IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Pierre Mendes-France (Grenoble)

Université Joseph Fourier (Grenoble)

Activity Report 2014

Project-Team MOAIS

PrograMming and scheduling design fOr Applications in Interactive Simulation

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 MOAIS

Keywords: Scheduling, Interactive Computing, Parallel And Distributed Algorithms, High Performance Computing, Fault Tolerance, Parallel Programming Model

Creation of the Project-Team: 2006 January 01.

1. Members

Research Scientists

Thierry Gautier [Inria, Researcher]
Bruno Raffin [Inria, Researcher, HdR]

Faculty Members

Jean-Louis Roch [Team leader, INP Grenoble, Associate Professor]
François Broquedis [INP Grenoble, Associate Professor]
Vincent Danjean [Univ. Grenoble I, Associate Professor]
Pierre-François Dutot [Univ. Grenoble II, Associate Professor]
Guillaume Huard [Univ. Grenoble I, Associate Professor]
Grégory Mounié [INP Grenoble, Associate Professor]
Clément Pernet [INP Grenoble, Associate Professor, visiting the ARIC Inria project during 2014, HdR]
Denis Trystram [INP Grenoble, Professor, HdR]
Frédéric Wagner [INP Grenoble, Associate Professor]

Engineers

Pierrick Brunet [Inria, from Oct 2013]
Youenn Corre [Inria, from May 2014]
Jérémy Jaussaud [Inria, until Oct 2014]
Xavier Martin [ENSIMAG apprentice at Inria, FP7 grant, until Sept 2014]
Pierre Neyron [CNRS]
Philippe Virouleau [Inria, from Oct 2013]

PhD Students

Vinícius Garcia Pinto [Grenoble INP and USP, from Oct 2014]
Alessandro Kraemer [Grenoble INP and UFPR, from Sep 2014]
David Beniamine [Univ. Grenoble I, from Sept 2015]
Raphaël Bleuse [Grenoble INP, DGA grant, from Sept 2013]
Damien Dosimont [Inria, from Feb 2012]
Matthieu Dreher [Inria, until Fev 2015]
Mathias Ettinger [Inria, until Oct 2014]
Joao Ferreira Lima [Univ. Grenoble and UFRGS, until May 2014]
David Glesser [Grenoble INP, Cifre grant from Bull, from Sept 2013]
Nicolas Kox [Univ. Grenoble I, from Oct 2014]
Stéfano Drimon Kurz Mor [Univ. Grenoble and UFRGS, until Mai 2015]
Fernando Machado Mendonca [Grenoble INP and USP]
Florence Monna [CNRS, until Oct 2014]
Vinicius Pinheiro [Grenoble INP and USP, until Feb 2014]
Millian Poquet [Univ. Grenoble I, from Oct 2014]
Maxime Puys [Univ. Grenoble I, from Oct 2014]
Adrien Roussel [Grenoble INP, IFPEN grant, from Oct 2014]
Marwa Sridi [CEA, from Apr 2013 until Mar 2016]
Abhinav Srivastav [Univ. Grenoble I, from Jan 2013 until Jan 2016]
Ziad Sultan [Univ. Grenoble I, from Oct 2012 until Sept 2015]

Julio Toss [Univ Grenoble I and UFRGS, from Nov 2014]

Post-Doctoral Fellow

Giorgio Lucarelli [INP Grenoble, ANR Grant, from Sep 2014]

Visiting Scientists

Pedro Luis Furio Raphael [Inria, from May 2014 until Jun 2014]
Adel Safi [Associate Professor ar ESSTT, until Jan 2014]
Andreï Tchernykh [Researcher at CICESE, from Apr 2014]
Guochuan Zhang [Professor at Zhejiang University, until Apr 2014]

Administrative Assistant

Annie Simon [Inria]

Others

Monica Liliana Hernandez Ariza [Inria, Master Student at University of Santander, Colombia, from Jan 2014 until l Apr 2014]
Christine Guimet [CNRS]

2. Overall Objectives

2.1. Introduction

The goal of the MOAIS team-project is to develop the scientific and technological foundations for parallel programming that enable to achieve provable performances on distributed parallel architectures, from multi-processor systems on chips to computational grids and global computing platforms. Beyond the optimization of the application itself, the effective use of a larger number of resources is expected to enhance the performance. This encompasses large scale scientific interactive simulations (such as immersive virtual reality) that involve various resources: input (sensors, cameras, ...), computing units (processors, memory), output (videoprojectors, images wall) that play a prominent role in the development of high performance parallel computing.

To reach this goal, MOAIS gathers experts in : algorithm design, scheduling, parallel programming (both low level and high level API), interactive applications. The research directions of the MOAIS team are focused on scheduling problems with a multi-criteria performance objective: precision, reactivity, resources consumption, reliability, ... The originality of the MOAIS approach is to use the application's adaptability to control its scheduling:

- the application describes synchronization conditions;
- the scheduler computes a schedule that verifies those conditions on the available resources;
- each resource behaves independently and performs the decision of the scheduler.

To enable the scheduler to drive the execution, the application is modeled by a macro data flow graph, a popular bridging model for parallel programming (BSP, Nesl, Earth, Jade, Cilk, Athapascan, Smarts, Satin, ...) and scheduling. A node represents the state transition of a given component; edges represent synchronizations between components. However, the application is malleable and this macro data flow is dynamic and recursive: depending on the available resources and/or the required precision, it may be unrolled to increase precision (e.g. zooming on parts of simulation) or enrolled to increase reactivity (e.g. respecting latency constraints). The decision of unrolling/enrolling is taken by the scheduler; the execution of this decision is performed by the application.

The MOAIS project-team is structured around four axis:

- **Scheduling**: To formalize and study the related scheduling problems, the critical points are: the modeling of an adaptive application; the formalization and the optimization of the multi-objective problems; the design of scalable scheduling algorithms. We are interested in classical combinatorial optimization methods (approximation algorithms, theoretical bounds and complexity analysis), and also in non-standard methods such as Game Theory.
- **Adaptive parallel and distributed algorithms**: To design and analyze algorithms that may adapt their execution under the control of the scheduling, the critical point is that algorithms are either parallel or distributed; then, adaptation should be performed locally while ensuring the coherency of results.
- **Programming interfaces and tools for coordination and execution**: To specify and implement interfaces that express coupling of components with various synchronization constraints, the critical point is to enable an efficient control of the coupling while ensuring coherency. We develop the **XKaapi** runtime software that manages the scheduling of multithreaded computations with billions of threads on a virtual architecture with an arbitrary number of resources; XKaapi supports node additions and resilience. XKaapi manages the *fine grain* scheduling of the computation part of the application. To enable parallel application execution and analysis. We develop runtime tools that support large scale and fault tolerant processes deployment (**TakTuk**), visualization of parallel executions on heterogeneous platforms (**Triva**).
- **Interactivity**: To improve interactivity, the critical point is scalability. The number of resources (including input and output devices) should be adapted without modification of the application. We develop the **FlowVR** middleware that enables to configure an application on a cluster with a fixed set of input and output resources. FlowVR manages the *coarse grain* scheduling of the whole application and the latency to produce outputs from the inputs.

Often, computing platforms have a dynamic behavior. The dataflow model of computation directly enables to take into account addition of resources. To deal with resilience, we develop softwares that provide **fault-tolerance** to dataflow computations. We distinguish non-malicious faults from malicious intrusions. Our approach is based on a checkpoint of the dataflow with bounded and amortized overhead.

3. Research Program

3.1. Scheduling

Participants: Pierre-François Dutot, Guillaume Huard, Grégory Mounié, Jean-Louis Roch, Denis Trystram, Frédéric Wagner.

The goal of this theme is to determine adequate multi-criteria objectives which are efficient (precision, reactivity, speed) and to study scheduling algorithms to reach these objectives.

In the context of parallel and distributed processing, the term *scheduling* is used with many acceptations. In general, scheduling means assigning tasks of a program (or processes) to the various components of a system (processors, communication links).

Researchers within MOAIS have been working on this subject for many years. They are known for their multiple contributions for determining the target dates and processors the tasks of a parallel program should be executed; especially regarding execution models (taking into account inter-task communications or any other system features) and the design of efficient algorithms (for which there exists a performance guarantee relative to the optimal scheduling).

Parallel tasks model and extensions. We have contributed to the definition and promotion of modern task models: parallel moldable tasks and divisible load. For both models, we have developed new techniques to derive efficient scheduling algorithms (with a good performance guaranty). We proposed recently some extensions taking into account machine unavailabilities (reservations).

Multi-objective Optimization. A natural question while designing practical scheduling algorithms is "which criterion should be optimized ?". Most existing works have been developed for minimizing the *makespan* (time of the latest tasks to be executed). This objective corresponds to a system administrator view who wants to be able to complete all the waiting jobs as soon as possible. The user, from his-her point of view, would be more interested in minimizing the average of the completion times (called *minsum*) of the whole set of submitted jobs. There exist several other objectives which may be pertinent for specific use. We worked on the problem of designing scheduling algorithms that optimize simultaneously several objectives with a theoretical guarantee on each objective. The main issue is that most of the policies are good for one criterion but bad for another one.
We have proposed an algorithm that is guaranteed for both *makespan* and *minsum*. This algorithm has been implemented for managing the resources of a cluster of the regional grid CIMENT. More recently, we extended such analysis to other objectives (makespan and reliability). We concentrate now on finding good algorithms able to schedule a set of jobs with a large variety of objectives simultaneously. For hard problems, we propose approximation of Pareto curves (best compromizes).

