures handling a large amount of users and/or data.

MESCAL deals with this problem using several complementary tracks:

- Fairness in large-scale distributed systems,
- Deployment and management tools,
- Scalable batch scheduler for clusters and grids.

3.2.1. Fairness in large-scale distributed systems

Large-scale distributed platforms (Grid computing platforms, enterprise networks, peer-to-peer systems) result from the collaboration of many people. Thus, the scaling evolution we are facing is not only dealing with the amount of data and the number of computers but also with the number of users and the diversity of their behavior. In a high-performance computing framework, the rationale behind this joining of forces is that most users need a larger amount of resources than what they have on their own. Some only need these resources for a limited amount of time. On the opposite some others need as many resources as possible but do not have particular deadlines. Some may have mainly tightly-coupled applications while some others may have mostly embarrassingly parallel applications. The variety of user profiles makes resources sharing a challenge. However resources have to be *fairly* shared between users, otherwise users will leave the group and join another one. Large-scale systems therefore have a real need for fairness and this notion is missing from classical scheduling models.

3.2.2. Tools to operate clusters

The MESCAL project-team studies and develops a set of tools designed to help the installation and the use of a cluster of PCs. The first version had been developed for the icluster1 platform exploitation. The main tools are a scalable tool for cloning nodes (KA-DEPLOY) and a parallel launcher based on the TAKTUK project (now developed by the MOAIS project-team). Many interesting issues have been raised by the use of the first versions among which we can mention environment deployment, robustness and batch scheduler integration. A second generation of these tools is thus under development to meet these requirements.

The new KA-DEPLOY has been retained as the primary deployment tool for the experimental national grid GRID'5000.

3.2.3. Simple and scalable batch scheduler for clusters and grids

Most known batch schedulers (PBS, LSF, Condor, ...) are of old-fashioned conception, built monolithically, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150000 lines of code for OpenPBS), offering a growing number of functions that are, most of the time, not used. In such a context, it becomes hard to control both the robustness and the scalability of the whole system.

OAR is an attempt to address these issues. Firstly, OAR is written in a very high level language (Perl) and makes intensive use of high level tools (MySql and TAKTUK), thereby resulting in a concise code (around 5000 lines of code) easy to maintain and extend. This small code as well as the choice of widespread tools (MySql) are essential elements that ensure a strong robustness of the system. Secondly, OAR makes use of SQL requests to perform most of its job management tasks thereby getting advantage of the strong scalability of most database management tools. Such scalability is further improved in OAR by making use of TAKTUK to manage nodes themselves.

3.3. Migration and resilience

Keywords: *Fault tolerance*, *distributed algorithms*, *migration*.

Participants: Yves Denneulin, Jean-François Méhaut.

Making a distributed system reliable has been and remains an active research domain. Nonetheless this has not so far lead to results usable in an intranet or federal architecture for computing. Most propositions address only a given application or service. This may be due to the fact that until clusters and intranet architectures arose, it was obvious that client and server nodes were independent. So, a fault or a predictable disconnection on most of the nodes didn't lead to a complete failure of the system. This is not the case in parallel scientific computing where a fault on a node can lead to a data loss on thousands of other nodes. The reliability of the system is hence a crucial point. MESCAL's work on this topic is based on the idea that each process in a parallel application will be executed by a group of nodes instead of a single node: when the node in charge of a process fails, another in the same group can replace it in a transparent way for the application.

There are two main problems to be solved in order to achieve this objective. The first one is the ability to migrate processes of a parallel, and thus communicating, application without enforcing modifications. The second one is the ability to maintain a group structure in a completely distributed way. The first one relies on a close interaction with the underlying operating systems and networks, since processes can be migrated in the middle of a communication. This can only be done by knowing how to save and replay later all ongoing communications, independently of the communications. Freezing a process to restore it on another node is also an operation that requires collaboration of the operating system and a good knowledge of its internals. The other main problem (keeping a group structure) belongs to the distributed algorithms domain and is of a much higher level nature.

The resulting software of this research topic is called SAMORY and is able to keep a set of processes alive on a given set of nodes, even in the presence of faults, hardware or software. It has been used on seismic applications (wave propagation) as a part of the RNTL IGGI project (http://iggi.imag.fr). SAMORY is freely available at http://iggi.imag.fr. It is composed of a Linux kernel module and a daemon that monitors the processes and their communication, and reacts when a fault is discovered or suspected. This includes a group management mechanism, to associate a group of replicates to every physical node, that aims at minimizing the number of members necessary to ensure resistance to a given number of faults. This group membership also creates a structured communication graph between the members of a group to route messages between them in a deterministic manner.

Future work will concern the behavior analysis of checkpoint systems in order to predict precisely critical operations to optimize resource usage (network and disk bandwidth).

3.4. Large scale data management

Keywords: *Fault tolerance*, *distributed algorithms*, *migration*.

Participants: Yves Denneulin, Vania Marangozova.

In order to use large data, it is necessary (but not always sufficient, as seen later) to efficiently store and transfer them to a given site (a set of nodes) where it is going to be used. The first step toward this achievement is the construction of a file system that is an extension of NFS for the grid environment. The second step is an efficient transfer tool that provides throughput close to optimal (*i.e.* the capacity of the underlying hardware).

3.4.1. Fast distributed storage over a cluster

Our goal here is to design a distributed file system for clusters that enables one to store data over a set of nodes (instead of a single one). It was designed to permit the usage of a set of disks to optimize memory allocations. It is important for performance and simplicity that this new file system has little overhead for access and updates. From a user point of view, it is used just as a classical NFS. From the server point of view, however, the storage is distributed over several nodes (possibly including the users).

The mounting point is only in charge of the meta-data, name, owner, access permissions, size, inodes, and etc., of the files while their content is stored on separate nodes. Every read or write request is received by the meta-server, the mounting point, which sends them to the relevant storage nodes, called IOD for Input/Output Daemon which will serve the request and send the result to the client.

Two implementations were done, one at the user level and one at the kernel level. Performances are good for read operations, for example 150MBs/sec for 16 IODs connected through a 100Mb/s for 16 clients. For write operations performances are limited by the bandwidth available for the meta-server which is a significant bottleneck.

3.4.2. Reliable distribution of data

Storage distribution on a large set of disks raises the reliability problem: more disks mean a higher fault rate. To address this problem we introduced in NFSP a redundancy on the IODs, the storage nodes by defining VIOD, Virtual IOD, which is a set of IODs that contain exactly the same data. So when an IOD fails another one can serve the same data and continuity of service is insured though. This doesn't modify the way the file-system is used by the clients: distribution and replication remain transparent. Several consistency protocols are proposed with various levels of performance; they all enforce at least the NFS consistency which is expected by the client.

3.4.3. Efficient transfer on grids

To efficiently transfer files across a grid, a "beowulf-like" solution consists in creating a set of point-to-point communications to parallelize the transfer of a file or a set of files. This approach was chosen, for instance, in gridftp [59]. It implies duplicating the data to transfer or distribute them on separate nodes before the transfer begins. We use the distributed storage property of NFSp to be able to do parallel transfer transparently. However, since a grid is heterogeneous from a hardware and a software point of view, we decided to build our own solution in a generic way, it can be used by any kind of data server: SAN, local file systems, NFS or NFSP. The component in charge of transfer across the grid is called GXFER, for Grid Transfer, its goal is to copy files between sites. A copy is done in a parallel way if both sender and receiver can handle it and have distributed storage capability. GXFER can be used as an external program, it will then behave like the classic `scp` command or can be used as a library inside an application.

GXFER performances are good, with a 1Gbytes file transferred in less than 10 seconds, 9.6s, between sites in Grenoble and Lyon connected with a 1Gbits/s link, with NFSp servers on both sides. Further experiments exhibited good scaling properties.

4. Application Domains

4.1. Introduction

Applications in the fields of numerical simulation, image synthesis, and processing are typical of the user demand for high performance computing. In order to confront our proposed solutions for parallel computing with real applications, the project-team is involved in collaborations with end-users to help them parallelize their applications.

4.2. Bioinformatics

Keywords: *heterogeneous collection of databanks*, *protein comparison*.

Participant: Arnaud Legrand.

This joint work involves the GRAAL project-team.

The problem of searching large-scale genomic sequence databases is an increasingly important bioinformatics problem. We have obtained results on the deployment of such applications in heterogeneous parallel computing environments. These results are based on the analysis of the GriPPS [61], [60] protein comparison application. The GriPPS framework is based on large databases of information about proteins; each protein is represented by a string of characters denoting the sequence of amino acids of which it is composed. Biologists need to search such sequence databases for specific patterns that indicate biologically homologous structures. The GriPPS software enables such queries in grid environments, where the data may be replicated across a distributed heterogeneous computing platform.

In fact, this application is a part of a larger class of applications, in which each task in the application workload exhibits an "affinity" for particular nodes of the targeted computational platform. In the genomic sequence comparison scenario, the presence of the required data on a particular node is the sole factor that constrains task placement decisions. In this context, task affinities are determined by location and replication of the sequence databanks in the distributed platform.

Such biological sequence comparison algorithms are however typically computationally intensive, embarrassingly parallel workloads. In the scheduling literature, this computational model is effectively a *divisible workload scheduling* problem with negligible communication overheads. This framework has enabled us to propose online scheduling algorithms whose output is *fair* and *efficient*: the slowdown experienced by every user due to the load incurred by the others is as uniform as possible.

4.3. On-demand Geographical Maps

Participant: Jean-Marc Vincent.

This joint work involves the UMR 8504 Géographie-Cité, LSR-IMAG, UMS RIATE and the Maisons de l'Homme et de la Société.

Improvements in the Web developments have opened new perspectives in interactive cartography. Nevertheless existing architectures have some problems to perform spatial analysis methods that require complex calculus over large data sets. Such a situation involves some limitations in the query capabilities and analysis methods proposed to users. The HyperCarte consortium with LSR-IMAG, Géographie-cité and UMR RIATE proposes innovative solutions to these problems. Our approach deals with various areas such as spatio-temporal modeling, parallel computing and cartographic visualization that are related to spatial organizations of social phenomena.

Nowadays, analysis are done on huge heterogeneous data set. For example, demographic data sets at nuts 5 level, represent more than 100.000 territorial units with 40 social attributes. Many algorithms of spatial analysis, in particular potential analysis are quadratic in the size of the data set. Then adapted methods are needed to provide "user real time" analysis tools.

4.4. Seismic simulations

Participant: Jean-François Méhaut.

Numerical modeling of seismic wave propagation in complex three-dimensional media is an important research topic in seismology. Several approaches will be studied, and their suitability with respect to the specific constraints of NUMA architectures shall be evaluated. These modeling approaches will rely on modern numerical schemes such as spectral elements, high-order finite differences or finite elements applied to realistic 3D models. The NUMASIS project (see Section 8.2.4) will focus on issues related to parallel algorithms (distribution, scheduling) in order to optimize computations based on such numerical schemes by taking advantage of execution frameworks developed for NUMA architectures.

These approaches will be tested and validated on applications related to seismic risk assessment. Recent seismic events as those in Asia have evidenced the crucial research and development needs in this field. Some regions in France may as well be prone to such risks (French Riviera, Alps, French Antilles,...) and the experiments in the NUMASIS project will be carried out using some of the available data from these regions.

4.5. The CIMENT project

Participant: Olivier Richard.

The CIMENT project (Intensive Computing, Numerical Modeling and Technical Experiments, http://ciment.ujf-grenoble.fr/) gathers a wide scientific community involved in numerical modeling and computing (from numerical physics and chemistry to astrophysics, mechanics, bio-modeling and imaging) and the distributed computer science teams from Grenoble. Among these various application domains, there is a huge demand to manage executions of large sets of independent jobs. These sets have between 10,000 to 100,000 jobs each. Providing a middleware able to steer such an amount of jobs is a challenge. The CiGri middleware project addresses this issue in a grid infrastructure.

