INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MESCAL

Middleware Efficiently SCALable

Grenoble - Rhône-Alpes

Theme : Distributed and High Performance Computing

Table of contents

1. Team 1
2. Overall Objectives 2
2.1. Presentation 2
2.2. Objectives 2
3. Scientific Foundations 3
3.1. Large System Modeling and Analysis 3
3.1.1. Behavior analysis of highly distributed systems 3
3.1.2. Simulation of distributed systems 4
3.1.2.1. SimGrid 4
3.1.2.2. Perfect Simulation 4
3.1.3. Fluid models and mean field limits 4
3.1.4. Markov Chain Decomposition 4
3.1.5. Discrete Event Systems 5
3.1.6. Game Theory Methods for Resolving Resource Contention 5
3.2. Management of Large Architectures 5
3.2.1. Fairness in large-scale distributed systems 6
3.2.2. Tools to operate clusters 6
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 7
3.4.1. Fast distributed storage over a cluster 7
3.4.2. Reliable distribution of data 7
4. Application Domains 8
4.1. Introduction 8
4.2. On-demand Geographical Maps 8
4.3. Nano simulations 8
4.4. Seismic simulations 8
4.5. Electromagnetic Fields simulations 9
4.6. The CIMENT project 9
5. Software 9
5.1. Tools for cluster management and software development 9
5.1.1. Triva 9
5.1.2. KA-Deploy: deployment tool for clusters and grids 10
5.1.3. Taktuk: parallel launcher 10
5.1.4. Generic trace and visualization: Paje 10
5.1.5. OAR: a simple and scalable batch scheduler for clusters and grids 10
5.2. Traces and tools for simulation 11
5.2.1. Failure Trace Archive 11
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 12
6. New Results 12
6.1. Simulation 12
6.1.1. Perfect sampling of Markov chains with piecewise homogeneous events 12
6.1.2. Fast and Scalable Simulation of Volunteer Computing Systems Using SimGrid 13
6.2. Tools for Performance Evaluation 13
6.2.1. Steady-state Property Verification 13
6.2.2. Queuing Based Performance Awareness in Autonomic Systems 13

6.2.3. Performance Evaluation of Streaming Applications 13
6.3. Distributed Computing Platforms: Measurements and Models 14
6.3.1. Availability Modeling of 100,000+ Node Systems. 14
6.3.2. Supporting Malleability in Parallel Architectures 14
6.3.3. Monetary Costs: Economics of Distributed Computing 14
6.4. Multi-User Systems 15
6.4.1. Optimal Mean-Field approximation 15
6.4.2. Mean-Field limits and differential inclusions 15
6.4.3. The Price of Forgetting in Parallel Queues 15
6.4.4. Self-optimizing Routing in MANETs with Multi-class Flows 16
6.4.5. Robust Control Problem with Infinitely Many Mobile Agents 16
6.4.6. Practical Implementation of Distributed Optimization Techniques 16
6.5. Programming Many-core Systems 16
6.5.1. Memory Issues on Multicore Architectures 16
6.5.2. Sharing and Mutualizing Clusters 17
6.6. Middleware and Experimental Testbeds 17
6.6.1. Charm++ on NUMA Platforms 17
6.6.2. Deadline-Constrained Checkpointing in a Batch Scheduler 18
6.6.3. Debugging Embedded Linux Kernel Through JTAG Port 18
7. Contracts and Grants with Industry **18**
7.1. Four CIFRE contracts with BULL 18
7.2. Two CIFRE contracts with France Télécom R&D 19
7.3. Four CIFRE with STMicroelectronics 19
7.4. Real-Time-At -Work 19
7.5. Selfnets Research action with Alcatel 19
7.6. CILOE with BULL, Compagnie des Signaux, TIMA, CEA-LETI, LIG, Edxact, Infiniscale, Probayes, SCelectronique, 06-10 20
8. Other Grants and Activities **20**
8.1. Regional initiatives 20
8.2. National initiatives 20
8.2.1. ADT HEMERA, 2010-2012 20
8.2.2. ADT INRIA, 2009-2011, SimGrid for Human Beings 21
8.2.3. NUMASIS, 2005-2010, ANR Calcul Intensif et Grilles de Calcul 21
8.2.4. Check-bound, 2007-2010 ANR SETIN 21
8.2.5. MEG, 2007-2010, ANR blanc 22
8.2.6. DOCCA, 2007-2011 ANR Jeunes Chercheurs 22
8.2.7. OMP2, 2008-2010, NANO 2012 22
8.2.8. Aladdin-G5K, 2008-2011, ADT 22
8.2.9. ALEAE, 2009-2010, ARC 23
8.2.10. PROHMPT, 2009-2011, ANR COSI 23
8.2.11. PEGASE, 2009-2011, ANR ARPEGE 23
8.2.12. USS Simgrid, 2009-2011, ANR SEGI 24
8.2.13. SPADES, 2009-2012, ANR SEGI 24
8.2.14. Clouds@home, 2009-2013 ANR Jeunes Chercheurs 24
8.3. International Initiatives 24
8.3.1. Europe 24
8.3.2. Africa 25
8.3.3. North America 25
8.3.4. South America 26
8.3.5. Pacific and South Asia 26
8.4. High Performance Computing Center 26

8.4.1. The ICluster2, the IDPot and the new Digitalis Platforms 26
8.4.2. The BULL Machine 26
8.4.3. GRID 5000 and CIMENT 27
9. Dissemination **27**
9.1. Leadership within the scientific community 27
9.1.1. Invited talks 27
9.1.2. Journal, Conference and Workshop Chairing 27
9.1.3. Program committees 28
9.1.4. Thesis defense 28
9.1.5. Thesis committees 28
9.1.6. Members of editorial board 28
9.1.7. Grenoble's Seminar on performance evaluation 28
9.1.8. Popular Science 29
9.2. Teaching 29
10. Bibliography **29**

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 Scientists

Bruno Gaujal [Team leader, Senior Researcher (DR) INRIA, HdR]
Derrick Kondo [Junior Researcher (CR) INRIA]
Corinne Touati [Junior Researcher (CR) INRIA]
Arnaud Legrand [Junior Researcher (CR) CNRS]

Faculty Members

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

Technical Staff

Jonatha Anselmi [Sceptre, 01/09- 04/10]
Paolo Ballarini [ANR Checkbound, 09/10- 12/10]
Bruno Bzeznik [Research Engineer]
Romain Cavagna [Engineer Assistant]
Augustin Degomme [Engineer Assistant]
Pere Manils [Engineer Assistant]
Pierre Navarro [Engineer Assistant]
Pierre Neyron [Research Engineer]
Joseph Emeras [2010, BDI-CNRS]

PhD Students

Hamza Adamou [2007, University of Yaoundé, Cameroon]
Rémi Bertin [2007-2010, ANR DOCCA]
Marcio Bastos Castro [2009, INRIA]
Rodrigue Chakode Noumowe [2008, Minalogic CILOE scholarship]
Pierre Coucheney [2008, INRIA-Alcatel Lucent scholarship]
Charbel El Kaed [2008, CIFRE France Télécom R&D scholarship]
Nicolas Gast [2007-2010, Allocation Couplée]
Kiril Georgiev [2009, CIFRE STMicroelectronics]
Yiannis Georgiou [2006-2010, CIFRE BULL scholarship]
Gael Gorgo [2010, CIFRE BULL]
Ahmed Harbaoui [2006, CIFRE France Télécom R&D scholarship]
Hussein Joumma [2006-2010, MNRT scholarship]
Patricia Lopez Cueva [2010, CIFRE STmicroelectronics scholarship]
Matthieu Ospici [2008, CIFRE BULL scholarship, co-tutelle]
Kevin Pouget [2010-2013]
Carlos Prada Rojas [2007, CIFRE STMicroelectronics]
Christiane Ribeiro [2008, Brazilian CAPES scholarship, co-tutelle]
Kelly Rosa Braghetto [2009, Brazilian CAPES scholarship]
Pedro Antonio Velho [2006, Brazilian CAPES scholarship]
Jérome Vienne [2006-2010, CIFRE BULL scholarship]
Blaise Yenké [2005, Ngaundere University scholarship]

Post-Doctoral Fellows

Nadir Farhi [ANR PEGASE 10/09 - 10/10]
Issam Al Azzoni [ANR 2010-2011]
Eric Heien [ANR 2010-2012]
Sascha Hunold [ANR 2010-2011]
Bahman Javadi-Jahantigh [ANR, 11/08 - 04/10]
Lucas Mello Schnorr [USS Simgrid, 11/09- 01/11]
Laurent Bobelin [09/10-09/11]
Sangho Yi [ARC ALEAE, 10/09 - 04/11]

Administrative Assistant

Annie Simon [Assistant (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 (algorithms, middleware and system services) for large distributed infrastructures.