Incertainties. Most of the new execution supports are characterized by a higher complexity in predicting the parameters (high versatility in desktop grids, machine crash, communication congestion, cache effects, etc.). We studied some time ago the impact of incertainties on the scheduling algorithms. There are several ways for dealing with this problem: First, it is possible to design robust algorithms that can optimized a problem over a set of scenarii, another solution is to design flexible algorithms. Finally, we promote semi on-line approaches that start from an optimized off-line solution computed on an initial data set and updated during the execution on the "perturbed" data (stability analysis).

Game Theory. Game Theory is a framework that can be used for obtaining good solution of both previous problems (multi-objective optimization and incertain data). On the first hand, it can be used as a complement of multi-objective analysis. On the other hand, it can take into account the incertainties. We are curently working at formalizing the concept of cooperation.

Scheduling for optimizing parallel time and memory space. It is well known that parallel time and memory space are two antagonists criteria. However, for many scientific computations, the use of parallel architectures is motivated by increasing both the computation power and the memory space. Also, scheduling for optimizing both parallel time and memory space targets an important multicriteria objective. Based on the analysis of the dataflow related to the execution, we have proposed a scheduling algorithm with provable performance.

Coarse-grain scheduling of fine grain multithreaded computations on heterogeneous platforms. Designing multi-objective scheduling algorithms is a transversal problem. Work-stealing scheduling is well studied for fine grain multithreaded computations with a small critical time: the speed-up is asymptotically optimal. However, since the number of tasks to manage is huge, the control of the scheduling is expensive. We proposed a generalized lock-free cactus stack execution mechanism, to extend previous results, mainly from Cilk, based on the *work-first principle* for strict multi-threaded computations on SMPs to general multithreaded computations with dataflow dependencies. The main result is that optimizing the sequential local executions of tasks enables to amortize the overhead of scheduling. This distributed work-stealing scheduling algorithm has been implemented in **XKaapi**.

3.2. Adaptive Parallel and Distributed Algorithms Design

Participants: François Broquedis, Pierre-François Dutot, Thierry Gautier, Guillaume Huard, Bruno Raffin, Jean-Louis Roch, Denis Trystram, Frédéric Wagner.

This theme deals with the analysis and the design of algorithmic schemes that control (statically or dynamically) the grain of interactive applications.

The classical approach consists in setting in advance the number of processors for an application, the execution being limited to the use of these processors. This approach is restricted to a constant number of identical resources and for regular computations. To deal with irregularity (data and/or computations on the one hand; heterogeneous and/or dynamical resources on the other hand), an alternate approach consists in adapting the potential parallelism degree to the one suited to the resources. Two cases are distinguished:

- in the classical bottom-up approach, the application provides fine grain tasks; then those tasks are clustered to obtain a minimal parallel degree.
- the top-down approach (Cilk, Cilk+, TBB, Hood, Athapascan) is based on a work-stealing scheduling driven by idle resources. A local sequential depth-first execution of tasks is favored when recursive parallelism is available.

Ideally, a good parallel execution can be viewed as a flow of computations flowing through resources with no control overhead. To minimize control overhead, the application has to be adapted: a parallel algorithm on p resources is not efficient on $q < p$ resources. On one processor, the scheduler should execute a sequential algorithm instead of emulating a parallel one. Then, the scheduler should adapt to resource availability by changing its underlying algorithm. This first way of adapting granularity is implemented by XKaapi (default work-stealing schedule based on work-first principle).

However, this adaptation is restrictive. More generally, the algorithm should adapt itself at runtime to improve its performance by decreasing the overheads induced by parallelism, namely the arithmetic operations and communications. This motivates the development of new parallel algorithmic schemes that enable the scheduler to control the distribution between computation and communication (grain) in the application to find the good balance between parallelism and synchronizations. MOAIS has exhibited several techniques to manage adaptivity from an algorithmic point of view:

- amortization of the number of global synchronizations required in an iteration (for the evaluation of a stopping criterion);
- adaptive deployment of an application based on on-line discovery and performance measurements of communication links;
- generic recursive cascading of two kind of algorithms: a sequential one, to provide efficient executions on the local resource, and a parallel one that enables an idle resource to extract parallelism to dynamically suit the degree of parallelism to the available resources.

The generic underlying approach consists in finding a good mix of various algorithms, what is often called a "poly-algorithm". Particular instances of this approach are Atlas library (performance benchmark are used to decide at compile time the best block size and instruction interleaving for sequential matrix product) and FFTW library (at run time, the best recursive splitting of the FFT butterfly scheme is precomputed by dynamic programming). Both cases rely on pre-benchmarking of the algorithms. Our approach is more general in the sense that it also enables to tune the granularity at any time during execution. The objective is to develop processor oblivious algorithms: similarly to cache oblivious algorithms, we define a parallel algorithm as *processor-oblivious* if no program variable that depends on architecture parameters, such as the number or processors or their respective speeds, needs to be tuned to minimize the algorithm runtime.

We have applied this technique to develop processor oblivious algorithms for several applications with provable performance: iterated and prefix sum (partial sums) computations, stream computations (cipher and hd-video transformation), 3D image reconstruction (based on the concurrent usage of multi-core and GPU), loop computations with early termination.

Extensions concern the development of algorithms that are both cache and processor oblivious on heterogeneous processors. The processor algorithms proposed for prefix sums and segmentation of an array are cache oblivious too.

3.3. Interactivity

Participants: Vincent Danjean, Pierre-François Dutot, Thierry Gautier, Bruno Raffin, Jean-Louis Roch.

The goal of this theme is to develop approaches to tackle interactivity in the context of large scale distributed applications.

We distinguish two types of interactions. A user can interact with an application having only little insight about the internal details of the program running. This is typically the case for a virtual reality application where the user just manipulates 3D objects. We have a "user-in-the-loop". In opposite, we have an "expert -in-the-loop" if the user is an expert that knows the limits of the progam that is being executed and that he can interacts with it to steer the execution. This is the case for instance when the user can change some parameters during the execution to improve the convergence of a computation.

3.3.1. User-in-the-loop

Some applications, like virtual reality applications, must comply with interactivity constraints. The user should be able to observe and interact with the application with an acceptable reaction delay. To reach this goal the user is often ready to accept a lower level of details. To execute such application on a distributed architecture requires to balance the workload and activation frequency of the different tasks. The goal is to optimize CPU and network resource use to get as close as possible to the reactivity/level of detail the user expect.

Virtual reality environments significantly improve the quality of the interaction by providing advanced interfaces. The display surface provided by multiple projectors in CAVE -like systems for instance, allows a high resolution rendering on a large surface. Stereoscopic visualization gives an information of depth. Sound and haptic systems (force feedback) can provide extra information in addition to visualized data. However driving such an environment requires an important computation power and raises difficult issues of synchronization to maintain the overall application coherent while guaranteeing a good latency, bandwidth (or refresh rate) and level of details. We define the coherency as the fact that the information provided to the different user senses at a given moment are related to the same simulated time.

Today's availability of high performance commodity components including networks, CPUs as well as graphics or sound cards make it possible to build large clusters or grid environments providing the necessary resources to enlarge the class of applications that can aspire to an interactive execution. However the approaches usually used for mid size parallel machines are not adapted. Typically, there exist two different approaches to handle data exchange between the processes (or threads). The synchronous (or FIFO) approach ensures all messages sent are received in the order they were sent. In this case, a process cannot compute a new state if all incoming buffers do not store at least one message each. As a consequence, the application refresh rate is driven by the slowest process. This can be improved if the user knows the relative speed of each module and specify a read frequency on each of the incoming buffers. This approach ensures a strong coherency but impact on latency. This is the approach commonly used to ensure the global coherency of the images displayed in multi-projector environments.The other approach, the asynchronous one, comes from sampling systems. The producer updates data in a shared buffer asynchronously read by the consumer. Some updates may be lost if the consumer is slower than the producer. The process refresh rates are therefore totally independent. Latency is improved as produced data are consumed as soon as possible, but no coherency is ensured. This approach is commonly used when coupling haptic and visualization systems. A fine tuning of the application usually leads to satisfactory results where the user does not experience major incoherences. However, in both cases, increasing the number of computing nodes quickly makes infeasible hand tuning to keep coherency and good performance.

We propose to develop techniques to manage a distributed interactive application regarding the following criteria :

- latency (the application reactivity);
- refresh rate (the application continuity);
- coherency (between the different components);
- level of detail (the precision of computations).

