

2025 Activity Report

RESEARCH CENTRE: Inria Lyon Centre

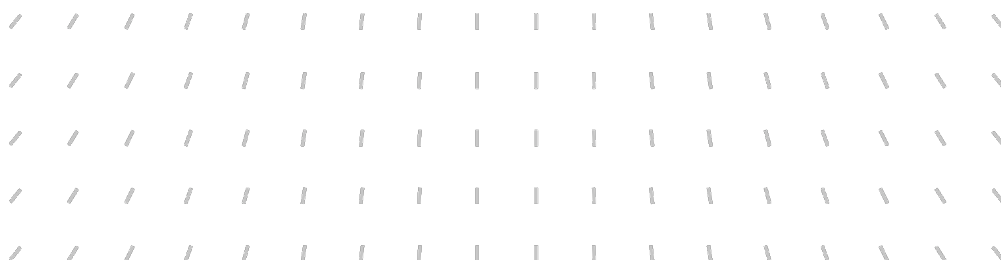
IN PARTNERSHIP WITH: CNRS, Université Claude Bernard (Lyon 1), Ecole normale supérieure de Lyon

Project-Team

ROMA

Optimisation des ressources : modèles, algorithmes
et ordonnancement

In collaboration with Laboratoire de l'Informatique du Parallélisme (LIP)



Project-Team ROMA

Creation of the Project-Team: 2015 January 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

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. – HPC for machine learning
- A6.3. – Computation-data interaction
- A7.1. – Algorithms
- A7.1.2. – Parallel 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

Contents

Project-Team ROMA	1
1 Team members, visitors, external collaborators	5
2 Overall objectives	6
3 Research program	6
3.1 Resilience for very large scale platforms	6
3.2 Multi-criteria scheduling strategies	7
3.3 Sparse direct solvers and sparsity in computing	7
4 Application domains	8
5 Social and environmental responsibility	8
5.1 Impact of research results	8
6 Highlights of the year	8
7 Latest software developments, platforms, open data	9
7.1 Latest software developments	9
7.1.1 MatchMaker	9
8 New results	9
8.1 Resilience for very large scale platforms	9
8.1.1 Fixed-Work vs. Fixed-Time Checkpointing on Large-Scale Failure-Prone Platforms	9
8.1.2 Partial Detectors Versus Replication To Cope With Silent Errors	10
8.1.3 Fault-tolerant numerical iterative algorithms at scale	10
8.2 Multi-criteria scheduling strategies	10
8.2.1 Green Scheduling on the Edge	10
8.2.2 Carbon-Aware Workflow Scheduling with Fixed Mapping and Deadline Constraint	11
8.2.3 Deadline-Aware Scheduling of Mixed-Criticality Tasks	11
8.2.4 Memory-aware Adaptive Scheduling of Scientific Workflows on Heterogeneous	
Architectures	12
8.2.5 A scheduler to foster data locality for GPU and out-of-core task-based linear algebra	
applications	12
8.2.6 Cache Management for Mixture-of-Experts LLMs	12
8.2.7 Leveraging Expert Usage to Speed up LLM Inference with Expert Parallelism . . .	13
8.2.8 Towards Parallel Transformer-Based Large Language Models for Fast Inference . .	13
8.2.9 Efficient and effective methods for variant selection	14
8.3 Sparse direct solvers and sparsity in computing	14
8.3.1 Efficient Parallel Sparse Tensor Contraction	14
8.3.2 Semi-Streaming Algorithms for Hypergraph Matching	15
8.3.3 Algorithms for symmetric Birkhoff-von Neumann decomposition of symmetric	
doubly stochastic matrices	15
8.3.4 SUperman: Efficient Permanent Computation on GPUs	15
8.3.5 Communication Lower Bounds and Optimal Algorithms for Symmetric Matrix	
Computations	16
8.3.6 Minimizing Communication for Parallel Symmetric Tensor Times Same Vector	
Computation	16
9 Bilateral contracts and grants with industry	16
9.1 Bilateral contracts with industry	16

10 Partnerships and cooperations	17
10.1 International initiatives	17
10.1.1 Inria associate team not involved in an IIL or an international program	17
10.1.2 Participation in other International Programs	17
10.2 International research visitors	19
10.2.1 Visits of international scientists	19
10.2.2 Visits to international teams	19
10.3 European initiatives	20
10.3.1 Horizon Europe	20
10.4 National initiatives	20
11 Dissemination	20
11.1 Promoting scientific activities	20
11.1.1 Scientific events: organisation	20
11.1.2 Scientific events: selection	20
11.1.3 Journal	21
11.1.4 Scientific expertise	21
11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach	21
11.2.1 Teaching	21
11.2.2 Supervision	22
11.2.3 Juries	22
12 Scientific production	22
12.1 Major publications	22
12.2 Publications of the year	23

1 Team members, visitors, external collaborators

Research Scientists

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

Faculty Members

- Anne Benoît [ENS DE LYON, Professor, HDR]
- Yves Robert [ENS DE LYON, Emeritus, HDR]

Post-Doctoral Fellows

- Antoine Jegou [SORBONNE UNIVERSITE]
- Maher Mallem [INRIA, Post-Doctoral Fellow]

PhD Students

- Joachim Cendrier [CNRS]
- Hadi Gholami [ENS DE LYON, from Nov 2025]
- Damien Lesens [ENS DE LYON, from Oct 2025]
- Michel Nicolis [INRIA, from Dec 2025]
- Adrien Obrecht [ENS DE LYON, from Sep 2025]
- Felix Wirth-Bonne [KOG, CIFRE, from Oct 2025]

Interns and Apprentices

- Damien Lesens [ENS DE LYON, Intern, until Jan 2025]

Administrative Assistant

- Chrystelle Mouton [INRIA]

Visiting Scientists

- Julien Langou [UNIV COLORADO, from Nov 2025]
- Julien Langou [UNIV COLORADO, until Jan 2025]

External Collaborator

- Theo Mary [CNRS]

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 several hundreds of thousands of cores, and edge servers represent many millions of resources. 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.

Combinatorial scientific computing (CSC) is a term 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 Social and environmental responsibility

5.1 Impact of research results

Within the framework of our collaboration with the University of Chicago (see Section 10.1.4), we explore novel scheduling algorithms that are able to adapt to dynamic power changes, to reduce carbon emissions, and to give priority to using green energy sources.

6 Highlights of the year

Anne Benoit was promoted full professor of ENS de Lyon.

