RESEARCH CENTRE

Inria Lyon Centre

IN PARTNERSHIP WITH:

CNRS, Université Claude Bernard (Lyon 1), Ecole normale supérieure de Lyon

2023

ACTIVITY REPORT

Project-Team

ROMA

Optimisation des ressources : modèles, algorithmes et ordonnancement

IN COLLABORATION WITH: Laboratoire de l'Informatique du Parallélisme (LIP)

DOMAIN

Networks, Systems and Services, Distributed Computing

THEME

Distributed and High Performance Computing

Inria

Contents

Project-Team ROMA

Creation of the Project-Team: 2015 January 01

Keywords

Computer sciences and digital sciences

A1.1.1. – Multicore, Manycore

A1.1.2. – Hardware accelerators (GPGPU, FPGA, etc.)

A1.1.3. – Memory models

A1.1.4. – High performance computing

A1.1.5. – Exascale

A1.1.9. – Fault tolerant systems

A1.6. – Green Computing

A6.1. – Methods in mathematical modeling

A6.2.3. – Probabilistic methods

A6.2.5. – Numerical Linear Algebra

A6.2.6. – Optimization

A6.2.7. – High performance computing

A6.3. – Computation-data interaction

A7.1. – Algorithms

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A8.7. – Graph theory

A8.9. – Performance evaluation

Other research topics and application domains

B3.2. – Climate and meteorology

B3.3. – Geosciences

B4. – Energy

B4.5.1. – Green computing

B5.2.3. – Aviation

B5.5. – Materials

1 Team members, visitors, external collaborators

Research Scientists

- Loris Marchal [Team leader, CNRS, Senior Researcher, HDR]
- Suraj Kumar [INRIA, ISFP]
- Bora Uçar [CNRS, Senior Researcher, HDR]
- Frédéric Vivien [INRIA, Senior Researcher, HDR]

Faculty Members

- Anne Benoît [ENS DE LYON, Associate Professor, HDR]
- Grégoire Pichon [UNIV LYON I, Associate Professor]
- Yves Robert [ENS DE LYON, Professor, HDR]

Post-Doctoral Fellows

- Redouane Elghazi [ENS DE LYON, from Oct 2023]
- Somesh Singh [UDL, Post-Doctoral Fellow, from Sep 2023, Inria until Feb 2023, CNRS from March to August 2023]

PhD Students

- Brian Bantsoukissa [INRIA, until Sep 2023]
- Joachim Cendrier [CNRS, from Oct 2023]
- Anthony Dugois [INRIA, until Aug 2023]
- Redouane Elghazi [UNIV FRANCHE-COMTE, until Sep 2023]
- Maxime Gonthier [INRIA, until Sep 2023]
- Lucas Perotin [ENS DE LYON, until Aug 2023]

Interns and Apprentices

- Yassine Koubaa [Univ. Lyon 1, Intern, from Apr 2023 until Jul 2023]
- Mathis Lamiroy [ENS DE LYON, from Oct 2023]
- Nicolas Malleton [INRIA, Intern, from Feb 2023 until Jul 2023]
- Corentin Munoz [INRIA, Intern, from May 2023 until Jul 2023]

Administrative Assistant

- Chrystelle Mouton [INRIA]

Visiting Scientist

- Julien Langou [INRIA, from Jun 2023 until Jul 2023, on an Inria International Chair]

External Collaborators

- Theo Mary [CNRS]
- Hongyang Sun [UNIV KANSAS]

2 Overall objectives

The ROMA project aims at designing models, algorithms, and scheduling strategies to optimize the execution of scientific applications.

Scientists now have access to tremendous computing power. For instance, the top supercomputers contain more than 100,000 cores, and volunteer computing grids gather millions of processors. Furthermore, it had never been so easy for scientists to have access to parallel computing resources, either through the multitude of local clusters or through distant cloud computing platforms.

Because parallel computing resources are ubiquitous, and because the available computing power is so huge, one could believe that scientists no longer need to worry about finding computing resources, even less to optimize their usage. Nothing is farther from the truth. Institutions and government agencies keep building larger and more powerful computing platforms with a clear goal. These platforms must allow to solve problems in reasonable timescales, which were so far out of reach. They must also allow to solve problems more precisely where the existing solutions are not deemed to be sufficiently accurate. For those platforms to fulfill their purposes, their computing power must therefore be carefully exploited and not be wasted. This often requires an efficient management of all types of platform resources: computation, communication, memory, storage, energy, etc. This is often hard to achieve because of the characteristics of new and emerging platforms. Moreover, because of technological evolutions, new problems arise, and fully tried and tested solutions need to be thoroughly overhauled or simply discarded and replaced. Here are some of the difficulties that have, or will have, to be overcome:

- Computing platforms are hierarchical: a processor includes several cores, a node includes several processors, and the nodes themselves are gathered into clusters. Algorithms must take this hierarchical structure into account, in order to fully harness the available computing power;
- The probability for a platform to suffer from a hardware fault automatically increases with the number of its components. Fault-tolerance techniques become unavoidable for large-scale platforms;
- The ever increasing gap between the computing power of nodes and the bandwidths of memories and networks, in conjunction with the organization of memories in deep hierarchies, requires to take more and more care of the way algorithms use memory;
- Energy considerations are unavoidable nowadays. Design specifications for new computing platforms always include a maximal energy consumption. The energy bill of a supercomputer may represent a significant share of its cost over its lifespan. These issues must be taken into account at the algorithm-design level.

We are convinced that dramatic breakthroughs in algorithms and scheduling strategies are required for the scientific computing community to overcome all the challenges posed by new and emerging computing platforms. This is required for applications to be successfully deployed at very large scale, and hence for enabling the scientific computing community to push the frontiers of knowledge as far as possible. The ROMA project-team aims at providing fundamental algorithms, scheduling strategies, protocols, and software packages to fulfill the needs encountered by a wide class of scientific computing applications, including domains as diverse as geophysics, structural mechanics, chemistry, electromagnetism, numerical optimization, or computational fluid dynamics, to quote a few. To fulfill this goal, the ROMA project-team takes a special interest in dense and sparse linear algebra.

3 Research program

The work in the ROMA team is organized along three research themes.

3.1 Resilience for very large scale platforms

For HPC applications, scale is a major opportunity. The largest supercomputers contain tens of thousands of nodes and future platforms will certainly have to enroll even more computing resources to enter the Exascale era. Unfortunately, scale is also a major threat. Indeed, even if each node provides an individual MTBF (Mean Time Between Failures) of, say, one century, a machine with 100,000 nodes will encounter a failure every 9 hours in average, which is shorter than the execution time of many HPC applications.

To further darken the picture, several types of errors need to be considered when computing at scale. In addition to classical fail-stop errors (such as hardware failures), silent errors (a.k.a silent data corruptions) must be taken into account. The cause for silent errors may be for instance soft errors in L1 cache, or bit flips due to cosmic radiations. The problem is that the detection of a silent error is not immediate, and that they only manifest later, once the corrupted data has propagated and impacted the result.

Our work investigates new models and algorithms for resilience at extreme-scale. Its main objective is to cope with both fail-stop and silent errors, and to design new approaches that dramatically improve the efficiency of state-of-the-art methods. Application resilience currently involves a broad range of techniques, including fault prediction, error detection, error containment, error correction, checkpointing, replication, migration, recovery, etc. Extending these techniques, and developing new ones, to achieve efficient execution at extreme-scale is a difficult challenge, but it is the key to a successful deployment and usage of future computing platforms.

