RESEARCH CENTRE

Inria Centre
at the University of Bordeaux

IN PARTNERSHIP WITH:

Institut Polytechnique de Bordeaux,
Université de Bordeaux

2023
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

Contents

Project-Team TOPAL 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3

3 Research program 4
3.1 Objectives . . . 4
3.2 Overall Positionning . . . 4
3.3 Research Axes . . . 5
3.3.1 Use of Runtime systems . . . 5
3.3.2 Design of compression techniques . . . 5
3.3.3 Energy minimization . . . 6
3.4 Main Research Topics . . . 6
3.4.1 Task-based Linear Algebra and Tensor Computations . . . 6
3.4.2 Multi-Linear Algebra and Tensor Decompositions . . . 7
3.4.3 Energy Minimization in Linear Solvers . . . 7
3.4.4 Task-based Approaches for Deep Learning . . . 7
3.4.5 Tensor Compression for Inference . . . 8
3.4.6 Carbon Saving and Energy-Efficient Training . . . 8

4 Application domains 8
4.1 Multi-Linear Algebra and Solvers . . . 8
4.2 Training and Inference for DNNs . . . 9

5 Social and environmental responsibility 9
5.1 Footprint of research activities . . . 9
5.2 Impact of research results . . . 10
5.2.1 Carbon Impact of Cloud Platforms . . . 10
5.2.2 Democratization of Large Models Training . . . 10

6 Highlights of the year 11
6.1 Awards . . . 11
6.2 Organization of Events . . . 11

7 New software, platforms, open data 11
7.1 New software . . . 11
7.1.1 Chameleon . . . 11
7.1.2 StarPart . . . 12
7.1.3 PaStiX . . . 13
7.1.4 rotor . . . 13
7.1.5 StarPU . . . 14
7.1.6 VITE . . . 15
7.1.7 pmtool . . . 16
7.1.8 rockmate . . . 16

8 New results 16
8.1 Toward a multilevel method for the Helmholtz equation . . . 17
8.2 Programming heterogeneous architectures using hierarchical tasks . . . 17
8.3 Task-based Parallel Programming for Scalable Matrix Product Algorithms . . . 17
8.4 Combining reduction with synchronization barrier on multi-core processors . . . 18
8.5 Data Distribution Schemes for Dense Linear Algebra Factorizations on Any Number of Nodes 18
8.6 Energy or performance: the impact of implementation on consumption . . . 18
8.7 On the Arithmetic Intensity of Distributed-Memory Dense Matrix Multiplication Involving a Symmetric Input Matrix (SYMM) . . . 18

8.8 Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch 19
8.9 H-Rockmate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks 19

9 Bilateral contracts and grants with industry 20
9.1 Bilateral Grants with Industry . . . 20
9.1.1 PhD Theses . . . 20
9.1.2 Research Engineers . . . 20

10 Partnerships and cooperations 21
10.1 International initiatives . . . 21
10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . 21
10.2 International research visitors . . . 21
10.2.1 Visits of international scientists . . . 21
10.2.2 Visits to international teams . . . 22
10.3 European initiatives . . . 22
10.3.1 H2020 projects . . . 22
10.4 National initiatives . . . 25
10.4.1 ANR . . . 25
10.4.2 Inria Challenge . . . 26

11 Dissemination 26
11.1 Promoting scientific activities . . . 26
11.1.1 Scientific events: organisation . . . 26
11.1.2 Scientific events: selection . . . 26
11.1.3 Journal . . . 27
11.1.4 Invited talks . . . 27
11.1.5 Scientific expertise . . . 27
11.1.6 Research administration . . . 27
11.2 Teaching - Supervision - Juries . . . 28
11.2.1 Supervision . . . 28
11.2.2 Juries . . . 29
11.3 Popularization . . . 29
11.3.1 Education . . . 29

12 Scientific production 29
12.1 Major publications . . . 29
12.2 Publications of the year . . . 30
12.3 Cited publications . . . 31

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, from Mar 2023, HDR]
- Lionel Eyraud Dubois [INRIA, Researcher]
- Yulia Gusak [INRIA, Researcher, from Oct 2023]
- Yulia Gusak [INRIA, Starting Research Position, from Mar 2023 until Sep 2023]

