RESEARCH CENTRE

**Inria Centre at the University of Bordeaux**

**IN PARTNERSHIP WITH:**

**Institut Polytechnique de Bordeaux, Université de Bordeaux, CNRS**

2024

# ACTIVITY REPORT

## Project-Team

# TOPAL

## Tools and Optimization for high Performance Applications and Learning

**IN COLLABORATION WITH: Laboratoire Bordelais de Recherche en Informatique (LaBRI)**

**DOMAIN**

**Networks, Systems and Services, Distributed Computing**

**THEME**

**Distributed and High Performance Computing**

Inria

# Contents

# Project-Team TOPAL

*Creation of the Project-Team: 2023 March 01*

## Keywords

### Computer sciences and digital sciences

A1.1.4. – High performance computing

A1.1.5. – Exascale

A1.6. – Green Computing

A2.6.4. – Ressource management

A3.4.4. – Optimization and learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A6.2.5. – Numerical Linear Algebra

A6.2.7. – High performance computing

A7.1. – Algorithms

A7.1.2. – Parallel algorithms

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.9. – Distributed AI, Multi-agent

### Other research topics and application domains

B4.2.2. – Fusion

B9.5.1. – Computer science

B9.5.2. – Mathematics

# 1 Team members, visitors, external collaborators

## Research Scientists

- Olivier Beaumont [Team leader, INRIA, Senior Researcher]
- Lionel Eyraud Dubois [INRIA, Researcher]
- Yulia Gusak [INRIA, Researcher]
- Laercio Lima Pilla [CNRS, Researcher, from Dec 2024]

## Faculty Members

- Aurélien Esnard [UNIV BORDEAUX, Associate Professor]
- Mathieu Faverge [BORDEAUX INP, Associate Professor]
- Abdou Guermouche [UNIV BORDEAUX, Associate Professor]
- Pierre Ramet [UNIV BORDEAUX, Professor]
- Philippe Swartvagher [BORDEAUX INP, Associate Professor]

## Post-Doctoral Fellow

- Esragul Korkmaz [INRIA, Post-Doctoral Fellow, until Sep 2024]

## PhD Students

- Adrien Aguila–Multner [INRIA, from Oct 2024]
- Abel Anas Calluaud [CEA, CIFRE]
- Jean Francois David [INRIA]
- Alycia Lisito [BULL]
- Brieuc Nicolas [INRIA, from Oct 2024]
- Clément Richefort [CEA, until Oct 2024]
- Hayfa Tayeb [INRIA]
- Dimitri Walther [CEA, CIFRE, from Oct 2024]
- Xunyi Zhao [INRIA, until Nov 2024]

## Technical Staff

- Pierre Estérie [INRIA, Engineer]

### Interns and Apprentices

- Adrien Aguila–Multner [INRIA, Intern, from Feb 2024 until Aug 2024]
- Raphael Bourgouin [INRIA, Intern, from Sep 2024]
- Raphael Bourgouin [INRIA, Intern, from May 2024 until Jun 2024]
- Enrique Galvez [INRIA, Intern, from Sep 2024]
- Ana Hourcau [INRIA, Intern, from Mar 2024 until Sep 2024]
- Mohamed Kherraz [INRIA, Intern, until May 2024]
- Brieuc Nicolas [INRIA, Intern, from Mar 2024 until Aug 2024]
- Camille Ordronneau [INRIA, Intern, from May 2024 until Aug 2024]

### Administrative Assistant

- Catherine Cattaert Megrat [INRIA]

## 2 Overall objectives

The expertise of the team is at the heart of the issues between numerical simulations, training and HPC. In this context, the ability to effectively use the ever-increasing power of machines for numerical simulations (the shift to exascale for the next few years) is always central. These new platforms are characterized by their huge size (in terms of number of cores) and the heterogeneity of computing resources, with most of the computational power based on accelerators. We have largely anticipated these evolutions, and in particular, the different members of the team have been making efforts for several years to promote the use of dynamic runtimes such as StarPU, through a long-running collaboration with Storm project team. Runtime systems allow heterogeneous resources to be used transparently and allow some placement and scheduling decisions to be made dynamically, without the need to make static planning in advance. Indeed, such a fully static allocation would not be able to cope with the uncertainties of task and communication durations in increasingly complex environments and with increasingly shared resources. The question of scaling up these solutions, their use in (Neural Network) training and the effective management of large-scale distributed machines in particular, remains largely open.

As in many other fields, Machine Learning is changing the landscape at many levels. Training of large networks represents a new application for HPC because of the huge computational and memory needs it generates. Training has become a major source of use for converged HPC systems such as the Jean Zay supercomputer at IDRIS. If considered as an HPC workflow, it is an application that is quite different from traditional numerical simulation applications, because the calculations are tensor-based rather than matrix-based and because the nature of the dependencies makes parallelization more difficult and more intertwined with memory management issues.

On the other hand, ML plays a central role in the analysis of data, particularly data produced by large scientific instruments and large numerical simulations. In this context, it is important to bridge the data placement, resource allocation and computational scheduling strategies that are used to perform simulations and to perform data analysis. There again, we believe that dynamic runtime schedulers, coupled with static data placement strategies, are a relevant and promising tool. Finally, training represents a very important market, has a strong and growing influence on processor architectures, their accuracy and their arithmetics. This requires to further adapt the algorithms, the management of ever-increasing heterogeneity and the control of computational accuracy, both for classical numerical kernels and training deep neural networks.

Another major concern is the control of energy and carbon footprint minimizations. HPC is not naturally and historically an area of energy sobriety, but energy is a critical issue. Firstly, energy is a major subject because the race towards exascale has highlighted the difficulty of electrically powering

all these resources, and the increasing presence of dark silicon in computing resources makes resource allocation and power management problems extremely difficult. Furthermore, the minimization of our carbon footprint is a major societal issue and must be an axis of evaluation for our research. In this context, we believe that the solution cannot only be at the architecture and system levels, but that it is necessary to rethink parallel numerical kernels and algorithms in such a way as to allow prolonged use of the computing resources

Overall, the objective of the project is to transfer our historical expertise in linear algebra, runtime systems and combinatorial optimization (resource allocation, scheduling) to new problems (decompositions and tensor algebra, training in DNNs) which require a change of scale and new algorithms for new computing platforms (with different number representations and an ever increasing heterogeneity of computing resources). In addition, these new applications and new platforms require a central focus on data, since the gap between the costs (in energy and time) of storing and moving data compared to the costs of computation is always growing, which encourages innovative solutions (compression, redundant computation) that can in turn contribute to increasing the duration of use of computing resources.

# 3 Research program

## 3.1 Objectives

We propose to structure our research around two main application fields (see Section 4): **linear multi-dimensional algebra and solvers** on the one hand, and **training** in particular of deep learning networks on the other hand. In these two domains, our contributions will be organized around three main research axes (see Section 3.3): **the use of task based runtime systems** (to provide robust solutions and to increase the portability in the context of heterogeneous large scale platforms), **the use of compression** (to limit memory footprint and data transfers) and **the minimization of energy consumption and carbon impact** (using an approach of rewriting algorithms and placement strategies to limit data movements). This matrix organization of our activities (see Section 3.4) is intended to maximize the interactions between the different researchers of the team and facilitate knowledge sharing and joint participation in projects.

In these topics, the use of task based runtime systems and the design of efficient linear algebra kernels and solvers belong to the historical expertise of the team and is shared by all team members, especially in the context of linear algebra kernels. Our goal is to build on this expertise to extend the use of task based runtime systems to other types of applications such as training and to use the precise knowledge of these linear algebra kernels to incorporate new criteria such as energy minimization. The application to training (and interference) in deep neural networks and data compression are subjects we have been interested in for a few years, typically during the last HiePACS evaluation period and within the Inria Challenge of AI, HPC and Big Data led by Bruno Raffin. The extension of the techniques developed in linear algebra to tensor algebra and tensor decompositions is natural, given the proximity of the fields and the practical importance of the subject, but it is more recent and reinforced by the arrival of Julia Gusak, who is an expert in the field. Finally, the objective of energy and carbon footprint minimization, at the algorithmic and software levels rather than at the architecture level, is a field that we wish to emphasize in our research, both because of its own fundamental importance and because we believe that our expertise and the techniques that we have developed in recent years are well adapted to it and that the approach we propose is original.

## 3.2 Overall Positionning

The general positioning of the team is to **produce tools** for users, academic or industrial, in the form of algorithms and software libraries. These users can work either in numerical simulation or in training. Nevertheless, as our experiences in simulation and training have already demonstrated, this interaction cannot be carried out in the form of providing black boxes and it is crucial for us to work directly with the users of our software to understand their needs and adapt our algorithms and codes to the characteristics of their data. This interaction will be particularly critical to work on data representation and compression, which requires a strong interaction with numerical methods and machine learning in order to understand the application requirements and the characteristics of data, based on their significance.

At the other end of the spectrum, it is also essential for us to maintain close relationships with both the architecture and system communities. Indeed, the very rapid growth of machine learning applications has also renewed the landscape of computing resources with the emergence of very original solutions, at the architectural and arithmetic level. Even if we cannot influence on these evolutions, it is very important to propose solutions that make the best use of them. We also decided several years ago to rely on task based runtime systems to implement our software developments. This decision has many implications on our developments and requires an extremely close collaboration with their designers. In this context, we have co-supervised several PhD theses related to StarPU with the Storm project team and we will pursue this strategy, which is crucial in particular to take into account the challenges ahead of us: the transition to exascale, the integration of the energy, the extension to training applications and the ever increasing heterogeneity of computing resources.

## 3.3 Research Axes

### 3.3.1 Use of Runtime systems

**Participants:** Olivier Beaumont, Aurélien Esnard, Lionel Eyraud Dubois, Mathieu Faverge, Abdou Guermouche, Laércio Lima Pilla, Philippe Swartvagher.

In previous works, our main goal was to study the methodology needed to efficiently exploit the new generation of high-performance computers with all the constraints that it induces (number of cores, heterogeneity, co-scheduling effects, etc.). To achieve this goal, we successfully proposed a methodology based on the use of modern task-based runtime systems to ensure both portability and performance portability (the ability to achieve high performance by only tuning few parameters of the application). This work was done in the context of several projects (ANR Solhar, ANR SOLHARIS, Projet Région HPC Scalable Ecosystem, etc.). The work done mainly targeted single multicore nodes equipped with several accelerator devices and the extension of these techniques to the multi-node case will be the focus of our future works, especially with the arrival of Philippe Swartvagher in the team. Indeed, it has been observed that in the context of distributed nodes, the placement strategies of runtime systems are insufficient and generate too much communication. In this context, it is therefore crucial to develop efficient placement strategies [40, 34]. The extension of these mixed (static/dynamic) strategies in the case of tensors is largely open.

### 3.3.2 Design of compression techniques

**Participants:** Abdou Guermouche, Yulia Gusak, Mathieu Faverge, Pierre Ramet, Philippe Swartvagher.

The memory consumption of the applications has been and will remain an important challenge for solving larger problems that will lead to exascale computations. In the recent years we have demonstrated the interest of data compression techniques in linear solvers, both to save space and computations. Increasingly complex compression schemes require programming models to evolve to properly express the parallelism of these formats and to accommodate the increasing irregularity of applications. In TOPAL, we propose to continue the study of data compression techniques (low-rank, mixed precision, ...) in the context of solvers, but also in the context of training and multi-linear algebra. This part will be a very pertinent field for the study of applications over runtime systems, because of the strong irregularities that make the load balancing more complicated. At the same time, it is an original and promising approach for energy reduction. Representing convolutional / fully-connected weights in tensor formats is an effective way to reduce the parameters/FLOP in neural networks. However, post-quantization (reduction of parameters precision, for example, from float32 to int8) of networks with factorized weights yields a significant drop in accuracy. Due to memory/power consumption limitations of real devices, the quantization step is necessary, when pre-trained models are deployed. Therefore, our goal is to find

algorithms that build tensorized neural networks, where weight factors are directly contain elements in low-precision format. Efficient implementation of operations on tensors represented in low-bit format will be required, as well as development of regularization techniques to tackle instability issues when training deep learning models with low-bit weights.

#### 3.3.3 Energy minimization

**Participants:** Olivier Beaumont, Lionel Eyraud Dubois, Mathieu Faverge, Abdou Guermouche, Laércio Lima Pilla.

Running computations with resource frugality is an important challenge, both for the upcoming exascale shift and for generally reducing the carbon impact of scientific computing. In addition to the usual objective of making computations run faster, we thus intend to design and evaluate our techniques and algorithms with the purpose of limiting their carbon footprint. In particular, given the lasting trend that the time and energy costs of computing are becoming ever lower than the costs of accessing and communicating data, we want to explore the tradeoffs of trading more computation for less data movements. This can be achieved in several ways: compression techniques as described above, replication of some computations, or use of lower precision. We are planning to work on this issue from two points of views: more frugal numerical algorithms, and energy-aware scheduling techniques. As for the embedded architectures in the phone, but also in the latest generation of laptops (Apple M1 Pro and Max chips), we are starting to see the emergence of Big-Little type technologies in the design of HPC oriented chips. In general, thermal design power (TDP) constraints push architects to increase the diversity and number of energy efficient circuits, even if they cannot all be powered simultaneously. If this hardware solution is very debatable from the point of view of carbon impact, it raises difficult and original questions about the optimization of computing performance under energy constraints. This kind of approach opens new perspectives, both from the point of view of scheduling algorithms but also in the design of computational kernels in linear algebra. We are also seeing the emergence of new processors (ARM or RISC-V technologies, Rhea from the SiPearl company within the EPI consortium, which should seriously compete with the supremacy of x86 architectures (Intel and AMD) with Nvidia accelerator cards in the search for a compromise between pure performance and energy sobriety.