3.2 Multi-criteria scheduling strategies

In this theme, we focus on the design of scheduling strategies that finely take into account some platform characteristics beyond the most classical ones, namely the computing speed of processors and accelerators, and the communication bandwidth of network links. Our work mainly considers the following two platform characteristics:

Energy consumption. Power management in HPC is necessary due to both monetary and environmental constraints. Using dynamic voltage and frequency scaling (DVFS) is a widely used technique to decrease energy consumption, but it can severely degrade performance and increase execution time. Part of our work in this direction studies the trade-off between energy consumption and performance (throughput or execution time). Furthermore, our work also focuses on the optimization of the power consumption of fault-tolerant mechanisms. The problem of the energy consumption of these mechanisms is especially important because resilience generally requires redundant computations and/or redundant communications, either in time (re-execution) or in space (replication), and because redundancy consumes extra energy.

Memory usage and data movement. In many scientific computations, memory is a bottleneck and should be carefully considered. Besides, data movements, between main memory and secondary storages (I/Os) or between different computing nodes (communications), are taking an increasing part of the cost of computing, both in term of performance and energy consumption. In this context, our work focuses on scheduling scientific applications described as task graphs both on memory constrained platforms, and on distributed platforms with the objective of minimizing communications. The task-based representation of a computing application is very common in the scheduling literature but meets an increasing interest in the HPC field thanks to the use of runtime schedulers. Our work on memory-aware scheduling is naturally multi-criteria, as it is concerned with both memory consumption, performance and data-movements.

3.3 Sparse direct solvers and sparsity in computing

In this theme, we work on various aspects of sparse direct solvers for linear systems. Target applications lead to sparse systems made of millions of unknowns. In the scope of the PaStiX solver, co-developed with the Inria HiePACS team, there are two main objectives: reducing as much as possible memory requirements and exploiting modern parallel architectures through the use of runtime systems.

A first research challenge is to exploit the parallelism of modern computers, made of heterogeneous (CPUs+GPUs) nodes. The approach consists of using dynamic runtime systems (in the context of the PASTIX solver, PARSEC or STARPU) to schedule tasks.

Another important direction of research is the exploitation of low-rank representations. Low-rank approximations are commonly used to compress the representation of data structures. The loss of information induced is often negligible and can be controlled. In the context of sparse direct solvers, we exploit the notion of low-rank properties in order to reduce the demand in terms of floating-point operations and memory usage. To enhance sparse direct solvers using low-rank compression, two orthogonal approaches are followed: (i) integrate new strategies for a better scalability and (ii) use preprocessing steps to better identify how to cluster unknowns, when to perform compression and which blocks not to compress.

CSC is a term (coined circa 2002) for interdisciplinary research at the intersection of discrete mathematics, computer science, and scientific computing. In particular, it refers to the development, application, and analysis of combinatorial algorithms to enable scientific computing applications. CSC's deepest roots are in the realm of direct methods for solving sparse linear systems of equations where graph theoretical models have been central to the exploitation of sparsity, since the 1960s. The general approach is to identify performance issues in a scientific computing problem, such as memory use, parallel speed up, and/or the rate of convergence of a method, and to develop combinatorial algorithms and models to tackle those issues. Most of the time, the research output includes experiments with real life data to validate the developed combinatorial algorithms and fine tune them.

In this context, our work targets (i) the preprocessing phases of direct methods, iterative methods, and hybrid methods for solving linear systems of equations; (ii) high performance tensor computations. The core topics covering our contributions include partitioning and clustering in graphs and hypergraphs, matching in graphs, data structures and algorithms for sparse matrices and tensors (different from partitioning), and task mapping and scheduling.

4 Application domains

Sparse linear system solvers have a wide range of applications as they are used at the heart of many numerical methods in computational science: whether a model uses finite elements or finite differences, or requires the optimization of a complex linear or nonlinear function, one often ends up solving a system of linear equations involving sparse matrices. There are therefore a number of application fields: structural mechanics, seismic modeling, biomechanics, medical image processing, tomography, geophysics, electromagnetism, fluid dynamics, econometric models, oil reservoir simulation, magneto-hydro-dynamics, chemistry, acoustics, glaciology, astrophysics, circuit simulation, and work on hybrid direct-iterative methods.

Tensors, or multidimensional arrays, are becoming very important because of their use in many data analysis applications. The additional dimensions over matrices (or two dimensional arrays) enable gleaning information that is otherwise unreachable. Tensors, like matrices, come in two flavors: dense tensors and sparse tensors. Dense tensors arise usually in physical and simulation applications: signal processing for electroencephalography (also named EEG, electrophysiological monitoring method to record electrical activity of the brain); hyperspectral image analysis; compression of large grid-structured data coming from a high-fidelity computational simulation; quantum chemistry etc. Dense tensors also arise in a variety of statistical and data science applications. Some of the cited applications have structured sparsity in the tensors. We see sparse tensors, with no apparent/special structure, in data analysis and network science applications. Well known applications dealing with sparse tensors are: recommender systems; computer network traffic analysis for intrusion and anomaly detection; clustering in graphs and hypergraphs modeling various relations; knowledge graphs/bases such as those in learning natural languages.

5 Highlights of the year

5.1 Awards

Anne Benoit has been elected a Senior Member of Institut Universitaire de France.

6 New software, platforms, open data

6.1 New software

6.1.1 MatchMaker

Name: Maximum matchings in bipartite graphs

Keywords: Graph algorithmics, Matching

Scientific Description: The implementations of ten exact algorithms and four heuristics for solving the problem of finding a maximum cardinality matching in bipartite graphs are provided.

Functional Description: This software provides algorithms to solve the maximum cardinality matching problem in bipartite graphs.

URL: https://gitlab.inria.fr/bora-ucar/matchmaker

Publications: hal-00786548, hal-00763920

Contact: Bora Uçar

Participants: Kamer Kaya, Johannes Langguth

6.1.2 PaStiX

Name: Parallel Sparse matriX package

Keywords: Direct solvers, Parallel numerical solvers, Linear Systems Solver

Scientific Description: PaStiX is based on an efficient static scheduling and memory manager, in order to solve 3D problems with more than 50 million of unknowns. The mapping and scheduling algorithm handles a combination of 1D and 2D block distributions. A dynamic scheduling can also be applied to take care of NUMA architectures while taking into account very precisely the computational costs of the BLAS 3 primitives, the communication costs and the cost of local aggregations.

Functional Description: PaStiX is a scientific library that provides a high performance parallel solver for very large sparse linear systems based on block direct and block ILU(k) methods. It can handle low-rank compression techniques to reduce the computation and the memory complexity. Numerical algorithms are implemented in single or double precision (real or complex) for LLt, LDLt and LU factorization with static pivoting (for non symmetric matrices having a symmetric pattern). The PaStiX library uses the graph partitioning and sparse matrix block ordering packages Scotch or Metis.

The PaStiX solver is suitable for any heterogeneous parallel/distributed architecture when its performance is predictable, such as clusters of multicore nodes with GPU accelerators or KNL processors. In particular, we provide a high-performance version with a low memory overhead for multicore node architectures, which fully exploits the advantage of shared memory by using a hybrid MPI-thread implementation.

The solver also provides some low-rank compression methods to reduce the memory footprint and/or the time-to-solution.

