IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2013

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

Creation of the Project-Team: 2006 January 01.

1. Members

Research Scientists

Bruno Gaujal [Team leader, Inria, Senior Researcher, HdR]
Arnaud Legrand [CNRS, Researcher]
Panayotis Mertikopoulos [CNRS, Researcher]
Corinne Touati [Inria, Researcher]

Faculty Members

Yves Denneulin [Grenoble INP, Professor, HdR]
Florence Perronnin [Univ. Grenoble I, Associate Professor]
Olivier Richard [Univ. Grenoble I, Associate Professor]
Jean-Marc Vincent [Univ. Grenoble I, Associate Professor]

External Collaborators

Bruno Bzeznik [Univ. Grenoble I, Engineer]
Vania Martin [Univ. Grenoble I, Associate Professor]
Jean-François Mehaut [Univ. Grenoble I, Professor, HdR]
Brice Videau [CNRS, Temporary Researcher]

Engineers

Marcio Bastos Castro [Inria, granted by Préfecture de la Région Rhône-Alpes, until May 2013]
Elodie Bertoncello [Inria]
Maxime Boutserin [Inria, from Oct 2013]
Romain Cavagna [Univ. Grenoble I]
Augustin Degomme [CNRS]
Sheng Di [Inria, from Dec 2013]
Salem Harrache [Inria, from Oct 2013]
Michaël Mercier [Inria]
Pierre Neyron [CNRS]
Generoso Pagano [Inria, granted by Préfecture de la Région Rhône-Alpes]
Christian Seguy [CNRS]
Matthieu Volat [Inria, until Mar 2013]

PhD Students

Poliana Correa de Oliveira [Inria, granted by Préfecture de la Région Rhône-Alpes, from May 2013 until Jul 2013]
Alexis Martin [Inria, granted by Préfecture de la Région Rhône-Alpes, until Jul 2013]
Erick Ramon Meneses Cuadros [Univ. Grenoble, granted by CIFRE]
Cristian Camilo Ruiz Sanabria [Inria]
Luka Stanisic [Univ. Grenoble I]
Rodrigo Virote Kassick [Cotutelle with UFRGS]
Francieli Zanon-Boito [Cotutelle with UFRGS]

Post-Doctoral Fellows

Mohamed-Slim Bouguerra [Inria, until Aug 2013]
Joseph Emeras [Inria, until Sep 2013]

Lucas Mello Schnorr [CNRS, until Jan 2013]

Visiting Scientist

Rhonda Righter [Inria, from May 2013 until May 2013]

Administrative Assistant

Annie Simon [Inria]

Others

Marion Dalle [Inria, intern, from May 2013 until Aug 2013]
Stéphane Durand [ENS Lyon, intern]
Sergio Gelvez Cortes [Inria, intern, from Jun 2013 until Sep 2013]
Valentin Gledel [Inria, intern, from Jun 2013 until Jul 2013]
Wagner Kolberg [Inria, intern, until Jan 2013]
Thomas Messi Nguele [Inria, intern, from Feb 2013 until May 2013]
Arnaud Panaiotis [Inria, intern, from Feb 2013 until Jul 2013]
Baptiste Roziere [Inria, intern, from Jun 2013 until Jul 2013]

2. Overall Objectives

2.1. Presentation

MESCAL is a project-team of Inria jointly with UJF and Grenoble INP universities and CNRS, created in 2006 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 and renewed for another 4 years.

2.2. Objectives

The recent evolutions in network and computer technology, as well as their diversification, go 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, BOINC) as well as clouds (Amazon, Google clouds) and communication networks (3G, LTE and Wifi networks).

Application domains concern intensive scientific computations and low power high performance computing. We are also 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.

3. Research Program

3.1. Large System Modeling and Analysis

Participants: Bruno Gaujal, Arnaud Legrand, Panayotis Mertikopoulos, Florence Perronnin, 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 unfeasible 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 specifically targets the simulation of large distributed systems (grids, clusters, peer-to-peer systems, volunteer computing systems, clouds) from the perspective of applications. It enables to obtain repeatable results and 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 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 tool 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: 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 of 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 queries 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.

4. Application Domains

4.1. Cloud, Grid, High Performance and Desktop Computing

Participants: 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 and in cloud infrastructures. 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. We can also mention the work conducted in SimGrid and which has also allowed to simulate HPC applications and platforms very accurately. Last, we should mention the continuous work on OAR and G5K, in particular on the experiment reconstructability aspect.

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. 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 computations 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, analyses 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.

5. Software and Platforms

5.1. Tools for cluster management and software development

Participant: Olivier Richard [correspondent].

The KA-Tools is a software suite developed 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 [correspondent].

The OAR project (see http://oar.imag.fr) 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). OAR is perfectly suited to such developments thanks to its highly modular architecture. Moreover, this development is used for the CiGri grid middleware project.

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

5.3. CiGri: Computing resource Reaper

Participant: Olivier Richard [correspondent].

CiGri (see http://cigri.imag.fr/) is a middleware which gathers the unused computing resource from intranet infrastructure and makes it available for the processing of 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 is s associated to the OAR resource management system (batch scheduler). Users can easily monitor and control their set of jobs through a web portal. CiGri provides mechanisms to identify job error causes, to isolate faulty components and to resubmit jobs in a safer context.

5.4. FTA: Failure Trace Archive

The Failure Trace Archive [11] is available at http://fta.inria.fr. Since Derrick Kondo left on sabbatical, the Failure Trace Archive has been migrated to University of Western Sidney, Australia (http://fta.scem.uws.edu.au/), which allows an easier management by his colleagues Bahman Javadi who was working as a post-doc in the MESCAL team while initializing the FTA.

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/Inria, 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 [correspondent], Lucas Mello Schnorr, Luka Stanisic, Augustin Degomme.

SimGrid (see http://simgrid.gforge.inria.fr/) 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.

5.6. TRIVA: interactive trace visualization

Participants: Lucas Mello Schnorr [correspondent], Arnaud Legrand.

TRIVA (see http://triva.gforge.inria.fr/) 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.

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

Participant: Jean-Marc Vincent [correspondent].

ψ and ψ^2 (see http://psi.gforge.inria.fr) 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.

5.8. GameSeer: simulation of game dynamics

Participant: Panayotis Mertikopoulos [correspondent].

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 [correspondent], Joseph Emeras.

Kameleon is a tool developed to facilitate the building and rebuilding of software environment. It helps the 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.

5.10. Platforms

5.10.1. Grid'5000

The MESCAL project-team is involved in development and management of Grid'5000 platform. The Digitalis and IDPot clusters are integrated in Grid'5000 as well as of CIMENT.