Anne Benoit, Loris Marchal and Bora Uçar each spent the first semester of 2025 in sabbatical to build new collaborations and acquire new expertise. Anne Benoit and Bora Uçar were hosted at the Georgia Institute of Technology, and Loris Marchal at the ILLS laboratory of the École de Technologie Supérieure de Montréal.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 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.

Release Contributions: This version includes the initialization algorithm/heuristic 5 and moves files in suitable directories.

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

Publications: [hal-02463717](#), [hal-00786548](#), [hal-00763920](#)

Contact: Bora Uçar

Participants: Ioannis Panagiotas, Bora Uçar, Kamer Kaya, Johannes Langguth

8 New results

8.1 Resilience for very large scale platforms

The ROMA team has been working on resilience problems for several years. In 2025, we have focused on the following problems.

8.1.1 Fixed-Work vs. Fixed-Time Checkpointing on Large-Scale Failure-Prone Platforms

Participants: Quentin Barbut, Lucas Perotin (*Vanderbilt University*), Anne Benoit, Thomas Hérault (*University of Tennessee, Knoxville*), Yves Robert, Frédéric Vivien.

Consider a High-Performance Computing (HPC) application executing on a large-scale failure-prone platform. The Fixed-Work Checkpointing (FWC) problem consists in minimizing the expected time to execute a fixed amount of work (namely a fraction or the totality of the application). Strategies for the FWC problem have received considerable attention and are well-understood. On the contrary, the dual problem, namely the Fixed-Time Checkpointing (FTC) problem, has been considered only very recently. The FTC problem consists in maximizing the expected work achieved during a fixed amount of time (namely the duration of a reservation granted to the application). This work provides a comparative overview of both problems. First we review existing strategies for the FWC problem and extend them to stochastic checkpoints, i.e., when the checkpoint is no longer a deterministic constant but obeys some probability distribution law instead. Then we provide a comprehensive study of the FTC problem. The problem turns out to be surprisingly difficult, even when restricting to taking one or two checkpoints. We provide a threshold-based heuristic to solve

the general instance of the problem with an arbitrary number of checkpoints, and we have to resort to time discretization to provide an optimal strategy. We further extend this latter strategy to stochastic checkpoints.

This work has been published in the International Journal of High Performance Computing Applications [8].

8.1.2 Partial Detectors Versus Replication To Cope With Silent Errors

Participants: Anne Benoit, Thomas Hérault (*Inria Bordeaux*), Yves Robert, Alix Tremodeux.

This work studies an iterative algorithm running on an error-prone platform, where silent errors strike each iteration with some probability. A detector verifies correctness before taking a checkpoint but may fail to detect errors. Specifically, an error at iteration I is detected only after iteration $(I - 1) + X$, where X follows a bounded probability distribution like a truncated geometric distribution. Intuitively, the error silently amplifies during some iterations before it can be detected at distance X or higher, and there is the risk of missing an error that has struck recently but cannot be detected yet. X is bounded by D , the maximum detection latency. To mitigate undetected errors during verification, a simple strategy keeps two checkpoints and divides the execution into $D - 1$ iteration segments, each followed by verification and checkpoint. In steady state: (i) if verification succeeds, the oldest checkpoint is erased and replaced; (ii) if it fails, rollback occurs to the oldest verified checkpoint. This work explores whether this scheme outperforms replication and determines the optimal number of checkpoints and segment lengths, both theoretically and via Monte Carlo simulations.

This work was presented at the EuroPar 2025 conference [18].

8.1.3 Fault-tolerant numerical iterative algorithms at scale

Participants: Alix Tremodeux, Anne Benoit, Emmanuel Agullo (*Inria Bordeaux*), Thomas Hérault (*Inria Bordeaux; University of Tennessee, Knoxville*), Luc Giraud (*Inria Bordeaux*), Yves Robert.

This work investigates how to protect numerical iterative algorithms from all types of errors that can strike at scale: fail-stop errors (a.k.a. failures) and silent errors, striking both as computation errors and memory bit-flips. We combine various techniques: detectors for computation errors, checksums for memory errors, and checkpoint/restart for failures. The objective is to minimize the expected time per iteration of the algorithm. We design a hierarchical pattern that combines and interleaves all these fault-tolerance mechanisms, and we determine the optimal periodic pattern that achieves this objective. We instantiate these results for the performance analysis of the Preconditioned Conjugate Gradient (PCG) algorithm: we report several scenarios where the optimal pattern dramatically decreases the overhead due to error mitigation.

This work has been published in the International Journal of High Performance Computing Applications [13].

8.2 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.

8.2.1 Green Scheduling on the Edge

Participants: Joachim Cendrier, Rajini Wijayawardana (*University of Chicago*), Anne Benoit, Yves Robert, Frédéric Vivien, Andrew A. Chien (*University of Chicago*).

This work aims at designing and evaluating scheduling algorithms that minimize carbon cost on edge platforms. When a job is released to some edge server, difficult scheduling questions arise: should the job be executed on that server? If yes, when? If no, which other edge server should the job be transferred to? Typically, jobs are submitted online, and have a deadline to enforce. Online scheduling problems are already difficult without accounting for different energy sources, so one should not expect any optimal solution. Still, an important research goal is to revisit standard algorithms such as Earliest Completion Time (ECT) and Earliest Deadline First (EDF) in order to design and evaluate carbon-aware variants. This work introduces several new algorithms that use sophisticated scheduling policies to efficiently decrease carbon cost; these algorithms maximize the use of green energy both on local and remote edge servers, by re-evaluating previous decisions whenever needed to accommodate newly released jobs. We provide a comprehensive simulation campaign based on actual platform/job data and carbon traces and report an average gain of 42% over standard approaches.

This work was presented at the EuroPar 2025 conference [19].

8.2.2 Carbon-Aware Workflow Scheduling with Fixed Mapping and Deadline Constraint

Participants: Dominik Schweisgut (*Humboldt University Berlin, Karlsruhe Institute of Technology*), Anne Benoit, Yves Robert, Henning Meyerhenke (*Karlsruhe Institute of Technology*).

Large data and computing centers consume a significant share of the world's energy consumption. A prominent subset of the workloads in such centers are workflows with interdependent tasks, usually represented as directed acyclic graphs (DAGs). To reduce the carbon emissions resulting from executing such workflows in centers with a mixed (renewable and non-renewable) energy supply, it is advisable to move task executions to time intervals with sufficient green energy when possible. To this end, we formalize the above problem as a scheduling problem with a given mapping and ordering of the tasks. We show that this problem can be solved in polynomial time in the uniprocessor case. For at least two processors, however, the problem becomes NP-hard. Hence, we propose a heuristic framework called CaWoSched that combines several greedy approaches with local search. To assess the 16 heuristics resulting from different combinations, we also devise a simple baseline algorithm and an exact ILP-based solution. Our experimental results show that our heuristics provide significant savings in carbon emissions compared to the baseline.