URL: https://gitlab.inria.fr/solverstack/pastix

Publications: inria-00346017, inria-00346018, hal-01485507, hal-01824275, hal-03361299

Contact: Pierre Ramet

Participants: Alycia Lisito, Grégoire Pichon, Mathieu Faverge, Pierre Ramet

7 New results

7.1 Resilience for very large scale platforms

The ROMA team has been working on resilience problems for several years. In 2023, we have focused on several problems.

7.1.1 Checkpointing à la Young/Daly.

Participants: Anne Benoit, Leonardo Bautista-Gomez *(Barcelona Supercomputing Center, Spain)*, Sheng Di *(Argonne National Laboratory, USA)*, Thomas Herault *(University of Tennessee, Knoxville, USA)*, Yves Robert, Hongyang Sun *(University of Kansas, USA)*.

The Young/Daly formula provides an approximation of the optimal checkpoint period for a parallel application executing on a supercomputing platform. The Young/Daly formula was originally designed for preemptible tightly-coupled applications. In an invited publication at IC3 2022, we had provided some background and various application scenarios to assess the usefulness and limitations of the formula. In 2023, we have considerably extended the scope of our survey and we have submitted this contribution to the special issue of FGCS scheduled for 2024 and which will focus on JLESC collaboration results.

7.1.2 When to checkpoint at the end of a fixed-length reservation?

Participants: Anne Benoit, Quentin Barbut, Thomas Herault *(University of Tennessee, Knoxville, USA)*, Yves Robert, Frédéric Vivien.

This work considers an application executing for a fixed duration, namely the length of the reservation that it has been granted. The checkpoint duration is a stochastic random variable that obeys some well-known probability distribution law. The question is when to take a checkpoint towards the end of the execution, so that the expectation of the work done is maximized. We address two scenarios. In the first scenario, a checkpoint can be taken at any time; despite its simplicity, this natural problem has not been considered yet (to the best of our knowledge). We provide the optimal solution for a variety of probability distribution laws modeling checkpoint duration. The second scenario is more involved: the application is a linear workflow consisting of a chain of tasks with IID stochastic execution times, and a checkpoint can be taken only at the end of a task. First, we introduce a static strategy where we compute the optimal number of tasks before the application checkpoints at the beginning of the execution. Then, we design a dynamic strategy that decides whether to checkpoint or to continue executing at the end of each task. We instantiate this second scenario with several examples of probability distribution laws for task durations.

This work has been published in FTXS'2023, a workshop co-located with SC'2023 [17].

7.1.3 Checkpointing strategies to tolerate non-memoryless failures on HPC platforms

Participants: Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien.

This work studies checkpointing strategies for parallel applications subject to failures. The optimal strategy to minimize total execution time, or makespan, is well known when failure inter-arrival times obey an Exponential distribution, but it is unknown for non-memoryless failure distributions. We explain why the latter fact is misunderstood in recent literature. We propose a general strategy that maximizes the expected efficiency until the next failure, and we show that this strategy achieves an asymptotically optimal makespan, thereby establishing the first optimality result for arbitrary failure distributions. Through extensive simulations, we show that the new strategy is always at least as good as the Young/Daly strategy for various failure distributions. For distributions with a high infant mortality (such as LogNormal with shape parameter $k = 2.51$ or Weibull with shape parameter 0.5), the execution time is divided by a factor 1.9 on average, and up to a factor 4.2 for recently deployed platforms.

This work has been published in the ACM TOPC journal [8].

7.2 Multi-criteria scheduling strategies

We report here the work undertaken by the ROMA team in multi-criteria strategies, which focuses on taking into account energy and memory constraints, but also budget constraints or specific constraints for scheduling online requests.

7.2.1 Risk-aware scheduling algorithms for variable capacity resources

Participants: Anne Benoit, Andrew A. Chien *(University of Chicago)*, Lucas Perotin, Yves Robert, Rajini Wijayawardana *(University of Chicago)*, Chaojie Zhang *(Microsoft Research)*.

The drive to decarbonize the power grid to slow the pace of climate change has caused dramatic variation in the cost, availability, and carbon-intensity of power. This has begun to shape the planning and operation of datacenters. This work focuses on the design of scheduling algorithms for independent jobs that are submitted to a platform whose resource capacity varies over time. Jobs are submitted online and assigned on a target machine by the scheduler, which is agnostic to the rate and amount of resource variation. The optimization objective is the goodput, defined as the fraction of time devoted to effective computations (re-execution does not count). We introduce several novel algorithms that: (i) decide which fraction of the resources can be used safely; (ii) maintain a risk index associated to each machine; and (iii) achieves a global load balance while mapping longer jobs to safer machines. We assess the performance of these algorithms using one set of actual workflow traces together with three sets of synthetic traces. The goodput achieved by our algorithms increases up to 10% compared to standard first-fit approaches, while we never experience any loss in complementary metrics such as the maximum or average stretch.

This work has been published in PMBS'2023, a workshop co-located with SC'2023 [22].

7.2.2 Concealing compression-accelerated I/O for HPC applications through In Situ task scheduling

Participants: Franck Cappello *(Argonne National Laboratory)*, Sheng Di *(Argonne National Laboratory)*, Sian Jin *(Indiana University)*, Yves Robert, Dingwen Tao *(Indiana University)*, Frédéric Vivien, Daoce Wang *(Indiana University)*.

Lossy compression and asynchronous I/O are two of the most effective solutions for reducing storage overhead and enhancing I/O performance in large-scale high-performance computing (HPC) applications. However, current approaches have limitations that prevent them from fully leveraging lossy compression, and they may also result in task collisions, which restrict the overall performance of HPC applications. To address these issues, we propose an optimization approach for the task scheduling problem that encompasses computation, compression, and I/O. Our algorithm adaptively selects the optimal compression and I/O queue to minimize the performance degradation of the computation. We also introduce an intra-node I/O workload balancing mechanism that evenly distributes the workload across

different processes. Additionally, we design a framework that incorporates fine-grained compression, a compressed data buffer, and a shared Huffman tree to fully benefit from our proposed task scheduling. Experimental results with up to 16 nodes and 64 GPUs from ORNL Summit, as well as real-world HPC applications, demonstrate that our solution reduces I/O overhead by up to 3.8×× and 2.6×× compared to non-compression and asynchronous I/O solutions, respectively.

This work will appear in EuroSys'2024 [20].

7.2.3 Energy-aware mapping and scheduling strategies for real-time workflows under reliability constraints

Participants: Li Han *(East China Normal University)*, Jing Liu *(East China Normal University)*, Yves Robert, Frédéric Vivien, Zhiwei Wu *(East China Normal University)*.

This work focuses on energy minimization for the mapping and scheduling of real-time workflows under reliability constraints. Workflow instances are input periodically to the system. Each instance is composed of several tasks and must complete execution before the arrival of the next instance, and with a prescribed reliability threshold. While the shape of the dependence graph is identical for each instance, task execution times are stochastic and vary from one instance to the next. The reliability threshold is met by executing several replicas for each task. The target platform consists of identical processors equipped with Dynamic Voltage and Frequency Scaling (DVFS) capabilities. A different frequency can be assigned to each task replica to save energy, but it may have negative effect on the deadline and reliability target.

This difficult tri-criteria mapping and scheduling problem (energy, deadline, reliability) has been studied only recently for workflows with arbitrary dependence constraints. We investigate new mapping and scheduling strategies based upon layers in the task graph. These strategies better balance replicas across processors, thereby decreasing the time overlap between different replicas of a given task, and saving energy. We compare these strategies with two state-of-the-art approaches and a reference baseline on a variety of benchmark workflows. Our best heuristics achieve an average energy gain of 60% over the competitors and of 82% over the baseline.