5.10.2. The ICluster-2, 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.

5.10.3. 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.

6. New Results

6.1. Simulation

6.1.1. Simulation of Parallel Computing Systems

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 [13] we focus on the validation of state-of-the-art flow-level network models of TCP communication on Wide Area Networks, 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.

The previous results assume steady-state and provide thus a reasonable model when message size is very large. Although, such assumptions may be reasonable when studying grid applications, when simulating HPC applications message sizes are often much smaller and phenomenon like slow-start or how communications and computations overlap have to be accurately modeled. Simulation and modeling for performance prediction and profiling is yet essential for developing and maintaining HPC code that is expected to scale for next-generation exascale systems. In [15], [34] we describe an implementation of a flow-based hybrid network model that accounts for factors such as network topology and contention, which are commonly ignored by the LogP models. Although, this may seem like a strange choice, we focus on large-scale, Ethernet-connected systems, as these currently compose 37.8% of the TOP500 index, and this share is expected to increase as higher-speed 10 and 100GbE become more available. Furthermore, the European Mont-Blanc project to study exascale computing by developing prototype systems with low-power embedded devices will also use Ethernet-based interconnect [28]. Our model is implemented within SMPI, an open-source MPI implementation that connects real applications to the SimGrid simulation framework. SMPI provides implementations of collective communications based on current versions of both OpenMPI and MPICH. SMPI and SimGrid also provide methods for easing the simulation of large-scale systems, including shadow execution, memory folding, and support for both online and offline (i.e., post-mortem) simulation. We validate our proposed model by comparing traces produced by SMPI with those from real world experiments, as well as with those obtained using other established network models. Our study shows that SMPI has a consistently better predictive power than classical LogP-based models for a wide range of scenarios including both established HPC benchmarks and real applications.

6.1.2. Perfect Simulation

Perfect simulation 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. In [7], we consider Jackson queueing networks (JQN) with finite buffer constraints and analyze the efficiency of sampling from their stationary distribution. In the context of exact sampling, the monotonicity structure of JQNs ensures that such efficiency is of the order of the coupling time (or meeting time) of two extremal sample paths. In the context of approximate sampling, it is given by the mixing time. Under a condition on the drift of the stochastic process underlying a JQN, which we call *hyper-stability*, in our main result we show that the coupling time is

polynomial in both the number of queues and buffer sizes. Then, we use this result to show that the mixing time of JQNs behaves similarly up to a given precision threshold. Our proof relies on a recursive formula relating the coupling times of trajectories that start from network states having 'distance one', and it can be used to analyze the coupling and mixing times of other Markovian networks, provided that they are monotone. An illustrative example is shown in the context of JQNs with blocking mechanisms.

In [35], we extend the technique to handle situations with infinite space state. We consider open JQN with losses with mixed finite and infinite queues and analyze the efficiency of sampling from their exact stationary distribution. Although the underlying Markov chain may have an infinite state space, we show that perfect sampling is possible. The main idea is to use a JQN with infinite buffers (that has a product form stationary distribution) to bound the number of initial conditions to be considered in the coupling from the past scheme. We also provide bounds on the sampling time of this new perfect sampling algorithm for acyclic or hyperstable networks. These bounds show that the new algorithm is considerably more efficient than existing perfect samplers even in the case where all queues are finite. We illustrate this efficiency through numerical experiments. We also extend our approach to non-monotone networks such as queueing networks with negative customers.

6.2. Interactive Analysis and Visualization of Large Distributed Systems

6.2.1. Interactive Visualization

High performance applications are composed of many processes that are executed in large-scale systems with possibly millions of computing units. A possible way to conduct a performance analysis of such applications is to register in trace files the behavior of all processes belonging to the same application. The large number of processes and the very detailed behavior that we can record about them lead to a trace size explosion both in space and time dimensions. The performance visualization of such data is very challenging because of the quantities involved and the limited screen space available to draw them all. 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.

In [33], we detail data aggregation techniques that are fully configurable by the user to control the level of details in both space and time dimensions. We also present two visualization techniques that take advantage of the aggregated data to scale. These features are part of the Viva and Triva open-source tools and framework.

The performance of parallel and distributed applications is also highly dependent on the characteristics of the execution environment. In such environments, the network topology and characteristics directly impact data locality and movements as well as contention, which are key phenomena to understand the behavior of such applications and possibly improve it. Unfortunately few visualizations available to the analyst are capable of accounting for such phenomena. In [26], we propose 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. We claim that such kind of visualization enables to explore and understand non trivial behavior that are impossible to grasp with classical visualization techniques. We also claim that the combination of multi-scale aggregation and dynamic graph layout allows our visualization technique to scale seamlessly to large distributed systems. We support these claims through a detailed analysis of a high performance computing scenario and of a grid computing scenario.

6.2.2. Entropy Based Analysis

Although the previous approaches already improve upon state of the art and are useful on current scenarios, it is clear that at very large scale they would probably not be as effective, which led us to change perspective and to investigate how entropy can help building tractable macroscopic descriptions. Indeed, data aggregation can provide such abstractions by partitioning the systems dimensions into aggregated pieces of information. This process leads to information losses, so the partitions should be chosen with the greatest caution, but in an acceptable computational time. While the number of possible partitions grows exponentially with the size of the system, we propose in [25] an algorithm that exploits exogenous constraints regarding the system semantics to find best partitions in a linear or polynomial time. We detail two constrained sets of partitions that

are respectively applied to temporal and spatial aggregation of an agent-based model of international relations. The algorithm succeeds in providing meaningful high-level abstractions for the system analysis.

Our approach is able to evaluate geographical abstractions used by the domain experts in order to provide efficient and meaningful macroscopic descriptions of the world global state [23]. We also successfully applied this technique to identify international media events by spatially and temporally aggregating RSS Flows of Newspapers [22], in particular with the case of the Syrian civil war between May 2011 and December 2012 [31], [21].

We also applied this technique to the analysis of large distributed systems and combined it with the treemap visualization technique [40], [14]. These features have been integrated in the Viva and Triva open-source tools and framework.

6.3. Trace Management and Analysis

6.3.1. Embedded Systems

The growing complexity of embedded system hardware and software makes their behavior analysis a challenging task. In this context, tracing provides relevant information about the system execution and appears to be a promising solution. However, trace management and analysis are hindered by several issues like the diversity of trace formats, the incompatibility of trace analysis methods, the problem of trace size and its storage as well as by the lack of visualization scalability. In [42], [27], [41], we present FrameSoC, a new trace management infrastructure that solves all the above issues together. It provides generic solutions for trace storage and defines interfaces and plugin mechanisms for integrating diverse analysis tools. We illustrate the benefit of FrameSoC with a case study of a visualization module that enables representation scalability of large traces by using an aggregation algorithm. Temporal aggregation techniques based on entropy are also currently integrated to the FrameSoC framework.