The aim of the CiGri project is to gather the unused computing resource from intranet infrastructure and to make it available for large scale applications. This grid is based on two software tools. The CiGri server software is based on a database and offers a user interface for launching grid computations (scripts and web tools). It interacts with the computing clusters through a batch scheduler software. CiGri is compatible with classical batch systems like PBS, but an efficient batch software (OAR, http://oar.imag.fr/) has been developed by the MESCAL and MOAIS project-teams for the easy integration and testing of scheduling tools.

5. Software

5.1. Tools for cluster management and software development

The large-sized clusters and grids show serious limitations in many basic system softwares. Indeed, the launching of a parallel application is a slow and significant operation in heterogeneous configurations. The broadcast of data and executable files is widely under the control of users. Available tools do not scale because they are implemented in a sequential way. They are mainly based on a single sequence of commands applied over all the cluster nodes. In order to reach a high level of scalability, we propose a new design approach based on a parallel execution. We have implemented a parallelization technique based on spanning trees with a recursive starting of programs on nodes. Industrial collaborations were carried out with Mandrake, BULL, HP and Microsoft.

5.1.1. KA-Deploy: deployment tool for clusters and grids

KA-DEPLOY is an environment deployment toolkit that provides automated software installation and reconfiguration mechanisms for large clusters and light grids. The main contribution of KA-DEPLOY 2 toolkit is the introduction of a simple idea, aiming to be a new trend in cluster and grid exploitation: letting users concurrently deploy computing environments tailored exactly to their experimental needs on different sets of nodes. To reach this goal KA-DEPLOY must cooperate with batch schedulers, like OAR, and use a parallel launcher like TAKTUK (see below).

5.1.2. Taktuk: parallel launcher

TAKTUK is a tool to launch or deploy efficiently parallel applications on large clusters, and simple grids. Efficiency is obtained thanks to the overlap of all independent steps of the deployment. We have shown that this problem is equivalent to the well known problem of the single message broadcast. The performance gap between the cost of a network communication and of a remote execution call enables us to use a work stealing algorithm to realize a near-optimal schedule of remote execution calls. Currently, a complete rewriting based on a high level language (precisely Perl script language) is under progress. The aim is to provide a light and robust implementation. This development is lead by the MOAIS project-team.

5.1.3. NFSp and Gxfer: parallel file system

When deploying a cluster of PCs there is a lack of tools to give a global view of the available space on the drives. This leads to a suboptimal use of most of this space. To address this problem NFSP was developed, as an extension to NFS that divides file system handling in two components: one responsible for the data stored and the other for the metadata, like inodes, access permission.... They are handled by a server, fully NFS compliant, which will contact associated data servers to access information inside the files. This approach enables a full compatibility, for the client side, with the standard in distributed file systems, NFS, while permitting the use of the space available on the clusters nodes. Moreover efficient use of the bandwidth is done because several data servers can send data to the same client node, which is not possible with a usual NFS server. The prototype has now reached a mature state. Sources are available at http://nfsp.imag.fr.

5.1.4. aIOLi

Modern distributed software uses and creates huge amounts of data with typical parallel I/O access patterns. Several issues, like *out-of-core limitation* or *efficient parallel input/output access* already known in a local context (on SMP nodes for example), have to be handled in a distributed environment such as a cluster.

We have designed AIOLI, an efficient I/O library for parallel access to remote storage in SMP clusters. Its SMP kernel features provide parallel I/O without inter-processes synchronization mechanisms as well as a simple interface based on the classic UNIX system calls (create/open/read/write/close). The AIOLI solution allows us to achieve performance close to the limits of the remote storage system. This was done in several steps:

- Build a local framework that can do aggregation of requests at the application level. This is done by putting a layer between the application and the kernel in charge of delaying individual requests in order to merge them and thus improve performances. The key factor here is to control the delay that should be large enough to discover aggregation patterns but with a limit to avoid excessive waiting times.
- Schedule all I/O requests on a cluster in a global way in order to avoid congestion on a server that leads to bad performances.
- Schedule I/O requests locally on the server so that methods of aggregation and mixing of client requests can be used to improve performances. For that reason AIOLI had to be ported to the kernel and placed at both the VFS level and the lower file system one.

Today, AIOLI compares favorably with the best MPI/IO implementation without any modification of the applications [62] sometimes with a factor of 4. AIOLI can be downloaded from the address http://aioli.imag.fr, both the user library and the Linux kernel module versions.

5.1.5. SAMORY

SAMORY *is an architecture to provide resiliency to parallel applications running on top of virtual clusters, typically built from an intranet or an enterprise network.*

SAMORY is a runtime aiming at providing resiliency to high performance computing applications running on a virtual cluster, typically hosts of an intranet. It is composed of a Linux kernel module that must be loaded at runtime and a distributed architecture for monitoring, checkpointing and restarting communicating processes. The size of the replication group for a process, the number of copies that will be done for a process, is a parameter that can be fixed, and modified, at runtime depending on the availability of the hosts. The communications between processes are also taken into account by SAMORY and, when a process fails, all pending communications will be transferred transparently to the site where the corresponding backup will resume.

The main advantage of SAMORY is its total transparency with respect to the applications: starting the monitoring of an application is solely telling the runtime to do so and the applications are not changed in any way. This can be done at runtime. Introduction of the checkpointing hinders performances but in a reasonable way, the cost of checkpointing a process is directly proportional to the amount of memory it uses with, for example, a checkpoint time of 400ms for a 100Mbytes process on a PC with a Pentium 4 at 2Ghz and 512Mbs of RAM. Virtual memory management, and the appearance of faults, increases these values for large processes, 10s for a 400Mbytes process. By saving only the state related part of the process, excluding code and shared library for example, this cost can be reduced by 75%. The time necessary to restart a process is low, typically milliseconds, since most data will be loaded when necessary and so the execution can resume soon but will generate page faults later. Since the final time will heavily depend on the behavior of the applications it is not possible to give generic performance results for this step. The overhead on the communications is negligible for small amounts of data and becomes significant for messages of size 8Mbytes.

5.1.6. Gedeon

Gedeon is a middleware for data management on grids. It handles metadata, lists of records made of (attribute, value) pairs, stored in a distributed manner on a grid. Advanced requests can be done on them, using regular expression, and they can be combined in traditional ways, aggregation for example, or used through join operations to federate various sources.

5.1.7. Generic trace and visualization: Paje

This software was formerly developed by members of the Apache project-team. Even if no real research effort is anymore done on this software, many members of the MESCAL project-team use it in their everyday research and promote its use. This software is now mainly maintained by Benhur Stein from Federal University Santa Monica (UFSM), Brazil.

PAJE allows applications programmers to define what is visualized and how new objects should be drawn. To achieve such flexibility, the hierarchy of events and the visualization commands may be defined by the programmers inside the applications. The visualization of parallel execution of ATHAPASCAN applications was achieved without any new addition into PAJE software. Inserting few events trace into the ATHAPASCAN runtime allows the visualization of different facets of the program: application computation time but also user task graph management and scheduling of these tasks. PAJE is also, among others, used to visualize Java program execution and large cluster monitoring. PAJE is actively used by the SIMGRID users' community and the NUMASIS project (see Section 8.2.4).

5.1.8. OAR: a simple and scalable batch scheduler for clusters and grids

OAR is a batch scheduler that emphasizes simplicity, extensibility, modularity, efficiency, robustness and scalability. It is based on a high level conception that reduces drastically its software complexity. Its internal architecture is built on top of two main components: a generic and scalable tool for the administration of the cluster (launch, nodes administration, ...) and a database as the only way to share information between its internal modules. Completely written in Perl, OAR is also extremely modular and straightforward to extend. Thus, it constitutes a privileged platform to develop and evaluate several scheduling algorithms and new kinds of services.

Most known batch schedulers (PBS, LSF, Condor, ...) are of old-fashioned conception, built monolithically, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150000 lines of code for OpenPBS), offering a growing number of functions that are, most of the time, not used. In such a context, it becomes hard to control both the robustness and the scalability of the whole system.

The OAR project focuses on robust and highly scalable batch scheduling for clusters and grids. Its main objectives are the validation of grid administration tools such as TAKTUK, the development of new paradigms for grid scheduling and the experimentation of various scheduling algorithms and policies.

The grid development of OAR has already started with the integration of best effort jobs whose purpose is to take advantage of idle times of the resources. Managing such jobs requires a support of the whole system from the highest level (the scheduler has to know which tasks can be canceled) down to the lowest level (the execution layer has to be able to cancel awkward jobs). The OAR architecture is perfectly suited to such developments thanks to its highly modular architecture. Moreover, this development is used for the CiGri grid middleware project.

The OAR system can also be viewed as a platform for the experimentation of new scheduling algorithms. Current developments focus on the integration of theoretical batch scheduling results into the system so that they can be validated experimentally.

5.2. Simulation tools

5.2.1. SimGrid: simulation of distributed applications

SIMGRID implements realistic fluid network models that enable very fast yet precise simulations. SIMGRID enables the simulation of distributed scheduling agents, which has become critical for current scheduling research in large-scale platforms.

Sources and documentations of SIMGRID are available at the following address http://simgrid.gforge.inria.fr/.

5.2.2. ψ and ψ^2: perfect simulation of Markov Chain stationary distribution

ψ and ψ^2 are two software implementing perfect simulation of Markov Chain stationary distributions using the coupling from the past technique. ψ starts from the transition kernel to derive the simulation program while ψ^2 uses a monotone constructive definition of a Markov chain. They are available at http://www-id.imag.fr/Logiciels/psi/.

5.2.3. PEPS

The main objective of PEPS is to facilitate the solution of large discrete event systems, in situations where classical methods fail. PEPS may be applied to the modeling of computer systems, telecommunication systems, road traffic, or manufacturing systems. The software is available at http://www-id.imag.fr/Logiciels/peps/.

5.3. HyperAtlas

The Hyperatlas software has been jointly developed with LSR-IMAG in the framework of the ESPON European project part 3.1 and 3.2. It includes visualization and analysis of socio-economical data in Europe at Nuts 1, Nuts 2 or Nuts 3 level providing analysis of dependence and spatial interaction. This software is available for European partners at http://www-lsr.imag.fr/HyperCarte/.

6. New Results

6.1. Hybrid Systems

Participants: Bruno Gaujal, Brigitte Plateau, Florence Perronnin, Jean-Marc Vincent.

As explained in Section 3.1.4, some systems cannot be modeled with classical approaches due to their size and their dynamic. By mixing fluid models and discrete models, it is possible to alleviate the combinatorial explosion of such systems. This year, we have successfully used this approach in the two following settings.

6.1.1. Open-Loop Control of Stochastic Fluid Systems and Applications to Storage and Ruin Problems

This is a collaborative work with Landy Rabehasaina (LMC Laboratory).

We used recent results on admission control to solve problems coming from fluid flow models and from risk theory in [12]. More precisely we consider a stochastic process $Q(t)$ satisfying a linear differential equation or a stochastic differential equation, driven by a jump process which is modulated by a binary sequence. We prove multi-modular properties related to the process $Q(t)$ and show that the expectation of a functional of the jump process is minimized by a randomized bracket sequence, when that sequence has to satisfy a constraint on its Cesaro limit. Applications related to optimal strategies in storage systems models and ruin problems are given.