This work has been published in the JPDC journal [15].

7.2.4 Asymptotic Performance and Energy Consumption of SLACK

Participants: Anne Benoit, Redouane Elghazi, Louis-Claude Canon *(Univ. Besançon)*, Pierre-Cyrille Héam *(Univ. Besançon)*.

Scheduling n independent tasks onto m identical processors in order to minimize the makespan has been widely studied. As an alternative to classical heuristics, the SLACK algorithm groups tasks by packs of m tasks of similar execution times, and schedules first the packs with the largest differences. It turns out to be very performant in practice, but only few studies have been conducted on its theoretical properties. We derive novel analytical results for SLACK, and in particular, we study the performance of this algorithm from an asymptotical point of view, under the assumption that the execution times of the tasks follow a given probability distribution. The study is building on a comparison of the most heavily loaded machine compared to the least loaded one. Furthermore, we extend the results when the objective is to minimize the energy consumption rather than the makespan, since reducing the energy consumption of the computing centers is an ever-growing concern for economical and ecological reasons. Finally, we perform extensive simulations to empirically assess the performance of the algorithms with both synthetic and realistic execution time distributions.

This work appeared in the proceedings of EuroPar 2023 [18]

7.2.5 Mapping Tree-shaped Workflows on Systems with Different Memory Sizes and Processor Speeds

Participants: Svetlana Kulagina *(Humboldt University of Berlin)*, Henning Meyerhenke *(Humboldt University of Berlin)*, Anne Benoit.

Directed acyclic graphs are commonly used to model scientific workflows, by expressing dependencies between tasks, as well as the resource requirements of the workflow. As a special case, rooted directed trees occur in several applications, for instance in sparse matrix computations. Since typical workflows are modeled by large trees, it is crucial to schedule them efficiently, so that their execution time (or makespan) is minimized. Furthermore, it is usually beneficial to distribute the execution on several compute nodes, hence increasing the available memory, and allowing us to parallelize parts of the execution. To exploit the heterogeneity of modern clusters in this context, we investigate the partitioning and mapping of tree-shaped workflows on two types of target architecture models: in AM1, each processor can have a different memory size, and in AM2, each processor can also have a different speed (in addition to a different memory size). We design a three-step heuristic for AM1, which adapts and extends previous work for homogeneous clusters. The changes we propose concern the assignment to processors (accounting for the different memory sizes) and the availability of suitable processors when splitting or merging subtrees. For AM2, we extend the heuristic for AM1 with a two-phase local search approach. Phase A is a swap-based hill climber, while (the optional) Phase B is inspired by iterated local search. We evaluate our heuristics for AM1 and AM2 with extensive simulations, and we demonstrate that exploiting the heterogeneity in the cluster significantly reduces the makespan, compared to the state of the art for homogeneous processors.

This is the extension of a previous work published in 2022 in the HeteroPar workshop, in conjunction with EuroPar, where we were targeting only AM1 (same-speed processors). It has been published in CCPE [13].

7.2.6 Resource-Constrained Scheduling Algorithms for Stochastic Independent Tasks With Unknown Probability Distribution

Participants: Yiqin Gao *(Shangaï Jiao Tong University)*, Yves Robert, Frédéric Vivien.

This work introduces scheduling algorithms to maximize the expected number of independent tasks that can be executed on a parallel platform within a given budget and under a deadline constraint. The main motivation for this problem comes from imprecise computations, where each job has a mandatory part and an optional part, and the objective is to maximize the number of optional parts that are successfully executed, in order to improve the accuracy of the results. The optional parts of the jobs represent the independent tasks of our problem. Task execution times are not known before execution; instead, the only information available to the scheduler is that they obey some (unknown) probability distribution. The scheduler needs to acquire some information before deciding for a cutting threshold: instead of allowing all tasks to run until completion, one may want to interrupt long-running tasks at some point. In addition, the cutting threshold may be reevaluated as new information is acquired when the execution progresses further. This work presents several algorithms to determine a good cutting threshold, and to decide when to re-evaluate it. In particular, we use the Kaplan-Meier estimator to account for tasks that are still running when making a decision. The efficiency of our algorithms is assessed through an extensive set of simulations with various budget and deadline values, and ranging over 13 probability Scheduling Stochastic Tasks With Unknown Probability Distribution distributions. In particular, the AutoPerSurvival(40%,0.005) strategy is proved to have a performance of 77% compared to the upper bound even in the worst case. This shows the robustness of our strategy.

This work was published in the journal *Algorithmica* [11].

7.2.7 Scheduling requests in replicated key-value stores

Key-value stores distribute data across several storage nodes to handle large amounts of parallel requests. Proper scheduling of these requests impacts the quality of service, as measured by achievable throughput

and (tail) latencies. In addition to scheduling, performance heavily depends on the nature of the workload and the deployment environment. It is, unfortunately, difficult to evaluate different scheduling strategies consistently under the same operational conditions. Moreover, such strategies are often hard-coded in the system, limiting flexibility. We have proposed Hector, a modular framework for implementing and evaluating scheduling policies in Apache Cassandra. Hector enables users to select among several options for key components of the scheduling workflow, from the request propagation via replica selection to the local ordering of incoming requests at a storage node. We have demonstrated the capabilities of Hector by comparing strategies in various settings. For example, we found that leveraging cache locality effects may be of particular interest: we proposed a new replica selection strategy, called Popularity-Aware, that can support 6 times the maximum throughput of the default algorithm under specific key access patterns. We have also shown that local scheduling policies have a significant effect when parallelism at each storage node is limited.

This work was presented at the ICPP 2023 conference [19].

7.2.8 Scheduling task graphs in runtime systems for data locality

Hardware accelerators, such as GPUs, now provide a large part of the computational power used for scientific simulations. GPUs come with their own (limited) memory and are connected to the main memory of the machine via a bus with limited bandwidth. Scientific simulations often operate on very large data, to the point of not fitting in the limited GPU memory. In this case, one has to turn to out-of-core computing: data are kept in the CPU memory, and moved back and forth to the GPU memory when needed for the computation. This out-of-core situation also happens when processing on multicore CPUs with limited memory huge datasets stored on disk. In both cases, data movement quickly becomes a performance bottleneck. Task-based runtime schedulers have emerged as a convenient and efficient way to manage large applications on such heterogeneous platforms. They are in charge of choosing which tasks to assign on which processing unit and in which order they should be processed. In this work, we have focused on this problem of scheduling for a task-based runtime to improve data locality in an out-of-core setting, in order to reduce data movements. We have designed a data-aware strategy for both task scheduling and data eviction from limited memories. We have compared this to existing scheduling techniques in runtime systems. Using the StarPU runtime, we have shown that our strategy achieves significantly better performance when scheduling tasks on multiple GPUs with limited memory, as well as on multiple CPU cores with limited main memory.

This work has been submitted for publication and is available as a research report [30]. A preliminary version of this work has been published in the journal FGCS [12].

7.3 Sparse direct solvers and sparsity in computing

We continued our work on the optimization of sparse solvers by concentrating on data locality when mapping tasks to processors, and by studying the tradeoff between memory and performance when using low-rank compression. We worked on combinatorial problems arising in sparse matrix and tensors computations. The computations involved direct methods for solving sparse linear systems and tensor factorizations. The combinatorial problems were based on matchings on bipartite graphs, partitionings, and hyperedge queries.

