INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MESCAL

Middleware Efficiently SCALable

Grenoble - Rhône-Alpes

Theme : Distributed and High Performance Computing

Activity Report

2009

Table of contents

1. **Team** 1
2. **Overall Objectives** 2
   - 2.1. Presentation 2
   - 2.2. Objectives 2
3. **Scientific Foundations** 2
   - 3.1. Large System Modeling and Analysis 2
     - 3.1.1. Behavior analysis of highly distributed systems 3
     - 3.1.2. Simulation of distributed systems 3
     - 3.1.3. Perfect Simulation 4
     - 3.1.4. Fluid models and mean field limits 4
     - 3.1.5. Markov Chain Decomposition 4
     - 3.1.6. Discrete Event Systems 4
     - 3.1.7. Game Theory Methods for Resolving Resource Contention 4
   - 3.2. Management of Large Architectures 5
     - 3.2.1. Fairness in large-scale distributed systems 5
     - 3.2.2. Tools to operate clusters 5
     - 3.2.3. Simple and scalable batch scheduler for clusters and grids 6
   - 3.3. Migration and resilience 6
   - 3.4. Large scale data management 6
     - 3.4.1. Fast distributed storage over a cluster 7
     - 3.4.2. Reliable distribution of data 7
4. **Application Domains** 7
   - 4.1. Introduction 7
   - 4.2. On-demand Geographical Maps 7
   - 4.3. Seismic simulations 8
   - 4.4. The CIMENT project 8
5. **Software** 8
   - 5.1. Tools for cluster management and software development 8
     - 5.1.1. KA-Deploy: deployment tool for clusters and grids 8
     - 5.1.2. Taktuk: parallel launcher 9
     - 5.1.3. NFSp: parallel file system 9
     - 5.1.4. aIOLi 9
     - 5.1.5. Gedeon 9
     - 5.1.6. Generic trace and visualization: Paje 10
     - 5.1.7. OAR: a simple and scalable batch scheduler for clusters and grids 10
   - 5.2. Traces and tools for simulation 10
     - 5.2.1. Failure Trace Archive 10
     - 5.2.2. SimGrid: simulation of distributed applications 11
     - 5.2.3. ψ and ψ^2: perfect simulation of Markov Chain stationary distribution 11
     - 5.2.4. PEPS 11
   - 5.3. HyperAtlas 11
6. **New Results** 11
   - 6.1. Perfect Simulation 11
     - 6.1.1. Different Monotonicity Definitions in Stochastic Modelling 11
     - 6.1.2. Perfect simulation and non-monotone Markovian systems 11
   - 6.2. Tools for Performance Evaluation 12
     - 6.2.1. Model Checking 12
     - 6.2.2. Achieving automatic performance modelling of black boxes for self-sizing 12
     - 6.2.3. Stochastic Automata Networks 12

6.3. Distributed Computing Platforms: Measurements and Models 12
6.3.1. Network Models for Simulation and Emulation 13
6.3.2. Resources Availability for Large Systems 13
6.3.3. Economic Models for Cloud Computing 13
6.4. Scheduling 13
6.4.1. Mean-Field Analysis 14
6.4.2. Work Stealing for Streaming Systems 14
6.4.3. Scheduling of Computing Services on Virtual Clusters 14
6.5. Multi-User Systems 14
6.5.1. Strategies in Allocation Games 14
6.5.2. A Fair User-Network Association Algorithm for Wireless Networks 15
6.6. Programming Many-core Systems 15
6.6.1. Memory Affinity for Hierarchical Shared Memory Multiprocessors 15
6.6.2. Component-based Observations of Systems-on-a-Chip 16
6.7. Middleware and Experimental Testbeds 16
6.8. On-demand Geographical Maps 16
6.9. Discrete Structures 16
7. Contracts and Grants with Industry **17**
7.1. CIFRE with BULL, 06-09 17
7.2. CIFRE with France Télécom R&D, 06-09 17
7.3. CIFRE with STMicroelectronics, 06-10 17
7.4. Sceptre with STMicroelectronics, (Divisions STS and HEG), INRIA Rhone-Alpes (MOAIS, Mescal, Arenaire, CompSys), TIMA/SLS, Verimag, CAPS-Entreprise and IRISA (CAPS) 06-10 17
7.5. Real-Time-At -Work 17
7.6. CILOE with BULL, Compagnie des Signaux, TIMA, CEA-LETI, LIG, Edxact, Infiniscale, Probayes, SCelectronique, 06-10 18
8. Other Grants and Activities **18**
8.1. Regional initiatives 18
8.1.1. CIMENT 18
8.1.2. Grappe200 project 19
8.1.3. Cluster Région 19
8.2. National initiatives 19
8.2.1. DSLLab, 2005-2009, ANR Jeunes Chercheurs 19
8.2.2. NUMASIS, 2005-2009, ANR Calcul Intensif et Grilles de Calcul 19
8.2.3. ALPAGE, 2005-2009, ANR Masses de Données 20
8.2.4. SMS, 2005-2009, ANR 20
8.2.5. Check-bound, 2007-2009 ANR SETIN 20
8.2.6. MEG, 2007-2010, ANR blanc 21
8.2.7. DOCCA, 2007-2011 ANR Jeunes Chercheurs 21
8.2.8. POPEYE, 2008-2009, ARC 22
8.2.9. OMP2, 2008-2010, NANO 2012 22
8.2.10. Aladdin-G5K, 2008-2011, ADT 22
8.2.11. ALEAE, 2009-2010, ARC 23
8.2.12. PROHMPT, 2009-2011, ANR COSI 23
8.2.13. PEGASE, 2009-2011, ANR ARPEGE 23
8.2.14. USS Simgrid, 2009-2011, ANR SEGI 23
8.2.15. SPADES, 2009-2012, ANR SEGI 23
8.2.16. Clouds@home, 2009-2013 ANR Jeunes Chercheurs 24
8.3. International Initiatives 24
8.3.1. Europe 24

8.3.2. Africa 24
8.3.3. North America 25
8.3.4. South America 25
8.3.5. Pacific and South Asia 25
8.4. High Performance Computing Center 25
8.4.1. The ICluster2, the IDPot and the new Digitalis Platforms 25
8.4.2. The BULL Machine 25
8.4.3. GRID 5000 and CIMENT 25
9. Dissemination **26**
9.1. Leadership within the scientific community 26
9.1.1. Tutorials 26
9.1.2. Journal, Conference and Workshop Chairing 26
9.1.3. Program committees 26
9.1.4. Thesis defense 26
9.1.5. Thesis committees 26
9.1.6. Members of editorial board 27
9.1.7. Grenoble’s Seminar on performance evaluation 27
9.2. Teaching 27
10. Bibliography **27**