We developed a programming environment, called FlowVR, that enables the expression and realization of loosen but controlled coherency policies between data flows. The goal is to give users the possibility to express a large variety of coherency policies from a strong coherency based on a synchronous approach to an uncontrolled coherency based on an asynchronous approach. It enables the user to loosen coherency where it is acceptable, to improve asynchronism and thus performance. This approach maximizes the refresh rate and minimizes the latency given the coherency policy and a fixed level of details. It still requires the user to tune many parameters. In a second step, we are planning to explore auto-adaptive techniques that enable to decrease the number of parameters that must be user tuned. The goal is to take into account (possibly dynamically) user specified high level parameters like target latencies, bandwidths and levels of details, and to have the system automatically adapt to reach a trade-off given the user wishes and the resources available. Issues include multi-criterion optimizations, adaptive algorithmic schemes, distributed decision making, global stability and balance of the regulation effort.

3.3.2. Expert-in-the-loop

Some applications can be interactively guided by an expert who may give advices or answer specific questions to hasten a problem resolution. A theoretical framework has been developed in the last decade to define precisely the complexity of a problem when interactions with an expert is allowed. We are studying these interactive proof systems and interactive complexity classes in order to define efficient interactive algorithms dedicated to scheduling problems. This, in particular, applies to load-balancing of interactive simulations when a user interaction can generate a sudden surge of imbalance which could be easily predicted by an operator.

3.4. Adaptive middleware for code coupling and data movements

Participants: François Broquedis, Vincent Danjean, Thierry Gautier, Clément Pernet, Bruno Raffin, Jean-Louis Roch, Frédéric Wagner.

This theme deals with the design and implementation of programming interfaces in order to achieve an efficient coupling of distributed components.

The implementation of interactive simulation application requires to assemble together various software components and to ensure a semantic on the displayed result. To take into account functional aspects of the computation (inputs, outputs) as well as non functional aspects (bandwidth, latency, persistence), elementary actions (method invocation, communication) have to be coordinated in order to meet some performance objective (precision, quality, fluidity, *etc*). In such a context the scheduling algorithm plays an important role to adapt the computational power of a cluster architecture to the dynamic behavior due to the interactivity. Whatever the scheduling algorithm is, it is fundamental to enable the control of the simulation. The purpose of this research theme is to specify the semantics of the operators that perform components assembling and to develop a prototype to experiment our proposals on real architectures and applications.

3.4.1. Application Programming Interface

The specification of an API to compose interactive simulation application requires to characterize the components and the interaction between components.The respect of causality between elementary events ensures, at the application level, that a reader will see the *last* write with respect to an order. Such a consistency should be defined at the level of the application to control the events ordered by a chain of causality. For instance, one of the result of Athapascan was to prove that a data flow consistency is more efficient than other ones because it generates fewer messages. Beyond causality based interactions, new models of interaction should be studied to capture non predictable events (delay of communication, capture of image) while ensuring a semantic.

Our methodology is based on the characterization of interactions required between components in the context of an interactive simulation application. For instance, criteria could be coherency of visualization, degree of interactivity. Beyond such characterization we hope to provide an operational semantic of interactions (at least well suited and understood by usage) and a cost model. Moreover they should be preserved by composition to predict the cost of an execution for part of the application.

The main result relies on a computable representation of the future of an execution; representations such as macro data flow are well suited because they explicit which data are required by a task. Such a representation can be built at runtime by an interpretation technique: the execution of a function call is differed by computing beforehand at runtime a graph of tasks that represents the (future) calls to execute.

3.4.2. Kernel for Asynchronous, Adaptive, Parallel and Interactive Application

Managing the complexity related to fine grain components and reaching high efficiency on a cluster architecture require to consider a dynamic behavior. Also, the runtime kernel is based on a representation of the execution: data flow graph with attributes for each node and efficient operators will be the basis for our software. This kernel has to be specialized for the considered applications. The low layer of the kernel has features to transfer data and to perform remote signalization efficiently. Well known techniques and legacy code have to be reused. For instance, multithreading, asynchronous invocation, overlapping of latency by computing, parallel communication and parallel algorithms for collective operations are fundamental techniques to reach performance. Because the choice of the scheduling algorithm depends on the application and the architecture, the kernel will provide an *causally connected representation* of the system that is running. This allows to specialize the computation of a good schedule of the data flow graph by providing algorithms (scheduling algorithms for instance) that compute on this (causally connected) representation: any modification of the representation is turned into a modification on the system (the parallel program under execution). Moreover, the kernel provides a set of basic operators to manipulate the graph (*e.g.* computes a partition from a schedule, remapping tasks, ...) to allow to control a distributed execution.

4. New Software and Platforms

4.1. XKaapi

Participants: Thierry Gautier [correspondant], François Broquedis, Vincent Danjean, Joao Ferreira Lima.

- ACM: D.1.3
- License: CeCILL
- OS/Middelware: Unix (Linux, MacOSX, ...)
- Programming language: C/C++, Fortran
- Characterization of Software : A-3 / SO-4 / SM-3 / EM-3 / SDL-4
- Own Contribution: DA-4 / CD-4 / MS-4 / TPM-4
- Additional information:

  XKaapi (http://kaapi.gforge.inria.fr, coordinator T. Gautier) is a library for high performance applications running on multi-cores/multi- processors with support for multi-GPUs. Publicly available at http://kaapi.gforge.inria.fr under CeCILL licence. XKaapi provides ABI compliant implementations of libGOMP (GCC runtime for OpenMP) and was one of the target runtime of the K'Star compiler (http://kstar.gforge.inria.fr). Direct competitors with : OMPSs (BSC), OpenMP, StarPU (Inria RUNTIME)

4.2. FlowVR

Participants: Bruno Raffin [correspondant MOAIS], Matthieu Dreher, Jérémy Jaussaud.

- ACM: D.1.3
- License: GPL and LGPL
- OS/Middelware: Unix (Linux, MacOSX, ...)
- Programming language: C/C++
- Characterization of Software : A-3 / SO-4 / SM-3 / EM-3 / SDL-4
- Own Contribution: DA-4 / CD-3 / MS-3 / TPM-4
- Additional information: FlowVR (http://flowvr.sf.net, coordinator B. Raffin) is an open source middelware to augment parallel simulations running on thousands of cores with in situ processing capabilities and live steering. FlowVR offers a very flexible environment while enabling high performance asynchronous in situ and in transit processing.

FlowVR was initially used for large scale virtual reality applications like real-time multicamera 3D modeling or telepresence. We recently retargeted FlowVR at in situ processing with development efforts focused on optimizing FlowVR performance at large scale and easing its usage in supercomputer environments.

4.3. TakTuk - Adaptive large scale remote execution deployment

Participants: Guillaume Huard [correspondant], Pierre Neyron.

- Characterization of Software : A-2 / SO-3 / SM-5 / EM-3 / SDL-4
- Own Contribution: DA-4 / CD-4 / MS-4 / TPM-4
- Additional information:
  - web site: http://taktuk.gforge.inria.fr, Coordinator G. Huard
  - Objective of the software: TakTuk is a tool for deploying parallel remote executions of commands to a potentially large set of remote nodes. It spreads itself using an adaptive algorithm and sets up an interconnection network to transport commands and perform I/Os multiplexing/demultiplexing. The TakTuk mechanics dynamically adapt to environment (machine performance and current load, network contention) by using a reactive work-stealing algorithm that mixes local parallelization and work distribution.
  - Users community: TakTuk is a research open source project available in the Debian GNU/Linux distribution (package taktuk) used in lower levels of Grid5000 software architectures (nodes monitoring in OAR, environment diffusion in Kadeploy). The community is small : developers and administrators for large scale distributed platforms, but active.
  - Positioning: main competing tools are pdsh (but uses linear deployment) and gexec (not fault tolerant, requires installation), for more details : B. Claudel, G. Huard and O. Richard. TakTuk, Adaptive Deployment of Remote Executions. In Proceedings of the International Symposium on High Performance Distributed Computing (HPDC), 2009. TakTuk is the only tool to provide to deployed processes a communication layer (just like an MPIrun, but not tied to a specific environment) and synchronization capabilities.

4.4. Triva

Participant: Guillaume Huard [correspondant].

- Characterization of Software : A-2 / SO-4 / SM-5 / EM-3 / SDL-3
- Own Contribution: DA-4 / CD-3 / MS-3 / TPM-3
- Additional information:
  - web site: http://triva.gforge.inria.fr/, Coordinator, Lucas Schnorr
  - Objective of the software: Triva is an open-source tool used to analyze traces (in the pajé format) registered during the execution of parallel applications. The tool serves also as a sandbox to the development of new visualization techniques.
  - Users community: Research open source project, applications developers, especially parallel applications.
  - Positioning: Main competing tools are Vampir (classical 2D Gantt charts) and Tau (less advanced agregation techniques), more details in : A Hierarchical Aggregation Model to achieve Visualization Scalability in the analysis of Parallel Applications. Lucas Mello Schnorr, Guillaume Huard, Philippe Olivier Alexandre Navaux. Parallel Computing. Volume 38, Issue 3, March 2012.

4.5. OAR

Participants: Pierre Neyron [correspondant MOAIS], Grégory Mounié.

