IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2012

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).

Creation of the Project-Team: January 01, 2006 .

1. Members

Research Scientists

Bruno Gaujal [Team leader, Senior Researcher (DR) Inria, HdR]
Derrick Kondo [Junior Researcher (CR), on leave since Aug. 2012]
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]
Florence Perronnin [Associate Professor, UJF]
Olivier Richard [Associate Professor, UJF]
Jean-Marc Vincent [Associate Professor, UJF]

Engineers

Romain Cavagna [2010-, Engineer Assistant, Inria]
Elodie Bertoncello [2012, Engineer]
Generoso Pagano [2012, Engineer]
Philippe Le Brouster [2011-2012, Engineer Assistant, UJF]
Pierre Navarro [2009-2011, Engineer Assistant, Inria]
Pierre Neyron [Research Engineer, CNRS]
Augustin Degomme [2012, ANR SONGS, CNRS]

PhD Students

Marcio Bastos Castro [2009-2012, Inria]
Rodrigue Chakode-Noumowe [2008-, Minalogic CILOE scholarship, Inria]
Pierre Coucheney [2008-2011, Inria-Alcatel Lucent scholarship]
Francieli Zanon-Boito [2012- ,co-tutelle Col.]
Erick Meneses-Cuadros [2012- ,co-tutelle Col.]
Charbel El Kaed [2008-2012, CIFRE France Télécom R&D scholarship]
Joseph Emeras [2010-, CNRS BDI scholarship]
Kiril Georgiev [2009-2012, CIFRE STMicroelectronics scholarship]
Patricia Lopez Cueva [2010-, CIFRE STmicroelectronics scholarship]
Kevin Pouget [2010-, CIFRE STMicroelectronics scholarship]
Christian Camilo Ruiz Sanabria [2011-, Inria]

Post-Doctoral Fellows

Sheng Di [2011-2012, Inria]
Slim Mohammed Bouguerra [2012, long visit in JPLC]
Lucas Schnorr [2012, ANR SONGS, CNRS]

Administrative Assistant

Annie Simon [Assistant (SAER), Inria]

Others

Laurent Bobelin [2010-2012, 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. MESCAL was evaluated again in October 2012. The feedback from the evaluation committee was not available at the time when this report was written.

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.

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).

Application domains concern wireless networks, intensive scientific computations and low power high performance computing. 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 of the Year

- Brigitte Plateau received the *Grand Prix des sciences de l'informatique et de leurs applications* of the EADS foundation.
- Panayotis Mertikopoulos received the best paper award at NETGCOOP 2012.

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.

Markov chains, Queuing networks, Mean field approximation, Simulation, Performance evaluation, Discrete event dynamic systems.

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. 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, Olivier Richard, Corinne Touati.

Administration, Deployment, Peer-to-peer, Clusters, Grids, Clouds, Job scheduler

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; Large scale data management

Participant: Yves Denneulin.

Fault tolerance, migration, distributed algorithms.

Most propositions to improve reliability 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. 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.

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).

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).

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. 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.

4.3. Cloud and Desktop Computing

Participants: Derrick Kondo, Arnaud Legrand, Olivier Richard.

The research of MESCAL on desktop grids has been very active and fruitful during the evaluation period. The main achievements concern the collection and statistical exploitation of traces in volunteer computing systems. Such models have enabled to optimize the behavior of volunteer computing systems or to extend the scope of their applicability. Such traces have also been used in SIMGRID to simulate volunteer computing systems at unprecedented scale.

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.

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).

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. CiGri: Computing ressource Reaper

Participant: Olivier Richard [correspondant].

CiGri is a middleware which gather the unused computing resource from intranet infrastructure and to make it available for large set of tasks. It manages the execution of large sets of parametric tasks on lightweight grid by submitting individual jobs to each batch scheduler. It's associated to the OAR resource management system (batch scheduler). Users can easily monitor and control their set of jobs through a web portal. System provides mechanisms to identify job error causes, to isolate faulty components and to resubmit job in a safer context. See also the web page http://cigri.imag.fr/

5.4. FTA: Failure Trace Archive

Participant: Derrick Kondo [correspondant].

The Failure Trace Archive is available at http://fta.inria.fr.