In the field of training, a complementary opportunity is available. Indeed, contrary to classical HPC, the renewal of computational resources is often linked to the need to run larger models (and data with a better resolution to a lesser extent), rather than by the acceleration of computations. In this context, the possibility offered by tools such as Rotor 7.1.4 to limit memory requirements contributes to limiting the carbon footprint. Our goal is to extend the scope of these techniques, including to other fields of application than training. Our collaboration with Qarnot Computing is consistent with this objective. The co-design environment of the TextaRossa and Eupex projects 10 are also great avenues to explore these questions.

### 3.4 Main Research Topics

The list of our contributions can be read at the intersection of the research domains described in Section 4 and research axes described in Section 3.3 as shown in the following table:

| | Axis 3.3.1 – Runtime | Axis 3.3.2 – Compression | Axis 3.3.3 – Energy |
|---|---|---|---|
| Domain 4.1 – Linear Algebra, Tensors | Topic 3.4.1 | Topic 3.4.2 | Topic 3.4.3 |
| Domain 4.2 – Training of DNNs | Topic 3.4.4 | Topic 3.4.5 | Topic 3.4.6 |

#### 3.4.1 Task-based Linear Algebra and Tensor Computations

**Participants:** Olivier Beaumont, Aurélien Esnard, Lionel Eyraud Dubois, Mathieu Faverge, Abdou Guermouche, Pierre Ramet, Philippe Swartvagher.

We plan to continue our activity on task-based linear algebra to find solutions for expressing high level algorithms in an elegant way while ensuring high performance. First, we want to consider the expressivity of the algorithms for large scale distributed architectures while considering the specific problems of scheduling, data and task mapping, and data granularity. This work will be done in tight collaboration with the Storm and Tadaam teams and is a key objective of the ANR SOLHARIS project. Moreover, the foundations of this topic fall back to the HiePACS project. Thus, we plan to collaborate and exchange with the CONCACE team on topics which are of interest to both teams (mainly expressivity and scalability). Second, as mentioned above, we plan to study data compression techniques in linear algebra [47, 52, 56], which brings new algorithmic schemes that are outside of the scope of the classical programming model used until now. As mid and long term objectives, we would like to find new ways to express these linear algebra algorithms to efficiently exploit large heterogeneous architectures. A second research topic focuses on the extension of the techniques developed in the framework of linear algebra, in particular with the Chameleon library, to multi-linear algebra and tensors. The idea is to build on the expertise we have in the field of compression and in the use of runtimes to use heterogeneous resources in particular.

Another challenge would be to redesign the graph partitioning & matrix ordering algorithms in a task-based runtime, in order to facilitate the integration of this basic building block in modern tasked-based solvers. This work has already been initiated in the StarPart 7.1.2 project.

### 3.4.2 Multi-Linear Algebra and Tensor Decompositions

**Participants:** Olivier Beaumont, Lionel Eyraud Dubois, Mathieu Faverge, Abdou Guermouche, Julia Gusak, Pierre Ramet.

Tensor decompositions are a natural extension of SVD-type decompositions in linear algebra. Unlike linear algebra, there are several types of decompositions, which play an important role in the analysis of large data and in the compression of networks, in particular to increase the efficiency of inference. The arrival of Julia Gusak in the project allows us to reinforce this competence. In addition to the basic kernels to be integrated in Chameleon proposed in the Topic 3.4.1, we will propose distributed tensor decomposition algorithms compression algorithms, focusing mainly on the case of small but large tensors, which is the most common in the context of neural networks.

### 3.4.3 Energy Minimization in Linear Solvers

**Participants:** Mathieu Faverge, Abdou Guermouche.

We plan to investigate how to reduce the energy consumption of linear algebra libraries (either sparse or dense). To do so we will rely on an algorithmic approach rather than a system approach. The idea, in a first step, is to consider several implementations of a same kernel and select the implementation while taking into account energy consumption [32, 31, 33]. For instance a low-rank implementation of a given operation will be slower than a regular high-performance implementation but it will tend to require less energy. In the longer term, we plan also to investigate how to design energy efficient implementations of basic kernels. They will then be used within higher level algorithms in order to find a better trade-off between energy consumption and high performance. In the context of developing linear algebra solvers using compression techniques, a research axis we would like to develop is the energy consumption study of these solvers: is it possible to provide computation kernels with different energy consumption levels that can be easily exchanged to lower the final energy consumption of the application while keeping the same numerical accuracy. Low-rank compression techniques, as well as mixed-precision solution are envisioned toward this objective.

#### 3.4.4 Task-based Approaches for Deep Learning

**Participants:** Olivier Beaumont, Lionel Eyraud Dubois, Mathieu Faverge, Julia Gusak, Abdou Guermouche, Laércio Lima Pilla, Pierre Ramet, Philippe Swartvagher.

In popular Deep Learning frameworks like TensorFlow or PyTorch, the parallelization of the training process is performed with a large granularity, mostly relying on Data Parallelism. Specialized frameworks have been proposed to explore finer parallel schemes, like PipeDream for model parallelism [58]. These implementations are however very static and require explicit and error-prone data management policies. We believe that our expertise in using task-based runtime systems can be used to provide much simpler approaches for a finer grain control on the execution of the corresponding task graphs and communications patterns, for both training and inference phases. We plan to design a prototype implementation that would allow to easily use clever scheduling and optimization techniques to improve the performance of inference. In the longer term, we expect that this approach will provide better scalability and flexibility, and unlock new opportunities for optimization, for a wide range of deep learning applications.

#### 3.4.5 Tensor Compression for Inference

**Participants:** Olivier Beaumont, Julia Gusak.

We envision a more exploratory research activity around the use of tensor compression for inference. Initially, the objective is to use tensor compression techniques and quantization to allow inference to be performed with little memory or low latency. These techniques can also be further extended in the context of online training performed after installation on the device itself, which requires in particular memory-saving approaches. Finally, an even more ambitious goal would be to combine these approaches with techniques for designing neural networks that are inherently efficient in terms of memory needs, such as extensions of RevNets [48, 51, 60].

#### 3.4.6 Carbon Saving and Energy-Efficient Training

**Participants:** Olivier Beaumont, Lionel Eyraud Dubois, Julia Gusak, Laércio Lima Pilla.

The training phase of Deep Neural Networks is notoriously very resource-hungry, especially regarding its energy consumption. In the last years, we have proposed several algorithmic solutions (re-materialization [6], offloading [37], their combination [35], pipelining [38]) to reduce the resource consumption of this training phase, with a focus on reducing the training time. We plan to broaden the scope of these studies, by also taking into account the energy usage. A heterogeneous context and a flexible runtime system, as planned in Topic 3.4.4, may also be an opportunity to reduce energy consumption by allocating some tasks, typically the non-critical ones, to the most efficient resources for them, or by selecting a different implementation with better energy efficiency. This can be seen as a generalization of mixed-precision techniques, which are also very popular in this context to help achieving a better frugality. However, care must be taken to not degrade the convergence of the training phase. Moreover, the carbon footprint comes essentially from the manufacturing [59, 50] of the computing resources (GPUs) and the main goal is to facilitate their non-renewal, as enabled by memory saving techniques.

# 4 Application domains

## 4.1 Multi-Linear Algebra and Solvers

**Participants:** Olivier Beaumont, Aurélien Esnard, Lionel Eyraud Dubois, Mathieu Faverge, Abdou Guermouche, Julia Gusak, Pierre Ramet, Philippe Swartvagher.

At the core of a large number of simulation tools, the resolution of large linear systems often represents the dominant part of the computing time. These linear solvers rely on a wide variety of numerical methods and algorithms. Massively parallel versions are required to support advances in multi-physics and multi-scale simulations, especially when targeting exascale platforms. The aim is therefore to address the major challenge of designing and building numerically robust solvers on top of runtime systems that can scale up and push back the limits of existing industrial codes by making full use of all computing resources such as CPUs, GPUs and other accelerator units. Following the ANR project SOLHARIS (and previously SOLHAR), we now have experience of strong/weak scalability of sparse direct solvers on large scale, distributed memory, heterogeneous computers. These solvers already rely on asynchronous task-based parallelism [29, 30, 55, 28], rather than traditional and widely adopted message-passing and multithreading techniques. Indeed, the use of modern runtime systems have proven to be good tools for the development of scientific computing applications [57, 42, 63], in particular in combination with compression [43, 62, 61, 39, 53] and communication avoiding techniques [40, 34]. This work can be extended naturally to multi-dimensional objects such as tensors. In the tensor case, we propose to extend the data distribution strategies to minimize communication and the use of system runtimes to handle the variability and heterogeneity of computational resources. Finally, we have focused so far on minimizing the execution time, whereas energy efficiency is becoming a critical element. We therefore plan to revisit the algorithms and methods we developed in linear algebra, and those we propose to design for handling tensors, to allow the optimal use of the available hardware in order to guarantee the performance of the computations within a fixed energy budget.

## 4.2 Training and Inference for DNNs

**Participants:** Olivier Beaumont, Lionel Eyraud Dubois, Julia Gusak, Laércio Lima Pilla, Pierre Ramet, Philippe Swartvagher.

The training phase in Deep Neural Networks has become an important source of HPC resource usage and it is crucial to perform it efficiently on parallel architectures. Until today, data parallelism is the most widely used method, but the associated requirement to replicate all the weights on all computing resources causes memory issues at the level of each node and of collective communications at the level of the platform.

In general, the overall shape of the dependency graphs associated with the feed forward training phase has characteristics (long dependencies) that generate a lot of memory needs and data exchange. However, there are multiple opportunities to address these problems by combining [35] re-computations [44, 6, 41, 54, 49], offloading [37], compression and different parallelism strategies (image, filter, kernel, model parallelism [38, 58, 36, 51]). It is also promising to consider other more radical techniques to go beyond feed forward training, such as the use of multigrid reduction in time (MGRIT) [45, 46] that come from the field of numerical simulations and that we already address in other contexts.

Within this general framework, the minimization of carbon footprint is obviously a major concern that must guide strategies. Tools to train complex and deep network on otherwise obsolete hardware using memory saving techniques are already a strong contribution in this direction to increase the lifetime of computing resources. and our goal is to extend these techniques in terms of efficiency and in terms of scope, which has consumed a little more energy associated with the computations. As in the case of linear algebra, energy optimization also requires the use of heterogeneous computation resources (CPUs, GPUs, TPUs, FPGAs). Conversely, this heterogeneity hinders scalability because of difficulties in predicting task durations and makes the use of dynamic runtime schedulers necessary. Finally, the use of these dynamic runtimes also poses the problem of knowing what needs to be decided statically and dynamically in terms of resource allocation and scheduling.

# 5 Social and environmental responsibility

## 5.1 Footprint of research activities

As part of our research activities, we use local computing resources such as PlaFRIM and the national computing resources of IDRIS and the TGCC.

The environmental impact of using these platforms is significant, whether for numerical simulation or training applications. However, the positioning of the team, which produces simulation and training tools but does not directly perform simulations and training, is relatively limited. For example, in the case of training, we have so far concentrated on techniques that do not modify the architecture of the networks and the computations that are performed, so that the number of epochs and the final accuracy are not impacted. In this way, it is possible to validate our developments to accelerate training on a single batch (at full machine scale) and then to extrapolate the acceleration at the whole training scale. Similarly, the techniques developed in linear algebra in the team often do not depend (typically for dense approaches) on the numerical properties of the matrices, so that acceleration (for a given problem size) can be validated without heavy experimental campaigns, beyond what is necessary to obtain valid experimental results in complex environments where performance varies from one experiment to another.

In this context, the use of simulation as opposed to direct experimentation is also a tool that enables us to limit the impact of our research on power consumption, since simulation can save several orders of magnitude in power consumption compared with direct experimentation. In this context, it is crucial to produce simulation tools that are as precise and generic as possible, and the team has been actively collaborating for many years in the development of simulation tools such as SimGRID.

Nevertheless, the tools we produce are used on a large scale in terms of computation resources and simulation/training time, and the associated energy consumption issue is therefore indirectly crucial. In this context, we are developing original solutions for reusing the heat dissipated by computation resources, in particular as part of the Inria-Qarnot Computing Pulse challenge (see Section 5.2). We have also added a research axis aimed at minimizing energy consumption for a given kernel (Section 3.3.3).

TOPAL has also signed the "Labos en transitions" Charter of Commitment for research facilities on the Bordeaux university site whose preamble states that "Faced with contemporary environmental and societal challenges, and the urgent need for systemic transformation to meet them, the academic world has a particular responsibility: to promote responsible research, aware of environmental issues and respectful of the people who produce it, which contributes to transitions and enables us to understand and guide current and future societal transformations". In exchange for this commitment, the establishments undertake to provide us with an estimate of the impact of our research activities (including the purchase of equipment and missions). At this stage, this information is difficult to aggregate at team level, but making it available will enable us to measure our progress and involvement.

## 5.2 Impact of research results

### 5.2.1 Carbon Impact of Cloud Platforms

To limit the environmental impact of Qarnot focuses on re-using the heat produced by computations in heat circuits or boilers. As part of the Pulse Inria challenge, we are working with Qarnot on algorithms for placing computations on their infrastructure, so as to maximize the use of reusable heat sources, depending on computation demand and task characteristics. The aim is to enable users of the Qarnot platform to specify their objective function on the (carbon footprint, time, cost) axes, and to be able to meet it.

### 5.2.2 Democratization of Large Models Training

In the context of training, at one end of the spectrum we see the provision of computing resources, such as the Jean Zay supercomputer, whose efficient use requires large-scale parallel training algorithms and frameworks to optimize resource utilization and accelerate time to discovery. At the other end of the spectrum, we see the importance of enabling researchers from different communities to use the resources at their disposal (often just a few GPUs) to develop original models without being constrained by hardware limitations. In particular, recent transformer-based models are very heavy-weight, and

techniques must be employed to run them on GPUs that are only a few years old, without compromising data quality, computational accuracy, or model size. In particular, the Topal team has been working for several years on memory-saving strategies to enable the training of large models on limited-capacity resources (re-materialization and offloading), and on software 7 such as Rotor and Rockmate, which are recognized and visible in the AI applications community and enable researchers with access to limited capacity resources to train large models.