MESCAL's target applications are intensive scientific computations such as particle detection, combinatorial optimization, Monte Carlo simulations, and others with a recent focus on nano-simulations. 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 systems 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,
- Mean field 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.

[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.

3.1.2. 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.

3.1.2.1. SimGrid

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).

SimGrid is a toolkit that provides core functionalities for the simulation of distributed applications in heterogeneous distributed environments. The specific goal of the project is to facilitate research in the area of distributed and parallel application scheduling on distributed computing platforms ranging from simple network of workstations to Computational Grids.

We have released two new major versions of SimGrid. 3.4 in April 2010 and 3.5 in December 2010. Both versions include our current work on visualization and analysis of large scale distributed systems. Version 3.5 also includes more recent work on platform modeling that enables highly scalable (in term of memory) simulations. With such improvements, it is now possible to simulate, in an extremely precise and realistic way, Volunteer computing platforms comprising dozens of thousands.

3.1.2.2. 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.

3.1.3. 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 evaluation of work stealing in large systems.

3.1.4. 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.

3.1.5. 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.

3.1.6. 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, Jean-François Méhaut.

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. Nano simulations

Participant: Jean-François Méhaut.

we have analyzed an electronic structure simulation application. The simulation of the structure and of the material property and molecules is based on quantum mechanics and more specifically on Shrodinger's equation. Our aim was to characterize as accurately as possible the performance and the behavior of this application so as to determine its optimal platform configuration. The cluster nodes are hierarchical multi-core SMPs and share a hierarchical memory (NUMA). These experiments have been conducted on two types of NUMA SMPs based either on Intel Itanium or on AMD Opteron CPUs using three different Fortran compilers (two commercial ones and a free one).

4.4. Seismic simulations

Participant: Jean-François Méhaut.

Numerical modeling of seismic wave propagation in complex three-dimensional media is an important research topic in seismology. Several approaches will be studied, and their suitability with respect to the specific constraints of NUMA architectures shall be evaluated. These modeling approaches will rely on modern numerical schemes such as spectral elements, high-order finite differences or finite elements applied to realistic 3D models. The NUMASIS project (see Section 8.2.3) will focus on issues related to parallel algorithms (distribution, scheduling) in order to optimize computations based on such numerical schemes by taking advantage of execution frameworks developed for NUMA architectures.

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

4.5. Electromagnetic Fields simulations

Participant: Yves Denneulin.

We study scaling properties in electromagnetism simulation applications and grids. This work is the main objective of our ANR MEG project. We have shown how to deploy computational electromagnetic applications on grid computing architectures [5]. We also have designed Parallelization of the Scale Changing Technique in Grid Computing environment for the Electromagnetic Simulation of Multi-scale Structures [8].

4.6. 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. Triva

TRIVA is an open-source tool used to analyze traces (in the pajé format) registered during the execution of parallel applications. The tool serves also as a sandbox to the development of new visualization techniques. Some features include:

- Temporal integration using dynamic time-intervals
- Spatial aggregation through hierarchical traces
- Scalable visual analysis with Squarified Treemaps
- A Custom Graph Visualization

We have released three stable versions of Triva this year: 1.0 in April 2010, 1.1 in June 2010, and 1.2 in October 2010.

5.1.2. 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).

5.1.3. 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.

5.1.4. 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 8.2.3).

5.1.5. 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).

With the increasing functionality, scale, and complexity of distributed systems, resource failures are inevitable. While numerous models and algorithms for dealing with failures exist, the lack of public trace data sets and tools has prevented meaningful comparisons. To facilitate the design, validation, and comparison of fault-tolerant models and algorithms, we led the creation of the Failure Trace Archive (FTA), an online public repository of availability traces taken from diverse parallel and distributed systems.

While several archives exist, the FTA differs in several respects. First, it defines a standard format that facilitates the use and comparison of traces. Second, the archive contains traces in that format for over 20 diverse systems over a time span of 10 years. Third, it provides a public toolbox for failure trace interpretation, analysis, and modeling [38], [55], [50].

Our conference paper [37] on the FTA received best paper award at CCGrid 2010. The FTA was recently released in November 2009. It has received over 11,000 hits since then. The FTA has had national and international impact. Several published works have already cited and benefited from the traces and tools of the FTA. Simulation toolkits for distributed systems, such as SimGrid (CNRS, France) and GridSim (University of Melbourne, Australia), have incorporated the traces to allow for simulations with failures.

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. 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. Perfect sampling of Markov chains with piecewise homogeneous events

Perfect simulation is very efficient if all the events in the system have monotonicity property. Indeed, if one considers a Jackson queueing networks (JQN) with finite capacity constraints and analyze the temporal computational complexity of sampling from their stationary distribution. In the context of perfect sampling, the monotonicity of JQNs ensures that it is the *coupling time* of extremal sample-paths. In the context of approximate sampling, it is given by the *mixing time*. We give a sufficient condition to prove that the coupling time of JQNs is $\Theta(\sum_i C_i)$, where C_i denotes the capacity (buffer size) of queue i. This condition lets us deal with networks having arbitrary topology, for which the best bound known is exponential in C_i. Then, we use this result to show that the mixing time of JQNs is $\lceil \log_2 \frac{1}{\epsilon} \rceil O(\sum_i C_i)$, where ϵ is a precision threshold. The main idea of our proof relies on a recursive formula on the coupling times of special sample-paths and provides a methodology for analyzing the coupling and mixing times of several monotone Markovian networks.

However, in the general (non-monotone) case, the perfect sampling technique needs to consider the whole state space, which limits its application only to chains with a state space of small cardinality. We propose here a new approach for the general case that only needs to consider two trajectories. Instead of the original chain, we use two bounding processes (envelopes) and we show that, whenever they couple, one obtains a sample under the stationary distribution of the original chain. We show that this new approach is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed. We further show that most Markovian queueing networks have this property and we propose efficient algorithms for some of them [20], [59].

In particular, we constructed the envelopes for phase-type servers [58] as well as for work stealing networks [35] and we showed that the envelope approach for perfect sampling is very efficient.

Another improvement to speed up perfect sampling of Markov chains can be obtained by skipping passive events during the simulation. We show that this can be done without altering the distribution of the samples. This technique is particularly efficient for the simulation of Markov chains with different time scales such as queueing networks where certain servers are much faster than others. In such cases, the coupling time of the Markov chain can be arbitrarily large while the runtime of the skipping algorithm remains bounded. This is further illustrated by several experiments that also show the role played by the entropy of the system in the performance of our algorithm [59].

