IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble 1)

Activity Report 2011

Project-Team MESCAL

Middleware Efficiently SCALable

IN COLLABORATION WITH: Laboratoire d'Informatique de Grenoble (LIG)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Distributed and High Performance Computing

Table of contents

Project-Team MESCAL

Keywords: High Performance Computing, Game Theory, Grid'5000, Scheduling, Stochastic Modeling

MESCAL is a common project-team also supported by CNRS, INPG, UJF, member of LIG laboratory (UMR 5217).

1. Members

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]
Panayotis Mertikopoulos [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, CEA Detachment, HdR]
Florence Perronnin [Associate Professor, UJF]
Olivier Richard [Associate Professor, UJF]
Jean-Marc Vincent [Associate Professor, UJF]

Technical Staff

Romain Cavagna [2010-, Engineer Assistant, Inria]
Augustin Degomme [2009-2011, Engineer Assistant, Inria]
Yann Genevois [2011-, Inria]
Philippe Le Brouster [2011, Engineer Assistant, UJF]
Pere Manils [2010-2011, Engineer Assistant, Inria]
Pierre Navarro [2009-2011, Engineer Assistant, Inria]
Pierre Neyron [Research Engineer, CNRS]

PhD Students

Marcio Bastos Castro [2009-, Inria]
Rodrigue Chakode-Noumowe [2008-, Minalogic CILOE scholarship, Inria]
Pierre Coucheney [2008-2011, Inria-Alcatel Lucent scholarship]
Charbel El Kaed [2008-, CIFRE France Télécom R&D scholarship]
Joseph Emeras [2010-, CNRS BDI scholarship]
Kiril Georgiev [2009-, CIFRE STMicroelectronics scholarship]
Gaël Gorgo [2010-, CIFRE Bull scholarship]
Ahmed Harbaoui [2006-2011, CIFRE France Télécom R&D scholarship]
Patricia Lopez Cueva [2010-, CIFRE STmicroelectronics scholarship]
Matthieu Ospici [2008-, CIFRE Bull scholarship]
Kevin Pouget [2010-, CIFRE STMicroelectronics scholarship]
Carlos Prada Rojas [2007-2011, CIFRE STMicroelectronics scholarship]
Pedro Antonio Velho [2006-2011, Brazilian CAPES scholarship]
Christiane Vilaca Pousa Ribeiro [2008-2011, Brazilian CAPES scholarship]
Kelly Rosa Braghetto [2007-2011, Brazilian CAPES scholarship]
Blaise Yenké [2005-2011, Ngaundere University scholarship]

Post-Doctoral Fellows

Eric Heien [2010-2011, ANR Clouds@home, Inria]
Sascha Hunold [2011, Université de Reims]
Lucas Mello Schnorr [2009-, ANR Simgrid, CNRS]

Administrative Assistant
Annie Simon [Assistant (SAER), Inria]

Others
Laurent Bobelin [2010-, ATER, UJF]
Francois Broquedis [2010-2011, ATER, Grenoble INP]

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 activities and objectives were evaluated by Inria in 2008. The MESCAL project-team received positive evaluations and useful feedback. The project-team was extended for another 4 years by the Inria evaluation commission.

2.2. Objectives

The recent evolutions in network and 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 the users.

This race towards *large scale* questions many assumptions underlying parallel and distributed algorithms as well as 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.

One MESCAL's target application is intensive scientific computations (with a recent focus on nano-simulations). Initially executed on large dedicated clusters (CRAY, IBM, COMPAQ), they have been recently deployed on collections of many-core architectures. MESCAL's target infrastructures are aggregations of commodity components and/or commodity clusters at metropolitan, national or international scale such as grids obtained through sharing of available resources inside autonomous computing services, lightweight grids (such as the local CIMENT Grid), clusters of intranet resources (Condor) or aggregation of Internet resources (SETI@home, XtremWeb) as well as clouds (Amazon, Google clouds).

Another application domain concerns wireless networks. We are designing algorithms and middleware for SON (Self Organizing Networks) with implementations in wireless devices and base stations.

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

2.3. Highlights

- Brigitte Plateau was nominated “Chevalier de la légion d’honneur” for her remarkable scientific contributions and her dedication to the influence of Grenoble in the scientific community.
- Derrick Kondo was the recipient of a Google award in 2011 for his work on the prediction of idleness in data-centers.
- Bruno Gaujal, Gaël Gorgo and Jean-Marc Vincent received the best paper award at the ASMTA conference (see Section 6.1 for a detailed account of their contribution).
- The software RTaW-Pegase has received the "Best Tool Demo Award" at the workshop "RTSS@work" at the RTSS conference. This tool is being developped by RTaW, a Start-up company of Inria Lorraine with consulting contributions by Bruno Gaujal.

BEST PAPER AWARD :
[37] **18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA’11)**. B. GAUJAL, G. GORGO, J.-M. VINCENT.

3. Scientific Foundations

3.1. Large System Modeling and Analysis

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

3.1.1. 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*.

3.1.1.1. Flow Simulations

To make simulations of large systems efficient and trustful, we have used flow simulations (where streams of packets are abstracted into flows). SIMGRID is a simulation platform that not only enable one to get repeatable results but also make it possible to explore wide ranges of platform and application scenarios.

3.1.1.2. Perfect Simulation

Using a constructive representation of a Markovian queuing network based on events (often called GSMPs), we have designed a perfect simulation algorithms computing samples distributed according to the stationary distribution of the Markov process with no bias. The tools based on our algorithms (ψ) can sample 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.

3.1.2. 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 behavior. One such tools is mean field analysis and fluid limits, that can be used at a modeling and simulation level. Proving that large discrete dynamic systems can be approximated by continuous dynamics uses the theory of stochastic approximation pioneered by Michel Benaïm or population dynamics introduced by Thomas Kurtz and others. We have extended the stochastic approximation approach to take into account discontinuities in the dynamics as well as to tackle optimization issues.

Recent applications include call centers and peer to peer systems. where the mean field approach 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 and to model central/local controllers as well as knitting systems.

3.1.3. 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.

The worst case analysis of networks can also be done using the (max,plus) machinery, called *network calculus* or *real time calculus* in this context.

3.1.4. Game Theory

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 result 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 natural to seek in fully distributed systems 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.

3.2.1. Instrumentation, analysis and prediction tools

To understand complex distributed systems, one has to provide reliable measurements together with accurate models before applying this understanding to improve system design.

Our approach for instrumentation os distributed systems (embedded systems as well as multi-core machines or distributed systems) relies on quality of service criteria. In particular, we focus on non-obtrusiveness and experimental reproducibility.

Our approach for analysis is to use statistical methods with experimental data of real systems to understand their normal or abnormal behavior. With that approach we are able to predict availability of very large systems (with more than 100,000 nodes), to design cost-aware resource management (based on mathematical modeling and performance evaluation of target architectures), and to propose several scheduling policies tailored for unreliable and shared resources.

3.2.2. 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.

3.2.3. 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.

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

3.2.4. Simple and scalable batch scheduler for clusters and grids

Most known batch schedulers (PBS, LSF, Condor, ...) are of old-fashioned conception, built in a monolithic way, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150,000 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 did not 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 communication pattern. 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.

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).

