RESEARCH CENTRE

Inria Centre
at Université Grenoble Alpes

IN PARTNERSHIP WITH:

Université de Grenoble Alpes, CNRS

2023
ACTIVITY REPORT

Project-Team

DATAMOVE

Data Aware Large Scale Computing

IN COLLABORATION WITH: Laboratoire d'Informatique de Grenoble (LIG)

DOMAIN

Networks, Systems and Services, Distributed Computing

THEME

Distributed and High Performance Computing

Contents

Project-Team DATAMOVE

Creation of the Project-Team: 2017 November 01

Keywords

Computer sciences and digital sciences

A1.1.4. – High performance computing

A1.1.5. – Exascale

A1.3.6. – Fog, Edge

A1.6. – Green Computing

A2.6.2. – Middleware

A2.6.4. – Ressource management

A7.1.1. – Distributed algorithms

A7.1.2. – Parallel algorithms

A8.2.1. – Operations research

A9.9. – Distributed AI, Multi-agent

Other research topics and application domains

B3.3. – Geosciences

B6.4. – Internet of things

1 Team members, visitors, external collaborators

Research Scientists

- Bruno Raffin [Team leader, INRIA, Senior Researcher, HDR]
- Fanny Dufosse [INRIA, Researcher]

Faculty Members

- Christophe Cerin [UNIV PARIS, Professor, until Aug 2023, in delegation]
- Yves Denneulin [GRENOBLE INP, Professor, HDR]
- Pierre Dutot [UGA, Associate Professor]
- Grégory Mounié [UGA, Associate Professor]
- Kim Thang Nguyen [GRENOBLE INP, Professor]
- Olivier Richard [UGA, Associate Professor]
- Denis Trystram [GRENOBLE INP, Professor, HDR]
- Frederic Wagner [GRENOBLE INP, Associate Professor]

Post-Doctoral Fellows

- Danilo Carastan Dos Santos [UGA, Post-Doctoral Fellow]
- Marc Schouler [UGA, Post-Doctoral Fellow, until Nov 2023]

PhD Students

- Luc Angelelli [UGA]
- Abdessalam Benhari [ATOS, CIFRE]
- Louis Boulanger [UGA]
- Louis Closson [BERGER-LEVRAULT, CIFRE]
- Anderson Da Silva [RYAX]
- Yoann Dupas [ORANGE, CIFRE]
- Sofya Dymchenko [INRIA]
- Vincent Fagnon [UGA, until Aug 2023]
- Ernest Foussard [UGA]
- Dorian Goepp [UGA, from Oct 2023]
- Amal Gueroudji [CEA, until Jun 2023]
- Mathilde Jay [UGA]
- Eniko Kevi [UGA]
- Yannick Malot [CEA]
- Lucas Meyer [EDF, until Oct 2023]

- Tuan Nguyen [UGA]
- Guillaume Raffin [BULL, CIFRE, from Mar 2023]
- Hamza Safri [BERGER-LEVRAULT, CIFRE]
- Miguel Silva Vasconcelos [UGA]
- Paul Youssef [UGA, until Jan 2023]

Technical Staff

- Robert Caulk [INRIA, Engineer, until Oct 2023]
- Adrien Faure [UGA, Engineer]

Interns and Apprentices

- Samuel Brun [UGA, Intern, from Apr 2023 until Jul 2023]
- Xico Fernandez Lozano [INRIA, Intern, from Feb 2023 until Jun 2023]
- Maxime Leroy [UGA, Intern, from May 2023 until Sep 2023]
- Alexandre Lithaud [UGA, Intern, from Apr 2023 until Jul 2023]

Administrative Assistant

- Annie Simon [INRIA]

Visiting Scientist

- Luis Alejandro Torres Nino [Univ Santander, from May 2023 until Nov 2023]

External Collaborator

- Christophe Cerin [UNIV PARIS, from Sep 2023]

2 Overall objectives

Moving data on large supercomputers is becoming a major performance bottleneck, and the situation is expected to worsen even more at exascale and beyond. Data transfer capabilities are growing at a slower rate than processing power ones. The profusion of flops available will be difficult to use efficiently due to constrained communication capabilities. Moving data is also an important source of power consumption. The DataMove team focuses on **data aware large scale computing**, investigating approaches to reduce data movements on large scale HPC machines. We will investigate data aware scheduling algorithms for job management systems. The growing cost of data movements requires adapted scheduling policies able to take into account the influence of intra-application communications, IOs as well as contention caused by data traffic generated by other concurrent applications. At the same time experimenting new scheduling policies on real platforms is unfeasible. Simulation tools are required to probe novel scheduling policies. Our goal is to investigate how to extract information from actual compute centers traces in order to replay job allocations and executions with new scheduling policies. Schedulers need information about the jobs behavior on the target machine to actually make efficient allocation decisions. We will research approaches relying on learning techniques applied to execution traces to extract data and forecast job behaviors. In addition to traditional computation intensive numerical simulations, HPC platforms also need to execute more and more often data intensive processing tasks like data analysis. In particular, the ever growing amount of data generated by numerical simulation calls for a tighter integration between the simulation and the data analysis. The goal is to reduce the data traffic and to speed-up result analysis by processing results in-situ, i.e. as closely as possible to the locus and time of data generation. Our goal is here to investigate how to program and schedule such analysis workflows in the HPC context, requiring the development of adapted resource sharing strategies, data structures and parallel analytics schemes. To tackle these issues, we will intertwine theoretical research and practical developments to elaborate solutions generic and effective enough to be of practical interest. Algorithms with performance guarantees will be designed and experimented on large scale platforms with realistic usage scenarios developed with partner scientists or based on logs of the biggest available computing platforms. Conversely, our strong experimental expertise will enable to feed theoretical models with sound hypotheses, to twist proven algorithms with practical heuristics that could be further retro-feeded into adequate theoretical models.

3 Research program

3.1 Motivation

Today's largest supercomputers are composed of few millions of cores, with performances reaching 1 ExaFlops [1] for the largest machines. Moving data in such large supercomputers is becoming a major performance bottleneck, and the situation is expected to worsen even more at exascale and beyond. The data transfer capabilities are growing at a slower rate than processing power ones. The profusion of available flops will very likely be underused due to constrained communication capabilities. It is commonly admitted that data movements account for 50% to 70% of the global power consumption. Thus, data movements are potentially one of the most important source of savings for enabling supercomputers to stay in the commonly adopted energy barrier of 20 MegaWatts. In the mid to long term, non volatile memory (NVRAM) is expected to deeply change the machine I/Os. Data distribution will shift from disk arrays with an access time often considered as uniform, towards permanent storage capabilities at each node of the machine, making data locality an even more prevalent paradigm.

The proposed DataMove team will work on **optimizing data movements for large scale computing** mainly at two related levels:

- Resource allocation
- Integration of numerical simulation and data analysis

The resource and job management system (also called batch scheduler or RJMS) is in charge of allocating resources upon user requests for executing their parallel applications. The growing cost of data movements requires adapted scheduling policies able to take into account the influence of intra-application communications, I/Os as well as contention caused by data traffic generated by other concurrent applications. Modelling the application behavior to anticipate its actual resource usage on such architecture is known to be challenging, but it becomes critical for improving performances (execution time, energy, or any other relevant objective). The job management system also needs to handle new types of workloads: high performance platforms now need to execute more and more often data intensive processing tasks like data analysis in addition to traditional computation intensive numerical simulations. In particular, the ever growing amount of data generated by numerical simulation calls for a tighter integration between the simulation and the data analysis. The challenge here is to reduce data traffic and to speed-up result analysis by performing result processing (compression, indexation, analysis, visualization, etc.) as closely as possible to the locus and time of data generation. This emerging trend called *in-situ analytics* requires to revisit the traditional workflow (loop of batch processing followed by postmortem analysis). The application becomes a whole including the simulation, in-situ processing and I/Os. This motivates the development of new well-adapted resource sharing strategies, data structures and parallel analytics schemes to efficiently interleave the different components of the application and globally improve the performance.

[1] 10^{18} floating point operations per second

3.2 Strategy

DataMove targets HPC (High Performance Computing) at Exascale. But such machines and the associated applications are expected to be available only in 5 to 10 years. Meanwhile, we expect to see a growing number of petaflop machines to answer the needs for advanced numerical simulations. A sustainable exploitation of these petaflop machines is a real and hard challenge that we will address. We may also see in the coming years a convergence between HPC and Big Data, HPC platforms becoming more elastic and supporting Big Data jobs, or HPC applications being more commonly executed on cloud like architectures. This is the second top objective of the 2015 US Strategic Computing Initiative: *Increasing coherence between the technology base used for modelling and simulation and that used for data analytic computing*. We will contribute to that convergence at our level, considering more dynamic and versatile target platforms and types of workloads.

Our approaches should entail minimal modifications on the code of numerical simulations. Often large scale numerical simulations are complex domain specific codes with a long life span. We assume these codes as being sufficiently optimized. We will influence the behavior of numerical simulations through resource allocation at the job management system level or when interleaving them with analytics code.

To tackle these issues, we propose to intertwine theoretical research and practical developments in an agile mode. Algorithms with performance guarantees will be designed and experimented on large scale platforms with realistic usage scenarios developed with partner scientists or based on logs of the biggest available computing platforms (national supercomputers like Curie, or the BlueWaters machine accessible through our collaboration with Argonne National Lab). Conversely, a strong experimental expertise will enable to feed theoretical models with sound hypotheses, to twist proven algorithms with practical heuristics that could be further retro-feeded into adequate theoretical models.