This work was presented at the ICPP'25 conference [24].

8.2.3 Deadline-Aware Scheduling of Mixed-Criticality Tasks

Participants: Maxime Gonthier (*University of Chicago*), Kyle Chard (*University of Chicago*), Ian Foster (*University of Chicago*), Loris Marchal, Frédéric Vivien.

High-performance computing centers and cloud providers host a wide variety of workloads, ranging from routine calibration tasks with no strict timing requirements to urgent real-time computations that must be completed within hard deadlines. Traditional approaches reserve resources for high-criticality tasks or preempt and kill lower-criticality tasks when necessary, resulting in wasted compute time and longer turnaround times for lower-criticality tasks. We suggest that a better solution is to interleave the execution of critical and non-critical tasks. We formulate a bi-objective optimization problem: guarantee that all critical tasks meet their deadlines, and minimize the maximum flow, defined as the time a task spends in the system, of non-critical tasks. We introduce a formal model, derive an approximation algorithm and a lower bound, and develop several heuristics based on the approximation framework. Through extensive simulations, based on synthetic and real-world workloads, we show that one of our heuristics reduces the maximum flow of non-critical tasks by up to 14% compared to static resource partitioning.

This work was presented at the ICPP'25 conference [21].

8.2.4 Memory-aware Adaptive Scheduling of Scientific Workflows on Heterogeneous Architectures

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

The analysis of massive scientific data often happens in the form of workflows with interdependent tasks. When such a scientific workflow needs to be scheduled on a parallel or distributed system, one usually represents the workflow as a directed acyclic graph (DAG). The vertices of the DAG represent the tasks, while its edges model the dependencies between the tasks (usually data to be communicated to successor tasks). When executed, each task requires a certain amount of memory and if that exceeds the available memory, the execution fails. The typical goal is to execute the workflow without failures (i.e., satisfying the memory constraints) and with the shortest possible execution time (i.e., to minimize its makespan). To address this problem, we investigate the memory-aware scheduling of DAG-shaped workflows on heterogeneous platforms, where each processor can have a different speed and a different memory size. We propose a variant of HEFT (Heterogeneous Earliest Finish Time) that, in contrast to the original, accounts for memory and includes eviction strategies for cases when it might be beneficial to remove some data from memory in order to have enough memory to execute other tasks. Furthermore, while HEFT assumes perfect knowledge of the execution time and memory usage of each task, the actual values might differ upon execution. Thus, we propose an adaptive scheduling strategy, where a schedule is recomputed when there has been a significant variation in terms of execution time or memory. The scheduler has been closely integrated with a runtime system, allowing us to perform a thorough experimental evaluation on real-world workflows. The runtime system warns the scheduler when the task parameters have changed, and a schedule can be recomputed on the fly. The memory-aware strategy allows us to schedule task graphs that would run out of memory with a state-of-the-art scheduler, and the adaptive setting allows us to significantly reduce the makespan.

This work was presented at the CCGrid 2025 conference [22].

8.2.5 A scheduler to foster data locality for GPU and out-of-core task-based linear algebra applications

Participants: Maxime Gonthier (*University of Chicago*), Loris Marchal, Samuel Thibault (*Inria Bordeaux*).

Hardware accelerators like GPUs now provide a large part of the computational power used for scientific simulations. Despite their efficacy, GPUs possess limited memory and are connected to the main memory of the machine via a bandwidth limited bus. Scientific simulations often operate on very large data, that surpasses the GPU's memory capacity. Therefore, one has to turn to out-of-core computing: data is kept in a remote, slower memory (CPU memory), and moved back and forth from/to the device memory (GPU memory), a process also present for multicore CPUs with limited memory. 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. We propose a scheduler for task-based runtimes that improves data locality for out-of-core linear algebra computations, to reduce data movement. We design a data-aware strategy for both task scheduling and data eviction from limited memories. We compare this scheduler to existing schedulers in runtime systems. Using StarPU, we show that our new scheduling strategy achieves comparable performance when memory is not a constraint, and significantly better performance when application input data exceeds memory, on both GPUs and CPU cores.

This work has been published in the Journal of Parallel and Distributed Computing [10].

8.2.6 Cache Management for Mixture-of-Experts LLMs

Participants: Spyros Angelopoulos (*IRL ILLS*), Loris Marchal, Adrien Obrecht, Bertrand Simon (*CNRS IN2P3*).

Large language models (LLMs) have demonstrated remarkable capabilities across a variety of tasks. One of the main challenges towards the successful deployment of LLMs is memory management, since they typically involve billions of parameters. To this end, architectures based on Mixture-of-Experts have been proposed, which aim to reduce the size of the parameters that are activated when producing a token. This raises the equally critical issue of efficiently managing the limited cache of the system, in that frequently used experts should be stored in the fast cache rather than in the slower secondary memory. In this work, we introduce and study a new paging problem that models expert management optimization. Our formulation captures both the layered architecture of LLMs and the requirement that experts are cached efficiently. We first present lower bounds on the competitive ratio of both deterministic and randomized algorithms, which show that under mild assumptions, LRU-like policies have good theoretical competitive performance. We then propose a layer-based extension of LRU that is tailored to the problem at hand. Extensive simulations on both synthetic datasets and actual traces of MoE usage show that our algorithm outperforms policies for the classic paging problem, such as the standard LRU.

This work was presented at the EuroPar 2025 conference [16].

8.2.7 Leveraging Expert Usage to Speed up LLM Inference with Expert Parallelism

Participants: Olivier Beaumont (*Inria Bordeaux*), Raphaël Bourgoïn (*Inria Bordeaux*), Maxime Darrin (*IRL ILLS*), Loris Marchal, Pablo Piantanida (*IRL ILLS*).

Large language models have become indispensable for many text-processing applications. Their inference, i.e., their use to generate text, is a time-consuming task since tokens have to be generated one after the other, even if the computational load has been reduced by model sparsification, e.g., by using a Mixture of Experts (MoE) models. In the MoE context, a subset of experts is selected at each stage. Note that not all subsets of experts (pairs of experts in most cases) in a given layer have the same probability of being selected. When experts are mapped to different GPUs, there is a risk of load imbalance if the selected experts end up on a small number of GPUs. This work proposes to leverage this heterogeneity in expert usage to map experts of popular subsets onto distinct GPUs, allowing them to be processed in parallel and thus reducing the time needed for inference. Even though this mapping problem is NP-complete, it is possible to design simple greedy strategies that significantly reduce the need for sequential expert processing. Our proof-of-concept confirms that our mapping strategies effectively reduce inference time on the Mixtral model.