7.3.1 Engineering fast algorithms for the bottleneck matching problem

Participants: Grégoire Pichon, Somesh Singh, Bora Uçar.

We investigate the maximum bottleneck matching problem in bipartite graphs. Given a bipartite graph with nonnegative edge weights, the problem is to determine a maximum cardinality matching in which the minimum weight of an edge is the maximum. To the best of our knowledge, there are two widely used solvers for this problem based on two different approaches. There exists a third known approach in the literature, which seems inferior to those two which is presumably why there is no implementation of

it. We take this third approach, make theoretical observations to improve its behavior, and implement the improved method. Experiments with the existing two solvers show that their run time can be too high to be useful in many interesting cases. Furthermore, their performance is not predictable, and slight perturbations of the input graph lead to considerable changes in the run time. On the other hand, the proposed solver's performance is much more stable; it is almost always faster than or comparable to the two existing solvers, and its run time always remains low. This work is published at a conference [21], and the codes are made available online with the CeCILL-B license.

7.3.2 Open problems in (hyper)graph partitioning

Participants: Bora Uçar.

We have worked on a conjecture of ours about partitioning 5-point stencils for perfect balance and minimum volume of communication, should the associated matrix undergoes matrix-vector multiplication in a distributed memory parallel computing. A month-long experimentation with Gurobi solver confirmed that the conjecture holds, however non-computational proofs are still needed. We stated this in a Dagstuhl report [26].

7.3.3 Parallel algorithms for dense computations

We also worked on the design of parallel and communication optimal algorithms for several dense matrix and tensor computations.

7.3.4 Parallel Memory-Independent Communication Bounds for SYRK

Participants: Hussam Al Daas *(Rutherford Appleton Laboratory, UK)*, Grey Ballard *(Wake Forest University, USA)*, Laura Grigori *(EPFL, Switzerland)*, Suraj Kumar, Kathryn Rouse *(Inmar Intelligence, USA)*.

In this work, we focus on the parallel communication cost of multiplying a matrix with its transpose, known as a symmetric rank-k update (SYRK). SYRK requires half the computation of general matrix multiplication because of the symmetry of the output matrix. Recent work (Beaumont et al., SPAA '22) has demonstrated that the sequential I/O complexity of SYRK is also a constant factor smaller than that of general matrix multiplication. Inspired by this progress, we establish memory-independent parallel communication lower bounds for SYRK with smaller constants than general matrix multiplication, and we show that these constants are tight by presenting communication-optimal algorithms. The crux of the lower bound proof relies on extending a key geometric inequality to symmetric computations and analytically solving a constrained nonlinear optimization problem. The optimal algorithms use a triangular blocking scheme for parallel distribution of the symmetric output matrix and corresponding computation.

This work has been published in SPAA 2023 [16].

7.3.5 Communication Lower Bounds and Optimal Algorithms for Multiple Tensor-Times-Matrix Computation

Participants: Hussam Al Daas *(Rutherford Appleton Laboratory, UK)*, Grey Ballard *(Wake Forest University, USA)*, Laura Grigori *(EPFL, Switzerland)*, Suraj Kumar, Kathryn Rouse *(Inmar Intelligence, USA)*.

Multiple Tensor-Times-Matrix (Multi-TTM) is a key computation in algorithms for computing and operating with the Tucker tensor decomposition, which is frequently used in multidimensional data

analysis. We establish communication lower bounds that determine how much data movement is required (under mild conditions) to perform the Multi-TTM computation in parallel. The crux of the proof relies on analytically solving a constrained, nonlinear optimization problem. We also present a parallel algorithm to perform this computation that organizes the processors into a logical grid with twice as many modes as the input tensor. We show that with correct choices of grid dimensions, the communication cost of the algorithm attains the lower bounds and is therefore communication optimal. Finally, we show that our algorithm can significantly reduce communication compared to the straightforward approach of expressing the computation as a sequence of tensor-times-matrix operations when the input and output tensors vary greatly in size.

This work will appear in the SIMAX journal [9].

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

2023 was the second year of the CHALRESIL associate team. CHALRESIL stands for *Challenges in resilience at scale* and is operated between ROMA (PI Yves Robert) and the Innovative Computing Laboratory of the University of Tennessee Knoxville, USA (PI Thomas Herault). Many fundamental challenges in the resilience field have yet to be addressed, and CHALRESIL focuses on some critical ones.

The year 2023 was quite productive for ChalResil. In addition to the several joint results reported elsewhere in this document, we have organized the 16th Scheduling for Large Scale Systems Workshop, which was held at The University of Tennessee in Knoxville, May 22-24, 2023. The workshop was organized by George Bosilca (UTK) and Yves Robert (Inria). The other three permanent members of ChalResil participated in the workshop, and gave a presentation. There was a total 25 participants, out of them 5 from Inria. Further details can be found on the workshop webpage online.

8.1.2 Inria associate team not involved in an IIL or an international program

The MODS associate team has started in 2023. MODS stands for *Match and Order: improving direct solvers for cardiac simulations* and is operated between ROMA (PI Grégoire Pichon) and SIMULA Laboratory, Oslo, Norway (PI Johannes Langguth). The goal of the MODS project is to enhance robustness, scalability, and performance of sparse direct solvers by developing novel parallel matching and ordering algorithms. The results will be tested on and applied to simulations of cardiac electrophysiology developed by SIMULA.

During the year 2023, Grégoire Pichon and Bora Uçar (ROMA) have visited SIMULA during a week in September. A visit from three SIMULA member is planned for January 2024. Further details on this associate team can be found online.

The PEACHTREE associated team online has reached to its completion. The work was carried out locally by Somesh Singh and Bora Uçar, as the partner was not available.

8.1.3 Participation in other International Programs

Homeland Homeland project is funded by PHC Procope programme. This project investigates various problems related graphs and hypergraphs (such as clustering, streaming partitioning, orientation, maximum independent set).

Participants: Uçar Bora.

Title: Homeland: Heidelberg & Lyon do machine learning for graph decomposition

Partner Institution(s):

- Heidelberg University, Germany

Date/Duration: 2022–2023.

Collaboration with U. Chicago

Participants: Benoit Anne, Cendrier Joachim, Robert Yves, Vivien Frédéric.

- 2022-2024: FACCTS research collaboration with A. Chien at U. Chicago: Foundational Models and Efficient Algorithms for Scheduling with Variable Capacity Resources, funded by the France Chicago Center (see website).
- 2023-2024: Additional funding as part of our collaboration with A. Chien to organize two workshops on the topic Bridging Communities — Scheduling Variable Capacity Resources for Sustainability, in the U. Chicago center in Paris (2023-24). These workshops are funded by the International Institute of Research in Paris. The first workshop was organized in March 2023, see website.
- 2023-2026: U. Chicago-CNRS collaboration on efficient and environment-friendly scheduling and resource management algorithms at the edge: Funding secured for the PhD thesis of Joachim Cendrier.

8.2 International research visitors

8.2.1 Visits of international scientists

Inria International Chair

Participants: Julien Langou.

Julien Langou, professor at the University Denver (USA) has been awarded an Inria International Chair to visit the ROMA team in the period 2023–2026. He spent 1.5 months in the team in June-July 2023 to start collaborations with researchers in ROMA.

8.3 National initiatives

8.3.1 ANR Project SOLHARIS (2019-2023), 4 years.

Participants: Maxime Gonthier, Grégoire Pichon, Loris Marchal, Bora Uçar.