# 6 Highlights of the year

In 2024, the Topal team continued its involvement in European EuroHPC projects. The Textarossa project came to a positive conclusion, while the Eupex project was extended to take account of delays in the provision of hardware. We applied for the Specific Grant Agreement (SGA) for the development of a large-scale European initiative for HPC ecosystem based on RISC-V as part of the DARE consortium, which was selected to start in early 2025.

As part of the collaboration with Qarnot Computing and the Pulse challenge, we have proposed a new algorithm for selecting which set of tasks to run on a Qarnot boiler and which set of tasks to delegate to conventional computing resources, so as to maximize the useful heat produced. The corresponding paper [7], submitted to the Euro-Par 2024 conference and co-authored by researchers from Inria and Qarnot Computing, was named best paper finalist.

In terms of scientific organization, Julia Gusak organized the second edition of the WANT Workshop during the ICML 2024 conference (the first edition took place during NeurIPS 2023 conference). This series of Workshops, at the intersection between AI and HPC, are dedicated to Computational Efficiency, Scalability, and Resource Optimization in large neural networks (both for inference and training). They attracted more than 500 offline participants (+ online), 100+ contributed papers, 150+ reviewers involved.

In 2024, Thomas Herault, Research Assistant Professor at the University of Knoxville in Jack Dongarra's ICL laboratory, has applied for a position as DR in the Topal team, and will take up his post on 1/1/2025. He will bring us his skills in task-based runtimes, communication libraries and fault tolerance. In addition, Laercio Lima Pilla, CR CNRS at LaBRI and with whom we have been collaborating for many years, in particular as part of the Pulse Inria Qarnot Computing challenge, joined the team in December. Laercio will bring us his skills in scheduling, energy optimization and Federated Learning in particular.

# 7 New software, platforms, open data

## 7.1 New software

### 7.1.1 Chameleon

**Keywords:** Runtime system, Task-based algorithm, Dense linear algebra, HPC, Task scheduling

**Scientific Description:** Chameleon is part of the MORSE (Matrices Over Runtime Systems @ Exascale) project. The overall objective is to develop robust linear algebra libraries relying on innovative runtime systems that can fully benefit from the potential of those future large-scale complex machines.

We expect advances in three directions based first on strong and closed interactions between the runtime and numerical linear algebra communities. This initial activity will then naturally expand to more focused but still joint research in both fields.

1. Fine interaction between linear algebra and runtime systems. On parallel machines, HPC applications need to take care of data movement and consistency, which can be either explicitly managed at the level of the application itself or delegated to a runtime system. We adopt the latter approach in order to better keep up with hardware trends whose complexity is growing exponentially. One major task in this project is to define a proper interface between HPC applications and runtime systems in order to maximize productivity and expressivity. As mentioned in the next section, a widely used approach consists in abstracting the application as a DAG that the runtime system is in charge of scheduling. Scheduling such a DAG over a set of heterogeneous processing units

introduces a lot of new challenges, such as predicting accurately the execution time of each type of task over each kind of unit, minimizing data transfers between memory banks, performing data prefetching, etc. Expected advances: In a nutshell, a new runtime system API will be designed to allow applications to provide scheduling hints to the runtime system and to get real-time feedback about the consequences of scheduling decisions.

2. Runtime systems. A runtime environment is an intermediate layer between the system and the application. It provides low-level functionality not provided by the system (such as scheduling or management of the heterogeneity) and high-level features (such as performance portability). In the framework of this proposal, we will work on the scalability of runtime environment. To achieve scalability it is required to avoid all centralization. Here, the main problem is the scheduling of the tasks. In many task-based runtime environments the scheduler is centralized and becomes a bottleneck as soon as too many cores are involved. It is therefore required to distribute the scheduling decision or to compute a data distribution that impose the mapping of task using, for instance the so-called "owner-compute" rule. Expected advances: We will design runtime systems that enable an efficient and scalable use of thousands of distributed multicore nodes enhanced with accelerators.

3. Linear algebra. Because of its central position in HPC and of the well understood structure of its algorithms, dense linear algebra has often pioneered new challenges that HPC had to face. Again, dense linear algebra has been in the vanguard of the new era of petascale computing with the design of new algorithms that can efficiently run on a multicore node with GPU accelerators. These algorithms are called "communication-avoiding" since they have been redesigned to limit the amount of communication between processing units (and between the different levels of memory hierarchy). They are expressed through Direct Acyclic Graphs (DAG) of fine-grained tasks that are dynamically scheduled. Expected advances: First, we plan to investigate the impact of these principles in the case of sparse applications (whose algorithms are slightly more complicated but often rely on dense kernels). Furthermore, both in the dense and sparse cases, the scalability on thousands of nodes is still limited, new numerical approaches need to be found. We will specifically design sparse hybrid direct/iterative methods that represent a promising approach.

Overall end point. The overall goal of the MORSE associate team is to enable advanced numerical algorithms to be executed on a scalable unified runtime system for exploiting the full potential of future exascale machines.

**Functional Description:** Chameleon is a dense linear algebra software relying on sequential task-based algorithms where sub-tasks of the overall algorithms are submitted to a Runtime system. A Runtime system such as StarPU is able to manage automatically data transfers between not shared memory area (CPUs-GPUs, distributed nodes). This kind of implementation paradigm allows to design high performing linear algebra algorithms on very different type of architecture: laptop, many-core nodes, CPUs-GPUs, multiple nodes. For example, Chameleon is able to perform a Cholesky factorization (double-precision) at 80 TFlop/s on a dense matrix of order 400 000 (i.e. 4 min 30 s).

**Release Contributions:** Chameleon includes the following features:

- BLAS 3, LAPACK one-sided and LAPACK norms tile algorithms - Support QUARK and StarPU runtime systems and PaRSEC since 2018 - Exploitation of homogeneous and heterogeneous platforms through the use of BLAS/LAPACK CPU kernels and cuBLAS/MAGMA CUDA kernels - Exploitation of clusters of interconnected nodes with distributed memory (using OpenMPI)

**URL:** https://gitlab.inria.fr/solverstack/chameleon

**Contact:** Mathieu Faverge

**Participants:** Samuel Thibault, Emmanuel Agullo, Florent Pruvost, Mathieu Faverge

**Partners:** Innovative Computing Laboratory (ICL), King Abdullha University of Science and Technology, University of Colorado Denver

#### 7.1.2 StarPart

**Functional Description:** StarPart is a flexible and extensible framework that integrates state-of-the-art methods for graph partitioning and sparse matrix ordering. More precisely, StarPart is a framework that offers a uniform API to manipulate graph, hypergraph and mesh structures. It is designed to be easily extensible by adding new methods and to plug all these methods into a comprehensive framework. It is initially designed to provide graph partitioning and sparse matrix ordering methods, that come from sate-of-the-art software such as Metis, Scotch, Patoh, Zoltan, etc. Besides, it provides some facilities for IO, diagnostic, benchmark, visualization (VTK, SVG, ...). StarPart is the core of the MetaPart project. It is built upon the LibGraph library.

**URL:** https://gitlab.inria.fr/metapart/starpart

**Contact:** Aurélien Esnard

**Participant:** Aurélien Esnard

#### 7.1.3 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, hal-04527103

**Contact:** Pierre Ramet

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

#### 7.1.4 rotor

**Name:** Re-materializing Optimally with pyTORch

**Keywords:** Deep learning, Optimization, Python, GPU, Automatic differentiation

**Scientific Description:** This software implements in PyTorch a new activation checkpointing method which allows to significantly decrease memory usage when training Deep Neural Networks with the back-propagation algorithm. Similarly to checkpointing techniques coming from the literature on Automatic Differentiation, it consists in dynamically selecting the forward activations that are saved during the training phase, and then automatically recomputing missing activations from those previously recorded. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs (this uses more memory, but requires fewer recomputations in the backward phase), and we provide in https://hal.inria.fr/hal-02352969 an algorithm to compute the optimal computation sequence for this model.

Our PyTorch implementation processes the entire chain, dealing with any sequential DNN whose internal layers may be arbitrarily complex and automatically executing it according to the optimal checkpointing strategy computed given a memory limit. In https://hal.inria.fr/hal-02352969, through extensive experiments, we show that our implementation consistently outperforms existing checkpoint-ing approaches for a large class of networks, image sizes and batch sizes.

**Functional Description:** Allows to train very large convolutional networks on limited memory by optimally selecting which activations should be kept and which should be recomputed. This code is meant to replace the checkpoint.py utility available in pytorch, by providing more efficient rematerialization strategies. The algorithm is easier to tune: the only required parameter is the available memory, instead of the number of segments.

**URL:** https://gitlab.inria.fr/hiepacs/rotor

**Publication:** hal-02352969

**Contact:** Lionel Eyraud Dubois

**Participants:** Olivier Beaumont, Alena Shilova, Alexis Joly, Lionel Eyraud Dubois, Julien Herrmann

#### 7.1.5 StarPU

**Name:** The StarPU Runtime System

**Keywords:** Runtime system, High performance computing

**Scientific Description:** Traditional processors have reached architectural limits which heterogeneous multicore designs and hardware specialization (eg. coprocessors, accelerators, ...) intend to address. However, exploiting such machines introduces numerous challenging issues at all levels, ranging from programming models and compilers to the design of scalable hardware solutions. The design of efficient runtime systems for these architectures is a critical issue. StarPU typically makes it much easier for high performance libraries or compiler environments to exploit heterogeneous multicore machines possibly equipped with GPGPUs or Cell processors: rather than handling low-level issues, programmers may concentrate on algorithmic concerns.Portability is obtained by the means of a unified abstraction of the machine. StarPU offers a unified offloadable task abstraction named "codelet". Rather than rewriting the entire code, programmers can encapsulate existing functions within codelets. In case a codelet may run on heterogeneous architectures, it is possible to specify one function for each architectures (eg. one function for CUDA and one function for CPUs). StarPU takes care to schedule and execute those codelets as efficiently as possible over the entire machine. In order to relieve programmers from the burden of explicit data transfers, a high-level data management library enforces memory coherency over the machine: before a codelet starts (eg. on an accelerator), all its data are transparently made available on the compute resource.Given its expressive interface and portable scheduling policies, StarPU obtains portable performances by efficiently (and easily) using all computing resources at the same time. StarPU also takes advantage of the heterogeneous nature of a machine, for instance by using scheduling strategies based on auto-tuned performance models.

StarPU is a task programming library for hybrid architectures.

The application provides algorithms and constraints: - CPU/GPU implementations of tasks, - A graph of tasks, using StarPU's rich C API.

StarPU handles run-time concerns: - Task dependencies, - Optimized heterogeneous scheduling, - Optimized data transfers and replication between main memory and discrete memories, - Optimized cluster communications.

Rather than handling low-level scheduling and optimizing issues, programmers can concentrate on algorithmic concerns!

**Functional Description:** StarPU is a runtime system that offers support for heterogeneous multicore machines. While many efforts are devoted to design efficient computation kernels for those architectures (e.g. to implement BLAS kernels on GPUs), StarPU not only takes care of offloading such kernels (and implementing data coherency across the machine), but it also makes sure the kernels are executed as efficiently as possible.

**Release Contributions:** StarPU is a runtime system that offers support for heterogeneous multicore machines. While many efforts are devoted to design efficient computation kernels for those architectures (e.g. to implement BLAS kernels on GPUs), StarPU not only takes care of offloading such kernels (and implementing data coherency across the machine), but it also makes sure the kernels are executed as efficiently as possible.

**URL:** https://starpu.gitlabpages.inria.fr/

**Publications:** tel-04213186, inria-00326917, inria-00378705, inria-00384363, inria-00411581, inria-00421333, inria-00467677, inria-00523937, inria-00547614, inria-00547616, inria-00547847, inria-00550877, inria-00590670, inria-00606195, inria-00606200, inria-00619654, hal-00643257, hal-00648480, hal-00654193, hal-00661320, hal-00697020, hal-00714858, hal-00725477, hal-00772742, hal-00773114, hal-00773571, hal-00773610, hal-00776610, tel-00777154, hal-00803304, hal-00807033, hal-00824514, hal-00851122, hal-00853423, hal-00858350, hal-00911856, hal-00920915, hal-00925017, hal-00926144, tel-00948309, hal-00966862, hal-00978364, hal-00978602, hal-00987094, hal-00992208, hal-01005765, hal-01011633, hal-01081974, hal-01101045, hal-01101054, hal-01120507, hal-01147997, tel-01162975, hal-01180272, hal-01181135, hal-01182746, hal-01223573, tel-01230876, hal-01283949, hal-01284004, hal-01284136, hal-01284235, hal-01316982, hal-01332774, hal-01353962, hal-01355385, hal-01361992, hal-01372022, hal-01386174, hal-01387482, hal-01409965, hal-01410103, hal-01473475, hal-01474556, tel-01483666, hal-01502749, hal-01507613, hal-01517153, tel-01538516, hal-01616632, hal-01618526, hal-01718280, tel-01816341, hal-01842038, tel-01959127, hal-02120736, hal-02275363, hal-02296118, hal-02403109, hal-02421327, hal-02872765, hal-02914793, hal-02933803, hal-02943753, hal-02970529, hal-02985721, hal-03144290, hal-03273509, hal-03290998, hal-03298021, hal-03318644, hal-03348787, hal-03552243, hal-03609275, hal-03623220, hal-03773486, hal-03773985, hal-03789625, hal-03936659, tel-03989856, hal-04005071, hal-04088833, hal-04115280, hal-04146714, hal-04236246, tel-04260094, tel-04316145, hal-04548787, hal-04646530, hal-04668550, hal-04690154

**Contact:** Nathalie Furmento

**Participants:** Cedric Augonnet, Olivier Aumage, Nathalie Furmento, Samuel Thibault, Simon Archipoff, Bérenger Bramas, Alfredo Buttari, Jérôme Clet-Ortega, Terry Cojean, Nicolas Collin, Camille Coti, Ludovic Courtes, Alexandre Denis, Lionel Eyraud Dubois, Maxime Gonthier, Amina Guermouche, Kun He, Sylvain Henry, Andra Hugo, Antoine Jego, LoÏc Jouans, Mehdi Juhoor, Yanis Khorsi, Xavier Lacoste, Romain Lion, Benoit Lize, Gwenole Lucas, Mariem Makni, Thomas Morin, Raymond Namyst, Cyril Roelandt, Corentin Salingue, Lucas Schnorr, Marc Sergent, Luka Stanisic, Ludovic Stordeur, Philippe Swartvagher, François Tessier, Leo Villeveygoux, Philippe Virouleau, Pierre Wacrenier

#### 7.1.6 VITE

**Name:** Visual Trace Explorer

**Keywords:** Visualization, Execution trace

**Functional Description:** ViTE is a trace explorer. It is a tool made to visualize execution traces of large parallel programs. It supports Pajé, a trace format created by Inria Grenoble, and OTF and OTF2 formats, developed by the University of Dresden and allows the programmer a simpler way to analyse, debug and/or profile large parallel applications.