This work was presented at the EuroPar 2025 conference [17].

8.2.8 Towards Parallel Transformer-Based Large Language Models for Fast Inference

Participants: Gaël Dellaleau (*Kog company*), Morgan Giraud (*Kog company*), Loris Marchal, Félix Wirth (*Kog company*).

Transformer-based models have reached quality levels that make them attractive for a wide range of applications, including those demanding faster generation speeds such as multimedia artifact generation. However, traditional parallelism techniques have struggled to meet these speed requirements due to the inherently sequential structure of Transformers, which limits parallel execution. In this work, we have proposed a simple modeling to the parallel execution of Transformer, which identifies synchronization times across inference devices as the culprit. Leveraging real world figures from GPU manufacturers, we highlight the increasing cost of this synchronization time with ever more powerful hardware. In response, we introduce Tensor Parallel Blocks, a naturally parallelized architecture. To maintain quality, we incorporate multiple enhancements, such as delayed inter-block communication, hybrid architecture, and different parallelism gateways, which mitigate potential degradation in model output. We design and validate a dedicated training protocol tailored to this architecture and conduct a preliminary hyperparameter study on 150M models, proving hardware-tied architecture can achieve relevant results. Using insights from these trials, we establish a set of rules for scaling the model to larger size, theorizing a 4.2x speedup in decoding time compared to an equivalently sized LLaMA based model on the target inference platform.

This work is detailed in a research report [33] describing the work initiated during the 6 month internship of Félix Wirth in the ROMA team and continued in collaboration with Gael Delalleau and Morgan Giraud from the Kog company.

8.2.9 Efficient and effective methods for variant selection

Participants: Srinivas Aluru (*Georgia Institute of Technology*), Anne Benoit, Bora Uçar.

Variation graphs succinctly capture genetic variations among individuals within a species or a target population. The use of variation graphs instead of a single reference genome is credited with reducing bias and increasing the accuracy of sequence mapping algorithms. However, complete variation graphs that comprehensively incorporate all genetic variations are often found to be ineffective and inaccurate in practice due to the presence of a combinatorially explosive number of paths in the graph that do not correspond to any observed genome. Thus, a balance is struck in carefully selecting a subset of variants to be incorporated, for which mathematical frameworks have recently been developed. We advance the mathematical framework proposed by Jain et al., where integer linear programming formulations were developed for optimal variant selection. We propose novel graph-based formulations and develop exact and fast algorithms for certain cases, approximation methods for some others, and empirically close to optimal results in all cases. The primary advantage of the algorithms designed here is that they provide near-optimal results at orders of magnitude faster run time of an ILP solver.

This work was presented at the 16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics [15].

8.3 Sparse direct solvers and sparsity in computing

We continued our work sparse tensors by looking at the tensor contraction operation, which is a higher order analogue of the sparse matrix sparse matrix multiplication operation. We worked on combinatorial problems arising in sparse tensor models. Further work on graph algorithms and Birkhoff–von Neumann decomposition using methods from the signal processing domain were also conducted. We have also investigated communication lower bounds and algorithms achieving those bounds for certain (dense) matrix and tensor kernels.

8.3.1 Efficient Parallel Sparse Tensor Contraction

Participants: Somesh Singh, Bora Uçar.

We investigate the performance of algorithms for sparse tensor-sparse tensor multiplication (SpGETT). This operation, also called sparse tensor contraction, is a higher order analogue of the sparse matrix-sparse matrix multiplication (SpGEMM) operation. Therefore, SpGETT can be performed by first converting the input tensors into matrices, then invoking high performance variants of SpGEMM, and finally reconvert the resultant matrix into a tensor. Alternatively, one can carry out the scalar operations underlying SpGETT in the realm of tensors without matrix formulation. We discuss the building blocks in both approaches and formulate a hashing-based method to avoid costly search or redirection operations. We present performance results with the current state-of-the-art SpGEMM-based approaches, existing SpGETT approaches, and a carefully implemented SpGETT approach with a new fine-tuned hashing method, proposed in this work. We evaluate the methods on real world tensors, contracting a tensor with itself along various dimensions. Our proposed hashing-based method for SpGETT consistently outperforms the state-of-the-art method, achieving a 25% reduction in sequential execution time on average and a 21% reduction in parallel execution time on average across a variety of input instances.

This work has been published in the IEEE Transactions on Parallel and Distributed Systems journal [12]. Codes are available [here](#).

8.3.2 Semi-Streaming Algorithms for Hypergraph Matching

Participants: Henrik Reinstädler (*Heidelberg University*), S.M. Ferdous (*PNNL - Pacific Northwest National Laboratory*), Alex Pothen (*Purdue University*), Bora Uçar, Christian Schulz (*Heidelberg University*).

We propose two one-pass streaming algorithms for the \mathcal{NP} -hard hypergraph matching problem. The first algorithm stores a small subset of potential matching edges in a stack using dual variables to select edges. It has an approximation guarantee of $\frac{1}{d(1+\epsilon)}$ and requires $O((\frac{n}{\epsilon}) \log^2 n)$ bits of memory, where n is the number of vertices in the hypergraph, d is the maximum number of vertices in a hyperedge, and $\epsilon > 0$ is a parameter to be chosen. The second algorithm computes, stores, and updates a single matching as the edges stream, with an approximation ratio dependent on a parameter α . Its best approximation guarantee is $\frac{1}{(2d-1)+2\sqrt{d(d-1)}}$, and it requires only $O(n)$ memory.

We have implemented both algorithms and compared them with respect to solution quality, memory consumption, and running times on two diverse sets of hypergraphs with a non-streaming greedy and a naive streaming algorithm. Our results show that the streaming algorithms achieve much better solution quality than naive algorithms when facing adverse orderings. Furthermore, these algorithms reduce the memory required by a factor of 13 in the geometric mean on our test problems, and also outperform the offline Greedy algorithm in running time.

This work was presented at the ESA 2025 conference [23]. Codes are available [here](#).

8.3.3 Algorithms for symmetric Birkhoff-von Neumann decomposition of symmetric doubly stochastic matrices

Participants: Damien Lesens, Jérémy E. Cohen (*CNRS, Lyon*), Bora Uçar.