6.1.2. Fast and Scalable Simulation of Volunteer Computing Systems Using SimGrid

Advances in inter-networking technology and the decreasing cost-performance ratio of commodity computing components have enabled Volunteer Computing (VC). VC platforms aggregate tens or hundreds of thousands of hosts. These hosts are typically volatile, which raises difficult research questions. Most research in this area relies on simulation. The main issue when developing VC simulators is scalability: How to perform simulations of large-scale VC platforms with reasonable amounts of memory and reasonably fast? To achieve scalability, state-of-the-art VC simulators employ simplistic simulation models and/or target on narrow platform and application scenarios. In this paper we enable VC simulations using the general-purpose SIMGRID simulation framework, which provides significantly more realistic and flexible simulation capabilities than the aforementioned simulators. Our key contribution is a set of improvements to SIMGRID so that it brings these benefits to VC simulations while achieving good scalability. Thanks to careful algorithmic and implementation optimizations, SIMGRID simulations prove even much faster than state-of-the-art simulators [27].

This set of improvements has been integrated in stable public version 3.3.4 (Dec. 2009).

6.2. Tools for Performance Evaluation

Participants: Brigitte Plateau, Jean-Marc Vincent.

6.2.1. Steady-state Property Verification

Model checking of probabilistic models can be done either by numerical analysis or by simulation and statistical methods. In this paper, we compare the efficiency and the scalability of different model checking approaches when they are applied to the verification of steady-state properties of large models. We provide an experimental comparison study between the statistical model checking using perfect sampling implemented in psi2 and the numerical method implemented in PRISM, for the verification of CSL steady-state properties. We show that the proposed statistical approach lets us to consider very large models [28].

6.2.2. Queuing Based Performance Awareness in Autonomic Systems

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

We advocate for the introduction of performance awareness in autonomic systems. The motivation is to be able to predict the performance of a target configuration when a self-* feature is planning a system reconfiguration.

We propose a global and partially automated process based on queues and queuing networks models. This process includes decomposing a distributed application into black boxes, identifying the queue model for each black box and assembling these models into a queuing network according to the candidate target configuration. Finally, performance prediction is performed either through simulation or analysis.

This paper sketches the global process and focuses on the black box model identification step. This step is automated thanks to a load testing platform enhanced with a workload control loop. Model identification is then based on statistical tests. The model identification process is illustrated by experimental results. The paper [36] got the best paper award at ICAS 2010.

6.2.3. Performance Evaluation of Streaming Applications

In this paper, we investigate how to compute the throughput of probabilistic and replicated streaming applications. We are given (i) a streaming application whose dependence graph is a linear chain; (ii) a one-to-many mapping of the application onto a fully heterogeneous target platform, where a processor is assigned at most one application stage, but where a stage can be replicated onto a set of processors; and (iii) a set of IID (Independent and Identically-Distributed) variables to model each computation and communication time in the mapping. How can we compute the throughput of the application, i.e., the rate at which data sets can be processed? We consider two execution models, the STRICT model where the actions of each processor are sequentialized, and the OVERLAP model where a processor can compute and communicate in parallel.

The problem is easy when application stages are not replicated, i.e., assigned to a single processor: in that case the throughput is dictated by the critical hardware resource. However, when stages are replicated, i.e., assigned to several processors, the problem becomes surprisingly complicated: even in the deterministic case, the optimal throughput may be lower than the smallest internal resource throughput. Our first main contribution is to provide a general method to compute the throughput when mapping parameters are constant or follow IID exponential laws. The second main contribution is to provide bounds for the throughput when stage parameters are arbitrary IID and NBUE (New Better than Used in Expectation) variables: the throughput is bounded from below by the exponential case and bounded from above by the deterministic case [18].

6.3. Distributed Computing Platforms: Measurements and Models

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

6.3.1. Availability Modeling of 100,000+ Node Systems.

Clearly, distributed systems (collections of Grids, data centers, volunteer computing systems, and hybrid systems) on the scale of 100,000+ nodes exhibit uncertain availability. We developed ways to statistically model the availability of individual hosts in 100,000+ node systems. We modeled relatively deterministic patterns and relatively random patterns, each of which requires different techniques. We also applied machine learning techniques to predict online the availability of collections of hosts.

This work differs from others in terms of what is measured and what is modeled. In terms of measurements, this work differs by the types of resources measured (Internet end hosts versus only hosts within an enterprise) and the scale and duration (hundreds of thousands of hosts over 1.5 years versus hundreds over weeks). In terms of modeling, most related works focus on modeling the system as a whole instead of individual resources, or the modeling does not capture the temporal structure of availability. Yet models of individual resources are essential for effective resource selection and scheduling.

Our modeling and prediction techniques are useful for stochastic scheduling in terms of both static analysis and online predictions. Our techniques have been implemented in BOINC (the software infrastructure of UC Berkeley for SETI@home, EINSTEIN@home, climateprediction.net, and other projects) for resource management. In total, these works on availability modeling have been cited over 400 times according to Google Scholar. Our traces of several volunteer computing systems such as SETI@home have been made publicly available at the Failure Trace Archive and have been extensively used by other researchers [7].

6.3.2. Supporting Malleability in Parallel Architectures

Current parallel architectures take advantage of new hardware evolution, like the use of multicore machines in clusters and grids. The availability of such resources may also be dynamic. Therefore, some kind of adaptation is required by the applications and the resource manager to perform a good resource utilization. Malleable applications can provide a certain flexibility, adapting themselves on-the-fly, according to variations in the amount of available resources. However, to enable the execution of this kind of applications, some support from the resource manager is required, thus introducing important complexities like special allocation and scheduling policies. Under this context, we investigate some techniques to provide malleable behavior on MPI applications and the impact of this support upon a resource manager. Our study deals with two approaches to obtain malleability: dynamic CPUSETsmapping and dynamic MPI, using the OAR resource manager. The validation experiments were conducted upon Grid5000 platform. The testbed associates the charge of real workload traces and the execution of MPI benchmarks. Our results show that a dynamic approach using malleable jobs can lead to almost 25 % of improvement in the resources utilization, when compared to a non-dynamic approach. Furthermore, the complexity of the malleability support, for the resource manager, seems to be overlapped by the improvement reached [21].

6.3.3. Monetary Costs: Economics of Distributed Computing

We have studied the tradeoffs in terms of monetary costs, reliability, and performance of different distributed systems, such as Amazon's Elastic Compute Cloud, Amazon's Spot Instances and SETI@home. Also, with my collaborators, we have investigated how these tradeoffs can be leveraged in hybrid computing systems that combine Clouds with Grids or Volunteer Computing Systems.

Previously, the exact quantitative costs for Cloud Computing Systems and Volunteer Computing Systems for high-throughput applications had not been known, nor had these costs been evaluated in the context of the performance and reliability of these systems. Yet, the monetary costs and the associated performance and reliability levels provided by different types of distributed systems are essential inputs for their users.

Our cost decision models give users of these systems insights about what users get for their money in terms of performance and reliability. The models also show how mixtures of the platforms (for instance dedicated Cloud node versus shared volunteer nodes) can affect cost, performance, or reliability of the systems. [14], [51], [13], [54].

6.4. Multi-User Systems

Participants: Bruno Gaujal, Arnaud Legrand, Jean-François Méhaut, Corinne Touati.

6.4.1. Optimal Mean-Field approximation

We studied the convergence of Markov Decision Processes made of a large number of objects to optimization problems on ordinary differential equations (ODE). We show that the optimal reward of such a Markov Decision Process, satisfying a Bellman equation, converges to the solution of a continuous Hamilton-Jacobi-Bellman (HJB) equation based on the mean field approximation of the Markov Decision Process. We give bounds on the difference of the rewards, and a constructive algorithm for deriving an approximating solution to the Markov Decision Process from a solution of the HJB equations. We illustrate the method on three examples pertaining respectively to investment strategies, population dynamics control and scheduling in queues. They are used to illustrate and justify the construction of the controlled ODE and to show the gain obtained by solving a continuous HJB equation rather than a large discrete Bellman equation. [6] This work was applied to study the performance behavior of working stealing over very large heterogeneous platforms [32].