6.3.2. Jobs Resource Utilization

In HPC community the System Utilization metric enables to determine if the resources of the cluster are efficiently used by the batch scheduler. This metric considers that all the allocated resources (memory, disk, processors, etc) are full-time utilized. To optimize the system performance, we have to consider the effective physical consumption by jobs regarding the resource allocations. This information gives an insight into whether the cluster resources are efficiently used by the jobs. In [20], [30], we propose an analysis of production clusters based on the jobs resource utilization. The principle is to collect simultaneously traces from the job scheduler (provided by logs) and jobs resource consumption. The latter has been realized by developing a job monitoring tool, whose impact on the system has been measured as lightweight (0.35% speed-down). The key point is to statistically analyze both traces to detect and explain underutilization of the resources. This could enable to detect abnormal behavior, bottlenecks in the cluster leading to a poor scalability, and justifying optimizations such as gang scheduling or best effort scheduling. This method has been applied to two medium sized production clusters on a period of eight months.

6.4. Reconstructing the Software Environment of an Experiment

In the scientific experimentation process, an experiment result needs to be analyzed and compared with several others, potentially obtained in different conditions. Thus, the experimenter needs to be able to redo the experiment. 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. The task can quickly become complex for experimenters, particularly on research platforms dedicated to scientific experimentation, where both hardware and software are in constant rapid evolution. In [19] we discuss the concept of the reconstructability of software environments and propose a tool, Kameleon, for dealing with this problem.

6.5. Performance Evaluation

6.5.1. Computing the Throughput of Probabilistic and Replicated Streaming Applications

In [8], we investigate how to compute the throughput of probabilistic and replicated streaming applications. We are given (i) a streaming application whose dependence graph is a linear chain; (ii) a one-to-many mapping of the application onto a fully heterogeneous target platform, where a processor is assigned at most one application stage, but where a stage can be replicated onto a set of processors; and (iii) a set of random variables modeling the computation and communication times in the mapping. We show how to compute the throughput of the application, i.e., the rate at which data sets can be processed, under two execution models, the Strict model where the actions of each processor are sequentialized, and the Overlap model where a processor can compute and communicate in parallel. The problem is easy when application stages are not replicated, i.e., assigned to a single processor: in that case the throughput is dictated by the critical hardware resource. However, when stages are replicated, i.e., assigned to several processors, the problem becomes surprisingly complicated: even in the deterministic case, the optimal throughput may be lower than the smallest internal resource throughput. The first contribution of the paper is to provide a general method to compute the throughput when mapping parameters are constant or follow I.I.D. exponential laws. The second contribution is to provide bounds for the throughput when stage parameters (computation and communication times) form associated random sequences, and are N.B.U.E. (New Better than Used in Expectation) variables: the throughput is bounded from below by the exponential case and bounded from above by the deterministic case. An extensive set of simulation allows us to assess the quality of the model, and to observe the actual behavior of several distributions.

6.5.2. Optimization of Cloud Task Processing with Checkpoint-Restart Mechanism

In [17], we explain how to optimize fault-tolerance techniques based on a checkpointing/restart mechanism, in the context of cloud computing. Our contribution is three-fold. (1) We derive a fresh formula to compute the optimal number of checkpoints for cloud jobs with varied distributions of failure events. Our analysis is not only generic with no assumption on failure probability distribution, but also attractively simple to apply in practice. (2) We design an adaptive algorithm to optimize the impact of checkpointing regarding various costs like checkpointing/restart overhead. (3) We evaluate our optimized solution in a real cluster environment with hundreds of virtual machines and Berkeley Lab Checkpoint/Restart tool. Task failure events are emulated via a production trace produced on a large-scale Google data center. Experiments confirm that our solution is fairly suitable for Google systems. Our optimized formula outperforms Young's formula by 3-10 percent, reducing wallclock lengths by 50-100 seconds per job on average.

6.6. Game Theory and Applications

6.6.1. Fair Scheduling in Large Distributed Computing Sytems

Fairly sharing resources of a distributed computing system between users is a critical issue that we have investigated in two ways.

Our first proposal specifically addresses the question of designing a distributed sharing mechanism. A possible answer resorts to Lagrangian optimization and distributed gradient descent. Under certain conditions, the resource sharing problem can be formulated as a global optimization problem, which can be solved by a distributed self-stabilizing demand and response algorithm. In the last decade, this technique has been applied to design network protocols (variants of TCP, multi-path network protocols, wireless network protocols) and even distributed algorithms for smart grids. In [9], we explain how to use this technique for scheduling Bag-of-Tasks (BoT) applications on a Grid since until now, only simple mechanisms have been used to ensure a fair sharing of resources amongst these applications. Although the resulting algorithm is in essence very similar to previously proposed algorithms in the context of flow control in multi-path networks, we show using carefully designed experiments and a thorough statistical analysis that the grid context is surprisingly more difficult than the multi-path network context. Interestingly, we can show that, in practice, the convergence of the algorithm is hindered by the heterogeneity of application characteristics, which is completely overlooked

in related theoretical work. Our careful investigation provides enough insights to understand the true difficulty of this approach and to propose a set of non-trivial adaptations that enable convergence in the grid context. The effectiveness of our proposal is proven through an extensive set of complex and realistic simulations.

Our second proposal is centralized but more fine grain as it does drop the steady-state hypothesis and considers sequences of campaigns. Campaign Scheduling is characterized by multiple job submissions issued from multiple users over time. The work in [18] presents a new fair scheduling algorithm called OStrich whose principle is to maintain a virtual time-sharing schedule in which the same amount of processors is assigned to each user. The completion times in the virtual schedule determine the execution order on the physical processors. Then, campaigns are interleaved in a fair way by OStrich. For independent sequential jobs, we show that OStrich guarantees the stretch of a campaign to be proportional to campaign's size and the total number of users. The theoretical performance of our solution is assessed by simulating OStrich compared to the classical FCFS algorithm, issued from synthetic workload traces generated by two different user profiles. This is done to demonstrate how OStrich benefits both types of users, in contrast to FCFS.