6.1.2. Perfect simulation of stochastic hybrid systems with an application to peer to peer systems participants

In [38], [11], we propose an algorithm for performing perfect simulation of a class of stochastic hybrid systems made of deterministic differential equations and random discrete jumps. We first define the class of hybrid systems at hand and exhibit two large scale peer-to-peer (P2P) applications that can be accurately modeled by such systems, namely Squirrel and KaZaA. We then show how to construct a simulation of such a stochastic hybrid system that provides perfect samples of its asymptotic behavior based on the extension to continuous state-space of coupling-from-the-past techniques introduced by Foss and Tweedie (1998) and using suitable envelope trajectories to tackle non-monotonicity. The approach is then applied to the two above systems, thereby illustrating the usefulness of the approach in practical cases.

This work has also led to the following perspectives that are currently work in progress in our group: one direction is to generalize the sup/inf approach to more general systems; another perspective is the analytical proof of the convergence of the discrete system to its fluid model limit (mean-field convergence). Finally, the need for analyzing other large-scale P2P systems using this kind of techniques led to the hiring of a new PhD student, Rémi Bertin.

6.2. Perfect Simulation

Participants: Bruno Gaujal, Brigitte Plateau, Florence Perronnin, Jean-Marc Vincent.

Perfect simulation enables one to compute samples distributed according to the stationary distribution of the Markov process with no bias. The following sections summarize the various new results obtained using this technique, or on this technique.

6.2.1. Perfect Simulation and Monotone Stochastic Bounds

In [41], we combine monotone bounds of Markov chains and coupling from the past to obtain an exact sampling of a strong stochastic bound of the steady-state distribution for a Markov chain. Stochastic bounds are sufficient to bound any positive increasing rewards on the steady-state, such as the loss rates and the average size or delay. We show the equivalence between st-monotonicity and event monotonicity when the state space is endowed with a total ordering, and we provide several algorithms to transform a system into a set of monotone events. As we deal with monotone systems, the coupling technique requires less computational efforts for one iteration. Numerical examples show that we can obtain very important speedups.

6.2.2. Perfect simulation of index based routing queuing networks

Monotonicity properties of index based routing queuing networks are established and developed in the perfect simulation framework psi. This has been applied in the context of grid scheduling to compare several scheduling heuristics.

6.2.3. Backward coupling for perfect simulation of free-choice nets

This is a collaborative work with Anne Bouillard (IRISA).

We show how to design a perfect sampling algorithm for stochastic Free-Choice Petri nets by backward coupling. For Markovian event graphs, the simulation time can be greatly reduced by using extremal initial states, namely blocking marking, although such nets do not exhibit any natural monotonicity property. Another approach for perfect simulation of non-Markovian event graphs is based on a (max,plus) representation of the system and the theory of (max,plus) stochastic systems. Next, we show how to extend this approach to one-bounded free choice nets to the expense of keeping all states. Finally, experimental runs show that the (max,plus) approach needs a larger simulation time than the Markovian approach.

6.2.4. Perfect simulation of non-monotone models

The approach proposed here generalizes what has been done in [11] to simulate non-monotone Markov chains. Its main advantage is that it does not need any preliminary assumption on the structure of the Markov chain. If $\mathcal{S}$ is a lattice, then the idea is to replace the classical transition function of a Markov chain, $X_{n+1} = \phi\left(X_n, U_{n+1}\right)$, by a new two-dimensional transition function computing *envelopes* of the trajectories of the chain, starting from a maximal and a minimal state M and m: $\mathcal{U}(M, m, u) \overset{\text{def}}{=} \sup_{m \leq s \leq M} \Phi(s, u)$ and $\mathcal{L}(M, m, u) \overset{\text{def}}{=} \inf_{m \leq s \leq M} \Phi(s, u)$.

In general, this approach may not gain over the general non-monotonous coupling techniques because of three problems:

- (P_1) The envelope may never couple,
- (P_2) even if they do, the coupling time may become prohibitively large,
- (P_3) the time needed to compute $\mathcal{U}(M, m, u)$ and $\mathcal{L}(M, m, u)$ might depend on the number of states between m and M which amounts to simulating all trajectories.

However, the envelopes can be used to simulate efficiently rather general classes of queuing networks, for example networks of N finite queues (of capacity C) with general index routing and batch arrivals (which break the monotonicity property). In that case, envelopes always couple w.p.1 (Problem (P_1)). Problem (P_2) is solved by using a partial split of the trajectories when the states reached by the lower and upper envelopes get close in a queue. Problem (P_3) is solved by constructing an algorithm computing ϕ with complexity $O(N \log(C))$.

Other examples are networks of N finite queues with negative customers and/or with fork and join nodes, which are not monotone.

6.3. Scheduling

Participants: Bruno Gaujal, Derrick Kondo, Arnaud Legrand, Corinne Touati, Jean-Marc Vincent.

6.3.1. Polling Systems

This is a collaborative work with Dinard Van der Laan (Vrije University) and Arie Hordijk (Leiden University).

In [9], we consider deterministic (both fluid and discrete) polling systems with N queues with infinite buffers, and we show how to compute the best polling sequence (minimizing the average total workload). With two queues, we show that the best polling sequence is always periodic when the system is stable and forms a regular sequence. The fraction of time spent by the server in the first queue is highly non continuous in the parameters of the system (arrival rate and service rate) and shows a fractal behavior. Moreover, convexity properties are shown and are used in a generalization of the computation of the optimal control policy (in open-loop) for the stochastic exponential case.

6.3.2. Optimal routing for end-to-end guarantees: the price of multiplexing

We show in [20] how Network Calculus can be used to compute the optimal route for a flow (w.r.t. end-to-end guarantees on the delay or the backlog) in a network in the presence of cross-traffic. When cross-traffic is independent, the computation is shown to boil down to a functional shortest path problem. When cross-traffic perturbs the main flow over more than one node, then the "Pay Multiplexing Only Once" phenomenon makes the computation more involved. We provide an efficient algorithm to compute the service curve available for the main flow and show how to adapt the shortest path algorithm in this case. This paper got the best paper award at the Valuetools conference.

6.3.3. Dynamic Voltage Scaling under EDF Revisited

This is a collaborative work with Nicolas Navet (INRIA-LORIA).

Basic algorithms have been proposed in the field of low-power , which compute the minimum energy-schedule for a set of non-recurrent tasks (or jobs) scheduled under EDF on a dynamically variable voltage processor. In [10], we propose improvements upon existing algorithms with lower average and worst-case complexities. They are based on a new EDF feasibility test that helps to identify the "critical intervals". The complexity of this feasibility test depends on structural characteristics of the set of jobs. More precisely, it depends on how tasks are included one in the other. The first step of the algorithm is to construct the Hasse diagram of the set of tasks where the partial order is defined by the inclusion relation on the tasks. Then, the algorithm constructs the shortest path in a geometrical representation at each level of the Hasse diagram. The optimal processor speed is chosen according to the maximal slope of each path.

6.3.4. Index Routing Policies for Grids

This is a collaborative work with Vandy Berten (Université Libre de Bruxelles).

We show in [19], [8] how dynamic brokering for batch allocation in grids based on bi-dimensional indices can be used in practice in computational grids, with or without knowing the sizes of the jobs. We provide a fast algorithm (with a quadratic complexity) which can be used off-line or even on-line to compute the index tables. We also report numerous simulations providing numerical evidence of the great efficiency of our index routing policy as well as its robustness with respect to parameter changes. The value of information is also assessed by comparing the performance of indexes when the sizes of the jobs are known and when they are not known.

6.3.5. Balanced structures

In [37], we study reduction on finite or periodic Sturmian words which are balanced structures in dimension 1. Links between reductions, usual morphisms and the Nearest Integer Continued Fraction expansions are exhibited. This leads to special factorizations of mechanical words used to check in linear time if a word is mechanical. One natural way to generalize balanced structures in dimension 1 is to consider trees of complexity $n + 1$ but in doing so, one looses many optimization properties.

In [36] we study non-planar trees. In this case, we show that strongly balanced trees are exactly mechanical trees. Moreover we will see that they are also aperiodic trees of minimal complexity, so that the three equivalent definitions of Sturmian words are almost preserved for non planar trees.

6.3.6. Divisible Workload Scheduling

This is collaborative work with Yang Yang (University of California at San Diego), Henri Casanova (University of Hawaii at Manoa), Maciej Drozdowski (Poznan University of Technology)

The work in [58] is geared towards divisible loads, which are computations that can be arbitrarily divided into independent "chunks". Each of these chunks can then be processed in parallel. We study master-worker scheduling of divisible loads in heterogeneous distributed systems. In multi-round scheduling, load is sent to each worker as several chunks rather than as a single one. Solving the divisible load scheduling (DLS) problem entails determining the subset of workers that should be used, the sequence of communication to these workers, and the sizes of each load chunk. We first state and establish an optimality principle in the general case. Then we establish a new complexity result by showing that a DLS problem, whose complexity has been open for a long time, is in fact NP-hard, even in the one-round case. We also show that this problem is pseudo-polynomially solvable under certain special conditions. Finally, we present a deep survey on algorithms and heuristics for solving the multi-round DLS problem.

6.3.7. Multi-Job Scheduling in a Volunteer Computing System

This is collaborative work with David P. Anderson (University of California at Berkeley).

In [45], we evaluate the multi-job scheduler of a real-world desktop grid system, namely the Berkeley Open Infrastructure for Network Computing (BOINC). BOINC, a middleware system for desktop grids, allows hosts to be attached to multiple projects. Each host periodically requests jobs from project servers and executes the jobs. This process involves three interrelated policies:

- of the runnable jobs on a host, which to execute?
- when and from what project should a host request more work?
- what jobs should a server send in response to a given request?
- how to estimate the remaining runtime of a job?

We consider several alternatives for each of these policies. Using simulation, we study various combinations of policies, comparing them on the basis of several performance metrics and over a range of parameters such as job length variability, deadline slack, and number of attached projects.

6.3.8. Resource Selection in Desktop Grids

Desktop Grids are popular platforms for high throughput applications, but due to their inherent resource volatility it is difficult to exploit them for applications that require rapid turnaround. Efficient desktop Grid execution of short-lived applications is an attractive proposition and we claim in [13] that it is achievable via intelligent resource selection. We propose three general techniques for resource selection: resource prioritization, resource exclusion, and task duplication. We use these techniques to instantiate several scheduling heuristics. We evaluate these heuristics through trace-driven simulations of four representative desktop Grid configurations. We find that ranking desktop resources according to their clock rates, without taking into account their availability history, is surprisingly effective in practice. Our main result is that a heuristic that uses the appropriate combination of resource prioritization, resource exclusion, and task replication can achieve performance within a factor of 1.7 of optimal in practice.

6.3.9. Characterizing Error Rates in Internet Desktop Grids

Desktop grids use the free resources in Intranet and Internet environments for large-scale computation and storage. While desktop grids offer a high return on investment, one critical issue is the validation of results returned by participating hosts. Several mechanisms for result validation have been previously proposed. However, the characterization of errors is poorly understood.

In [47], to study error rates, we implemented and deployed a desktop grid application across several thousand hosts distributed over the Internet. We then analyzed the results to give quantitative and empirical characterization of errors stemming from input or output (I/O) failures. We find that in practice, error rates are widespread across hosts but occur relatively infrequently. Moreover, we find that error rates tend to not be stationary over time nor correlated between hosts. In light of these characterization results, we evaluated state-of-the-art error detection mechanisms and describe the trade-offs for using each mechanism.

In [46], we conduct a benefit analysis of a mechanism for result validation that we proposed recently for the detection of errors in long-running applications. This mechanism is based on using the digest of intermediate checkpoints, and we show in theory and simulation that the relative benefit of this method compared to the state-of-the-art is as high as 45%.

6.4. Middleware and Experimental Testbeds

Participants: Olivier Richard, Yves Denneulin, Jean-François Méhaut.