- Characterization of Software : A-5 / SO-3 / SM-4 / EM-4 / SDL-5
- Own Contribution: DA-3 / CD-2 / MS-1 / TPM-1
- Additional information: OAR (http://oar.imag.fr, Coordinator O. Richard, Inria MESCAL) is a batch scheduler. The MOAIS team develops the central automata and the scheduling module that includes successive evolutions and improvements of the policy.OAR is used to schedule jobs both on the CiGri (Grenoble region) and Grid50000 (France) grids. CiGri is a production grid that federates about 500 heterogeneous resources of various Grenoble laboratories to perform computations in physics. MOAIS has also developed the distributed authentication for access to Grid5000.

4.6. LinBox

Participants: Clément Pernet [correspondant], Thierry Gautier.

- Characterization of Software : A-3 / SO-4 / SM-2 / EM-3 / SDL-5
- Own Contribution: DA-4 / CD-3 / MS-3 / TPM-4
- Additional information:
  - web site: http://linalg.org
  - Objective of the software: LinBox is an open-source C++ template library for exact, high-performance linear algebra computations. It is considered as the reference library for numerous computations (such as linear system solving, rank, characteristic polynomial, Smith normal forms,...) over finite fields and integers with dense, sparse, and structured matrices.
  - The LinBox group is an international collaboration (USA: NCSU, UDel; Canada: U Waterloo, U Calgary; France: LIP, LIRMM, LJK and LIG). Articles related to the library have been published in the main Conferences of the area: ISSAC, ICMS. MOAIS contributes to its development and more specifically to its parallelization in the context of ANR HPAC project. It is currently experiencing a major change of design, to better integrate parallelism.
  - Users community: mostly researchers doing computational mathematics (number theory, cryptology, group theory, persistent homology. They use the library by either linking against it directly (the library is packaged in Debian, Fedora, etc) or withing the general purpose math software Sage (sagemath.org very broad diffusion) which includes LinBox as a kernel for exact linear algebra.

4.7. K'Star

Participants: Thierry Gautier [correspondant], François Broquedis, Pïerrick Brunet, Philippe Virouleau, Olivier Aumage [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Samuel Thibault [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Nathalie Furmento [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Samuel Pitoiset [RUNTIME project-team, Inria Bordeaux - Sud-Ouest].

- ACM: D.1.3
- License: CeCILL
- OS/Middelware: Unix (Linux, MacOSX, ...)
- Programming language: C/C++
- Characterization of Software : A-3 / SO-2 / SM-3 / EM-2 / SDL-4
- Own Contribution: DA-4 / CD-4 / MS-4 / TPM-4
- Additional information:

The K'Star project (http://kstar.gforge.inria.fr) supports the development of Klang, a source-to-source compiler that turns C programs with OpenMP pragmas to C programs with calls to either the StarPU or the XKaapi runtime system. K'Star is a collaboration with the EPI RUNTIME/STORM.

4.8. Kastors

Participants: Thierry Gautier [correspondant], François Broquedis, Pïerrick Brunet, Philippe Virouleau, Olivier Aumage [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Samuel Thibault [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Nathalie Furmento [RUNTIME project-team, Inria Bordeaux - Sud-Ouest].

- ACM: D.1.3
- License: CeCILL
- OS/Middelware: Unix (Linux, MacOSX, ...)
- Programming language: C/C++
- Characterization of Software : A-3 / SO-2 / SM-3 / EM-2 / SDL-4
- Own Contribution: DA-4 / CD-4 / MS-4 / TPM-4
- Additional information:

  The KaStORS benchmarks suite (http://kastors.gforge.inria.fr) has been designed to evaluate the implementation of the OpenMP dependent task paradigm, introduced as part of the OpenMP 4.0 specification. KaStORS is a collaboration with the EPI RUNTIME/STORM.

4.9. Platforms

4.9.1. Multi-camera Platforms Grimage and Kinovis

MOAIS has managed with the LJK-Inria Morpheo team the Grimage platfrom (http://grimage.inrialpes.fr) dedicated to off-line and on-line 3D modeling from multiple cameras and telepresence. In 2012, we received an Equipex funding, Kinovis (http://kinovis.inrialpes.fr), to renew this platform. Kinovis will be operational by early 2015 and will consist of 68 cameras, a compute cluster and a large acquisition space. FlowVR is the software backbone of both platforms for live processing. MOAIS is participating to the FP7 infrastructure project Visionair to enable European research teams to experiment on both platforms.

4.9.2. HPC Platforms Grid'5000 and Ciment

MOAIS is involved in the national platform Grid'5000, the regional mezzo center Ciment and obtained in 2014 with the Mescal and Erods team a grant (FAIRE from Grenoble-INP and LIG) to buy various large NUMA nodes and accelerators that will be integrated into the Grid'5000 infrastructure.

5. New Results

5.1. Scheduling semi-malleable jobs to minimize mean flow time

This paper [9] deals with the problem of scheduling n_A malleable and n_B non-malleable jobs to be executed together on two parallel identical machines to minimize mean flow time. We propose a set of dominant schedules for this problem, and a dynamic programming algorithm that finds an optimal schedule in this dominant set in time $O(n_A^2 n_B)$.

5.2. Elements of Design for Containers and Solutions in the LinBox Library

We describe in this paper [12] new design techniques used in the exact linear algebra library LinBox, intended to make the library safer and easier to use, while keeping it generic and efficient. First, we review the new simplified structure for containers, based on our *founding scope allocation* model. We explain design choices and their impact on coding: unification of our matrix classes, clearer model for matrices and submatrices,... Then we present a variation of the *strategy* design pattern that is comprised of a controller-plugin system: the controller (solution) chooses among plug-ins (algorithms) that always call back the controllers for subtasks. We give examples using the solution `mul`. Finally we present a benchmark architecture that serves two purposes: Providing the user with easier ways to produce graphs; Creating a framework for automatically tuning the library and supporting regression testing.

5.3. Scheduling Data Flow Program in XKaapi: A New Affinity Based Algorithm for Heterogeneous Architectures

Efficient implementations of parallel applications on heterogeneous hybrid architectures require a careful balance between computations and communications with accelerator devices. Even if most of the communication time can be overlapped by computations, it is essential to reduce the total volume of communicated data. The literature therefore abounds with ad hoc methods to reach that balance, but these are architecture and application dependent. We propose [12] here a generic mechanism to automatically optimize the scheduling between CPUs and GPUs, and compare two strategies within this mechanism: the classical Heterogeneous Earliest Finish Time (HEFT) algorithm and our new, parametrized, Distributed Affinity Dual Approximation algorithm (DADA), which consists in grouping the tasks by affinity before running a fast dual approximation. We ran experiments on a heterogeneous parallel machine with twelve CPU cores and eight NVIDIA Fermi GPUs. Three standard dense linear algebra kernels from the PLASMA library have been ported on top of the XKaapi runtime system. We report their performances. It results that HEFT and DADA perform well for various experimental conditions, but that DADA performs better for larger systems and number of GPUs, and, in most cases, generates much lower data transfers than HEFT to achieve the same performance.

5.4. Evaluation of OpenMP Dependent Tasks with the KASTORS Benchmark Suite

The recent introduction of task dependencies in the OpenMP specification provides new ways of synchronizing tasks. Application programmers can now describe the data a task will read as input and write as output, letting the runtime system resolve fine-grain dependencies between tasks to decide which task should execute next. Such an approach should scale better than the excessive global synchronization found in most OpenMP 3.0 applications. As promising as it looks however, any new feature needs proper evaluation to encourage application programmers to embrace it. This paper [26] introduces the KASTORS benchmark suite designed to evaluate OpenMP tasks dependencies. We modified state-of-the-art OpenMP 3.0 benchmarks and data-flow parallel linear algebra kernels to make use of tasks dependencies. Learning from this experience, we propose extensions to the current OpenMP specification to improve the expressiveness of dependencies. We eventually evaluate both the GCC/libGOMP and the CLANG/libIOMP implementations of OpenMP 4.0 on our KASTORS suite, demonstrating the interest of task dependencies compared to taskwait-based approaches.

5.5. Sparse Polynomial Interpolation Codes and their decoding beyond half the minimal distance

We present [21] algorithms performing sparse univariate polynomial interpolation with errors in the evaluations of the polynomial. Based on the initial work by Comer, Kaltofen and Pernet [Proc. ISSAC 2012], we define the sparse polynomial interpolation codes and state that their minimal distance is precisely the length divided by twice the sparsity. At ISSAC 2012, we have given a decoding algorithm for as much as half the

minimal distance and a list decoding algorithm up to the minimal distance. Our new polynomial-time list decoding algorithm uses sub-sequences of the received evaluations indexed by a linear progression, allowing the decoding for a larger radius, that is, more errors in the evaluations while returning a list of candidate sparse polynomials. We quantify this improvement for all typically small values of number of terms and number of errors, and provide a worst case asymptotic analysis of this improvement. For instance, for sparsity T = 5 with up to 10 errors we can list decode in polynomial-time from 74 values of the polynomial with unknown terms, whereas our earlier algorithm required 2T (E + 1) = 110 evaluations. We then propose two variations of these codes in characteristic zero, where appropriate choices of values for the variable yield a much larger minimal distance: the length minus twice the sparsity.