Faculty Members

- Aurélien Esnard [UNIV BORDEAUX, Associate Professor]
- Mathieu Faverge [BORDEAUX INP, Associate Professor]
- Abdou Guermouche [UNIV BORDEAUX, Associate Professor, HDR]
- Pierre Ramet [UNIV BORDEAUX, Associate Professor, HDR]
- Philippe Swartvagher [BORDEAUX INP, Associate Professor, from Sep 2023]

Post-Doctoral Fellow

- Esragul Korkmaz [INRIA, Post-Doctoral Fellow, from Mar 2023]

PhD Students

- Abel Anas Calluaud [CEA, CIFRE, from Mar 2023]
- Jean Francois David [INRIA, from Mar 2023]
- Aboul-Karim Mohamed El Maarouf [IFPEN, from Mar 2023 until Apr 2023]
- Clément Richefort [CEA, from Mar 2023]
- Hayfa Tayeb [INRIA, from Sep 2023]
- Xunyi Zhao [INRIA, from Mar 2023]

Technical Staff

- Ahmed Abdourahman Mahamoud [INRIA, Engineer, from Mar 2023 until Nov 2023]
- Pierre Estérie [INRIA, Engineer, from Nov 2023]
- Zhe Li [INRIA, Engineer, from Mar 2023 until Sep 2023]
- Alycia Lisito [INRIA, Engineer, from Mar 2023]
- Marc Sergent [ATOS, Engineer, from Mar 2023]

Interns and Apprentices

- Jean Alexandre Collin [INRIA, Intern, from Mar 2023]
- Mohamed Kherraz [INRIA, Intern, from Oct 2023]
- Mohamed Kherraz [INRIA, Intern, from Jun 2023 until Jul 2023]
- Bastien Lugato [INRIA, Intern, from Nov 2023 until Nov 2023]
- Hicham Nekt [INRIA, Intern, from Jun 2023 until Sep 2023]
- Brieuc Nicolas [INRIA, Intern, from Mar 2023 until May 2023]

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 multidimensional 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, 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 [33, 26]. 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, 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.

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 [40, 45, 49], 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 [24, 23, 25]. 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, Abdou Guermouche, 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 [51]. 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 [41, 44, 53].

3.4.6 Carbon Saving and Energy-Efficient Training

Participants: Olivier Beaumont, Lionel Eyraud Dubois, Julia Gusak.

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 [27], offloading [30], their combination [28], pipelining [31]) 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 [52, 43] 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 [21, 22, 48, 20], 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 [50, 35, 56], in particular in combination with compression [36, 55, 54, 32, 46] and communication avoiding techniques [33, 26]. 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, 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 [28] re-computations [37, 27, 34, 47, 42], offloading [30], compression and different parallelism strategies (image, filter, kernel, model parallelism [31, 51, 29, 44]). 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) [38, 39] 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

6.1 Awards

- Philippe Swartvagher received the accessit price of the GDR RSD (Research Group about Networks and Distributed Systems) prix de thèse
- Xunyi Zhao gave an Oral Presentation at ICML of the paper "Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch".

6.2 Organization of Events

- Julia Gusak and Olivier Beaumont were part of the organization committee of the Workshop on Advancing Neural Network Training (WANT) at the NeurIPS 2023 conference, focusing on Computational Efficiency, Scalability, and Resource Optimization. The workshop gathered more than 200 participants.
- In the Student Cluster Competition of the SuperComputing'23 conference, the TOPAL team participated in the organization of the Reproducibility Challenge, based on a paper published at SC22 on the Symmetric Block Cyclic distribution. The Reproducibility Challenge was one of the 3 applications that the students had to work with. The objective was to reproduce the findings of the paper, on their own hardware, using the Chameleon library.
- With the Tadaam team, TOPAL organized the 15th JLESC workshop in Talence from March 21st to March 23rd. It gathered 128 participants from the different JLESC institutions.