6.6.2. Fundamentals of Continuous Games

We have made the following contributions:

1. Continuous-time game dynamics are typically first order systems where payoffs determine the growth rate of the players' strategy shares. In [12], we investigate what happens beyond first order by viewing payoffs as higher order forces of change, specifying e.g., the acceleration of the players' evolution instead of its velocity (a viewpoint which emerges naturally when it comes to aggregating empirical data of past instances of play). To that end, we derive a wide class of higher order game dynamics, generalizing first order imitative dynamics, and, in particular, the replicator dynamics. We show that strictly dominated strategies become extinct in n-th order payoff-monotonic dynamics n orders as fast as in the corresponding first order dynamics; furthermore, in stark contrast to first order, weakly dominated strategies also become extinct for $n \geq 2$. All in all, higher order payoff-monotonic dynamics lead to the elimination of weakly dominated strategies, followed by the iterated deletion of strictly dominated strategies, thus providing a dynamic justification of the well-known epistemic rationalizability process of Dekel and Fudenberg. Finally, we also establish a higher order analogue of the folk theorem of evolutionary game theory, and we show that convergence to strict equilibria in n-th order dynamics is n orders as fast as in first order.
2. In [37] we introduce a new class of game dynamics made of a pay-off replicator-like term modulated by an entropy barrier which keeps players away from the boundary of the strategy space. We show that these *entropy-driven* dynamics are equivalent to players computing a score as their on-going exponentially discounted cumulative payoff and then using a quantal choice model on the scores to pick an action. This dual perspective on *entropy-driven* dynamics helps us to extend the folk theorem on convergence to quantal response equilibria to this case, for potential games. It also provides the main ingredients to design a discrete time effective learning algorithm that is fully distributed and only requires partial information to converge to QRE. This convergence is resilient to stochastic perturbations and observation errors and does not require any synchronization between the players.

6.6.3. Application to Wireless Networks

We have made the following contributions:

1. Starting from an entropy-driven reinforcement learning scheme for multi-agent environments, we develop in [36] a distributed algorithm for robust spectrum management in Gaussian multiple-input, multiple-output (MIMO) uplink channels. In continuous time, our approach to optimizing the transmitters' signal distribution relies on the method of matrix exponential learning, adjusted by an entropy-driven barrier term which generates a distributed, convergent algorithm in discrete time. As opposed to traditional water-filling methods, the algorithm's convergence speed can be controlled by tuning the users' learning rate; accordingly, entropy-driven learning algorithms in MIMO systems converge arbitrarily close to the optimum signal covariance profile within a few iterations (even for large numbers of users and/or antennas per user), and this convergence remains robust even in the

presence of imperfect (or delayed) measurements and asynchronous user updates.

2. Consider a wireless network of transmitter-receiver pairs where the transmitters adjust their powers to maintain a target SINR level in the presence of interference. In [46], we analyze the optimal power vector that achieves this target in large, random networks obtained by "erasing" a finite fraction of nodes from a regular lattice of transmitter-receiver pairs. We show that this problem is equivalent to the so-called Anderson model of electron motion in dirty metals which has been used extensively in the analysis of diffusion in random environments. A standard approximation to this model is the so-called coherent potential approximation (CPA) method which we apply to evaluate the first and second order intra-sample statistics of the optimal power vector in one- and two-dimensional systems. This approach is equivalent to traditional techniques from random matrix theory and free probability, but while generally accurate (and in agreement with numerical simulations), it fails to fully describe the system: in particular, results obtained in this way fail to predict when power control becomes infeasible. In this regard, we find that the infinite system is always unstable beyond a certain value of the target SINR, but any finite system only has a small probability of becoming unstable. This instability probability is proportional to the tails of the eigenvalue distribution of the system which are calculated to exponential accuracy using methodologies developed within the Anderson model and its ties with random walks in random media. Finally, using these techniques, we also calculate the tails of the system's power distribution under power control and the rate of convergence of the Foschini-Miljanic power control algorithm in the presence of random erasures.

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 Nancy-Grand Est 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 Roca. 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) that 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.

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.

The Digitalis and IDPot clusters and the Bull Machine are integrated in the CIMENT Grid. More precisely, their unused resources may be exploited to execute jobs from partners of the CIMENT project. Mescal is also involved in CIMENT through the development or OAR and CiGri.

8.2. National Initiatives

8.2.1. Inria Large Scale Initiative

- *HEMERA, 2010-2012* Leading action "Completing challenging experiments on Grid'5000 (Methodology)" (see https://www.grid5000.fr/Hemera).

  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 controlled 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.

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*. Partners: Inria Grenoble (MESCAL, MOAIS), Inria Lyon (GRAAL), Inria Saclay (GRAND-LARGE).

  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.
- *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 exploit 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 led 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.

- *ANR MARMOTE, 2013-2016*. Partners: Inria Sophia (MAESTRO), Inria Rocquencourt (DIOGEN), PRiSM laboratory from University of Versailles-Saint-Quentin, Telecom SudParis (SAMOVAR), University Paris-Est Créteil (*Spécification et vérification de systèmes*), Université Pierre-et-Marie-Curie/LIP6.

  The project aims at realizing a software prototype dedicated to Markov chain modeling. It gathers seven teams that will develop advanced resolution algorithms and apply them to various domains (reliability, distributed systems, biology, physics, economy).

- *ANR NETLEARN, 2013-2015*. Partners: PRiSM laboratory from University of Versailles-Saint-Quentin, Telecom ParisTech, Orange Labs, LAMSADE/University Paris Dauphine, Alcatel-Lucent, Inria (MESCAL).

  The main objective of the project is to propose a novel approach of distributed, scalable, dynamic and energy efficient algorithms for managing resources in a mobile network. This new approach relies on the design of an orchestration mechanism of a portfolio of algorithms. The ultimate goal of the proposed mechanism is to enhance the user experience, while at the same time to better utilize the operator resources. User mobility and new services are key elements to take into account if the operator wants to improve the user quality of experience. Future autonomous network management and control algorithms will thus have to deal with a real-time dynamicity due to user mobility and to traffic variations resulting from various usages. To achieve this goal, we focus on two central aspects of mobile networks (the management of radio resources at the Radio Access Network level and the management of the popular contents users want to get access to) and intend to design distributed learning mechanisms in non-stationary environments, as well as an orchestration mechanism that applies the best algorithms depending on the situation.

8.2.4. National Organizations

Jean-Marc Vincent is member of the scientific committees of the CIST (Centre International des Sciences du Territoire).

8.3. European Initiatives

8.3.1. FP7 Projects

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

Type: FP7 Programme

Objectif: ICT-2011.9.13 Exa-scale computing, software and simulation

Duration: October 2011 - October 2014

Coordinator: Alex Ramirez

Partner: BSC (Barcelone), Bull, ARM (UK), Julich (Germany), Genci, CINECA (Italy), CNRS (LIRMM, LIG)

Inria contact: Arnaud Legrand

Abstract: There is a continued need for higher computing 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 1.7 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.