A central scientific question is to make the relevant choices for optimizing performance (in a broad sense) in a reasonable time. HPC architectures and applications are increasingly complex systems (heterogeneity, dynamicity, uncertainties), which leads to consider the **optimization of resource allocation based on multiple objectives**, often contradictory (like energy and run-time for instance). Focusing on the optimization of one particular objective usually leads to worsen the others. The historical positioning of some members of the team who are specialists in multi-objective optimization is to generate a (limited) set of trade-off configurations, called *Pareto points*, and choose when required the most suitable trade-off between all the objectives. This methodology differs from the classical approaches, which simplify the problem into a single objective one (focus on a particular objective, combining the various objectives or agglomerate them). The real challenge is thus to combine algorithmic techniques to account for this diversity while guaranteeing a target efficiency for all the various objectives.

The DataMove team aims to elaborate generic and effective solutions of practical interest. We will make our new algorithms accessible through the team flagship software tools, **the OAR batch scheduler and the Ensemble run online data processing framework Melissa**. We will maintain and enforce strong links with teams closely connected with large architecture design and operation (CEA DAM, BULL, Argonne National Lab), as well as scientists of other disciplines, in particular computational biologists, with whom we will elaborate and validate new usage scenarios (IBPC, CEA DAM, EDF).

3.3 Research Directions

DataMove research activity is organized around three directions:

1. When a parallel job executes on a machine, it triggers data movements through the input data it needs to read, the results it produces (simulation results as well as traces) that need to be stored in the file system, as well as internal communications and temporary storage (for fault tolerance related data for instance). Modeling in details the simulation and the target machines to analyze scheduling policies is not feasible at large scales. We propose to investigate alternative approaches, including learning approaches, to capture and model the influence of data movements on the performance metrics of each job execution to develop **Data Aware Batch Scheduling** models and algorithms (Sec. 4.1).

2. Experimenting new scheduling policies on real platforms at scale is unfeasible. Theoretical per-

formance guarantees are not sufficient to ensure a new algorithm will actually perform as expected on a real platform. An intermediate evaluation level is required to probe novel scheduling policies. The second research axe focuses on the **Empirical Studies of Large Scale Platforms** (Sec. 4.2). The goal is to investigate how we could extract from actual computing centers traces information to replay the job allocations and executions on a simulated or emulated platform with new scheduling policies. Schedulers need information about jobs behavior on target machines to actually be able to make efficient allocation decisions. Asking users to characterize jobs often does not lead to reliable information.

3. The third research direction **Integration of High Performance Computing and Data Analytics** (Sec. 4.3) addresses the data movement issue from a different perspective. New data analysis techniques on the HPC platform introduce new type of workloads, potentially more data than compute intensive, but could also enable to reduce data movements by directly enabling to pipeline simulation execution with a live (in situ) analysis of the produced results. Our goal is here to investigate how to program and schedule such analysis workflows in the HPC context.

4 Application domains

4.1 Data Aware Batch Scheduling

Large scale high performance computing platforms are becoming increasingly complex. Determining efficient allocation and scheduling strategies that can adapt to technological evolutions is a strategic and difficult challenge. We are interested in scheduling jobs in hierarchical and heterogeneous large scale platforms. On such platforms, application developers typically submit their jobs in centralized waiting queues. The job management system aims at determining a suitable allocation for the jobs, which all compete against each other for the available computing resources. Performances are measured using different classical metrics like maximum completion time or slowdown. Current systems make use of very simple (but fast) algorithms that however rely on simplistic platform and execution models, and thus, have limited performances.

For all target scheduling problems we aim to provide both theoretical analysis and complementary analysis through simulations. Achieving meaningful results will require strong improvements on existing models (on power for example) and the design of new approximation algorithms with various objectives such as stretch, reliability, throughput or energy consumption, while keeping in focus the need for a low-degree polynomial complexity.

4.1.1 Algorithms

The most common batch scheduling policy is to consider the jobs according to the First Come First Served order (FCFS) with backfilling (BF). BF is the most widely used policy due to its easy and robust implementation and known benefits such as high system utilization. It is well-known that this strategy does not optimize any sophisticated function, but it is simple to implement and it guarantees that there is no starvation (i.e. every job will be scheduled at some moment).

More advanced algorithms are seldom used on production platforms due to both the gap between theoretical models and practical systems and speed constraints. When looking at theoretical scheduling problems, the generally accepted goal is to provide polynomial algorithms (in the number of submitted jobs and the number of involved computing units). However, with millions of processing cores where every process and data transfer have to be individually scheduled, polynomial algorithms are prohibitive as soon as the polynomial degree is too large. The model of *parallel tasks* simplifies this problem by bundling many threads and communications into single boxes, either rigid, rectangular or malleable. Especially malleable tasks capture the dynamicity of the execution. Yet these models are ill-adapted to heterogeneous platforms, as the running time depends on more than simply the number of allotted resources, and some of the common underlying assumptions on the speed-up functions (such as monotony or concavity) are most often only partially verified.

In practice, the job execution times depend on their allocation (due to communication interferences and heterogeneity in both computation and communication), while theoretical models of parallel jobs usually consider jobs as black boxes with a fixed (maximum) execution time. Though interesting and powerful, the classical models (namely, synchronous PRAM model, delay, LogP) and their variants (such as hierarchical delay), are not well-suited to large scale parallelism on platforms where the cost of moving data is significant, non uniform and may change over time. Recent studies are still refining such models in order to take into account communication contentions more accurately while remaining tractable enough to provide a useful tool for algorithm design.

Today, all algorithms in use in production systems are oblivious to communications. One of our main goals is to **design a new generation of scheduling algorithms fitting more closely job schedules according to platform topologies**.

4.1.2 Locality Aware Allocations

Recently, we developed modifications of the standard back-filling algorithm taking into account platform topologies. The proposed algorithms take into account locality and contiguity in order to hide communication patterns within parallel tasks. The main result here is to establish good lower bounds and small approximation ratios for policies respecting the locality constraints. The algorithms work in an online fashion, improving the global behavior of the system while still keeping a low running time. These improvements rely mainly on our past experience in designing approximation algorithms. Instead of relying on complex networking models and communication patterns for estimating execution times, the communications are disconnected from the execution time. Then, the scheduling problem leads to a trade-off: optimizing locality of communications on one side and a performance objective (like the makespan or stretch) on the other side.

In the perspective of taking care of locality, other ongoing works include the study of schedulers for platforms whose interconnection network is a static structured topology (like the 3D-torus of the BlueWaters platform we work on in collaboration with the Argonne National Laboratory). One main characteristic of this 3D-torus platform is to provide I/O nodes at specific locations in the topology. Applications generate and access specific data and are thus bounded to specific I/O nodes. Resource allocations are constrained in a strong and unusual way. This problem is close for actual hierarchical platforms. The scheduler needs to compute a schedule such that I/O nodes requirements are filled for each application while at the same time avoiding communication interferences. Moreover, extra constraints can arise for applications requiring accelerators that are gathered on the nodes at the edge of the network topology.

While current results are encouraging, they are however limited in performance by the low amount of information available to the scheduler. We look forward to extend ongoing work by progressively increasing application and network knowledge (by technical mechanisms like profiling or monitoring or by more sophisticated methods like learning). It is also important to anticipate on application resource usage in terms of compute units, memory as well as network and I/Os to efficiently schedule a mix of applications with different profiles. For instance, a simple solution is to partition the jobs as "communication intensive" or "low communications". Such a tag could be achieved by the users them selves or obtained by learning techniques. We could then schedule low communications jobs using leftover spaces while taking care of high communication jobs. More sophisticated options are possible, for instance those that use more detailed communication patterns and networking models. Such options would leverage the work proposed in Section 4.2 for gathering application traces.

4.1.3 Data-Centric Processing

Exascale computing is shifting away from the traditional compute-centric models to a more data-centric one. This is driven by the evolving nature of large scale distributed computing, no longer dominated by pure computations but also by the need to handle and analyze large volumes of data. These data can be large databases of results, data streamed from a running application or another scientific instrument (collider for instance). These new workloads call for specific resource allocation strategies.

Data movements and storage are expected to be a major energy and performance bottleneck on next generation platforms. Storage architectures are also evolving, the standard centralized parallel file system being complemented with local persistent storage (Burst Buffers, NVRAM). Thus, one data producer can stage data on some nodes' local storage, requiring to schedule close by the associated analytics tasks to limit data movements. This kind of configuration, often referred as *in-situ analytics*, is expected to become common as it enables to switch from the traditional I/O intensive workflow (batch-processing followed by *post mortem* analysis and visualization) to a more storage conscious approach where data are processed as closely as possible to where and when they are produced (in-situ processing is addressed in details in section 4.3). By reducing data movements and scheduling the extra processing on resources not fully exploited yet, in-situ processing is expected to have also a significant positive energetic impact. Analytics codes can be executed in the same nodes than the application, often on dedicated cores commonly called helper cores, or on dedicated nodes called stagging nodes. The results are either forwarded to the users for visualization or saved to disk through I/O nodes. In-situ analytics can also take benefit of node local disks or burst buffers to reduce data movements. Future job scheduling strategies should take into account in-situ processes in addition to the job allocation to optimize both energy consumption and execution time. On the one hand, this problem can be reduced to an allocation problem of extra asynchronous tasks to idle computing units. But on the other hand, embedding analytics in applications brings extra difficulties by making the application more heterogeneous and imposing more constraints (data affinity) on the required resources. Thus, the main point here is to develop efficient algorithms for dealing with heterogeneity without increasing the global computational cost.