**URL:** https://solverstack.gitlabpages.inria.fr/vite/

**Publications:** hal-00707236, hal-04725983

**Contact:** Mathieu Faverge

**Participants:** Mathieu Faverge, Philippe Swartvagher

#### 7.1.7 pmtool

**Keywords:** Scheduling, Task scheduling, StarPU, Heterogeneity, GPGPU, Performance analysis

**Functional Description:** Analyse post-mortem the behavior of StarPU applications. Provide lower bounds on makespan. Study the performance of different schedulers in a simple context. Provide implementations of many scheduling algorithms from the literature

**URL:** https://gitlab.inria.fr/eyrauddu/pmtool

**Publications:** hal-01386174, hal-01878606

**Contact:** Lionel Eyraud Dubois

**Participant:** Lionel Eyraud Dubois

#### 7.1.8 rockmate

**Name:** rockmate

**Keywords:** Deep learning, Optimization, Python, Pytorch, GPU, Automatic differentiation

**Scientific Description:** We propose Rockmate to control the memory requirements when training PyTorch DNN models. Rockmate is an automatic tool that starts from the model code and generates an equivalent model, using a predefined amount of memory for activations, at the cost of a few re-computations. Rockmate automatically detects the structure of computational and data dependencies and rewrites the initial model as a sequence of complex blocks. We show that such a structure is widespread and can be found in many models in the literature (Transformer based models, ResNet, RegNets,...). This structure allows us to solve the problem in a fast and efficient way, using an adaptation of Checkmate (too slow on the whole model but general) at the level of individual blocks and an adaptation of Rotor (fast but limited to sequential models) at the level of the sequence itself. We show through experiments on many models that Rockmate is as fast as Rotor and as efficient as Checkmate, and that it allows in many cases to obtain a significantly lower memory consumption for activations (by a factor of 2 to 5) for a rather negligible overhead (of the order of 10% to 20%). Rockmate is open source and available at https://github.com/topal-team/rockmate.

Complete paper: https://openreview.net/pdf?id=wLAMOoL0KD

**Functional Description:** Given a PyTorch model, a sample input, and a GPU memory budget, Rockmate builds a new torch.nn.Module, which performs forward and backward pass while keeping the memory of activations under the given budget.

The new model produces the same outputs and gradients as the original one. Training the model with a lower memory than PyTorch Autodiff is achieved by re-computing some of the activations instead of storing them for gradient calculation. Based on the budget, Rockmate determines automatically which activations should be recomputed.

**URL:** https://github.com/topal-team/rockmate

**Contact:** Lionel Eyraud Dubois

# 8 New results

As explained in Section 3.4, our contributions can be read at the intersection of the research domains described in Section 4 and research axes described in Section 3.3 as shown in the following table:

| | Axis 3.3.1 – Runtime | Axis 3.3.2 – Compression | Axis 3.3.3 – Energy |
|---|---|---|---|
| Domain 4.1 – Linear Algebra, Tensors | Topic 3.4.1 | Topic 3.4.2 | Topic 3.4.3 |
| Domain 4.2 – Training of DNNs | Topic 3.4.4 | Topic 3.4.5 | Topic 3.4.6 |

## 8.1 Enhancing sparse direct solver scalability through runtime system automatic data partition (Topic 3.4.1)

**Participants:** Alycia Lisito, Mathieu Faverge, Pierre Ramet.

In [14], with the ever-growing number of cores per node, it is critical for runtime systems and applications to adapt the task granularity to scale on recent architectures. Among applications, sparse direct solvers are a time-consuming step and the task granularity is rarely adapted to large many-core systems. In this paper, we investigate the use of runtime systems to automatically partition tasks in order to achieve more parallelism and refine the task granularity. Experiments are conducted on the new version of the PaStiX solver, which has been completely rewritten to better integrate modern task-based runtime systems. The results demonstrate the increase in scalability achieved by the solver thanks to the adaptive task granularity provided by the StarPU runtime system.

This work has been presented in the Workshop on Asynchronous Many-Task Systems and Applications, Feb 2024, Knoxville, United States.

## 8.2 Toward an algebraic multigrid method for the indefinite Helmholtz equation (Topic 3.4.2)

**Participants:** Clement Richefort, Pierre Ramet.

In [17], it is well known that multigrid methods are very competitive in solving a wide range of SPD problems. However achieving such performance for non-SPD matrices remains an open problem. In particular, three main issues may arise when solving a Helmholtz problem : some eigenvalues may be negative or even complex, requiring the choice of an adapted smoother for capturing them, and because the near-kernel space is oscillatory, the geometric smoothness assumption cannot be used to build efficient interpolation rules. Moreover, the coarse correction is not equivalent to a projection method since the indefinite matrix does not define a norm. We present some investigations about designing a method that converges in a constant number of iterations with respect to the wavenumber. The method builds on an ideal reduction-based framework and related theory for SPD matrices to improve an initial least squares minimization coarse selection operator formed from a set of smoothed random vectors. A new coarse correction is proposed to minimize the residual in an appropriate norm for indefinite problems. We also present numerical results at the end of the paper.

This paper has been submitted to SIAM SISC.

## 8.3 Comparative Study of Mixed-Precision and Low-Rank Compression Techniques in Sparse Direct Solvers (Topic 3.4.2)

**Participants:** Brieuc Nicolas, Mohamed Kherraz, Alycia Lisito, Mathieu Faverge, Pierre Ramet.

In [18], sparse direct solvers play a crucial role in numerical simulation and are one of the most time-consuming steps in many applications. Recently, many efforts have been made to reduce the complexity of dense and sparse direct solvers by introducing low-rank compression techniques. These techniques allow applications to reduce the amount of information stored in the matrix, depending on the quality of the solution being sought, and greatly reduce both memory requirements and computational complexity. Another solution driven by the computational capabilities of the hardware is the use of mixed-precision computations. This has been continuously in vogue with new generations of hardware providing single to double performance ratios of more than two. In our study, we extend the sparse direct solver PaStiX to use reduced-precision factorization and compare it to its low-rank strategy in terms of time to solution, numerical stability, memory consumption, and energy consumption. The goal of this study is to evaluate whether the tradeoff between computational speed and solution accuracy is worthwhile, and if so, which strategy, low-rank or mixed-precision, is better.

This work has been presented in SIAM Conference on Applied Linear Algebra, May 2024, Paris, France.

## 8.4 Scalable and Portable LU Factorization with Partial Pivoting on top of Runtime Systems (Topic 3.4.2)

**Participants:** Alycia Lisito, Mathieu Faverge, Pierre Ramet.

In [25, 13], Task-based runtime systems have demonstrated efficiency in leveraging the capabilities of large, heterogeneous architectures. Many linear algebra algorithms and applications have been implemented on top of runtime systems to increase their performance. However, the High Performance Linpack (HPL) benchmark, used by the TOP500 to rank supercomputers, has not yet been successfully implemented using task-based runtime systems. In this paper, we explore solutions to implement efficient LU factorization with partial pivoting using the sequential task flow programming model. We show that, due to the pivoting strategy, this algorithm generates a large number of very small tasks, which usually overload the runtime system and make it inefficient. We propose two solutions to improve the efficiency and reduce the number of tasks. First, we apply well-known blocking strategies in the context of task-based algorithms. Secondly, we explore batching techniques to reduce the number of tasks submitted to the runtime system. Moreover, in distributed architectures, partial pivoting generates many reductions on the critical path throughout the factorization which needs to be carefully handled to reach high performance. Two task-based reduction algorithms are proposed to express these operations and improve the runtime reactivity on the critical path. These proposals have been implemented in the dense linear algebra library `Chameleon` on top of the `StarPU` runtime system. Experiments conducted on our cluster with these optimizations show that our LU with partial pivoting asymptotically reaches the performance of the non-pivoting algorithm.

This paper has been accepted in IPDPS 2025 conference.

## 8.5 H-Rockmate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks (Topic 3.4.5)

**Participants:** Xunyi Zhao, Lionel Eyraud-Dubois, Yulia Gusak, Olivier Beaumont.

Training modern neural networks poses a significant memory challenge, as storing intermediate results during the forward and backward passes demands substantial memory resources. To address this issue while maintaining model accuracy, re-materialization techniques have been introduced to recompute selected intermediate results rather than storing them, thereby adhering to peak memory constraints. The main algorithmic problem is to compute a re-materialization schedule that minimizes the computational overhead within a given memory budget. In [21], we proposed an H-Rockmate framework that builds upon existing Rockmate solution and overcomes its limitation to work with sequential block structures by proposing a hierarchical approach. The framework performs an automatic decomposition

of the data-flow graph into a hierarchy of small-scale subgraphs, and finds a re-materialization schedule for the whole graph by recursively solving optimization problems for each subgraph. H-Rockmate allows users to transform their PyTorch models into nn.Modules that execute forward and backward passes efficiently within the specified memory budget. This framework can handle neural networks with diverse data-flow graph structures, including U-Nets and encoder-decoder Transformers. H-Rockmate consistently outperforms existing re-materialization approaches both in terms of average training iteration time and peak memory trade-offs, demonstrating superior memory efficiency in training modern neural networks.

## 8.6 Approximation Algorithms for Scheduling with/without Deadline Constraints where Rejection Costs are Proportional to Processing Times (Topic 3.4.3)

**Participants:** Olivier Beaumont, Lionel Eyraud-Dubois, Esragul Korkmaz, Laércio Lima Pilla.

We address two offline job scheduling problems, where jobs can either be processed on a limited supply of energy-efficient machines on the edge, or offloaded to an unlimited supply of energy-inefficient machines on the cloud (called rejected in our context). The goal is to minimize the total energy consumed in processing all tasks. We consider a first scheduling problem with no due date (or deadline) constraints, and we formulate it as a scheduling problem with rejection, where the cost of rejecting a job is directly proportional to its processing time. In [7] (code in [27]), we introduce a novel $5/4(1+\epsilon)$ approximation algorithm BEKP by associating it with a Multiple Subset Sum problem for this version. Our algorithm is an improvement over the existing literature, which provides a $(3/2-1/2m)$ approximation for scenarios with arbitrary rejection costs. In [19], we also cover a second scheduling problem, where jobs have due date (or deadline) constraints, and the goal is to minimize the weighted number of late jobs. In this context, if a job is late, it is offloaded (rejected) to an energy-inefficient machine on the cloud, which incurs a cost directly proportional to its processing time of the job. We position this problem in the literature, and introduce a novel $(1-(m-1)^m/m^m)$-approximation algorithm MDP for this version, where we got our inspiration from an algorithm for the interval selection problem with a $(1-m^m/(m+1)^m)$ approximation ratio for arbitrary rejection costs. We evaluate and discuss the effectiveness of our approaches through a series of experiments, comparing them to existing algorithms.

## 8.7 Tightening I/O Lower Bounds through the Hourglass Dependency Pattern (Topic 3.4.1)

**Participants:** Lionel Eyraud-Dubois.

When designing an algorithm, one cares about arithmetic/computational complexity, but data movement (I/O) complexity plays an increasingly important role that highly impacts performance and energy consumption. For a given algorithm and a given I/O model, scheduling strategies such as loop tiling can reduce the required I/O down to a limit, called the I/O complexity, inherent to the algorithm itself. The objective of I/O complexity analysis is to compute, for a given program, its minimal I/O requirement among all valid schedules. In [10], we consider a sequential execution model with two memories, an infinite one, and a small one of size S on which the computations retrieve and produce data. The I/O is the number of reads and writes between the two memories. We identify a common "hourglass pattern" in the dependency graphs of several common linear algebra kernels. Using the properties of this pattern, we mathematically prove tighter lower bounds on their I/O complexity, which improves the previous state-of-the-art bound by a parametric ratio. This proof was integrated inside the IOLB automatic lower bound derivation tool.

### 8.8 StarONNX: a Dynamic Scheduler for Low Latency and High Throughput Inference on Heterogeneous Resources (Topic 3.4.4)

**Participants:** Olivier Beaumont, Lionel Eyraud-Dubois, Jean Francois David.

Efficient inference of Deep Neural Network (DNN) models on heterogeneous processors is challenging, not only because of the heterogeneity between CPUs and hardware accelerators, but also because the problem is fundamentally bi-objective in many contexts, since both latency (time to perform an inference) and throughput (number of inferences per unit time) need to be optimized. In [9, 16], we present StarONNX, a solution based on integrating ONNX Runtime in StarPU, which aims to optimize the distribution of inference tasks and resource management on heterogeneous architectures. This strategy relies on (i) the efficient execution of deep learning models by ONNX Runtime to maximize individual resource efficiency, and (ii) the orchestration of heterogeneous resources by StarPU to provide sophisticated scheduling and overlapping strategies for computation and communication. An essential point of the framework is the ability to split a DNN into two parts, one running on the GPU and the other on the CPU, thus increasing throughput by using all possible resources with a minimal degradation of worst case latency. We show that integrating the ONNX Runtime into StarPU does not introduce significant overhead. We also evaluated our approach against the Triton Inference Server and showed a significant improvement in resource utilization and reduced latency.