6.4.1. Lightweight Emulation to Study Peer-to-Peer Systems

The current methods used to test and study peer-to-peer systems (namely modeling, simulation, or execution on real testbeds) often show limits regarding scalability, realism and accuracy. In [17] we design and evaluate P2PLab, a framework to study peer-to-peer systems by combining emulation (use of the real studied application within a configured synthetic environment) and virtualization. P2PLab is scalable (it uses a distributed network model) and has good virtualization characteristics (many virtual nodes can be executed on the same physical node by using process level virtualization). Experiments with the BitTorrent file sharing system complete this work and demonstrate the usefulness of this platform.

6.4.2. Grid'5000: a large scale and highly reconfigurable experimental Grid testbed.

Large scale distributed systems like Grids are difficult to study only from theoretical models and simulators. Most Grids deployed at large scale are production platforms that are inappropriate research tools because of their limited reconfiguration, control and monitoring capabilities. Grid5000 is a 5000 CPUs nation-wide infrastructure for research in Grid computing. Grid5000 is designed to provide a scientific tool for computer scientists similar to the large-scale instruments used by physicists, astronomers, and biologists. In [39]We describe the motivations, design considerations, architecture, control, and monitoring infrastructure of this experimental platform. We present configuration examples and performance results for the reconfiguration subsystem.

6.4.3. Experiment Engine for Lightweight Grids

In [52], we present a case study conducted on the Grid'5000 platform, a lightweight grid. The goal was to make a rather simple experiment, and study how difficult it was to carry out correctly. This means it had to be correct, reproducible and efficient.

The work shows that despite the precautions taken, many parameters that could have an effect on the result were at first overlooked. It also shows that benchmarking plays a key role on making an experiment correct and reproducible. The process is in the end extremely tedious, and stresses the need for new tools to help users. We have thus presented a methodology to get correct results on grid architecture, to identify relevant problems and to propose an infrastructure that answers part of the problems encountered during experiments. Additionally, pieces of this infrastructure have been built.

6.4.4. Data grid middleware

We have created Gedeon: a data management middleware specifically designed for scientific data management on grids. Gedeon comprises a low-level I/O library, a remote access broker and various data access interfaces. Gedeon relies on a hierarchy of caches of data and requests to provide efficient data-access. The performance evaluation of this approach has been presented in [54]. Bio-informatics applications and microscope image databases are among the set of targeted applications.

6.4.5. Multi-Scale Scheduling for Electromagnetic Problems

The goal of this problem is to build a task scheduler for grids architecture that aims at matching the various scales of electromagnetism problems on the architectural scales on a grid. In [43], [44] we introduce an analysis of the electromagnetic problems that suits a scale decomposition and is so the first step towards building this scheduler. This work has been done within the MEG ACI (see Section 8.2.11).

6.4.6. High Performances I/O

Distributed applications, especially the ones being I/O intensive, often access the storage subsystem in a non-sequential way (stride requests). Since such behaviors lower the overall system performance, many applications use parallel I/O libraries such as ROMIO to gather and reorder requests. In the meantime, as cluster usage grows, several applications are often executed concurrently, competing for access to storage subsystems and, thus, potentially canceling optimizations brought by Parallel I/O libraries. The aIOLi project aims at optimizing the I/O accesses within the cluster and providing a simple POSIX API. We present an extension of aIOLi to address the issue of disjoint accesses generated by different concurrent applications in a cluster. In such a context, good trade-off has to be assessed between performance, fairness and response time. To achieve this, an I/O scheduling algorithm together with a *requests aggregator* that considers both application access patterns and global system load, have been designed and merged into aIOLi. This improvement led to the implementation of a new generic framework pluggable into any I/O file system layer. A test composed of two concurrent IOR benchmarks has shown improvements on read accesses by a factor ranging from 3.5 to 35 with POSIX calls and from 3.3 to 5 with ROMIO, both reference benchmarks have been executed on a traditional NFS server without any additional optimizations.

The leveraging of existing storage space in a cluster is a desirable characteristic of a parallel file system. While undoubtedly an advantage from the point of view of resource management, this possibility may face the administrator with a wide variety of alternatives for configuring the file server, whose optimal layout is not always easy to devise. Given the diversity of parameters such as the number of processors on each node and the capacity and topology of the network, decisions regarding the locality of server components like metadata servers and I/O servers have a direct impact on performance and scalability. We explore the capabilities of the dNFSp file system on a large cluster installation, observing how scalable the system behaves in different scenarios and comparing it to a dedicated parallel file system. Our obtained results show that the design of dNFSp allows for a scalable and resource-saving configuration for clusters with a large number of nodes.

In [40] we further developed this experiment by measuring the ratio between metadata and data transfer in order to better quantify their respective role and the care that should be taken in their management.

6.5. Tools for Performance Evaluation

Participants: Jean-François Méhaut, Brigitte Plateau, Jean-Marc Vincent.

6.5.1. PSI2: a Software Tool for the Perfect Simulation of Finite Queuing Networks

Markovian networks of finite capacity queues are widely used models for performance evaluation of systems and networks. Unfortunately, except in some specific situations, these models are not tractable analytically. Approximation techniques, aggregation, fluidification have been proposed to capture the behavior of such systems. But, in most complex cases, simulation remains the only tool to estimate the steady state of the system.

Classical simulation iterates from an initial state and estimates the steady-state on a long run trajectory via the ergodic theorem. The first problem of this direct simulation is the simulation control of the burn-in time period that ensures that the process have reached the steady state. The second difficulty is related to the auto-correlation of a one trajectory sample. Approximations or regenerative arguments should be used to compute confidence intervals. For large state-space systems regeneration arguments fail and should be adapted with a high knowledge on the system.

Perfect simulation provides a new technique to sample steady-state and avoids the burn-in time period. When the simulation algorithm stops, the returned state value is in steady-state. Initiated by Propp and Wilson in the context of statistical physics, this technique is based on a coupling from the past scheme that, provided some conditions on the system, ensures convergence in a finite time to steady-state. This approach have been successfully applied in various domains, stochastic geometry, interacting particle systems, statistical physics, networking, and etc.

We applied this technique first to Markov chain with sparse transition matrix, and to queuing networks with finite capacities and complex routing strategies. In [53], we describe software that we have developed to validate this simulation approach in the context of low probability event estimation. This software is available freely at http://psi.gforge.inria.fr/

6.5.2. Split: a flexible and efficient algorithm to vector-descriptor product

Many Markovian stochastic structured modeling formalisms like Petri nets, automata networks and process algebra represent the infinitesimal generator of the underlying Markov chain as a descriptor instead of a traditional sparse matrix. A descriptor is a compact and structured storage based on a sum of tensor

(Kronecker) products of small matrices that can be handled by many algorithms allowing affordable stationary and transient solutions even for very large Markovian models. One of the most efficient algorithms used to compute iterative solutions of descriptors is the Shuffle algorithm which is used to perform the multiplication by a probability vector. In [25] we propose an alternative algorithm called Split, since it offers a flexible solution between the pure sparse matrix approach and the Shuffle algorithm using a hybrid solution. The Split algorithm puts the Shuffle approach in perspective by presenting a faster execution time for many cases and at least the same efficiency for the worst cases. The Split algorithm is applied to solve two SAN models based on real problems.

6.5.3. Stochastic Automata Networks

With excellent cost/performance trade-off and good scalability, multiprocessor systems are becoming attractive alternatives when high performance, reliability and availability are needed. They are now more popular in universities, research labs and industries. In these communities, life-critical applications requiring high degrees of precision and performance are executed and controlled. Thus, it is important to the developers of such applications to analyze during the design phase how hardware, software and performance related failures affect the quality of service delivered to the users. This analysis can be conducted using modeling techniques such as transition systems. However, the high complexity of such systems (large state space) makes them difficult to analyze.

In [21], we have presented an efficient tool (PEPS2007) to model and analyze multiprocessor system in a structured and compact manner using a Stochastic Automata Network (SAN). A SAN is a high-level formalism for modeling very large and complex Markov chains. The formalism permits complete systems to be represented as a collection of interacting sub-systems. The basic concept which renders SAN powerful is the use of tensor algebra for its representation and analysis. Furthermore, a new modeling alternative has been recently incorporated into SANs: the use of phase-type distributions, which remains a desirable objective for the more accurately modeling of numerous real phenomena such as the repair and service time in multiprocessor systems.

6.5.4. Product form for Stochastic Automata Networks

We consider in [34] Stochastic Automata Networks (SAN) in continuous time and we prove a sufficient condition for the steady-state distribution to have product form. We consider SAN without synchronizations where the transitions of one automaton may depend of the states of the other automata. Even with this restriction, this sufficient condition is quite simple and this theorem generalizes former results on SAN but also on modulated Markovian queues, such as the Boucherie's theory on competing Markov chain, or on reversible queues considered by Kelly. The sufficient condition and the proof are purely algebraic.

6.6. Network Measurements and Models

Participants: Yves Denneulin, Derrick Kondo, Arnaud Legrand, Jean-François Méhaut, Olivier Richard.

6.6.1. Resources availability for Peer to Peer systems

In [14], we present application-level traces of four real desktop grids that can be used for simulation and modeling purposes. In addition, we describe aggregate and per host statistics that reflect the heterogeneity and volatility of desktop grid resources. Finally, we apply our characterization to develop a performance model for desktop grid applications for various task granularities, and then use a cluster equivalence metric to quantify the utility of the desktop grid relative to that of a dedicated cluster for task-parallel applications.

6.6.2. SAN Communication Modeling

SMP clusters are one of the most common HPC platform used by scientific applications. The nodes of SMP cluster contain several computing elements. Scientific applications may be executed over a large number of such nodes introducing complex communication behaviors. Using for instance MPI, communications on a same node with a common interval time create concurrent accesses to resources of nodes. On SMP nodes, concurrent access implies resource sharing depending on the underlying network architecture and the MPI implementation used. In [15], [51], we present a model to predict communication times of simultaneous MPI communications over SMP clusters. This model considers the concurrency over resources of nodes and network predicting accurately communication time for many communications in conflict.

6.6.3. *High Bandwidth Communication Models*

The performance of the intensive communications in clusters and grids is critical during the execution time of the parallel programs. In [18] we try to characterize the behavior of communication links in order to improve bandwidth use and, so, performances. We have detected anomalies during test on the Grid 5000 infrastructure as losses of bandwidth. We have done tests and, based on the results, proposed a model for bandwidth lost in intensive data transfer on high scalable architectures.

6.6.4. *Application-level Network Mapping*

This is a collaborative work with Frédéric Vivien (GRAAL project-team), Lionel Eyraud-Dubois (SCALAPPLIX project-team) and Martin Quinson (Algorille project-team) and is part of the ANR ALPAGE project.

To fully harness Grids, users or middleware must have some knowledge on the topology of the platform interconnection network. As such knowledge is usually not available, one must uses tools which automatically build a topological network model through some measurements. In [28], we have defined a methodology to assess the quality of these network model building tools, and we have applied this methodology to representatives of the main classes of model builders and to two new algorithms. We have shown that none of the main existing techniques build models that enable to accurately predict the running time of simple application kernels for actual platforms. However some of the new algorithms we have proposed give excellent results in a wide range of situations.

6.7. Game Theory

Participants: Arnaud Legrand, Corinne Touati.

6.7.1. *Pareto Efficiency Measure*

In the context of applied game theory in networking environments, a number of concepts have been introduced to measure both efficiency and optimality of resource allocations, the most famous certainly being the price of anarchy and the Jain index. We compared in [48], [55] these different measures and showed their limitations in terms of analysis of the efficiency of an equilibria. We also proposed a novel measure of efficiency, defined as the distance to the Pareto border (the optimality criterion in multi-users optimization) and named it the SDF (Selfish Degradation Factor). This measure is coherent with the notion of ϵ-approximation proposed by Papadimitriou and Yannanakis and is thus a natural extension of the classical approximation ratio concept to the multi-criteria setting.