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. On-demand Geographical Maps

Participant: Jean-Marc Vincent.

This joint work involves the UMR 8504 Géographie-Cité, LIG, 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 LIG, 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.

4.2. 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).

4.3. Seismic simulations

Participant: Jean-François Méhaut.

Numerical modeling of seismic wave propagation in complex three-dimensional media is an important research topic in seismology. Several approaches should be studied, and their suitability with respect to the specific constraints of NUMA architectures should be evaluated. These modeling approaches rely on modern numerical schemes such as spectral elements, high-order finite differences or finite elements applied to realistic 3D models. The NUMASIS project focused 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 have been 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 has been carried out using some of the available data from these regions.

4.4. Electromagnetic Fields simulations

Participant: Yves Denneulin.

We study scaling properties in electromagnetism simulation applications and grids. We have shown how to deploy computational electromagnetic applications on grid computing architectures. We have also designed a parallelization of the scale changing technique in Grid computing environment for the electromagnetic simulation of multi-scale structures.

4.5. Embedded Systems

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

Embedded computing is shifting to multi/many-core designs to boost performance due to unacceptable power consumption and operating temperature increase of fast single-core CPU's. Hence, we help embedded system designers to face several big challenges, namely: instrumentation of heterogeneous platforms, bottleneck identification and support for a variety of concurrent applications.

4.6. Wireless Networks

Participants: Bruno Gaujal, Corinne Touati, Panayotis Mertikopoulos.

MESCAL is involved in the common laboratory between Inria and Alcatel-Lucent. Bruno Gaujal is leading the Selfnets research action. This action was started in 2008 and was renewed for four more years (from 2012 to 2016). In our collaboration with Alcatel we use game theory techniques as well as evolutionary algorithms to compute optimal configurations in wireless networks (typically 3G or LTE networks) in a distributed manner. One patent has been taken in 2010 and a second one has been filled in 2011.

5. Software

5.1. Tools for cluster management and software development

Participant: Olivier Richard [correspondant].

The KA-Tools is a software suite developped by MESCAL for exploitation of clusters and grids. It uses a parallelization technique based on spanning trees with a recursive starting of programs on nodes. Industrial collaborations were carried out with Mandrake, BULL, HP and Microsoft.

5.1.1. KA-Deploy

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

5.1.2. Taktuk

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.

5.2. OAR: Batch scheduler for clusters and grids

Participant: Olivier Richard [correspondant].

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.

See also the web page http://oar.imag.fr.

5.3. FTA: Failure Trace Archive

Participant: Derrick Kondo [correspondant].

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 on-line 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. The FTA was 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.

5.4. SimGrid: simulation of distributed applications

Participants: Arnaud Legrand [correspondant], Lucas Schnorr, Pierre Navarro, Sascha Hunold, Laurent Bobelin.

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 one new major version (3.6) of SimGrid (June 2011) and two minor versions (June and October 2011). These versions include our current work on visualization, analysis of large scale distributed systems, and extremely scalable simulation. See also the web page http://simgrid.gforge.inria.fr/.

5.5. TRIVA: interactive trace visualization

Participants: Lucas Schnorr [correspondant], Arnaud Legrand.

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 for 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.

See also the web page http://triva.gforge.inria.fr/.

5.6. ψ and ψ^2: perfect simulation of Markov Chain stationary distributions

Participant: Jean-Marc Vincent [correspondant].

ψ and ψ^2 are two software tools implementing perfect simulation of Markov Chain stationary distributions using *coupling from the past*. 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/.

6. New Results

6.1. Perfect simulation

We have proposed a new approach for sampling the stationary distribution of general Markov chains that only needs to consider two trajectories. We show that this new approach is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed [26]. We further show that most Markovian queuing networks have this property and we propose efficient algorithms for some of them, in particular when the rates of events range over several orders of magnitude [45]. We also provided a novel approach for efficient sampling of queues with phase type servers [37] (this paper has received the best paper award at ASMTA 2011) and Markov chains with infinite state spaces (but with a known bounding process). Perfect sampling has been used for model checking of probabilistic models in [14].

6.2. Economic models for clouds

Recently introduced spot instances in the Amazon Elastic Compute Cloud (EC2) offer low resource costs in exchange for reduced reliability; these instances can be revoked abruptly due to price and demand fluctuations. Mechanisms and tools that deal with the cost-reliability trade-offs under this scheme are of great value for users seeking to lessen their costs while maintaining high reliability. We study how mechanisms, namely, checkpointing and migration, can be used to minimize the cost and volatility of resource provisioning. Based on the real price history of EC2 spot instances, we compare several adaptive checkpointing schemes in terms of monetary costs and improvement of job completion times. We evaluate schemes that apply predictive methods for spot prices. Furthermore, we also study how work migration can improve task completion in the midst of failures while maintaining low monetary costs. Trace-based simulations show that our schemes can reduce significantly both monetary costs and task completion times of computation on spot instance [25].

6.3. Game theory and networks

We studied the traffic routing problem in networks whose users try to minimize their latencies by employing a distributed learning rule inspired by the replicator dynamics of evolutionary game theory. The stable states of these dynamics coincide with the network's (Wardrop) equilibrium points. Despite this abundance of stable states, we find that (almost) every solution trajectory converges to an equilibrium point at an exponential rate. When network latencies fluctuate unpredictably we show that the time-average of the traffic flows of sufficiently patient users is still concentrated in a neighborhood of evolutionarily stable equilibria and we estimate the corresponding stationary distribution and convergence times [42].

We also analyzed the distributed power allocation problem in parallel multiple access channels (MAC) by studying an associated non-cooperative game which admits an exact potential function. We show that the parallel MAC game admits a unique equilibrium almost surely. Furthermore, if the network's users employ a distributed learning scheme based on the replicator dynamics, we show that they converge to equilibrium from almost any initial condition, even though users only have local information at their disposal [41].

Using a large deviations approach we calculate the probability distribution of the mutual information of MIMO channels in the limit of large antenna numbers. We calculate the full distribution, including its tails which strongly deviate from the Gaussian behavior near the mean. This calculation provides us with a tool to obtain outage probabilities analytically at any point in the parameter space, as long as the number of antennas is not too small [20].

6.4. Mean field analysis for networks

We have studied the deterministic limits of Markov processes made of several interacting objects. While most classical results assume that the limiting dynamics has Lipschitz properties, we show 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 way 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 [65].

6.5. Idleness and failure prediction in large infrastructures

We have proposed a method to discover statistical models of availability in large distributed systems and applied it to run an enlightening study of SETI@home [19]. This was also used to make long-term availability predictions for groups of desktop grid resources [21]. We have used statistically based models of heterogeneous failures in parallel systems and assessed their tolerance [39]. A similar approach was used to design correlated resource models of Internet end hosts [38], [17].

6.6. Scheduling and Game Theory

A stochastic model of failures has been used to optimize the scheduling of checkpoints on desktop grids [28].

We have also shown that non-cooperative scheduling can be considered harmful in collaborative volunteer computing environments [33]

Optimal scheduling and route selection have been investigated using a novel approach based on Lagrangian optimization. This result is inspired from flow control in multi-path networks and was used for multiple mag-of-tasks application scheduling on grids [61].