### 8.9 Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory (Topic 3.4.5)

**Participants:** Olivier Beaumont, Lionel Eyraud-Dubois.

Training in Feed Forward Deep Neural Networks is a memory-intensive operation which is usually performed on GPUs with limited memory capacities. This may force data scientists to limit the depth of the models or the resolution of the input data if data does not fit in the GPU memory. The re-materialization technique, whose idea comes from the checkpointing strategies developed in the Automatic Differentiation literature, allows data scientists to limit the memory requirements related to the storage of intermediate data (activations), at the cost of an increase in the computational cost. In [6], we introduce a new strategy of re-materialization of activations that significantly reduces memory usage. It consists in selecting which activations are saved and which activations are deleted during the forward phase, and then recomputing the deleted activations when they are needed during the backward phase. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs. This paper focuses on the fully heterogeneous case, where the computation time and the memory requirement of each layer is different. We prove that finding the optimal solution is NP-hard and that classical techniques from Automatic Differentiation literature do not apply. Moreover, the classical assumption of memory persistence of materialized activations, used to simplify the search of optimal solutions, does not hold anymore. Thus, we propose a weak memory persistence property and provide a Dynamic Program to compute the optimal sequence of computations. This algorithm is made available through the Rotor software, a PyTorch plug-in dealing with any network consisting of a sequence of layers, each of them having an arbitrarily complex structure. Through extensive experiments, we show that our implementation consistently outperforms existing re-materialization approaches for a large class of networks, image sizes and batch sizes.

### 8.10 Towards Modern linear Algebra Libraries (Topic 3.4.1)

**Participants:** Mathieu Faverge, Abdou Guermouche, Pierre Esterie.

In the context of the NumPEx research program, some needs raised from the HPC community about having modern and generic linear algebra software libraries. The current librairies available in the Topal team software stack propose state of the art performances and a high degree of domain specific knowledges. Mainly written within the C language, they can lack some expressiveness and thus, making it difficult for application developers or different language communities to use them. To answer these limitations, we work on making the Chameleon library available in the C++ ecosystem. The current work targets the Chameleon dense linear algebra library for heterogeneous architectures as it provides a strong and essential building block for HPC application developers and also fits within the upcoming C++ standard support for linear algebra. Our challenges are to be interoperable with the current and further C++ standards while keeping our libraries open to new design opportunities.

## 8.11 Exploiting Processor Heterogeneity to Improve Throughput and Reduce Latency for Deep Neural Network Inference (Topic 3.4.4)

**Participants:** Olivier Beaumont, Lionel Eyraud-Dubois, Jean Francois David.

The growing popularity of Deep Neural Networks (DNNs) in a variety of domains, including computer vision, natural language processing, and predictive analytics, has led to an increase in the demand for computing resources. Graphics Processing Units (GPUs) are widely used for training and inference of DNNs. However, this exclusive use can quickly lead to saturation of GPU resources while CPU resources remain underutilized. In [8], we propose a performance evaluation of a solution that exploits processor heterogeneity by combining the computational power of GPUs and CPUs. A solution is proposed for distributing the computational load across the different processors to optimize their utilization and achieve better performance. A solution for partitioning a DNN model with different computational resources is proposed. This solution transfers part of the load from the GPUs to the CPUs when necessary to reduce latency and increase throughput. The partitioning of DNN models is performed using METIS to balance the computational load to be distributed among the different resources while minimizing communication. The experimental results show that latency and throughput are improved for a number of DNN models. Potential applications include real-time processing systems such as autonomous vehicles, drones, and video surveillance systems where minimizing latency and maximizing throughput are critical.

## 8.12 OffMate: full fine-tuning of LLMs on a single GPU by re-materialization and offloading (Topic 3.4.5)

**Participants:** Xunyi Zhao, Lionel Eyraud-Dubois, Yulia Gusak, Olivier Beaumont.

In [23], we present OFFMATE, an efficient memory-reducing framework to enable fine-tuning large language models on a single GPU. In the same way that PyTorch Dynamo takes a model and automatically changes it to reduce the execution time, OFFMATE takes a model and automatically modifies it to fit memory constraints (e.g. GPU VRAM), while keeping the same numerical results without approximation. OFFMATE uses integer linear programming to combine re-materialization (deleting some intermediate activations and recomputing them when needed), weight and activation offloading (moving data to CPU memory), and CPU optimization in a holistically optimized way, ensuring an efficient usage of available resources. With 10%-50% execution time overhead, OFFMATE has achieved up to 10× GPU memory reduction on billion-size models including Llama, Phi, Bloom and Mistral from HuggingFace. OFFMATE

is also designed to be compatible with reduced precision and parameter-efficient fine-tuning techniques, so that the memory benefits can be combined.

### 8.13 Multi-Criteria Mesh Partitioning for an Explicit Temporal Adaptive Task-Distributed Finite-Volume Solver (Topic 3.4.1)

**Participants:** Alice Lasserre, Abdou Guermouche.

The aerospace industry is one of the largest users of numerical simulation, which is an essential tool in the field of aerodynamic engineering, where many fluid dynamics simulations are involved. In order to obtain the most accurate solutions, some of these simulations use unstructured finite volume solvers that cope with irregular meshes by using explicit time-adaptive integration methods. Modern parallel implementations of these solvers rely on task-based runtime systems to perform fine-grained load balancing and to avoid unnecessary synchronizations. Although such implementations greatly improve performance compared to a classical fork-join MPI+OpenMP variants, it remains a challenge to keep all cores busy throughout the simulation loop. In [12], we first investigate the origins of this lack of parallelism. We emphasize that the irregular structure of the task graph plays a major role in the inefficiency of the computation distribution. Our main contribution is to improve the shape of the task graph by using a new mesh partitioning strategy. The originality of our approach is to take the temporal level of mesh cells into account during the mesh partitioning phase. We evaluate our approach by integrating our solution in an ArianeGroup production code used by Airbus. We show that our partitioning ethod leads to a more balanced task graph. The resulting task scheduling is up to two times faster for meshes ranging from 200,000 to 12,000,000 components.

### 8.14 Dynamic Tasks Scheduling with Multiple Priorities on Heterogeneous Computing Systems (Topic 3.4.1)

**Participants:** Hayfa Tayeb, Mathieu Faverge, Abdou Guermouche.

The efficient utilization of heterogeneous computing systems is crucial for scientists and industrial organizations to execute computationally intensive applications. ask-based programming has emerged as an effective approach for harnessing the processing power of these systems. However, effective scheduling of task-based applications is critical for achieving high performance. Typically, these applications are represented as directed acyclic graphs (DAGs), which can be optimized through careful scheduling to minimize execution time and maximize resource utilization. In [15], we introduce MultiPrio, a dynamic task scheduler that aims to minimize the overall completion time of parallelized task-based applications. The goal is to find a trade-off between resource affinity, task criticality, and workload balancing on the resources. To this end, we compute scores for each task and manage the available tasks in the system with a data structure based on a set of priority queues. Tasks are assigned to available resources according to these scores, which are dynamically computed by heuristics based on task affinity and criticality. We also consider workload balancing across resources and data locality awareness. To evaluate the scheduler, we study the performance of dense and sparse linear algebra task-based applications and task-based FMM application using the StarPU runtime system on heterogeneous nodes. Our scheduler shows interesting results compared to other state-of-the-art schedulers in StarPU for regular applications, and excels at optimizing irregular workloads, improving performance by up to 31

### 8.15 Optimizing Parallel System Efficiency: Dynamic Task Graph Adaptation with Recursive Tasks (Topic 3.4.1)

**Participants:** Thomas Morin, Abdou Guermouche.

Task-based programming models significantly improve the efficiency of parallel systems. The Sequential Task Flow (STF) model focuses on static task sizes within task graphs, but determining optimal granularity during graph submission is tedious. To overcome this, in [11], we extend StarPU's STF recursive tasks model, enabling dynamic transformation of tasks into subgraphs. Early evaluations on homogeneous shared memory reveal that this just-in-time adaptation enhances performance.

# 9 Bilateral contracts and grants with industry

## 9.1 Bilateral Grants with Industry

**Participants:** Olivier Beaumont, Lionel Eyraud-Dubois, Mathieu Faverge, Abdou Guermouche, Yulia Gusak, Pierre Ramet.

Some on the ongoing PhD thesis are developed within bilateral contract with industry for PhD advisory:

- Airbus (2022-). This collaboration concerns the parallelization and optimization of the Flusepa application, which models the separation of boosters for space launchers at Airbus Safran Launchers. Flusepa combines computational fluid mechanics, algorithms (AMR) and task-based parallelism based on the StarPU runtime system. We are involved in the supervision of the PhD. of Alice Lasserre in this context.

- CEA-Cesta for the PhD of Clément Richefort. The aim of this thesis is to open a research work on an alternative method to domain decomposition. The basic principle of multigrid method is to use a collection of coarser problems which permit to accelerate the convergence to the fine solution. These methods are iterative with an optimal linear scalability. However, they are not efficient for oscillatory kernel problems such as electromagnetism or acoustic, which lead to indefinite matrices. The aim of this thesis is to draw up a first analysis of this method applied to indefinite Helmholtz equation, then to find the appropriates operators and finally to adapt them to Maxwell equations.

- CEA-Cesta for the PhD of Abel Calluaud. A direct solver developed at CEA relies on the approximation by hierarchical matrices to reduce both computational and memory costs. Although these developments have met a growing demand for increased simulation accuracy, there are still open problems to pursue these research efforts in an HPC context. In this thesis, we propose to develop and compare several approaches to adapt the granularity of hierarchical tasks and extract parallelism to exploit the multicore computational nodes associated with massively parallel architectures such as GPUs.

- CEA-Cesta for the PhD of Dimitri Walther. In the context of numerical simulation of electromagnetism, integral methods are among the most widely used because of their power. These methods lead to the solution of dense linear problems and are therefore very expensive. For this reason, hierarchical compression methods have been developed that drastically reduce the cost associated with these matrices. They are based on a hierarchical partitioning of the matrix, and therefore of the mesh, and the efficiency of the compression depends on this partitioning. In this context, the aim of the thesis is to develop efficient and scalable hierarchical partitioners to optimise the compression of the matrix.

- For over two years, we have been collaborating with Eviden on the development of an HPL benchmark on top of runtime systems. This work is continued as part of Alycia Lisito's thesis funded by a CIFRE contract. To guarantee a high level of flexibility and portability, it is possible to use a task-based implementation through an executive support (or runtime). This programming model has

already proved its effectiveness in the implementation of various parallel algorithms, in particular for dense linear algebra (LU decomposition, Cholesky decomposition, QR, etc.). In this thesis, we will use Inria's existing software stack, through the dense linear algebra library Chameleon and the executive support StarPU. These reference libraries for runtime linear algebra will be studied to enable the scaling up of more complex algorithms such as HPL.

# 10 Partnerships and cooperations

## 10.1 International initiatives

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

ELF Associate Team on on Efficient deep Learning Frameworks.
Partners

- TOPAL
- California Institute of Technology (Caltech)

Nowadays, Deep Learning (DL) and Artificial Intelligence (AI) technologies are incorporated in more and more areas to solve various problems of video, audio, natural language processing, content generation, etc. Frameworks based on neural networks, which are core modules of deep learning models, have been already successfully used for action recognition, weather forecasting, robotic surgery and other inspiring applications [24, 44, 48]. The drawbacks of modern neural networks are that they usually require a significant amount of data and a lot of GPU devices to be trained, which makes them expensive in terms of energy and money costs, and harmful in terms of air emissions [27]. The general question we are going to address during the work of the associate team is: given your application and your computation platform, how to perform the model training efficiently in terms of time/energy?

## 10.2 International research visitors

### 10.2.1 Visits of international scientists

**Other international visits to the team**

- Jean Kossaifi and Nick Kovachki (Caltech and NVidia) visited the team for one week in November 2024 to work on the parallelization of training and the minimization of memory consumption as part of the associated ELF team and for networks based on Neural Operators.
- From May 16 to May 24, 2024, Thomas Herault, research assistant professor from the University of Tennessee, visited the Centre Inria de l'Université Bordeaux Sud-Ouest, following up on his 2023 visit, where he initiated discussions on applying fault-tolerance techniques from HPC. During this visit, he further interacted with researchers from Topal, laying the foundation for future collaborations and advancing his research project as he applied to join the TOPAL project-team. Technical discussions included AI workloads and parallel algorithms, checkpointing and rollback-recovery in the StarPU runtime system, and tolerating memory silent data corruption in SVD problems. He also used this time to refine his research project presentation for the Inria audition.
- As part of Clément Richefort's thesis defense on multigrid method for the indefinite Helmholtz equation, we had the visit of Rob Falgout (Lawrence Livermore National Laboratory) and Edmond Chow (Georgia Institute of Technology) for a week, from November 25 to 29, 2024. They participated in the day around H-matrices organized with colleagues from ENSTA (in particular the POEMS team of Stéphanie Chaillat with Xavier Claeys, Luiz Faria and Pierre Marchand). This day followed a first meeting between researchers from the TOPAL and POEMS teams, organized on October 2, 2024, in order to identify possible collaborations around our software developments for H-matrices.