6.4.2. Mean-Field limits and differential inclusions

We have extended the applicability of deterministic limits of Markov processes made of several interacting objects. While most classical results assume that the limiting dynamics has Lipschitz properties, we have shown that these conditions are not necessary to prove convergence to a deterministic system.

We show that under mild assumptions, the stochastic system converges to the set of solutions of a differential inclusion and we provide simple ways to compute the limiting inclusion. When this differential inclusion satisfies a one-sided Lipschitz condition, there exists a unique solution of this differential inclusion and we show convergence in probability with explicit bounds.

This extends the applicability of mean field techniques to systems exhibiting threshold dynamics such as queuing systems with boundary conditions or controlled dynamics. This is illustrated by applying our results to several types of systems: fluid limits of priority queues, best response dynamics in games, push-pull queues with a large number of sources and a large number of servers and self-adapting computing systems [33].

6.4.3. The Price of Forgetting in Parallel Queues

We consider a broker-based network of non-observable parallel queues and analyze the problem of finding the minimum response time and the optimal routing policy. The broker has the memory of its previous routing decisions. We provide lower bounds on the optimal response time by means of convex programming that are tight, as follows by a numerical comparison with a proposed routing scheme. We introduce the "Price of Forgetting" (PoF), the ratio between the optimal response times achieved by a probabilistic broker and a broker with memory, is shown to strongly depend on the coefficient of variation of the service time distributions. In the case of general service times, the PoF can be unbounded or arbitrarily close to one. In the case of exponential service times, the PoF is bounded from above by two, which is tight in heavy-traffic, and independent of the network size and heterogeneity. These properties yield a simple engineering product-form approximating tightly the optimal response time.

Finally, we put our results in the context of game theory revisiting the "Price of Anarchy" for such systems: It is bounded from above by the product of the PoA achieved by probabilistic brokers (already well understood) and the PoF [16]. The work [17] has won the best poster award at sigmetrics 2010.

6.4.4. Self-optimizing Routing in MANETs with Multi-class Flows

We show [24] how game theory and Gibbs sampling techniques can be used to design a self-optimizing algorithm for minimizing end-to-end delays for all flows in a multi-class mobile ad hoc network (MANET).

This is an improvement over the famed Ad-Hoc On-demand Distance Vector (AODV) protocol, that computes the routes with minimal number of hops for each flow in a multi-flow ad-hoc network. Here, the load of each flow is taken into account to choose the best route (in terms of delays) among a fixed number of routes. The algorithm can be implemented in a fully distributed and asynchronous way and is guaranteed to converge to the global optimal configuration. Numerous numerical experiments show that the gain over AODV, computed over a large number of networks, is quite substantial.

6.4.5. Robust Control Problem with Infinitely Many Mobile Agents

In [11], we consider a two-hop routing delay-tolerant network in the presence of a malicious intruder. When the source encounters a mobile then it transmits, with some probability, a file to that mobile, with the probability itself being a decision variable. The number of mobiles is not fixed, with new mobiles arriving at some constant rate. The file corresponds to some software that is needed for offering some service to some clients, which themselves may be mobile or fixed. We assume that mobiles have finite life time due to limited energy, but that the rate at which they die is unknown (and could be controlled by an adversary). We formulate this problem as a robust control problem which we transform into a zero-sum differential game, and obtain its value as well as the saddle-point policies for both players.

6.4.6. Practical Implementation of Distributed Optimization Techniques

In [19], we illustrate practical issues arising in the development of efficient implementation of distributed algorithms that solve a general (concave) constrained maximization problem. Such optimizations arise in many situations. One typical example is those of resource allocation in computer networks, where the system aims at maximizing some global function of the users individual throughput subject to link capacity constraints.

6.5. Programming Many-core Systems

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

6.5.1. Memory Issues on Multicore Architectures

Transactional Memory (TM) is a new programming paradigm that offers an alternative to traditional lock-based concurrency mechanisms. It offers a higher-level programming interface and promises to greatly simplify the development of correct concurrent applications on multicore architectures. However, simplicity often comes with an important performance deterioration and given the variety of TM implementations it is still a challenge to know what kind of applications can really take advantage of TM. In order to gain some insight on these issues, helping developers to understand and improve the performance of TM applications, we propose a generic approach for collecting and tracing relevant information about transactions. Our solution can be applied to different Software Transactional Memory (STM) libraries and applications as it does not modify neither the target application nor the STM library source codes. We show that the collected information can be helpful in order to comprehend the performance of TM applications [57].

Nowadays, on Multi-core Multiprocessors with Hierarchical Memory (Non-Uniform Memory Access (NUMA) characteristics), the number of cores accessing memory banks is considerably high. Such accesses produce stress on the memory banks, generating load-balancing issues, memory contention and remote accesses. In this context, how to manage memory accesses in an efficient fashion remains an important concern. To reduce memory access costs, developers have to manage data placement on their application assuring memory affinity. The problem is: How to guarantee memory affinity for different applications/NUMA

platforms and assure efficiency, portability, minimal or none source code changes (transparency) and fine control of memory access patterns? Our research have led to the proposal of Minas: an efficient and portable memory affinity management framework for NUMA platforms. Minas provides both explicit memory affinity management and automatic one with good performance, architecture abstraction, minimal or none application source code modifications and fine control. We have evaluated its efficiency and portability by performing some experiments with numerical scientific HPC applications on NUMA platforms. The results have been compared with other solutions to manage memory affinity [45].

We have evaluated our solution on two NUMA platforms using two geophysics parallel applications. The results show some performance improvements in comparison with other solutions available for Linux [42].

We also proposed a portable preprocessor to enhance memory affinity on multi-core NUMAs for OpenMP applications. Our preprocessor extracts information from the application at compile time (e.g. shared variables, mode of access in a parallel section) and transforms the application source code to map variables in function of hardware characteristics. We have evaluated our preprocessor by performing experiments with some benchmarks on two multi-core NUMA platforms. Our results show that benchmarks optimized with the preprocessor can present performance gains [43].

Finally, it is important to get a good understanding of memory access patterns and what are the influences of data placement on such patterns. We have investigated memory accesses behavior of microbenchmarks and benchmarks over a ccNUMA platform with multi-core processors. Additionally, we have evaluated a set of memory policies that were used to place data among the machine memory banks. Our results have shown that an appropriate selection of data placement, considering the memory accesses, can generated great improvement gains [60].

6.5.2. Sharing and Mutualizing Clusters

For software vendors who need to provide their softwares as services via Internet, an infrastructure of high performance computing (HPC) such as clusters is required. However, for small and medium enterprises (SMEs) and/or startup businesses, owning a cluster is generally out of reach. Indeed, the cost of a cluster can be very high, since enterprises have to deal with acquisition costs, as well as many operating costs (engineering, power supply, air conditioning, etc.). The emergence of infrastructure providers like Amazon, Google, IBM, Sun, etc. allows businesses to use remote infrastructures. However, in the long term, renting those infrastructures can also be expensive. In order to lower the cost, an alternative solution might be for small businesses to join in order to purchase and maintain a common infrastructure that would be shared among them. In this case, each partner has to have the guarantee that their use of the infrastructure would be equitable, rational and proportional to their investment. Additionally, customers would expect the service/application to be cheaper and to have a good performance. In principle, infrastructure sharing is not simple to manage. In this context, we have defined two approaches concerning the equitable sharing of a cluster among several concurrent softwares hosted as services. The first approach is based on the static partitioning of resources, and the second approach is based on the dynamic resource allocation with dynamic priorities among applications. This work has been carried out in cooperation with an industrial project named CILOE1. Such a project aims at providing a shared computing cluster to small editors of electronic design automation (EDA) and embedded softwares [22].