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: Cédric Castagnede, 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

Keywords: High performance computing, HPC, Parallel computing, Graph algorithmics, Graph, Hypergraph

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: Aurelien Esnard

Participant: Aurelien 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

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-04236246

Contact: Olivier Aumage

Participants: Corentin Salingue, Andra Hugo, Benoît Lize, Cédric Augonnet, Cyril Roelandt, François Tessier, Jérôme Clet-Ortega, Ludovic Courtes, Ludovic Stordeur, Marc Sergent, Mehdi Juhoor, Nathalie Furmento, Nicolas Collin, Olivier Aumage, Pierre Wacrenier, Raymond Namyst, Samuel Thibault, Simon Archipoff, Xavier Lacoste, Terry Cojean, Yanis Khorsi, Philippe Virouleau, LoÏc Jouans, Leo Villeveygoux, Maxime Gonthier, Philippe Swartvagher, Gwenole Lucas, Romain Lion

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/`

Contact: Mathieu Faverge

Participant: Mathieu Faverge

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:

	Acis 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 Toward a multilevel method for the Helmholtz equation

Participants: Clement Richefort, Pierre Ramet.

In [13] and [14], 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, two main issues may arise when solving a Helmholtz problem. Some eigenvalues become negative or even complex, requiring the choice of an adapted smoothing method for capturing them. Moreover, since the near-kernel space is oscillatory, the geometric smoothness assumption cannot be used to build efficient interpolation rules. 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 correct an initial least squares minimization coarse selection operator formed from a set of smoothed random vectors. We also present numerical results at the end of the paper.

8.2 Programming heterogeneous architectures using hierarchical tasks

Participants: Mathieu Faverge, Abdou Guermouche, Gwenole Lucas.

Task-based systems have become popular due to their ability to utilize the computational power of complex heterogeneous systems. A typical programming model used is the Sequential Task Flow (STF) model [16], which unfortunately only supports static taskgraphs. This can result in submission overhead and a static task graph that is not well-suited for execution on heterogeneous systems. A common approach is to find abalance between the granularity needed for accelerator devices and the granularityrequired by CPU cores to achieve optimal performance. To address these issues, we have extended the STF model in the STARPU runtime system in [8] by introducing the concept of hierarchical tasks. This allows for a more dynamic task graph and, when combined with an automatic data manager, it is possible to adjust granularity at runtime to best match the targeted computing resource. That data manager makes it possible to switch between various data layout without programmer input and allows us to enforce the correctness of the DAG as hierarchical tasks alter it during runtime. Additionally, submission overhead is reduced by using large-grain hierarchical tasks, as the submission process can now be done in parallel. We have shown that the hierarchical task model is correct and have conducted an early evaluation on shared memory heterogeneous systems using the CHAMELEON dense linear algebra library.

8.3 Task-based Parallel Programming for Scalable Matrix Product Algorithms

Participants: Abdou Guermouche.

Task-based programming models have succeeded in gaining the interest of the high-performance mathematical software community because they relieve part of the burden of developing and implementing distributed- memory parallel algorithms in an efficient and portable way. In increasingly larger, more heterogeneous clusters of computers, these models appear as a way to maintain and enhance more complex algorithms. However, task-based programming models lack the flexibility and the features that are necessary to express in an elegant and compact way scalable algorithms that rely on advanced communication patterns. We showed in [6] that the Sequential Task Flow paradigm can be extended to write compact yet efficient and scalable routines for linear algebra computations. Although, this work focuses on dense General Matrix Multiplication, the proposed features enable the implementation of more complex algorithms. We describe the implementation of these features and of the resulting

GEMM operation. Finally, we present an experimental analysis on two homogeneous supercomputers showing that our approach is competitive up to 32,768 CPU cores with state-of-the-art libraries and may outperform them for some problem dimensions.

8.4 Combining reduction with synchronization barrier on multi-core processors

Participants: Abdou Guermouche, Aboul-Karim Mohamed El Maarouf.