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

1. Team

Research Scientist

Bruno Gaujal [Team leader, Research Director (DR) INRIA, HdR]
Corinne Touati [Research Associate (CR) INRIA]
Derrick Kondo [Research Associate (CR) INRIA]
Arnaud Legrand [Research Associate (CR) CNRS]

Faculty Member

Yves Denneulin [Professor, Grenoble INP, HdR]
Brigitte Plateau [Professor, Grenoble INP, HdR]
Vania Marangozova-Martin [Associate Professor, UJF]
Jean-François Méhaut [Professor, UJF, HdR]
Florence Perronnin [Associate Professor,UJF]
Olivier Richard [Associate Professor UJF, INRIA Delegation]
Jean-Marc Vincent [Associate Professor, UJF]
Jean-Michel Fourneau [Professor, HdR]

Technical Staff

Romain Cavagna [Engineer Assistant]
Augustin Degomme [Engineer Assistant]
Joseph Emeras [Engineer Assistant]
Pierre Navarro [Engineer Assistant]

PhD Student

Hamza Adamou [2007, University of Yaoundé, Cameroon]
Carlos Jaime Barrios Hernandez [2005, EGIDE, co-tutelle]
Rémi Bertin [2007, ANR DOCCA]
Marcio Bastos Castro [2009, INRIA, co-tutelle]
Kelly Rosa Braghetto [2009, Brazilian CAPES scholarship]
Léonardo Brenner [2004, Brazilian CAPES scholarship]
Marcia Cristina Cera [2008, Brazilian CAPES scholarship, co-tutelle]
Rodrigue Chakode Noumowe [2008, Minalogic CILOE scholarship]
Pierre Coucheney [2008, INRIA-Alcatel Lucent scholarship]
Charbel El Kaed [2008, CIRFE France Télécom R&D scholarship]
Nicolas Gast [2007, AC]
Kiril Georgiev [2009, CIFRE STMicroelectronics]
Yiannis Georgiou [2006, CIFRE BULL scholarship]
Ahmed Harbaoui [2006, CIFRE France Télécom R&D scholarship]
Hussein Joumma [2006, MNRT scholarship]
Benjamin Negrevergne [2008, MNRT scholarship]
Lucas Nussbaum [2005, BDI-CNRS MNRT scholarship]
Matthieu Ospici [2008, CIFRE BULL scholarship, co-tutelle]
Carlos Prada Rojas [2007, CIFRE STMicroelectronics]
Christiane Ribeiro [2008, Brazilian CAPES scholarship, co-tutelle]
Afonso Sales [2005, Brazilian CAPES scholarship]
Nazha Touati [2004, Rhone-Alpes scholarship]
Pedro Antonio Velho [2006, Brazilian CAPES scholarship]
Jérome Vienne [2006, CIFRE BULL scholarship]
Brice Videau [2005, MNRT scholarship]

Blaise Yenké [2004, Ngaundere University scholarship]

Post-Doctoral Fellow

Nadir Farhi [ANR PEGASE, October 2009]
Bahman Javadi-jahantigh [INRIA, November 2008]
Lucas Mello Schnorr [USS Simgrid, November 2009]
Sangho Yi [ARC ALEAE, October 2009]

Administrative Assistant

Annie Simon [Secretary (SAR) INRIA]

2. Overall Objectives

2.1. Presentation

MESCAL is a project-team of INRIA jointly with UJF and INPG universities and CNRS, created in 2005 as an offspring of the former APACHE project-team, together with MOAIS.

MESCAL's research progress and objective were evaluated by INRIA in 2008. The MESCAL project-team received positive evaluations and useful feedback. As such, the project-team was extended for another 4 years by the INRIA evaluation commission.

2.2. Objectives

The recent evolutions in computer technology, as well as their diversification, goes with a tremendous change in the use of these architectures: 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 and heterogeneity of processors, 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 many-core architectures. The experience showed that such sytemrs 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 mathematical modeling and performance evaluation of target architectures, software layers and applications.

3. Scientific Foundations

3.1. Large System Modeling and Analysis

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, analysis and optimization),
- Decomposition of the state space,
- Structural and qualitative analysis,
- Game theory methods for resolving auto-optimization problems.

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 scenarios.

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).

[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.

3.1.3. 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.

3.1.4. Fluid models and mean field limits

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 can be used on a modeling and simulation level. One recent significant application is call centers where . 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. Another application concerns task brokering in desktop grids taking into account statistical features of tasks as well as of the availability of the processors. Mean field has also been applied to the performance evaluatin of work stealing in large systems.

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

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

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

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.

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

Participants: Yves Denneulin, Vania Marangozova-Martin.

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.

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. 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.

4.3. 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 8.2.2) 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.

4.4. 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: 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 [53] 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. 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.

5.1.6. 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 ATHA-PAS-CAN applications was achieved without any new addition into PAJE software. Inserting few events trace into the ATHA-PAS-CAN 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 8.2.2).

5.1.7. 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. Traces and tools for simulation

5.2.1. Failure Trace Archive

The Failure Trace Archive (FTA, http://fta.inria.fr) is centralized public repository of availability traces of distributed systems, and tools for their analysis. The purpose of this archive is to facilitate the design, validation, and comparison of fault-tolerant models and algorithms. In particular, the FTA contains the following:

- availability traces of distributed systems, differing in scale, volatility, and usage
- a standard format for failure traces
- scripts and tools for analyzing these traces

5.2.2. 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/.

5.2.3. 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/.

5.2.4. 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. 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.

6.1.1. Different Monotonicity Definitions in Stochastic Modelling

In [32], we discuss different monotonicity definitions applied in stochastic modelling. Obviously, the relationships between the monotonicity concepts depends on the relation order that we consider on the state space. In the case of total ordering, the stochastic monotonicity used to build bounding models and the realizable monotonicity used in perfect simulation are equivalent to each other while in the case of partial order there is only implication between them. Indeed, there are cases of partial order, where we cannot move from the stochastic monotonicity to the realizable monotonicity. This is why we will try to find the conditions for which there are equivalences between these two notions.

6.1.2. Perfect simulation and non-monotone Markovian systems

Perfect simulation, or coupling from the past, is an efficient technique for sampling the steady state of non-monotone discrete time Markov chains over lattices. Indeed, one only needs to consider two trajectories corresponding to inf. and sup. trajectories. We have shown that these envelopes can be efficiently computed for piece-wise space homogenenous Markov chains. In particular for Almost Sapce Homogeneous Events (ASHEs), closed form formulas for the envelopes are known. This can be extended to events with polytopic piece-wise zones where linear programming can be used to compute envelopes. Linear progamming can be replaced by a fast computations of Minkowsky sums of polytopes and cubes up to a slight loss in the tightness of the bounds. These approaches are useful for example, to simulate queuing networks with “ join the shortest queue” routing schemes.

6.2. Tools for Performance Evaluation

Participants: Jean-Michel Fourneau, Brigitte Plateau, Jean-Marc Vincent.