5.6. A Spatiotemporal Data Aggregation Technique for Performance Analysis of Large-scale Execution Traces

Analysts commonly use execution traces collected at runtime to understand the behavior of an application running on distributed and parallel systems. These traces are inspected post mortem using various visualization techniques that, however, do not scale properly for a large number of events. This issue, mainly due to human perception limitations, is also the result of bounded screen resolutions preventing the proper drawing of many graphical objects. This paper [21] proposes a new visualization technique overcoming such limitations by providing a concise overview of the trace behavior as the result of a spatiotemporal data aggregation process. The experimental results show that this approach can help the quick and accurate detection of anomalies in traces containing up to two hundred million events.

5.7. Scheduling independent tasks on multi-cores with GPU accelerators

More and more computers use hybrid architectures combining multi-core processors and hardware accelerators like GPUs (Graphics Processing Units). We present in this paper [3] a new method for scheduling efficiently parallel applications with m CPUs and k GPUs, where each task of the application can be processed either on a core (CPU) or on a GPU. The objective is to minimize the maximum completion time (makespan). The corresponding scheduling problem is NP-hard, we propose an efficient approximation algorithm which achieves an approximation ratio of $4/3 + 1/3k$. We first detail and analyze the method, based on a dual approximation scheme, that uses dynamic programming to balance evenly the load between the heterogeneous resources. Then, we present a faster approximation algorithm for a special case of the previous problem, where all the tasks are accelerated when affected to GPU, with a performance guarantee of $3/2$ for any number of GPUs. We run some simulations based on realistic benchmarks and compare the solutions obtained by a relaxed version of the generic method to the one provided by a classical scheduling algorithm (HEFT). Finally, we present an implementation of the $4/3$-approximation and its relaxed version on a classical linear algebra kernel into the scheduler of the XKaapi runtime system.

5.8. A Flexible Framework for Asynchronous In Situ and In Transit Analytics for Scientific Simulations

High performance computing systems are today composed of tens of thousands of processors and deep memory hierarchies. The next generation of machines will further increase the unbalance between I/O capabilities and processing power. To reduce the pressure on I/Os, the in situ analytics paradigm proposes to process the data as closely as possible to where and when the data are produced. Processing can be embedded in the simulation code, executed asynchronously on helper cores on the same nodes, or performed in transit on staging nodes dedicated to analytics. Today, software environnements as well as usage scenarios still need to be investigated before in situ analytics become a standard practice. In this paper [3] we introduce a framework for designing, deploying and executing in situ scenarios. Based on a component model, the scientist designs analytics workflows by first developing processing components that are next assembled in a dataflow graph through a Python script. At runtime the graph is instantiated according to the execution context, the framework taking care of deploying the application on the target architecture and coordinating the analytics workflows

with the simulation execution. Component coordination, zero-copy intra-node communications or inter-nodes data transfers rely on per-node distributed daemons. We evaluate various scenarios performing in situ and in transit analytics on large molecular dynamics systems simulated with Gromacs using up to 1664 cores. We show in particular that analytics processing can be performed on the fraction of resources the simulation does not use well, resulting in a limited impact on the simulation performance (less than 6%). Our more advanced scenario combines in situ and in transit processing to compute a molecular surface based on the Quicksurf algorithm.

5.9. Generic Deterministic Random Number Generation in Dynamic-Multithreaded Platforms

On dynamic multithreaded platforms with on-line scheduling such as work-stealing, randomized computations raise the issue of reproducibility. Compliant with de facto standard sequential Deterministic Random Number Generators (DRNGs) noted R, we propose [23] a parallel DRNG implementation for finite computations that provides deterministic parallel execution. It uses the stateless sub-stream approach, enabling the use of efficient DRNG such as Mersenne Twister or Linear Congruential. We demonstrate that if R provides fast jump ahead in the random sequence, the re-seeding overhead is small, polylog in expectation, independently from the parallel computation's depth. Experiments benchmark the performance of randomized algorithms employing our solution against the stateful DRNG DotMix, tailored to the Cilk Plus dynamic multithreading runtime. The overhead of our implementation ParDRNG compares favorably to the linear overhead of DotMix re-seedings.

6. Bilateral Contracts and Grants with Industry

6.1. Bilateral Contracts with Industry

- Contract with Bull (2013–2016). Multiobjective scheduling on supercomputer towards exascale. Associated to a CIFRE PhD grant (David Glesser, started in 4/2013). Partners: Inria - LIG Moais, Bull
- Contract with CEA Saclay (2014): Hierarchical work stealing for Eruoplexus. Partners: Moais and CEA Saclay.
- Contract with CEA Bruyères-le-Châtel. (2014): Multi-core and many-core parallelization for scientific visualization. Partners: Moais and CEA Bruyères-le-Châtel.
- Contract with CEA Saclay. (2015): Multi-core and many-core parallelization of EPX code. Partners: Moais and CEA Saclay.
- Contract with Bulll (2014-2016): Multi-objective batch scheduling. Partners: Moais and Bull.
- Contract with Incas-ITSec (2014): IPSec with pre-shared key for MISTIC security module. Partners: Moais, Privatics and Incas-ITSec
- Contract with XYALIS (2014): remote software distributed protection (internship support). Partners: Floralis (Moais and UJF-IF [P. Elbaz-Vincent] and XYALIS.

6.2. Bilateral Grants with Industry

- Contract with EDF (2010-2014). High performance scientific visualization. Funds 1 PhD (Mathias Ettinger). Partners: Inria (MOAIS and EVASION), EDF R&D.
- Collaboration with CEA (2012-2015): Europlexus Parallelization with XKaapi. Partners: Inria Rhônes-Alpes and CEA Saclay (CEA funds the PhD of Marwa Sridi started in 4/2013).
- Contract with IFPEN (2014-2017). Multi-CPUs-Multi-GPUs parallelization of numerical solvers. Funds 1 PhD (Adrien Roussel). Partners: Inria (MOAIS), IFPEN.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. ANR

- **ANR grant MOEBIUS (2013-2015).** Multi-objective scheduling for large computing platforms. Coordinator: Grenoble INP (Moais team). Partners: Grenoble INP, Inria, BULL.
- **ANR grant EXAVIZ (2011-2015).** Large Scale Interactive Visual Analysis for Life Science. Partners: Inria Rhône-Alpes, Université d'Orléans, the LBT lab from IBPC, the LIMSI from Université d'Orsay, and the CEMHTI labs from CNRS.
- **ANR HPAC (2012-2015)**. High Performance Algebraic Computing. Coordinator: UJF (LJK/CASYS team). Partners: project-team MOAIS (Grenoble), project-team ARENAIRE (LIP, Lyon), project-team SALSA (LIP6, Paris), the ARITH group (LIRMM lab, Montpellier).
- **Equipex Kinovis (2012-2017)**. 2.6 Meuros. Large scale multi-camera platform (extension of the Grimage platform to 60 cameras, depth and X-ray cameras). Coordinator E Boyer, LJK Inria MORPHEO team. Partners: Inria Rhône-Alpes and the LJK, LIG, LADAF and GIPSA labs.
- **ANR-11-LABX-0025 PERSYVAL-Lab** funds the following PhD in collaboration with other labs:
  - in collaboration with Verimag: Multi-objective optimization for resource management on multicore systems, (PhD Abhinav Srivastav, since 9/2012)
  - In collaboration with Gipsa-lab and Inria BiBop: Simulations of Fibrous Materials. (PhD Gilles Daviet, since 9/2013)
  - in collaboration with Inria Privatics and Verimag: Secure Outsourcing (PhD Amrit Kumar, since 11/2013)

7.1.2. Competitivity Clusters

- **SoC-Trace**, Minalogic 2011-2014 contract. This project aims the development of tools for the monitoring and debug of mumticore systems on chip. Leader: ST-Microelectonic. Partners: Inria (Mescal, Moais); UJF (TIMA, LIG/Hadas); Magilem, ProBayes. Moais contributes with technics and tools for visual aggregation of application traces. The contract funds 1 phD thesis (Damien Dosimont) and 1 year engineer.
- **ARAMIS, PIA contract n°P3342-146798 (2014-2017)**: Architecture Robuste pour les Automates et Mate´riels des Infrastructures Sensibles. Coordinator: ATOS-WorldGrid; Partners: CEA, SecLab, UJF. The UJF gathers the folowing teams: LIG (Moais, Drakkar, Vasco); LJK (Casys); IF; Verimag (DCS). BPI funds UJF with 775 ke (funds 4 PhDs and 5 years egineers), among which 410ke for LIG. Moais co-advises two PhD Thesis: Nicolas Kox with LIG-VASCO team (Rupture de protocole avec garanties de se´curite´ pour les syste`mes de contro^le-commande); Maxime Puys with VERIMAG-DCS (Processus de ge´ne´ration de filtres certifie´s pour les syste`mes de contro^le-commande).
- **PIA ELCI (2014-2017)**. Environnement Logiciel pour le Calcul Intensif. Coordinator BULL. Partners: BULL, CEA, Inria, SAFRAB, UVSQ.

7.1.3. National ADT

- ADT K'STAR with cooperation between EPIs MOAIS and RUNTIME (Bordeaux). Coordinator: T. Gautier. https://gforge.inria.fr/projects/kstar. The main objective is to provide OpenMP-3.1 with some extension from OpenMP-4.0 standard to perform OpenMP programs on multi-CPUs multi-GPUs by using XKaapi and StarPU runtimes.