With the rise of multicore processors with a large number of cores, the need for shared memory reduction that performs efficiently on a large number of cores is more pressing. Efficient shared memory reduction on these multicore processors will help share memory programs be more efficient. In [9], we propose a reduction method combined with a barrier method that uses SIMD read/write instructions to combine barrier signaling and reduction value to minimize memory/cache traffic between cores, thereby reducing barrier latency. We compare different barriers and reduction methods on three multicore processors and show that the proposed combining barrier/reduction methods are 4 and 3.5 times faster than respectively GCC 11.1 and Intel 21.2 OpenMP 4.5 reduction.

8.5 Data Distribution Schemes for Dense Linear Algebra Factorizations on Any Number of Nodes

Participants: Olivier Beaumont, Jean Alexandre Collin, Lionel Eyraud-Dubois.

In [12], we consider the problem of distributing the tiles of a dense matrix onto a set of homogeneous nodes. We consider both the case of non-symmetric (LU) and symmetric (Cholesky) factorizations. The efficiency of the well-known 2D Block-Cyclic (2DBC) distribution degrades significantly if the number of nodes P cannot be written as the product of two close numbers. Similarly, the recently introduced Symmetric Block Cyclic (SBC) distribution is only valid for specific values of P. In both contexts, we propose generalizations of these distributions to adapt them to any number of nodes. We show that this provides improvements to existing schemes (2DBC and SBC) both in theory and in practice, using the flexibility and ease of programming induced by task-based runtime systems like Chameleon and StarPU.

8.6 Energy or performance: the impact of implementation on consumption

Participants: Abdou Guermouche, Hicham Nekt.

In [19] we analyze the energy profile of several computational kernels. We chose kernels that perform basic operations, such as matrix product. Our goal is to study the impact of different implementations on both performance and energy consumption. The different variants covered aspects such as vectorization, accuracy, etc. In order to generalize the results, the tests are performed on different machines equipped with Intel processors, but of different types. In order to provide an in-depth answer to the question of the relationship between speed and energy efficiency, our study was based on two HPC computing application profiles, both "compute-bound" and "memory-bound" application. This approach allowed us to observe different possible energy behaviors.

8.7 On the Arithmetic Intensity of Distributed-Memory Dense Matrix Multiplication Involving a Symmetric Input Matrix (SYMM)

Participants: Lionel Eyraud-Dubois, Mathieu Faverge, Abdou Guermouche.

Dense matrix multiplication involving a symmetric input matrix (SYMM) is implemented in reference distributed-memory codes with the same data distribution as its general analogue (GEMM). We show that, when the symmetric matrix is dominant, such a 2D block-cyclic (2D BC) scheme leads to a lower arithmetic intensity (AI) of SYMM than that of GEMM by a factor of 2. We proposed in [11] alternative data distributions preserving the memory benefit of SYMM of storing only half of the matrix while achieving up to the same AI as GEMM. We also show that, in the case we can afford the same memory footprint as GEMM, SYMM can achieve a higher AI. We propose a task-based design of SYMM independent of the data distribution. This design allows for scalable A-stationary SYMM with which all discussed data distributions, may they be very irregular, can be easily assessed. We have integrated the resulting code in a reduction dimension algorithm involving a randomized singular value decomposition dominated by SYMM. An experimental study shows a compelling impact on performance.

8.8 Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch

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

In [15], we propose Rockmate 7.1.8 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 on GitHub.

8.9 H-Rockmate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks

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 [18], 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.

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.

9.1.1 PhD Theses

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.

For over a year, we have been collaborating with Eviden on the development of an HPL benchmark on top of runtime systems. This work will be continued next year as part of Alycia Lisito's thesis funded by a CIFRE contract.

9.1.2 Research Engineers

We are also involved in a bilateral collaboration with Atos as part of the recovery plan, which has led in particular to the recruitment of Marc Sergent and Ahmed Abdourahmane as research engineers.