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

5.5. SimGrid: simulation of distributed applications

Participants: Arnaud Legrand [correspondant], Lucas Schnorr, Pierre Navarro, Degomme Augustin, 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/.

5.6. 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/.

5.7. 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/.

5.8. GameSeer: simulation of game dynamics

Participant: Panayotis Mertikopoulos [correspondant].

Mathematica toolbox (graphical user interface and functions library) for efficient, robust and modular simulations of game dynamics.

5.9. Kameleon: environment for experiment reproduction

Participants: Olivier Richard [correspondant], Joseph Emeras.

Kameleon is a tool developed to facilitate the building and rebuilding of software environment. It helps experimenter to manage his experiment's software environment which can include the operating system, libraries, runtimes, his applications and data. This tool is an element in the experimental process to obtain repeatable experiments and therefore reproducible results.

6. New Results

6.1. Analysis and Control of Large Stochastic Systems

Perfect sampling is a very efficient technique that uses coupling arguments to provide a sample from the stationary distribution of a Markov chain in a finite time without ever computing the distribution. Even though, the general (non-monotone) case needs to consider the whole state space, we developed a new approach for the general case that only needs to consider two trajectories, an approach which is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed [8]. Importantly, we also showed that perfect sampling is possible in Jackson networks, even though the underlying Markov chain has a large or even infinite state space and illustrated the efficiency of our approach via numerical simulations [17]. In a similar vein, given that the analysis of a system's dynamics relies on the collection and the description of events, we developed in [37] a new approach to reduce the descriptional complexity of a system by aggregating events' properties, such as their Shannon entropy, entropy gain, divergence etc. These measures were applied to the evaluation of geographic aggregations in the context of news analysis and they allowed us to determine which abstractions one should prefer depending on the task to perform.

In the study of Markov decision processes composed of a large number of objects, we showed that the optimal reward satisfies a Bellman equation, which converges to the solution of a continuous Hamilton-Jacobi-Bellman (HJB) equation based on the mean field approximation of the Markov decision process [10]. We also gave bounds on the difference of the rewards and an algorithm for deriving an approximating solution to the Markov decision process from a solution of the HJB equations. Furthermore, we also studied deterministic limits of Markov processes with discontinuous drifts and showed that under mild assumptions, the stochastic system is a constant-step stochastic approximation algorithm which converges to a differential inclusion obtained by convexifying the rescaled drift of the Markov chain [9].

Finally, in terms of performance evaluation and its applications, we also studied resource-aware business process models by defining a new framework that allows the generation of analytical models. We showed that the analysis of the generated SAN model provides several performance indices we showed that these indices can be easily calculated by a business specialist with no skills in stochastic modeling [7].

6.2. Game Theory and Applications

As far as results in pure game theory are concerned, we studied in [12] a general framework of systems wherein there exists a Pareto optimal allocation that is Pareto superior to an inefficient Nash equilibrium and defined a 'Nash proportionately fair' Pareto optima. In this context, we provided conditions for the existence of a Pareto-optimal allocation that is, truly or most closely, proportional to a Nash equilibrium – an approach with applications in non-cooperative flow-control problems in communication networks.

In a learning context, we also explored what happens beyond the standard first-order framework of continuous time game dynamics and introduced in [42] a class of higher order game dynamics, extending all first order imitative dynamics, and, in particular, the replicator dynamics to higher orders. In stark contrast to the first order case, we showed that weakly dominated strategies become eliminated in all n-th order payoff-monotonic dynamics for all $n > 1$ and strictly dominated strategies become extinct in n-th order dynamics n orders as fast as in first order. Finally, we also established a higher order analogue of the folk theorem of evolutionary game theory which shows that higher order accelerate the rate of convergence to equilibria in games.

In terms of applications, we also examined the distribution of traffic in networks whose users try to minimise their delays by adhering to a simple learning scheme inspired by the replicator dynamics of evolutionary game theory. A major challenge occurs in this context when the users' delays fluctuate unpredictably due to random external factors, but we showed that if users are not too greedy in their learning scheme, then the long-term averages of the users' traffic flows converge to the vicinity of an equilibrium [43].

6.3. Wireless networks

Power and energy considerations in wireless networks have brought to the forefront the need for efficient power allocation and handover policies.