The ANR Project SOLHARIS was launched in November 2019, for a duration of 48 months. It gathers five academic partners (the HiePACS, ROMA, RealOpt, STORM and TADAAM INRIA project-teams, and CNRS-IRIT) and two industrial partners (CEA/CESTA and Airbus CRT). This project aims at producing scalable methods for direct methods for the solution of sparse linear systems on large scale and heterogeneous computing platforms, based on task-based runtime systems.

The proposed research is organized along three distinct research thrusts. The first objective deals with the development of scalable linear algebra solvers on task-based runtimes. The second one focuses on the deployment of runtime systems on large-scale heterogeneous platforms. The last one is concerned with scheduling these particular applications on a heterogeneous and large-scale environment.

8.3.2 ANR Project SPARTACLUS (2023-2027), 4 years.

Participants: Loris Marchal, Grégoire Pichon, Bora Uçar, Frédéric Vivien.

The ANR Project SPARTACLUS was launched in January 2023 for a duration of 48 months. This is a JCJC project lead by Grégoire Pichon and including other participants of the ROMA team. This project aims at building new ordering strategies to enhance the behavior of sparse direct solvers using low-rank compression.

The objective of this project is to end up with a common tool to perform the ordering and the clustering for sparse direct solvers when using low-rank compression. We will provide statistics that are currently missing and that will help understanding the compressibility of each block. The objective is to enhance sparse direct solvers, in particular targeting larger problems. The benefits will directly apply to academic or industrial applications using sparse direct solvers.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

Workshop on Scheduling Variable Capacity Resources for Sustainability Anne Benoit, Andrew A. Chien (University of Chicago and Argonne National Laboratory, USA) and Yves Robert have co-organized a workshop in the Paris Center of the University of Chicago, on March 29-31, 2023. This workshop gathered scheduling leaders from the United States, France, Europe, and Asia to discuss the research challenges facing the scheduling community, arising from the increasing fluctuations of renewable energy in the power grid. In such an environment, scheduling to reduce carbon-emissions, to reduce power cost, help decarbonize the power grid, or just to stabilize the grid all requires scheduling with awareness of variable capacity.

The Workshop Report outlining the research challenges documented by this group of researchers has been completed, and is available online. The full workshop notes, including position papers from all attendees, as well as workshop organization, are available at online.

16th Workshop on Scheduling for Large Scale Systems George Bosilca (University Tennessee Knoxville, USA) and Yves Robert have organized the 16th Workshop on Scheduling for Large Scale Systems in Knoxville in May 2023. Further details can be found on the workshop webpage.

Member of the organizing committees

- Anne Benoit was a member of the organizing committee of SIAM ACDA 2023.
- Bora Uçar organized Minisymposia at ICIAM 2023 on Applied and Computational Discrete Algorithms.

9.1.2 Scientific events: selection

Chair of conference program committees

- Anne Benoit was Program area co-chair (parallel and distributed algorithms for computational science) of IPDPS'2023

Member of the conference program committees

- Anne Benoit was a member of the program committee of EuroPar'23, and of the workshops program committee of SC'23.
- Loris Marchal wa a membre of the program committee of EuroPar'23.

- Grégoire Pichon was a member of the program committee of research posters for SC'23.
- Yves Robert was a member of the program committees of FTXS'23, SCALA'23, PMBS'23, SuperCheck'23 (co-located with SC'23) and Resilience (co-located with Euro-Par'23).
- Bora Uçar served in the PC of IPDPS24, SIAM PP2024, HPC Asia 2024, CCGrid 23, SC'23, IPDPS'23.
- Frédéric Vivien was a member of the program committees of IEEE EuroPar'23, IEEE BigData 2023, IPDPS'23 and SC'23.

Reviewer

- Suraj Kumar has reviewed papers for the conference SPAA 2023, 35th ACM Symposium on Parallelism in Algorithms and Architectures.
- Bora Uçar has reviewed papers for SODA 2024 and PPoPP2023.

9.1.3 Journal

Member of the editorial boards

- Anne Benoit was Associate Editor in Chief of Elsevier JPDC, the Journal of Parallel and Distributed Computing, and of Elsevier ParCo, the journal of Parallel Computing: Systems and Applications.
- Yves Robert is a member of the editorial board of the International Journal of High Performance Computing (IJHPCA) and the Journal of Computational Science (JOCS).
- Bora Uçar is a member of the editorial board of IEEE Transactions on Parallel and Distributed Systems (IEEE TPDS); IEEE TPDS Reproducibility Editorial Board; SIAM Journal on Scientific Computing (SISC); SIAM Journal on Matrix Analysis and Applications (SIMAX); Parallel Computing.
- Frédéric Vivien is a member of the editorial board of the Journal of Parallel and Distributed Computing and of ACM Transactions on Parallel Computing.

Reviewer - reviewing activities

- Suraj Kumar has reviewed manuscripts for the journals: Transactions on Parallel and Distributed Systems, Journal of Parallel and Distributed Computing.
- Grégoire Pichon has reviewed manuscripts for the journals: Journal of Parallel and Distributed Computing, Transactions on Mathematical Software.
- Bora Uçar has reviewed papers for ACM TOMS (2x), IJHPC (2x).

9.1.4 Invited talks

- Yves Robert has given a keynote talk at the 2023 IEEE CloudNet conference in Hoboken, NJ, USA.

9.1.5 Leadership within the scientific community

- Anne Benoit is the IEEE TCPP Chair (IEEE Technical Community on Parallel Processing).
- Yves Robert serves in the steering committee of IPDPS and HCW.
- Bora Uçar is the program director of the SIAM Activity Group on Applied and Computational Discrete Algorithms (ACDA). He is also in a member of the steering committee of SEA.

9.1.6 Scientific expertise

- Loris Marchal has evaluated a project proposal for the Natural Sciences and Engineering Research Council of Canada.
- Yves Robert was the Chair of the 2023 IEEE Charles Babbage Award Committee.
- Bora Uçar was a member of the prize committee of SIAM's George Polya Prize in Applied Combinatorics. He was also in the 2023 and 2022 IEEE TCPP Outstanding Service and Contributions Award Committee.
- Bora Uçar has evaluated a project proposal for Israeli Ministry of Innovation, Science and Technology.
- Frédéric Vivien is an elected member of the scientific council of the École normale supérieure de Lyon.
- Frédéric Vivien is an elected member of INRIA *commision d'évaluation*.
- Frédéric Vivien is a member of the scientific council of the IRMIA labex.

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

- Anne Benoit, Chair of the Computer Science department at ENS Lyon, France, up to Sep 30, 2023
- Yves Robert, Chair of the Computer Science department at ENS Lyon, France, from Oct 1, 2023
- Licence: Anne Benoit, Algorithmique avancée, 48h, L3, ENS Lyon, France
- Master: Anne Benoit, Parallel and Distributed Algorithms and Programs, 42h, M1, ENS Lyon, France
- Master: Suraj Kumar, Data-aware algorithms for matrix and tensor computations, 26h, M2, ENS Lyon, France.
- Master: Grégoire Pichon, Bibliographie, étude de cas, projet, certifications, 12h, M2, Univ. Lyon 1, France
- Master: Grégoire Pichon, Compilation / traduction des programmes, 22.5h, M1, Univ. Lyon 1, France
- Master: Grégoire Pichon, Systèmes avancés, 19.5h, M1, Univ. Lyon 1, France
- Master: Grégoire Pichon, Réseaux, 12h, M1, Univ. Lyon 1, France
- Licence: Grégoire Pichon, Programmation concurrente, 33h, L3, Univ. Lyon 1, France
- Licence: Grégoire Pichon, Réseaux, 36h, L3, Univ. Lyon 1, France
- Licence: Grégoire Pichon, Système d'exploitation, 27h, L2, Univ. Lyon 1, France
- Licence: Grégoire Pichon, Référent pédagogique, 30h, L1/L2/L3, Univ. Lyon 1, France
- Licence: Yves Robert, Probabilités et algorithmes randomisés, 48h, L3, ENS Lyon, France
- Agrégation Informatique: Yves Robert, Algorithmique, NP-complétude et algorithmes d'approximation, probabilités, graphes, structures de données, 75h, ENS Lyon, France
- Master: Frédéric Vivien, Parallel algorithms, 10h, M1/M2, ECNU, Shanghaï, Chine, 2023 (remote teaching).