In the similar context of broker-based networks of non-observable parallel queues, we provide lower bounds on the minimum response time. We introduce the “Price of Forgetting” (PoF), the ratio between the minimum response times achieved by a probabilistic broker and a broker with memory, that is shown to be unbounded or arbitrarily close to one depending on the coefficient of variation of the service time distributions. We also put our results in the context of game theory revisiting the “Price of Anarchy” (PoA) of parallel queues: It can be decomposed into the product of the PoA achieved by a probabilistic broker (already well understood) and the PoF [10].

6.7. Validity study of flow-based network models.

Researchers in the area of distributed computing conduct many of their experiments in simulation. While packet-level simulation is often used to study network protocols, it can be too costly to simulate network communications for large-scale systems and applications. The alternative chosen in SimGrid and a few other simulation frameworks is to simulate the network based on less costly flow-level models. Surprisingly, in the literature, validation of these flow-level models is at best a mere verification for a few simple cases. Consequently, although distributed computing simulators are widely used, their ability to produce scientifically meaningful results is in doubt. In [9], [70] we focus on the validation of state-of-the-art flow-level network models of TCP communication, via comparison to packet-level simulation. While it is straightforward to show cases in which previously proposed models lead to good results, instead we systematically seek cases that lead to invalid results. Careful analysis of these cases reveal fundamental flaws and also suggest improvements. One contribution of this work is that these improvements lead to a new model that, while far from being perfect, improves upon all previously proposed models. A more important contribution, perhaps, is provided by the pitfalls and unexpected behaviors encountered in this work, leading to a number of enlightening lessons. In particular, this work shows that model validation cannot be achieved solely by exhibiting (possibly many) ”good cases.” Confidence in the quality of a model can only be strengthened through an invalidation approach that attempts to prove the model wrong.

6.8. Vizualisation

We have proposed a methodology for detecting resource usage anomalies in large scale distributed systems. The methodology relies on four functionalities: characterized trace collection, multi-scale data aggregation, specifically tailored user interaction techniques, and visualization techniques. We have shown the efficiency of this approach through the analysis of simulations of the volunteer computing Berkeley Open Infrastructure for Network Computing architecture (BOINC). Three scenarios have been analyzed in [48], [23]: analysis of the resource sharing mechanism, resource usage considering response time instead of throughput, and the evaluation of input file size on Berkeley Open Infrastructure for Network Computing architecture. The results show that our methodology enables to easily identify resource usage anomalies, such as unfair resource sharing, contention, moving network bottlenecks, and harmful short-term resource sharing. Triva, the resulting software, has been demonstrated at the SuperComputing conference.

We also have investigated how to use trace-based visualization to understand applications I/O performance [49] and how to visually compare two traces [70] and highlight differences.

6.9. Experimental methodology

In the scientific experimentation process, an experiment result needs to be analyzed and compared with several others, potentially obtained in different conditions. Several tools are dedicated to the control of the experiment input parameters and the experiment replay. In parallel, concurrent and distributed systems, experiment conditions are not only restricted to the input parameters, but also to the software environment in which the experiment was carried out. It is therefore essential to be able to reconstruct this type of environment. This can quickly become complex for experimenters, particularly on research platforms dedicated to scientific experimentation, where both hardware and software are in constant rapid evolution. We study the concept of the reconstructability of software environments and propose a tool for dealing with this problem in [64].

We have also started investigating the systematic use of Design of Experiments to computer studies (see [61]). Nonetheless such approach provides results that are much more trustworthy than what is generally done in the parallel and distributed computing community but it also enables to shorten the experiments cycle and to use less computing resources.

6.10. Multi-core platforms

We have used memory access traces to map threads on hierarchical multi-core platforms [13]. We have also used software transactional memory to analyze and trace applications running on multi-core architectures [30].

An approach based on machine learning was used to map threads on transactional memory applications in [31].

The impact of CPU and memory affinity on multi-core platforms was investigated in [46] using numerical scientific multi-threaded applications as a typical case study. This resulted in improvement of the performance of parallel systems using a NUMA-aware load balancer [68].

We have also carried a performance evaluation of WiNoCs for parallel workloads based on collective communications [43] as well as for Infiniband networks [40].

6.11. High performance computing

We have developed a runtime system, named SGPU 2, that enable large applications to run on clusters of hybrid nodes [44].

BigDFT is a parallel simulator of the matter at the nano scale. It uses Daubechies Wavelets for High Performance Electronic Structure Calculations [16]. This tool is shown to make efficient use of massive parallel hybrid architectures [57].

6.12. Input-Output

Atmospheric models usually demand high processing power and generate large amounts of data. As the degree of parallelism grows, the I/O operations may become the major impacting factor of their performance. In [27], we evaluate the Ocean-Land-Atmosphere Model (OLAM) on the PVFS file system in order to point the I/O characteristics of the application. We show that storing the files on PVFS has lower performance than using the local disks of the cluster nodes due to file creation and network concurrency. Additionally, we study the performance of a new version of OLAM that used MPI associated with OpenMP and show that the combined strategy presents I/O times 20 times shorter than the original MPI-only version and 9 times shorter on total execution time. Finally, a survey on I/O Characterization of several applications is given in [51].

7. Contracts and Grants with Industry

7.1. Contracts with Industry

7.1.1. 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 http://www.RealTimeAtWork.com.

7.1.2. CILOE with Bull, Compagnie des Signaux, TIMA, CEA-LETI, LIG, Edxact, Infiniscale, Probayes, SCelectronique

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 start-ups) 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 association with 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.

7.1.3. ADR Selfnets 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 analyze fully distributed algorithms based on evolutionary games that exploit the benefits of vertical handover by finding fair and efficient user-network association schemes.

A patent on a simplified version of our algorithm has been taken by the common lab in 2010.

In 2011, a new patent has been filled on new algorithms that are robust to noise on measurements as well as to several revision scenarios (mobiles change their connections to base stations simultaneously or asynchronously).

7.2. Grants with Industry

7.2.1. CIFRE contracts with Bull

- 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.

7.2.2. CIFRE contracts with Orange Labs

- Ahmed Harbaoui did his PhD thesis with France Télécom R&D company. He worked in load injection and performance evaluation issues in networks. He defended his thesis in October 2011.
- Charbel El Kaed is doing his PhD thesis in France Télécom on the usage of communication devices.

7.2.3. CIFRE contracts with STMicroelectronics

- Carlos Prada Rojas did his PhD thesis with STMicroelectronics. He started in September 2007 and defended in June 2011. The objective of his thesis was 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.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. CIMENT

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

8.1.2. High Performance Computing Center

- 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 Digitalis cluster is also meant to replace the Grimage platform in which the MOAIS project-team is very involved.

- The Bull Machine

In the context of our collaboration with Bull the MESCAL project-team exploits 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.

- 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.

8.2. National Initiatives

8.2.1. "Action d'envergure"

- *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 explores 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. ARC Inria

- *Meneur 2011-2013*:

Partners: EPI Dionysos, EPI Maestro, EPI MESCAL, EPI Comore, GET/Telecom Bretagne, FTW, Vienna (Forschungszentrum Telekommunikation Wien), Columbia University, USA, Pennsylvania State University, USA, Alcatel-Lucent Bell Labs France, Orange Labs.