6.2.1. Model Checking

This is collaborative work with Stavros Tripakis (Cadence Research Laboratories)

Exhaustive verification often suffers from the state-explosion problem, where the reachable state space is too large to fit in main memory. For this reason, and because of disk swapping, once the main memory is full very little progress is made, and the process is not scalable. To alleviate this, partial verification methods have been proposed, some based on randomized exploration, mostly in the form of random walks. In , we enhance partial, randomized state-space exploration methods with the concept of resource-awareness: the exploration algorithm is made aware of the limits on resources, in particular memory and time. We present a memory-aware algorithm that by design never stores more states than those that fit in main memory. We also propose criteria to compare this algorithm with similar other algorithms. We study properties of such algorithms both theoretically on simple classes of state spaces and experimentally on some preliminary case studies.

6.2.2. Achieving automatic performance modelling of black boxes for self-sizing

This is a collaborative work with Nabila Salmi (France Télécom), Bruno Dillenseger (France Télécom)

Modern distributed systems are characterized by a growing complexity of their architecture, functionalities and workload. This complexity, and in particular significant workloads, often lead to quality of service loss, saturation and sometimes unavailability of on-line services. To avoid troubles caused by important workloads and fulfill a given level of quality of service (such as response time), systems need to *self-manage*, for instance by tuning or strengthening one tier through replication. This autonomic feature requires performance modelling of systems. In [29], we developed an automatic identification process providing a queuing model for a part of distributed system considered as black box. This process is a part of a general approach targetting self-sizing for distributed systems and is based on a theoretical and experimental approach. We show how to derive automatically the performance model of one black box considered as a constituent of a distributed system, starting from load injection experiments. This model is determined progressively, using self-regulated test injections, from statistical analysis of measured metrics, namely response time. This process is illustrated through experimental results.

6.2.3. Stochastic Automata Networks

With excellent cost/performance trade-offs 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 for 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 [37], we present a new approach to obtain the Reachable State Space (RSS) of a structured model which uses functional transitions. We use Multi-valued Decision Diagrams (MDD) to store sets of reachable spaces and Stochastic Automata Networks (SAN) formalism to describe structured models. We propose a method to generate a compact MDD description taking advantage of the modular structure of SAN formalism, which also allows one to represent the transition rate matrix of a continuous-time Markov chain by means of a sum of generalized Kronecker products.

6.3. Distributed Computing Platforms: Measurements and Models

Participants: Yves Denneulin, Derrick Kondo, Jean-François Méhaut, Olivier Richard, Jean-Marc Vincent.

6.3.1. Network Models for Simulation and Emulation

Studies in distributed systems generally resort to simulations, which enable reproducible results and make it possible to explore wide ranges of platform and application scenarios. In this context, network simulation is certainly the most critical part. Many packet-level network simulators are available and enable high-accuracy simulation but they lead to prohibitively long simulation times. Therefore, many simulation frameworks have been developed that simulate networks at higher levels, thus enabling fast simulation but losing ac- curacy. One such framework, SimGrid, uses a low-level approach that approximates the behavior of TCP networks, including TCP's bandwidth sharing properties. A prelimliminary study of the accuracy loss by comparing it to popular packet-level simulators has been proposed previously, and regimes in which SimGrid's accuracy is comparable to that of these packet-level simulators are identified. In [41], we come back on this study, reproduce these experiments and provide a deeper analysis that enables us to greatly improve SimGrid's range of validity.

Between discrete event simulation and evaluation within real networks, network emulation is a useful tool to study and evaluate the behaviour of applications. Using a real network as a basis to simulate another network's characteristics, it enables researchers to perform experiments in a wide range of conditions. After giving an overview of the various available network emulators, we compare and contrast in [35] three freely available and widely used network link emulators: Dummynet, NISTNet, and the Linux Traffic Control subsystem. We start by comparing their features, then focus on the accuracy of their latency and bandwidth emulation, and discuss the way they are affected by the time source of the system. We expose several problems that cannot be ignored when using such tools. We also outline differences in their user interfaces, such as the interception point, and discuss possible solutions. This work aims at providing a complete overview of the different solutions for network emulation.

6.3.2. Resources Availability for Large Systems

In the age of cloud, Grid, P2P, and volunteer distributed computing, large-scale systems with tens of thousands of unreliable hosts are increasingly common. Invariably, these systems are composed of heterogeneous hosts whose individual availability often exhibit different statistical properties (for example stationary versus non-stationary behavior) and fit different models (for example Exponen- tial, Weibull, or Pareto probability distributions). In [30], we describe an effective method for discovering subsets of hosts whose availability have similar statistical properties and can be modelled with similar probability distributions. We apply this method with about 230,000 host availability traces obtained from a real large-scale Internet-distributed system, namely SETI@home. We find that about 34% of hosts exhibit availability that is a truly random process, and that these hosts can often be modelled accurately with a few distinct distributions from different families. We believe that this characterization is fundamental in the design of stochastic scheduling algorithms across large-scale systems where host availability is uncertain.

6.3.3. Economic Models for Cloud Computing

Cloud Computing has taken commercial computing by storm. However, adoption of cloud computing platforms and services by the scientific community is in its infancy as the performance and monetary cost-benefits for scien- tific applications are not perfectly clear. This is especially true for desktop grids (aka volunteer computing) applications. In [34] , we compare and contrast the performance and monetary cost-benefits of clouds for desktop grid applications, ranging in computational size and storage. We address the following questions: (i) What are the performance trade- offs in using one platform over the other? (ii) What are the specific resource requirements and monetary costs of creating and deploying applications on each platform? (iii) In light of those monetary and performance cost-benefits, how do these platforms compare? (iv) Can cloud computing platforms be used in combination with desktop grids to improve cost-effectiveness even further? We examine those questions using performance measurements and monetary expenses of real desktop grids and the Amazon elastic com- pute cloud.

6.4. Scheduling

Participants: Jean-Michel Fourneau, Bruno Gaujal, Arnaud Legrand, Jean-François Méhaut.

6.4.1. *Mean-Field Analysis*

In [28], we study the limit behavior of Markov decision processes (MDPs) made of independent particles evolving in a common environment, when the number of particles goes to infinity. In the finite horizon case or with a discounted cost and an infinite horizon, we show that when the number of particles becomes large, the optimal cost of the system converges almost surely to the optimal cost of a deterministic system (the ”optimal mean field”). Convergence also holds for optimal policies. We further provide insights on the speed of convergence by proving several central limits theorems for the cost and the state of the Markov decision process with explicit formulas for the variance of the limit Gaussian laws. Then, our framework is applied to a brokering problem in grid computing. The optimal policy for the limit deterministic system is computed explicitly. Several simulations with growing numbers of processors are reported. They compare the performance of the optimal policy of the limit system used in the finite case with classical policies (such as Join the Shortest Queue) by measuring its asymptotic gain.