6.6. Middleware and Experimental Testbeds

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

6.6.1. Charm++ on NUMA Platforms

Cache-coherent Non-Uniform Memory Access (NUMA) platforms based on multi-core chips are now a common resource in High Performance Computing. To overcome scalability issues in such platforms, the shared memory is physically distributed among several memory banks. Its memory access costs may vary depending on the distance between processing units and data. The main challenge of a NUMA platform is to manage efficiently threads, data distribution and communication over all the machine nodes. Charm++ is

a parallel programming system that provides a portable programming model for platforms based on shared and distributed memory. In this work, we revisit some of the implementation decisions currently featured on Charm++ on the context of ccNUMA platforms. First, we studied the impact of the new – shared-memory based – inter-object communication scheme utilized by Charm++. We show how this shared-memory approach can impact the performance of Charm++ on ccNUMA machines. Second, we conduct a performance evaluation of the CPU and memory affinity mechanisms provided by Charm++ on ccNUMA platforms. Results show that SMP optimizations and affinity support can improve the overall performance of our benchmarks in up to 75%. Finally, in light of these studies, we have designed and implemented a NUMA-aware load balancing algorithm that address the issues found. The performance evaluation of our prototype showed results as good as the ones obtained by GreedyLB and significant improvements when compared to GreedyCommLB [56], [44].

6.6.2. Deadline-Constrained Checkpointing in a Batch Scheduler

We have integrated scheduling of deadline-constrained checkpointing in a batch scheduler for dynamic environments such as virtual clusters. The checkpointing scheduler implemented focuses on the parallel checkpointing on a unique server of long-running independent applications in a virtual cluster made up of free resources for long periods of an intranet network, assuming that the resources must be released within a delay T. As parallel checkpointing on a unique server can face bandwidth constraints, the checkpointing scheduler uses a function that gives the aggregated bandwidth suitable for the parallel checkpointing of m applications of aggregated size V to solve the deadline-constrained checkpointing problem within the deadline T . Specifically, we present the integration of the checkpointing scheduler in the batch scheduler OAR. This implementation uses data from the OAR database for the checkpointing scheduling. It is portable and can be easily modified to interact with any other batch scheduler, provided that the structure of the database is known and an estimator of the bandwidth of the system suitable for parallel checkpointing available. Experimental results obtained on a virtual cluster built on GRID5000 show that the implementation of the checkpointing scheduler does not induce a significant overhead on checkpointing mechanisms. As a consequence, this work aims at providing HPC platforms for a tool to enhance the quality of services offered to end users [41].

6.6.3. Debugging Embedded Linux Kernel Through JTAG Port

We have designed a JTAG-based debugger LKD with extended functionality for debugging the Linux kernel and its modules. LKD is especially designed to debug embedded systems based on Linux and to overcome the limits of conventional solutions like printk and KGDB. Embedded Linux systems are commonly extensions of the Linux kernel, either through modules or modified kernel functions to manage specific components of the embedded system. A debugging solution is fundamental in these cases to cope with such extensions.

LKD only requires peek and poke operations over JTAG port. LKD is composed, on one side of a Linux awareness layer on top of a standard GDB client, and on the other side a target specific layer using the JTAG port of a specific platform. The Linux awareness layer includes features like: virtual memory management, modules handling, user process awareness etc. These features largely reduces the time of the debugging phase and helps to respect time to market constraints [34].

7. Contracts and Grants with Industry

7.1. Four CIFRE contracts with BULL

- Yiannis Georgiou has done his PhD thesis in a CIFRE contract with the BULL company. His work started in September 2006, and he defended his thesis in November 2010. The focus of his research is batch scheduling on Grids.
- Jérôme Vienne did his phD thesis with BULL from September 2006 to July 2010. He worked on scheduling tasks on multicore computers.

- Gaël Gorgo started his PhD with BULL in October 2010. He works on performance models for new computer architectures
- Mathieu Ospici started his PhD with BULL in 2008. He works on the bigDFT project.

7.2. Two CIFRE contracts with France Télécom R&D

- 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 defend in thesis in 2011. He works in load injection and performance evaluation issues in networks.
- Charbel El Kaed is doing his PhD thesis in France Télécom on the usage of communication devices.

7.3. Four CIFRE with STMicroelectronics

- Carlos Rojas is doing his PhD thesis under a CIFRE contract with STMicroelectronics. He started in September 2007 and will finish in 2011. The objective of his thesis is to develop methods and tools for multiprocessor embedded applications.
- Kiril Georgiev is doing his PhD with STMicroelectronics on distributed file systems.
- Patricia Cueva has started her PhD with STMicroelectronics on high performance computing.
- Kevin Pouget has started his PhD with STMicroelectronics on multi-core computers.

7.4. Real-Time-At -Work

RealTimeAtWork.com is a startup from INRIA Lorraine created in December 2007. Bruno Gaujal is a scientific partner and a founding member of 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.

7.5. Selfnets Research action with Alcatel

Selfnets is an ADR (action de recherche) of the common laboratory between INRIA and Alcatel Lucent Bell Labs. Bruno Gaujal is co-leading the action with Vincent Rocca. Selfnets is mainly concerned with self-optimizing wireless networks (Wifi, 3G, LTE). Eight INRIA teams are participating in Selfnets. As for Mescal, we mainly work on recent mobile equipment (e.g. using the norm IEEE 802.21) can freely switch between different technologies (vertical handover). This allows for some flexibility in resource assignment and, consequently, increases the potential throughput allocated to each user. We develop and analyse fully distributed algorithms based on evolutionary games that exploit the benefits of vertical handover by finding fair and efficient user-network association schemes.

Our contributions are on four different levels.

1. On a modelisation level - we propose a multi-population asymmetric game model for the association problem in heterogeneous wireless networks.

2. On a dynamic point of view - we show that, in our case, the replicator dynamics always converges to Nash equilibria. Furthermore, the limit points corresponds to monomorphic populations (where all packets of a given mobile use a single cell).

3. On an algorithmic level - we propose a fully distributed algorithm that converges to the limit points of the replicator dynamics. This algorithm is not based on a gradient discretization of the replicator dynamics but rather on a constant stepsize stochastic approximation, so as to avoid numerous (costly) handovers operated by the algorithm.

4. On an experimental point of view - we provide an extensive numerical study of several scenarios assessing the quality of our approach in terms of: scalability of the convergence speed with respect to the size of the system.

A patent on a simplified version of our algorithm has been taken by the common lab.

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.

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.

8.2. National initiatives

8.2.1. ADT HEMERA, 2010-2012

Leading action "Completing challenging experiments on Grid'5000 (Methodology)"

Experimental platforms like Grid'5000 or PlanetLab provide an invaluable help to the scientific community, by making it possible to run very large-scale experiments in con- trolled environment. However, while performing relatively simple experiments is generally easy, it has been shown that the complexity of completing more challenging experiments (involving a large number of nodes, changes to the environment to introduce heterogeneity or faults, or instrumentation of the platform to extract data during the experiment) is often underestimated.

This working group will explore different complementary approaches, that are the basic building blocks for building the next level of experimentation on large scale experimental platforms. This encompasses several aspects.

8.2.2. ADT INRIA, 2009-2011, SimGrid for Human Beings

Partners: INRIA Grand Est. Two young engineers have been allotted by the INRIA to the SimGrid project to help with the software maintenance and with the transfer of research ideas and prototypes from the ANR USS SimGrid to public stable versions.

8.2.3. NUMASIS, 2005-2010, 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.

8.2.4. Check-bound, 2007-2010 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.

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

8.2.5. 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.

8.2.6. 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.

8.2.7. 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.

8.2.8. 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.

[3] Design and Optimization of Collaborative Computing Architectures

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.

8.2.9. 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 availability 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.