In [13], we analyze the power allocation problem for orthogonal multiple access channels by means of a non-cooperative potential game in which each user distributes his power over the channels available to him. When the channels are static, we show that this game possesses a unique optimum point; moreover, if the network's users follow a distributed learning scheme based on the replicator dynamics of evolutionary game theory, then they converge to this optimum exponentially fast.

On the other hand, in case the network users have access to multiple-antenna technologies (as most smarphone users do nowadays, we also analyze the problem of finding the optimal signal covariance matrix for MIMO multiple access channels by using an approach based on "exponential learning" – a novel optimization method which applies more generally to (quasi-)convex problems defined over sets of positive-definite matrices (with or without trace constraints) [24]. Furthermore, by using a Riemannian-geometric approach, we devise a distributed optimization algorithm which converges to the optimum signal distribution exponentially fast: users attain an ϵ-neighborhood of the system's optimum configuration in time which is at most $\mathcal{O}(\log(1/\epsilon))$ (and, in practice, within only a few iterations, even for large numbers of users) [25].

In the context of heterogeneous wireless networks where vertical handovers are allowed, we also studied a control problem for a new joint admission and resource allocation controller. To account for multi-objective optimization, we considered the maximization of an objective subject to a set of constraints, and we turned this constrained problem into an unconstrained one that we solved numerically using the Semi-Markovian Decision Process (SMDP) framework [19].

6.4. Scheduling

The parallel computing platforms available today are increasingly larger, so it is necessary to develop efficient strategies providing safe and reliable completion for parallel applications. In [6], we proposed a performance model that expresses formally the checkpoint scheduling problem by exhibiting the tradeoff between the impact of the checkpoints operations and the lost computation due to failures. In particular, we proved that the checkpoint scheduling problem is NP-hard even in the simple case of uniform failure distribution and also presented a dynamic programming scheme for determining the optimal checkpointing times in all variants of the problem. On a similar issue, we proposed in [35] a fair scheduling algorithm that handles the problem of fair scheduling by adopting processor fair-share as a strategy for user fairness. Our approach showed that a parallel machine can give a similar type of performance guarantee as a round-robin scheduler, without requiring job preemption been required.

From a network calculus perspective, we presented in [16] a new formalism for data packetization in networks, the "packet curves". Indeed, a more precise knowledge of the packet characteristics can be efficiently exploited to get tighter performance bounds, for example for aggregation of flows, packet-based service policies and shared buffers; finally, we also gave a model for a wormhole switch and showed how our results can be used to get efficient delay bounds.

6.5. Multi-Core Systems

Modern multi-core platforms feature complex topologies with different cache levels and hierarchical memory subsystems, so thread and data placement become crucial to achieve good performance. In [14], we evaluate CPU and memory affinity strategies for numerical scientific multithreaded benchmarks on multi-core platforms and analyzed hardware performance event counters in order to acquire a better understanding of such impact. Likewise, thread mapping is an appealing approach to efficiently exploit the potential of modern chip-multiprocessors, so we proposed in [18] a dynamic thread mapping approach to automatically infer a suitable thread mapping strategy for transactional memory applications composed of multiple execution phases with potentially different transactional behavior in each phase. Our results showed that the proposed dynamic approach presents performance improvements up to 31% compared to the best static solution. esp From an optimization perspective, the asymmetry in memory access latencies may reduce the overall performance of the system. Therefore, to achieve scalable performance in this environment, we exploited in [28] the machine architecture while taking into account the application communication patterns. Specifically, we introduced a topology-aware asymptotically optimal load balancing algorithm named HwTopoLB which combines the machine topology characteristics with the communication patterns of the application to equalize the application load on the available cores while reducing latencies. We also introduced in [27] a topology-aware load balancer called NucoLB that focuses on redistributing work while reducing communication costs among and within compute nodes, thus leading to performance improvements of up to 20% when compared to state-of-the-art load balancers.

6.6. Cloud Computing

Even though a new era of Cloud Computing has emerged, the characteristics of Cloud load in data centers is not perfectly clear. In [20], we characterized the job/task load and host load in a real-world production data center at Google Inc. by using a detailed trace of over 25 million tasks across over 12,500 hosts. We found that the Google data center exhibits finer resource allocation with respect to CPU and memory than that of Grid/HPC systems and Google jobs are always submitted with much higher frequency and they are much

shorter than Grid jobs, leading to higher variance and noise. Moreover, as far as prediction is concerned, we designed in [21] a Bayes model to predict the mean load over a long-term time interval, as well as the mean load in consecutive future time intervals. Real-world experiments showed that our Bayes method achieved high accuracy with a mean squared error of 0.0014 and that it improves the load prediction accuracy by 5.6-50% compared to other state-of-the-art methods based on moving averages, auto-regression, and/or noise filters.