6.4.2. *Work Stealing for Streaming Systems*

In [16], we study the performance of parallel stream computations on a multiprocessor architecture using a work-stealing strategy. Incoming tasks are split in a number of jobs allocated to the processors and whenever a processor becomes idle, it steals a fraction (typically half) of the jobs from a busy processor. We propose a new model for the performance analysis of such parallel stream computations. This model takes into account both the algorithmic behavior of work-stealing as well as the hardware constraints of the architecture (synchronizations and bus contentions). Then, we show that this model can be solved using a recursive formula. We further show that this recursive analytical approach is more efficient than the classic global balance technique. However, our method remains computationally impractical when tasks split in many jobs or when many processors are considered. Therefore, bounds are proposed to efficiently solve very large models in an approximate manner. Experimental results show that these bounds are tight and robust so that they immediately find applications in optimization studies. An example is provided for the optimization of energy consumption with performance constraints. In addition, our framework is flexible and we show how it adapts to deal with several stealing strategies.

6.4.3. *Scheduling of Computing Services on Virtual Clusters*

In [13], we consider a context where the available resources of the Intranet of a company are used as a virtual cluster for scientific computation, during the idle periods (nights, weekends, holidays,...). Generally, these idle periods do not permit one to carry out completely the computations. For instance, a workstation mobilized during the night must be released in the morning to make it available for the employee, even if the application running on it is not completed. It is therefore necessary to save the context of uncompleted applications for a possible restart. Hereafter, we assume that the computations running on the workstations are independent from each other. The checkpointing mechanism which ensures the continuity of applications is subject to resource constraints : the network bandwidth, the disk bandwidth and the delay T imposed for releasing the workstations. We first show that the designing of a scheduling strategy which optimizes resource consumption while taking into account the above constraints, can be formalized as a variant of the classical 0/1 knapsack problem. Then, we propose an algorithm whose implementation does not have a significant overhead on checkpointing mechanisms. Experiments carried out on a real cluster show that this algorithm performs better than the naive scheduling algorithm which selects applications one after the other in order of decreasing amount of resource consumption.

6.5. Multi-User Systems

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

6.5.1. *Strategies in Allocation Games*

In [25], we consider allocation games and we investigate the following question: under what conditions does the replicator dynamics select a pure strategy? By definition, an allocation game is a game such that the payoff of a player when she takes an action only depends on the set of players who also take the same action. Such a game can be seen as a set of users who share a set of resources, a choice being an allocation to a resource. A companion game (with modified utilities) is introduced. From the payoffs of an allocation game, we define the reper- cussion utilities: for each player, her repercussion utility is her payoff minus the decrease in marginal payoff that her presence causes to all other players. The corresponding allocation game with repercussion utilities is the game whose payoffs are the repercussion utilities. A simple characterization of those games is given. In such games, if the players select their strategy according to a stochastic approximation of the replicator dynamics, we show that it converges to a Nash equilibrium of the game that is a locally optimal for the initial game. The proof is based on the construction of a potential function for the game. Furthermore, a spectral study of the dynamics shows that no mixed equilibrium is stable, so that the strategies of all players converge to a set of Nash equilibria. Then, martingale argument prove the convergence of the stochastic approximation to a pure point. A discussion of the global/local optimality of the limit points is also included.

There are several approaches of sharing resources among users. There is a noncooperative approach wherein each user strives to maximize its own utility. The most common optimality notion is then the Nash equilibrium. Nash equilibria are generally Pareto inefficient. On the other hand, we consider a Nash equilibrium to be fair as it is defined in a context of fair competition without coalitions (such as cartels and syndicates). In [33], we show a general framework of systems wherein there exists a Pareto optimal allocation that is Pareto superior to an inefficient Nash equilibrium. We consider this Pareto optimum to be ”Nash equilibrium based fair”. We futher define a ”Nash proportion- ately fair” Pareto optimum. We then provide conditions for the existence of a Pareto-optimal allocation that is, truly or most closely, proportional to a Nash equilibrium. As examples that fit in the above framework, we consider noncooperative flow- control problems in communication networks, for which we show the conditions on the existence of Nash-proportionately fair Pareto optimal allocations.

6.5.2. A Fair User-Network Association Algorithm for Wireless Networks

Recent mobile equipment (as well as the norm IEEE 802.21) now offers the possibility for users to switch from one technology to another (vertical handover). This allows flexibility in resource assignments and, consequently, increases the potential throughput allocated to each user. In [24], we design a fully distributed algorithm based on trial and error mechanisms that exploits the benefits of vertical handover by finding fair and efficient assignment schemes. On the one hand, mobiles gradually update the fraction of data packets they send to each network based on the rewards they receive from the stations. On the other hand, network stations send rewards to each mobile that represent the impact each mobile has on the cell throughput. This reward function is closely related to the concept of marginal cost in the pricing literature. Both the station and the mobile algorithms are simple enough to be implemented in current standard equipment. Based on tools from evolutionary games, potential games and replicator dynamics, we analytically show the convergence of the algorithm to solutions that are efficient and fair in terms of throughput. Moreover, we show that after convergence, each user is connected to a single network cell which avoids costly repeated vertical handovers. Several simple heuristics based on this algorithm are proposed to achieve fast convergence. Indeed, for implementation purposes, the number of iterations should remain in the order of a few tens. We also compare, for different loads, the quality of their solutions.

6.6. Programming Many-core Systems

Participants: Jean-François Méhaut, Vania Marangozova-Martin.

6.6.1. Memory Affinity for Hierarchical Shared Memory Multiprocessors

Currently, parallel platforms based on large scale hierarchical shared memory multiprocessors with Non-Uniform Memory Access (NUMA) are becoming a trend in scientific High Performance Computing (HPC). Due to their memory access constraints, these platforms require a very careful data distribution. Many solutions were proposed to resolve this issue. However, most of these solutions did not include optimizations for numerical scientific data (array data structures) and portability issues. Besides, these solutions provide a restrict set of memory policies to deal with data placement. In [39], we describe an user-level interface named Memory Affinity interface (MAi), which allows memory affinity control on Linux based cache-coherent NUMA (ccNUMA) platforms. Its main goals are, fine data control, flexibility and portability. The performance of

MAi is evaluated on three ccNUMA platforms using numerical scientific HPC applications, the NAS Parallel Benchmarks and a Geophysics application. The results show important gains (up to 31%) when compared to the Linux default solution.

In [27], we apply this memory affinity interface to an application that simulations seismic wave propagation. First, we parallelize the application using OpenMP. Then, we improve the OpenMP solution using the MAI (Memory Affinity Interface) library, which allows a control of memory allocation in NUMA machines. The results show that the optimization of memory allocation leads to significant performance gains over the pure OpenMP parallel solution.