- ATOS within the framework of French Plan de Relance: Supercomputers are equipped with gas pedals to meet computing capacity requirements. Most of these accelerators are GPUs, and the computations they perform concern traditional scientific applications, but also more recently artificial intelligence training, in particular deep neural networks, and the recognition of collective communication patterns such as broadcasts. With regard to artificial intelligence and the training of deep neural networks on supercomputers, the use of the Horovod library makes it possible to distribute artificial intelligence codes based on TensorFlow2 or PyTorch on different computing nodes. Horovod then relies on an underlying MPI communication library to optimize communications during training. Given the complexity of the memory hierarchy of supercomputers, it is therefore necessary to propose and develop new collective communications adapted to these machines and capable of accelerating the training of deep-type artificial intelligence. As part of the stimulus package, ATOS and the Inria-TOPAL project-team have set up a collaboration to develop improvements to hierarchical collective communications within the Open MPI communication library, in order to accelerate the training of deep-type artificial intelligence on supercomputers with multiple types of accelerator.

10 Partnerships and cooperations

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

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

Thomas Hérault

Status : researcher

Institution of origin: ICL/UTK

Country: US

Dates: 24 / 09 / 23 to 10 / 10 / 23

Context of the visit: Visit in the context of the Phd Defenses of M. Gonthier and G. Lucas

Mobility program/type of mobility: research stay

10.2.2 Visits to international teams

Mathieu Faverge

Visited institution: ICL / UTK

Country: US

Dates: 18 / 05 / 23 to 26 / 05 / 23

Context of the visit: Research Visit and Participation to the 16^{th} Scheduling for Large Scale Workshop

Mobility program/type of mobility: research stay

Mathieu Faverge

Visited institution: ICL / UTK

Country: US

Dates: 18 / 05 / 23 to 26 / 05 / 23

Context of the visit: Research Visit and participation to the 16^{th} Scheduling for Large Scale Workshop

Mobility program/type of mobility: research stay

Lionel Eyraud Dubois

Visited institution: University of Denver

Country: US

Dates: 11 / 12 / 23 to 15 / 12 / 23

Context of the visit: Research Visit to Julien Langou (UC Denver)

Mobility program/type of mobility: research stay

Xunyi Zhao

Visited institution: National University of Singapore

Country: US

Dates: 17 / 09 / 23 to 30 / 09 / 23

Context of the visit: Research Visit to Yang You's group at the National University of Singapore.

Mobility program/type of mobility: research stay

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:

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

SASHIMI**: Sparse Direct Solver using Hierarchical Matrices**

Duration: 2018 – 2023

Coordinator: Mathieu Faverge

Summary: Nowadays, the number of computational cores in supercomputers has grown largely to a few millions. However, the amount of memory available has not followed this trend, and the memory per core ratio is decreasing quickly with the advent of accelerators. To face this problem, the SaSHiMi project wants to tackle the memory consumption of linear solver libraries used by many major simulation applications by using low-rank compression techniques. In particular, the direct solvers which offer the most robust solution to strategy but suffer from their memory cost. The project will especially investigate the super-nodal approaches for which low-rank compression techniques have been less studied despite the attraction of their large parallelism and their lower memory cost than for the multi-frontal approaches. The results will be integrated in the PaStiX solver that supports distributed and heterogeneous architectures.

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 –

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

- Olivier Beaumont and Emmanuel Jeannot (Tadaam team) organized the 15th JLESC workshop in Talence from March 21st to March 23rd. It gathered 128 participants from the different JLESC institutions (Inria, BSC, Jülisch, Riken, ANL, U.Tennessee, NCSA). It featured discussions and exchanges on: Artificial intelligence, Big Data, I/O and in-situ visualization, Numerical methods and algorithms, Resilience, Performance tools, Programming Languages, Advanced architectures, among others.

- Yulia Gusak and Olivier Beaumont co-organized the Workshop on Advancing Neural Network Training WANT on Computational Efficiency, Scalability, and Resource Optimization. The workshop gathered around 250 participants and focused on practically addressing challenges to enhance computational efficiency, scalability, and resource optimization, with invited talks given by researches from Oakridge, DeepSpeed, ColossalAI, NVidia and Cerebras.

11.1.2 Scientific events: selection

Member of the conference program committees

- Olivier Beaumont was involved in the following programm committees: HPDC'23, ISC'23, IPDPS'23, PPAM'23

- Philippe Swartvagher was involved in the following program committees: Bench 2023 and tutorials of SC 23.

- Lionel Eyraud Dubois was part of the program committee of Euro-Par 2023 and IPDPS 2023.