8.2.10. 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 recently emerged. These new and heterogeneous architectures require new application 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.

8.2.11. 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.

8.2.12. 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 instantiation. 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.

8.2.13. 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.

8.2.14. 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.

FP7 EDGI (European Desktop Grid Initiative) 2010-2012

Partners: SZTAKI insitute (Hungary), CIEMAT (Spain), Univ. Coimbra (Portugal), Univ Cardi (UK), Univ Westminster (UK), AlmereGrid (NL), IN2P3 (FR), INRIA (GRAAL, MESCAL)

EDGI is an FP7 European project whose goal is to build a Grid infrastructure composed of "Desktop Grids", such as BOINC or XtremWeb, where computing resources are provided by Internet volunteers, and "Service Grids", where computing resources are provided by institutional Grid such as EGEE, gLite, Unicore and "Clouds systems" such asOpenNebula and Eucalyptus, where resources are provided on-demand. The EDGI infrastructure will consist of Service Grids that are extended with public and institutional Desktop Grids and Clouds. Our partners include SZTAKI insitute (Hungary), CIEMAT (Spain), Univ. Coimbra (Portugal), Univ Cardi (UK), Univ Westminster (UK), AlmereGrid (NL), IN2P3 (FR) and more.

European Exascale Software Initiative (EESI) The objective of this Support Action, co-funded by the European Commission is to build a european vision and roadmap to address the challenges of the new generation of massively parallel systems composed of millions of heterogeneous cores which will provide Petaflop performances in 2010 and Exaflop performances in 2020 (the speed of a supercomputer is measured in "FLOPS" (FLoating Point Operations Per Second)), "Petascale" supercomputers can process one quadrillion (1015) (1000 trillion) FLOPS, Exascale is computing performance is one quintillion (1018) FLOPS (one million teraflops).

http://www.eesi-project.eu/pages/menu/homepage.php

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.

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). Arnaud Legrand visited David Anderson at UC Berkeley and Alan Snavely at the San Diego Supercomputer Center for a week.

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

Amazon Inc., 35000 server hours: (2010-2011) The overall goal is to integrate G5K with Amazon Inc's Elastic Compute Cloud (EC2) such that workload, especially during peak periods, can be rerouted to EC2. So we would like to adapt OAR for an on-demand cloud infrastructure. We envision an OAR server, running within G5K, that manages sites within G5K and remote instances in EC2.

This will require the several additions to OAR:

- Enable deployments over EC2 instances with OAR.
- Enabling job data management.
- Enable rerouting of jobs from G5K to EC2.

Common Laboratory INRIA-UIUC and NCSA Several members of Mescal are partners of this laboratory, and have paid several visits to Urbana-Champaign. The next workshop of the laboratory will be organized by Jean-François Méhaut in Grenoble.

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, Arnaud Legrand, 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 Grenoble INP correspondent for student exchanges with Japan. She spent two months as a visiting researcher in the University of Tsukuba and Kyoto University and she tutored two ENSIMAG students that spent their second year of master in Kyoto University.

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

Brigitte Plateau has been appointed director of Grenoble-INP ENSIMAG.

9.1.1. Invited talks

Researchers of the MESCAL project-team have been invited to give plenary talks on their research subjects in international conferences:

- Arnaud Legrand was the keynote speaker at the Performance Evaluation of Large Scale Distributed Systems at CLCAR 2010.
- SimGrid and Triva have been presented for two years at the SuperComputing conference (SC 2009 and 2010) within the INRIA booth.
- Jean-Marc Vincent gave an invited lecture at NetCoop 2010.
- Nicolas Gast was an invited speaker at NetCoop 2010 and at YEQT 2010.
- Bruno Gaujal gave an invited lecture at AlgoGT 2010.

9.1.2. Journal, Conference and Workshop Chairing

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

- Co-panel chair on Energy-aware, Power-aware, and Green Computing for Large Distributed Systems and Applications at the International Conference on High Performance Computing & Simulation (HPCS), Caen, France, June 2010 (Derrick Kondo)
- General co-chair of the 4th Workshop of Large-Scale and Volatile Desktop Grids (PCGRID'10), Melbourne, Australia, May 2010. (Derrick Kondo)
- Spring School: SimGrid User Days, Cargese, April 13-15, 2010 - Cargese The SimGrid project has organized a workshop as part of the USS-SimGrid ANR project. These few days have been an opportunity for advanced users to meet developers and discuss the latest features of SimGrid and currently explored research areas as well as to influence future developments. This event has gathered participants from all around the France and from Switzerland (CERN, Neuchatel), Belgium (Antwerpe), Brazil (Porto Alegre), and Italy (Piemonte Orientale).
- Chairing committee of the national journal Technique et sciences informatiques (Ed: Hermes/Lavoisier)
- General co-chair of the Workshop on Algorithmic Game Theory: Dynamics and Convergence in Distributed Systems, Bordeaux, France (Corinne Touati).

9.1.3. Program committees

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

- Program Committee member for the 2010 International Workshop on HPC and Grid Applications (IWHGA2010) IEEE
- Program Committee member for the 10th International Symposium on Cluster Computing and the Grid (CCGrid'10), Melbourne, Australia, May 2010.
- Program Committee member for the 3rd International Joint Conference on Computational Sciences and Optimization (CSO'10), HuangShan, China, May 2010.
- Program Committee member for the 2nd Workshop on Service-Oriented P2P Networks and Grid Systems (ServP2P'10), Melbourne, Australia, May 2010.
- Program Committee member for the Workshop on Bio-Inspired Algorithms for Distributed Systems (BADS'10), Washington, DC, USA, June 2010.
- Program Committee member for the International Workshop on Volunteer Resource Computing, Cloud Computing and Internet of Things (VCI'10), Nanjing, China, November 2010.
- Program Committee member of IPDPS 2010.
- Rencontres Francophones du Parallélisme 2010, RenPar.

9.1.4. Thesis defense

- Nicolas Gast
- Hussien Joumaa
- Jérôme Vienne
- Yannis Georgiou

9.1.5. Thesis committees

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

- Derrick Kondo served as a thesis committee of Fatiha Bouabache and Paul Malécot.
- Bruno Gaujal served on the thesis committee of Louis-Claude Canon (Rapporteur) and Jean-Philippe Gayon (HDR).
- Jean-François Méhaut was in the thesis committee of J. Perez (rapporteur), F. Broquedis (rapporteur), J. Bigot (rapporteur), F. Diakhate, F. Dupros. He was president of the jury for the defense of J. Arnaud, Y. Falcone and L. Touseau.
- Olivier Richard served as a referee the thesis committee of Paul Macelotand and Ghislain Charrier.
- Corinne Touati served as a referee in the thesis committee of Walid Saad.

9.1.6. Members of editorial board

- Co-guest editor of Journal of Future Generation Computer Systems, Special issue on Desktop Grids and Volunteer Computing, 2010.

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.

9.1.8. Popular Science

MESCAL team actively promotes science to young and non-scientific audience by participating to the "Fête de la Science" 2010 (booth "Fermez des routes, vous irez plus vite").

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 ENS Lyon** Corinne Touati gives a course on game theory for networks
- **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)**
- Coordinator of the RICM 1st year
- Coordinator of the "Réseaux informatiques et Communication multimédia" section, University Joseph Fourier, Grenoble

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, Springer-Verlag, 2003, n° 1829.

[2] N. GAST, B. GAUJAL. *A Mean Field Approach for Optimization in Discrete Time*, in "Journal of Discrete Event Dynamic Systems", 2010, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/jdeds2010.pdf.

[3] B. JAVADI, D. KONDO, J.-M. VINCENT, D. ANDERSON. *Discovering Statistical Models of Availability in Large Distributed Systems: An Empirical Study of SETI@home*, in "IEEE Transactions on Parallel and Distributed Systems", 2010.