6.6.2. Component-based Observations of Systems-on-a-Chip

We propose in [44] a component-based approach to provide a well-suited solution to the programming and deployment problems of systems on chip (SoC) that can become increasingly complex and heterogeneous. Focusing on the aspect of observation, we show, from system to application, that components help in observing all software levels. We present the EMBera prototype and relate our experience in implementing it on two different platforms: a Linux-based 16-core SMP machine and a 5-core embedded system developed by STMicroelectronics.

6.7. Middleware and Experimental Testbeds

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

6.7.1. TakTuk: Middelware for adaptive deployment of remote execution

In [21], we describe TakTuk, which a middleware that deploys efficiently parallel remote executions on large scale grids (thousands of nodes). This tool is mostly intended for interactive use, in particular distributed machines administration and parallel applications development. Thus, it has to minimize the time required to complete the whole deployment process.

To achieve this minimization, we propose and validate a remote execution deployment model inspired by the real world behavior of standard remote execution protocols (rsh and ssh). From this model and from existing works in networking, we deduce an optimal deployment algorithm for the homogeneous case. Unfortunately, this optimal algorithm does not translate directly to the heterogeneous case.

Therefore, we derive from the theoretical solution a heuristic based on dynamic work-stealing that adapts to heterogeneities (processors, links, load, ...). The underlying principle of this heuristic is the same as the principle of the optimal algorithm: to deploy nodes as soon as possible. Experiments assess TakTuk's efficiency and show that TakTuk scales well to thousands of nodes. Compared to similar tools, TakTuk ranks among the best while offering more features and versatility. In particular, TakTuk is the only tool well-suited for remote execution deployment on grids or more heterogeneous platforms.

6.8. On-demand Geographical Maps

Participant: Jean-Marc Vincent.

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

- The potential methods have been developed in the HyperSmooth software and applied in the European ESPON project .
- The HyperSmooth software architecture has been presented in China .

6.9. Discrete Structures

Participants: Yves Denneulin, Bruno Gaujal.

6.9.1. Distributing Labels on Infinite Trees

Sturmian words are infinite binary words with many equivalent definitions. They have a minimal factor complexity among all aperiodic sequences. Also, they are balanced sequences (the labels 0 and 1 are as evenly distributed as possible) and they can be constructed using a mechanical definition. All these properties make them good candidates for being extremal points in scheduling problems over two processors.

In [11], we study infinite unordered d-ary trees with nodes labeled by 0, 1. We introduce the notions of rational and Sturmian trees along with the definitions of (strongly) balanced trees and mechanical trees, and study the relations among them. In particular, we show that (strongly) balanced trees exist and coincide with mechanical trees in the irrational case, providing an effective construction. Such trees also have a minimal factor complexity, hence are Sturmian. We also give several examples illustrating the inclusion relations between these classes of trees.

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 Rhone-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. CILOE 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, Rhone-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, GRAAL 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. DSLLab, 2005-2009, 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.

8.2.2. NUMASIS, 2005-2009, 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.

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

8.2.3. ALPAGE, 2005-2009, 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.

8.2.4. SMS, 2005-2009, 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 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.

8.2.5. 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.

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.

8.2.6. MEG, 2007-2010, ANR blanc

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.

8.2.7. 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.

[3] Design and Optimization of Collaborative Computing Architectures

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.

8.2.8. POPEYE, 2008-2009, ARC

Partners: INRIA Maestro, INRIA TOSCA, INRA, UMPC, LIA, Polytech Nice Sophia-Antipolis.

The MESCAL project-team participates in the Popeye INRIA ARC, lead by Eltan Altman of the INRIA Maestro project-team. The project focuses on the behavior of large complex systems that involve interactions among one or more populations. By population we mean a large set of individuals, that may be modeled as individual agents, but that we will often model as consisting of a continuum of non-atomic agents. The project brings together researchers from different disciplines: computer science and network engineering, applied mathematics, economics and biology. This interdisciplinary collaborative research aims at developing new theoretical tools as well as at their applications to dynamic and spatial aspects of populations that arise in various disciplines, with a particular focus on biology and networking.

8.2.9. OMP2, 2008-2010, NANO 2012

Rapid advances in multi-core technologies have been incorporated in general-purpose processors from Intel, IBM, Sun, and AMD, and special-purpose graphics processors from NVIDIA and ATI. This technology will soon be introduced to the next generation of processors in embedded systems. The increase in the number of cores per processor will introduce critical challenges for the access of data stored in memory. The synchronization of memory accesses is often done using the use of locks for shared variables. As the number of threads increases, the cost of synchronization also increases due to increased access to these shared variables. Transactional memory is currently an approach being actively investigated. The goal of this project is to improve the programability and performance of parallel systems using the approach of transactional memory in the context of embedded systems.

8.2.10. Aladdin-G5K, 2008-2011, ADT

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

After the success of the Grid'5000 project of the ACI Grid initiative led by the French ministry of research, INRIA is launching the ALADDIN project to further develop the Grid'5000 infrastructure and foster scientific research using the infrastructure.

ALADDIN will build on Grid'5000's experience to provide an infrastructure enabling computer scientists to conduct experiments on large scale computing and produce scientific results that can be reproduced by others. ALADDIN focus on the following challenges :

1. Transparent, safe and efficient large scale system utilization and programming
2. Providing service agreement to users in large scale parallel and distributed systems
3. Providing confidence to the user about the infrastructure
4. Efficient exploitation of highly heterogeneous and hierarchical large-scale systems
5. Efficient and scalable composition and orchestration of services
6. Modeling of large scale systems and validation of their simulators
7. Scalable applications for large scale systems
8. Dynamic interconnection of autonomous and heterogeneous resources
9. Efficiently manage very large volumes of information (search, mining, classification, secure storage and access, etc) for a wide spectrum of applications areas (web applications, image processing, health, environment, etc).

Mescal members are particularly involved in topics 1, 3, 4, and 6.

8.2.11. ALEAE, 2009-2010, ARC

Partners: INRIA ALGORILLE, INRIA GRAAL, INRIA MESCAL, TU Delft.

The MESCAL project-team participates in the ALEAE project of the INRIA ARC program. This project is led by Emmanuel Jeannot of the INRIA ALGORILLE project-team, who recently moved to the RUNTIME project-team.

The project's goal is to provide models and algorithmic solutions in the field of resource management that cope with uncertainties in large-scale distributed systems. This work is based on the Grid Workloads Archive designed at TU Delft, Netherlands. Resulting from this collaboration, we have created the Failure Trace Archive, which is a repository of availabilty traces of distributed systems, and analytical tools. Moreover, we are conducting trace-driven experiments to test our solutions, to validate the proposed models, and to evaluate the algorithms. These experiments are being conducted using simulators and large-scale environments such as Grid'5000 in order to improve both models and algorithms.

