RESEARCH CENTRE

Inria Center
at Université Grenoble Alpes

IN PARTNERSHIP WITH:

Université de Grenoble Alpes, CNRS

2022

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

Inria

Contents

Project-Team DATAMOVE 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3

3 Research program 4
3.1 Motivation . . . 4
3.2 Strategy . . . 4
3.3 Research Directions . . . 5

4 Application domains 6
4.1 Data Aware Batch Scheduling . . . 6
4.1.1 Algorithms . . . 6
4.1.2 Locality Aware Allocations . . . 7
4.1.3 Data-Centric Processing . . . 7
4.1.4 Learning . . . 8
4.1.5 Multi-objective Optimization . . . 8
4.2 Empirical Studies of Large Scale Platforms . . . 9
4.2.1 Workload Traces with Resource Consumption . . . 9
4.2.2 Simulation . . . 10
4.2.3 Job and Platform Models . . . 10
4.2.4 Emulation and Reproducibility . . . 11
4.3 Integration of High Performance Computing and Data Analytics . . . 11
4.3.1 Programming Model and Software Architecture . . . 11
4.3.2 Resource Sharing . . . 12
4.3.3 Co-Design with Data Scientists . . . 13

5 Social and environmental responsibility 13

6 Highlights of the year 14
6.1 Awards . . . 14

7 New software and platforms 14
7.1 New software . . . 14
7.1.1 FlowVR . . . 14
7.1.2 OAR . . . 14
7.1.3 MELISSA . . . 15
7.1.4 Batsim . . . 15
7.2 New platforms . . . 16
7.2.1 SILECS/Grid'5000 and Meso Center Ciment . . . 16

8 New results 16
8.1 Data Aware Batch Scheduling . . . 16
8.1.1 Energy Saving and sustainability of information technologies . . . 16
8.1.2 On-line Scheduling Using Resource Augmentation . . . 16
8.1.3 Scheduling for Edge Infrastructures . . . 17
8.1.4 Gradient algorithms for online decentralized optimization . . . 17
8.2 Empirical Studies of Large Scale Platforms . . . 17
8.2.1 A Methodology to Scale Containerized HPC Infrastructures in the Cloud . . . 17
8.2.2 Model-free control for resource harvesting in computing grids . . . 18
8.2.3 Painless Transposition of Reproducible Distributed Environments with NixOS Compose . . . 18
8.2.4 Reproducibility . . . 18
8.3 Integration of High Performance Computing and Data Analytics . . . 18

8.3.1 Data analysis for a single simulation run. . . . 19
8.3.2 Data analysis for ensemble simulation runs. . . . 19

9 Bilateral contracts and grants with industry 19
9.1 Bilateral grant with industry . . . 20

10 Partnerships and cooperations 20
10.1 International initiatives . . . 20
10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . 20
10.1.2 STIC/MATH/CLIMAT AmSud projects . . . 20
10.2 International research visitors . . . 21
10.2.1 Visits of international scientists . . . 21
10.3 European initiatives . . . 21
10.3.1 H2020 projects . . . 21
10.4 National initiatives . . . 25
10.4.1 BPI . . . 26
10.4.2 ANR . . . 26
10.4.3 INRIA . . . 26
10.4.4 Univ. Grenoble Alpes . . . 26

11 Dissemination 26
11.1 Promoting scientific activities . . . 26
11.1.1 Scientific events: organisation . . . 26
11.1.2 Journal . . . 27
11.1.3 Invited talks . . . 27
11.1.4 Leadership within the scientific community . . . 27
11.2 Teaching - Juries . . . 28
11.2.1 Teaching . . . 28
11.2.2 Juries . . . 28
11.2.3 Internal or external Inria responsibilities . . . 28

12 Scientific production 28
12.1 Major publications . . . 28
12.2 Publications of the year . . . 29
12.3 Other . . . 32

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

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

B3.3.2. – Water: sea & ocean, lake & river

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

- Christoph Cérin [Université de Paris, Professor, Délégation]
- Yves Denneulin [GRENOBLE INP, Professor, HDR]
- Pierre-François Dutot [UGA, Associate Professor]
- Grégory Mounié [GRENOBLE INP, Associate Professor]
- Kim Thang Nguyen [GRENOBLE INP, Professor, from Sep 2022]
- 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]
- Marc Schouler [UGA]
- Alena Shilova [INRIA]

PhD Students

- Luc Angelelli [UGA]
- Abdessalam Benhari [ATOS, CIFRE, from May 2022]
- Louis Boulanger [UGA]
- Louis Closson [BERGER-LEVRAULT, CIFRE]
- Anderson Da Silva [RYAX, CIFRE]
- Yoann Dupas [Orange, CIFRE, from Nov 2022]
- Sofya Dymchenko [INRIA, from Mar 2022]
- Vincent Fagnon [UGA]
- Ernest Foussard [UGA]
- Amal Gueroudji [CEA]
- Quentin Guilloteau [Inria]
- Mathilde Jay [UGA]
- Eniko Kevi [UGA]
- Yannick Malot [CEA, from Oct 2022]

- Lucas Meyer [EDF, CIFRE]
- Tuan Anh Nguyen Huu [UGA]
- Hamza Safri [BERGER-LEVRAULT, CIFRE, from Mar 2022]
- Miguel Silva Vasconcelos [UGA]
- Paul Youssef [UGA]

Technical Staff

- Jonathan Bleuzen [UGA, Engineer]
- Robert Caulk [INRIA, Engineer, from Sep 2022]
- Adrien Faure [UGA, Engineer, until Sep 2022]
- Millian Poquet [UGA, Engineer]

Administrative Assistant

- Annie Simon [INRIA]

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 [2]. 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.