The goal of this project is to study the interest of network neutrality, a topic that has recently gained a lot of attention. The project aims at elaborating mathematical models that will be analyzed to investigate its impact on users, on social welfare and on providers' investment incentives, among others, and eventually propose how (and if) network neutrality should be implemented. It brings together experts from different scientific fields, telecommunications, applied mathematics, economics, mixing academy and industry, to discuss those issues. It is a first step towards the elaboration of a European project.

8.2.3. ADT Inria (2)

- *SimGrid for Human Beings, 2009-2011*:

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.

- *Aladdin-G5K, 2008-2011*

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

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

ALADDIN built on Grid'5000's experience to provide an infrastructure enabling computer scientists to conduct experiments on large scale computing and produced scientific results that can be reproduced by others.

MESCAL members are particularly involved in efficient large scale system utilization, providing confidence to the user about the infrastructure and modeling of large scale systems and validation of their simulators.

8.2.4. 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.

8.2.5. ANR Jeunes Chercheurs et Jeunes Chercheuses (2)

- *DOCCA, 2007-2011*

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:

1. 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;
2. 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 to 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(Design and Optimization of Collaborative Computing Architectures) project are the following:

First, we want 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.

We also want to design a P2P architecture based on the algorithms designed in the second step, and to create a novel P2P collaborative computing system.

- *Clouds@home, 2009-2013*

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.

8.2.6. ANR COSI

- PROHMPT, 2009-2011

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.

8.2.7. ANR ARPEGE

- PEGASE, 2009-2011

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.

8.2.8. ANR SEGI (2)

- *USS Simgrid, 2009-2011*

Partners: Inria Nancy, Inria Sophia, Inria Bordeaux, University of Reims, IN2P3, University of Hawaii at Manoa

The goal of the USS-SimGrid project is to enable scalable and accurate simulations by means of the SimGrid simulation toolkit. This toolkit is widely used for simulation of Grid 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 are three main thrusts in this project. First, we improve the models used in SimGrid, increasing their scalability and easing their instantiation. Second, we develop tools that ease the analysis of detailed and large simulation results, and aid the management of simulation deployments. Third, we improve the scalability of simulations using parallelization and optimization methods. A mid-term report summarizing our findings has been published in [59].

- *SPADES, 2009-2012*

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.

8.3. European Initiatives

8.3.1. FP7 EDGI (European Desktop Grid Initiative)

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

Years: 2010-2012

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.

8.3.2. FP7 Mont-Blanc project: European scalable and power efficient HPC platform based on low-power embedded technology

FP7 Programme: ICT-2011.9.13 Exa-scale computing, software and simulation

Mont-Blanc Partners: BSC (Barcelone), Bull, ARM (UK), Julich (Germany), Genci, CINECA (Italy), CNRS (LIRMM, LIG)

Duration: 3 Years from 1/10/2011

There is a continued need for higher compute performance: scientific grand challenges, engineering, geophysics, bioinformatics, etc. However, energy is increasingly becoming one of the most expensive resources and the dominant cost item for running a large supercomputing facility. In fact, the total energy cost of a few years of operation can almost equal the cost of the hardware infrastructure. Energy efficiency is already a primary concern for the design of any computer system and it is unanimously recognized that Exascale systems will be strongly constrained by power.

The analysis of the performance of HPC systems since 1993 shows exponential improvements at the rate of one order of magnitude every 3 years: One petaflops was achieved in 2008, one exaflops is expected in 2020. Based on a 20 MW power budget, this requires an efficiency of 50 GFLOPS/Watt. However, the current leader in energy efficiency achieves only 1.7n GFLOPS/Watt. Thus, a 30x improvement is required.

In this project, the partners believe that HPC systems developed from today's energy-efficient solutions used in embedded and mobile devices are the most likely to succeed. As of today, the CPUs of these devices are mostly designed by ARM. However, ARM processors have not been designed for HPC, and ARM chips have never used in HPC systems before, leading to a number of significant challenges.

8.3.3. HPC-GA project: High Performance Computing for Geophysics Applications

FP7 programme: Marie Curie Actions, International Research Staff Exchange Scheme (IRSES)

Partners: Inria (Grenoble, Bordeaux, Pau), BCAM (Bilbao), UFRGS (Brazil), UNAM (Mexico), BRGM (France), UJF (France)

Duration: 3 years from 1/1/2012

PI: Inria (Grenoble and Bordeaux)

Simulating large-scale geophysics phenomenon represents, more than ever, a major concern for our society. Recent seismic activity worldwide has shown how crucial it is to enhance our understanding of the impact of earthquakes. Numerical modeling of seismic 3D waves obviously requires highly specific research efforts in geophysics and applied mathematics, leveraging a mix of various schemes such as spectral elements, high-order finite differences or finite elements. But designing and porting geophysics applications on top of nowadays supercomputers also requires a strong expertise in parallel programming and the use of appropriate runtime systems able to efficiently deal with heterogeneous architectures featuring many-core nodes typically equipped with GPU accelerators. The HPC-GA project aims at evaluating the functionalities provided by current runtime systems in order to point out their limitations. It also aims at designing new methods and mechanisms for an efficient scheduling of processes/threads and a clever data distribution on such platforms.

8.3.4. Collaborations in European Programs, except FP7

- 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 LIG, 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.

- European Exascale Software Initiative (EESI)

The objective of this Support Action, co-funded by the European Commission is to build a European vision and road-map 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.

8.4. International Initiatives

8.4.1. Inria Associate Teams

8.4.1.1. Cloud Computing at Home

Title: Cloud Computing over Internet Volunteer Resources

Inria principal investigator: Derrick Kondo

International Partner:

Institution: University of California Berkeley (United States)

Laboratory: Space Sciences Laboratory

Researcher: David P.

Duration: 2009 - 2011

See also: http://abenaki.imag.fr/cloudcomputing/pmwiki.php

Recently, a new vision of cloud computing has emerged where the complexity of an IT infrastructure is completely hidden from its users. At the same time, cloud computing platforms provide massive scalability, 99.999% reliability, and speedy performance at relatively low costs for complex applications and services. In this proposed collaboration, we investigate the use of cloud computing for large-scale and demanding applications and services over the most unreliable but also most powerful resources in the world, namely volunteered resources over the Internet. The motivation is the immense collective power of volunteer resources (evident by FOLDING@home's 3.9 PetaFLOPS

system), and the relatively low cost of using such resources. We will address these challenges drawing on the experience of the BOINC team 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).

8.4.1.2. DIODEA

Title: France/Brazil Associated research team on Parallel Computing

Inria principal investigator: Bruno Raffin

International Partner:

Institution: Universidade Federal do Rio Grande do Sul (Brazil)

Laboratory: UFRGS

Researcher: Philippe Olivier Alexandre Navaux

Duration: 2009 - 2011

See also: http://diodea.imag.fr/

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.

8.4.2. Inria International Partners

- MESCAL has strong connections with both UFRGS (Porto Alegre, Brazil) and USP (Sao Paulo, Brazil). This year, Jean-François Méhaut visited both laboratories in July. The creation of the LICIA common laboratory (see next section) will make this collaboration even tighter.
- MESCAL has strong bounds with the University of Illinois Urbana Champaign, within the (Joint Laboratory on Petascale Computing (see next section).
- MESCAL also has long lasting collaborations with University of California in Berkeley and a new one with Google. Derrick Kondo is being visiting them in October and November.
- Vania Martin has been visiting the Pontifícia Universidade Catolica de Minas Gerais (Belo Horizonte, Brazil).