4.1.4 Learning

Another important issue is to adapt the job management system to deal with the bad effects of uncertainties, which may be catastrophic in large scale heterogeneous HPC platforms (jobs delayed arbitrarly far or jobs killed). A natural question is then: *is it possible to have a good estimation of the job and platform parameters in order to be able to obtain a better scheduling ?* Many important parameters (like the number or type of required resources or the estimated running time of the jobs) are asked to the users when they submit their jobs. However, some of these values are not accurate and in many cases, they are not even provided by the end-users. In DataMove, we propose to study new methods for a better prediction of the characteristics of the jobs and their execution in order to improve the optimization process. In particular, the methods well-studied in the field of big data (in supervised Machine Learning, like classical regression methods, Support Vector Methods, random forests, learning to rank techniques or deep learning) could and must be used to improve job scheduling in large scale HPC platforms. This topic received a great attention recently in the field of parallel and distributed processing. A preliminary study has been done recently by our team with the target of predicting the job running times (called wall times). We succeeded to improve significantly in average the reference EASY Back Filling algorithm by estimating the wall time of the jobs, however, this method leads to big delay for the stretch of few jobs. Even if we succeed in determining more precisely hidden parameters, like the wall time of the jobs, this is not enough to determine an optimized solution. The shift is not only to learn on dedicated parameters but also on the scheduling policy. The data collected from the accounting and profiling of jobs can be used to better understand the needs of the jobs and through learning to propose adaptations for future submissions. The goal is to propose extensions to further improve the job scheduling and improve the performance and energy efficiency of the application. For instance preference learning may enable to compute on-line new priorities to back-fill the ready jobs.

4.1.5 Multi-objective Optimization

Several optimization questions that arise in allocation and scheduling problems lead to the study of several objectives at the same time. The goal is then not a single optimal solution, but a more complicated mathematical object that captures the notion of trade-off. In broader terms, the goal of multi-objective optimization is not to externally arbitrate on disputes between entities with different goals, but rather to explore the possible solutions to highlight the whole range of interesting compromises. A classical tool for studying such multi-objective optimization problems is to use *Pareto curves*. However, the full description of the Pareto curve can be very hard because of both the number of solutions and the hardness of computing each point. Addressing this problem will opens new methodologies for the analysis of algorithms.

To further illustrate this point here are three possible case studies with emphasis on conflicting interests measured with different objectives. While these cases are good representatives of our HPC context, there are other pertinent trade-offs we may investigate depending on the technology evolution in the coming years. This enumeration is certainly not limitative.

Energy versus Performance. The classical scheduling algorithms designed for the purpose of performance can no longer be used because performance and energy are contradictory objectives to some extent. The scheduling problem with energy becomes a multi-objective problem in nature since the energy consumption should be considered as equally important as performance at exascale. A global constraint on energy could be a first idea for determining trade-offs but the knowledge of the Pareto set (or an approximation of it) is also very useful.

Administrators versus application developers. Both are naturally interested in different objectives: In current algorithms, the performance is mainly computed from the point of view of administrators, but the users should be in the loop since they can give useful information and help to the construction of better schedules. Hence, we face again a multi-objective problem where, as in the above case, the approximation of the Pareto set provides the trade-off between the administrator view and user demands. Moreover, the objectives are usually of the same nature. For example, *max stretch* and *average stretch* are two objectives based on the slowdown factor that can interest administrators and users, respectively. In this case the study of the norm of stretch can be also used to describe the trade-off (recall that the L_1-norm corresponds to the average objective while the L_∞-norm to the max objective). Ideally, we would like to design an algorithm that gives good approximate solutions at the same time for all norms. The L_2 or L_3-norm are useful since they describe the performance of the whole schedule from the administrator point of view as well as they provide a fairness indication to the users. The hard point here is to derive theoretical analysis for such complicated tools.

Resource Augmentation. The classical resource augmentation models, i.e. speed and machine augmentation, are not sufficient to get good results when the execution of jobs cannot be frequently interrupted. However, based on a resource augmentation model recently introduced, where the algorithm may reject a small number of jobs, some members of our team have given the first interesting results in the non-preemptive direction. In general, resource augmentation can explain the intuitive good behavior of some greedy algorithms while, more interestingly, it can give ideas for new algorithms. For example, in the rejection context we could dedicate a small number of nodes for the usually problematic rejected jobs. Some initial experiments show that this can lead to a schedule for the remaining jobs that is very close to the optimal one.

4.2 Empirical Studies of Large Scale Platforms

Experiments or realistic simulations are required to take into account the impact of allocations and assess the real behavior of scheduling algorithms. While theoretical models still have their interest to lay the groundwork for algorithmic designs, the models are necessarily reflecting a purified view of the reality. As transferring our algorithm in a more practical setting is an important part of our creed, we need to ensure that the theoretical results found using simplified models can really be transposed to real situations. On the way to exascale computing, large scale systems become harder to study, to develop or to calibrate because of the costs in both time and energy of such processes. It is often impossible to convince managers to use a production cluster for several hours simply to test modifications in the RJMS. Moreover, as the existing RJMS production systems need to be highly reliable, each evolution requires several real scale test iterations. The consequence is that scheduling algorithms used in production systems are mostly outdated and not customized correctly. To circumvent this pitfall, we need to develop tools and methodologies for alternative empirical studies, from analysis of workload traces, to job models, simulation and emulation with reproducibility concerns.

4.2.1 Workload Traces with Resource Consumption

Workload traces are the base element to capture the behavior of complete systems composed of submitted jobs, running applications, and operating tools. These traces must be obtained on production platforms to provide relevant and representative data. To get a better understanding of the use of such systems, we need to look at both, how the jobs interact with the job management system, and how they use the allocated resources. We propose a general workload trace format that adds jobs resource consumption to the commonly used Standard Workload Format workload trace format. This requires to instrument the platforms, in particular to trace resource consumptions like CPU, data movements at memory, network and I/O levels, with an acceptable performance impact. In a previous work we studied and proposed a dedicated job monitoring tool whose impact on the system has been measured as lightweight (0.35% speed-down) with a 1 minute sampling rate. Other tools also explore job monitoring, like TACC Stats. A unique feature from our tool is its ability to monitor distinctly jobs sharing common nodes.

Collected workload traces with jobs resource consumption will be publicly released and serve to provide data for works presented in Section 4.1. The trace analysis is expected to give valuable insights to define models encompassing complex behaviours like network topology sensitivity, network congestion and resource interferences.

We expect to join efforts with partners for collecting quality traces (ATOS/Bull, Ciment meso center, Joint Laboratory on Extreme Scale Computing) and will collaborate with the INRIA team POLARIS for their analysis.

4.2.2 Simulation

Simulations of large scale systems are faster by multiple orders of magnitude than real experiments. Unfortunately, replacing experiments with simulations is not as easy as it may sound, as it brings a host of new problems to address in order to ensure that the simulations are closely approximating the execution of typical workloads on real production clusters. Most of these problems are actually not directly related to scheduling algorithms assessment, in the sense that the workload and platform models should be defined independently from the algorithm evaluations, in order to ensure a fair assessment of the algorithms' strengths and weaknesses. These research topics (namely platform modeling, job models and simulator calibration) are addressed in the other subsections.

We developed an open source platform simulator within DataMove (in conjunction with the OAR development team) to provide a widely distributable test bed for reproducible scheduling algorithm evaluation. Our simulator, named Batsim, allows to simulate the behavior of a computational platform executing a workload scheduled by any given scheduling algorithm. To obtain sound simulation results and to broaden the scope of the experiments that can be done thanks to Batsim, we did not chose to create a (necessarily limited) simulator from scratch, but instead to build on top of the SimGrid simulation framework.

To be open to as many batch schedulers as possible, Batsim decouples the platform simulation and the scheduling decisions in two clearly-separated software components communicating through a complete and documented protocol. The Batsim component is in charge of simulating the computational resources behaviour whereas the scheduler component is in charge of taking scheduling decisions. The scheduler component may be both a resource and a job management system. For jobs, scheduling decisions can be to execute a job, to delay its execution or simply to reject it. For resources, other decisions can be taken, for example to change the power state of a machine i.e. to change its speed (in order to lower its energy consumption) or to switch it on or off. This separation of concerns also enables interfacing with potentially any commercial RJMS, as long as the communication protocol with Batsim is implemented. A proof of concept is already available with the OAR RJMS.

Using this test bed opens new research perspectives. It allows to test a large range of platforms and workloads to better understand the real behavior of our algorithms in a production setting. In turn, this opens the possibility to tailor algorithms for a particular platform or application, and to precisely identify the possible shortcomings of the theoretical models used.