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

Publications of the year

Articles in International Peer-Reviewed Journal

[5] C. BARRIOS HERNANDEZ, F. KHALIL, Y. DENNEULIN, H. AUBERT, F. COCCETTI, R. PLANA. *Deployment of computational electromagnetics applications on grid computing architectures*, in "Acta Científica Venezolana", 2010.

[6] N. GAST, B. GAUJAL. *A Mean Field Approach for Optimization in Discrete Time*, in "Journal of Discrete Event Dynamic Systems", 2010, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/jdeds2010.pdf.

[7] B. JAVADI, D. KONDO, J.-M. VINCENT, D. ANDERSON. *Discovering Statistical Models of Availability in Large Distributed Systems: An Empirical Study of SETI@home*, in "IEEE Transactions on Parallel and Distributed Systems", 2010.

[8] F. KHALIL, H. AUBERT, F. COCCETTI, P. LORENZ, R. PLANA, C. BARRIOS HERNANDEZ, Y. DENNEULIN. *Parallelization of the Scale Changing Technique in Grid Computing environment for the Electromagnetic Simulation of Multi-scale Structures*, in "International Journal of Numerical modeling", 2010.

[9] L. M. SCHNORR, G. HUARD, P. O. A. NAVAUX. *Triva: Interactive 3D visualization for performance analysis of parallel applications*, in "Future Generation Computer Systems", 2010, vol. 26, n^{o} 3, p. 348 - 358.

Invited Conferences

[10] J.-M. VINCENT. *Scheduling Strategies in Large Scale Heterogeneous Grids: Monotonicity and Perfect Sampling*, in "NETCOOP", Ghent, December 2010.

International Peer-Reviewed Conference/Proceedings

[11] E. ALTMAN, A. ARAM, T. BASAR, C. TOUATI, S. SARKAR. *Robust Control in Sparse Mobile Ad-Hoc Networks*, in "Conference on Decision and Game Theory for Security (GameSec)", 2010.

[12] A. ANDRZEJAK, D. KONDO, D. P. ANDERSON. *Exploiting non-dedicated resources for cloud computing*, in "NOMS", 2010, p. 341-348.

[13] A. ANDRZEJAK, D. KONDO, D. P. ANDERSON. *Exploiting non-dedicated resources for cloud computing*, in "IEEE/IFIP Network Operations and Management Symposium (NOMS)", 2010, p. 341-348.

[14] A. ANDRZEJAK, D. KONDO, S. YI. *Decision Model for Cloud Computing under SLA Constraints*, in "MASCOTS", 2010, p. 257-266.

[15] A. ANDRZEJAK, D. KONDO, S. YI. *Decision Model for Cloud Computing under SLA Constraints*, in "IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)", 2010, p. 257-266.

[16] J. ANSELMI, B. GAUJAL. *Optimal Routing in Parallel, non-Observable Queues and the Price of Anarchy Revisited*, in "22nd International Teletraffic Congress (ITC)", Amsterdam, 2010, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/anselmi10ITC.pdf.

[17] J. ANSELMI, B. GAUJAL. *The Price of Anarchy in Parallel Queues Revisited*, in "ACM sigmetrics", New-York, 2010, Short paper.

[18] A. BENOIT, F. DUFOSSÉ, M. GALLET, B. GAUJAL, Y. ROBERT. *Computing the throughput of probabilistic and replicated streaming applications*, in "22nd Symposium on Parallelism in Algorithms and Architectures (SPAA)", Santorini, Greece, 2010.

[19] R. BERTIN, P. COUCHENEY, A. LEGRAND, C. TOUATI. *Practical Implementation Issues of Lagrangian Based Distributed Optimization Algorithms*, in "12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (Synasc)", 2010.

[20] A. BUSIC, B. GAUJAL, G. GORGO, J.-M. VINCENT. *PSI2 : Envelope Perfect Sampling of Non Monotone Systems*, in "International Conference on Quantitative Evaluation of Systems (QEST)", Williamsburg, USA, 2010, Tool presentation paper.

[21] M. CERA, Y. GEORGIOU, O. RICHARD, N. MAILLARD, P. O. A. NAVAUX. *Supporting Malleability in Parallel Architectures with Dynamic CPUSETs Mapping and Dynamic MPI*, in "10th International Conference on Distributed Computing and Networking (ICDCN 2010)", Lecture Notes in Computer Science (LNCS), Springer, 2010, p. 242-257.

[22] R. CHAKODE, J.-F. MÉHAUT, F. CHARLET. *High Performance Computing on Demand: Sharing and Mutualizing Clusters*, in "24th IEEE International Conference on Avanced Information Networking and Applications (AINA'2010)", Perth, Australia, 2010, p. 126-133.

[23] D. CORDEIRO, G. MOUNIÉ, S. PERARNAU, D. TRYSTRAM, J.-M. VINCENT, F. WAGNER. *Random graph generation for scheduling simulations*, in "Proceedings of 3rd International ICST Conference on Simulation Tools and Techniques (SIMUTools 2010)", Malaga Spain, ICST, March 2010, 10, http://hal.archives-ouvertes.fr/hal-00471255/PDF/ggen.pdf.

[24] P. COUCHENEY, B. GAUJAL, C. TOUATI. *Self-optimizing Routing in MANETs with Multi-class Flows*, in "21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)", Istanbul, 2010, invited paper, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/pimrc2010.pdf.

[25] G. DA-COSTA, M. DIAS DE ASSUNCAO, J.-P. GELAS, Y. GEORGIOU, L. LEFEVRE, A.-C. ORGERIE, J.-M. PIERSON, O. RICHARD, A. SAYAH. *Multi-facet approach to reduce energy consumption in clouds and grids: The GREEN-NET Framework*, in "e-Energy 2010 : First International Conference on Energy-Efficient Computing and Networking", Passau, Germany, April 2010.

[26] Y. DENNEULIN, C. LABBÉ, L. D'ORAZIO, C. RONCANCIO. *Merging File Systems and Data Bases to Fit the Grid*, in "Globe, Lecture Notes in Computer Science", 2010, vol. 6265, p. 13-25.

[27] B. DONASSOLO, H. CASANOVA, A. LEGRAND, P. VELHO. *Fast and Scalable Simulation of Volunteer Computing Systems Using SimGrid*, in "Workshop on Large-Scale System and Application Performance (LSAP)", 2010.

[28] D. ELRABIH, G. GORGO, N. PEKERGIN, J.-M. VINCENT. *Steady-state Property Verification: a Comparison Study*, in "4th International Workshop on Verification and Evaluation of Computer and Communication Systems, VECoS", Paris, France, 2010.

[29] G. FEDAK, J.-P. GELAS, T. HERAULT, V. INIESTA, D. KONDO, L. LEFEVRE, P. MALÉCOT, L. NUSSBAUM, A. REZMERITA, O. RICHARD. *DSL-Lab: A Low-Power Lightweight Platform to Experiment on Domestic Broadband Internet*, in "ISPDC", 2010, p. 141-148.

[30] M. GALLET, N. YIGITBASI, B. JAVADI, D. KONDO, A. IOSUP, D. EPEMA. *A Model for Space-Correlated Failures in Large-Scale Distributed Systems*, in "Euro-Par (1)", 2010, p. 88-100.

[31] M. GALLET, N. YIGITBASI, B. JAVADI, D. KONDO, A. IOSUP, D. EPEMA. *A Model for Space-Correlated Failures in Large-Scale Distributed Systems*, in "European Conference on Parallel and Distributed Computing (Euro-Par)", 2010, p. 88-100.

[32] N. GAST, B. GAUJAL. *A Mean Field Model of Work Stealing in Large-Scale Systems*, in "ACM sigmetrics", New-York, 2010, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/sigmetrics2010.pdf.