Reviewer The members of the TOPAL project have also performed reviewing for the following list of conferences: Cluster 23, HPDC 23, SBAC-PAD 23.

11.1.3 Journal

Member of the editorial boards

- Olivier Beaumont is Associate Editor in Chief of the Journal of Parallel and Distributed Computing (JPDC, Elsevier)

Reviewer - reviewing activities The members of the TOPAL project have performed reviewing for IEEE Transactions on Parallel and Distributed Systems (Mathieu Faverge), ACM Transactions on Parallel Computing (Mathieu Faverge), Journal of Parallel and Distributed Computing (Mathieu Faverge, Lionel Eyraud Dubois, Abdou Guermouche), Journal of Computational and Applied Mathematics (Pierre Ramet), International Journal of High Performance Computing Applications (Pierre Ramet), ACM Transactions on Mathematical Software (Mathieu Faverge), Parallel Computing (Abdou Guermouche).

11.1.4 Invited talks

- Olivier Beaumont gave an invited talk at the 24th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing IPDPS workshop, entitled « Exotic Data Distributions for (Symmetric Dense) Linear Algebra Kernels »
- Mathieu Faverge gave an ICL Lunch Talk on March 24th entitled Programming Heterogeneous Architectures using Hierarchical Tasks.
- Lionel Eyraud Dubois was an invited speaker at the WANT workshop of NeurIPS 2023.
- Lionel Eyraud Dubois gave a talk at the Department of Mathematical and Statistical Sciences Seminar at CU Denver, on December 4th, entitled “Optimal rematerialization algorithms for Memory-efficient learning”
- Olivier Beaumont gave an invited talk entitled "Memory Saving Techniques for Training" at the 16th Scheduling for Large Scale Systems Workshop held at The University of Tennessee in Knoxille Tennessee, May 22-May 24 2023
- Mathieu Faverge gave an invited talk entitled "Programming Heterogeneous Architectures using Hierarchical Tasks" at the 16th Scheduling for Large Scale Systems Workshop held at The University of Tennessee in Knoxille Tennessee, May 22-May 24 2023

11.1.5 Scientific expertise

- Pierre Ramet is Scientific Advisor at the CEA-DAM CESTA.
- Olivier Beaumont participated in the HCERES evaluation committee of the Mathématiques et Systèmes (MathSys) unit at the Ecole de Mines de Paris.
- Olivier Beaumont participated to the evaluation committee of the Inno4scale EuroHPC call.

11.1.6 Research administration

- Aurélien Esnard is responsible for the second year of the computer science bachelor degree (L2 Informatique), which involves managing about 200 students each year.
- Aurélien Esnard is in charge of the *Numerical Transformation* mission at the College ST (Science & Technology) of the University of Bordeaux. In this context, he assists the management of the College in its decisions, informs the College’s advisors about the current issues in various projects, and leads a working group to propose improvements to business software for education and its administration.

- Abdou Guermouche is member of Scientific comittee of the LaBRI.
- Mathieu Faverge is member of the Study committee of Bordeaux INP.
- Pierre Ramet is the head of the CNRS Satanas department.
- Pierre Ramet is member of Scientific comittee of the LaBRI.

11.2 Teaching - Supervision - Juries

- Undergraduate level/Licence:
  - Aurélien Esnard: Network (54h), Software technologies (80h) at Bordeaux University.
  - Lionel Eyraud Dubois: Graphs and Algorithms (18h).
  - Mathieu Faverge: Programming environment (26h), Numerical algorithmic (25h), C projects (25h) at Bordeaux INP (ENSEIRB-MATMECA).
  - Abdou Guermouche: System programming 36h 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 (44h) at Bordeaux INP (ENSEIRB-MATMECA).
- Post graduate level/Master:
  - Aurélien Esnard: Network management (24h), Network security (24h) at Bordeaux University.
  - Lionel Eyraud Dubois: Approximation and BigData (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.
  - Abdou Guermouche: Network management 92h, Network security 64h, Operating system 24h at Bordeaux University.
  - Yulia Gusak: Deep Learning Frameworks, at Bordeaux INP (ENSEIRB-MATMECA), 20h.
  - Olivier Beaumont: Parallel Algorithms (24h) at Bordeaux INP (ENSEIRB-MATMECA).
  - Pierre Ramet: Cryptography 20h and Numerical algorithms 40h at Bordeaux INP (ENSEIRB-MATMECA).
  - Philippe Swartvagher: Parallel Algorithms (13h) at Bordeaux INP (ENSEIRB-MATMECA).