4.2.3 Job and Platform Models

The central purpose of the Batsim simulator is to simulate job behaviors on a given target platform under a given resource allocation policy. Depending on the workload, a significant number of jobs are parallel applications with communications and file system accesses. It is not conceivable to simulate individually all these operations for each job on large plaforms with their associated workload due to implied simulation complexity. The challenge is to define a coarse grain job model accurate enough to reproduce parallel application behavior according to the target platform characteristics. We will explore models similar to the BSP (Bulk Synchronous Program) approach that decomposes an application in local computation supersteps ended by global communications and a global synchronization. The model parameters will be established by means of trace analysis as discussed previously, but also by instrumenting some parallel applications to capture communication patterns. This instrumentation will have a significant impact on the concerned application performance, restricting its use to a few applications only. There are a lot of recurrent applications executed on HPC platform, this fact will help to reduce the required number of instrumentations and captures. To assign each job a model, we are considering to adapt the concept of application signatures as proposed in. Platform models and their calibration are also required. Large parts of these models, like those related to network, are provided by Simgrid. Other parts as the filesystem and energy models are comparatively recent and will need to be enhanced or reworked to reflect the HPC platform evolutions. These models are then generally calibrated by running suitable benchmarks.

4.2.4 Emulation and Reproducibility

The use of coarse models in simulation implies to set aside some details. This simplification may hide system behaviors that could impact significantly and negatively the metrics we try to enhance. This issue is particularly relevant when large scale platforms are considered due to the impossibility to run tests at nominal scale on these real platforms. A common approach to circumvent this issue is the use of emulation techniques to reproduce, under certain conditions, the behavior of large platforms on smaller ones. Emulation represents a natural complement to simulation by allowing to execute directly large parts of the actual evaluated software and system, but at the price of larger compute times and a need for more resources. The emulation approach was chosen in to compare two job management systems from workload traces of the CURIE supercomputer (80000 cores). The challenge is to design methods and tools to emulate with sufficient accuracy the platform and the workload (data movement, I/O transfers, communication, applications interference). We will also intend to leverage emulation tools like Distem from the MADYNES team. It is also important to note that the Batsim simulator also uses emulation techniques to support the core scheduling module from actual RJMS. But the integration level is not the same when considering emulation for larger parts of the system (RJMS, compute node, network and filesystem).

Replaying traces implies to prepare and manage complex software stacks including the OS, the resource management system, the distributed filesystem and the applications as well as the tools required to conduct experiments. Preparing these stacks generate specific issues, one of the major one being the support for reproducibility. We propose to further develop the concept of reconstructability to improve experiment reproducibility by capturing the build process of the complete software stack. This approach ensures reproducibility over time better than other ways by keeping all data (original packages, build recipe and Kameleon engine) needed to build the software stack.

In this context, the Grid’5000 (see Sec. 7.2) experimentation infrastructure that gives users the control on the complete software stack is a crucial tool for our research goals. We will pursue our strong implication in this infrastructure.

4.3 Integration of High Performance Computing and Data Analytics

Data produced by large simulations are traditionally handled by an I/O layer that moves them from the compute cores to the file system. Analysis of these data are performed after reading them back from files, using some domain specific codes or some scientific visualisation libraries like VTK. But writing and then reading back these data generates a lot of data movements and puts under pressure the file system. To reduce these data movements, **the in situ analytics paradigm proposes to process the data as closely as possible to where and when the data are produced**. Some early solutions emerged either as extensions of visualisation tools or of I/O libraries like ADIOS. But significant progresses are still required to provide efficient and flexible high performance scientific data analysis tools. Integrating data analytics in the HPC context will have an impact on resource allocation strategies, analysis algorithms, data storage and access, as well as computer architectures and software infrastructures. But this paradigm shift imposed by the machine performance also sets the basis for a deep change on the way users work with numerical simulations. The traditional workflow needs to be reinvented to make HPC more user-centric, more interactive and turn HPC into a commodity tool for scientific discovery and engineering developments. In this context DataMove aims at investigating programming environments for in situ analytics with a specific focus on task scheduling in particular, to ensure an efficient sharing of resources with the simulation.

4.3.1 Programming Model and Software Architecture

In situ creates a tighter loop between the scientist and her/his simulation. As such, an in situ framework needs to be flexible to let the user define and deploy its own set of analysis. A manageable flexibility requires to favor simplicity and understandability, while still enabling an efficient use of parallel resources. Visualization libraries like VTK or Visit, as well as domain specific environments like VMD have initially been developed for traditional post-mortem data analysis. They have been extended to support in situ processing with some simple resource allocation strategies but the level of performance, flexibility and ease of use that is expected requires to rethink new environments. There is a need to develop a middleware and programming environment taking into account in its fundations this specific context of high performance scientific analytics.

Similar needs for new data processing architectures occurred for the emerging area of Big Data Analytics, mainly targeted to web data on cloud-based infrastructures. Google Map/Reduce and its successors like Spark or Stratosphere/Flink have been designed to match the specific context of efficient analytics for large volumes of data produced on the web, on social networks, or generated by business applications. These systems have mainly been developed for cloud infrastructures based on commodity architectures. They do not leverage the specifics of HPC infrastructures. Some preliminary adaptations have been proposed for handling scientific data in a HPC context. However, these approaches do not support in situ processing.

Following the initial development of FlowVR, our middleware for in situ processing, we will pursue our effort to develop a programming environment and software architecture for high performance scientific data analytics. Like FlowVR, the map/reduce tools, as well as the machine learning frameworks like TensorFlow, adopted a dataflow graph for expressing analytics pipe-lines. We are convinced that this dataflow approach is both easy to understand and yet expresses enough concurrency to enable efficient executions. The graph description can be compiled towards lower level representations, a mechanism that is intensively used by Stratosphere/Flink for instance. Existing in situ frameworks, including FlowVR, inherit from the HPC way of programming with a thiner software stack and a programming model close to the machine. Though this approach enables to program high performance applications, this is usually too low level to enable the scientist to write its analysis pipe-line in a short amount of time. The data model, i.e. the data semantics level accessible at the framework level for error check and optimizations, is also a fundamental aspect of such environments. The key/value store has been adopted by all map/reduce tools. Except in some situations, it cannot be adopted as such for scientific data. Results from numerical simulations are often more structured than web data, associated with acceleration data structures to be processed efficiently. We will investigate data models for scientific data building on existing approaches like Adios or DataSpaces.

4.3.2 Resource Sharing

To alleviate the I/O bottleneck, the in situ paradigm proposes to start processing data as soon as made available by the simulation, while still residing in the memory of the compute node. In situ processings include data compression, indexing, computation of various types of descriptors (1D, 2D, images, etc.). Per se, reducing data output to limit I/O related performance drops or keep the output data size manageable is not new. Scientists have relied on solutions as simple as decreasing the frequency of result savings. In situ processing proposes to move one step further, by providing a full fledged processing framework enabling scientists to more easily and thoroughly manage the available I/O budget.

The most direct way to perform in situ analytics is to inline computations directly in the simulation code. In this case, in situ processing is executed in sequence with the simulation that is suspended meanwhile. Though this approach is direct to implement and does not require complex framework environments, it does not enable to overlap analytics related computations and data movements with the simulation execution, preventing to efficiently use the available resources. Instead of relying on this simple time sharing approach, several works propose to rely on space sharing where one or several cores per node, called *helper cores*, are dedicated to analytics. The simulation responsibility is simply to handle a copy of the relevant data to the node-local in situ processes, both codes being executed concurrently. This approach often lead to significantly beter performance than in-simulation analytics.

For a better isolation of the simulation and in situ processes, one solution consists in offloading in situ tasks from the simulation nodes towards extra dedicated nodes, usually called *staging nodes*. These computations are said to be performed *in-transit*. But this approach may not always be beneficial compared to processing on simulation nodes due to the costs of moving the data from the simulation nodes to the staging nodes.

FlowVR enables to mix these different resources allocation strategies for the different stages of an analytics pile-line. Based on a component model, the scientist designs analytics workflows by first developing processing components that are next assembled in a dataflow graph through a Python script. At runtime the graph is instantiated according to the execution context, FlowVR taking care of deploying the application on the target architecture, and of coordinating the analytics workflows with the simulation execution.

But today the choice of the resource allocation strategy is mostly ad-hoc and defined by the programmer. We will investigate solutions that enable a cooperative use of the resource between the analytics and the simulation with minimal hints from the programmer. In situ processings inherit from the parallelization scale and data distribution adopted by the simulation, and must execute with minimal perturbations on the simulation execution (whose actual resource usage is difficult to know a priori). We need to develop adapted scheduling strategies that operate at compile and run time. Because analysis are often data intensive, such solutions must take into consideration data movements, a point that classical scheduling strategies designed first for compute intensive applications often overlook. We expect to develop new scheduling strategies relying on the methodologies developed in Sec. 4.1.5. Simulations as well as analysis are iterative processes exposing a strong spatial and temporal coherency that we can take benefit of to anticipate their behavior and then take more relevant resources allocation strategies, possibly based on advanced learning algorithms or as developed in Section 4.1.

In situ analytics represent a specific workload that needs to be scheduled very closely to the simulation, but not necessarily active during the full extent of the simulation execution and that may also require to access data from previous runs (stored in the file system or on specific burst-buffers). Several users may also need to run concurrent analytics pipe-lines on shared data. This departs significantly from the traditional batch scheduling model, motivating the need for a more elastic approach to resource provisioning. These issues will be conjointly addressed with research on batch scheduling policies (Sec. 4.1).

4.3.3 Co-Design with Data Scientists

Given the importance of users in this context, it is of primary importance that in situ tools be co-designed with advanced users, even if such multidisciplinary collaborations are challenging and require constant long term investments to learn and understand the specific practices and expectations of the other domain.

We will tightly collaborate with scientists of some application domains, like molecular dynamics or fluid simulation, to design, develop, deploy and assess in situ analytics scenarios.