## 10.3 European initiatives

### 10.3.1 H2020 projects

**EUPEX** EUPEX project on cordis.europa.eu

**Title:** EUROPEAN PILOT FOR EXASCALE

**Duration:** From January 1, 2022 to December 31, 2025

**Partners:**

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- GRAND EQUIPEMENT NATIONAL DE CALCUL INTENSIF (GENCI), France
- VSB - TECHNICAL UNIVERSITY OF OSTRAVA (VSB - TU Ostrava), Czechia
- FORSCHUNGSZENTRUM JULICH GMBH (FZJ), Germany
- COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES (CEA), France
- IDRYMA TECHNOLOGIAS KAI EREVNAS (FOUNDATION FOR RESEARCH AND TECHNOLOGYHELLAS), Greece
- SVEUCILISTE U ZAGREBU FAKULTET ELEKTROTEHNIKE I RACUNARSTVA (UNIVERSITYOF ZAGREB FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING), Croatia
- UNIVERSITA DEGLI STUDI DI TORINO (UNITO), Italy
- CYBELETECH (Cybeletech), France
- UNIVERSITA DI PISA (UNIPI), Italy
- GRAN SASSO SCIENCE INSTITUTE (GSSI), Italy
- ISTITUTO NAZIONALE DI ASTROFISICA (INAF), Italy
- UNIVERSITA DEGLI STUDI DEL MOLISE, Italy
- E 4 COMPUTER ENGINEERING SPA (E4), Italy
- UNIVERSITA DEGLI STUDI DELL'AQUILA (UNIVAQ), Italy
- CONSIGLIO NAZIONALE DELLE RICERCHE (CNR), Italy
- JOHANN WOLFGANG GOETHE-UNIVERSITAET FRANKFURT AM MAIN (GUF), Germany
- EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS (ECMWF), United Kingdom
- BULL SAS (BULL), France
- POLITECNICO DI MILANO (POLIMI), Italy
- EXASCALE PERFORMANCE SYSTEMS - EXAPSYS IKE, Greece
- ALMA MATER STUDIORUM - UNIVERSITA DI BOLOGNA (UNIBO), Italy
- PARTEC AG (PARTEC), Germany
- ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA, Italy
- CINECA CONSORZIO INTERUNIVERSITARIO (CINECA), Italy
- SECO SPA (SECO SRL), Italy
- CONSORZIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA (CINI), Italy

**Inria contact:** Olivier Beaumont

**Coordinator:** Jean-Robert Bacou (Eviden)

**Summary:** The EUPEX consortium aims to design, build, and validate the first EU platform for HPC, covering end-to-end the spectrum of required technologies with European assets: from the architecture, processor, system software, development tools to the applications. The EUPEX prototype will be designed to be open, scalable and flexible, including the modular OpenSequana-compliant platform and the corresponding HPC software ecosystem for the Modular Supercomputing Architecture. Scientifically, EUPEX is a vehicle to prepare HPC, AI, and Big Data processing communities for upcoming European Exascale systems and technologies. The hardware platform is sized to be large enough for relevant application preparation and scalability forecast, and a proof of concept for a modular architecture relying on European technologies in general and on European Processor Technology (EPI) in particular. In this context, a strong emphasis is put on the system software stack and the applications.

Being the first of its kind, EUPEX sets the ambitious challenge of gathering, distilling and integrating European technologies that the scientific and industrial partners use to build a production-grade prototype. EUPEX will lay the foundations for Europe's future digital sovereignty. It has the potential for the creation of a sustainable European scientific and industrial HPC ecosystem and should stimulate science and technology more than any national strategy (for numerical simulation, machine learning and AI, Big Data processing).

The EUPEX consortium – constituted of key actors on the European HPC scene – has the capacity and the will to provide a fundamental contribution to the consolidation of European supercomputing ecosystem. EUPEX aims to directly support an emerging and vibrant European entrepreneurial ecosystem in AI and Big Data processing that will leverage HPC as a main enabling technology.

**TEXTAROSSA** TEXTAROSSA project on cordis.europa.eu

**Title:** Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale

**Duration:** From April 1, 2021 to March 31, 2024

**Partners:**

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- IN QUATTRO SRL (in quattro), Italy
- FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV (FHG), Germany
- UNIVERSITA DEGLI STUDI DI TORINO (UNITO), Italy
- INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK, Poland
- AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE (ENEA), Italy
- UNIVERSITA DI PISA (UNIPI), Italy
- E 4 COMPUTER ENGINEERING SPA (E4), Italy
- CONSIGLIO NAZIONALE DELLE RICERCHE (CNR), Italy
- UNIVERSITE DE BORDEAUX (UBx), France
- BULL SAS (BULL), France
- POLITECNICO DI MILANO (POLIMI), Italy
- BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS), Spain
- CONSORZIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA (CINI), Italy
- ISTITUTO NAZIONALE DI FISICA NUCLEARE (INFN), Italy

**Inria contact:** Olivier BEAUMONT

**Coordinator:**

**Summary:** To achieve high performance and high energy efficiency on near-future exascale computing systems, a technology gap needs to be bridged: increase efficiency of computation with extreme efficiency in HW and new arithmetics, as well as providing methods and tools for seamless integration of reconfigurable accelerators in heterogeneous HPC multi-node platforms. TEXTAROSSA aims at tackling this gap through applying a co-design approach to heterogeneous HPC solutions, supported by the integration and extension of IPs, programming models and tools derived from European research projects, led by TEXTAROSSA partners. The main directions for innovation are towards: i) enabling mixed-precision computing, through the definition of IPs, libraries, and compilers supporting novel data types (including Posits), used also to boost the performance of AI accelerators; ii) implementing new multilevel thermal management and two-phase liquid cooling; iii) developing improved data movement and storage tools through compression; iv) ensure secure HPC operation through HW accelerated cryptography; v) providing RISC-V based IP for fast task scheduling and IPs for low-latency intra/inter-node communication. These technologies will be tested on the Integrated Development Vehicles mirroring and extending the European Processor Initiative ARM64-based architecture, and on an OpenSequana testbed. To drive the technology development and assess the impact of the proposed innovations TEXTAROSSA will use a selected but representative number of HPC, HPDA and AI demonstrators covering challenging HPC domains such as general-purpose numerical kernels, High Energy Physics (HEP), Oil & Gas, climate modelling, and emerging domains such as High Performance Data Analytics (HPDA) and High Performance Artificial Intelligence (HPC-AI).

## 10.4 National initiatives

### 10.4.1 ANR

SOLHARIS**: SOLvers for Heterogeneous Architectures over Runtime systems, Investigating Scalability**

**Duration:** 2018 – 2023

**Coordinator:** Alfredo Buttari (IRIT)

**Local contact:** Abdou Guermouche

**Partners:**

- IRIT Institut de Recherche en Informatique de Toulouse
- Inria Bordeaux - Sud-Ouest and Lyon
- Airbus Central R&T
- CEA Commissariat à l'énergie atomique et aux énergies alternatives

**Summary:** The SOLHARIS project aims at addressing the issues related to the development of fast and scalable linear solvers for large-scale, heterogeneous supercomputers. Because of the complexity and heterogeneity of the targeted algorithms and platforms, this project intends to rely on modern runtime systems to achieve high performance, programmability and portability. By gathering experts in computational linear algebra, scheduling algorithms and runtimes, SOLHARIS intends to tackle these issues through a considerable research effort for the development of numerical algorithms and scheduling methods that are better suited to the characteristics of large scale, heterogeneous systems and for the improvement and extension of runtime systems with novel features that more accurately fulfill the requirements of these methods. This is expected to lead to fundamental research results and software of great interest for researchers of the scientific computing community.

#### 10.4.2 Inria Challenge

**Challenge PULSE: Pushing low-carbon services towards the Edge**

**Duration:** 2022 – 2026

**Coordinator:** Romain Rouvoy

**Local contact:** Olivier Beaumont & Lionel Eyraud Dubois

**Partners:** Qarnot Computing, ADEME

**Inria teams:**

- Avalon
- Ctrl-A
- Spirals
- Stack
- Storm
- Topal

**Summary:** The Pulse challenge aims to develop and promote best practices in geo-repaired hardware and software infrastructures for more environmentally friendly intensive computing. The idea is to analyze which solutions are the most relevant, and which levers need to be focused on, to reduce the impact of infrastructures while maximizing the usefulness of their emissions. To this end, the challenge is structured around two complementary research axes to address this technological and environmental issue: holistic analysis of the environmental impact of intensive computing, and implementing more virtuous edge services.

# 11 Dissemination

## 11.1 Promoting scientific activities

### 11.1.1 Scientific events: organisation

The Workshop on Advancing Neural Network Training WANT Workshop at ICML'2024 focused on improving efficiency, scalability, and resource optimization in neural network training. Organized by experts from INRIA, NVIDIA, Oak Ridge National Laboratory, and others, it features 42 accepted papers (including 7 oral presentations) reviewed by 77 reviewers and 7 area chairs. The workshop highlights topics like efficient data management, scaling across devices, hardware and cluster optimization, and advanced techniques such as quantization, pruning, and low-rank adaptation, with applications in computer vision, NLP, reinforcement learning, and scientific domains. Paper contributors represented a wide range of expertise, with 130 affiliations from academia and 57 from industry, showcasing a strong collaborative effort. The hybrid-format event included offline and online activities like panel discussions, poster presentations, and networking sessions via Gather Town and Discord. Key speakers included experts from Hugging Face, NVIDIA, Carnegie Mellon University, and TogetherAI, who shared insights on large-scale training, hardware design, and optimization techniques. With a focus on fostering impactful research, the workshop is an essential platform for advancing computationally efficient AI solutions.

**General chair, scientific chair**
Julia Gusak acted as a general and program chair of the WANT Workshop at ICML'2024

### 11.1.2 Scientific events: selection

**Member of the conference program committees**

- Olivier Beaumont was involved in the following program committes: SC24 (Algorithms) HPDC24 IPDPS24 (Experiments) ICML24 (Workshop Selection Committee) Euro-PAR24 (Scheduling) ISC24 (tutorials)
- Lionel Eyraud Dubois was involved the the program committee of EuroPar 2024.
- Philippe Swartvagher was involved in the following program committees: Cluster 2024, Bench 2024, PMBS 2024 workshop, ComPAS 2024 and reproducibility committee of SC 24.
- Abdou Guermouche was involved in the following program committees : SC24 (System Software and Cloud Computing), IPDPS24 (System Software).
- Mathieu Faverge was involved in the program committee of : SSBAC-PAD 2024 (Parallel Applications and Algorithms).
- Julia Gusak was a program chair of the WANT Workshop at ICML'2024.

**Reviewer**
The members of the TOPAL project have also performed reviewing for the following list of conferences: HPDC 2024, IPDPS'25, SC 24, NeurIPS'24, ICML'24

### 11.1.3 Journal

**Member of the editorial boards**

- Olivier Beaumont is Associate Editor in Chief for the Journal of Parallel and Distributed Computing Elsevier JPDC
- Olivier Beaumont is Guest Editor for a Special Issue of IEEE Internet Computing with Shadi Ibrahim et al. on Serverless Computing.

**Reviewer - reviewing activities**
The members of the TOPAL project have performed reviewing for Journal of Parallel and Distributed Computing (Lionel Eyraud Dubois, Philippe Swartvagher, Abdou Guermouche), ACM Transactions on Mathematical Software (Pierre Ramet, Abdou Guermouche)), IEEE Transactions on Parallel and Distributed Systems (Lionel Eyraud Dubois, Abdou Guermouche, Mathieu Faverge), SoftwareX (Abdou Guermouche).

### 11.1.4 Invited talks

- Lionel Eyraud Dubois gave a lecture at the New Trends in Computing Summer School, entitled "Scheduling on heterogeneous machines".
- Philippe Swartvagher gave an invited talk at the ComPAS 2024 conference and at the FOSDEM HPC, Bigdata and Data Science devroom, entitled "Making reproducible and publishable experiments".
- Philippe Swartvagher gave a talk at the Journées non-thématiques du GDR RSD, entitled "On the Interactions between HPC Task-based Runtime Systems and Communication Libraries".
- Julia Gusak gave a talk at GAP'24 (Grenoble) entitled "Neural ODEs, neural operators, and their efficiency".
- Julia Gusak gave a talk at LoRAINNE'24 (Nancy) entitled "Tensor Methods in Deep Learning and their Efficiency".
- Julia Gusak gave a talk at AI4Industry'24 entitled "Efficient Training of Neural Networks".

#### 11.1.5 Leadership within the scientific community

- Olivier Beaumont is a member of the IEEE CS Babbage Award selection commitee
- Olivier Beaumont is the Inria representative on the EuroHPC Mirror Group, which helps define the position and strategy of the French Ministry of Research at the EuroHPC Governing Board.

#### 11.1.6 Scientific expertise

- Olivier Beaumont acts as Cross Reader for the European Innovation Council and the Pathfinder Open program.
- Olivier Beaumont acted as external evaluator for several EuroHPC calls: Inno4Scale, Energy, FFPlus
- Pierre Ramet is Scientific Advisor at the CEA-DAM CESTA.
- Pierre Ramet participated in the HCERES evaluation committee of the IRFU (Institut de recherche sur les lois fondamentales de l'Univers) at CEA Saclay.

#### 11.1.7 Research administration

- Pierre Ramet is the head of the CNRS Satanas department.
- Pierre Ramet is member of Scientific comittee of the LaBRI.
- Philippe Swartvagher is the communication referent for the NumPEx/Exa-SofT project.
- Philippe Swartvagher is the point of contact in Bordeaux for Grid5000/SLICES-FR infrastructure.
- Philippe Swartvagher is the representative of the TOPAL project at the Bordeaux CUMI.
- Philippe Swartvagher is elected member of the school council of the ENSEIRB-MATMECA engineering school.
- Abdou Guermouche is the scientific lead of the numerical library work package of the ExaSoft project (PEPR NumPEx).
- Abdou Guermouche is member of the Scientific Committee of LaBRI.
- Julia Gusak is a PI of the ELF associate team between Topal and Caltech.