6.7.2. *Price Wars*

In [27], we consider the problem where competing providers offer queued services with QoS guarantees. Users have quasi-linear utility functions that depends on delay and the price paid. We first analyze the equilibrium market shares of various providers. We assume that each provider has a service queue (such as an M/M/1 FIFO or an M/G/1 processor sharing) with a given capacity. Each provider offers delay guarantees and announces prices. Its objective is to maximize its revenue. We first observe that a competitive equilibrium does not exist in such a market in general. However, we show that the (price, delay) pair offered by a provider is unique. We then model the interaction between the providers as a Stackelberg game. Our analysis seems to indicate non-existence of an equilibrium of this game. In fact, we observe a *price-war* phenomenon also seen in many other contexts. However, we conjecture that the allocations that the providers offer in the Stackelberg game model, are in the core of the competitive exchange economy.

6.7.3. *Fair Steady-State Scheduling on Large-Scale Distributed Systems*

This is a collaborative work with Yves Robert (GRAAL project-team), Olivier Beaumont (SCALAPPLIX project-team) and Larry Carter and Jeanne Ferrante (University of California San Diego).

Many applications (cellular micro-physiology, protein conformations, particle detection or others) are constituted of a very large set of independent, equal-sized tasks. In [7], we study the situation where a small number of users each having a large number of tasks compete for resources.

We consider the problem of scheduling applications to ensure fair and efficient execution on a distributed network of processors. We limit our study to the case where communication is restricted to a tree embedded in the network, and the applications consist of a large number of independent tasks (Bags of Tasks) that originate at the tree's root. The tasks of a given application all have the same computation and communication requirements, but these requirements can vary for different applications. The goal of scheduling is to maximize throughput of each application while ensuring a fair sharing of resources between applications.

We can find the optimal asymptotic rates by solving a linear programming problem that expresses all necessary problem constraints, and we show how to construct a periodic schedule from any linear program solution. For single-level trees, the solution is characterized by processing tasks with larger communication-to-computation ratios at children with larger bandwidths. For multi-level trees, this approach requires global knowledge of all application and platform parameters. For large-scale platforms, such global coordination by a centralized scheduler may be unrealistic. Thus, we also investigate decentralized schedulers that use only local information at each participating resource. We assess their performance via simulation, and compare to an optimal centralized solution obtained via linear programming.

The best of our decentralized heuristics achieves the same performance on about two-thirds of our test cases, but is far worse in a few cases. While our results are based on simple assumptions and do not explore all parameters (such as the maximum number of tasks that can be held on a node), they provide insight into the important question of fairly and optimally scheduling heterogeneous applications on heterogeneous grids. In particular, this study suggests that max-min fairness is probably not suited to this context. Indeed, such a fairness seems hard to obtain in a fully distributed way. Moreover, as applications may have very different characteristics, a small decrease in the throughput of a single application may enable to greatly increase the throughput of others, hence a need for a less strict fairness objective.

6.7.4. Non-Cooperative Scheduling

We have also studied the situation where multiple applications execute concurrently on an heterogeneous platforms competing for CPU and network resources. In [49] we analyze the behavior of K non-cooperative schedulers using the optimal strategy that maximize their efficiency. Meanwhile fairness is ensured at a system level ignoring applications characteristics. We limit our study to simple single-level master-worker platforms and the case where applications consist of a large number of independent tasks. The tasks of a given application all have the same computation and communication requirements, but these requirements can vary from one application to another. Therefore, each scheduler aims at maximizing its throughput. We give closed-form formula of the equilibrium reached by such a system and study its performances. We characterize the situations where this Nash equilibrium is Pareto-optimal and show that even though no catastrophic situation (Braess-like paradox) can occur, such an equilibrium can be arbitrarily bad for any classical performance measure.

6.8. On-demand Geographical Maps

Participants: Jean-Marc Vincent, Saïd Oulahal.

The new results regarding on-demand geographical maps are threefold.

- The Hyperatlas software has been presented in a plenary session of the European Union parliament (département thématique: politique structurelle et de cohésion). The software has been distributed to the members of the parliament on a CD.
- The software environment has been adapted to environmental data with Environment European Agency and we are now able to mix social economical data with environmental data.
- The potential methods have been developed in the HyperSmooth software and applied in the European ESPON project [50]

7. Contracts and Grants with Industry

7.1. CIFRE with BULL, 06-09

Yiannis Georgiou is doing his PhD thesis in a CIFRE contract with the BULL company. His work started in September 2006, and he will finish in September 2009. The focus of his research is batch scheduling on Grids.

7.2. CIFRE with France Télécom R&D, 06-09

Ahmed Harbaoui is doing his PhD thesis in a CIFRE contract with the France Télécom R&D company. His work started in September 2006, and he will finish in September 2009. He is interested in load injection and performance evaluation issues in networks.

7.3. CIFRE with STMicroelectronics, 06-10

Carlos Rojas is doing his PhD thesis under a CIFRE contract with STMicroelectronics. He started in September 2007 and will finish in September 2010. The objective of his thesis is to develop methods and tools for multiprocessor embedded applications.

7.4. Sceptre with STMicroelectronics, (Divisions STS and HEG), INRIA Rhône-Alpes (MOAIS, Mescal, Arenaire, CompSys), TIMA/SLS, Verimag, CAPS-Entreprise and IRISA (CAPS) 06-10

Sceptre is a minalogic project, supported by the Pole de Competitivite Minalogic. Global competitiveness cluster Minalogic fosters research-led innovation in intelligent miniaturized products and solutions for industry. Located in Grenoble, France, the cluster channels in a single physical location a range of highly-specialized skills and resources from knowledge creation to the development and production of intelligent miniaturized services for industry. Sceptre main objective is to provide SoC implementation techniques, using novel approaches originating from both multiprocessor programming and reconfigurable processors. The application domain is distributed multimedia code optimization.

Our work is focused on tools and methods to develop embedded systems. The main working directions are software and hardware integration, scalable and configurable architectures, real time constraints, heterogeneous multiprocessing, and load-balancing.

7.5. Real-Time-At -Work

RealTimeAtWork.com is a startup from INRIA Lorraine created in December 2007. Some members of Mescal are scientific partners in the startup. Its main target is to provide software tools for solving real time constraints in embedded systems, particularly for superposition of periodic flows. Such flows are typical in automotive and avionics industries who are the privileged potential users of the technologies developed by RealTimeAtWork.com

7.6. CILEO with BULL, Compagnie des Signaux, TIMA, CEA-LETI, LIG, Edxact, Infiniscale, Probayes, SCelectronique, 06-10

The increasingly miniaturization of components and the ever-increasing complexity of electronic circuits for communication systems requires a set of sophisticated tools for design and simulation. These tools in turn often require immense computational resources, sometimes more than several orders of magnitude above the performance of a desktop PC or a workstation. These tools are so compute-intensive that they require supercomputers, clusters and grids. However, these types of computing resources are often not within the reach of PME's (relatively small companies or startups) in the semiconductor industry and sometimes even large companies, not only because of the cost of infrastructure, but also because of the lack of adequate methods and technologies for high performance computing.

In the association of Minalogic, there are about twenty PME's that develop CAD software, and other companies in the field of embedded systems, the design of electronic circuits, and the simulation process. The most advanced companies utilize high performance computing, and the others will have to do so in 2 or 3 years. All of these companies are confronted with a notable lack of services and facilities for intensive computing, which heavily affect their competitiveness and speed of development.

It is in this context that the partners of this CILOE project propose to design and develop a complete computational infrastructure, including methodologies, software, and security mechanisms. This infrastructure will contribute decisively to the development and visibility of the international PME partners in the project. It will be an essential tool for a sustainable boost in the sector of electronic CAD, embedded software and high-performance simulation and moreover, facilitate growth for all companies in the electronics industry in Alpes region.

This project has three main objectives that will allow industry to leverage large-scale compute-intensive platforms:

- Reduce the delay in the development of reliable software of the industry partners (Jivaro for Edxact, ProBayes-BT for Probayes, Stressio for SC Online). The validation of software improvements requires numerous test cases of modest size but also test cases of much larger size. For example, the biggest test case (15 GB approximately) for the software Jivaro of the company Edxact requires computation on the order of days. Often, the long duration of these computations can delay the validation of software. The goal here is to improve the competitiveness of local companies so that they can provide more quickly new versions of their software that has been completely validated in a number of tests.
- Develop highly parallel versions of software of the PME/PMI partners. The targeted architectures here are clusters of multi-core machines and specialized processors (system-on-a-chip multiprocessors, NoC-; Cell). This technological gain for business partners (Edxact, ProBayes) will enhance their competitiveness.
- Experiment with services for enabling resource access by applications. This would be based on the principles of IaaS (Infrastructure as a Service) and SaaS (Software as a Service). In the models of IaaS and SaaS, customers of the PME partners do not have to pay for the construction and maintenance of the entire infrastructure and software licenses. Instead, the customers only pay for their direct use. Once the infrastructure and services are deployed, customer access is enabled through a simple Web interface, which will allow PME's to cheaply target a global market.

8. Other Grants and Activities

8.1. Regional initiatives

8.1.1. CIMENT

The CIMENT project (Intensive Computing, Numerical Modeling and Technical Experiments, http://ciment.ujf-grenoble.fr/) gathers a wide scientific community involved in numerical modeling and computing (from numerical physics and chemistry to astrophysics, mechanics, bio-modeling and imaging) and the distributed computer science teams from Grenoble. Several heterogeneous distributed computing platforms were set up (from PC clusters to IBM SP or alpha workstations) each being originally dedicated to a scientific domain. More than 600 processors are available for scientific computation. The MESCAL project-team provides expert skills in high performance computing infrastructures.

8.1.2. Grappe200 project

MENRT-UJF-INPG, Rhône-Alpes Region, INRIA , ENS-Lyon have funded a cluster composed of 110 bi-processors Itanium2 connected with a Myrinet (donation of MyriCom) high performance network. This project is lead by MESCAL, MOAIS, ReMaP and SARDES. It is part of the CIMENT project which aims at building high performance distributed grids between several research labs (see above).

8.1.3. Cluster Région

The MESCAL project-team is a member of the regional “cluster” project on computer science and applied mathematics, the focus of its participation is on handling large amount of data large scale architecture. Other members of this subproject are the INRIA GRAAL project-team, the LSR-IMAG and IN2P3-LAPP laboratories.

8.2. National initiatives

8.2.1. GEDEON, 04-07, ACI Masse de Données

Partners (IMAG-LSR).

File systems (FS) are commonly used to store data. Especially, they are intensively used in the community of large scientific computing (astronomy, biology, weather prediction) which needs the storage of large amounts of data in a distributed manner. In a GRID context (cluster of clusters), traditional distributed file systems have been adapted to manage a large number of hosts (like the Andrew File System). However, such file systems remain inadequate to manage huge data. They are suited for traditional Unix (small) files. Thus, the grain of distribution is typically an entire file and not a piece of file which is essential for large files. Furthermore, the tools for managing data (e.g, interrogation, duplication, consistency) are unsuited for large sizes.

Database Management Systems (DBMS) provides different abstraction layers, high level languages for data interrogation and manipulation etc. However, the imposed data structure, the low distribution, and the usually monolithic architecture of DBMSs limit their utilization in the scientific computing context.

The main idea of the Gedeon project is to merge the functions of file systems and DBMS, focusing on structuring of meta-data, duplication and coherency control. Our goal is NOT to build a DBMS describing a set of files. We will study how database management services can be used to improve the efficiency of file access and to increase the functionality provided to scientific programmers.