7.1.4. Inria Project Lab

7.1.4.1. C2S@Exa - Computer and Computational Scienecs at Exascale

Participants: Olivier Aumage [RUNTIME project-team, Inria Bordeaux - Sud-Ouest], Jocelyne Erhel [SAGE project-team, Inria Rennes - Bretagne Atlantique], Philippe Helluy [TONUS project-team, Inria Nancy - Grand-Est], Laura Grigori [ALPINE project-team, Inria Saclay - Île-de-France], Jean-Yves L'excellent [ROMA project-team, Inria Grenoble - Rhône-Alpes], Thierry Gautier [MOAIS project-team, Inria Grenoble - Rhône-Alpes], Luc Giraud [HIEPACS project-team, Inria Bordeaux - Sud-Ouest], Michel Kern [POMDAPI project-team, Inria Paris - Rocquencourt], Stéphane Lanteri [Coordinator of the project], François Pellegrini [BACCHUS project-team, Inria Bordeaux - Sud-Ouest], Christian Perez [AVALON project-team, Inria Grenoble - Rhône-Alpes], Frédéric Vivien [ROMA project-team, Inria Grenoble - Rhône-Alpes].

Since January 2013, the team is participating to the C2S@Exa http://www-sop.inria.fr/c2s_at_exa Inria Project Lab (IPL). This national initiative aims at the development of numerical modeling methodologies that fully exploit the processing capabilities of modern massively parallel architectures in the context of a number of selected applications related to important scientific and technological challenges for the quality and the security of life in our society. At the current state of the art in technologies and methodologies, a multidisciplinary approach is required to overcome the challenges raised by the development of highly scalable numerical simulation software that can exploit computing platforms offering several hundreds of thousands of cores. Hence, the main objective of C2S@Exa is the establishment of a continuum of expertise in the computer science and numerical mathematics domains, by gathering researchers from Inria project-teams whose research and development activities are tightly linked to high performance computing issues in these domains. More precisely, this collaborative effort involves computer scientists that are experts of programming models, environments and tools for harnessing massively parallel systems, algorithmists that propose algorithms and contribute to generic libraries and core solvers in order to take benefit from all the parallelism levels with the main goal of optimal scaling on very large numbers of computing entities and, numerical mathematicians that are studying numerical schemes and scalable solvers for systems of partial differential equations in view of the simulation of very large-scale problems.

T. Gautier is coordinator of the Pole 4: Programming Models.

7.2. European Initiatives

7.2.1. FP7 & H2020 Projects

7.2.1.1. VISIONAIR

Type: FP7

Defi: NC

Instrument: Combination of COLLABORATIVE PROJECTS and COORDINATION and SUPPORT ACTIONS

Objectif: NC

Duration: February 2011 - January 2015

Coordinator: Frederic Noël, Grenoble INP.

Partner: Gather 27 European Partners.

Inria contact: G. Dumont

Abstract: Federation of European Virtual Reality and Scientific Visualization Platforms opened to European researchers. Moais involved through the Grimage Platform.

7.3. International Initiatives

7.3.1. Inria Associate Teams

7.3.1.1. ANOMALIES@EXASCALE

Title: Anomalies Detection and Handling towards Exascale Platforms

International Partner (Institution - Laboratory - Researcher):

University of Chicago (ÉTATS-UNIS)

Duration: 2014 - 2016

See also: http://intra-id.imag.fr/

The Anomalies@exascale project intends to prospect new scheduling solutions for very large parallel computing platforms. In particular, we consider the new problems related to fault tolerance raising with the developments of exascale platforms. We expect to define new ways to detect both execution failures and more transient performance anomalies. Information gathered from the detectors will then be taken into account by schedulers to implement corrective measures.

7.3.1.2. ExaSE

Title: Exascale Computing Scheduling and Energy

International Partner (Institution - Laboratory - Researcher):

UFRGS, PUC Minas and UPS (Brazil)

Duration: 2014 - 2016

See also: https://team.inria.fr/exase/

The main scientific context of this project is high performance computing on Exascale systems: large-scale machines with billions of processing cores and complex hierarchical structures. This project intends to explore the relationship between scheduling algorithms and techniques and the energy constraints present on such exascale systems.

7.3.2. Participation to other International Programs

7.3.2.1. LICIA

Title: International Laboratory in High Performance and Ubiquitous Computing

International Partner (Institution - Laboratory - Researcher):

UFRGS (Brazil)

Duration: 2011 - 2018

See also: http://licia-lab.org/

The LICIA is an Internacional Laboratory and High Performance and Ubiquitous Computing born in 2011 from the common desire of members of Informatics Institute of the Federal University of Rio Grande do Sul and of Laboratoire d'Informatique de Grenoble to enhance and develop their scientific parternship that started by the end of the 1970. LICIA is an Internacional Associated Lab of the CNRS, a public french research institution. It has support from several brazilian and french research funding agencies, such as CNRS, Inria, ANR, European Union (from the french side) and CAPES, CNPq, FAPERGS (from the Brazilian side). Moais is deeply invovled in the creation and animation of LICIA. Bruno Raffin is LICIA associate director.

7.3.2.2. CAPES/COFECUB StarShip

Title: Scalable Tools and Algorithms para Resilient, Scalable, Hybrid Interactive Processing

International Partner (Institution - Laboratory - Researcher):

UFRGS (Brazil)

Duration: 2013 - 2016

7.4. International Research Visitors

7.4.1. Visits of International Scientists

- Guochuan Zhang. Professor at Zhejiang University, China, one month stay at Moais in 2014.

- Adel Safi, Associate Professor ar ESSTT, Tunisia, 2 weeks stay at Moais in 2014.
- Andreï Tchernykh, Researcher at CICESE, Mexico, one month stay at Moais in 2014.
- Monica Liliana Hernandez Ariza, Master Student at University of Santander, Colombia, 4 months stay at Moais in 2014.

8. Dissemination

8.1. Promoting Scientific Activities

Moais has an intensive reviewing activity for a large variety of Conferences and Journals we do not detail here.

8.1.1. Scientific events organisation

8.1.1.1. General chair, scientific chair

- 11th bi-annual Workshop on new challenges in scheduling theory 2014. Chair: D. Trystram.

8.1.1.2. Member of the organizing committee

- EGPGV (Eurographics Symposium on Parallel Rendering and Visualization), Member of the steering committee of : B. Raffin.
- French Day of Visualization (visu2014.imag.fr). Member of the steering committee of : B. Raffin.

8.1.2. Scientific events selection

8.1.2.1. Member of the conference program committee

- LDAV 2014 (IEEE Symposium on Large-Scale Data Anlysis and Visualization), B. Raffin.
- ICCS 2014 (International Conference on Computational Science), B. Raffin.
- SEARIS 2014 (Workshop on Software Engineering and Architectures for Realtime Interactive Systems), B. Raffin.
- JVRC 2014 (International Join Virtual Reality Conference), B. Raffin
- COMPAS, april 22-25, 2013, Neuchatel, Switzerland, D. Trystram
- IPDPS 2014 (25th IEEE International Parallel & Distributed Processing Symposium) may 20-22, Phoenix, USA, D. Trystram.
- HCW'2014 (23th IEEE Heterogeneous Computing Workshop) may 19, 2014, Phoenix, USA, D. Trystram.
- HICOMB'2014 (13th IEEE Internat Workshop on High Performance Computational Biology) may 19, Phoenix, USA,D. Trystram.
- HICOMB'2014 (13th IEEE Internat Workshop on High Performance Computational Biology) may 19, Phoenix, USA. D. Trystram.
- ISPDC (13th Internat Symposium on Parallel and Distributed Computing) june 24-27, Porquerolles Island, France, D. Trystram.
- PMAA'14 (8th internat. workshop on Parallel Matrix Algorithms and Applications) july 2-4, Lugano, Switzerland, D. Trystram.
- OPTIM'14 (Workshop on Opt. Issues in Energy Efficient Distributed Systems) july 21-25, 2014, Bologna, Italy, D. Trystram.
- EuroMPI/ASIA sept. 9-12, 2014, Kyoto, Japan, D. Trystram.
- 26th SBAC-PAD, october 22-24, 2014, Paris, France, D. Trystram.
- CARI (12th conf on research in computer science and applied Maths), oct. 23-26, 2014, Saint Louis, Senegal, D. Trystram.

8.1.3. Journal

8.1.3.1. Member of the editorial board

- Associate Editor of the Parallel Computing journal PARCO, D. Trystram.
- Member of the Editorial Board of JPDC, D. Trystram.
- Member of the Editorial Board of Computational Methods in Science and Technology, D. Trystram.
- Member of the Editorial Board of ARIMA (revue africaine de recherche en informatique et maths appliquées), D. Trystram
- Member of the Editorial Board of IEEE Trans. Parallel and Distributed Systems TPDS, D. Trystram.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Master: J-L. Roch co-director (Grenoble-INP) with P Elbaz-Vincent (Université Joseph Fourier, Math. Dept) of the Master "SCCI Security, Cryptology and Coding of Information Systems" (M2) joined between UJF and INP Grenoble Universities. This Master, started in 2001, is taught in English from sept 2007 (international Master).