### 11.2 Teaching - Supervision - Juries

- Undergraduate level/Licence:
  - Aurélien Esnard: Network (54h), Software technologies (80h) at Bordeaux University.
  - Pierre Ramet: System programming 24h, Databases 32h, Object programming 48h, Distributed programming 16h, Cryptography 16h, Introduction to AI Deep Learning and Data Analytics 16h at Bordeaux University.
  - Philippe Swartvagher: C Programming (46h), Web Programming (36h), Tools for Programming and C project (28h) at Bordeaux INP (ENSEIRB-MATMECA).
  - Abdou Guermouche System programming 36h at Bordeaux University.
  - MathieuFaverge : Programming environment (26h), Numerical algorithmic (25h), C projects (25h) at Bordeaux INP (ENSEIRB-MATMECA).
- Post graduate level/Master:
  - Aurélien Esnard: Network management (24h), Network security (24h) at Bordeaux University.
  - Lionel Eyraud Dubois: Graphs and Algorithms (20h), Complexity and Approximation (20h) at Bordeaux University.

- Olivier Beaumont: Parallel Algorithms, 20h at Bordeaux INP.
- Pierre Ramet: Cryptography 20h and Numerical algorithms 40h at Bordeaux INP (ENSEIRB-MATMECA).
- Philippe Swartvagher: Parallel Algorithms (15h), Project of network and system programming (25h) at Bordeaux INP (ENSEIRB-MATMECA).
- Abdou Guermouche Network management 92h, Network security 64h, Operating system 24h at Bordeaux University.
- Mathieu Faverge : System programming: lecture, practice and project (54h), Linear Algebra for high Performance Computing (9h) at Bordeaux INP (ENSEIRB-MATMECA). He is also in charge of the master 2 internship for the Computer Science department at Bordeaux INP (Enseirb-MatMeca) and he is in charge, with Raymond Namyst, of the High Performance Computing - High Performance Data Analytics specialty at Enseirb-MatMeca. This is a common training curriculum between the Computer Science and the MatMeca departments at Bordeaux INP and with the Bordeaux University in the context of the Computer Science Research Master.
- Julia Gusak: Efficient Deep Learning (Outils pour l'apprentissage) (19h) at Bordeaux INP (ENSEIRB-MATMECA).

### 11.2.1 Supervision

- Defended PhD: Clément Richefort; Development of an algebraic multigrid solver for the indefinite Helmholtz equation ; defended November 2024; advisor Pierre Ramet.
- Defended PhD: Xunyi Zhao; Optimizing Memory Usage when Training Deep Neural Networks; defended December 2024; advisors Olivier Beaumont, Yulia Gusak, Lionel Eyraud Dubois.
- PhD in progress: Abel Calluaud; Combined compiler and runtime approach for a direct hierarchical solver; started Nov. 2022; advisors Pierre Ramet, Mathieu Faverge.
- PhD in progress: Jean-François David; Dynamic Scheduling for Inference in Deep Neural Networks; advisors Olivier Beaumont, Lionel Eyraud Dubois.
- PhD in progress: Alycia Lisito; Design and implementation of a portable linear algebra benchmark on runtime systems for performance evaluation of heterogeneous Exascale architectures ; started Nov. 2023; advisors Pierre Ramet, Mathieu Faverge, Matthieu Kuhn (Eviden).
- PhD in progress: Dimitri Walther; ; started Nov. 2024; advisors Pierre Ramet, Mathieu Faverge, M. Lecouvez (CEA Cesta).
- Mathieu Faverge and Philippe Swartvagher supervised the internship of Camille Ordronneau [26].
- PhD in progress: Hayfa Tayeb; Optimization of high-performance applications on heterogeneous computing nodes; started Nov. 2021; A. Guermouche , B. Bramas , M. Faverge. Defense March 25th, 2025.
- PhD in progress: Albert D'Aviau de Piolant; started October 2023; Energy aware scheduling for exascale architectures. Advisors: Abdou Guermouche and Amina Guermouche.
- PhD in progress: Thomas Morin; started October 2023; Scheduling recursive task graphs. Advisors: Abdou Guermouche, Samuel Thibault, Pierre-André Wacrenier.
- PhD in progress : Alice Lasserre; Started Oct. 2022; Optimization of a task-based simulation code on a distributed supercomputer; Advisors: Jean-Marie Couteyen-Carpaye, Raymond Namyst and Abdou Guermouche.
- PhD in progress: Adrien Aguilla–Multner, Started October 2024; Efficient Training of Neural Networks. Advisors: Julia Gusak, Olivier Beaumont.
- Internship in progress: Enrique Galves, started in September 2024; Task-based systems for efficient deep learning. Advisors: Julia Gusak, Olivier Beaumont.

## 11.3 Popularization

### 11.3.1 Productions (articles, videos, podcasts, serious games, ...)

Olivier Beaumont participated with Remi Bouzel (Qarnot Computing) to the podcast "Déassemblons le numérique" (in French) to present the joint Inria/Qarnot Challenge "Pulse" (podcast)

### 11.3.2 Participation in Live events

- As part of the "Circuit Scientifique Bordelais", Philippe Swartvagher presented to high school pupils from the Lycée Bastide at Villeneuve-sur-Lot what is research in computer science and how to become a researcher.
- Olivier Beaumont participated in several internal events (closed doors, unité ou café) to present the activities of the Inria Bordeaux center teams at the interface between HPC and AI.
- As part of Maths en Jeans, Olivier Beaumont worked with groups of students from Andernos high school on combinatorial problems linked to data distribution for Linear Algebra problems.
- On several occasions, we have welcomed 3rd and 2nd grade students into the team, with the participation of Topal's PhD students, for periods of 2 hours to half a day.

# 12 Scientific production

## 12.1 Major publications

[1] O. Beaumont, P. Duchon, L. Eyraud-Dubois, J. Langou and M. Vérité. 'Symmetric Block-Cyclic Distribution: Fewer Communications Leads to Faster Dense Cholesky Factorization'. In: SC 2022 - Supercomputing. Dallas, Texas, United States, 13th Nov. 2022. URL: https://inria.hal.science/hal-03768910.

[2] O. Beaumont, L. Eyraud-Dubois, M. Vérité and J. Langou. 'I/O-Optimal Algorithms for Symmetric Linear Algebra Kernels'. In: ACM Symposium on Parallelism in Algorithms and Architectures. Philadelphie, United States, 11th July 2022. URL: https://inria.hal.science/hal-03580531.

[3] M. Faverge, N. Furmento, A. Guermouche, G. Lucas, R. Namyst, S. Thibault and P.-a. Wacrenier. 'Programming Heterogeneous Architectures Using Hierarchical Tasks'. In: *Concurrency and Computation: Practice and Experience* 35.25 (2023). DOI: 10.1002/cpe.7811. URL: https://hal.science/hal-04088833.

[4] C. Richefort, M. Lecouvez, R. Falgout and P. Ramet. 'Toward a multilevel method for the Helmholtz equation'. In: 21st SIAM Copper Mountain Conference on Multigrid Method. Copper Mountain, CO, United States, 16th Apr. 2023. URL: https://hal.science/hal-04046622.

[5] X. Zhao, T. Le Hellard, L. Eyraud-Dubois, J. Gusak and O. Beaumont. 'Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch'. In: ICML 2023. Honolulu (HI), United States, 23rd July 2023. URL: https://hal.science/hal-04095305.

## 12.2 Publications of the year

**International journals**

[6] O. Beaumont, L. Eyraud-Dubois, J. Herrmann, A. Joly and A. Shilova. 'Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory'. In: *ACM Transactions on Mathematical Software* (2024). URL: https://inria.hal.science/hal-02352969. In press (cit. on pp. 8, 9, 20).

**International peer-reviewed conferences**

[7] O. Beaumont, R. Bouzel, L. Eyraud-Dubois, E. Korkmaz, L. Lima Pilla and A. van Kempen. 'A 1.25(1+$\varepsilon$)-Approximation Algorithm for Scheduling with Rejection Costs Proportional to Processing Times'. In: International European Conference on Parallel and Distributed Computing (Euro-Par). Vol. 14801. Lecture Notes in Computer Science. Madrid, Spain: Springer Nature Switzerland, 26th Aug. 2024, pp. 225–238. DOI: 10.1007/978-3-031-69577-3_16. URL: https://hal.science/hal-04670834 (cit. on pp. 11, 19).

[8] O. Beaumont, J.-F. David, L. Eyraud-Dubois and S. Thibault. 'Exploiting Processor Heterogeneity to Improve Throughput and Reduce Latency for Deep Neural Network Inference'. In: SBAC-PAD 2024 - IEEE 36th International Symposium on Computer Architecture and High Performance Computing. Hilo, Hawaii, United States, 13th Nov. 2024. URL: https://hal.science/hal-04690154 (cit. on p. 21).

[9] O. Beaumont, J.-F. David, L. Eyraud-Dubois and S. Thibault. 'StarONNX: a Dynamic Scheduler for Low Latency and High Throughput Inference on Heterogeneous Resources'. In: HeteroPar 2024 - 22ND INTERNATIONAL WORKSHOP Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms. HeteroPar'24 Proceedings. Madrid, Spain, 2024. URL: https://inria.hal.science/hal-04646530 (cit. on p. 20).

[10] L. Eyraud-Dubois, G. Iooss, J. Langou and F. Rastello. 'Tightening I/O Lower Bounds through the Hourglass Dependency Pattern'. In: SPAA 2024 - 36th ACM Symposium on Parallelism in Algorithms and Architectures. Nantes, France, Apr. 2024, pp. 1–34. URL: https://inria.hal.science/hal-04555744 (cit. on p. 19).

[11] N. Furmento, A. Guermouche, G. Lucas, T. Morin, S. Thibault and P.-A. Wacrenier. 'Optimizing Parallel System Efficiency: Dynamic Task Graph Adaptation with Recursive Tasks'. In: WAMTA 2024 - Workshop on Asynchronous Many-Task Systems and Applications 2024. Knoxville, United States: https://wamta24.icl.utk.edu/, 14th Feb. 2024. URL: https://inria.hal.science/hal-04548787 (cit. on p. 23).

[12] A. Lasserre, J. M. Couteyen Carpaye, A. Guermouche and R. Namyst. 'Multi-Criteria Mesh Partitioning for an Explicit Temporal Adaptive Task-Distributed Finite-Volume Solver'. In: The 25th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC 2024). San Francisco, United States, 31st May 2024, p. 10. URL: https://inria.hal.science/hal-04403209 (cit. on p. 22).

[13] A. Lisito. 'Efficient HPL on top of runtime systems'. In: compas 2024 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Nantes, France, 2nd July 2024. URL: https://inria.hal.science/hal-04603576 (cit. on p. 18).

[14] A. Lisito, M. Faverge, G. Pichon and P. Ramet. 'Enhancing sparse direct solver scalability through runtime system automatic data partition'. In: WAMTA 2024 - Workshop on Asynchronous Many-Task Systems and Applications 2024. Vol. 14626. Lecture Notes in Computer Science. Knoxville, United States: Springer Nature Switzerland, 14th Mar. 2024, pp. 105–110. DOI: 10.1007/978-3-031-61763-8_10. URL: https://inria.hal.science/hal-04527103 (cit. on p. 17).

[15] H. Tayeb, B. Bramas, M. Faverge and A. Guermouche. 'Dynamic Tasks Scheduling with Multiple Priorities on Heterogeneous Computing Systems'. In: *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 38th IEEE International Parallel & Distributed Processing Symposium. San franscisco, CA, United States, 2024, pp. 31–40. DOI: 10.1109/IPDPSW63119.2024.00014. URL: https://hal.science/hal-04498634 (cit. on p. 22).

**Conferences without proceedings**

[16] O. Beaumont, J.-F. David, L. Eyraud-Dubois and S. Thibault. 'StarONNX : Un ordonanceur dynamique pour une inférence rapide et à haut débit sur des ressources hétérogènes'. In: Compas 2024 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Nantes, France, 2nd July 2024. URL: https://inria.hal.science/hal-04668550 (cit. on p. 20).

[17] R. Falgout, M. Lecouvez, P. Ramet and C. Richefort. 'Slides 18th CMCIM 2024, Toward an algebraic multigrid method for the indefinite Helmholtz equation'. In: 18th Copper Mountain Conference On Iterative Methods. Frisco (Co), United States, 14th Apr. 2024. URL: https://cea.hal.science/cea-04620993 (cit. on p. 17).

[18] B. Nicolas, M. Faverge, P. Ramet, A. Lisito and M. Kherraz. 'Comparative Study of Mixed-Precision and Low-Rank Compression Techniques in Sparse Direct Solvers'. In: 2024 SIAM Conference on Applied Linear Algebra. Paris, France, 13th May 2024. URL: https://hal.science/hal-04585047 (cit. on p. 18).

**Reports & preprints**

[19] O. Beaumont, R. Bouzel, L. Eyraud-Dubois, E. Korkmaz, L. Lima Pilla and A. van Kempen. *Approximation Algorithms for Scheduling with/without Deadline Constraints where Rejection Costs are Proportional to Processing Times*. 15th Oct. 2024. URL: https://hal.science/hal-04745701 (cit. on p. 19).

[20] R. D. Falgout, M. Lecouvez, P. Ramet and C. Richefort. *Toward an algebraic multigrid method for the indefinite Helmholtz equation*. 28th June 2024. URL: https://cea.hal.science/cea-04620991.

[21] J. Gusak, X. Zhao, T. Le Hellard, Z. Li, L. Eyraud-Dubois and O. Beaumont. *HiRemate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks*. July 2024. URL: https://hal.science/hal-04403844 (cit. on p. 18).

[22] A. d'Aviau de Piolant, H. Tayeb, B. Bramas, M. Faverge, A. Guermouche and A. Guermouche. *Improving energy efficiency of HPC applications using unbalanced GPU power capping*. 11th Oct. 2024. URL: https://inria.hal.science/hal-04883872.