8.3.1.2. Network of Excellence in Wireless COMmunications

Type: FP7 Programme

Objectif: 1.1 Future Networks

Duration: November 2012 - October 2015

Coordinator: Marco Louise

Partner: CNIT (IT), Aalborg University (DK), Bilkent University (TK), CNRS (FR), CTTC (ES), IASA (GR), INOV (P), Poznan University of Technology (PL), Technion (IL), Technische Universitaet Dresden (D), University of Cambridge (UK), Université de Louvain (BE), OulunYliopisto (FIN), Technische Universitaet Wien (A).

Inria contact: Panayotis Mertikopoulos

Abstract: The NEWCOM researchers will pursue long-term, interdisciplinary research on the most advanced aspects of wireless communications like Finding the Ultimate Limits of Communication Networks, Opportunistic and Cooperative Communications, Energy- and Bandwidth-Efficient Communications and Networking.

8.3.2. Collaborations in European Programs, except FP7

8.3.2.1. ESPON

Program: ESPON

Project acronym: HyperATLAS

Duration: 2007-2013

Coordinator: European Community

Abstract: 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.

8.3.2.2. CROWN

Program: European Community and Greek General Secretariat for Research and Technology

Project acronym: CROWN

Project title: Optimal Control of Self Organized Wireless Networks

Duration: 2012-2015

Coordinator: Tassiulas Leandros

Other partners: Thales, University of Thessaly, National and Kapodistrian University of Athens, Athens University of Economics and Business

Abstract: Wireless networks are rapidly becoming highly complex systems with large numbers of heterogeneous devices interacting with each other, often in a harsh environment. In the absence of central control, network entities need to self-organize to reach an efficient operating state, while operating in a distributed fashion. Depending on whether the operating criteria are individual or global, nodes interact in an autonomic or coordinated way. Despite recent progress in autonomic networks, the fundamental understanding of the operational behaviour of large-scale networks is still lacking. This project will address these emergent network properties, by introducing new tools and concepts from other disciplines.

We will first analyze how imperfect network state information can be harvested and distributed efficiently through the network using machine learning techniques. We will design flexible methodologies to shape the competition between autonomous nodes for resources, with aim to maintain robust social optimality. Both cooperating and non-cooperating game-theoretic models will be used. We also consider networks with nodes coordinating to achieve a joint task, e.g., global optimization. Using algorithms inspired from statistical physics, we will address two representative paradigms in the context of wireless ad hoc networks, namely connectivity optimization and the localization of a network of primary sources from a sensor network.

Finally, we will explore delay tolerant networks as a case study of an emerging class of networks that, while sharing most of the characteristics of traditional autonomic or coordinated networks, they present unique challenges, due to the intermittency and constant fluctuations of the connectivity. We will study tradeoffs involving delay, the impact of mobility on information transfer, and the optimal usage of resources by using tools from information theory and stochastic evolution theory.

8.3.3. Collaborations with Major European Organizations

University of Athens: Panayotis Mertikopoulos was an invited professor for 4 months.

EPFL: Laboratoire pour les communications informatiques et leurs applications 2, Institut de systèmes de communication ISC, Ecole polytechnique fédérale de Lausanne (Switzerland). We collaborate with Jean-Yves Leboudec and Nicolas Gast on fluid limits.

BCAM: Basque Center for Applied Mathematics, Bilbao (Spain). Bruno gaujal was invited to teach several time and collaborates with Jonatha Anselmi on perfect simulation.

TU Wien: Research Group Parallel Computing, Technische Universität Wien (Austria). We collaborate with Sascha Hunold on experimental methodology and reproducibility of experiments in HPC.

8.4. International Initiatives

8.4.1. Inria Associate Teams

8.4.1.1. CLOUDSHARE

Title: Guaranteed Application Performance on Idle Data Center Resources

Inria principal investigator: Arnaud Legrand

International Partner (Institution - Laboratory - Researcher):

Walfredo Cirne (Google Inc. (United States))

David P. Anderson (University of California Berkeley - Space Sciences Laboratory)

Duration: 2009 - 2014

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

Data centers are often 85% idle as they must over-provision to ensure service level agreements. At the same time, high data center utilization is essential for efficient resource usage and optimal revenue. One way to improve utilization is for low-priority applications to use the idle resources of data centers, allowing high-priority applications to preempt them at any time. While users benefit from the lower costs of using these idle resources, parallel applications such as Map-Reduce can suffer severe overheads and unpredictable performance due to unexpected preemption and unavailability. The goal of this project is to enable complex applications to utilize idle data center resources with guaranteed performance. Our approach will be as follows. First, we will investigate novel statistical methods to predict the execution time of complex batch applications. Second, we will apply machine learning methods to predict idleness in data centers. Third, we will craft fair scheduling algorithms for multiple applications that compete for idle data center resources. The collaboration bridges experts in statistical modeling and simulation from the Inria MESCAL team with system and scheduling experts in the Berkeley BOINC team and the Google Infrastructure team.

8.4.2. Inria International Partners

8.4.2.1. Declared Inria International Partners

- MESCAL has strong connections with both UFRGS (Porto Alegre, Brazil) and USP (Sao Paulo, Brazil). The creation of the LICIA common laboratory (see next section) has made 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. Arnaud Legrand visited Berkeley and the Inria Grenoble hosted the yearly BOINC workshop in 2013.

8.4.3. Inria International Labs

8.4.3.1. North America

- 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 done several visits to Urbana-Champaign or NCSA. One Mescal Postdoc (Slim Bougherra) spent one year in Urbana-Champaign.
- Associated Team with Berkeley. MESCAL is thus involved in the Inria@SiliconValley program.

8.4.4. Participation In other International Programs

8.4.4.1. 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, both from the MESCAL team.

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

  More information can be found at http://www.ufrgs.br/sisinfo/?ai1ec_event=terceira-reuniao-do-licia&instance_id=.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- Wenjing Wu (Chinese Academy of Science) visited MESCAL for two weeks in September.
- Sergio Gelvez Cortes (Universidad Industrial de Santander Bucaramanga, Colombia) visited MESCAL for two months.

8.5.1.1. Internships

- Wagner Kolberg (MSc UFRGS) made a 4 months internship in MESCAL.

8.5.2. Visits to International Teams

- Panayotis Mertikopoulos was invited to work for 3 weeks at Universidade de Chile (14/01 -> 2/02)
- Panayotis Mertikopoulos was invited to work for 4 months at University of Athens (01/03 -> 30/06)
- Jean-Marc Vincent was invited to work for 3 weeks at UFRGS and PUC-RS, Porto Alegre

9. Dissemination

9.1. Scientific Animation

- 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).
- Arnaud Legrand is mandated by the LIG for representing the Networking and Parallel and Distributed System teams of the LIG.
- Panayotis Mertikopoulos is mandated by the LIG to supervise PhD students of the laboratory.