The classical Birkhoff–von Neumann (BvN) decomposition expresses a given doubly stochastic matrix as a convex combination of permutation matrices. We investigate the BvN decomposition of symmetric doubly stochastic matrices where the permutation matrices in the decomposition are also symmetric, called SymBvN decomposition. This decomposition is not always possible. Two pioneering theoretical works establish the conditions under which such a decomposition is possible using graph terminology. We propose a practical algorithm by combining these two works. A simple transformation converts any given symmetric doubly stochastic matrix, with possibly nonzero diagonal elements, to be the adjacency matrix of an edge-weighted undirected graph. The adjacency matrix of the resulting graph admits a SymBvN decomposition if and only if the given matrix does so. The practicality of the proposed algorithm allows us to implement it, release its source code, and report the first set of experiments ever performed for the SymBvN decomposition. Our experiments suggest that the proposed algorithm is as effective as the state-of-the-art algorithms for the classical BvN decomposition.

This work has been accepted to be published in SIAM Journal on Matrix Analysis and Applications in Dec 2025 [11]. Codes are available [here](#).

8.3.4 Superman: Efficient Permanent Computation on GPUs

Participants: Deniz Elbek (*Sabancı University*), Fatih Taşyaran (*Sabancı University*), Bora Uçar, Kamer Kaya (*Sabancı University*).

The *permanent* is a function, defined for a square matrix, with applications in various domains including quantum computing, statistical physics, complexity theory, combinatorics, and graph theory. Its formula is similar to that of the determinant, however unlike the determinant, its exact computation is $\#P$ -complete, i.e., there is no algorithm to compute the permanent in polynomial time unless $P=NP$. For an $n \times n$ matrix, the fastest algorithm has a time complexity of $O(2^{n-1}n)$. Although supercomputers have been employed for

permanent computation before, there is no work and more importantly, no publicly available software that leverages cutting-edge, yet widely accessible, High-Performance Computing accelerators such as GPUs. In this work, we designed, developed, and investigated the performance of SUPERMAN, a complete software suite that can compute matrix permanents on multiple nodes/GPUs on a cluster while handling various matrix types, e.g., real/complex/binary and sparse/dense etc., with a unique treatment for each type. Compared to a state-of-the-art parallel algorithm on 44 cores, SUPERMAN can be 86× faster on a single Nvidia A100 GPU. Combining multiple GPUs, we also showed that SUPERMAN can compute the permanent of a 56×56 matrix which is the largest reported in the literature.

This work has been recently accepted to be published in Computer Physics Communications [30].

8.3.5 Communication Lower Bounds and Optimal Algorithms for Symmetric Matrix Computations

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

In this work, we focus on the communication costs of three symmetric matrix computations: i) multiplying a matrix with its transpose, known as a symmetric rank-k update (SYRK) ii) adding the result of the multiplication of a matrix with the transpose of another matrix and the transpose of that result, known as a symmetric rank-2k update (SYR2K) iii) performing matrix multiplication with a symmetric input matrix (SYMM). All three computations appear in the Level 3 Basic Linear Algebra Subroutines (BLAS) and have wide use in applications involving symmetric matrices. We establish communication lower bounds for these kernels using sequential and distributed-memory parallel computational models, and we show that our bounds are tight by presenting communication-optimal algorithms for each setting. Our lower bound proofs rely on applying a geometric inequality for symmetric computations and analytically solving constrained nonlinear optimization problems. The symmetric matrix and its corresponding computations are accessed and performed according to a triangular block partitioning scheme in the optimal algorithms.

This work has been published in the ACM Transactions on Parallel Computing journal [9].

8.3.6 Minimizing Communication for Parallel Symmetric Tensor Times Same Vector Computation

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

In this work, we focus on the parallel communication cost of multiplying the same vector along two modes of a 3-dimensional symmetric tensor. This is a key computation in the higher-order power method for determining eigenpairs of a 3-dimensional symmetric tensor and in gradient-based methods for computing a symmetric CP decomposition. We establish communication lower bounds that determine how much data movement is required to perform the specified computation in parallel. We demonstrate that the communication lower bounds are tight by presenting an optimal algorithm where the data distribution is a natural extension of the triangle block partition scheme for symmetric matrices to 3-dimensional symmetric tensors.

This work was presented at the SPAA 2025 conference [14].

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Participants: Loris Marchal, Félix Wirth-Bonne.

- Contrat de collaboration entre la société Kog et l'équipe-projet ROMA pour le co-encadrement du stage de Félix Wirth (2200€).

10 Partnerships and cooperations

10.1 International initiatives

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

MODS

Participants: Bora Uçar.

Title: Match and Order: improving direct solvers for cardiac simulations

Duration: 2023 -> 2025.

Coordinator: Johannes Langguth (langguth@simula.no)

Partners:

- University of Bergen (Norway)

Inria contact: Bora Uçar

Summary: 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.

10.1.2 Participation in other International Programs

JLESC — Joint Laboratory on Extreme Scale Computing.

Participants: Anne Benoit, Yves Robert, Frédéric Vivien.

The University of Illinois at Urbana-Champaign, INRIA, the French national computer science institute, Argonne National Laboratory, Barcelona Supercomputing Center, Jülich Supercomputing Centre and the Riken Advanced Institute for Computational Science formed the Joint Laboratory on Extreme Scale Computing, a follow-up of the Inria-Illinois Joint Laboratory for Petascale Computing. The Joint Laboratory is based at Illinois and includes researchers from INRIA, and the National Center for Supercomputing Applications, ANL, BSC and JSC. It focuses on software challenges found in extreme scale high-performance computers.

Research areas include:

- Scientific applications (big compute and big data) that are the drivers of the research in the other topics of the joint-laboratory.
- Modeling and optimizing numerical libraries, which are at the heart of many scientific applications.
- Novel programming models and runtime systems, which allow scientific applications to be updated or reimaged to take full advantage of extreme-scale supercomputers.
- Resilience and Fault-tolerance research, which reduces the negative impact when processors, disk drives, or memory fail in supercomputers that have tens or hundreds of thousands of those components.

- I/O and visualization, which are important parts of parallel execution for numerical simulations and data analytics
- HPC Clouds, that may execute a portion of the HPC workload in the near future.

Several members of the ROMA team are involved in the JLESC joint lab through their research on scheduling and resilience.

Collaboration with Humboldt University Berlin/KIT

Participants: Anne Benoit, Yves Robert.

Title: FONDA

Partner Institution(s): • Humboldt University Berlin and Karlsruhe Institute of Technology, Germany

Date/Duration: 2024–2028

Additional info/keywords: Anne Benoit is a Mercator Fellow for the Collaborative Research Center FONDA. This project aims at studying the foundations of workflows for large-scale scientific data analysis.