5 Social and environmental responsibility

DataMove is environmentally involved at different levels:

- Pursuing research on energy optimization of large scale distributed compute infrastructures
- Intend to include in publications the total amount of compute hours required for running all associated experiments, especially when using supercomputers, to, in a first step, get a measure of the impact of our experimentation activity.
- Lead and participate to different local LIG and INRIA groups in charge of evaluating, proposing and implementing solutions to limit our environmental impact in the lab.

- Take actions for lowering our carbon impact (extend laptop, smart phones, servers life to 5-8 years, favor fixing equipment rather then replacing them, put priority on train rather than plane)
- Bicycle is just our favorite, very low carbon, way for commuting.

6 Highlights of the year

DataMove organized the second Journées de Recherche en Apprentissage Frugal at Grenoble.

7 New software, platforms, open data

7.1 New software

7.1.1 OAR

Keywords: HPC, Cloud, Clusters, Resource manager, Light grid

Scientific Description: This batch system is based on a database (PostgreSQL (preferred) or MySQL), a script language (Perl) and an optional scalable administrative tool (e.g. Taktuk). It is composed of modules which interact mainly via the database and are executed as independent programs. Therefore, formally, there is no API, the system interaction is completely defined by the database schema. This approach eases the development of specific modules. Indeed, each module (such as schedulers) may be developed in any language having a database access library.

Functional Description: OAR is a versatile resource and task manager (also called a batch scheduler) for HPC clusters, and other computing infrastructures (like distributed computing experimental testbeds where versatility is a key).

URL: http://oar.imag.fr

Contact: Olivier Richard

Participants: Bruno Bzeznik, Olivier Richard, Pierre Neyron

Partners: LIG, CNRS, Grid'5000, CIMENT, UAR GRICAD

7.1.2 MELISSA

Keywords: Sensitivity Analysis, HPC, Data assimilation, Exascale

Functional Description: Melissa is a middleware framework for on-line processing of data produced from large scale ensemble runs (parameter sweep data analysis). Initial developments focused on sensibility analysis, Melissa relying on iterative statistics to provide a file avoiding, fault tolerant and elastic framework. Largest runs so far involved up to 30k core, executed 80 000 parallel simulations, and generated 288 TB of intermediate data that did not need to be stored on the file system. Melissa was next extended to large scale data assimilation, with the integration of a simulation (or member) virtualization mechanism that enables to abstract the number of members from the actual resource allocations, further improving execution efficiency and elasticity. Latest contribution is support for training deep surrogate models with support for Phytorch and Tensorflow

URL: https://gitlab.inria.fr/melissa

Publications: hal-04145897, hal-04213978, hal-04102400, hal-01383860, hal-01607479, hal-03017033, hal-03927612, hal-03842106

Authors: Theophile Terraz, Bruno Raffin, Alejandro Ribes, Bertrand Iooss

Contact: Bruno Raffin

Partner: Edf

7.1.3 NixOS-Compose

Keywords: Infrastructure software, Deployment, High performance computing, Distributed computing

Functional Description: NixOS-Compose simplifies the process of setting up ephemeral distributed systems by utilizing Nix's functional package management and NixOS's declarative configuration management. The tool facilitates testing, development, infrastructure prototyping, benchmarking, and advanced experiments in high-performance computing by providing easy and reproducible software stack deployment.

URL: `https://gitlab.inria.fr/nixos-compose/nixos-compose`

Publication: hal-03723771

Contact: Olivier Richard

Partners: LIG, CNRS, UGA

7.1.4 Batsim

Keywords: Simulation, Distributed systems

Functional Description: BatSim is a Resource and Job Management System (RJMS) framework simulator based on SimGrid. It aims at taking into account platform's hardware capabilities and impacts in simulations. Also, schedulers parts are plugable through a comprehensive API and they are seen as external component of the framework.

Release Contributions: see https://batsim.readthedocs.io/en/latest/changelog.html

URL: `https://batsim.readthedocs.io/en/latest/`

Contact: Olivier Richard

Partner: IRIT

7.2 New platforms

Participants: Olivier Richard.

7.2.1 SILECS/Grid'5000 and Meso Center Ciment

We are very active in promoting the factorization of compute resources at a regional and national level. We have a three level implication, locally to maintain a pool of very flexible experimental machines (hundreds of cores), regionally through the CIMENT meso center, and nationally by contributing to the SILECS/Grid'5000 platform, our local resources being included in this platform. Olivier Richard is member of SILECS/Grid'5000 scientific committee. The OAR scheduler in particular is deployed on both infrastructures. DataMove is hosting several ingineers dedicated to Grid'5000 support.

8 New results

8.1 Algorithms for Resource Allocations

Participants: Denis Trystram, Fanny Dufossé, Gregory Mounié, Pierre-François Dutot, Kim Thang Nguyen.

8.1.1 Multi-objective Scheduling Policy for Serverless-based Edge-Cloud Continuum

The cloud is extended towards the edge to form a computing continuum while managing resources' heterogeneity. The serverless technology simplified how to build cloud applications and use resources, becoming a driving force in consolidating the continuum with the deployment of small functions with short execution. However, the adaptation of serverless to the edge-cloud continuum brings new challenges mainly related to resource management and scheduling. Standard cloud scheduling policies are based on greedy algorithms that do not efficiently handle platforms' heterogeneity nor deal with problems such as cold start delays. This work introduces a new scheduling policy that tries to address these issues. It is based on multi-objective optimization for data transfers and makespan while considering heterogeneity. Using simulations that vary workloads, platforms, and heterogeneity levels, we study the system utilization, the trade-offs between the targets, and the impacts of considering platforms' heterogeneity. We perform comparisons with a baseline inspired by a Kubernetes-based policy, representing greedy algorithms. Our experiments show considerable gaps between the efficiency of a greedy-based scheduling policy and a multi-objective-based one. The last outperforms the baseline by reducing makespan, data transfers, and system utilization by up to two orders of magnitudes in relevant cases for the edge-cloud continuum [8].

8.1.2 Primal-Dual Algorithms with Predictions for Online Bounded Allocation and Ad-Auctions Problems

Designing online algorithms with predictions is a recent technique for various practically relevant online problems (scheduling, caching (paging), clustering, ski rental, etc.). The unified approach through a primal-dual framework for linear covering problems extends the online primal-dual method by incorporating predictions, and its performance goes beyond the worst-case analysis. Following this research line, we developped competitive algorithms with predictions for non-linear covering problems, generalizing the previous technique. We illustrate the applicability of our algorithms through experiments on energy minimization, congestion management, and submodular minimization problems [20].

8.1.3 Optimization Metrics for the Evaluation of Batch Schedulers in HPC

Machine Learning techniques are taking a prominent position in the design of system softwares. In HPC, many work are proposing to use such techniques (specifically Reinforcement Learning) to improve the performance of batch schedulers. Their main limitation is the lack of transparency of their decision. This underlines the importance of choosing correctly the optimization criteria when evaluating these solutions. In this work, we discuss bias and limitations of the most frequent optimization metrics in the literature. We provide elements on how to evaluate performance when studying HPC batch scheduling. We also propose a new metric: the standard deviation of the utilization, which we believe can be used when the utilization reaches its limits. We then experimentally evaluate these limitations by focusing on the use-case of runtime estimates. One of the information that HPC batch schedulers use to schedule jobs on the available resources is user runtime estimates: an estimation provide by the user of how long their job will run on the machine. These estimates are known to be inaccurate, hence many work have focused on improving runtime prediction [10].

8.2 Energy and Environmental Impact

Participants: Olivier Richard, Christophe Cerin, Denis Trystram, Fanny Dufossé, Kim Thang Nguyen.

8.2.1 Comparing Software-based Power Meters

The global energy demand for digital activities is constantly growing. Computing nodes and cloud services are at the heart of these activities. Understanding their energy consumption is an important step towards reducing it. On one hand, physical power meters are very accurate in measuring energy but they are expensive, difficult to deploy on a large scale, and are not able to provide measurements at the service level. On the other hand, power models and vendor-specific internal interfaces are already available or can be implemented on existing systems. Plenty of tools, called software-based power meters, have been developed around the concepts of power models and internal interfaces, in order to report the power consumption at levels ranging from the whole computing node to applications and services. However, we have found that it can be difficult to choose the right tool for a specific need. In this work, we qualitatively and experimentally compare several software-based power meters able to deal with CPU or GPU-based infrastructures. For this purpose, we evaluate them against high-precision physical power meters while executing various intensive workloads. We extend this empirical study to highlight the strengths and limitations of each software-based power meter [16].

8.2.2 Optimal Sizing of a Globally Distributed Low Carbon Cloud Federation

The carbon footprint of IT technologies has been a significant concern in recent years. This concern mainly focuses on the electricity consumption of data centers; many cloud suppliers commit to using 100% of renewable energy sources. However, this approach neglects the impact of device manufacturing. We consider the question of dimensioning the renewable energy sources of a geographically distributed cloud with considering the carbon impact of both the grid electricity consumption in the considered locations and the manufacturing of solar panels and batteries. We design a linear program to optimize cloud dimensioning over one year, considering worldwide locations for data centers, real-life workload traces, and solar irradiation values. Our results show a carbon footprint reduction of about 30% compared to a cloud fully supplied by solar energy and of 85% compared to the 100% grid electricity model [19].