In a similar vein, the exploitation of Best Effort Distributed Computing Infrastructures (BE-DCIs) allows operators to maximize the utilization of the infrastructures, and users to access the unused resources at relatively low cost. Profiling the execution of Bag-of-Tasks (BoT) applications on several kinds of BE-DCIs demonstrates that their task completion rate drops near the end of the execution. In [33], we presented the SpeQuloS service which enhances the QoS of BoT applications executed on BE-DCIs by reducing the execution time, improving its stability, and reporting to users a predicted completion time. We presented the design and development of the framework and several strategies to decide when and how Cloud resources should be provisioned; moreover, performance evaluation using simulations showed that SpeQuloS fulfill its objectives in speeding up the execution of BoTs, in the best cases by a factor greater than 2, while offloading less than 2.5% of the workload to the Cloud. These topics were also further explored in the book chapter [30].

6.7. Experimentation and Visualization in Large Systems

Despite a widespread belief regarding the simulation of large-scale computing systems, we showed in [15] that achieving high scalability does not necessarily require to resort to overly simple models and ignore important phenomena. In fact, by relying on a modular and hierarchical platform representation while taking advantage of regularity when possible, we were able to model systems such as data and computing centers, peer-to-peer networks, grids, or clouds in a scalable way. Finally, in [34], we examined the ability to conduct consistent, controlled, and repeatable large-scale experiments in areas of computer science where availability, repeatability, and open sharing of electronic products are still difficult to achieve.

We also discussed in [22] the concept of the reconstructability of software environments and we proposed a tool for dealing with this problem. In a similar vein, we developed Expo [41], a tool for conducting experiments on distributed platforms. Our experiments confirmed that Expo is a promising tool to help the user with two primary concerns: how to perform a large scale experiment efficiently and easily, together with its reproducilibity.

The exponential number of processes that are executed in high performance applications and the very detailed behavior that we can record about them lead to a trace size explosion both in space and time dimensions. Thus, if the amount of data is not properly treated for visualization, the analysis may give the wrong idea about the behavior registered in the traces. We dealt with this issue in [38] in two ways: first, by detailing data aggregation techniques that are fully configurable by the user to control the level of details in both space and time dimensions, and second, by presenting two visualization techniques that take advantage of the aggregated data to scale.

Furthermore, given that the performance of parallel and distributed applications is highly dependent on the characteristics of the execution environment, the network topology and characteristics directly impact data locality and movements as well as contention. Unfortunately few visualization available to the analyst are capable of accounting for such phenomena, so we proposed in [39] an interactive topology-based visualization technique based on data aggregation that enables to correlate network characteristics, such as bandwidth and topology, with application performance traces. Such visualization techniques enable us to explore and understand non-trivial behaviors that are impossible to grasp otherwise and the combination of multi-scale aggregation and dynamic graph layout allows us to scale the visualization seamlessly to large distributed systems.

7. Bilateral 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. 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.

7.2. Grants with Industry

7.2.1. CIFRE contracts with STMicroelectronics

- Kiril Georgiev has done his PhD with STMicroelectronics on distributed file systems and defended in Dec. 2012.

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. ANR

- *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.

- *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.

- *ANR SONGS, 2012-2015*

Partners: Inria Nancy (Algorille), Inria Sophia (MASCOTTE), Inria Bordeaux (CEPAGE, HiePACS, RunTime), Inria Lyon (AVALON), University of Strasbourg, University of Nantes

The last decade has brought tremendous changes to the characteristics of large scale distributed computing platforms. Large grids processing terabytes of information a day and the peer-to-peer technology have become common even though understanding how to efficiently such platforms still raises many challenges. As demonstrated by the USS SimGrid project funded by the ANR in 2008, simulation has proved to be a very effective approach for studying such platforms. Although even more challenging, we think the issues raised by petaflop/exaflop computers and emerging cloud infrastructures can be addressed using similar simulation methodology.