9.2.2 Supervision

- Anne Benoit and Yves Robert were co-supervisor of the thesis of Lucas Perotin. Lucas defended on June 29, 2023, and he worked on scheduling algorithms to optimize the performance, energy consumption and robustness of HPC applications.
- Anne Benoit was a co-supervisor of the thesis of Redouane Elghazi, with Pierre-Cyrille Héam and Louis-Claude Canon. Redouane defended on October 9, 2023, and he worked on theoretical bounds for scheduling problems and their application to asymptotic analysis and energy consumption minimization.
- Anne Benoit is co-supervising the thesis of Svetlana Kulagina, with Henning Meyerhenke (Humboldt-Universitat zu Berlin, Germany). Svetlana works on scalable algorithms for decentralized scheduling and load balancing of large-scale scientific workflows. She is funded by the FONDA CRC project (Foundations of Workflows for Large-Scale Scientific Data Analysis).
- Anne Benoit and Frédéric Vivien are co-supervising the thesis of Joachim Cendrier. Joachim works on efficient and environment-friendly scheduling and resource management algorithms at the edge, in collaboration with Andrew Chien from U. Chicago. He is funded by a CNRS - U. Chicago project.
- Loris Marchal has co-supervised the thesis of Anthony Dugois, with Louis-Claude Canon. Anthony defended his thesis on September 28, 2023.
- Loris Marchal has co-supervised the thesis of Maxime Gonthier, with Samuel Thibault. Maxime defended his thesis on September 25, 2023.
- Bora Uçar was a co-supervisor of two Master degree students in Heidelberg University.

9.2.3 Juries

- Anne Benoit was a member of a COS for hiring an associate professor of CNU for Université Grenoble Alpes, 2023.
- Loris Marchal was a referee and a member of the PhD committee for the defense of Paweł Żuk, at the University of Warsaw, in October 2023.
- Loris Marchal was a referee and a member of the PhD committee for the defense of Anderson Andrei Da Silva, at the University of Grenoble, in December 2023.
- Bora Uçar was a member of a COS for hiring of a professor (Article recrutement number: 46.3), Section 27 (Informatique) of CNU for Université de Bordeaux, 2023.
- Bora Uçar was a rapporteur of the PhD committee of Hubert Hirtz, Paris-Saclay University, 12 December 2023; he was an examinateur of the PhD committee of Matthias Beaupère, Sorbonne University, 2023.

9.3 Popularization

- Yves Robert, together with George Bosilca, Aurélien Bouteiller and Thomas Herault, gave a full-day tutorial at SC'23 in Denver, CO, USA, on *Fault-tolerant techniques for HPC and Big Data: theory and practice*.
- Anne Benoit and Yves Robert have prepared a dissemination article on resilience and silent errors, in French, for the book on computing and architectures edited by Mokrane Bouzeghoub, Michel Daydé, Elisa Godet and Christian Jutten, to be published by CNRS editions in 2024.

10 Scientific production

10.1 Major publications

[1] A. Benoit, T. Hérault, V. Le Fèvre and Y. Robert. 'Replication Is More Efficient Than You Think'. In: *SC 2019 - International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'19)*. Denver, United States, Nov. 2019. URL: https://hal.inria.fr/hal-02273142.

[2] A. Benoit, L. Perotin, Y. Robert and F. Vivien. 'Checkpointing strategies to tolerate non-memoryless failures on HPC platforms'. In: *ACM Transactions on Parallel Computing* (Sept. 2023). DOI: 10.1145/3624560. URL: https://inria.hal.science/hal-04215283.

[3] M. Bougeret, H. Casanova, M. Rabie, Y. Robert and F. Vivien. 'Checkpointing strategies for parallel jobs.' In: *SuperComputing (SC) - International Conference for High Performance Computing, Networking, Storage and Analysis, 2011*. United States, 2011, pp. 1–11. URL: https://hal.archives-ouvertes.fr/hal-00738504.

[4] J. Dongarra, T. Hérault and Y. Robert. 'Fault Tolerance Techniques for High-Performance Computing'. In: *Fault-Tolerance Techniques for High-Performance Computing*. Ed. by T. Hérault and Y. Robert. Springer, May 2015, p. 83. URL: https://hal.inria.fr/hal-01200488.

[5] F. Dufossé and B. Uçar. 'Notes on Birkhoff-von Neumann decomposition of doubly stochastic matrices'. In: *Linear Algebra and its Applications* 497 (Feb. 2016), pp. 108–115. DOI: 10.1016/j.laa.2016.02.023. URL: https://hal.inria.fr/hal-01270331.

[6] L. Eyraud-Dubois, L. Marchal, O. Sinnen and F. Vivien. 'Parallel scheduling of task trees with limited memory'. In: *ACM Transactions on Parallel Computing* 2.2 (July 2015), p. 36. DOI: 10.1145/2779052. URL: https://hal.inria.fr/hal-01160118.

[7] L. Marchal, B. Simon and F. Vivien. 'Limiting the memory footprint when dynamically scheduling DAGs on shared-memory platforms'. In: *Journal of Parallel and Distributed Computing* 128 (Feb. 2019), pp. 30–42. DOI: 10.1016/j.jpdc.2019.01.009. URL: https://hal.inria.fr/hal-02025521.

10.2 Publications of the year

International journals

[8] A. Benoit, L. Perotin, Y. Robert and F. Vivien. 'Checkpointing strategies to tolerate non-memoryless failures on HPC platforms'. In: *ACM Transactions on Parallel Computing* (Sept. 2023). DOI: 10.1145/3624560. URL: https://inria.hal.science/hal-04215283.

[9] H. A. Daas, G. Ballard, L. Grigori, S. Kumar and K. Rouse. 'Communication Lower Bounds and Optimal Algorithms for Multiple Tensor-Times-Matrix Computation'. In: *SIAM Journal on Matrix Analysis and Applications* (4th Aug. 2023). URL: https://inria.hal.science/hal-03950359.

[10] Y. Gao, G. Pallez, Y. Robert and F. Vivien. 'Dynamic Scheduling Strategies for Firm Semi-Periodic Real-Time Tasks'. In: *IEEE Transactions on Computers* 72.1 (1st Jan. 2023), pp. 55–68. DOI: 10.1109/TC.2022.3208203. URL: https://inria.hal.science/hal-03778357.

[11] Y. Gao, Y. Robert and F. Vivien. 'Resource-Constrained Scheduling Algorithms for Stochastic Independent Tasks With Unknown Probability Distribution'. In: *Algorithmica* 85.8 (Aug. 2023), pp. 2363–2394. DOI: 10.1007/s00453-023-01100-8. URL: https://inria.hal.science/hal-04214825.