8.2.3 A Methodology and a Toolbox to Explore Dataset related to the Environmental Impact of HTTP Requests

EcoIndex has been proposed to evaluate the absolute environmental performance of a given URL using a score ranging from 0 to 100 (the higher, the better). We make a critical analysis of the initial approach and propose alternatives that no longer calculate a plain score but allow the query to be situated among other queries. The generalized critiques come with statistics and rely on extensive experiments (first contribution). Then, we move on to low-cost Machine Learning (ML) approaches (second contribution) and a transition before obtaining our final results (third contribution). Our research aims to extend the initial idea of analytical computation, i.e., a relation between three variables, in the direction of algorithmic ML computations. The fourth contribution corresponds to a discussion on our implementation, available on a GitHub repository. We also examine the question: What attributes make sense for our problem?, or equivalently, what is a relevant data policy for studying digital environmental impacts? Beyond computational questions, it is important for the scientific community to focus on this question in particular. We currently promote using well established ML techniques because of their potential. However, we also question techniques for their frugality or otherwise. Our data science project is still at the data exploration stage. We also want to encourage synergy between technical expertise and business knowledge because this is fundamental for advancing the data project [12].

8.2.4 Social and Environmental Effects of Post-COVID-19 Computer Science Virtual Conferencing: The Euro-Par Case

Conferencing is one of the main pillars of Computer Science research activity regarding career and networking, with conference publications playing a more pronounced role compared to other disciplines. The COVID-19 pandemic forced us to switch to virtual conferencing, and many works have shown the benefits of virtual conferencing in terms of inclusivity and reduction of Green House Gas emissions. We are moving toward the usual conferencing format as it appears that the pandemic is increasingly under control. However, the changes imposed during the period of the pandemic brought many essential lessons regarding conferencing social and environmental effects. A crucial task is to gather these community experiences to give directions on how to keep the learned lessons post-COVID-19. We used the Euro-Par conference to combine these lessons in the Computer Science case. We show practical results that reinforce the marginal emissions of virtual conferencing compared to in-person conference travel. We also open the debate that rethinking the conference utility according to our objectives (scientific and ecologic) and being aware of social/geographical biases are essential factors in participating and organizing post-COVID-19 conferences [11].

8.3 Workflows and Data Efficient Programming

Participants: Bruno Raffin, Frederic Wagner, Yves Denneulin.

8.3.1 Taenite : Seamless Persistant Memory Management with Rust

We introduce Taenite [21], a transactional programming library designed for persistent memory. Utilizing the copy-on-write principle, it enables us to circumvent the necessity for logs and, simultaneously, affirms the atomicity of all modifications at the conclusion of a transaction. Data consistency is assured at all times, and recovery after failures is instantaneous. Its programming interface, closely resembling that of the Rust standard library, facilitates simple and transparent usage. We elaborate on its operational principle and, in particular, its memory allocator. Starting from elementary components, we establish the construction of intricate persistent structures such as tree structures, for example, hitchhiking trees or DAGs.

8.3.2 Deisa: In Transit Data Analytics with Dask

We are working with CEA (PhD of Amal Gueroudji) to enable in situ processing for the Dask distributed task programming environment. We developped a hybrid model called DEISA, that supports coupling MPI parallel codes with analyses written using Dask. This implementation requires minimal modifications of both the simulation and analysis codes compared to their post hoc counterpart, while giving access to parallel versions of Numpy, Pandas and scikit-learn [15, 27].

8.3.3 Deep Surrogate Training

Recent years have seen a surge in deep learning approaches to accelerate numerical solvers, which provide faithful but computationally intensive simulations of the physical world. These deep surrogates are generally trained in a supervised manner from limited amounts of data slowly generated by the same solver they intend to accelerate. We extend the Melissa framewrok to enable the online training of these models from a large ensemble run of simulations. It leverages multiple levels of parallelism to generate rich datasets. The framework avoids I/O bottlenecks and storage issues by directly streaming the generated data. A training reservoir mitigates the inherent bias of streaming while maximizing GPU throughput. Experiment on training a fully connected network as a surrogate for the heat equation shows the proposed approach enables training on 8TB of data in 2 hours with an accuracy improved by 47% and a batch throughput multiplied by 13 compared to a traditional offline procedure [17, 7]. Other experiments with various neural architectures indicate that more dataset diversity, up to hundreds of GB, can increase deep surrogate generalization capabilities. Fully connected neural networks, Fourier Neural Operator (FNO), and Message Passing PDE Solver prediction accuracy is improved by 68%, 16% and 7%, respectively [18].

Following up on this approach of online deep surrogate training, we are also investigating active learning methods for training neural networks from synthetic input samples that can be generated on-demand. This includes Physics Informed Neural Networks (PINNs), simulation-based inference, deep surrogates and deep reinforcement learning. An adaptive process observes the training progress and steers the data generation with the goal of speeding up and increasing the quality of training. We proposed a novel adaptive sampling method that concentrates samples close to the areas showing high loss values. Compared to the state-ofthe-art R3 sampling our algorithm converges to a validation loss of 0.5 in 6000 iterations, while it takes 25000 iterations to reach a loss of 0.7 for the R3 algorithm when training a PINN with the Allen Cahn equation [14].

9 Bilateral contracts and grants with industry

Participants: Bruno Raffin, Denis Trystram, Olivier Richard.

9.1 Bilateral grant with industry

- **EDF R&D (2020-2023)**. PhD grant (Lucas Meyer). 160K euros.
- **Ryax Technologies (2020-2023)**. PhD grant (Anderson Andrei Da Silva). 170K euros.
- **Berger-Levrault (2022-2025)**. PhD grant (Halmza Safri). 170K euros
- **ATOS (2022-2026)**. PhD grants (Abdessalam Benharii and Guillaume Raffin). 340K euros
- **Orange (2023-2026)**. Phd grant (Yoann Dupas). 170K euros.

10 Partnerships and cooperations

Participants: Denis Trystram, Fanny Dufossé, Gregory Mounié, Pierre-François Dutot, Kim Thang Nguyen, Olivier Richard, Christophe Cerin, Denis Trystram, Yves Denneulin.

10.1 European initiatives

10.1.1 LIGHTAIGE

Program: SKŁODOWSKA-CURIE ACTIONS - Individual Fellowship

Duration: November 2023 - December 2023

Followship Recipient: Danilo Carastan-Santos

Abstract: The goal of the LIGHTAIGE project is to approach the orchestration and simulation problems of Edge Intelligence with crossdiscipline, lightweight ang transparent methods to minimize the GHG emissions – measured as the amount of equivalent CO2 emissions – of using Edge Intelligence.

10.1.2 REGALE

Title: REGALE (An open architecture to equip next generation HPC applications with exascale capabilities)

Program: EuroHPC Horizon 2020 research and innovation program

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

Partners:

- Institut National de Recherche en Informatique et Automatique (INRIA), France
- Institut Polytechnique de Grenoble (INP Grenoble), France
- Ryax Technologies (Ryax Technologies), France
- Électricité de France (EDF), France
- Ethnicon Metsovion Polytechnion (National Technical Univesity of Athens - NTUA), Greece
- Technische Universitaet Muenchen (TUM), Germany

- Gioumpitek Meleti Schediasmos Ylopoiisi Kai Polisi Ergon Pliroforikis Etaireia Periorismenis Efthynis (Ubitech Design Planning Implementation and Sale of Information Works), Greece
- Bayerische Akademie Der Wissenschaften (BADW), Germany
- Université Grenoble Alpes (UGA), France
- E 4 Computer Engineering SPA (E4), Italy
- BULL SAS (BULL), France
- TWT GMBH Science and Innovation (TWT GMBH Science and Innovation), Germany
- Alma Mater Studiorum - Universita di Bologna (UNIBO), Italy
- SCIO IKE (SCIO P.C.), Greece
- Cineca Consorzio Interuniversitario (CINECA), Italy
- Erevnitiko Panepistimiako Institouto Systimaton Epikoinonion Kai Ypologiston (Research University Institute of Communication and Computer Systems), Greece
- Andritz Hydro GMBH, Austria
- Barcelona Supercomputing Center, Centro Nacional de Supercomputacion (BSC CNS), Spain

Coordinator for UGA/INRIA: Pierre-François Dutot

Datamove Budget: 680K euros.

Summary: With exascale systems almost outside our door, we need now to turn our attention on how to make the most out of these large investments towards societal prosperity and economic growth. REGALE aspires to pave the way of next-generation HPC applications to exascale systems. To accomplish this we define an open architecture, build a prototype system and incorporate in this system appropriate sophistication in order to equip supercomputing systems with the mechanisms and policies for effective resource utilization and execution of complex applications.

REGALE brings together leading supercomputing stakeholders, prestigeous academics, top European supercomputing centers and end users from five critical target sectors, covering the entire value chain in system software and applications for extreme scale technologies.

10.2 National initiatives

10.2.1 PEPR NUMPEX

Goals: The main objective of the NumPEx (Numeric for Exascale) program in France is to develop state-of-the-art skills and infrastructures in the field of exascale computing.

Duration: From 2023 to 2030

Web site: NUMPEX

Datamove implication:

- Exa-DoST (Data-oriented Software and Tools for the Exascale): Co-lead WP3.
- Exa-AToW (Architectures and Tools for Large-Scale Workflows): Co-lead WP5.
- Exa-DI (Development and integration): CO-lead WP3.

Datamove budget: 1.295 M euros.