9.1.1. Invited Talks

- Bruno Gaujal was a keynote speaker at ICPP'13 and an invited speaker at the Dagstuhl Seminar on Exascale computing.
- Panayotis Mertikopoulos was an invited speaker to
  - the 30 years congress of the Société de Mathématiques Appliquées et Industrielles
  - the Evolutionnary Dynamics and Market Behavior workshop, Haussdorf Research
  - the Institute for Mathematics , Bonn, Germany
  - Erice 2013 (Stochastich Methods in Game Theory), Sicile, Italy
  - ADGO 2013 (Algorithms and Dynamics for Games and Optimization), Playa Blanca, Chile
- Arnaud Legrand has given an invited talk at TU Wien (April), at the JLPC at NCSA (November). He was a keynote speaker at ERADS (Porto Alegre, Brazil) in March 2013 and at SimuTools (Nice, Cannes) in March 2013.

9.1.2. Journal, Conference and Workshop Organization

- Arnaud Legrand has organized the *SimGrid user days* in Lyon (June 2013).
- Arnaud Legrand has organized the *BOINC workshop* in Grenoble (September 2013).
- Corinne Touati and Panayotis Mertikopoulos have organized the *Algo-GT workshop* in Grenoble (July 2013).

9.1.3. Program Committees

- Panayotis Mertikopoulos has been Publication chair of WiOpt'13, TPC member of ValueTools'13. He is a regular reviewer for ISIT'13, ITW'13, Games on Economic Behavior, Journal of Economic Theory, Advances on Applied Probability, IEEE Trans. on Information Theory, IEEE Trans. on Wireless Communications, IEEE Trans, on Signal Processing.
- Bruno Gaujal has been a TPC member of Wodes, IPDPS and SigMetrics.
- Olivier Richard has initiated a reproducible research track in Compas entitled Realis. He is a regular reviewer for Parallel Computing, TSI and CLCAR.
- Jean-Marc Vincent has been in the steering committee of AMSTA and a TPC member of IPDPS'13, SimuTools'13, ValueTools'13 and SimulTech'13.
- Arnaud Legrand has been a TPC member of IPDPS'13, ICPP'13, PPAM and is a regular reviewer for IJHPCA, SIMPAT and TPDS.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Several members of MESCAL are university professors and comply with their recurrent teaching duties. We only list here lectures (i.e., not tutorials or practical sessions) at the master level or above.

Master : Florence Perronnin, Probabilities for computer science, 23h Eq. TD, M1, ENSIMAG

Master : Florence Perronnin, Performance Evaluation, 23h Eq. TD, M1, Polytech

Master : Arnaud Legrand and Jean-Marc Vincent, Performance Evaluation, 32h Eq. TD , M2, UJF

Master : Arnaud Legrand, Parallel Systems, 47h Eq. TD, M2, UJF

Master : Bruno Gaujal, Discrete Events, 18 Eq. TD, M2, MPRI/ Paris Diderot.

Master : Bruno Gaujal, Mean field Approximation, 20 Eq. TD, M2, BCAM (Bilbao)

Doctorat : Bruno Gaujal, Mean field Approximation 30 Eq. TD, Toulouse PhD Program.

Master : Panayotis Mertikopoulos, Game theory for the working economist, 55 Eq. TD, M2, University of Athens.

Master : Olivier Richard, Networking, 33 Eq. TD, M1, Polytech

Master : Jean-Marc Vincent, Probabilities and Simulation, 23h Eq. TD, M1, Polytech

- Olivier Richard is also responsible of the organization of the RICM4 (M1 Polytech) and of multi-disciplinary projects.

9.2.2. Supervision

PhD : Joseph Emeras, *Workload Traces Analysis and Replay in Large Scale Distributed Systems*, Université de Grenoble, 01 octobre 2013 [5].

PhD : Robin Lamarche Perrin, *Building Meaningful Macroscopic Descriptions of Large-scale Complex Systems*, 14 octobre 2013 [6].

9.2.3. Juries

- Bruno Gaujal has been president of the junior researcher selection committee in Inria Bordeaux.
- Bruno Gaujal has been member of the professor selection committee in University of Avignon.
- Bruno Gaujal has been member of the PhD thesis committee of Alexandre Salch (G-SCOP laboratory, Grenoble).
- Arnaud Legrand has been member of the PhD thesis committee of Sorina Camarasu Pop (Creatis laboratory, Lyon).
- Arnaud Legrand has been a reviewer of the PhD thesis of Javier Celaya Alastrué (University of Zaragoza, Spain).

9.3. Popularization

9.3.1. Popular Science

- MESCAL actively promotes science to young and non-scientific audience. This year Corinne Touati participated the *Fête de la science* and animated a workshop on game theory. theory [48].
- Jean-Marc Vincent contributed to the national initiative for introducing computer science to high school professors in mathematics. He was responsible of the high school professors training, of the corresponding university diploma and of the online training provided by the rectorat. Jean-Marc also participated to the steering committee of CERVIN (http://flet.fr/cervin/) for Inria.
- Olivier Richard is involved in training *Classe Préparatoires* professors to Python. He is also responsible of the Polytech fablab, which he presented at the Toulouse Hacker Space Factory.

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] J. EMERAS. , *Analyse et rejeu de traces de charge dans les grands systèmes de calcul distribués*, Université de Grenoble, October 2013, http://hal.inria.fr/tel-00940055

[5] J. EMERAS. , *Workload Traces Analysis and Replay in Large Scale Distributed Systems*, Université de Grenoble, 10 2013

[6] R. LAMARCHE-PERRIN. , *Building Meaningful Macroscopic Descriptions of Large-scale Complex Systems*, Université de Grenoble, 10 2013

Articles in International Peer-Reviewed Journals

[7] J. ANSELMI, B. GAUJAL. *Efficiency of simulation in monotone hyper-stable queueing networks*, in "Queueing Systems", March 2013, http://hal.inria.fr/hal-00801437

[8] A. BENOIT, M. GALLET, B. GAUJAL, Y. ROBERT. *Computing the throughput of probabilistic and replicated streaming applications*, in "Algorithmica", March 2013, http://hal.inria.fr/hal-00800083

[9] R. BERTIN, S. HUNOLD, A. LEGRAND, C. TOUATI. *Fair scheduling of bag-of-tasks applications using distributed Lagrangian optimization*, in "Journal of Parallel and Distributed Computing", August 2013 [*DOI :* 10.1016/J.JPDC.2013.08.011], http://hal.inria.fr/hal-00872473