The goal of the SONGS project (Simulation of Next Generation Systems) is to extend the applicability of the SimGrid simulation framework from Grids and Peer-to-Peer systems to Clouds and High Performance Computation systems. Each type of large-scale computing system will be addressed through a set of use cases and lead by researchers recognized as experts in this area.

Any sound study of such systems through simulations relies on the following pillars of simulation methodology: Efficient simulation kernel; Sound and validated models; Simulation analysis tools; Campaign simulation management.

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. 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.

8.4. International Initiatives

8.4.1. Inria Associated 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: 2012 - 2013

See also: http://mescal.imag.fr/membres/derrick.kondo/ea/ea.html

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).

8.4.2. 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). Slim Bouguerra is visiting JLPC for an extended period (one year).

- MESCAL also has long lasting collaborations with University of California in Berkeley and a new one with Google. Bruno Gaujal, Derrick Kondo and Arnaud Legrand visited Berkeley in 2012.

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 for 2011-2012 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.
- 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. Slim Bougherra (Mescal Postdoc) is visiting JLPC for one year, starting Jan. 2012.

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 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 president of Grenoble-INP.
- Yves Denneulin is the director of Grenoble-INP ENSIMAG.
- 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).
- Bruno Gaujal has been a member of selection committees in Grenoble.

9.1.1. Invited Talks

- Bruno Gaujal gave an invited talk at *Piecewise Deterministic Markov Process Days*, Marne la Vallée, 2012. He was invited to give a series of lectures at the Workshop *Weak Kam Dynamics and mean field games and control*, Warwick Mathematics Dept., 2012. He wa also invited to give a talk at the *Complex systems workshop*,Rescom, Paris, Nov. 2012.
- Bruno Gaujal and Corinne Touati are invited to give a one hour lecture each at the autumn school on *Analyse D'Algorithmes et Modèles Aléatoires* (ADAMA), Monastir, October 2012 on game theory for networks.
- Panayotis Mertikopoulos was invited to give talks at ENSEA, Wireless Networks Seminar, March 2012, Ecole Polytechnique, Economic Theory Seminar, March 2012, Games and Strategy in Paris (International Conference for Sylvain Sorin's 60th birthday), June 2012, Paris Game Theory Seminar, May 2012, University of Athens, Wireless Systems Seminar, October 2012.
- Arnaud Legrand has given an invited talk at JLPC (Argone), November 2012, at the workshop *New challenges in scheduling theory*, Frejus, Oct. 2012, and invited seminars at the PUC Rio and at UFRGDS, Brazil.

9.1.2. Journal, Conference and Workshop Organization

- Panayotis Mertikopoulos is the publications chair of Valuetools 2012 (the 6th International ICST Conference on Performance Evaluation Methodologies and Tools). He is also the co-organizer of the Paris Working Group on Evolutionary Games.
- Jean-Marc Vincent has been the general chair of the ASMTA conference, June 2012.
- Corinne Touati has been the general chair of ValueTools 2012, Cargese, Oct. 2012.
- Bruno Gaujal, Panayotis Mertikopoulos and Corinne Touati organized the Rescom summer school on Game Theory for networks, Vittel, June 2012.
- Panayotis Mertikopoulos was a publicity chair of the 11th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks.
- Arnaud Legrand has organized the *simgrid user days* in Lyon.

9.1.3. Program Committees

- Bruno Gaujal was the Technical Program Committee Chair of ValueTools 2012.
- Bruno Gaujal was program committee member of Wodes 2012.
- Panayotis Mertikopoulos was program committee member of ValueTools 2012.
- Arnaud Legrand was program committee member of IPDPS 2012.

9.1.4. Thesis Defense

- Marcio Bastos Castro, *Improving the Performance of Transactional Memory Applications on Multicores: A Machine Learning-based Approach*, Dec. 3, 2012.
- Kiril Geogiev, *Débogage des systèmes embarqués multiprocesseur basé sur la ré-exécution déterministe et partielle*, Dec. 4, 2012 [5].
- Slim Bouguerra, *Tolérance aux pannes dans des environnements de calcul parèlles et distribués: optimisation des stratégies de sauvegarde/reprise et ordonnancement*, Apr. 12, 2012 [4].