Master: C. Pernet and Denis Trystram are responsible of the first year (M1) of the international Master of Science in Informatics at Grenoble (MOSIG-M1).

Master: V Danjean: course "Parallel Programming" (M2), Grenoble University,.

Master: J-L. Roch, "Security models" 24h (M2), Grenoble University

Master: D. Trystram, P.-F. Dutot, J.-L. Roch, "Complexity, approximation theory and randomization" master course (M2) at Grenoble University

Master: François Broquedis. 192 hours per year. 192 hours per year. Engineering school Grenoble-INP/Ensimag, 1st year/L3 and Master (M1/2nd year and M2/3rd year).

Master: Vincent Danjean. 242 hours per year. Licence (third year) and Master (first and second year) at Joseph Fourier University. First to third year of engineering school at Polytech' Grenoble.

Master: Pierre-François Dutot. 226 hours per year. Licence (first and second year) at IUT2/UPMF (Institut Universitaire Technologique de l'Université Pierre Mendès-France) and 9 hours Master M2R-ISC Informatique-Systèmes-Communication at Joseph Fourier University.

Master: Guillaume Huard. 242 hours per year. Licence (first and third year) and Master (first year) at Joseph Fourier University.

Master: Grégory Mounié. 242 hours per year. Master (first year) and Computer Science for Non Computer Scientist Post-Master at Engineering school ENSIMAG and Dept TELECOM, Grenoble-INP.

Master: Clement Pernet. 210 hours per year. University J. Fourier. Master (first year and second year) and Licence (3rd year).

Licence: Bruno Raffin. 38 hours per year. Engineering school ENSIMAG.

Master: Jean-Louis Roch. 242 hours per year. Engineering school Grenoble-INP/Ensimag and Master (M1/2nd year and M2/3rd year)

Master: Denis Trystram. 200 hours per year in average, mainly at first level of Engineering School.

Master: Frédéric Wagner. 220 hours per year. Engineering school ENSIMAG, Grenoble-INP (M1/2nd year and M2/3rd year) (190h), Master DESS/M2-P SCCI Security (30h).

8.2.2. Supervision

Adrien Roussel. Mutli-CPUs/Multi-GPUs parallelization of numerical solvers. Co-advised by Jean-Marc Gratien (IFPEN) and Thierry Gautier.

David Beniamine (since 2013). Parallelisation Patterns and Scheduling for Real-Time Physics Simulations. Co-advised by Guillaume Huard and Bruno Raffin.

Amrit Kumar (since 2013). Analysis of work-stealing and adpative algorithms. Labex-Persyval co-advised PhD by Pascal Lafourcade (Verimag lab), Cedric Lauradoux (Inria Privatics team) and Jean-Louis Roch.

Gilles Daviet (since 2013). Parallel Macroscopic Simulations of Fibrous Materials. Co-advised by Florence Bertails-Descoubes and Bruno Raffin.

Julio Toss (since 2013). Parallel Algorithms and Data Structures for Physically Based Simulation of Deformable Objects. Join Ph.D with UFRGS, Brazil. Co-advised by Joao Comba and Bruno Raffin.

Marwa Sridi (since 2013). Parallel Algorithm Composition for Transient Mechanic Simulations. Join Ph.D with CEA, France. Co-advised by Vincent Faucher, Thierry Gautier and Bruno Raffin.

Ziad Sultan (since 2012). Hig-performance algebraic computations. Co-advised Ph.D by Jean-Guillaume Dumas (LJK Lab) and Clément Pernet.

Stefano Mor (since 2011). Analysis of work-stealing and adpative algorithms. Join Ph.D with UFRGS, Brazil. Co-advised by Jean-Louis Roch, Nicolas Maillard and Bruno Raffin

Mathieu Dreher (since 2011). In-Situ Visualization for Molecular Dynamics. Advised by Bruno Raffin.

David Glesser. Energy Aware Resource Management for HPC. Co-advised by Yianis Georgiou (BULL) and Denis Trystram.

Raphaël Bleuse. Affinity Scheduling. Co-advised by Gregory Mounié and Denis Trystram.

Millian Poquet. Enhanced Data Movements for HPC. Co-advised by Pierre-François Dutot and Denis Trystram.

Abhinav Srivastav. Persyval. Multi-objective Scheduling. Co-advised by Oded Maler (VERIMAG) and Denis Trystram.

Fernando Machado Mendonca. Locality Aware Scheduling. Co-advised by Frederic Wagner and Denis Trystram.

Alessandro Kraemer. Scheduling in the Cloud. Join Ph.D with UFPR, Brazil. Co-advised by Olivier Richard, and Denis Trystram.

8.2.3. Juries

- Vincent Faucher, HDR, INSA Lyon, June 2014.
- Laurent Gourvés, HDR, December 2014
- Florian Jamain,PhD, June 2014
- Rohit Babbar, PhD, October 2014
- Abdelkader Kersani, PhD, October 2014
- Florence Monna, PhD, November 2014
- Dounia Zaidouni, PhD, Decembre 2014
- Georgos Zois, PhD, December 2014)
- Marc Labadens, PhD, Ecole Polytechnique, December 2014.
- Fabien Coubard, PhD, Université Paris-Est, October 2014.

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] C. PERNET. *High Performance and Reliable Algebraic Computing*, Université Jospeh Fourier, Grenoble 1, November 2014, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-01094212

Articles in International Peer-Reviewed Journals

[2] J. BLAZEWICZ, N. CHERIERE, P.-F. DUTOT, J. MUSIAL, D. TRYSTRAM. *Novel dual discounting functions for the Internet shopping optimization problem: new algorithms*, in "Journal of Scheduling", 2014 [*DOI :* 10.1007/S10951-014-0390-0], https://hal.archives-ouvertes.fr/hal-01101708

[3] R. BLEUSE, S. KEDAD-SIDHOUM, F. MONNA, G. MOUNIÉ, D. TRYSTRAM. *Scheduling independent tasks on multi-cores with GPU accelerators*, in "Concurrency: Practice and Experience", 2014, forthcoming [*DOI :* 10.1002/CPE.3359], https://hal.archives-ouvertes.fr/hal-01081625

[4] M. S. BOUGUERRA, D. KONDO, F. MENDONCA, D. TRYSTRAM. *Fault-tolerant scheduling on parallel systems with non-memoryless failure distributions*, in "Journal of Parallel and Distributed Computing", May 2014, vol. 74, n° 5, pp. 2411-2422 [*DOI :* 10.1016/J.JPDC.2014.01.005], https://hal.archives-ouvertes.fr/hal-01101712

[5] J. COHEN, D. CORDEIRO, D. TRYSTRAM. *Coordination mechanisms for decentralized parallel systems*, in "Concurrency and Computation: Practice and Experience", 2014, forthcoming [*DOI :* 10.1002/CPE.3298], https://hal.archives-ouvertes.fr/hal-01101724

[6] M. DREHER, J. PREVOTEAU-JONQUET, M. TRELLET, M. PIUZZI, M. BAADEN, B. RAFFIN, N. FÉREY, S. ROBERT, S. LIMET. *ExaViz: a Flexible Framework to Analyse, Steer and Interact with Molecular Dynamics Simulations*, in "Faraday Discussions of the Chemical Society", May 2014, vol. 169, pp. 119-142 [*DOI :* 10.1039/C3FD00142C], https://hal.inria.fr/hal-00942627

[7] D. EL MOSTAFA, F. WAGNER, D. TRYSTRAM. *Scheduling 2-Dimensional grids with large communication delays*, in "RAIRO Operations Research", 2014, forthcoming [*DOI :* 10.1051/RO/2014048], https://hal.archives-ouvertes.fr/hal-01101728

[8] V. FAUCHER, P. GALON, A. BECCANTINI, F. CROUZET, F. DEBAUD, T. GAUTIER. *Hybrid parallel strategy for the simulation of fast transient accidental situations at reactor scale*, in "Annals of Nuclear Energy", 2014, forthcoming [*DOI :* 10.1016/J.ANUCENE.2014.07.049], https://hal.archives-ouvertes.fr/hal-01101743

[9] Y. HENDEL, W. KUBIAK, D. TRYSTRAM. *Scheduling semi-malleable jobs to minimize mean flow time*, in "Journal of Scheduling", 2014 [*DOI :* 10.1007/S10951-013-0341-1], https://hal.archives-ouvertes.fr/hal-01101783

[10] L. PILLA, C. POUSA RIBEIRO, P. COUCHENEY, F. BROQUEDIS, B. GAUJAL, P. NAVAUX, J.-F. MEHAUT. *A Topology-aware Load Balancing Algorithm for Clustered Hierarchical Multi-core Machines*, in "Future Generation of Computing Systems (FGCS)", 2014, vol. 30, n° 1, pp. 191-201 [*DOI :* 10.1016/J.FUTURE.2013.06.023], https://hal.inria.fr/hal-00953132