[12] M. Gonthier, L. Marchal and S. Thibault. 'Taming data locality for task scheduling under memory constraint in runtime systems'. In: *Future Generation Computer Systems* (2023). DOI: 10.1016/j.future.2023.01.024. URL: https://inria.hal.science/hal-03623220.

[13] S. Kulagina, H. Meyerhenke and A. Benoit. 'Mapping tree-shaped workflows on systems with different memory sizes and processor speeds'. In: *Concurrency and Computation: Practice and Experience* 35.25 (6th July 2023). DOI: 10.1002/cpe.7842. URL: https://inria.hal.science/hal-04397633.

[14] L. Marchal, S. McCauley, B. Simon and F. Vivien. 'Minimizing I/Os in Out-of-Core Task Tree Scheduling'. In: *International Journal of Foundations of Computer Science* 34.01 (Jan. 2023), pp. 51–80. DOI: 10.1142/s0129054122500186. URL: https://hal.science/hal-03758021.

[15] Z. Wu, L. Han, J. Liu, Y. Robert and F. Vivien. 'Energy-aware mapping and scheduling strategies for real-time workflows under reliability constraints'. In: *Journal of Parallel and Distributed Computing* 176 (June 2023), pp. 1–16. DOI: 10.1016/j.jpdc.2023.02.004. URL: https://inria.hal.science/hal-04214810.

International peer-reviewed conferences

[16] H. Al Daas, G. Ballard, L. Grigori, S. Kumar and K. Rouse. 'Parallel Memory-Independent Communication Bounds for SYRK'. In: *Proceedings of the 35th ACM Symposium on Parallelism in Algorithms and Architectures*. SPAA '23 - ACM Symposium on Parallelism in Algorithms and Architectures. Orlando, United States, 16th June 2023. DOI: 10.1145/3558481.3591072. URL: https://inria.hal.science/hal-04076513.

[17] Q. Barbut, A. Benoit, T. Herault, Y. Robert and F. Vivien. 'When to checkpoint at the end of a fixed-length reservation?' In: Fault Tolerance for HPC at eXtreme Scales (FTXS) Workshop. Denver, United States, 12th Nov. 2023. URL: https://inria.hal.science/hal-04215554.

[18] A. Benoit, L.-C. Canon, R. Elghazi and P.-C. Heam. 'Asymptotic Performance and Energy Consumption of SLACK'. In: Euro-Par. Vol. 14100. Lecture Notes in Computer Science. Limassol, Cyprus: Springer Nature Switzerland, 24th Aug. 2023, pp. 81–95. DOI: 10.1007/978-3-031-39698-4_6. URL: https://inria.hal.science/hal-04397726.

[19] L.-C. Canon, A. Dugois, L. Marchal and E. Rivière. 'Hector: A Framework to Design and Evaluate Scheduling Strategies in Persistent Key-Value Stores'. In: *ICPP '23: Proceedings of the 52nd International Conference on Parallel Processing*. ICPP 2023 - 52nd International Conference on Parallel Processing. Salt Lake City, United States, 7th Aug. 2023. URL: https://hal.science/hal-04158577.

[20] S. Jin, S. Di, F. Vivien, D. Wang, Y. Robert, D. Tao and F. Cappello. 'Concealing Compression-accelerated I/O for HPC Applications through In Situ Task Scheduling'. In: EuroSys 2024. Athens, Greece, 22nd Apr. 2024. URL: https://inria.hal.science/hal-04225758.

[21] I. Panagiotas, G. Pichon, S. Singh and B. Uçar. 'Engineering fast algorithms for the bottleneck matching problem'. In: ESA 2023 - The 31st Annual European Symposium on Algorithms. Amsterdam (Hollande), Netherlands, 29th June 2023. URL: https://inria.hal.science/hal-04146298.

[22] L. Perotin, C. Zhang, R. Wijayawardana, A. Benoit, Y. Robert and A. A. Chien. 'Risk-Aware Scheduling Algorithms for Variable Capacity Resources'. In: PMBS Workshop - SC-W 2023: Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis. Denver, CO, United States: ACM, Nov. 2023, pp. 1306–1315. DOI: 10.1145/3624062.3624194. URL: https://inria.hal.science/hal-04397574.

National peer-reviewed Conferences

[23] M. Gonthier. 'Exploiting data locality to maximize the performance of data-sharing tasksets'. In: ComPAS 2023 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Annecy, France, 4th July 2023. URL: https://inria.hal.science/hal-04090634.

Doctoral dissertations and habilitation theses

[24] A. Dugois. 'Scheduling in Distributed Storage Systems'. Ecole normale supérieure de lyon - ENS LYON, 28th Sept. 2023. URL: https://theses.hal.science/tel-04379876.

[25] M. Gonthier. 'Scheduling Under Memory Constraint in Task-based Runtime Systems'. Ecole normale supérieure de lyon - ENS LYON, 25th Sept. 2023. URL: https://theses.hal.science/tel-04260094.

Reports & preprints

[26] D. Ajwani, R. H. Bisseling, K. Casel, Ü. V. Çatalyürek, C. Chevalier, F. Chudigiewitsch, M. F. Faraj, M. Fellows, L. Gottesbüren, T. Heuer, G. Karypis, K. Kaya, J. Lacki, J. Langguth, X. S. Li, R. Mayer, J. Meintrup, Y. Mizutani, F. Pellegrini, F. Petrini, F. Rosamond, I. Safro, S. Schlag, C. Schulz, R. Sharma, D. Strash, B. D. Sullivan, B. Uçar and A.-J. Yzelman. *Open Problems in (Hyper)Graph Decomposition*. 18th Oct. 2023. DOI: 10.4230/LIPIcs. URL: https://hal.science/hal-04251953.

[27] A. Benoit, L.-C. Canon, R. Elghazi and P.-C. Heam. *Asymptotic Performance and Energy Consumption of SLACK*. RR-9501. Inria Lyon, May 2023, p. 27. URL: https://inria.hal.science/hal-04021482.

[28] A. Benoit, A. A. Chien and Y. Robert. *Scheduling Variable Capacity Resources for Sustainability Workshop*. ROMA (INRIA Rhône-Alpes / LIP Laboratoire de l'Informatique du Parallélisme); University of Chicago, July 2023. URL: https://inria.hal.science/hal-04159509.

[29] A. Benoit, T. Herault, L. Perotin, Y. Robert and F. Vivien. *Revisiting I/O bandwidth-sharing strategies for HPC applications*. RR-9502 v3. INRIA, Mar. 2023, p. 57. URL: https://inria.hal.science/hal-04038011.

[30] M. Gonthier, S. Thibault and L. Marchal. *A generic scheduler to foster data locality for GPU and out-of-core task-based applications*. 30th June 2023. DOI: 10.1145/nnnnnnn.nnnnnnn. URL: https://inria.hal.science/hal-04146714.

Other scientific publications

[31] M. Gonthier, L. Marchal and S. Thibault. 'Memory-Aware Scheduling Of Tasks Sharing Data On Multiple GPUs'. In: ISC 2023 - ISC High Performance 2023. Hamburg, Germany, 21st May 2023. URL: https://inria.hal.science/hal-04090595.

[32] M. Gonthier, L. Marchal and S. Thibault. *Memory-Aware Scheduling of Tasks Sharing Data on Multiple GPUs*. Bordeaux, France, 21st Mar. 2023. URL: https://inria.hal.science/hal-04090612.