9.1.5. Thesis Committees

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

- Bruno Gaujal served on the thesis commitees of Laurent Jouhet (president), Adrien Brandejsky (reviewer), Xavier Koegler (reviewer).

9.1.6. Popular Science

- MESCAL 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.

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 2012.

- Performance evaluation. Two new lectures on system modeling and performance evaluation started in Ensimag in 2012 at M1 and M2 levels.
- Game Theory and Applications, 24 hours, M2R, ENS de Lyon, France.
- Stochastic System Optimisation, 36 hours, MPRI, Paris.
- "Research School" (open to M1, M2 and PhD students) on Game Theory for Networks, Rescom Summer School, Vittel, France.
- Olivier Richard was involved in the creation of a new "fab lab" in the University Joseph Fourier.
- Arnaud Legrand is responsible of the Parallel systems track at the MOSIG.

10. Bibliography

Major publications by the team in recent years

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

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

[3] B. JAVADI, D. KONDO, J.-M. VINCENT, D. 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] M. S. BOUGUERRA. *Tolérance aux pannes dans des environnements de calcul parallèle et distribué : optimisation des stratégies de sauvegarde/reprise et ordonnancement*, University of Grenoble, April 2012.

[5] K. GEORGIEV. *Débogage des Systèmes Embarqués Multiprocesseur basé sur la Ré-exécution Déterministe et Partielle*, Université de Grenoble, Ecole Doctorale MSTII, December 2012.

Articles in International Peer-Reviewed Journals

[6] M. S. BOUGUERRA, D. TRYSTRAM, F. WAGNER. *Complexity Analysis of Checkpoint Scheduling with Variable Costs*, in "IEEE Transactions on Computers", 2012, vol. 99, n° PrePrints, http://doi.ieeecomputersociety.org/10.1109/TC.2012.57.

[7] K. R. BRAGHETTO, J. E. FERREIRA, J.-M. VINCENT. *Performance Evaluation of Resource-Aware Business Processes Using Stochastic Automata Networks*, in "IJICIC", July 2012, vol. 8, n^{o} 7(B), p. 5295-5380, http://www.ijicic.org/vol-8(7)b.htm.

[8] A. BUSIC, B. GAUJAL, F. PIN. *Perfect Sampling of Markov Chains with Piecewise Homogeneous Events*, in "Performance Evaluation", 2012, http://www.sciencedirect.com/science/article/pii/S0166531612000120.

[9] N. GAST, B. GAUJAL. *Markov chains with discontinuous drifts have differential inclusion limits*, in "Performance Evaluation", 2012, to appear.

[10] N. GAST, B. GAUJAL, J.-Y. LE BOUDEC. *Mean field for Markov Decision Processes: from Discrete to Continuous Optimization*, in "IEEE Transactions on Automatic Control", 2012, vol. 57, n^{o} 8, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06144708.

[11] B. GAUJAL, L. GULYAS, Y. MANSURI, E. THIERRY. *Can heterogeneous agent-based models be validated? A discrete Markov chain approach*, in "Journal of Economic Dynamics & Control", 2012, accepted for publication.

[12] H. KAMEDA, E. ALTMAN, C. TOUATI, A. LEGRAND. *Nash Equilibrium Based Fairness*, in "Mathematical Methods of Operations Research", 2012, vol. 76, n^{o} 1, http://www.springerlink.com/content/25u129434k887rkk/.

[13] P. MERTIKOPOULOS, E. BELMEGA, A. MOUSTAKAS, S. LASAULCE. *Distributed learning policies for power allocation in multiple access channels*, in "IEEE Journal on Selected Advances on Communications", January 2012, vol. 30, n^{o} 1, p. 96-106.

[14] C. POUSA RIBEIRO, M. CASTRO, V. MARANGOZOVA-MARTIN, J.-F. MEHAUT, H. COTA DE FREITAS, C. AUGUSTO PAIVA DA SILVA MARTINS. *Evaluating CPU and Memory Affinity for Numerical Scientific Multi-threaded Benchmarks on Multi-cores*, in "IADIS International Journal on Computer Science and Information Systems (IJCSIS)", 2012, http://www.iadis.net/dl/Search_list_open.asp?code=7225.