8.4.3. Participation In International Programs

8.4.3.1. Africa

- SARIMA and IDASCO / LIRIMA (Cameroon)

MESCAL takes part in the SARIMA (Soutien aux Activités de Recherche Informatique et Mathématiques en Afrique http://www-direction.inria.fr/international/AFRIQUE/sarima.html) 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.

8.4.3.2. North America

- Google Derick Kondo has received a Google Research Award in 2011 for his proposal on predicting idleness in data centers. The technical goal of the proposed work is to give probabilistic guarantees on when data centers are idle. The implication of such predictions is improved data center utilization, while reducing and amortizing monetary costs. The general goal of this award is to facilitate collaboration between Google Inc. and academic researchers. Google Inc. provides the award as an unrestricted gift without constraints on intellectual property.
- Amazon (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.
- JLPC (Joint Laboratory on Petascale Computing) (with University of University of Illinois Urbana Champaign. Several members of MESCAL are partners of this laboratory, and have paid several visits to Urbana-Champaign. The latest workshop of the laboratory has been organized by Jean-François Méhaut in Grenoble.

8.4.3.3. South America

- LICIA. The CNRS, Inria, the Universities of Grenoble, Grenoble INP and Universidade Federal do Rio Grande do Sul have created the LICIA (*laboratoire International de Calcul intensif et d'Informatique Ambiante*). On the French side, the laboratory is co-directed by Yves Denneulin and Jean-Marc Vincent.

  The grand opening workshop has taken place in Porto Alegre, Brazil from Oct. 31st to Nov. 1st.

  The main themes are artificial intelligence, high performance computing, information representation, interfaces and visualization as well as distributed systems.

  More information can be found on http://www.inf.ufrgs.br/licia/.

9. Dissemination

9.1. Animation of the scientific community

- Brigitte Plateau is the director of Grenoble-INP ENSIMAG.
- Yves Denneulin has been appointed deputy director of the LIG.
- Corinne Touati is the Grenoble INP correspondent for international relations with Japan.
- Yves Denneulin and Jean-Marc Vincent are co-directors of the LICIA (Franco-Brazilian Laboratory).
- Jean-François Méhaut has been appointed as expert for the action HPC PME (Genci, Inria, Oseo) as well as for the ANR "COSINUS" program and the "Blanc" program.
- Jean-François Méhaut was a member of two selection committees in Joseph Fourier University and Grenoble INP.
- Arnaud Legrand was a member of the selection committee in University of Bordeaux 1 (LABRI).
- Bruno Gaujal was a member of the selection committee in ENS de Lyon.
- Jean-François Méhaut is a member of the HDR commission of Joseph Fourier University.

9.1.1. Invited Talks

- The 2011 International Conference on Mathematical Aspects of Game Theory and Applications (MAGTA 2011) (Panayotis Mertikopoulos)

- Network Games, Control and Optimization Conference, Paris (NetgCoop 2011) (Bruno Gaujal).
- Twenty years of LaBRI (Bordeaux). Jean-François Méhaut gave an invited talk entitled “twenty years of multithreading”.

9.1.2. Journal, Conference and Workshop Organization

- Chair and organizer of the JLPC (Joint Laboratory on Petascale Computing) workshop (J-F. Méhaut)
- Co-chair of the Workshop on Algorithmic Game Theory: Dynamics and Convergence in Distributed Systems (AlgoGT 2011) (C. Touati). Local organization committee led by Annie Simon.
- Co-chair of the 1st International Workshop on Optimization for Green Wireless Communication Networks (GreenNet 2011) (C. Touati)
- Web chair of the International Conference on NETwork Games, COntrol and OPtimization (NetGCoop 2011) (C. Touati)
- Local chair for Europar 2011 (Prediction and Performance Evaluation Topic) (D. Kondo).
- Co-general chair of PCGrid 2011 (D. Kondo).

9.1.3. Program Committees

- D. Kondo was a program committee member of CCGrid 2011, ServP2P 2011, and BADS 2011.
- 7th International ICQT Workshop on Advanced Internet Charging and QoS Technology (ICQT’11) (C. Touati)
- Valuetools 2011 (B. Gaujal and J.-M. Vincent).
- A. Legrand was program committee member of CCgrid 2011 and IPDPS 2011 and 2012.

9.1.4. Thesis Defense

- Pierre Coucheney *Optimisation des réseaux sans fil: une approche par la théorie des jeux* (August 31)
- Ahmed Harbaoui *Vers une modélisation et un dimensionnement automatique des applications réparties* (October 21)
- Carlos Prada Rojas *Une approche à base de composants logiciels pour l’observation de systèmes embarqués* (June 24)
- Kelly Rosa Braghetto *Modeling Techniques for Business Process Performance Analysis* (September 21)
- Pedro Velho *Accurate and Fast Simulations of Large-Scale Distributed Computing Systems* (July 4)
- Christiane Vilaca Pousa Ribeiro *Contributions on Memory Affinity Management for Hierarchical Shared Memory Multi-core Platforms* (June 29)
- Blaise Yenke *Ordonnancement des sauvegardes/reprises d’applications de calcul haute performance dans les environnements dynamiques* (January 7)

9.1.5. Thesis Committees

Researchers of the MESCAL project-team have served on the following "habilitation" thesis (HDR) committees

- Brigitte Plateau served on the HDR thesis committee of J.-M. Menaud (Ecole des Mines de Nantes) and Agnès Front (LIG).

And researchers of the MESCAL project-team have served on the following PhD thesis committees

- Brigitte Plateau served on the thesis committee of Eric Simon (LIG) and presided the jury of Ahmed Harbaoui (Orange Labs).
- Bruno Gaujal served on the thesis committee of Euriell Le Corronc (Angers) as a reviewer.
- Jean-Françoids Méhaut presided the jury of the PhD defense of Benjamin Negrevergne (Grenoble). He was a member of the thesis committee of Cristian Rosa (Nancy) Cédric Augonnet (Bordeaux), Bogdan Cornea (Montbelliard), Khawar Sajjad (Versailles) and Paulin Melatagia (Yaoundé, Cameroon).
- Jean-Marc Vincent was a member of the thesis committee of Bogdan Cornea (Besançon) and Xavier Grehant (Paris).
- Yves Denneulin served on the thesis committee of Guilherme Piêgas Koslovski (Lyon), Nicolas Ferry (Nice) and Hiep-Thuan Do (Orléans) as a reviewer and presided the jury of the thesis of Willy Malveau (Grenoble) and the jury of the HDR of Vivien Quema (Grenoble).

9.1.6. Popular Science

- Several members of MESCAL were involved in “fete de la science” showing everyday practical use of game theory.
- MESCAL team actively promotes science to young and non-scientific audience. This year, Corinne Touati participated to the "stage Maths C2+" and the bi-annual "semaine de l’ingénieur" to promote the use of mathematics to junior high and high school students in Rhône-Alpes.
- Jean-Marc Vincent contributed to the national initiative for introducing computer science to high school professors in mathematics. This material was collected in a book [71].