[23] X. Zhao, L. Eyraud-Dubois, T. Le Hellard, J. Gusak and O. Beaumont. *OFFMATE: full fine-tuning of LLMs on a single GPU by re-materialization and offloading*. 24th July 2024. URL: https://hal.science/hal-04660745 (cit. on p. 21).

**Other scientific publications**

[24] A. Lasserre, J. M. Couteyen Carpaye, A. Guermouche and R. Namyst. 'Multi-Criteria Mesh Partitioning for an Explicit Temporal Adaptive Task-Distributed Finite-Volume Solver'. In: Doctoral students' day of the Mathematics and Computer Science doctoral school of the University of Bordeaux (EDMI). Bordeaux, France, 11th Apr. 2024. URL: https://inria.hal.science/hal-04895715.

[25] A. Lisito, M. Faverge, D. Goudin, M. Kuhn and P. Ramet. 'Scalable and Portable LU Factorization with Partial Pivoting on top of Runtime Systems'. In: journée calculs et données 2024. Bordeaux, France, 4th Nov. 2024. URL: https://inria.hal.science/hal-04867213 (cit. on p. 18).

[26] C. Ordronneau. 'Développement et maintenance du logiciel ViTE'. Talence: ENSEIRB-MATMECA, 21st Oct. 2024, p. 21. URL: https://inria.hal.science/hal-04725983 (cit. on p. 31).

**Software**

[27] [SW] O. Beaumont, L. Eyraud-Dubois, E. Korkmaz and L. Lima Pilla, *Experimental codes and results for the paper "A 5/4(1+eps)-Approximation Algorithm for Scheduling with Rejection Costs Proportional to Processing Times"* 22nd Mar. 2024. Inria & Labri, Univ. Bordeaux. LIC: CeCILL Free Software License Agreement v2.0. HAL: ⟨hal-04517532⟩, URL: https://inria.hal.science/hal-04517532, SWHID: ⟨swh:1:dir:53aa25178b70f7d119690440f64912c226521893;origin=https://hal.archives-ouvertes.fr/hal-04517532;visit=swh:1:snp:690ab9c97d792e39eb94c530093dc36623bb9dac;anchor=swh:1:rel:f7390f430900bcc0a290b2fc33d2ab5285922588;path=/⟩ (cit. on p. 19).

## 12.3 Cited publications


[28] E. Agullo, O. Aumage, M. Faverge, N. Furmento, F. Pruvost, M. Sergent and S. P. Thibault. 'Achieving High Performance on Supercomputers with a Sequential Task-based Programming Model'. In: *IEEE Transactions on Parallel and Distributed Systems* (2017), pp. 1–1. DOI: 10.1109/TPDS.2017.2766064 (cit. on p. 9).

[29] E. Agullo, A. Buttari, A. Guermouche and F. Lopez. 'Implementing Multifrontal Sparse Solvers for Multicore Architectures with Sequential Task Flow Runtime Systems'. In: *ACM Trans. Math. Softw.* 43.2 (Aug. 2016), 13:1–13:22. DOI: 10.1145/2898348. eprint: \url{https://hal.inria.fr/hal-01333645}. URL: http://doi.acm.org/10.1145/2898348 (cit. on p. 9).

[30] E. Agullo, A. Buttari, A. Guermouche and F. Lopez. 'Task-Based Multifrontal QR Solver for GPU-Accelerated Multicore Architectures.' In: *HiPC*. **Best paper award**. IEEE Computer Society, 2015, pp. 54–63. DOI: 10.1109/HiPC.2015.27. eprint: \url{https://hal.archives-ouvertes.fr/hal-01270145} (cit. on p. 9).

[31] P. Alonso, M. F. Dolz, F. D. Igual, R. Mayo and E. S. Quintana-Ortí. 'Reducing Energy Consumption of Dense Linear Algebra Operations on Hybrid CPU-GPU Platforms'. In: *2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications*. 2012, pp. 56–62. DOI: 10.1109/ISPA.2012.16 (cit. on p. 7).

[32] P. Alonso, M. F. Dolz, R. Mayo and E. S. Quintana-Ortí. 'Modeling power and energy consumption of dense matrix factorizations on multicore processors'. In: *Concurrency and Computation: Practice and Experience* 26.17 (2014), pp. 2743–2757. DOI: \url{https://doi.org/10.1002/cpe.3162}. eprint: \url{https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.3162}. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3162 (cit. on p. 7).

[33] H. Anzt, J. Dongarra and E. S. Quintana-Ortí. 'Adaptive Precision Solvers for Sparse Linear Systems'. In: *Proceedings of the 3rd International Workshop on Energy Efficient Supercomputing*. E2SC '15. Austin, Texas: Association for Computing Machinery, 2015. DOI: 10.1145/2834800.2834802. URL: https://doi.org/10.1145/2834800.2834802 (cit. on p. 7).

[34] O. Beaumont, P. Duchon, L. Eyraud-Dubois, J. Langou and M. Verite. 'Symmetric Block-Cyclic Distribution: Fewer Communications leads to Faster Dense Cholesky Factorization'. In: *SC'22: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. (best paper, Algorithm Track). IEEE and ACM. 2022 (cit. on pp. 5, 9).

[35] O. Beaumont, L. Eyraud-Dubois and A. Shilova. 'Efficient Combination of Rematerialization and Offloading for Training DNNs'. In: *NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems*. Virtual-only Conference, Dec. 2021. URL: https://hal.inria.fr/hal-03359793 (cit. on pp. 8, 9).

[36] O. Beaumont, L. Eyraud-Dubois and A. Shilova. 'MadPipe: Memory Aware Dynamic Programming Algorithm for Pipelined Model Parallelism'. In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2022 (cit. on p. 9).

[37] O. Beaumont, L. Eyraud-Dubois and A. Shilova. 'Optimal GPU-CPU Offloading Strategies for Deep Neural Network Training'. In: *Euro-Par 2020: Parallel Processing*. Ed. by M. Malawski and K. Rzadca. Cham: Springer International Publishing, 2020, pp. 151–166 (cit. on pp. 8, 9).

[38] O. Beaumont, L. Eyraud-Dubois and A. Shilova. 'Pipelined Model Parallelism: Complexity Results and Memory Considerations'. In: *Proceedings of Europar 2021*. Lisbon, Portugal: Springer, Aug. 2021. URL: https://hal.inria.fr/hal-02968802 (cit. on pp. 8, 9).

[39] O. Beaumont, L. Eyraud-Dubois and M. Verite. '2D Static Resource Allocation for Compressed Linear Algebra and Communication Constraints'. In: *2020 IEEE 27th International Conference on High Performance Computing, Data, and Analytics (HiPC)*. IEEE. 2020, pp. 181–191 (cit. on p. 9).

[40] O. Beaumont, L. Eyraud-Dubois, M. Vérité and J. Langou. 'I/O-Optimal Algorithms for Symmetric Linear Algebra Kernels'. In: *ACM Symposium on Parallelism in Algorithms and Architectures*. Association for Computing Machinery : SIGACT, SIGARCH. Philadelphie, United States, July 2022. URL: https://hal.inria.fr/hal-03580531 (cit. on pp. 5, 9).

[41] O. Beaumont, J. Herrmann, G. Pallez and A. Shilova. 'Optimal memory-aware backpropagation of deep join networks'. In: *Philosophical Transactions of the Royal Society A* 378.2166 (2020), p. 20190049 (cit. on p. 9).

[42] R. Carratalá-Sáez, M. Faverge, G. Pichon, E. S. Quintana-Ortí and G. Sylvand. 'Exploiting Generic Tiled Algorithms Toward Scalable H-Matrices Factorizations on Top of Runtime Systems'. In: *SIAM PP20-SIAM Conference on Parallel Processing for Scientific Computing*. 2020 (cit. on p. 9).

[43] R. Carratalá-Sáez, M. Faverge, G. Pichon, G. Sylvand and E. S. Quintana-Ortí. 'Tiled Algorithms for Efficient Task-Parallel ?-Matrix Solvers'. In: *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2020, pp. 757–766 (cit. on p. 9).

[44] T. Chen, B. Xu, C. Zhang and C. Guestrin. 'Training deep nets with sublinear memory cost'. In: *arXiv preprint arXiv:1604.06174* (2016) (cit. on p. 9).

[45] R. D. Falgout, S. Friedhoff, T. V. Kolev, S. P. MacLachlan and J. B. Schroder. 'Parallel time integration with multigrid'. In: *SIAM Journal on Scientific Computing* 36.6 (2014), pp. C635–C661 (cit. on p. 9).

[46] M. J. Gander and S. Vandewalle. 'Analysis of the parareal time-parallel time-integration method'. In: *SIAM Journal on Scientific Computing* 29.2 (2007), pp. 556–578 (cit. on p. 9).

[47] P. Ghysels, X. S. Li, F.-H. Rouet, S. Williams and A. Napov. 'An Efficient Multicore Implementation of a Novel HSS-Structured Multifrontal Solver Using Randomized Sampling'. In: *SIAM Journal on Scientific Computing* 38.5 (2016), S358–S384 (cit. on p. 7).

[48] A. N. Gomez, M. Ren, R. Urtasun and R. B. Grosse. 'The reversible residual network: Backpropagation without storing activations'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 2211–2221 (cit. on p. 8).

[49] A. Gruslys, R. Munos, I. Danihelka, M. Lanctot and A. Graves. 'Memory-efficient backpropagation through time'. In: *Advances in Neural Information Processing Systems*. 2016, pp. 4125–4133 (cit. on p. 9).

[50] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks and C.-J. Wu. 'Chasing Carbon: The Elusive Environmental Footprint of Computing'. In: *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2021, pp. 854–867 (cit. on p. 8).

[51] J. Gusak, D. Cherniuk, A. Shilova, A. Katrutsa, D. Bershatsky, X. Zhao, L. Eyraud-Dubois, O. Shlyazhko, D. Dimitrov, I. Oseledets and O. Beaumont. 'Survey on Large Scale Neural Network Training'. In: *The 31st International Joint Conference on Artificial Intelligence (IJCAI)*. 2022 (cit. on pp. 8, 9).

[52] A. Ida, T. Iwashita, T. Mifune and Y. Takahashi. 'Parallel Hierarchical Matrices with Adaptive Cross Approximation on Symmetric Multiprocessing Clusters'. In: *Journal of Information Processing* 22.4 (2014), pp. 642–650 (cit. on p. 7).

[53] E. Korkmaz, M. Faverge, G. Pichon and P. Ramet. *Deciding Non-Compressible Blocks in Sparse Direct Solvers using Incomplete Factorization*. Research Report RR-9396. Inria Bordeaux - Sud Ouest, 2021, p. 16. URL: https://hal.inria.fr/hal-03152932 (cit. on p. 9).

[54] N. Kukreja, A. Shilova, O. Beaumont, J. Huckelheim, N. Ferrier, P. Hovland and G. Gorman. 'Training on the Edge: The why and the how'. In: *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2019, pp. 899–903 (cit. on p. 9).

[55] X. Lacoste, M. Faverge, G. Bosilca, P. Ramet and S. Thibault. 'Taking Advantage of Hybrid Systems for Sparse Direct Solvers via Task-Based Runtimes'. In: *2014 IEEE International Parallel & Distributed Processing Symposium Workshops, Phoenix, AZ, USA, May 19-23, 2014*. IEEE Computer Society, 2014, pp. 29–38. DOI: 10.1109/IPDPSW.2014.9. URL: https://doi.org/10.1109/IPDPSW.2014.9 (cit. on p. 9).

[56] T. Mary. *Block Low-Rank multifrontal solvers: complexity, performance and scalability*. Université Toulouse 3 Paul Sabatier: Ph.D. Dissertation, 2017 (cit. on p. 7).

[57] S. Moustafa, F. Févotte, M. Faverge, L. Plagne and P. Ramet. 'Efficient Parallel Solution of the 3D Stationary Boltzmann Transport Equation for Diffusive Problems'. In: *Journal of Computational Physics* (Mar. 2019). DOI: 10.1016/j.jcp.2019.03.019. URL: https://hal.inria.fr/hal-02080624 (cit. on p. 9).

[58] D. Narayanan, A. Harlap, A. Phanishayee, V. Seshadri, N. R. Devanur, G. R. Ganger, P. B. Gibbons and M. Zaharia. 'PipeDream: generalized pipeline parallelism for DNN training'. In: *Proceedings of the 27th ACM Symposium on Operating Systems Principles*. 2019, pp. 1–15 (cit. on pp. 8, 9).

[59] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier and J. Dean. 'Carbon emissions and large neural network training'. In: *arXiv preprint arXiv:2104.10350* (2021) (cit. on p. 8).

[60] A.-H. Phan, K. Sobolev, K. Sozykin, D. Ermilov, J. Gusak, P. Tichavskỳ, V. Glukhov, I. Oseledets and A. Cichocki. 'Stable Low-rank Tensor Decomposition for Compression of Convolutional Neural Network'. In: ***European Conference on Computer Vision (ECCV)***. Springer. 2020, pp. 522–539 (cit. on p. 8).

[61] G. Pichon, E. Darve, M. Faverge, P. Ramet and J. Roman. 'Sparse supernodal solver using block low-rank compression: Design, performance and analysis'. In: *International Journal of Computational Science and Engineering* 27 (July 2018), pp. 255–270. DOI: 10.1016/J.JOCS.2018.06.007. URL: https://hal.inria.fr/hal-01824275 (cit. on p. 9).

[62] G. Pichon, M. Faverge and P. Ramet. 'Recent Developments Around the Block Low-Rank PaStiX Solver'. In: *SIAM Conference on Parallel Processing for Scientific Computing (SIAM PP 2020)*. 2020 (cit. on p. 9).

[63] D. Sukkari, H. Ltaief, D. Keyes and M. Faverge. 'Leveraging Task-Based Polar Decomposition Using PARSEC on Massively Parallel Systems'. In: *2019 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE. 2019, pp. 1–12 (cit. on p. 9).