11.2.1 Supervision

- Defended PhD: Aboul-Karim Mohamed El Maarouf; Parallel fine grain imcomplete LU factorization for the solution of sparse linear systems; defended March 2023; L. Giraud, A. Guermouche.
- Defended PhD: Gwénolé Lucas; Programmation des architectures hétérogènes à l'aide de tâches divisibles; defended Oct. 2023; A. Guermouche, R. Namyst, M. Faverge, N. Furmento, P.-A. Wacrenier.
- PhD in progress: Xunyi Zhao; Memory optimization for Deep Learning Applications; started Sept. 2020; L. Eyraud Dubois, O. Beaumont.
- PhD in progress: Jean Francois David; Task-based inference for heterogeneous architectures; started Sept. 2020; L. Eyraud Dubois, O. Beaumont.

- PhD in progress: Clément Richefort; Multigrid methods applied to electromagnetism problems; started Nov. 2021; P. Ramet, M. Lecouvez (CEA Cesta).
- PhD in progress: Hayfa Tayeb; Optimization of high-performance applications on heterogeneous computing nodes; started Nov. 2021; A. Guermouche, B. Bramas, M. Faverge.
- PhD in progress: Abel Calluaud; Combined compiler and runtime approach for a direct hierarchical solver; started Nov. 2022; P. Ramet, M. Faverge, D. Lugato (CEA Cesta).
- PhD in progress: Alice Lasserre; Optimisation d'un code de calcul écrit en tâche sur calculateur à mémoire distribuée; started Oct. 2022; A. Guermouche, R. Namyst (with Airbus)
- PhD in progress: Thomas Morin; Scheduling recursive task graphs; started Oct. 2023; A. Guermouche, S. Thibault

11.2.2 Juries

- Olivier Beaumont: Reviewer for the PhD Thesis of Matthias Beaupère, advised by Laura Grigori, Sorbonne University
- Olivier Beaumont: Reviewer for the PhD Thesis of Matthias Beaupère, entitled "Algorithmes parallèles pour le calcul des décompositions de rang faible des matrices et tenseurs", advised by Laura Grigori, Sorbonne University
- Olivier Beaumont: Reviewer for the PhD Thesis of Paul Youssef, entitled "Online Learning At The Edge", advised by Nguyen Kim Thang and Denis Trystram at the University of Grenoble Alpes.
- Olivier Beaumont: examiner for the PhD thesis of Redouane Elghazi "Theoretical bounds for scheduling problems and their application to asymptotic analysis and energy consumption minimization", advised by Louis-Claude Canon and Loris Marchal at ENS-Lyon.
- Aurélien Esnard: Member of the thesis jury for Hubert Hirtz (Mesh partitioning for load balancing multiphysics simulations, CEA & Univ. Paris-Saclay), defended on 12 December 2023 in Bruyères-le-Chatel.

11.3 Popularization

11.3.1 Education

- As part of the science festival and the Bordeaux scientific circuit, Philippe Swartvagher lead a workshop with high school pupils about being a citizen in a digital world.
- As part of the "Fête de la rentrée" of Bordeaux University, Philippe Swartvagher presented Inria activities to students.
- As part of the SIF (Société Informatique de France) Education Day 2023, Aurélien Esnard presented feedback on project-based learning in the computer science degree course at the University of Bordeaux.
- Olivier Beaumont participated to the "Maths En Jeans" program with a group of students from the Lycée d'Andernos.

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] E. Agullo, A. Buttari, A. Guermouche, J. Herrmann and A. Jego. 'Task-based parallel programming for scalable matrix product algorithms'. In: *ACM Transactions on Mathematical Software* (2023). DOI: 10.1145/3583560. URL: https://hal.science/hal-03936659.