8.2.12. PROHMPT, 2009-2011, ANR COSI

Partners: BULL SAS, CAPS entreprise, CEA CESTA, CEA INAC, INRIA RUNTIME, UVSQ PriSM

Processor architectures with many-core processors and special-purpose processors such as GPUS and the CELL processor have recenty emerged. These new and heterogeneous architectures require new applicaton programming methods and new programming models. The goal of the ProHMPT project is to address this challenge by focusing on the immense computing needs and requirements of real simulations for nanotechnologies. In order for nanosimulations to fully leverage heterogeneous computing architectures, project members will novel technologies at the compiler, runtime, and scientific kernely levels with proper abstractions and wide portability. This project brings experts from industry, in particular HPC hardware expertise from BULL and nanosimulation expertise from CEA.

8.2.13. PEGASE, 2009-2011, ANR ARPEGE

Partners: RealTimeAtWork, Thales, ONERA, ENS Cachan

The goal of this project to achieve performance guarantees for communicating embedded systems. Members will develop mathematical methods that give accurate bounds on maximum network delays in both space and aviation systems. The mathematical methods will be based on Network Calculus theory, which is type of queuing theory that deals with worst-case performance evaluation. The expected results will be novel models and software tools validated in mission-critical real-time embedded networks of the aerospace industry.

8.2.14. USS Simgrid, 2009-2011, ANR SEGI

Partners: INRIA Nancy, INRIA Saclay, INRIA Bordeaux, University of Reims, IN2P3, University of Hawaii at Manoa

The goal of the USS-SimGrid project is to allow scalable and accurate simulations by means of the SimGrid simulation toolkit. This toolkit is widely used for simulation of HPC systems. We aim to extend the functionality of the toolkit to enable the simulation of heterogeneous systems with more than tens of thousands of nodes.

There three main thrusts in this project. First, we will improve the models used in SimGrid, increasing their scalability and easing their instanciation. Second, we will develop tools that ease the analysis of detailed and large simulation results, and aid the management of simulation deployments. Third, we will improve the scalability of simulations using parallelization and optimization methods.

8.2.15. SPADES, 2009-2012, ANR SEGI

Partners: INRIA GRAAL, INRIA GRAND-LARGE, CERFACS, CNRS, INRIA PARIS, LORIA

Petascale systems consisting of thousands to millions of resources have emerged. At the same, existing infrastructure are not capable of fully harnessing the computational power of such systems. The SPADES project will address several challenges in such large systems. First, the members are investigating methods for service discovery in volatile and dynamic platforms. Second, the members creating novel models of reliability in PetaScale systems. Third, the members will develop stochastic scheduling methods that leverage these models. This will be done with emphasis on applications with task dependencies structured as graph.

8.2.16. Clouds@home, 2009-2013 ANR Jeunes Chercheurs

The overall objective of this project is to design and develop a cloud computing platform that enables the execution of complex services and applications over unreliable volunteered resources over the Internet. In terms of reliability, these resources are often unavailable 40% of the time, and exhibit frequent churn (several times a day). In terms of "real, complex services and applications", we refer to large-scale service deployments, such as Amazon's EC2, the TeraGrid, and the EGEE, and also applications with complex dependencies among tasks. These commercial and scientific services and applications need guaranteed availability levels of 99.999% for computational, network, and storage resources in order to have efficient and timely execution. As such we have the following goals:

- To research methods that guarantee performance for computation and storage across unreliable Internet volunteered resources using a combination of prediction and virtual machine techniques
- To design a cloud computing platform that allows complex services and applications to leverage this guaranteed computing and storage power

We are currently working in the following areas:

- Predictive models of availability of groups of volatile Internet resources
- Strategies for checkpointing applications using virtual machines (VM's) in low-bandwidth, volatile, and wide-area networks
- Methods for data management that ensure data durability, availability, and access performance
- Implementation of a cloud computing prototype with validation on an experimental platform such as PlanetLab.

8.3. International Initiatives

8.3.1. Europe

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. Cameroon student Blaise Yenké completed his PhD under the joint supervision of 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. In addition, MESCAL participates in the IDASCO joint project with the University of Yaoundé 1. This is part of the international LIRIMA laboratory, whose goal to develop novel methods and tools for collecting and analyzing massive data sets from biological or environmental domains.

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

8.3.3. North America

CloudComputing@home (2009-2011) is an Associate Team funded by INRIA between UC Berkeley and the MESCAL project-team. Members of this collaborative project focus on several challenges to achieve cloud computing over Internet hosts. They address these challenges drawing on the experience of the BOINC team at UC Berkeley which designed and implemented BOINC (a middleware for volunteer computing that is the underlying infrastructure for SETI@home), and the MESCAL team which designed and implemented OAR (an industrial-strength resource management system that runs across France's main 5000-node Grid called Grid'5000). This year Bahman Javadi and Derrick Kondo visited UC Berkeley for 2 weeks. Jeremy Cowles from UC Berkeley visited Grenoble for 1 week.

8.3.4. South America

- DIODE-A (2009-2011) 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 (Jean-François Méhaut, Bridgette Plateau, Jean-Marc Vincent) to enhance the existing collaborations and to form new ones.
- 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.

8.3.5. Pacific and South Asia

Corinne Touati is the INPG correspondent for student exchanges with Japan and has visited many Japanese universities to ease these exchanges.

8.4. High Performance Computing Center

8.4.1. The ICluster2, the IDPot and the new Digitalis 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) and another based on 34 bi-processor quad-core XEON (Digitalis) located at INRIA. The three of them are integrated in the Grid'5000 grid platform.

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 Icluster-2 has been stopped this year as it was getting obsolete and has been replaced by the Digitalis platform. The Digitalis cluster is also meant to replace the Grimage platform in which the MOAIS project-team is very involved.

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 Digitalis 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. Tutorials

Researchers of the MESCAL project-team have been invited to give tutorials on critical research subjects in international conferences:

- Jean-Marc Vincent gave a tutorial at SBAC 2009 on Visualization for peformance debugging of large scale parallel applications
- Bruno Gaujal was a keynote speaker at Bionetics (special day in the memory of Thomas Vincent).

9.1.2. Journal, Conference and Workshop Chairing

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

- Workshop on Desktop Grids and Volunteer Computing (Derrick Kondo, Co-general chair).
- Journal of Grid Computing, Special Issue on Volunteer Computing and Desktop Grids (Derrick Kondo, Editor)
- High Performance Computing and Simulation (Bahman Javadi, Derrick Kondo, Co-panel chairs)

9.1.3. Program committees

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

- Workshop on Desktop Grids and Volunteer Computing, Italy.
- IEEE International Symposium on Cluster, Cloud, and Grid Computing, Shanghai
- Workshop on Service-Oriented P2P Networks and Grid Systems, Shanghai
- International Conference on Performance Evaluation Methodologies and Tools (ValueTools), Athens.
- Simutools 2009, Italy.
- SPAA 2009, Nancy.