8.2.2. GRID 5000, 04-07, ACI GRID

Partners (INRIA FUTURS, INRIA Sophia, IRISA, LORIA, IRIT, LABRI, LIP, LIFL).

The foundations of Grid’5000 have emerged from a thorough analysis and numerous discussions about methodologies used for scientific research in the Grid domain. A report presents the rationale for Grid’5000. In addition to theory, simulators and emulators, there is a strong need for large scale testbeds where real life experimental conditions hold. The size of Grid’5000, in terms of number of sites and number of CPUs per site, was established according to the scale of the experiments and the number of researchers involved in the project.

8.2.3. DSLLab, 2005-2008, ANR Jeunes Chercheurs

Partners (INRIA-FUTURS).

DSLlab is a research project aiming at building and using an experimental platform about distributed systems running on DSL Internet. The objective is twofold:

- provide accurate and customized measures of availability, activity and performances in order to characterize and tune the models of the ASDL resources;
- provide a validation and experimental tool for new protocols, services and simulators and emulators for these systems.

DSLlab consists of a set of low power, low noise computers spread over the ASDL. These computers are used simultaneously as active probes to capture the behavior traces, and as operational nodes to launch experiments. We expect from this experiment a better knowledge of the behavior of the ASDL and the design of accurate models for emulation and simulation of these systems, which represents now a significant capability in terms of storage and computing power.

8.2.4. NUMASIS, 2005-2008, ANR Calcul Intensif et Grilles de Calcul

Future generations of multiprocessors machines will rely on a NUMA architecture featuring multiple memory levels as well as nested computing units (multi-core chips, multi-threaded processors, multi-modules NUMA, etc.). To achieve most of the hardware's performance, parallel applications need powerful software to carefully distribute processes and data so as to limit non-local memory accesses. The ANR NUMASIS[2] project aims at evaluating the functionalities provided by current operating systems and middleware in order to point out their limitations. It also aims at designing new methods and mechanisms for an efficient scheduling of processes and a clever data distribution on such platforms. These mechanisms will be implemented within operating systems and middleware. The target application domain is seismology, which is very representative of the needs of computer-intensive scientific applications. Two technical reports on this subject [57], [56] have been produced for this ANR.

8.2.5. ALPAGE, 2005-2008, ANR Masses de Données

The new algorithmic challenges associated with large-scale platforms have been approached from two different directions. On the one hand, the parallel algorithms community has largely concentrated on the problems associated with heterogeneity and large amounts of data. Algorithms have been based on a centralized single-node, responsible for calculating the optimal solution; this approach induces significant computing times on the organizing node, and requires centralizing all the information about the platform. Therefore, these solutions clearly suffer from scalability and fault tolerance problems.

On the other hand, the distributed systems community has focused on scalability and fault-tolerance issues. The success of file sharing applications demonstrates the capacity of the resulting algorithms to manage huge volumes of data and users on large unstable platforms. Algorithms developed within this context are completely distributed and based on peer-to-peer communications. They are well adapted to very irregular applications, for which the communication pattern is unpredictable. But in the case of more regular applications, they lead to a significant waste of resources.

The goal of the ALPAGE project is to establish a link between these directions, by gathering researchers (Mescal, LIP, LORIA, LaBRI, LIX, LRI) from the distributed systems and parallel algorithms communities. More precisely, the objective is to develop efficient and robust algorithms for some elementary applications, such as broadcast and multicast, distribution of tasks that may or may not share files, resource discovery. These fundamental applications correspond well to the spectrum of the applications that can be considered on large scale, distributed platforms.

8.2.6. SMS, 2005-2008, ANR

The ACI SMS, "Simulation et Monotonie Stochastique en évaluation de performances", is composed by two teams: Performance Evaluation team from PRiSM Laboratory (ACI Leader) and the MESCAL project-team. The main objective is to study monotonicity properties of computer systems models in order to speed up the simulations and estimate performance indexes more accurately.

The composition formalisms we have contributed to develop during the recent years allow to build large Markov chains associated to complex systems in order to analyze their performance. However, it is often impossible to solve the stationary or transient distributions. Analytical methods and simulations fail for different reasons.

However brute performances are not really useful. We need the proof that the system is better than an objective. Therefore it is natural to use comparison of random variables and sample-paths. Two important concepts appear: stochastic ordering and stochastic monotony. We chose to develop these two important concepts and apply them to perfect simulation, distributed simulation and product form queuing network. These concepts seem to appear frequently in various techniques in performance evaluation. Using the monotony property, one can reduce the computation time for perfect simulation with coupling from the past. Coupling from the past

[2] NUMASIS: Adapting and Optimizing Applicative Performance on NUMA Architectures: Design and Implementation with Applications in Seismology

allows to sample the steady-state distribution in a finite time. Thus we do not encounter the same stopping problem that holds for ordinary simulations. Furthermore, some results show that the monotony property is often present in queuing network even if they do not have product form. We simply have to renormalize them to let the property appear. Using both properties, it is also possible to derive distributed simulations which will be more efficient. We will develop two ideas: sample-path transformations to avoid rollback in optimistic simulations (and we compute a bound) and regenerative simulations.

Finally, these concepts can be used for product form queuing network to explain why some transformation applied on customer synchronization can provide product form solution, and also how we can compute a solution of the traffic equation when they are unstable.

8.2.7. COINC, 2006-2007, ARC

Partners (INRIA Sophia-Antipolis, INRIA Rhone-Alpes, INRIA Lorraine, LIP, University of Angers).

Most large distributed systems, such as embedded systems, systems on chip, Local Area Networks or sensor networks are meant to function under hard safety constraints. One may distinguish two main classes of constraints. The first set of constraints concerns the functional behavior of the system. The techniques used to make sure that the behavior of the system is correct range from formal proof theory and software verifications to testing tools. The second main set of constraints concerns the temporal behavior of the system. In a closed system where all events are controllable, these time guarantees are often given by scheduling techniques under real time constraints or model-checking on timed automata. However, most modern distributed embedded systems are more or less open systems where some exogenous non-controllable processes are involved. In this framework, one of the most promising techniques to provide deterministic guarantees is the Network Calculus. Having been developed in the last 10 years, the Network Calculus can be seen as a set of formulas and techniques that allow the designers of distributed systems to get certified bounds on response times as well as buffer capacities.

The objectives of this project are the following:

- To determine suitable classes of functions for which we will design algorithms implementing the main Network Calculus operations;
- To find ways to finitely describe these functions;
- To develop algorithms that estimate or bound the an arrival curve function, which in practice is infrequently known precisely;
- To investigate optimization problems in the context of the network calculus; and
- To design fast algorithms for different classes of discrete real functions (piecewise affine, ultimately periodic functions, convex/concave functions) using the appropriate data structures.

8.2.8. IFANY, 2006-2007, ARC

Partners (INRIA Sophia-Antipolis, INRIA Rocquencourt, INRIA ENS, France Telecom, IRISA, EURECOM, EPFL, University of Cyprus, University of Thessaly, GET/INT).

Information Theory (IT) is concerned with identifying, characterization and computation of basic limits of performance measures of communication systems. The classical question that it seeks to answer is what are the information rates in the multi-user setting that can be transmitted reliably over a given noisy channel. The transmission may involve various transmitter and various receivers. Rendering transmission over noisy channels reliable involves coding of the information to be transmitted. A given transmission rate vector C is said to be reliable if for any rate C’ < C and every epsilon > 0 there exists a code that allows transmission at the rate C’ with error rates smaller than epsilon. The set of reliable transmission rates C are called the capacity region.

We seek to create cooperation between specialists on classical information theory (IT) and other disciplines that are needed for the computations of new concepts of capacity that appear in the next generation networks in which the geometric aspects, mobility, power restriction and non-cooperative behavior play a key role on the capacity and connectivity. In view of the wealth of mathematical tools that are needed to study networks (and in particular SNs), the IFANY group focuses on inter-disciplinary research for networking dimensioning and design.

8.2.9. DOCCA, 2007-2011 ANR Jeunes Chercheurs

The race towards the design and development of scalable distributed systems offers new opportunities to applications, in particular as far as scientific computing, databases, and file sharing are concerned. Recently many advances have been done in the area of large-scale file-sharing systems, building upon the peer-to-peer paradigm that somehow seamlessly responds to the dynamicity and resilience issues. However, achieving a fair resource sharing amongst a large number of users in a distributed way is clearly still an open and active research field. For all previous issues there is a clear gap between

- widely deployed systems as peer-to-peer file-sharing systems (KaZaA, Gnutella, EDonkey) that are generally not very efficient and do not propose generic solutions that can be extended to other kind of usage;
- academic work with generally smart solutions (probabilistic routing in random graphs, set of node-disjoint trees, lagrangian optimization) that sometimes lack a real application.

Up until now, the main achievements based on the peer-to-peer paradigm mainly concern file- sharing issues. We believe that a large class of scientific computations could also take advantage of this kind of organization. Thus our goal is to design a peer-to-peer computing infrastructure with a particular emphasis on the fairness issues. In particular, the objectives of the ANR DOCCA[3] project are the following:

- to combine theoretical tools and metrics from the parallel computing community and from the network community, and to explore algorithmic and analytical solutions to the specific resource management problems of such systems.
- to design a P2P architecture based on the algorithms designed in the second step, and to create a novel P2P collaborative computing system.

We expect the following results from this project:

- to provide user synthetic models to the scientific community that can be used as an input in modeling, simulation and experimentation of P2P collaborative computing systems.
- to provide optimal strategies and resource management algorithms in P2P collaborative computing.
- to design a decentralized protocol that implements the optimal strategies for the target user models.
- to implement a prototype and validate the approach on an experimental platform.

8.2.10. Check-bound, 2007-2009 ANR SETIN

Partners (University of Paris I).

The increasing use of computerized systems in all aspects of our lives gives an increasing importance on the need for them to function correctly. The presence of such systems in safety-critical applications, coupled with their increasing complexity, makes indispensable their verification to see if they behaves as required . Thus the model checking which is the automated manner of formal verification techniques is of particular interest. Since verification techniques have become more efficient and more prevalent, it is natural to extend the range of models and specification formalisms to which model checking can be applied. Indeed the behavior of many real-life processes is inherently stochastic, thus the formalism has been extended to probabilistic model checking. Therefore, different formalisms in which the underlying system has been modeled by Markovian models have been proposed.

[3] Design and Optimization of Collaborative Computing Architectures

Stochastic model checking can be performed by numerical or statistical methods. In model checking formalism, models are checked to see if the considered measures are guaranteed or not. We apply Stochastic Comparison technique for numerical stochastic model checking. The main advantage of this approach is the possibility to derive transient and steady-state bounding distributions as well as the possibility to avoid the state-space explosion problem. For the statistical model checking we study the application of perfect simulation by coupling in the past. This method has been shown to be efficient when the underlying system is monotonous for the exact steady-state distribution sampling. We consider to extend this approach for transient analysis and to model checking by means of bounding models and the stochastic monotonicity. As one of the most difficult problems for the model checking formalism, we also study the case when the state space is infinite. In some cases, it would be possible to consider bounding models defined in finite state space.

8.2.11. ACI blanche MEG 2007-2010

The "ACI blanche” MEG, is composed of two teams: physicists working on electromagnetism from the LAAS (Toulouse) and the MESCAL project-team. The main objective is to study scaling properties in electromagnetism simulation applications and grids. The first results are promising. They demonstrate that the tools developed by Mescal on large data storage and middleware for deployment on clusters and grids are appropriate for that kind of application.

8.3. International Initiatives

8.3.1. Europe

CoreGrid: The project-team MESCAL participates in the CoreGrid Network Of Excellence.