International Conferences with Proceedings

[15] L. BOBELIN, A. LEGRAND, D. A. GONZÀLEZ MÀRQUEZ, P. NAVARRO, M. QUINSON, F. SUTER, C. THIERY. *Scalable Multi-Purpose Network Representation for Large Scale Distributed System Simulation*, in "Proceedings of the 12th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'12)", IEEE Computer Society Press, May 2012, http://hal.inria.fr/hal-00650233.

[16] A. BOUILLARD, N. FARHI, B. GAUJAL. *Packetization and packet curves in network calculus*, in "VALUETOOLS '12: Proceedings of the 6th International Conference on Performance Evaluation Methodologies and Tools", Cargèse, France, IEEE explore, 2012, Invited paper.

[17] A. BUSIC, B. GAUJAL, F. PERRONNIN. *Perfect Sampling of Networks with Finite and Infinite Capacity Queues*, in "19th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA) 2012", Grenoble, 2012.

[18] M. CASTRO, L. FABRÍCIO GÓES, L. GUSTAVO FERNANDES, J.-F. MEHAUT. *Dynamic Thread Mapping Based on Machine Learning for Transactional Memory Applications*, in "International European Conference

on Parallel and Distributed Computing (Euro-Par)", Rhodes Island, Greece, Lecture Notes in Computer Science (LNCS), Springer-Verlag, 2012, vol. 7484, p. 465-476.

[19] P. COUCHENEY, E. HYON, C. TOUATI. *Admission and Allocation Policies in Heterogeneous Wireless Networks with Handover*, in "VTC2012-Spring: 2012 IEEE 75th Vehicular Technology Conference", 2012.

[20] S. DI, D. KONDO, W. CIRNE. *Characterization and Comparison of Google Cloud Load versus Grids*, in "Proceedings of the IEEE Cluster Conference", 2012.

[21] S. DI, D. KONDO, W. CIRNE. *Host Load Prediction in a Google Compute Cloud with a Bayesian Model*, in "IEEE/ACM Supercomputing Conference (SC)", November 2012.

[22] J. EMERAS, O. RICHARD, B. BZEZNIK. *Reconstructing the Software Environment of an Experiment with Kameleon*, in "ACM Compute", January 2012.

[23] V. MARANGOZOVA-MARTIN, G. PAGANO. *SoC-TRACE: Handling the Challenge of Embedded Software Design and Optimization*, in "Proceedings of the ACM/IFIP/Usenix International Middleware Conference", Montreal, Canada, December 2012.

[24] P. MERTIKOPOULOS, E. BELMEGA, A. MOUSTAKAS. *Matrix Exponential Learning: Distributed Optimization in MIMO systems*, in "ISIT '12: Proceedings of the 2012 IEEE International Symposium on Information Theory", 2012.

[25] P. MERTIKOPOULOS. *Strange bedfellows: Riemann, Gibbs and vector Gaussian multiple access channels*, in "NetGCoop '12: Proceedings of the 6th International Conference on Network Games, Control and Optimization", 2012.

[26] L. PILLA, P. NAVAUX, J.-F. MEHAUT. *Communication Cost-Aware Load Balancing for Architectures with Asymetric Performance Behavior*, in "International Supercomputing Conference, ISC 2012", Hamburg, Germany, June 2012.

[27] L. PILLA, C. POUSA RIBEIRO, D. CORDEIRO, C. MEI, A. BHATELE, P. NAVAUX, F. BROQUEDIS, J.-F. MEHAUT, L. KALE. *A Hierarchical Approach for Load Balancing on Parallel Multi-core Systems*, in "Proceedings of the 41st International Conference on Parallel Processing, ICPP 2012", Pittsburgh, Pennsylvania, September 2012, http://dx.doi.org/10.1109/ICPP.2012.9.

[28] L. PILLA, C. POUSA RIBEIRO, P. NAVAUX, P. COUCHENEY, F. BROQUEDIS, B. GAUJAL, J.-F. MEHAUT. *Asymptotically Optimal Load Balancing for Hierarchical Multi-Core Systems*, in "Proceedings of the 18th IEEE International Conference on Parallel and Distributed Systems, ICPADS", Singapore, December 2012.

[29] K. POUGET, M. SANTANA, V. MARANGOZOVA-MARTIN, J.-F. MEHAUT. *Debugging Component-Based Embedded Applications*, in "Joint Workshop Map2MPSoC (Mapping of Applications to MPSoCs) and SCOPES (Software and Compilers for Embedded Systems)", St Goar, Germany, May 2012, published in the ACM library.