[33] N. GAST, B. GAUJAL. *Mean field limit of non-smooth systems and differential inclusions*, in "MAthematical performance Modeling and Analysis (MAMA)", New-York, 2010, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/mama2010.pdf.

[34] K. GEORGIEV, M. AUVRAY, S. DE-PAOLI, M. SANTANA, C. SMITH. *Debugging Embedded Linux Kernel Through JTAG*, in "Proceedings of the S4D (System, Software, Soc and Silicon Debug)", Southampton, England, September 2010.

[35] G. GORGO, J.-M. VINCENT. *Perfect Sampling of Load Sharing Policies in Large Scale Distributed Systems*, in "ASMTA, LNCS", Cardiff, 2010, n^{o} 6148, p. 174-188.

[36] A. HARBAOUI, N. SALMI, B. DILLENSEGER, J.-M. VINCENT. *Introducing Queuing Network-Based Performance Awareness in Autonomic Systems*, in "The Sixth International Conference on Autonomic and Autonomous Systems ICAS 2010", Cancun, Mexico, March 2010, Best paper award.

[37] D. KONDO, B. JAVADI, A. IOSUP, D. EPEMA. *The Failure Trace Archive: Enabling Comparative Analysis of Failures in Diverse Distributed Systems*, in "CCGRID", 2010, p. 398-407.

[38] D. KONDO, B. JAVADI, A. IOSUP, D. EPEMA. *The Failure Trace Archive: Enabling Comparative Analysis of Failures in Diverse Distributed Systems*, in "Proceedings of the IEEE International Symposium on Cluster Computing and the Grid (CCGRID)", 2010, p. 398-407.

[39] B. NEGREVERGNE, J.-F. MÉHAUT, A. TERMIER, T. UNO. *Découverte d'itemsets fréquents fermés sur architecture multicoeurs*, in "EGC (Extraction et Gestion des Connaissances)", 2010, p. 465-470.

[40] B. NEGREVERGNE, A. TERMIER, J.-F. MÉHAUT, T. UNO. *Discovering Closed Frequent Itemsets on Multicore: Parallelizing Computations and Optimizing Memory Accesses*, in "Proceedings of HPCS (Intl. Conference on High Performance Computing and Simulation), Special Session on High Performance Parallel and Distributed Data Mining", 2010, p. 521-528.

[41] B. OMER YENKE, J.-F. MÉHAUT, J. MICHEL NLONG II, R. CHAKODE. *Integrating Deadline-Constrained Checkpointing in a Batch Scheduler for Dynamic Environments*, in "Annual International Conference on Software Engineering (SE'2010)", Thailand, 2010.

[42] C. R. POUSA RIBEIRO, M. CASTRO, J.-F. MÉHAUT, A. CARISSIMI. *Improving Memory Affinity of Geophysics Applications on NUMA platforms Using Minas*, in "9th International Meeting High Performance Computing for Computational Science (VECPAR)", 2010.

[43] C. R. POUSA RIBEIRO, N. MAILLARD, I. STANGHERLINI, J.-F. MÉHAUT. *Compiling OpenMP Applications to Enhance Memory Affinity on Hierarchical Multi-Core Machines*, in "23rd International Workshop on Languages and Compilers for Parallel Computing", 2010, poster.

[44] C. R. POUSA RIBEIRO, M. MARTINASSO, J.-F. MÉHAUT. *NUMA Support for the charm++ Environment*, in "8th Workshop on Charm++ and its Applications", 2010.

[45] C. R. POUSA RIBEIRO, J.-F. MÉHAUT, A. CARISSIMI. *Memory Affinity Management for Numerical Scientific Applications over Multi-core Multiprocessors with Hierarchical Memory*, in "PhD Forum of 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS)", 2010.

[46] C. PRADA-ROJAS, M. SANTANA, S. DE-PAOLI, X. RAYNAUD. *Summarizing Embedded Execution Traces through a Compact View*, in "Conference on System Software, SoC and Silicon Debug (S4D)", Southampton, UK, 2010.

[47] K. ROSA BRAGHETTO, J. EDUARDO FERREIRA, J.-M. VINCENT. *Performance Analysis Modeling Applied to Business Processes*, in "SpringSim'10: Proceedings of the 2010 Spring Simulation Multiconference", The Society for Modeling and Simulation International, 2010, p. 12–19.

[48] M. TCHIBOUKDJIAN, N. GAST, D. TRYSTRAM, J.-L. ROCH, J. BERNARD. *A Tighter Analysis of Work Stealing*, in "The 21st International Symposium on Algorithms and Computation (ISAAC)", 2010, http://moais.imag.fr/membres/marc.tchiboukdjian/pub/isaac10.pdf.

[49] S. YI, D. KONDO, D. P. ANDERSON. *Toward Real-Time, Many-Task Applications on Large Distributed Systems*, in "Euro-Par (1)", 2010, p. 355-366.

[50] S. YI, D. KONDO, D. P. ANDERSON. *Toward Real-Time, Many-Task Applications on Large Distributed Systems*, in "European Conference on Parallel and Distributed Computing Euro-Par", 2010, p. 355-366.

[51] S. YI, D. KONDO, A. ANDRZEJAK. *Reducing Costs of Spot Instances via Checkpointing in the Amazon Elastic Compute Cloud*, in "3rd Conference on Cloud Computing, IEEE CLOUD", July 2010.

[52] S. YI, D. KONDO, B. KIM, G. PARK, Y. CHO. *Using replication and checkpointing for reliable task management in computational Grids*, in "HPCS", 2010, p. 125-131.

[53] S. YI, D. KONDO, B. KIM, G. PARK, Y. CHO. *Using replication and checkpointing for reliable task management in computational Grids*, in "IEEE International Conference on High Performance Computing & Simulatio (HPCS)", 2010, p. 125-131.

[54] S. YI, D. KONDO. *How Checkpointing Can Reduce Costs of Using Clouds*, in "EU-Korea Conference on Science and Technology (EKC)", July 2010.

[55] N. YIGITBASI, M. GALLET, D. KONDO, A. IOSUP, D. EPEMA. *Analysis and Modeling of Time-Correlated Failure in Large-Scale Distributed Systems*, in "IEEE/ACM International Conference on Grid Computing (GRID)", 2010, p. 355-366.

Workshops without Proceedings

[56] L. PILLA, C. POUSA RIBEIRO, D. CORDEIRO, J.-F. MÉHAUT. *Charm++ on NUMA Platforms: the impact of SMP Optimizations and a NUMA-aware Load Balancing*, in "The fourth workshop of the INRIA-Illinois Joint Laboratory on Petascale Computing", Urbana, USA, November 2010, https://wiki.ncsa.illinois.edu/download/attachments/17630761/INRIA-UIUC-WS4-llpilla.pdf?version=1&modificationDate=1290534501000.

Research Reports

[57] M. CASTRO, K. GEORGIEV, V. MARANGONZOVA-MARTIN, J.-F. MÉHAUT, L. G. FERNANDES, M. SANTANA. *Analyzing Software Transactional Memory Applications by Tracing Transactions*, INRIA, 2010, n^{o} 7334.

[58] B. GAUJAL, G. GORGO, J.-M. VINCENT. *Perfect Sampling of Phase-Type Servers using Bounding Envelopes*, INRIA, November 2010, n^{o} RR-7460, http://hal.inria.fr/inria-00540967/PDF/RR-7460.pdf.

[59] F. PIN, A. BUSIC, B. GAUJAL. *Perfect Sampling of Markov Chains with Piecewise Homogeneous Events*, arXiv, 2010, n^{o} 163442.

[60] C. POUSA RIBEIRO, A. CARISSIMI, J.-F. MÉHAUT. *Memory Access Characterization of OpenMP Workloads on a Multi-core NUMA Machine*, INRIA, 07 2010, n^{o} RR-7330.