Collaboration with U. Darmstadt

Participants: Anne Benoit, Hadi Gholami, Frédéric Vivien.

Title: ANR-DFG EnergyDoldrums

Partner Institution(s): • University of Darmstadt, Germany

Date/Duration: 2025–2028

Additional info/keywords: The aim of this project is to optimize the scheduling and execution of adaptive HPC jobs on platforms whose energy supply varies over time.

Collaboration with U. Chicago

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

Title: U. Chicago-CNRS project

Partner Institution(s): • University of Chicago, USA

Date/Duration: 2023–2026

Additional info/keywords: This project is funding the PhD of J. Cendrier, on efficient and environment-friendly scheduling and resource management algorithms at the edge

Participants: Loris Marchal, Frédéric Vivien.

Title: FACCTS project

Partner Institution(s): • University of Chicago, USA

Date/Duration: 2024–2026

Additional info/keywords: In this project, we focus on the heterogeneity in the workload submitted to computing centers and clouds. Some of the submitted jobs may be critical, with hard deadlines on their completion, while other may be scheduled on a best-effort basis to reduce their response time. We aim at designing scheduling strategies to accommodate both type of jobs with their different constraints and objectives on a shared computing platform.

10.2 International research visitors

10.2.1 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 4.5 months in the team in May-July 2025 and starting mid-November 2025.

10.2.2 Visits to international teams

Research stays abroad

Loris Marchal

Visited institution: École Supérieure de Technologie, Montréal

Country: Canada

Dates: 01–07/2025

Context of the visit: temporary assignement at the ILLS laboratory (CNRS, McGill, ETS, MILA).

Mobility program/type of mobility: sabbatical

Anne Benoit

Visited institution: Institute of Data Engineering and Science (IDEaS), Georgia Institute of Technology

Country: US

Dates: 02–07/2025

Context of the visit: Collaboration with S. Aluru

Mobility program/type of mobility: Long term stay abroad

Bora Uçar

Visited institution: Institute of Data Engineering and Science (IDEaS), Georgia Institute of Technology

Country: US

Dates: 02–07/2025

Context of the visit: Collaboration with S. Aluru

Mobility program/type of mobility: Long term stay abroad

10.3 European initiatives

10.3.1 Horizon Europe

Participants: Anne Benoit, Michel Nicolis, Frédéric Vivien.

HORIZON-INFRA project ODISSEE (2025–2027). The ODISSEE project aims to develop innovative technologies and methodologies to process the unprecedented volume of scientific data produced by research infrastructures such as CERN’s HL-LHC and SKAO. It notably plans to develop on-the-fly AI data processing, which is a major challenge in research in the physical sciences, and where the contribution of SLICES RI will be decisive. Coordinated by Damien Gratadour, researcher at the CNRS Laboratory for Instrumentation and Research in Astrophysics, ODISSEE is leveraging the European HPC ecosystem to open up a new era in science, helping to unravel fundamental mysteries such as the nature of dark matter. The three-year ODISSEE consortium brings together 14 partners and 2 associate partners from academia and industry. F. Vivien is the head of work-package 5.

10.4 National initiatives

ECLAT

Participants: Frédéric Vivien.

Partner Institution(s): CNRS, Inria, Eviden, Observatoire de la Côte d’Azur, Observatoire de Paris-PSL

Date/Duration: 2023-

Summary **ECLAT** is a joint laboratory gathering 14 laboratories. Its aim is to support the French contribution to the **SKAO** observatory.

11 Dissemination

11.1 Promoting scientific activities

18th Workshop on Scheduling for Large Scale Systems: Loris Marchal, Pablo Piantanida and Yves Robert have organized the 18th Workshop on Scheduling for Large Scale Systems in Montréal in July 2025. Further details can be found on the [workshop webpage](#).

11.1.1 Scientific events: organisation

Bora Uçar was the organizing committee co-chair of SIAM Conference on Applied and Computational Discrete Algorithms (ACDA25) Montréal, Canada. In SIAM conference systems this role corresponds to the general chair.

11.1.2 Scientific events: selection

Chair of conference program committees

- Anne Benoit was chair for Track B of ESA, the European Symposium on Algorithms, in conjunction with ALGO’2025, Warsaw, Poland, September 15-19, 2025.

Member of the conference program committees

- Anne Benoit was a member of the program committee of ALENEX'25, EuroPar'25, ICPP'25, SC'25 workshops and symposiums.
- Suraj Kumar was a member of the program committee of SC 2025.
- Loris Marchal was a member of the program committee of EuroPar 2025.
- Bora Uçar was a member of the program committee of 39th IEEE International Parallel & Distributed Processing Symposium, June 3–7, 2025, Milan, Italy.
- Frédéric Vivien was a member of the program committee of ESA 2025.

Reviewer Bora Uçar reviewed articles for ALENEX2025:SIAM Symposium on Algorithm Engineering and Experiments, New Orleans, Louisiana, USA.

11.1.3 Journal

Member of the editorial boards

- Anne Benoit is Editor in Chief of ParCo, the journal of Parallel Computing: Systems and Applications, and she is also a member of the editorial board of ACM TOPC (Transactions on Parallel Computing).
- 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 System, the journal of Parallel Computing (until May 2025), SIAM Journal on Scientific Computing (SISC) until December 2025, and SIAM Journal on Matrix Analysis and Applications (SIMAX).
- Frédéric Vivien is a member of the editorial board of the Journal of Parallel and Distributed Computing.

Reviewer - reviewing activities

- Anne Benoit has reviewed manuscripts for ACM Transactions on Parallel Computing, IEEE TPDS, IJHPCA.
- Suraj Kumar has reviewed manuscripts for Journal of Parallel and Distributed Computing, SIAM Journal on Scientific Computing, and Transactions on Parallel and Distributed Systems journal.
- Bora Uçar was a reviewer for ACM Transactions on Parallel Computing, The Journal of Supercomputing.

11.1.4 Scientific expertise

- Frédéric Vivien was a member of the evaluation committee for the Scientific Evaluation of Forschungszentrum Jülich (FZJ), Germany, in April 2025.
- Frédéric Vivien is an elected member of INRIA *commission d'évaluation*.
- Frédéric Vivien is a member of the scientific council of the [IRMIA labex](#).

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

11.2.1 Teaching

- Master: Anne Benoit, Parallel and Distributed Algorithms and Programs, 39h, M1, ENS Lyon, France
- Master: Suraj Kumar, Loris Marchal and Frédéric Vivien, Resource-aware computations on CPUs and GPUs, 30h, M2, ENS Lyon, France.