9.1.4. Thesis defense

- Thais Weber
- Leonardo Brenner
- Nazha Touati
- Afonso Sales
- Brice Videau

9.1.5. Thesis committees

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

- Bruno Gaujal served on the thesis committee of Sydney Rosario as a reviewer.
- Bruno Gaujal served on the thesis committee of Afonso Sales
- Bruno Gaujal served on the committee of the HDR defense of Johanne Cohen.
- Jean-Marc Vincent served on the thesis committee of Leonardo Brenner.
- Jean-Marc Vincent served on the thesis committee of Thais Weber.

9.1.6. Members of editorial board

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:

- **2nd year of Research Master of Paris (MPRI)** Bruno Gaujal gives a course on discrete event dynamic systems.
- **2nd year of International Research Master of Grenoble (MOSIG)** 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° 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° 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° 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° 2, 1997, p. 209-232.

Year Publications

Doctoral Dissertations and Habilitation Theses

[5] N. ABED. *Exploration probabiliste de larges espaces dâêtats pour la vérification*, Université Joseph Fourier, Grenoble, June 2009, Ph. D. Thesis.

[6] L. BRENNER. *Réseaux d'Automates Stochastiques: Analyse trasitoire en temps continu et Algèbre tensorielle pour une sémantique en temps discret*, Institut Polytechnique de Grenoble, France, September 2009, Ph. D. Thesis.

[7] A. SALES. *Réseaux d'Automates Stochastiques: Génération de l'espace d'états atteignables et Multiplication vecteur-descripteur pour une sémantique en temps discret*, Institut Polytechnique de Grenoble, France, September 2009, Ph. D. Thesis.

[8] B. VIDEAU. *Expérimentations sur les nouvelles architectures: Des processeurs multi-cœurs aux grilles de calcul*, Université Joseph Fourier, September 2009, Ph. D. Thesis.

[9] T. WEBBER. *Reducing the Impact of State Space Explosion in Stochastic Automata Networks*, Pontifícia Universidade Católica do Rio Grande do Sul, March 2009, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[10] L. BRENNER, P. FERNANDES, J.-M. FOURNEAU, B. PLATEAU. *Modelling Grid5000 point availability with SAN*, in "Electronic Notes In Theoretical Computer Science", vol. 232, March 2009, p. 165-178.

[11] N. GAST, B. GAUJAL. *Infnite Labeled Trees: from Rational to Sturmian Trees*, in "Theoretical Computer Science", 2009, Accepted for publication.

[12] L. GENOVESE, M. OSPICI, T. DEUTSCH, J.-F. MÉHAUT, A. NEELOV, S. GOEDECKER. *Density Functional Theory Calculation on many-cores Hybrid CPU-GPU architectures*, in "Journal of Chemical Physics", vol. 131, n^{o} 3, July 2009, 034103.

[13] B. YENKE, J.-F. MÉHAUT, M. TCHUENTÉ. *Scheduling of Computing Services on Intranet Networks*, in "IEEE Transactions on Computing Services", 2009, Accepted for publication.

Invited Conferences

[14] L. M. SCHNORR, G. HUARD, B. STEIN, J.-M. VINCENT. *Visualization for Performance Debugging of Large-Scale Parallel Applications*, in "SBAC (tutorial), Saõ-Paulo, Brazil", October 2009.

[15] J.-M. VINCENT. *Perfect Sampling of Queuing Networks with Complex Routing complexity and computational aspects*, in "RESCOM (tutorial), La Palmyre", June 2009.

International Peer-Reviewed Conference/Proceedings

[16] J. ANSELMI, B. GAUJAL. *Performance Evaluation of Work Stealing for Streaming Applications*, in "International Conference On Principles Of Distributed Systems (OPODIS), Nimes, France", 2009.

[17] A. BENOIT, M. GALLET, B. GAUJAL, Y. ROBERT. *Computing the throughput of replicated workflows on heterogeneous platforms*, in "Int. conf. on Parallel Processing, ICPP, Vienna, Austria", 2009.

[18] A. BOUILLARD, B. COTTENCEAU, B. GAUJAL, L. HARDOUIN, S. LAGRANGE, M. LHOMMEAU, E. THIERRY. *COINC Library : A toolbox for Network Calculus*, in "Fourth International Conference on Performance Evaluation Methodologies and Tools, Valuetools, Pisa, Italy", 2009.

[19] M. CASTRO, L. G. FERNANDES, C. POUSA, J.-F. MÉHAUT, M. AGUIAR. *NUMA-ICTM: A Parallel Version of ICTM Exploiting Memory Placement Strategies for NUMA Machines*, in "Proceedings of the 10th International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC), Roma, Italy", May 2009.

[20] M. CERA, Y. GEORGIOU, O. RICHARD, N. MAILLARD, P. NAVAUX. *Supporting MPI Malleable Applications upon the OAR Resource Manager*, in "Colloque d'Informatique INRIA Brésil, Coopérations, Avancées, Défis (COLIBRI)", 2009.

[21] B. CLAUDEL, G. HUARD, O. RICHARD. *TakTuk, Adaptive Deployment of Remote Executions*, in "Proceedings of the International Symposium on High Performance Distributed Computing (HPDC)", June 2009.

[22] P. COUCHENEY, E. HYON, C. TOUATI, B. GAUJAL. *Myopic versus clairvoyant admission policies in wireless networks*, in "3rd ICST/ACM International Workshop on Game Theory in Communication Networks, Pisa", ICST, 2009, invited paper.

[23] P. COUCHENEY, C. TOUATI, B. GAUJAL. *Different Dynamics for Optimal Association in Heterogeneous Wireless Networks*, in "The 5th workshop on Resource Allocation, Cooperation and Competition in Wireless Networks (RAWNET/WNC3)", June 2009.

[24] P. COUCHENEY, C. TOUATI, B. GAUJAL. *Fair and Efficient User-Network Association Algorithm for Multi-Technology Wireless Networks*, in "Proc. of the 28th conference on Computer Communications miniconference (INFOCOM)", 2009.

[25] P. COUCHENEY, C. TOUATI, B. GAUJAL. *Selection of Efficient Pure Strategies in Allocation Games*, in "Proc. of the International Conference on Game Theory for Networks (GameNets)", 2009.

[26] G. DA-COSTA, J.-P. GELAS, Y. GEORGIOU, L. LEFEVRE, A.-C. ORGERIE, J.-M. PIERSON, O. RICHARD, K. SHARMA. *The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems*, in "HPPAC 2009 : High Performance Power Aware Computing Workshop in conjunction with IPDPS 2009, Rome, Italy", May 2009.

[27] F. DUPROS, C. POUSA, A. CARISSIMI, J.-F. MÉHAUT. *Parallel Seismic Wave Propagation on NUMA architectures*, in "Proceedings of International Conference on Parallel Computing (ParCO), ENS Lyon", September 2009.