9.2. Teaching

Several members of mescal are university professors and comply with their recurrent teaching duties. In addition, here are more details on new lectures that were initiated by MESCAL members in 2011.

- Performance evaluation. Two new lectures on system modeling and performance evaluation started in Ensimag in 2011 at M1 and M2 levels.
- Game Theory and Applications, 24 hours, M2R, ENS de Lyon, France.
- A teaching unit has been opened in the University of Yaoundé 1, on distributed systems.
- "Research School" (open to M1, M2 and PhD students) on Game Theory for Networks, ENS de Lyon, France.

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^{o} 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. P. 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.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[4] P. COUCHENEY. *Optimisation des réseaux sans fil: une approche par la théorie des jeux*, Université de Grenoble, June 2011.

[5] A. HARBAOUI. *Vers une modélisation et un dimensionnement automatique des applications réparties*, Université de Grenoble, October 2011.

[6] B. NEGREVERGNE. *A Generic and Parallel Pattern Mining Algorithm for Multi-Core Architectures*, Université de Grenoble, November 2011.

[7] C. POUSA RIBEIRO. *Contributions on Memory Affinity Management for Hierarchical Shared Memory Multi-core Platforms*, Université de Grenoble, June 2011.

[8] C. PRADA-ROJAS. *Une approche à base de composants logiciels pour l'observarion de systèmes embarqués*, Université de Grenoble, June 2011.

[9] P. VELHO. *Accurate and Fast Simulations of Large-Scale Distributed Computing Systems*, Université de Grenoble, July 2011.

Articles in International Peer-Reviewed Journal

[10] J. ANSELMI, B. GAUJAL. *The Price of Forgetting in Parallel and Non-Observable Queues*, in "Performance Evaluation", 2011, to appear.

[11] F. ARAUJO, J. FARINHA, P. DOMINGUES, G. C. SILAGHI, D. KONDO. *A Maximum Independent Set Approach for Collusion Detection in Voting Pools*, in "Journal of Parallel and Distributed Computing", 2011, vol. 71, n^{o} 10.

[12] S. ASEERVATHAM, A. ANTONIADIS, E. GAUSSIER, M. BURLET, Y. DENNEULIN. *A sparse version of the ridge logistic regression for large-scale text categorization*, in "Pattern Recognition Letters", 2011, vol. 32, n^{o} 2, p. 101-106.

[13] E. CRUZ, C. POUSA RIBEIRO, M. ALVES, A. CARISSIMI, P. O. A. NAVAUX, J.-F. MEHAUT. *Using Memory Access Traces to Map Threads on Hierarchical Multi-core Platforms*, in "International Journal on Networking and Computing", 2011, Accepted.

[14] D. ELRABIH, G. GORGO, N. PEKERGIN, J.-M. VINCENT. *Steady state property verification of very large systems*, in "International Journal of Critical Computer-Based Systems", 2011, vol. 2, p. 309–331.

[15] N. GAST, B. GAUJAL, J.-Y. LE BOUDEC. *Mean field for Markov Decision Processes: from Discrete to Continuous Optimization*, in "IEEE Transaction on Automatic Control", 2011, accepted for publication.

[16] L. GENOVESE, B. VIDEAU, M. OSPICI, T. DEUTSCH, S. GODECKER, J.-F. MEHAUT. *Daubechies Wavelets for High Performance Electronice Structure Calculations: the BigDFT Project*, in "Comptes Rendus de l'Académie des Sciences", February 2011, vol. 339, n^{o} 2, p. 149-164, Special Issue on Intensive Computing.

[17] E. HEIEN, D. KONDO, D. P. ANDERSON. *A Correlated Resource Model of Internet End Hosts*, in "IEEE Transactions on Parallel and Distributed Systems", 2011, to appear.

[18] M. D. IHAB SBEITY, B. PLATEAU. *Stochastic Bounds for Microprocessor Systems Availability*, in "IAJIT-International Arab Journal of Information Technology", January 2011, vol. 8, n^{o} 1, http://www.ccis2k.org/iajit/PDF/vol.8,no.1/14.pdf.

[19] B. JAVADI, D. KONDO, J.-M. VINCENT, D. P. ANDERSON. *Discovering Statistical Models of Availability in Large Distributed Systems: An Empirical Study of SETI@home*, in "IEEE Transactions on Parallel and Distributed Systems", 2011, vol. 99, preprint, http://mescal.imag.fr/membres/derrick.kondo/pubs/javadi_tpds10.pdf.

[20] P. KAZAKOPOULOS, P. MERTIKOPOULOS, A. MOUSTAKAS, G. CAIRE. *Living at the edge: a large deviations approach to the outage MIMO capacity*, in "IEEE Transactions on Information Theory", April 2011, vol. 57, n^{o} 4, p. 1984-2007, http://arxiv.org/abs/0907.5024.

[21] D. LAZARO IGLESIAS, D. KONDO, J. MANUEL MARQUES PUIG. *Long-term Availability Prediction for Groups of Desktop Grid Resources*, in "Journal of Parallel and Distributed Systems", 2011, to appear.

[22] C. PAWLOWITSCH, P. MERTIKOPOULOS, N. RITT. *Neutral stability, drift, and the diversification of languages*, in "Journal of Theoretical Biology", July 2011, n^{o} 287, p. 1-12.

[23] L. M. SCHNORR, A. LEGRAND, J.-M. VINCENT. *Detection and analysis of resource usage anomalies in large distributed systems through multi-scale visualization*, in "Concurrency and Computation: Practice and Experience", 2011, http://dx.doi.org/10.1002/cpe.1885.

[24] B. YENKE, J.-F. MEHAUT, M. TCHUENTÉ. *Scheduling of Computing Services on Intranet Networks*, in "IEEE Transactions on Services Computing (TSC)", July-September 2011, vol. 4, n^{o} 3, p. 207-215.

[25] S. YI, A. ANDRZEJAK, D. KONDO. *Monetary Cost-Aware Checkpointing and Migration on Amazon Cloud Spot Instances*, in "IEEE Transactions on Services Computing", 2011, http://mescal.imag.fr/membres/derrick.kondo/pubs/yi_tsc10.pdf.

International Conferences with Proceedings

[26] J. ANSELMI, B. GAUJAL. *On the efficiency of perfect simulation in monotone queueing networks*, in "IFIP Performance: 29th International Symposium on Computer Performance, Modeling, Measurements and Evaluation", Amsterdam, ACM Performance Evaluation Review, October 2011.

[27] F. BOITO, R. KASSICK, L. L. PILLA, N. BARBIERI, C. SCHEPKE, P. O. A. NAVAUX, N. MAILLARD, Y. DENNEULIN, C. OSTHOFF, P. GRUNMANN, P. DIAS, J. PANETTA. *I/O Performance of a Large Atmospheric Model using PVFS*, in "Rencontres francophones du Parallélisme (RenPar'20)", 2011.

[28] M. S. BOUGUERRA, D. KONDO, D. TRYSTAM. *On the Scheduling of Checkpoints on Desktop Grids*, in "11th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid 2011)", May 2011, p. 305-313, http://mescal.imag.fr/membres/derrick.kondo/pubs/slim_ccgrid11.pdf.