10.2.2 BPI

- **Projet AMI Cloud OTPaaS (2021-2024)**. Aims at offering a new Cloud offer, compatible with Gaia-X and easy to use, that could favour the massive digital transition of companies. Datamove Budget: 110 Keuro.

10.2.3 ANR

- **PPR Océan et Climat MEDIATION (2022-2030)**. Methodological developments for a robust and efficient digital twin of the ocean. Pi: INRIA team AIRSEA. Partners: INRIA, CNRS, IFREMER, IRD, Université Aix-Marseille, Institut National Polytechnique de Toulouse, Ecole Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, Service Hygrodgraphique et Océanographique de la Marine, Université Grenoble Alpes, Météo-France-DESR-Centre National de Recherches Météorologiques. Total budget: 2,4 Meuros. Datamove Budget: 110 Keuros. CO-lead of the WP Leveraging AI and HPC for Digital Twins of the Ocean.

10.2.4 Univ. Grenoble Alpes

- Edge Intelligence chair of the Institute of Artificial Intelligence of Univ. Grenoble Alpes (MIA@Grenoble-Alpes) (2019-2023). PI: Denis Trystram. The challenges are to design new machine learning methods that fully exploit the distributed character of the edge and to develop algorithms and subsequent pieces of software that will allow the deployment of the edge/fog hybrid infrastructures. The research agenda is two-fold. In the first hand, we study new methods for distributed machine learning and data analytic. In the second hand, we develop the models and mechanisms for the orchestration of efficient local resource management. Budget: 335K euro
- **IRS SoSCloud.** Dimensioning of green energy in Clouds, 2020-2023. UGA Grant. PhD funding. Co-advised by D. Cordeiro, USP, Brasil. Budget : 120 Keuros

11 Dissemination

Participants: Denis Trystram, Fanny Dufossé, Gregory Mounié, Pierre-François Dutot, Kim Thang Nguyen, Olivier Richard, Christophe Cerin, Denis Trystram, Yves Denneulin.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

- Organization of the second Journées de Recherche en Apprentissage Frugal at Grenoble.
- Organizing committee of ISAV 2023. In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization. Workshop from SC23.

11.1.2 Scientific events: selection

Member of the conference program committees

- LDAV 2023.
- SC23 Workshops.

11.2 Scientific expertise

- Expertise of research project proposals for DOE,USA.

11.3 PhD Juries

- President of the PhD Jury of Nicolas Grenèche. Élasticité des infrastructures HPC conteneurisées : résonance entre le HPC et le Cloud. Novembre 2023.
- President of the PhD Jury of Amal Gueroudji. Distributed Task-Based In Situ Data Analytics for High-Performance Simulations. May 2023

- Reviewer and Jury Membrer for the PhD of Romain Peressoni. Large Scale Multidimensional Scaling for the study of biodiversity. June 2023.

11.4 Popularization

11.4.1 Articles and contents

- EcoIndex : que vaut cet outil qui mesure le score environnemental des sites web ? The Conversation. May 2023.
- « L'envers des mots » : Exascale. The Conversation. Nov. 2023.
- Les défis d'une IA frugale. Journal du CNRS. Nov. 2023.

11.4.2 Interventions

- *La RO face au réchauffement climatique : opportunité ou menace ?*. Congrès ROADEF, Rennes – 22/02/2023.
- *Paver la route pour une RO (plus) verte.*Keynote Green Days, Lyon – 28/03/2023.
- *Why current Batch Schedulers are not sustainable.* Workshop Chicago, Paris – 29/03/2023.
- *Impacts écologiques de l'IA.* Journée GdR RADIA IA et sobriété, Paris – 4/04/2023.
- *Transformations numériques éco-responsables.*Webminaire UGA, Grenoble – 6/04/2023.
- *L'apprentissage automatique opportunit 'e ou frein pour sortir de la crise climatique?* Séminaire FEMPTO, Besançon – 4/05/2023.
- *Doit-on croire dans le numérique pour sortir de la crise climatique ?* Séminaire LIP6, Paris – 26/05/2023
- *Quizz Eco Info.* Program off ICT4S, Rennes – 6/06/2023
- *Environmental issues of AI.* Course EFELIA, Grenoble –11/07/2023
- *Ecological impacts of AI.* Summer School RSD on distributed learning, Lyon – 19/09/2023
- *Atelier médiation.* Journée Eco-Info, Saint Malo – 10/10/2023
- *Enjeux écologiques de l'IA – Un état des lieux.* Journées thématiques nationales CUME, Paris – 22/11/2023

12 Scientific production

12.1 Major publications

[1] D. Carastan-Santos and R. Y. de Camargo. 'Obtaining Dynamic Scheduling Policies with Simulation and Machine Learning'. In: SC'17 -2 International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing). Denver, United States, 12th Nov. 2017. URL: https://inria.hal.science/hal-01618940.

[2] P.-F. Dutot, M. Mercier, M. Poquet and O. Richard. 'Batsim: a Realistic Language-Independent Resources and Jobs Management Systems Simulator'. In: *20th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP)*. 20th Workshop on Job Scheduling Strategies for Parallel Processing. Chicago, United States, 27th May 2016. URL: https://hal.archives-ouvertes.fr/hal-01333471.

[3] S. Friedemann and B. Raffin. 'An elastic framework for ensemble-based large-scale data assimilation'. In: *The international journal of high performance computing applications* 36 (28th June 2022), pp. 1–37. DOI: 10.1177/10943420221110507. URL: https://hal.inria.fr/hal-03017033.

[4] G. Lucarelli, B. Moseley, N. K. Thang, A. Srivastav and D. Trystram. 'Online Non-preemptive Scheduling on Unrelated Machines with Rejections'. In: SPAA 2018 - 30th ACM Symposium on Parallelism in Algorithms and Architectures. Vienna, Austria: ACM Press, 2018, pp. 291–300. DOI: 10.1145/3210377.3210402. URL: https://hal.archives-ouvertes.fr/hal-01986312.

[5] L. Meyer, M. Schouler, R. A. Caulk, A. Ribés and B. Raffin. 'High Throughput Training of Deep Surrogates from Large Ensemble Runs'. In: *SC '23: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. SC 2023 - The International Conference for High Performance Computing, Networking, Storage, and Analysis. Denver, CO, United States: ACM, 18th Nov. 2023, pp. 1–14. DOI: 10.1145/3581784.3607083. URL: https://hal.science/hal-04213978.

[6] M. F. Silva Vasconcelos, D. Cordeiro, G. da Costa, F. Dufossé, J.-M. Nicod and V. Rehn-Sonigo. 'Optimal sizing of a globally distributed low carbon cloud federation'. In: CCGrid 2023 - IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing. Bangalore, India, 2023, pp. 1–13. DOI: 10.1109/CCGrid57682.2023.00028. URL: https://hal.science/hal-04032094.

12.2 Publications of the year

International journals

[7] M. Schouler, R. A. Caulk, L. Meyer, T. Terraz, C. Conrads, S. Friedemann, A. Agarwal, J. M. Baldonado, B. Pogodziński, A. Sekuła, A. Ribes and B. Raffin. 'Melissa: coordinating large-scale ensemble runs for deep learning and sensitivity analyses'. In: *Journal of Open Source Software* 8.86 (June 2023), p. 5291. DOI: 10.21105/joss.05291. URL: https://inria.hal.science/hal-04145897.

International peer-reviewed conferences

[8] L. Angelelli, A. A. da Silva, Y. Georgiou, M. Mercier, G. Mounié and D. Trystram. 'Towards a Multi-objective Scheduling Policy for Serverless-based Edge-Cloud Continuum'. In: CCGrid 2023 - 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing. Bangalore, India: IEEE, 1st May 2023, pp. 485–497. DOI: 10.1109/CCGrid57682.2023.00052. URL: https://hal.science/hal-04177085.

[9] A. Berthelot, E. Caron, M. Jay and L. Lefèvre. 'Estimating the environmental impact of Generative-AI services using an LCA-based methodology'. In: *Procedia CIRP*. CIRP LCE 2024 - 31st Conference on Life Cycle Engineering. Turin, Italy, 2024, pp. 1–10. URL: https://inria.hal.science/hal-04346102.

[10] R. Boëzennec, F. Dufossé and G. Pallez. 'Optimization Metrics for the Evaluation of Batch Schedulers in HPC'. In: JSSPP 2023 - 26th edition of the workshop on Job Scheduling Strategies for Parallel Processing. St. Petersburg, Florida, United States, 23rd Mar. 2023, pp. 1–19. URL: https://inria.hal.science/hal-04042591.

[11] D. Carastan-Santos, K. Rzadca, L. Sousa and D. Trystram. 'Social and environmental effects of post-COVID-19 Computer Science virtual conferencing: The Euro-Par case'. In: 2023 International Conference on ICT for Sustainability (ICT4S). Rennes, France: IEEE, 5th June 2023, pp. 132–141. DOI: 10.1109/ICT4S58814.2023.00022. URL: https://hal.science/hal-03903632.

[12] C. Cérin, M. Jay, L. Lefèvre and D. Trystram. 'A Methodology and a Toolbox to Explore Dataset related to the Environmental Impact of HTTP Requests'. In: 2023 IEEE International Conference on Big Data (BigData) - 3rd International Workshop on Big Data Analytics for Sustainability. Sorrento (Naples), Italy: IEEE, 2024, pp. 1–10. DOI: 10.1109/BigData59044.2023.10386275. URL: https://inria.hal.science/hal-04386964.