International Conferences with Proceedings

[11] R. BLEUSE, T. GAUTIER, J. V. F. LIMA, G. MOUNIÉ, D. TRYSTRAM. *Scheduling Data Flow Program in XKaapi: A New Affinity Based Algorithm for Heterogeneous Architectures*, in "Euro-Par 2014 Parallel Processing - 20th International Conference", Porto, Portugal, August 2014, pp. 560 - 571 [*DOI :* 10.1007/978-3-319-09873-9_47], https://hal.archives-ouvertes.fr/hal-01081629

[12] B. BOYER, J.-G. DUMAS, P. GIORGI, C. PERNET, B. D. SAUNDERS. *Elements of Design for Containers and Solutions in the LinBox Library*, in "4th International Congress on Mathematical Software", Seoul, South Korea, Springer, August 2014, vol. 8592, 8 pages, https://hal.archives-ouvertes.fr/hal-01015138

[13] J. COHEN, D. CORDEIRO, P. L. F. RAPHAEL. *Energy-Aware Multi-Organization Scheduling Problem*, in "Euro-Par 2014 Parallel Processing :20th International Conference", Porto, Portugal, F. SILVA, I. DUTRA, V. S. COSTA (editors), Lecture Notes in Computer Science, Springer, August 2014, vol. 8632, pp. 186-197 [*DOI :* 10.1007/978-3-319-09873-9_16], https://hal.archives-ouvertes.fr/hal-01078280

[14] D. DOSIMONT, R. LAMARCHE-PERRIN, L. MELLO SCHNORR, G. HUARD, J.-M. VINCENT. *A Spatiotemporal Data Aggregation Technique for Performance Analysis of Large-scale Execution Traces*, in "IEEE Cluster 2014", Madrid, Spain, September 2014, https://hal.inria.fr/hal-01065093

[15] D. DOSIMONT, G. PAGANO, G. HUARD, V. MARANGOZOVA-MARTIN, J.-M. VINCENT. *Efficient Analysis Methodology for Huge Application Traces*, in "HPCS 2014 - The 2014 International Conference on High Performance Computing & Simulation", Bologna, Italy, July 2014, https://hal.inria.fr/hal-01065783

[16] M. DREHER, B. RAFFIN. *A Flexible Framework for Asynchronous In Situ and In Transit Analytics for Scientific Simulations*, in "14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing", Chicago, United States, IEEE Computer Science Press, May 2014, https://hal.inria.fr/hal-00941413

[17] J.-G. DUMAS, T. GAUTIER, C. PERNET, Z. SULTAN. *Parallel computation of echelon forms*, in "EuroPar-2014", Porto, Portugal, August 2014, 12 p. , https://hal.archives-ouvertes.fr/hal-00947013

[18] A. ESSAFI, D. TRYSTRAM, Z. ZAIDI. *An Efficient Algorithm for Scheduling Jobs in Volunteer Computing Platforms*, in "HCW - Proceedings of the 2014 IEEE International Parallel & Distributed Processing Symposium Workshops", Phoenix, United States, IEEE Computer Society, 2014, pp. 68–76 [*DOI :* 10.1109/IPDPSW.2014.13], https://hal.archives-ouvertes.fr/hal-01102180

[19] S. GUELTON, J. FALCOU, P. BRUNET. *Exploring the vectorization of Python constructs using Pythran and Boost SIMD*, in "WPMVP 2014 : Workshop on Programming models for SIMD/Vector processing", Orlando, United States, February 2014, pp. 79-86 [*DOI :* 10.1145/2568058.2568060], https://hal.archives-ouvertes.fr/hal-00981551

[20] L. JACQUIN, V. ROCA, J.-L. ROCH. *Too Big or Too Small? The PTB-PTS ICMP-based Attack against IPsec Gateways*, in "IEEE Global Communications Conference (GLOBECOM'14)", Austin, United States, T. RAPPAPORT (editor), IEEE, December 2014, https://hal.inria.fr/hal-01052994

[21] E. L. KALTOFEN, C. PERNET. *Sparse Polynomial Interpolation Codes and their decoding beyond half the minimal distance*, in "ISSAC - 39th International Symposium on Symbolic and Algebraic Computation", Kobe, Japan, July 2014, In Proceedings of the International Symposium on Symbolic and Algebraic Computation 2014 (ISSAC'14) [*DOI :* 10.1145/2608628.2608660], https://hal.inria.fr/hal-01068308

[22] S. KEDAD-SIDHOUM, F. MENDONCA, F. MONNA, G. MOUNIÉ, D. TRYSTRAM. *Fast Biological Sequence Comparison on Hybrid Platforms*, in "ICPP - 43rd International Conference on Parallel Processing", Minneapolis, United States, September 2014, pp. 501 - 509 [*DOI :* 10.1109/ICPP.2014.59], https://hal.archives-ouvertes.fr/hal-01102263

[23] S. MOR, J.-L. ROCH, N. MAILLARD. *Generic Deterministic Random Number Generation in Dynamic-Multithreaded Platforms*, in "Euro-Par 2014 Parallel Processing - 20th International Conference", Porto, Portugal, F. M. A. SILVA, I. DE CASTRO DUTRA, V. S. COSTA (editors), Lecture Notes in Computer Science, August 2014, vol. 8632, pp. 427 - 438 [*DOI :* 10.1007/978-3-319-09873-9_36], https://hal.inria.fr/hal-01102365

[24] M. SRIDI, V. FAUCHER, B. RAFFIN, T. GAUTIER. *Optimisation de l'utilisation du cache dans EUROPLEXUS*, in "ComPAS 2014 : Conférence en Parallélisme, Architecture et Systèmes", Neuchâtel, Switzerland, P. FELBER, L. PHILIPPE, E. RIVIERE, A. TISSERAND (editors), April 2014, https://hal.inria.fr/hal-01020500

[25] J. TOSS, L. D. COMBA, B. RAFFIN. *Parallel Shortest Path Algorithm for Voronoi Diagrams with Generalized Distance Functions*, in "XXVII SIBGRAPI, Conference on Graphics Patterns and Images", Rio de Janerio, Brazil, August 2014, https://hal.inria.fr/hal-01026159

[26] P. VIROULEAU, P. BRUNET, F. BROQUEDIS, N. FURMENTO, S. THIBAULT, O. AUMAGE, T. GAUTIER. *Evaluation of OpenMP Dependent Tasks with the KASTORS Benchmark Suite*, in "IWOMP - 10th International Workshop on OpenMP", Salvador, Brazil, France, 10th International Workshop on OpenMP, IWOMP2014, Springer, September 2014, pp. 16 - 29 [*DOI :* 10.1007/978-3-319-11454-5_2], https://hal.inria.fr/hal-01081974

[27] A. SRIVASTAV, D. TRYSTRAM. *Competitive analysis for non-preemptive average stretch*, in "ECCO XXVII – CO 2014 Joint Conference 27 th Conference of the European Chapter on Combinatorial Optimization", München, France, May 2014, https://hal.archives-ouvertes.fr/hal-01102274

National Conferences with Proceedings

[28] D. DOSIMONT, G. HUARD, J.-M. VINCENT. *Agrégation temporelle pour l'analyse de traces volumineuses*, in "10ème Atelier en Evaluation de Performances", Sophia Antipolis, France, June 2014, https://hal.inria.fr/hal-01065862

Conferences without Proceedings

[29] D. GLESSER, Y. GEORGIOU, M. HAUTREUX, D. TRYSTRAM. *Introducing Power-capping in Slurm scheduling*, in "Slurm User Group Meeting 2014", Lugano, Switzerland, 2014, https://hal.archives-ouvertes.fr/hal-01102285

[30] D. GLESSER, Y. GEORGIOU, D. TRYSTRAM. *Introducing Energy based fair-share scheduling*, in "Slurm User Group Meeting 2014", Lugano, Switzerland, 2014, https://hal.archives-ouvertes.fr/hal-01102291

Scientific Books (or Scientific Book chapters)

[31] J.-G. DUMAS, J.-L. ROCH, E. TANNIER, S. VARRETTE. *Foundations of Coding: Compression, Encryption, Error-Correction*, Wiley, 2015, 376 p. , https://hal.archives-ouvertes.fr/hal-00765802

Research Reports

[32] D. DOSIMONT, L. M. SCHNORR, G. HUARD, J.-M. VINCENT. *A Trace Macroscopic Description based on Time Aggregation*, April 2014, n^{o} RR-8524, Trace visualization; trace analysis; trace overview; time aggregation; parallel systems; embedded systems; information theory; scientific computation; multimedia application; debugging; optimization, https://hal.inria.fr/hal-00981020

[33] A. KUMAR, P. LAFOURCADE, C. LAURADOUX. *Performances of Cryptographic Accumulators*, May 2014, https://hal.archives-ouvertes.fr/hal-00999432

Other Publications

[34] D. DOSIMONT, R. LAMARCHE-PERRIN, L. M. SCHNORR, G. HUARD, J.-M. VINCENT. *Combining Data and Visual Aggregation Techniques to Build a Coherent Spatiotemporal Overview*, October 2014, https://hal.inria.fr/hal-01065853

[35] J.-G. DUMAS, T. GAUTIER, C. PERNET, J.-L. ROCH, Z. SULTAN. *Recursion based parallelization of exact dense linear algebra routines for Gaussian elimination*, November 2014, https://hal.archives-ouvertes.fr/hal-01084238