11.2.2 Supervision

- Anne Benoit and Frédéric Vivien are co-supervising the PhD 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.
- Anne Benoit and Frédéric Vivien are co-supervising the PhD of Hadi Gholami. Hadi works on resource allocation to jobs running on variable-capacity platforms.
- Anne Benoit is co-supervising the PhD of Svetlana Kulagina with Henning Meyerhenke (Humboldt-Universität zu Berlin, Germany), as part of the FONDA project, on the execution of large workflows in heterogeneous execution environments.
- Anne Benoit is co-supervising the PhD of Dominik Schweisgut with Henning Meyerhenke (now at Karlsruhe Institute of Technology, Germany), as part of the FONDA project, on algorithms for carbon-aware workflow scheduling.
- Frédéric Vivien and Anne Benoit are co-supervising the PhD of Michel Nicolis. Michel works on predictive maintenance.
- Anne Benoit and Loris Marchal are supervising the PhD of Adrien Obrecht on scheduling task graphs with application to large language models.
- Loris Marchal is supervising the PhD of Felix Wirth (Cifre collaboration with the Kog company) on the design of generative models with efficient parallel inference.
- Bora Uçar is co-supervising the PhD of Damien Lesens on nonnegative matrix factorization.

11.2.3 Juries

- Anne Benoit was a reviewer of the PhD of Diane Orhan, defended at Bordeaux University on December 9, 2025.
- Anne Benoit was an examiner for the PhD of Roblex Nana Tchakouté, defended at PSL University Paris on December 5, 2025.
- Loris Marchal was a reviewer of the PhD of Robin Boezennec, defended at Rennes University on December 10, 2025.
- Loris Marchal was a reviewer of the PhD of Huijun Wang, to be defended at Auckland university (New Zealand) in Spring 2026.
- Frédéric Vivien was a member of the jury for hiring INRIA CRCN and ISFP researchers for the Centre Inria de l'Université de Bordeaux.
- Frédéric Vivien was a member of the jury for hiring INRIA senior (DR2) researchers.

12 Scientific production

12.1 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1016/j.jpdc.2019.01.009). URL: <https://hal.inria.fr/hal-02025521>.

12.2 Publications of the year

International journals

- [8] Q. Barbut, L. Perotin, A. Benoit, T. Hérault, Y. Robert and F. Vivien. ‘Fixed-Work vs. Fixed-Time Checkpointing on Large-Scale Failure-Prone Platforms’. In: *International Journal of High Performance Computing Applications* 40.1 (Aug. 2025), pp. 96–114. DOI: [10.1177/10943420251379278](https://doi.org/10.1177/10943420251379278). URL: <https://inria.hal.science/hal-05232847> (cit. on p. 10).
- [9] H. A. Daas, G. Ballard, L. Grigori, S. Kumar, K. Rouse and M. Verite. ‘Communication Lower Bounds and Optimal Algorithms for Symmetric Matrix Computations’. In: *ACM Transactions on Parallel Computing* 12.2 (2025). DOI: [10.1145/3727344](https://doi.org/10.1145/3727344). URL: <https://inria.hal.science/hal-04701302> (cit. on p. 16).
- [10] M. Gonthier, L. Marchal and S. Thibault. ‘A scheduler to foster data locality for GPU and out-of-core task-based linear algebra applications’. In: *Journal of Parallel and Distributed Computing* 206 (11th Aug. 2025), p. 105170. DOI: [10.1016/j.jpdc.2025.105170](https://doi.org/10.1016/j.jpdc.2025.105170). URL: <https://inria.hal.science/hal-05226796> (cit. on p. 12).
- [11] D. Lesens, J. E. Cohen and B. Uçar. ‘Algorithms for symmetric Birkhoff-von Neumann decomposition of symmetric doubly stochastic matrices’. In: *SIAM Journal on Matrix Analysis and Applications* (Jan. 2025), pp. 1–30. URL: <https://inria.hal.science/hal-04877502>. In press (cit. on p. 15).
- [12] S. Singh and B. Uçar. ‘Efficient Parallel Sparse Tensor Contraction’. In: *IEEE Transactions on Parallel and Distributed Systems* 36.6 (2025), pp. 1206–1219. DOI: [10.1109/TPDS.2025.3557750](https://doi.org/10.1109/TPDS.2025.3557750). URL: <https://hal.science/hal-05047235> (cit. on p. 14).
- [13] A. Tremodeux, A. Benoit, E. Agullo, T. Hérault, L. Giraud and Y. Robert. ‘Fault-tolerant numerical iterative algorithms at scale’. In: *International Journal of High Performance Computing Applications* (2025). URL: <https://inria.hal.science/hal-05234063> (cit. on p. 10).

International peer-reviewed conferences

- [14] H. Al Daas, G. Ballard, L. Grigori, S. Kumar, K. Rouse and M. Verite. ‘Minimizing Communication for Parallel Symmetric Tensor Times Same Vector Computation’. In: *Brief Announcement: Minimizing Communication for Parallel Symmetric Tensor Times Same Vector Computation*. SPAA 2025 - ACM Symposium on Parallelism in Algorithms and Architectures. Portland, United States, 28th July 2025. DOI: [10.1145/3694906.3743332](https://doi.org/10.1145/3694906.3743332). URL: <https://inria.hal.science/hal-05130982> (cit. on p. 16).