Scientific Books (or Scientific Book chapters)

[30] A. ANDRZEJAK, D. KONDO. *Modeling and Optimizing Availability of Non-Dedicated Resources*, in "Desktop grid computing", C. CERIN, G. FEDAK (editors), Chapman & Hall/CRC numerical analysis and scientific computing, CRC Press, 2012, http://www.crcpress.com/product/isbn/9781439862148.

Books or Proceedings Editing

[31] K. AL-BEGAIN, D. FIEMS, J.-M. VINCENT (editors). *Analytical and Stochastic Modeling Techniques and Applications: 19th International Conference, ASMTA 2012, Grenoble, France, June 4-6, 2012*, 2012, Springer, May 2012.

[32] C. TOUATI, B. GAUJAL, A. JEAN-MARIE, E. JORSWIECK, A. SEURET (editors). *Performance Evaluation Methodologies and Tools: 6th International Conference, VALUETOOLS 2012*, 2012, Springer, Cargèse, France, October 2012.

Research Reports

[33] S. DELAMARE, G. FEDAK, D. KONDO, O. LODYGENSKY. *SpeQuloS: A QoS Service for BoT Applications Using Best Effort Distributed Computing Infrastructures*, Inria, February 2012, n^{o} RR-7890, http://hal.inria.fr/hal-00672046.

[34] F. DESPREZ, G. FOX, E. JEANNOT, K. KEAHEY, M. KOZUCH, D. MARGERY, P. NEYRON, L. NUSSBAUM, C. PÉREZ, O. RICHARD, W. SMITH, G. VON LASZEWSKI, J. VÖCKLER. *Supporting Experimental Computer Science*, Argonne National Lab, March 2012, n^{o} MCS Technical Memo 326, http://hal.inria.fr/hal-00720815.

[35] J. EMERAS, V. PINHEIRO, K. RZADCA, D. TRYSTRAM. *Fair Scheduling for Multiple Submissions*, LIG, Grenoble, France, 2012, n^{o} RR-LIG-033, http://rr.liglab.fr/research_report/RR-LIG-033_orig.pdf.

[36] N. GAST, B. GAUJAL. *Markov chains with discontinuous drifts have differential inclusions limits. Application to stochastic stability and mean field approximation*, Inria, March 2012, n^{o} RR-7315, http://hal.inria.fr/inria-00491859/en.

[37] R. LAMARCHE-PERRIN, J.-M. VINCENT, Y. DEMAZEAU. *Informational Measures of Aggregation for Complex Systems Analysis*, LIG, Grenoble, France, 2012, n^{o} RR-LIG-026, http://rr.liglab.fr/research_report/RR-LIG-026.pdf.

[38] L. MELLO SCHNORR, A. LEGRAND. *Visualizing More Performance Data Than What Fits on Your Screen*, Inria, September 2012, n^{o} RR-8079, 17, http://hal.inria.fr/hal-00737651.

[39] L. MELLO SCHNORR, A. LEGRAND, J.-M. VINCENT. *Interactive Analysis of Large Distributed Systems with Topology-based Visualization*, Inria, September 2012, n^{o} RR-8085, 24, http://hal.inria.fr/hal-00738321.

[40] G. PAGANO, V. MARANGOZOVA-MARTIN. *SoC-Trace Infrastructure*, Inria, July 2012, n^{o} RT-0427, http://hal.inria.fr/hal-00719745/.

[41] C. C. RUIZ SANABRIA, O. RICHARD, B. VIDEAU, I. OLEG. *Managing Large Scale Experiments in Distributed Testbeds*, Inria, October 2012, n^{o} RR-8106, http://hal.inria.fr/hal-00742582.

Other Publications

[42] R. LARAKI, P. MERTIKOPOULOS. *Higher Order Game Dynamics*, June 2012, http://arxiv.org/abs/1206.4181, http://arxiv.org/abs/1206.4181.

[43] P. MERTIKOPOULOS, A. MOUSTAKAS. *A learning approach to efficient routing and the effect of stochastic fluctuations*, 2012, http://arxiv.org/abs/0912.4012, http://arxiv.org/abs/0912.4012.