[13] P.-F. Dutot, Y.-S. Fu, N. Prasad and O. Sinnen. 'A Guaranteed Approximation Algorithm for Scheduling Fork-Joins with Communication Delay'. In: *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS). St. Petersburg, United States: IEEE, 2023, pp. 820–830. DOI: 10.1109/IPDPS54959.2023.00087. URL: https://hal.science/hal-04418074.

[14] S. Dymchenko and B. Raffin. 'Loss-driven sampling within hard-to-learn areas for simulation-based neural network training'. In: MLPS 2023 - Machine Learning and the Physical Sciences Workshop at NeurIPS 2023 - 37th conference on Neural Information Processing Systems. New Orleans, United States, 2023, pp. 1–5. URL: https://hal.science/hal-04305233.

[15] A. Gueroudji, J. Bigot, B. Raffin and R. Ross. 'Dask-Extended External Tasks for HPC/ML In Transit Workflows'. In: SC-W 2023 - Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis. Denver, United States: ACM, 12th Nov. 2023, pp. 831–838. DOI: 10.1145/3624062.3624151. URL: https://hal.science/hal-04409157.

[16] M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie and B. Fichel. 'An experimental comparison of software-based power meters: focus on CPU and GPU'. In: CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing. Bangalore, India: IEEE, 2023, pp. 1–13. DOI: 10.1109/CCGrid57682.2023.00020. URL: https://inria.hal.science/hal-04030223.

[17] L. Meyer, M. Schouler, R. A. Caulk, A. Ribés and B. Raffin. 'High Throughput Training of Deep Surrogates from Large Ensemble Runs'. In: *SC '23: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. SC 2023 - The International Conference for High Performance Computing, Networking, Storage, and Analysis. Denver, CO, United States: ACM, 18th Nov. 2023, pp. 1–14. DOI: 10.1145/3581784.3607083. URL: https://hal.science/hal-04213978.

[18] L. Meyer, M. Schouler, R. A. Caulk, A. Ribés and B. Raffin. 'Training Deep Surrogate Models with Large Scale Online Learning'. In: *Proceedings of Machine Learning Research*. International Conference on Machine Learning. Vol. 202. Honolulu (Hawai'i), United States, 29th July 2023, pp. 24614–24630. URL: https://hal.science/hal-04102400.

[19] M. F. Silva Vasconcelos, D. Cordeiro, G. da Costa, F. Dufossé, J.-M. Nicod and V. Rehn-Sonigo. 'Optimal sizing of a globally distributed low carbon cloud federation'. In: CCGrid 2023 - IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing. Bangalore, India, 2023, pp. 1–13. DOI: 10.1109/CCGrid57682.2023.00028. URL: https://hal.science/hal-04032094.

[20] N. K. Thắng and E. Kevi. 'Primal-Dual Algorithms with Predictions for Online Bounded Allocation and Ad-Auctions Problems'. In: ALT 2023 - 34th International Conference on Algorithmic Learning Theory. Singapore, Singapore, 2023, pp. 1–18. URL: https://hal.science/hal-03997203.

National peer-reviewed Conferences

[21] L. Boulanger, F. Wagner and Y. Denneulin. 'Taenite : Gestion transparente de la mémoire persistante en Rust'. In: COMPAS 2023 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Annecy, France, 2023, pp. 1–7. URL: https://hal.science/hal-04299508.

[22] E. Foussard, G. Bridonneau, M.-L. Espinouse, G. Mounié and M. Nattaf. 'Génération de colonnes pour le Bin-Packing avec seuils'. In: ROADEF 2023 - 24ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Rennes, France, 20th Feb. 2023. URL: https://hal.science/hal-04009192.

[23] Q. Guilloteau, A. Faure, M. Poquet and O. Richard. 'Comment rater la reproductibilité de ses expériences ?' In: ComPAS 2023 Conférence francophone en informatique. 1-9. Annecy, France, 4th July 2023, à paraître. URL: https://hal.science/hal-04132438.

Conferences without proceedings

[24] P.-F. Dutot. 'REGALE project: Oar, A Versatile Resource and JobManagement System'. In: CONCERTO 2024 - 2nd workshop on projeCts crOss-synergy iN advanCing Exascale platfoRms and quanTum cOmputing. Munich, Germany, 2024. URL: https://hal.science/hal-04418094.

[25] L. Rosa, D. Carastan-Santos and A. Goldman. ‘An experimental analysis of regression-obtained HPC scheduling heuristics’. In: Job Scheduling Strategies for Parallel Processing. Vol. 14283. Lecture Notes in Computer Science. St. Petersburg, United States: Springer Nature Switzerland, 15th Sept. 2023, pp. 116–136. DOI: 10.1007/978-3-031-43943-8_6. URL: https://inria.hal.science/hal-03979237.

Doctoral dissertations and habilitation theses

[26] V. Fagnon. ‘Prediction and resource augmentation come to the rescue of algorithmics.’ Université Grenoble Alpes [2020-....], 6th July 2023. URL: https://theses.hal.science/tel-04279180.

[27] A. Gueroudji. ‘Distributed Task-Based In Situ Data Analytics for High-Performance Simulations’. Université Grenoble Alpes [2020-....], 26th May 2023. URL: https://theses.hal.science/tel-04194958.

[28] Q. Guilloteau. ‘Control-based runtime management of HPC systems with support for reproducible experiments’. Université Grenoble Alpes, 11th Dec. 2023. URL: https://hal.science/tel-04389290.

Reports & preprints

[29] R. Boëzennec, F. Dufossé and G. Pallez. *Analyzing Qualitatively Optimization Objectives in the Design of HPC Resource Manager*. 21st Aug. 2023. URL: https://hal.science/hal-04187517.

[30] S. Friedemann, K. Keller, Y.-S. Lu, B. Raffin and L. Bautista Gomez. *Dynamic Load/Propagate/Store for Data Assimilation with Particle Filters on Supercomputers*. 2024. DOI: 10.1016/j.jocs.2024.102229. URL: https://inria.hal.science/hal-03927612.

[31] Q. Guilloteau. *Simulating a Multi-Layered Grid Middleware*. 19th May 2023. URL: https://hal.science/hal-04101015.

[32] Q. Guilloteau, O. Richard, R. Bleuse and E. Rutten. *Folding a Cluster containing a Distributed File-System*. 2023. URL: https://hal.science/hal-04038000.

[33] E. Kevi and N. K. Thắng. *Online Covering with Multiple Experts*. 21st Nov. 2023. URL: https://hal.science/hal-04297794.

[34] E. Kevi and N. K. Thắng. *Online Primal-Dual Algorithm with Predictions for Non-Linear Covering Problems*. 22nd Dec. 2023. URL: https://hal.science/hal-04361128.

[35] L. Lefèvre, A.-L. Ligozat, D. Trystram, S. Bouveret, A. Bugeau, J. Combaz, E. Frenoux, G. Guennebaud, J. Lefèvre, J.-P. Nicolaï and K. Dassas. *Environmental assessment of projects involving AI methods*. 4th Jan. 2023. URL: https://hal.science/hal-03922093.

[36] G. Raffin and D. Trystram. *Dissecting the software-based measurement of CPU energy consumption: a comparative analysis*. 26th Jan. 2024. URL: https://hal.science/hal-04420527.

[37] A. Shilova, T. Delliaux, P. Preux and B. Raffin. *Learning HJB Viscosity Solutions with PINNs for Continuous-Time Reinforcement Learning*. RR-9541. Inria Lille - Nord Europe, CRIStAL - Centre de Recherche en Informatique, Signal et Automatique de Lille - UMR 9189; Univ. Lille, CNRS, Centrale Lille, Inria UMR 9189 - CRIStAL,INRIA Lille Nord Europe, Villeneuve d’Ascq, France; Univ. Grenoble Alps, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France, 7th Feb. 2024, pp. 1–30. URL: https://inria.hal.science/hal-04445160.

Other scientific publications

[38] S. Brun. ‘Étude du phénomène Stat-Storm - Limitation des appels systèmes pour les systèmes de fichiers distribués de type store’. Université Grenoble Alpes, 1st Sept. 2023. URL: https://inria.hal.science/hal-04197724.

[39] A. A. Da Silva, Y. Georgiou, M. Mercier, G. Mounié and D. Trystram. *Towards an Energy-Aware Multi-objective Scheduling Policy for Server-based Edge-Cloud Continuum*. Annecy, France, 2023. URL: https://inria.hal.science/hal-04244625.

[40] A. A. Da Silva, Y. Georgiou, M. Mercier, G. Mounié and D. Trystram. *Towards Container-layer-aware Scheduling Policies for Serverless-based Edge-Cloud Continuum*. Brno, Czech Republic, June 2023. URL: https://inria.hal.science/hal-04244617.

[41] S. Delamare, D. Margery, P. Neyron and L. Nussbaum. *Évolution du matériel et des services logiciels disponibles dans Grid'5000*. 10th May 2023. URL: https://inria.hal.science/hal-04098050.

[42] S. Dymchenko and B. Raffin. 'Loss-driven sampling for online neural network training with large scale simulations'. In: LIG PhD day 2023. Grenobe, France, 2023, pp. 1–1. URL: https://hal.science/hal-04305159.

[43] A. Lithaud. 'Contribution au projet NixOS Compose'. Université Grenoble Alpes, 1st Sept. 2023, pp. 1–39. URL: https://inria.hal.science/hal-04197720.