[7] A. Denis, E. Jeannot, P. Swartvagher and S. Thibault. 'Tracing task-based runtime systems: Feedbacks from the StarPU case'. In: *Concurrency and Computation: Practice and Experience* (10th Oct. 2023), p. 24. DOI: 10.1002/cpe.7920. URL: https://inria.hal.science/hal-04236246.

[8] 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.

[9] A.-k. Mohamed El Maarouf, L. Giraud, A. Guermouche and T. Guignon. 'Combining reduction with synchronization barrier on multi-core processors'. In: *Concurrency and Computation: Practice and Experience* 35.1 (10th Jan. 2023), e7402. DOI: 10.1002/cpe.7402. URL: https://inria.hal.science/hal-03948901.

National journals

[10] A. Esnard and N. Bonichon. 'Retour d'expérience sur une UE Projet en licence informatique'. In: *1024 : Bulletin de la Société Informatique de France* 22 (Nov. 2023), pp. 73–87. DOI: 10.48556/SIF.1024.22.73. URL: https://inria.hal.science/hal-04302857.

International peer-reviewed conferences

[11] E. Agullo, A. Buttari, O. Coulaud, L. Eyraud-Dubois, M. Faverge, A. Franc, A. Guermouche, A. Jego, R. Peressoni and F. Pruvost. 'On the Arithmetic Intensity of Distributed-Memory Dense Matrix Multiplication Involving a Symmetric Input Matrix (SYMM)'. In: *International Parallel and Distributed Processing Symposium.* IPDPS 2023 - 37th International Parallel and Distributed Processing Symposium. St. Petersburg, FL, United States, June 2023, pp. 357–367. URL: https://inria.hal.science/hal-04093162.

[12] O. Beaumont, J.-A. Collin, L. Eyraud-Dubois and M. Vérité. 'Data Distribution Schemes for Dense Linear Algebra Factorizations on Any Number of Nodes'. In: *Proceedings of the 37th IEEE International Parallel & Distributed Processing Symposium.* IPDPS 2023 - 37th IEEE International Parallel & Distributed Processing Symposium. St. Petersburg, Florida, United States: IEEE, 15th May 2023. URL: https://inria.hal.science/hal-04013708.

[13] 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.

Conferences without proceedings

[14] R. Falgout, M. Lecouvez, P. Ramet and C. Richefort. 'Toward a Multigrid Method for the Indefinite Helmholtz Equation'. In: CSE 2023 - SIAM Conference on Computational Science and Engineering. Amsterdam, Netherlands, 26th Feb. 2023. URL: https://hal.science/hal-04046630.

[15] 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.

Doctoral dissertations and habilitation theses

[16] G. Lucas. 'On the Use of Hierarchical Task for Heterogeneous Architectures'. Université de Bordeaux, 10th Oct. 2023. URL: https://theses.hal.science/tel-04316145.

Reports & preprints

[17] N. Bonichon and A. Esnard. *Retour d'Expérience sur une UE Projet en Licence Informatique: Exposé à la SIF lors des Journées Enseignements 2023.* Université de bordeaux, 11th May 2023. URL: https://inria.hal.science/hal-04096066.

[18] J. Gusak, X. Zhao, T. Le Hellard, Z. Li, L. Eyraud-Dubois and O. Beaumont. *H-Rockmate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks.* May 2023. URL: https://hal.science/hal-04403844.

Other scientific publications

[19] H. Nekt. 'Énergie ou performance : impact de l'implémentation sur la consommation'. Bordeaux-INP, 26th Oct. 2023. URL: https://inria.hal.science/hal-04394261.

12.3 Cited publications

[20] 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.

[21] 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.

[22] 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}.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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*. Research Report RR-9302. Inria Bordeaux Sud-Ouest, Nov. 2019. URL: https://hal.inria.fr/hal-02352969.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] T. Mary. *Block Low-Rank multifrontal solvers: complexity, performance and scalability*. Université Toulouse 3 Paul Sabatier: Ph.D. Dissertation, 2017.

[50] 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.

[51] 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.

[52] 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).

[53] 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.

[54] 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.

[55] 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.

[56] 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.