[10] S. DI, D. KONDO, W. CIRNE. *Google hostload prediction based on Bayesian model with optimized feature combination*, in "J. Parallel Distrib. Comput.", 2014, vol. 74, n° 1, pp. 1820-1832 [*DOI :* 10.1016/J.JPDC.2013.10.001], http://hal.inria.fr/hal-00936829

[11] B. JAVADI, D. KONDO, A. IOSUP, D. EPEMA. *The Failure Trace Archive: Enabling the comparison of failure measurements and models of distributed systems*, in "Journal of Parallel and Distributed Computing", 2013, vol. 73, n° 8, pp. 1208 - 1223 [*DOI :* 10.1016/J.JPDC.2013.04.002], http://hal.inria.fr/hal-00925098

[12] R. LARAKI, P. MERTIKOPOULOS. *Higher Order Game Dynamics*, in "Journal of Economic Theory", 2013, vol. 148, n° 6, pp. 2666-2695 [*DOI :* 10.1016/J.JET.2013.08.002], http://hal.inria.fr/hal-00911531

[13] P. VELHO, L. SCHNORR, H. CASANOVA, A. LEGRAND. *On the Validity of Flow-level TCP Network Models for Grid and Cloud Simulations*, in "ACM Transactions on Modeling and Computer Simulation", October 2013, vol. 23, n° 3, http://hal.inria.fr/hal-00872476

Articles in National Peer-Reviewed Journals

[14] R. LAMARCHE-PERRIN, L. M. SCHNORR, J.-M. VINCENT, Y. DEMAZEAU. *Agrégation de traces pour la visualisation de grands systèmes distribués*, in "Technique et Science Informatiques (TSI)", 2013, http://hal.inria.fr/hal-00918432

International Conferences with Proceedings

[15] P. BEDARIDE, A. DEGOMME, S. GENAUD, A. LEGRAND, G. MARKOMANOLIS, M. QUINSON, M. STILLWELL, F. SUTER, B. VIDEAU. *Toward Better Simulation of MPI Applications on Ethernet/TCP Networks*, in "PMBS13 - 4th International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems", Denver, United States, November 2013, http://hal.inria.fr/hal-00919507

[16] S. DI, D. KONDO, F. CAPPELLO. *Characterizing Cloud Applications on a Google Data Center*, in "42nd International Conference on Parallel Processing (ICPP'13)", 2013, pp. 468-473 [*DOI :* 10.1109/ICPP.2013.56], http://hal.inria.fr/hal-00936827

[17] S. DI, Y. ROBERT, F. VIVIEN, D. KONDO, C.-L. WANG, F. CAPPELLO. *Optimization of Cloud Task Processing with Checkpoint-Restart Mechanism*, in "SC13 - Supercomputing - 2013", Denver, United States, ACM, November 2013 [*DOI :* 10.1145/2503210.2503217], http://hal.inria.fr/hal-00847635

[18] J. EMERAS, V. PINHEIRO, K. RZADCA, D. TRYSTRAM. *OStrich: Fair Scheduling for Multiple Submissions*, in "PPAM'2013", Warsaw, Poland, Springer, 2013, http://hal.inria.fr/hal-00918374

[19] J. EMERAS, O. RICHARD, B. BZEZNIK, G. YIANNIS, C. RUIZ. *Reconstructing the Software Environment of an Experiment with Kameleon*, in "Proc. of ACM Compute 2012", Pune, India, ACM, 2013 [*DOI :* 10.1145/2459118.2459134], http://hal.inria.fr/hal-00919836

[20] J. EMERAS, C. RUIZ, J.-M. VINCENT, O. RICHARD. *Analysis of the Jobs Resource Utilization on a Production System*, in "Job Scheduling Strategies for Parallel Processing", Boston, United States, W. CIRNE,

N. DESAI, E. FRACHTENBERG, U. SCHWIEGELSHOHN (editors), Lecture Notes in Computer Science, Springer, 2013, http://hal.inria.fr/hal-00918372

[21] T. GIRAUD, C. GRASLAND, R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Identification of International Media Events by Spatial and Temporal Aggregation of RSS Flows of Newspapers: Application to the Case of the Syrian Civil War between May 2011 and December 2012*, in "18th European Colloquium on Theoretical and Quantitative Geography (ECTQG'13)", Dourdan, France, 2013, pp. 112–114, http://hal.inria.fr/hal-00918434

[22] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Analysis of International Relations through Spatial and Temporal Aggregation*, in "11th International Conference on Practical Applications of Agents and Multi-Agent Systems, PAAMS'13", Springer Verlag, Salamanca, Unknown, 2013, vol. LNAI, pp. 296-299, http://hal.inria.fr/hal-00947931

[23] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *How to Build the Best Macroscopic Description of your Multi-agent System?*, in "11th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS'13)", Salamanca, Spain, 2013, pp. 157-169, http://hal.inria.fr/hal-00918435

[24] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *Multi-resolution Representations of Media Information*, in "23rd International Joint Conference on Artificial Intelligence", Beijing, China, 2013, http://hal.inria.fr/hal-00918437

[25] R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT. *The Best-partitions Problem: How to Build Meaningful Aggregations*, in "IEEE/WIC/ACM International Conference on Intelligent Agent Technology (Iat'13)", Atlanta, United States, 2013, http://hal.inria.fr/hal-00918433

[26] L. MELLO SCHNORR, A. LEGRAND, J.-M. VINCENT. *Interactive Analysis of Large Distributed Systems with Scalable Topology-based Visualization*, in "International Symposium on Performance Analysis of Systems and Software (ISPASS'13)", Austin, Texas, United States, IEEE Computer Society Press, 2013, http://hal.inria.fr/hal-00789436

[27] G. PAGANO, D. DOSIMONT, G. HUARD, V. MARANGOZOVA-MARTIN, J.-M. VINCENT. *Trace Management and Analysis for Embedded Systems*, in "Ieee 7th International Symposium on Embedded Multicore/Many-core SoCs", Tokyo, Japan, 2013, http://hal.inria.fr/hal-00918439

[28] L. STANISIC, B. VIDEAU, J. CRONSIOE, A. DEGOMME, V. MARANGOZOVA-MARTIN, A. LEGRAND, J.-F. MEHAUT. *Performance Analysis of HPC Applications on Low-Power Embedded Platforms*, in "DATE - Design, Automation & Test in Europe", Grenoble, France, March 2013, pp. 475-480 [*DOI :* 10.7873/DATE.2013.106], http://hal.inria.fr/hal-00872482