EuroNGI : The project-team MESCAL participates in EuroNGI (Next Generation Internet).

ESPON : The MESCAL project-team participates to the ESPON (European Spatial Planning Observation Network) http://www.espon.lu/ It is involved in the action 3.1 on tools for analysis of socio-economical data. This work is done in the consortium hypercarte including the laboratories LSR-IMAG (UMR 5526), Géographie-cité (UMR 8504) and RIATE (UMS 2414). The Hyperatlas tools have been applied to the European context in order to study spatial deviation indexes on demographic and sociological data at nuts 3 level.

8.3.2. Africa

Cameroon : MESCAL takes part in the SARIMA[4] project and more precisely with the University of Yaoundé 1. Two Cameroon students (Jean-Michel NLong 2 and Blaise Yenké) are preparing their PhD in cotutelle (joint and remote supervision) with Professor Maurice Tchuenté. SARIMA also funded Adamou Hamza to prepare his Master Thesis during three months in the MESCAL project-team. SARIMA proposed J-F Méhaut to give a course on Operating System and Networks at Master Research Students.

8.3.3. South America

- DIODE (2006-2008) Associate Team funded by INRIA with the MOAIS project-team of INRIA, and the Brazilian University UFRGS. The goal of this project is to design and develop programming tools for grid and clusters for virtual reality. This collaboration was initiated 10 years ago, and has greatly affected the activities (doctoral, publications and joint production software) of the Apache project-team, from which MOAIS and MESCAL were formed. In particular, four PhD Brazilian students have joined the MESCAL project-team as a result of this long-standing collaboration. This year, 3 members of the MESCAL project-team visited Brazil (Olivier Richard for 1 week, Corinne Touati for 1 month, and Jean-Marc Vincent for 2 weeks) to enhance the existing collaborations and to form new ones.

[4] Soutien aux Activités de Recherche Informatique et Mathématiques en Afrique http://www-direction.inria.fr/international/AFRIQUE/sarima.html

- PICS (2005-2007) CADIGE funded by the CNRS with the universities of Rio Grande do Sul, Brazil (UFRGS, UFSM, PUC, UNISINOS), around PC clusters, grid and performance evaluation tools.
- CAPES/COFECUB grant (2006-2008) with the UFRGS, Porto Alegre, Brazil around grid and PC clusters.
- ECOS grant (2007-2009) Colombia: joint project with the universities of Los Andes, Bogota, and UIS, Bucaramanga, on the topic of grids for computation and data management.

8.4. High Performance Computing Center

8.4.1. The ICluster2 and IDPot Platforms

The MESCAL project-team manages a cluster computing center on the Grenoble campus. The center manages different architectures: a 48 bi-processors PC (ID-POT), and the center is involved with a cluster based on 110 bi-processors Itanium2 (ICluster-2) located at INRIA.

More than 60 research projects in France have used the architectures, especially the 204 processors Icluster-2. Half of them have run typical numerical applications on this machine, the remainder has worked on middleware and new technology for cluster and grid computing.

The ICluster2 and IDPot platforms are now integrated the Grid'5000 grid platform.

8.4.2. The BULL Machine

In the context of our collaboration with BULL (LIPS, NUMASIS), the MESCAL project-team acquired a Novascale NUMA machine. The configuration is based on 8 Itanium II processors at 1.5 Ghz and 16 GB of RAM. This platform is mainly used by the BULL PhD students. This machine is also connected to the CIMENT Grid.

8.4.3. GRID 5000 and CIMENT

The MESCAL project-team is involved in development and management of Grid'5000 platform. The ICluster2 and IDPot clusters are integrated in Grid'5000. Moreover, these two clusters take part in CIMENT Grid. More precisely, their unused resources may be exploited to execute jobs from partners of CIMENT project (see Section 8.1.1).

9. Dissemination

9.1. Leadership within the scientific community

9.1.1. Conference and Workshop Chairing

Researchers of the MESCAL project-team have been chairs of the following conferences or workshops:

- Euro-Par (Bruno Gaujal, local topic chair)
- Workshop on Programming Models for Grid Computing (Arnaud Legrand, co-chair)
- Workshop on Desktop Grids and Volunteer Computing (Derrick Kondo, co-chair)

9.1.2. Program committees

Researchers of the MESCAL project-team have been program committee members of the following conferences or workshops:

- ValueTools
- Europar
- MSR
- High Performance Computing for Computational Science
- Workshop on Programming Models for Grid Computing
- Workshop on Desktop Grids and Volunteer Computing
- International Workshop on Global and Peer-to-Peer Computing
- EuroPVM/MPI

9.1.3. Thesis defense

- Yves Denneulin defended his "thèse d'habilitation à diriger des recherches" on December 7th, 2007 in Grenoble. Thesis committee: Jacques Mossière, Luc Bougé, Ian Foster, Sacha Krakowiak, Franck Cappello, Jean-Marc Geib [5].
- Maxime Martinasso defended his PhD on May 25th, 2007 in Grenoble. Thesis committee: Jean-Claude Fernandez, Frédéric Desprez, Raymond Namyst, Arnaud Legrand, Pascale Rossé, Jean-François Méhaut [6].

9.1.4. Thesis committees

Researchers of the MESCAL project-team have served on the following thesis committees:

- Bruno Gaujal served on the thesis committee of Ana Bušić (Versailles University). He was also on the committees of Thomas Fernique (University of Montpellier) and Vandy Berten (Free University of Brussels) as a reporter.
- Jean-François Méhaut served on the thesis committee of Samuel Thibault (University of Bordeaux 1) as a reporter.
- Brigitte Plateau served as the head of the thesis committee for Gaelle Calvary (HDR, LIG) and Nicolas Peltier (HDR, LIG). She also served on the committee for Émmanuel Jeannot (HDR, INRIA LORIA).
- Arnaud Legrand served on the thesis committee of Maxime Martinasso (University of Grenoble) and Feryal Moulai (University of Grenoble).

9.1.5. Members of editorial board

Bruno Gaujal is an editor of the special issue of the Journal of Discrete Event Dynamic Systems on Valuetools.

9.1.6. PAGE: Probabilities and Applications in Grenoble and its surroundings

This seminar on probabilities and applications is targeted toward computer scientists as well as mathematicians. One of the goals is to encourage collaborations between people from different laboratories with varied backgrounds.

9.1.7. Grenoble's Seminar on performance evaluation

This seminar is organized by Jean-Marc Vincent and Bruno Gaujal. It is tightly coupled with the PAGE group and its main goal is to organize meetings between the various researchers of Grenoble using the same kind of mathematical tools (stochastic models, queuing networks, Petri networks, stochastic automata, Markovian process and chains, (max,+) algebra, fluid systems, ...). On the long term, this seminar should lead to inter-laboratory working groups on precise themes. More information is available at http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/EPG/.

9.2. Teaching

Members of the MESCAL team are actively involved in teaching. Their activities are balanced between graduate students and post-graduate students. Here are a few examples of their responsibilities:

- **2$^{\text{nd}}$ year of Research Master (Grenoble): Operating Systems and Software** head of the SAP track (operating systems, parallel and distributed applications, networks and multimedia). Here is a list of courses taught by researchers of the MESCAL project-team:
  - Cluster architectures for high-performance computing and high throughput data management.
  - Data measurement and analysis for network and operating systems performance evaluation.

  - Modeling and simulation for network and operating systems performance evaluation.
  - Building parallel and distributed applications (contributor).
  - Algorithms and basic techniques for parallel computing (contributor).
- **2nd year of Research Master (Yaoundé)** Operating systems and networks.
- **Magistère d'informatique Licence (Université Joseph Fourier)**

10. Bibliography

Major publications by the team in recent years

[1] E. ALTMAN, B. GAUJAL, A. HORDIJK. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity*, LNM, n^{o} 1829, Springer-Verlag, 2003.

[2] K. ATIF, B. PLATEAU. *Stochastic Automata Network for modeling parallel systems*, in "IEEE Transactions on Software Engineering", vol. 17, n^{o} 10, October 1991.

[3] B. GAUJAL, S. HAAR, J. MAIRESSE. *Blocking a Transition in a Free Choice Net, and what it tells about its throughput*, in "Journal of Computer and System Sciences", vol. 66, n^{o} 3, 2003, p. 515-548.

[4] J.-M. VINCENT. *Some Ergodic Results on Stochastic Iterative Discrete Event Systems*, in "Discrete Event Dynamic Systems", vol. 7, n^{o} 2, 1997, p. 209-232.

Year Publications

Doctoral dissertations and Habilitation theses

[5] Y. DENNEULIN. *Intergiciel à haute performance pour architecture grande échelle*, Habilitation à diriger des recherches, Ph. D. Thesis, INPG, Grenoble, December 2007.

[6] M. MARTINASSO. *Analyse et Modélisation des Communications Concurrentes dans les Réseaux Haute Performance*, CIFRE Bull, Ph. D. Thesis, Université Joseph Fourier, école doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique, Grenoble, May 2007.

Articles in refereed journals and book chapters

[7] O. BEAUMONT, L. CARTER, J. FERRANTE, A. LEGRAND, L. MARCHAL, Y. ROBERT. *Centralized Versus Distributed Schedulers Multiple Bag-of-Tasks Applications*, in "IEEE Trans. Parallel Distributed Systems", 2007.

[8] V. BERTEN, B. GAUJAL. *Brokering strategies in computational Grids using stochastic prediction models*, in "Parallel Computing", Special Issue on Large Scale Grids, 2007.

[9] B. GAUJAL, A. HORDIJK, D. V. DER LAAN. *On the Optimal Open-Loop Control Policy for Deterministic and Exponential Polling Systems*, in "Probability in Engineering and Informational Sciences", vol. 21, 2007, p. 157-187.

[10] B. GAUJAL, N. NAVET. *Dynamic Voltage Scaling under EDF Revisited*, in "Real Time Systems", vol. 31, n^{o} 1, 2007, p. 77-97.

[11] B. Gaujal, F. Perronnin, R. Bertin. *Perfect simulation of a class of stochastic hybrid systems with an application to peer to peer systems*, in "Journal of Discrete Event Dynamic Systems", Special Issue on Hybrid Systems, 2007.

[12] B. Gaujal, L. Rabehasaina. *Open-Loop Control of stochastic Fluid Systems and Applications*, in "Operations Research Letters", vol. 35, 2007, p. 455-462.

[13] D. Kondo, A. A. Chien, H. Casanova. *Scheduling Task Parallel Applications for Rapid Application Turnaround on Enterprise Desktop Grids*, in "Journal of Grid Computing", vol. 5, n^{o} 4, December 2007, p. 379–405.

[14] D. Kondo, G. Fedak, F. Cappello, A. A. Chien, H. Casanova. *Characterizing Resource Availability in Enterprise Desktop Grids*, in "Journal of Future Generation Computer Systems", vol. 23, n^{o} 7, August 2007, p. 888-903.

[15] M. Martinasso, J.-F. Méhaut. *Modèles de Communications Concurrentes sur des Grappes SMP*, in "Techniques et Sciences Informatiques (TSI)", 2007.

[16] J.-F. Méhaut. *Architectures Informatiques pour la Simulation en Géosciences*, in "Géosciences", Numéro spécial: Terre Virtuelle: les systèmes d'information géoscientifiques, n^{o} 6, October 2007, p. 54–61.

[17] L. Nussbaum, O. Richard. *Lightweight emulation to study peer-to-peer systems*, in "Concurrency and Computation: Practice and Experience", 2007.