[28] N. GAST, B. GAUJAL. *A Mean Field Approach for Optimization in Particle Systems and Applications*, in "Fourth International Conference on Performance Evaluation Methodologies and Tools, Valuetools, Pisa", 2009.

[29] A. HARBAOUI, N. SALMI, B. DILLENSEGER, J.-M. VINCENT. *Achieving automatic performance modelling of black boxes for self-sizing*, in "HotAC IV, Barcelona", June 2009.

[30] B. JAVADI, D. KONDO, J.-M. VINCENT, D. P. ANDERSON. *Mining for Statistical Models of Availability in Large-Scale Distributed Systems: An Empirical Study of SETI@home*, in "17th IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)", September 2009.

[31] H. JOUMAA, Y. DEMAZEAU, J.-M. VINCENT. *Performance visualization of a transport multi-agent application*, in "7th International Conference on Practical Applications of Agents and Multi-Agent Systems, PAAMS09, Salamanca, Spain", Springer Verlag, March 2009, p. 188-196, AISC 55.

[32] I. KADI, N. PEKERGIN, J.-M. VINCENT. *Different Monotonicity Definitions in Stochastic Modelling*, in "ASMTA, Madrid", n^{o} 5513, LNCS, June 2009, p. 144-158.

[33] H. KAMEDA, E. ALTMAN, C. TOUATI, A. LEGRAND. *Nash Equilibrium Based Fairness*, in "Proc. of the International Conference on Game Theory for Networks (GameNets)", 2009.

[34] D. KONDO, B. JAVADI, P. MALÉCOT, F. CAPPELLO, D. P. ANDERSON. *Cost-Benefit Analysis of Cloud Computing versus Desktop Grids*, in "18th International Heterogeneity in Computing Workshop, Rome, Italy", May 2009, http://mescal.imag.fr/membres/derrick.kondo/pubs/kondo_hcw09.pdf.

[35] L. NUSSBAUM, O. RICHARD. *A Comparative Study of Network Link Emulators*, in "12 Communications and Networking Simulation Symposium (CNS'09), San Diego, USA", 2009.

[36] L. NUSSBAUM, O. RICHARD. *On Robust Covert Channels Inside DNS*, in "24th IFIP International Security Conference, Pafos, Cyprus", 2009.

[37] B. PLATEAU, A. SALES. *Reachable state space generation for structured models which use functional transitions*, in "Proceedings of the 6th International Conference on the Quantitative Evaluation of Systems (QEST'09), Budapest, Hungary", IEEE Computer Society, September 2009, p. 269-278.

[38] C. POUSA, M. CASTRO, F. DUPROS, A. CARISSIMI, L. G. FERNANDES, J.-F. MÉHAUT. *High Performance Applications on Hierarchical Shared Memory Multiprocessors*, in "Colloque d'Informatique INRIA Brésil, Coopérations, Avancées, Défis (COLIBRI), Bento Goncalves, Brazil", July 2009.

[39] C. POUSA, M. CASTRO, L. G. FERNANDES, A. CARISSIMI, J.-F. MÉHAUT. *Memory Affinity for Hierarchical Shared Memory Multiprocessors*, in "Proceedings of IEEE International Symposium on Computer Architectures and High Performance Computing (SBAC-PAD), Sao Paulo, Brazil", IEEE, October 2009.

[40] C. PRADA ROJAS, V. MARANGOZOVA, K. GEORGIEV, J.-F. MÉHAUT, M. SANTANA. *Towards a Component Based Observation of MPSoC*, in "Proceedings of the 4th IEEE International Symposium on Embedded Multicore Systems on Chips (MCSoC), Vienna, Austria", September 2009.

[41] P. VELHO, A. LEGRAND. *Accuracy Study and Improvement of Network Simulation in the SimGrid Framework*, in "SIMUTools'09, 2nd International Conference on Simulation Tools and Techniques", 2009, http://mescal.imag.fr/membres/arnaud.legrand/articles/simutools09.pdf.

[42] B. VIDEAU, E. SAULE, J.-F. MÉHAUT. *Parallel Runtime and Algorithms for Small Datasets*, in "Proceedings of the 2nd International IEEE Workshop on Multi-Core Computing Systems (MuCoCoS), Fukuoka, Japan", March 2009.

National Peer-Reviewed Conference/Proceedings

[43] M. Ospici, L. Genovese, T. Deutsch. *Exploitation et partage de GPU dans des grappes de calcul hybride*, in "Rencontres Francophones du Parallélisme (RenPar), Toulouse, France", September 2009.

[44] C. Prada Rojas, V. Marangozova, K. Georgiev, J.-F. Méhaut, M. Santana. *Observation de systèmes embarqués : une approche à base de composants*, in "Conférence Française sur les Systèmes en Exploitation (CFSE), Toulouse, France", September 2009.

Scientific Books (or Scientific Book chapters)

[45] F. Cappello, G. Fedak, D. Kondo, P. Malécot, A. Rezmerita. *Desktop Grids : From Volunteer Distributed Computing to High Throughput Computing Production Platforms*, in "Handbook of Research on Scalable Computing Technologies", K.-C. Li, C.-H. Hsu, L. Tianruo Yang, J. Dongarra, H. Zima (editors), IGI Global, 2009.

[46] B. Gaujal, J.-M. Vincent. *Comparison of stochastic task-ressource systems*, in "New trends in scheduling", Taylor and Francis publisher, 2009, to appear.

[47] A. Legrand, L. Eyraud. *Influence of Platform Models on Scheduling Techniques*, in "New Trends in Scheduling", Taylor and Francis publisher, 2009, To appear.

Research Reports

[48] N. Balacheff, J. Crowley, C. Garbay, F. Ouabdesselam, B. Plateau. *Proposition d'une unité mixte de recherche 2007-2010, CNRS-INP-INRIA-UJF-UPMF*, n^o RR-LIG-001, LIG, Grenoble, France, 2009, http://rr.liglab.fr/research_report/RR-LIG-001.pdf, Research Report.

[49] K. R. Braghetto, F. Joao Eduardo, J.-M. Vincent. *Comparison of Modeling Approaches to Business Process Performance Evaluation*, n^o 7065, INRIA, 2009, Technical report.

[50] N. Gast, B. Gaujal. *A Mean Field Approach for Optimization in Particles Systems and Applications*, n^o RR-6877, INRIA, 2009, http://hal.inria.fr/inria-00368011/en/, Research Report.

[51] A. Harbaoui, N. Salmi, B. Dillenseger, J.-M. Vincent. *Automatic performance modelling of black boxes targetting self-sizing*, n^o 7027, INRIA, 2009, Technical report.

[52] C. Prada Rojas, V. Marangozova, K. Georgiev, J.-F. Méhaut, M. Santana. *Towards a Component-based Observation of MPSoC*, n^o 6905, INRIA, April 2009, http://hal.inria.fr/INRIA-RHA/inria-00376759/en/, Research Report.

References in notes

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