[29] K. R. BRAGHETTO, J. E. FERREIRA, J.-M. VINCENT. *Performance Evaluation of Business Processes through a Formal Transformation to SAN*, in "Proceedings of the 8th European Performance Engineering Workshop (EPEW 2011)", Borrowdale, UK, LNCS, Springer, oct 2011, vol. 6977, p. 42 - 56, http://mescal.imag.fr/membres/jean-marc.vincent/papers/epew2011.pdf.

[30] M. CASTRO, K. GEORGIEV, V. MARANGOZOVA-MARTIN, J.-F. MEHAUT, L. GUSTAVO FERNANDES, M. SANTANA. *Analysis and Tracing of Applications Based on Software Transactional Memory on Multicore Architectures*, in "Euromicro International Conference on Parallel, Distributed and Network-Based Computing (PDP)", Ayia Napa, Cyprus, IEEE Computer Society, 2011, p. 199-206.

[31] M. CASTRO, L. GOES, C. POUSA RIBEIRO, M. CINTRA, J.-F. MEHAUT. *A Machine Learning-Based Approach for Thread Mapping on Transactional Memory Applications*, in "18th Annual and International Conference on High Performance Computing (HiPC)", Bengalore, India, December 2011, Accepted.

[32] R. CHAKODÉ, B. YENKE, J.-F. MEHAUT. *Resource Management of Virtual Infrastructure for On-demand SaaS Services*, in "The 1st International Conference on Cloud Computing and Services Science (CLOSER)", Noordwikerhout, Netherlands, May 2011, p. 352-361.

[33] B. DE MOURA DONASSOLO, A. LEGRAND, C. GEYER. *Non-Cooperative Scheduling Considered Harmful in Collaborative Volunteer Computing Environments*, in "Proceedings of the 11th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'11)", IEEE Computer Society Press, May 2011, http://mescal.imag.fr/membres/arnaud.legrand/articles/2011-ccgrid-donassolo.pdf.

[34] C. EL KAED, Y. DENNEULIN, F.-G. OTTOGALLI. *Dynamic Service Adaptation for Plug and Play Device Interoperability*, in "IEEE, 7th International Conference on Network and Service Management (CNSM 2011)", Paris, France, October 2011.

[35] C. EL KAED, Y. DENNEULIN, F.-G. OTTOGALLI. *On the Fly Proxy Generation for Home Devices Interoperability*, in "Proceedings of the 2011 IEEE 12th International Conference on Mobile Data Management", Washington, DC, USA, MDM '11, IEEE Computer Society, 2011, vol. 01, p. 299–302, http://dx.doi.org/10.1109/MDM.2011.47.

[36] C. EL KAED, L. PETIT, M. LOUVEL, A. CHAZALET, Y. DENNEULIN, F.-G. OTTOGALLI. *INSIGHT: Interoperability and Service Management for the Digital Home.*, in "Middleware '11: Proceedings of the ACM/IFIP/USENIX 2011 International Conference on Middleware", New York, NY, USA, Springer-Verlag New York, Inc., 2011.

[37] *Best Paper*
B. GAUJAL, G. GORGO, J.-M. VINCENT. *Perfect Sampling of Phase-Type Servers using Bounding Envelopes*, in "18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA'11)", Venise, LNCS, Springer-Verlag, 2011, Best Paper Award, http://mescal.imag.fr/membres/jean-marc.vincent/papers/asmta2011.pdf.

[38] E. HEIEN, D. KONDO, D. P. ANDERSON. *Correlated Resource Models of Internet End Hosts*, in "The 31st IEEE International Conference on Distributed Computing Systems (ICDCS'11)", June 2011, http://hal.inria.fr/inria-00538932/PDF/model_synth.pdf.

[39] E. HEIEN, D. KONDO, A. GAINARU, D. LAPINE, B. KRAMER, F. CAPPELLO. *Modeling and Tolerating Heterogeneous Failures on Large Parallel Systems*, in "IEEE/ACM Supercomputing Conference (SC)", November 2011.

[40] M. MARTINASSO, J.-F. MEHAUT. *A Contention-Aware Performance Model for the HPC-based networks: A Case Study of Infiniband Network*, in "17th European Conference on Parallel Computing (Europar)", Bordeaux, August 2011, p. 91-102.

[41] P. MERTIKOPOULOS, E. BELMEGA, A. MOUSTAKAS, S. LASAULCE. *Dynamic power allocation games in parallel multiple access channels*, in "ValueTools '11: ACM Proceedings of the 5th International Conference on Performance Evaluation Methodologies and Tools", 2011.

[42] P. MERTIKOPOULOS, A. MOUSTAKAS. *Selfish routing revisited: degeneracy, evolution and stochastic fluctuations*, in "ValueTools '11: ACM Proceedings of the 5th International Conference on Performance Evaluation Methodologies and Tools", 2011.

[43] P. OLIVEIRA, H. COTA DE FREITAS, C. POUSA RIBEIRO, M. CASTRO, V. MARANGOZOVA-MARTIN, J.-F. MEHAUT. *Performance Evaluation of WiNoCs for Parallel Workloads Based on Collective Communications*, in "IADIS International Conference on Applied Computing (AC)", Rio de Janeiro, Brazil, IADIS Press, 2011.

[44] M. OSPICI, D. KOMATITSCH, J.-F. MEHAUT, T. DEUTSCH. *SGPU 2: a runtime system for using of large applications on clusters of hybrid nodes*, in "Second Workshop on Hybrid Multi-core Computing, held in conjunction with HiPC 2011", Bangalore, India, December 2011.

[45] F. PIN, A. BUSIC, B. GAUJAL. *Acceleration of perfect sampling by skipping events*, in "Valuetools", Paris, 2011.

[46] C. POUSA RIBEIRO, M. CASTRO, J.-F. MEHAUT, V. MARANGOZOVA-MARTIN, H. COTA DE FREITAS, C. MARTINS. *Investigating the Impact of CPU and Memory Affinity on Multi-core Platforms: A Case Study of Numerical Scientific Multithreaded Applications*, in "IADIS International Conference on Applied Computing (AC)", Rio de Janeiro, Brazil, IADIS Press, 2011.

[47] C. PRADA, V. MARANGOZOVA-MARTIN, J.-F. MEHAUT, M. SANTANA. *A Generic Component-Based Approach to MPSoC Observation*, in "9th IEEE/IFIP International Conference on Embedded and Ubiquitous Computing (EUC 2011)", Melbourne, Australia, October 2011.

[48] L. M. SCHNORR, A. LEGRAND, J.-M. VINCENT. *Multi-scale analysis of large distributed computing systems*, in "Proceedings of the third international workshop on Large-scale system and application performance", LSAP '11, ACM, 2011, p. 27–34, http://mescal.imag.fr/membres/arnaud.legrand/articles/2011-lsap-schnorr.pdf.

[49] R. VIROTE KASSICK, C. OSTHOFF, P. O. A. NAVAUX, F. ZANON BOITO, C. SCHEPKE, N. MAILLARD, M. DIENER, Y. DENNEULIN. *Trace-based Visualization as a Tool to Understand Applications I/O Performance*, in "proceeding of the SBAC-PAD 2011 - WAMCA 2011 workshop", 2011.