Publications in Conferences and Workshops

[18] C. Barrios-Hernandez, Y. Denneulin, M. Riveill. *High Bandwidth Anomalies in Cluster and Grid Data Transfer*, in "Proceedings of the International Workshop on Advanced Topics in Network Computing, Technology, and Applications in conjuction with IFIP International Conference on Network and Parallel Computing (NPC), IEEE Computer Society, Dalian, China", September 2007.

[19] V. Berten, B. Gaujal. *Grid brokering for batch allocation using indexes*, in "Euro-FGI NET-COOP, Avignon, France", LNCS, June 2007.

[20] A. Bouillard, B. Gaujal, E. Thierry, S. Lagrange. *Optimal routing for end-to-end guarantees: the price of multiplexing*, in "Valuetools, 2nd International Conference on Performance Evaluation Methodologies and Tools, Nantes", 2007.

[21] L. Brenner, P. Fernandes, B. Plateau, I. Sbeity. *PEPS2007 - Stochastic Automata Networks Software Tool*, in "4th International Conference on the Quantitative Evaluation of SysTems (QEST) 2007, Edimbourgh, UK", September 2007.

[22] A. Busic, J.-M. Fourneau, K. Grochla, T. Czachorski. *Level Crossing Ordering of Markov Chains: Computing End to End Delays in an All Optical Network*, in "2nd International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2007, Nantes, France", ICTS, October 2007.

[23] A. Busic, J.-M. Fourneau, K. Grochla, T. Czachorski. *Simulation Analysis of Deflection Routing in Hypercube*, in "14th Polish Teletraffic Symposium, Zakopane, Poland", 2007.

[24] A. CARISSIMI, F. DUPROS, J.-F. MÉHAUT, R. V. POLANCZYK. *Aspectos de Programaçao en Maquinas NUMA*, in "Proceedings of VIII Workshop em Sistemas Computacionas de Alto Desempenho, Gramado - Rio Grande de Sul - Brasil", CDROM ISBN 857669153-1, Sociedade Brasileira de Computaçao, October 2007.

[25] R. CZEKSTER, P. FERNANDES, J.-M. VINCENT, T. WEBBER. *Split: a flexible and efficient algorithm to vector-descriptor product*, in "SMCtools", October 2007, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/smctools07Final.pdf.

[26] Y. DENNEULIN. *File systems for large scale architectures*, in "CLCAR'07: Conferencia latinoamericana de computacion de alto rendimiento", Santa Marta, Colombie, August 2007.

[27] P. DUBE, R. JAIN, C. TOUATI. *An Analysis of Oligopolistic Competitive Markets for Queued Services with QoS Guarantees*, in "INFORMS Conference on Applied Probabilities, Eindhoven", July 2007.

[28] L. EYRAUD, A. LEGRAND, M. QUINSON, F. VIVIEN. *A First Step Towards Automatically Building Network Representations*, in "Proceedings of EUROPAR'07), Rennes, France", May 2007.

[29] T. FERRANDIZ, V. MARANGOZOVA. *Managing Scheduling and Replication in the LHC Grid*, in "Proceedings of CoreGRID Workshop on Grid Middleware", July 2007.

[30] J.-M. FOURNEAU. *Closed G-networks with Resets: product form solution*, in "Fourth International Conference on the Quantitative Evaluation of Systems (QEST 2007), Edinburgh, UK", IEEE Computer Society, 2007.

[31] J.-M. FOURNEAU. *Discrete Time Stochastic Automata Networks: using structural properties and stochastic bounds to simplify the SAN*, in "2nd International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2007, Nantes, France", ICTS, October 2007.

[32] J.-M. FOURNEAU, N. PEKERGIN, S. YOUNÈS. *Censoring Markov Chains and Stochastic Bounds*, in "EPEW", 2007, p. 213–227.

[33] J.-M. FOURNEAU, N. PEKERGIN, S. YOUNÈS. *CUT: Combining stochastic ordering and censoring to bound steady-state rewards and first passage time*, in "Fourth International Conference on the Quantitative Evaluation of Systems (QEST 2007), Edinburgh, UK", IEEE Computer Society, 2007.

[34] J.-M. FOURNEAU, B. PLATEAU, W. STEWART. *Product form for Stochastic Automata Networks*, in "2nd International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2007, Nantes, France", ICTS, October 2007.

[35] J.-M. FOURNEAU, H. YAHIAOUI. *Modelling the effect of a worm propagation on a BGP router*, in "14th conference on analytical and stochastic modelling techniques and applications", SCS, 2007.

[36] N. GAST, B. GAUJAL. *Balanced labeled trees: density, complexity and mechanicity*, in "Words, 6th international conference on words, Marseille, France", 2007.

[37] B. GAUJAL, E. HYON. *Reductions of mechanical words and nearest integer continued fractions*, in "Words, 6th international conference on words, Marseille, France", 2007.

[38] B. GAUJAL, F. PERRONNIN. *Coupling from the past in hybrid models for file sharing peer to peer systems*, in "Proceedings of the 10th International Conference on HYBRID SYSTEMS: COMPUTATION AND CONTROL (HSCC'07), Pisa, Italy", April 2007.

[39] Y. GEORGIOU, N. CAPIT, O. RICHARD. *Evaluations of the lightweight grid CIGRI upon the Grid5000 platform*, in "3rd IEEE International Conference on e-Science and Grid Computing (eScience2007), Bangalore, India", December 2007.

[40] E. HERMMAN, R. KASSICK, R. AVILA, C. BARRIOS-HERNANDEZ, M. RIVEILL, Y. DENNEULIN, P. NAVAUX. *Performance Evaluation of Meta Data Transfer and Storage in Clusters*, in "Proceedings of the Latinamerican Conference of High Performance Computing", Santa Marta, Colombia, August 2007, p. 127–134.

[41] I. Y. KADI, J.-M. FOURNEAU, N. PEKERGIN, J. VIENNE, J.-M. VINCENT. *Perfect Simulation and Monotone Stochastic Bounds*, in "2nd International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2007, Nantes, France", ICTS, October 2007.

[42] I. KADI, J.-M. FOURNEAU, N. PEKERGIN, J.-M. VINCENT, J. VIENNE. *Perfect Simulation, Monotone Stochastic Bounds and application to all optical networks*, in "EuroFGI meeting on new modelling technics and Tools, Ghent, Belgium, June 2007,", 2007.

[43] F. KHALIL, H. AUBERT, F. COCCETTI, Y. DENNEULIN, B. MIEGEMOLLE, T. MONTEIL, H. LEGAY. *Electromagnetic Simulation of MEMS-Controlled Reflectarrays based on SCT in Grid Environment*, in "Proceedings of the IEEE International Symposium on Antennas and Propagation, Honolulu, Hawai, USA", June 2007, p. 49–52.

[44] F. KHALIL, F. COCCETTI, H. AUBERT, R. PLANA, Y. DENNEULIN, B. MIEGEMOLLE, T. MONTEIL. *Etude des potentialités du concept de Grille de Calcul pour la Simulation Electromagnétique de Micro-Systèmes Complexes*, in "15èmes Journées Nationales Micro-ondes, Toulouse, France", May 2007, 92.

[45] D. KONDO, D. ANDERSON, J. V. MCLEOD. *Performance Evaluation of Scheduling Policies for Volunteer Computing*, in "Proceedings of the 3rd IEEE International Conference on e-Science and Grid Computing e-Science'07, Bangalore, India", December 2007.

[46] D. KONDO, F. ARAUJO, P. DOMINGUES, L. M. SILVA. *Result Error Detection on Heterogeneous and Volatile Resources Via Intermediate Checkpointing*, in "Coregrid Integration Workshop, Crete, Greece", June 2007.

[47] D. KONDO, F. ARAUJO, P. MALECOT, P. DOMINGUES, L. M. SILVA, G. FEDAK, F. CAPPELLO. *Characterizing Error Rates in Internet Desktop Grids*, in "Proceedings of EUROPAR'07, Rennes, France", May 2007.

[48] A. LEGRAND, C. TOUATI. *How to measure efficiency?*, in "Proceedings of the 1st International Workshop on Game theory for Communication networks (Game-Comm'07)", 2007, http://www-id.imag.fr/Laboratoire/Membres/Touati_Corinne/Articles/gamecomm.pdf.

[49] A. LEGRAND, C. TOUATI. *Non-Cooperative Scheduling of Multiple Bag-of-Task Appplications*, in "Proceedings of the 25th Conference on Computer Communications (INFOCOM'07), Alaska, USA", May 2007.

[50] C. PLUMEJEAUD, J.-M. VINCENT, C. GRASLAND, J. GENSEL, H. MATHIAN, S. GUELTON, J. BOULIER. *HyperSmooth : calcul et visualisation de cartes de potentiel interactives*, in "SAGEO, Clermont-Ferrand", June 2007.

[51] L. A. STEFFENEL, M. MARTINASSO, D. TRYSTRAM. *Assessing Contention Effects on MPI-all-to-all Communications*, in "Proceedings of the 2nd International Conference on Grid and Pervasive Computing (GPC'2007), Paris", Lecture Notes in Computer Science, n^{o} 4459, Springer-Verlag, May 2007, p. 424–435.

[52] B. VIDEAU, C. TOUATI, O. RICHARD. *Toward an Experiment Engine for Lightweight Grids*, in "Proceedings of the MetroGrid workshop: Metrology for Grid Networks", 2007, http://www-id.imag.fr/Laboratoire/Membres/Touati_Corinne/Articles/MetroGrid-videau.pdf.

[53] J.-M. VINCENT, J. VIENNE. *PSI2 a Software Tool for the Perfect Simulation of Finite Queueing Networks*, in "QEST, Edinburgh", September 2007, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/vincent-psi2-soft-tool.pdf.

[54] L. D'ORAZIO, F. JOUANOT, Y. DENNEULIN, C. LABBÉ, C. RONCANCIO, O. VALENTIN. *Distributed semantic caching in grid middleware*, in "Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA'07), Regensburg, Allemagne", September 2007.

Internal Reports

[55] A. LEGRAND, C. TOUATI. *How to measure efficiency?*, Research Report, n^{o} 6216, INRIA, 06 2007, https://hal.inria.fr/inria-00153720.

[56] M. MARTINASSO, J.-F. MÉHAUT, V. DANJEAN. *Rapport d'étude et d'évaluation des outils de traçage et visualisation de programmes sur architectures NUMA*, Projet NUMASIS: Livrable L1.3.1, Technical report, Agence Nationale de la Recherche, Grenoble, July 2007.

[57] M. MARTINASSO, J.-F. MÉHAUT, V. DANJEAN, P.-A. WACRENIER, D. FOUEILLASSAR. *Rapport d'étude et d'évaluation des stratégies d'allocation mémoire adaptées aux machines hiérarchisées: utilisation de directives générées par l'application*, Projet NUMASIS: Livrable L1.2.1, Technical report, Agence Nationale de la Recherche, Grenoble, January 2007.

[58] Y. YANG, H. CASANOVA, M. DROZDOWSKI, M. LAWENDA, A. LEGRAND. *On the Complexity of Multi-Round Divisible Load Scheduling*, Research Report, n^{o} 6096, INRIA, 01 2007, https://hal.inria.fr/inria-00123711.

References in notes

[59] *The GridFTP Protocol and Software*, 2002, http://www.globus.org/.

[60] *GriPPS webpage at , http://gripps.ibcp.fr/*, 2005.

[61] C. BLANCHET, C. COMBET, C. GEOURJON, G. DELÉAGE. *MPSA: Integrated System for Multiple Protein Sequence Analysis with client/server capabilities*, in "Bioinformatics", vol. 16, n^{o} 3, 2000, p. 286-287.

[62] A. LEBRE, Y. DENNEULIN. *aIOLi: An Input/Output LIbrary for cluster of SMP*, in "Proceedings of CCGrid 2005, Cardiff, Pays de Galles", 2005.