- [15] S. Aluru, A. Benoit and B. Uçar. ‘Efficient and effective methods for variant selection’. In: *BCB’25: Proceedings of the 16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB 2025)*. 16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB 2025). Philadelphia, PA, United States, Dec. 2025. URL: <https://hal.science/hal-05237949> (cit. on p. 14).
- [16] S. Angelopoulos, L. Marchal, A. Obrecht and B. Simon. ‘Cache Management for Mixture-of-Experts LLMs’. In: *Euro-Par 2025: Parallel Processing. Euro-Par 2025. Lecture Notes in Computer Science*. 31st European Conference on Parallel and Distributed Processing (EURO-PAR 2025). Vol. 15902. Dresden, Germany, 2025, pp. 18–32. doi: [10.1007/978-3-031-99872-0_2](https://doi.org/10.1007/978-3-031-99872-0_2). URL: <https://hal.science/hal-04961621> (cit. on p. 13).
- [17] O. Beaumont, R. Bourgouin, M. Darrin, L. Marchal and P. Piantanida. ‘Leveraging Expert Usage to Speed up LLM Inference with Expert Parallelism’. In: *Lecture Notes in Computer Science. Euro-Par 2025: Parallel Processing*. Dresden, Germany, 25th Aug. 2025. URL: <https://hal.science/hal-04994839> (cit. on p. 13).
- [18] A. Benoit, T. Herault, Y. Robert and A. Tremodeux. ‘Partial Detectors Versus Replication To Cope With Silent Errors’. In: *Euro-Par 2025 - 31 st International European Conference on Parallel and Distributed Computing*. Dresden, Germany, 25th Aug. 2025. URL: <https://inria.hal.science/hal-05231852> (cit. on p. 10).
- [19] J. Cendrier, R. Wijayawardana, A. Benoit, Y. Robert, F. Vivien and A. A. Chien. ‘Green Scheduling on the Edge’. In: *Euro-Par 2025 - 31st International European Conference on Parallel and Distributed Computing*. Vol. 15900. Lecture Notes in Computer Science. Dresden, Germany: Springer Nature Switzerland, 25th Aug. 2025, pp. 380–394. doi: [10.1007/978-3-031-99854-6_26](https://doi.org/10.1007/978-3-031-99854-6_26). URL: <https://inria.hal.science/hal-05224558> (cit. on p. 11).
- [20] A. Chhabra, C. Schulz, B. Uçar and L. Wilwert. ‘Exact Minimum Cuts in Hypergraphs at Scale’. In: *SIAM Symposium on Algorithm Engineering and Experiments (ALENEX26)*. Vancouver, Canada, 11th Jan. 2026. URL: <https://hal.science/hal-05243769>.
- [21] M. Gonthier, K. Chard, I. Foster, L. Marchal and F. Vivien. ‘Deadline-Aware Scheduling of Mixed-Criticality Tasks’. In: *ICPP ’25: Proceedings of the 54th International Conference on Parallel Processing*. ICPP ’25 - 54th International Conference on Parallel Processing. San Diego, CA, United States, 8th Sept. 2025. URL: <https://inria.hal.science/hal-05190988> (cit. on p. 11).
- [22] S. Kulagina, A. Benoit and H. Meyerhenke. ‘Memory-aware Adaptive Scheduling of Scientific Workflows on Heterogeneous Architectures’. In: *CCGrid 2025 - 25th IEEE International Symposium on Cluster, Cloud and Internet Computing*. Tromsø, Norway, May 2025. URL: <https://inria.hal.science/hal-05231849> (cit. on p. 12).
- [23] H. Reinstädler, S. M. Ferdous, A. Pothen, B. Uçar and C. Schulz. ‘Semi-Streaming Algorithms for Hypergraph Matching’. In: *Leibniz International Proceedings in Informatics (LIPIcs)*. ESA 2025 - European Symposium on Algorithms. Warsaw (POLAND), Poland, 19th Feb. 2025. URL: <https://hal.science/hal-04966763> (cit. on p. 15).
- [24] D. Schweisgut, A. Benoit, Y. Robert and H. Meyerhenke. ‘Carbon-Aware Workflow Scheduling with Fixed Mapping and Deadline Constraint’. In: *International Conference on Parallel Processing (ICPP)*. ICPP 2025 - 54th International Conference on Parallel Processing. San Diego (CA), United States, 11th Aug. 2025. URL: <https://inria.hal.science/hal-05231854> (cit. on p. 11).

Reports & preprints

- [25] A. Benoit, J. Cendrier and F. Vivien. *Scheduling Jobs Under a Variable Number of Processors*. RR-9582. Inria, Nov. 2025. URL: <https://inria.hal.science/hal-05000632>.
- [26] A. Benoit, A. A. Chien and Y. Robert. *2nd Workshop on Scheduling Variable Capacity Resources for Sustainability*. Inria, Jan. 2026. URL: <https://inria.hal.science/hal-05457056>.
- [27] A. Benoit, T. Herault, Y. Robert and A. Tremodeux. *Partial Detectors Versus Replication To Cope With Silent Errors*. RR-9581. Inria, Mar. 2025. URL: <https://inria.hal.science/hal-04996292>.

- [28] R. H. Bisseling and B. Uçar. *Optimal partitioning of 2D meshes for stencil computations*. RR-9585. Inria Lyon, Sept. 2025. URL: <https://inria.hal.science/hal-05006840>.
- [29] J. Cendrier, R. Wijayawardana, A. Benoit, Y. Robert, F. Vivien and A. A. Chien. *Green Scheduling on the Edge*. RR-9580. Inria, Mar. 2025. URL: <https://inria.hal.science/hal-04994586>.
- [30] D. Elbek, F. Taşyaran, B. Uçar and K. Kaya. *SUPERman: Efficient Permanent Computation on GPUs*. RR-9578. Inria Lyon, Mar. 2025, pp. 1–25. URL: <https://hal.science/hal-04985427> (cit. on p. 16).
- [31] Q. Gao, L. Han, S. Hunold, Y. Robert and F. Vivien. *Coping with Silent Errors for Workflows of Moldable Tasks*. RR-9589. INRIA, June 2025. URL: <https://inria.hal.science/hal-05116521>.
- [32] A. Tremodeux, E. Agullo, A. Benoit, L. Giraud, T. Herault and Y. Robert. *Fault-tolerant numerical iterative algorithms at scale*. RR-9567. Inria Lyon, Jan. 2025. URL: <https://inria.hal.science/hal-04872041>.
- [33] F. Wirth, G. Delalleau and L. Marchal. *Towards Parallel Transformer-Based Large Language Models for Fast Inference*. RR-9573. Inria, Jan. 2025, p. 42. URL: <https://inria.hal.science/hal-04920049> (cit. on p. 14).

Software

- [34] [SW] K. Kaya, J. Langguth, I. Panagiotas and B. Uçar, *MatchMaker* 15th May 2025. LIC: CeCILL-B. HAL: [hal-05069079](https://hal.science/hal-05069079), URL: <https://hal.science/hal-05069079>, vcs: <https://gitlab.inria.fr/bora-ucar/matchmaker>, SWHID: [sw:1:dir:9baf1bc03292dbadf5b5b50a9d65bdb1d38c9981;origin=https://gitlab.inria.fr/bora-ucar/matchmaker.git;visit=sw:1:snip:7e6827b13c64ee91add458723f068e1b5031d947;anchor=sw:1:rev:50b09f54e9c4255c1a7becfce54d088940124149](https://sw.hic.cc/v1/urn:inria:2025-05-15-15:1:snip:7e6827b13c64ee91add458723f068e1b5031d947;anchor=sw:1:rev:50b09f54e9c4255c1a7becfce54d088940124149).