[50] S. YI, E. JEANNOT, D. KONDO, D. P. ANDERSON. *Towards Real-Time Volunteer Distributed Computing*, in "11th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid 2011)", May 2011, http://mescal.imag.fr/membres/derrick.kondo/pubs/yi_ccgrid11.pdf.

[51] F. ZANON BOITO, R. VIROTE KASSICK, P. O. A. NAVAUX, Y. DENNEULIN. *A Survey on Applications' I/O Characterization*, in "proceeding os the IX Workshop de Processamento Paralelo e Distribuído", Porto Alegre, 2011.

National Conferences with Proceeding

[52] R. DAVID, Y. GEORGIOU, S. POULAT, O. RICHARD. *Gestion des ressources de calcul : 4 points de vue*, in "Journées Réseaux", 2011.

[53] J. EMERAS, B. BZEZNIK, Y. GEORGIOU, P. LE BROUSTER, O. RICHARD. *Reconstruction des environnements logiciels des expériences avec Kameleon*, in "Renpar", 2011.

[54] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Observation macroscopique et émergence dans les SMA de très grande taille*, in "19es Journées Francophones des Systèmes Multi-Agents", Valenciennes, France, Cépaduès, October 2011, vol. JFSMA'11, p. 53-62.

[55] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Organisation, agrégation et visualisation d'informations médiatiques*, in "Colloque "Fonder les sciences du territoire"", Paris, France, CIST, October 2011, p. 240-246, http://mescal.imag.fr/membres/jean-marc.vincent/papers/Cist2011.pdf.

Conferences without Proceedings

[56] C. EL KAED, F.-G. OTTOGALLI, Y. DENNEULIN. *Generation de mandataire pour l'interoperabilite des services*, in "8eme Conference Francaise des Systemes d'exploitation, CFSE", St Malo, France, May 2011, p. 480-485.

Scientific Books (or Scientific Book chapters)

[57] L. GENOVESE, M. OSPICI, B. VIDEAU, T. DEUTSCH, J.-F. MEHAUT. *Wavelet-based Density Functional Theory Calculation on Massive Parallel Hybrid Architectures*, in "GPU Computing Gems", Moran Kaufmann, January 2011.

Research Reports

[58] N. BALACHEFF, B. BUCCIO, P. CHAPUIS, J. COUTIN, J. COUTAZ, J. CROWLEY, Y. DENNEULIN, L. DU BOUSQUET, A. DUDA, R. ECHAHED, M.-C. FAUVET, C. GARBAY, E. GAUSSIER, I. GUILLET, C. LAUGIER, Y. LEDRU, A. LEGRAND, N. MANDRAN, H. MARTIN, J.-F. MEHAUT, T. MORTURIER, B. PLATEAU, E. PONS, J. PREVOST, F. PROST, V. QUINT, P. REIGNIER, F. ROUSSEAU, M.-C. ROUSSET, E. RUTTEN, M. VACHER. *The First Four Years (2007-2010) and Beyond, Volume 1: Research Program and Activity Report*, LIG, Grenoble, France, 2011, n° RR-LIG-015, http://rr.liglab.fr/research_report/RR-LIG-015.pdf.

[59] O. BEAUMONT, L. BOBELIN, H. CASANOVA, P.-N. CLAUSS, B. DE MOURA DONASSOLO, L. EYRAUD-DUBOIS, S. GENAUD, S. HUNOLD, A. LEGRAND, M. QUINSON, C. ROSA, L. SCHNORR, M. STILLWELL, F. SUTER, C. THIERY, P. VELHO, J.-M. VINCENT, W. J. YOUNG. *Towards Scalable, Accurate, and Usable Simulations of Distributed Applications and Systems*, INRIA, October 2011, n° RR-7761, http://hal.inria.fr/inria-00631141/en.

[60] A. BENOIT, M. GALLET, B. GAUJAL, Y. ROBERT. *Computing the throughput of probabilistic and replicated streaming applications*, INRIA, January 2011, n° RR-7510, http://hal.inria.fr/inria-00555890/en.

[61] R. BERTIN, S. HUNOLD, A. LEGRAND, C. TOUATI. *From Flow Control in Multi-path Networks to Multiple Bag-of-tasks Application Scheduling on Grids*, INRIA, September 2011, n° RR-7745, http://hal.inria.fr/inria-00627532/en.

[62] A. BOUILLARD, N. FARHI, B. GAUJAL. *Packetization and Aggregate Scheduling*, INRIA, July 2011, n° RR-7685, http://hal.inria.fr/inria-00608852/en.

[63] K. R. BRAGHETTO, J. E. FERREIRA, J.-M. VINCENT. *From Business Process Model and Notation to Stochastic Automata Network.*, IME-USP, March 2011, n° RT-MAC-2011-03.

[64] J. EMERAS, O. RICHARD, B. BZEZNIK. *Reconstructing the Software Environment of an Experiment with Kameleon*, INRIA, October 2011, n° RR-7755, http://hal.inria.fr/inria-00630044/en/.

[65] N. GAST, B. GAUJAL. *Markov chains with discontinuous drifts have differential inclusions limits. Application to stochastic stability and mean field approximation.*, INRIA, April 2011, n° RR-7315, http://hal.inria.fr/inria-00491859/en.

[66] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Macroscopic Observation of Multiagent Systems*, LIG, Grenoble, France, 2011, n° RR-LIG-010, http://rr.liglab.fr/research_report/RR-LIG-010.pdf.

[67] B. NEGREVERGNE, A. TERMIER, M.-C. ROUSSET, J.-F. MEHAUT. *ParaMiner: a Generic Parallel Pattern Mining Algorithm*, LIG, Grenoble, France, 2011, n° RR-LIG-012, http://rr.liglab.fr/research_report/RR-LIG-012.pdf.

[68] L. L. PILLA, C. POUSA RIBEIRO, D. CORDEIRO, A. BHATELE, P. O. A. NAVAUX, J.-F. MEHAUT, L. V. KALÉ. *Improving Parallel System Performance with a NUMA-aware Load Balancer*, INRIA-Illinois Joint Laboratory on Petascale Computing, July 2011, n° TR-JLPC-11-02, http://hdl.handle.net/2142/25911.

[69] A. TERMIER, B. NEGREVERGNE, S. MARLOW, S. SINGH. *HLCM: a first experiment on parallel data mining with Haskell*, LIG, Grenoble, France, 2011, n° RR-LIG-009, http://rr.liglab.fr/research_report/RR-LIG-009.pdf.

[70] P. VELHO, L. SCHNORR, H. CASANOVA, A. LEGRAND. *Flow-level network models: have we reached the limits?*, INRIA, November 2011, n° RR-7821, http://hal.inria.fr/hal-00646896/en/.

Scientific Popularization

[71] J. P. ARCHAMBAULT, E. BACCELLI, S. BOLDO, D. BOUHINEAU, P. CÉGIELSKI, T. CLAUSEN, G. DOWEK, I. GUESSARIAN, S. LOPÈS, L. MOUNIER, B. NGUYEN, F. QUESSETTE, A. RASSE, B. ROZOY, C. TIMSIT, T. VIÉVILLE, J.-M. VINCENT. *Une introduction à la science informatique: Pour les enseignants de la discipline informatique au lycée*, CNDP-CRDP Eds, 2011, http://crdp.ac-paris.fr/Introduction-a-la-science.