National Conferences with Proceedings

[29] D. DOSIMONT, G. HUARD, J.-M. VINCENT. *La visualisation de traces, support à l'analyse, déverminage et optimisation d'applications de calcul haute performance*, in "Actes de l'atelier Visualisation d'informations, interaction et fouille de données (VIF) de la 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2013)", Toulouse, France, 2013, pp. 55–66, http://hal.inria.fr/hal-00918438

[30] J. EMERAS, C. RUIZ, J.-M. VINCENT, O. RICHARD. *Jobs Resource Utilization as a Metric for Clusters Comparison and Optimization*, in "ComPAS'2013 Proceedings", Grenoble, France, 2013, http://hal.inria.fr/hal-00916284

Conferences without Proceedings

[31] T. GIRAUD, C. GRASLAND, R. LAMARCHE-PERRIN, Y. DEMAZEAU, J.-M. VINCENT, T. THÉVENIN. *Identification of international media events by spatial and temporal aggregation of RSS flows of newspapers*, in "18th European Colloquium in Theoretical and Quantitative Geography (ECTQG)", Dourdan, France, September 2013, http://hal.inria.fr/hal-00881313

Scientific Books (or Scientific Book chapters)

[32] D. BALOUEK, A. CARPEN AMARIE, G. CHARRIER, F. DESPREZ, E. JEANNOT, E. JEANVOINE, A. LÈBRE, D. MARGERY, N. NICLAUSSE, L. NUSSBAUM, O. RICHARD, C. PÉREZ, F. QUESNEL, C. ROHR, L. SARZYNIEC. *Adding Virtualization Capabilities to the Grid'5000 Testbed*, in "Cloud Computing and Services Science", I. IVANOV, M. SINDEREN, F. LEYMANN, T. SHAN (editors), Communications in Computer and Information Science, Springer International Publishing, 2013, vol. 367, pp. 3-20 [*DOI :* 10.1007/978-3-319-04519-1_1], http://hal.inria.fr/hal-00946971

[33] L. MELLO SCHNORR, A. LEGRAND. *Visualizing More Performance Data Than What Fits on Your Screen*, in "Tools for High Performance Computing 2012", A. CHEPTSOV, S. BRINKMANN, J. GRACIA, M. M. RESCH, W. E. NAGEL (editors), Springer Berlin Heidelberg, 2013, pp. 149-162 [*DOI :* 10.1007/978-3-642-37349-7_10], http://hal.inria.fr/hal-00842761

Research Reports

[34] P. BEDARIDE, S. GENAUD, A. DEGOMME, A. LEGRAND, G. MARKOMANOLIS, M. QUINSON, M. STILLWELL, F. SUTER, B. VIDEAU. , *Improving Simulations of MPI Applications Using A Hybrid Network Model with Topology and Contention Support*, Inria, May 2013, n° RR-8300, 22 p. , http://hal.inria.fr/hal-00821446

[35] A. BUSIC, B. GAUJAL, F. PERRONNIN. , *Perfect sampling of Jackson Queueing Networks*, Inria, August 2013, n° RR-8332, 27 p. , http://hal.inria.fr/hal-00851331

[36] P. COUCHENEY, B. GAUJAL, P. MERTIKOPOULOS. , *Distributed Optimization in Multi-User MIMO Systems with Imperfect and Delayed Information*, Inria, December 2013, n° RR-8426, 19 p. , http://hal.inria.fr/hal-00918762

[37] P. COUCHENEY, B. GAUJAL, P. MERTIKOPOULOS. , *Entropy-driven dynamics and robust learning procedures in games*, Inria, February 2013, n° RR-8210, 33 p. , http://hal.inria.fr/hal-00790815

[38] S. DELAMARE, G. FEDAK, D. KONDO, O. LODYGENSKY, P. KACSUK, J. KOVACS, F. ARAUJO. , *Advanced QoS Prototype for the EDGI Infrastructure*, Inria, May 2013, n° RR-8295, http://hal.inria.fr/hal-00819907

[39] S. DELAMARE, G. FEDAK, D. KONDO, O. LODYGENSKY, P. KACSUK, J. KOVACS, F. ARAUJO. , *Intermediate QoS Prototype for the EDGI Infrastructure*, Inria, May 2013, n° RR-8294, http://hal.inria.fr/hal-00819903

[40] R. LAMARCHE-PERRIN, L. SCHNORR, Y. DEMAZEAU, J.-M. VINCENT. , *Evaluating Trace Aggregation Through Entropy Measures for Optimal Performance Visualization of Large Distributed Systems*, Inria, 2013, n^{o} RR-LIG-037, http://hal.inria.fr/hal-00872483

[41] G. PAGANO, D. DOSIMONT, G. HUARD, V. MARANGOZOVA-MARTIN, J.-M. VINCENT. , *Trace Management and Analysis for Embedded Systems*, Inria, May 2013, n^{o} RR-8304, 21 p. , http://hal.inria.fr/hal-00821907

[42] G. PAGANO, V. MARANGOZOVA-MARTIN. , *SoC-Trace Infrastructure Benchmark*, Inria, June 2013, n^{o} RT-0435, 25 p. , http://hal.inria.fr/hal-00830008

Other Publications

[43] H. CASANOVA, A. GIERSCH, A. LEGRAND, M. QUINSON, F. SUTER. , *SimGrid: a Sustained Effort for the Versatile Simulation of Large Scale Distributed Systems*, 2013, 4 pages, submission to WSSSPE'13, http://hal.inria.fr/hal-00926437

[44] J. KWON, P. MERTIKOPOULOS. , *A continuous-time approach to online optimization*, 2014, http://hal.inria.fr/hal-00937400

[45] R. LARAKI, P. MERTIKOPOULOS. , *Inertial game dynamics and applications to constrained optimization*, 2013, 31 p. , Submitted to the SIAM Journal on Control and Optimization, http://hal.inria.fr/hal-00920928

[46] A. L. MOUSTAKAS, P. MERTIKOPOULOS, N. BAMBOS. , *Power Optimization in Random Wireless Networks*, 2013, Submitted to IEEE Trans. Inform. Theory, http://hal.inria.fr/hal-00920648

[47] L. NUSSBAUM, P. NEYRON, O. RICHARD, E. JEANVOINE. *Grid'5000: A Production-grade Testbed for Experiment-driven Computer Science on HPC and Clouds*, in "Inria Booth at SC'13", Denver, United States, November 2013, Inria Booth at SC'13, http://hal.inria.fr/hal-00920389

[48] C. TOUATI. *Le jeu des Kirlis et des Gourlus*, in "Fête de la Science", Montbonnot, France, October 2013, Fête de la Science, http://hal.inria.fr/hal-00871607