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.

[1] 10^{18} floating point operations per second
[2] SciDAC Review 2010

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 performance 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.

Socially we are concerned that management issues are affecting INRIA's work environment and usual enthusiasm of being part of this research institute.

6 Highlights of the year

6.1 Awards

- Euro-Par 2022 Artefact Prize. Nicolas Grenèche, Tarek Menouer, Christophe Cérin, Olivier Richard. A Methodology to Scale Containerized HPC Infrastructures in the Cloud. Europar 2022, Aug 2022, Glasgow, United Kingdom.
- ATOSD/GENCI Joseph Fourier Price 2022, category AI. Danilo Carastan-Santos and Denis Trystram.

7 New software and platforms

7.1 New software

7.1.1 FlowVR

Scientific Description: FlowVR adopts the "data-flow" paradigm, where your application is divided as a set of components exchanging messages (think of it as a directed graph). FlowVR enables to encapsulate existing codes in components, interconnect them through data channels, and deploy them on distributed computing resources. FlowVR takes care of all the heavy lifting such as application deployment and message exchange.

The base entity, called a module or component, is an autonomous process, potentially multi-threaded with tools like OpenMP, TBB, or deferring computations to a GPU or Xeon Phi. This module processes data coming from input ports and write data on output ports. A module has no global insight on where the data comes from or goes to. The programming interface is designed to limit code refactoring, easing turning an existing code into a FlowVR component. The three main functions are:

wait(): Blocking function call that waits for the availability of new messages on input ports. get(): Retrieve a handle to access the message received at the previous wait() call on a given input port. put(): Notify FlowVR that a new message on a given output port is ready for dispatch. FlowVR manages data transfers. Intra-node communications between two components take place through a shared memory segment, avoiding copies. Once the sender has prepared the data in a shared memory segment, it simply handles a pointer to the destination that can directly access them. Inter-node communications extend this mechanism, FlowVR taking care of packing and transferring the data from the source shared memory segment to the destination shared memory segment.

Assembling components to build an application consists in writing a Python script, instanciate it according to the target machine. FlowVR will process it and prepare everything so that in one command line you can deploy and start your application.

Functional Description: FlowVR adopts the "data-flow" paradigm, where your application is divided as a set of components exchanging messages (think of it as a directed graph). FlowVR enables to encapsulate existing codes in components, interconnect them through data channels, and deploy them on distributed computing resources. FlowVR takes care of all the heavy lifting such as application deployment and message exchange.

URL: http://flowvr.sf.net

Contact: Bruno Raffin

Participants: Bruno Raffin, Clément Ménier, Emmanuel Melin, Jean Denis Lesage, Jérémie Allard, Jérémy Jaussaud, Matthieu Dreher, Sébastien Limet, Sophie Robert, Valérie Gourantou

7.1.2 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

7.1.3 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

News of the Year: Melissa has been extended for supporting large scale data assimilation (EnkF and Particle Filters), as well as online training of deep surrogate models

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

Publications: hal-01383860, hal-01607479, hal-03017033, hal-03927612, hal-03842106

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

Contact: Bruno Raffin

Partner: Edf

7.1.4 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 Data Aware Batch Scheduling

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

8.1.1 Energy Saving and sustainability of information technologies

The question of energy consumption of computing platforms has become a major topic for DataMove team in the last years.

The energy consumption of computing platforms is one of the major concerns for sustainability of information technology. A new study [14] has evaluated the energy consumption of numerous Deep Learning models considering their accuracy and duration of training. The question of energy consumption measurement of an on-going project [47] that compares the accuracy of different tools.

The question of energy consumption can also be studied considering its footprint. In [35], a scheduling problem is considered on a cloud geo-distributed platform supplied by both the electric grid and a photovoltaic farm. The objective is to allocate the workload to minimize the grid electricity consumption. This requests to predict the future solar irradiation and to avoid traffic congestion at communication network level.

The question of sustainability of computing science research community has been studied by [39] to evaluate the footprint of conferences travel and the impact of virtual conferences based on Euro-Par conference data.

8.1.2 On-line Scheduling Using Resource Augmentation

Resource augmentation is a well-established (and powerful) model for analyzing algorithms, particularly in online setting. It has been successfully used for providing theoretical evidence for several heuristics in scheduling with good performance in practice. According to this model, the algorithm is applied to a more powerful environment than that of the adversary. Several types of resource augmentation for scheduling problems have been proposed up to now, including speed augmentation, machine augmentation and more recently rejection. We presented a framework that unifies the various types of resource augmentation. It allows to generalize the notion of resource augmentation for other types of resources. Our framework is based on Dual Fitting. It consists of extending the domain of feasible solutions for the algorithm with respect to the domain of an adversary. This, in turn, allows the natural concept of duality for mathematical programming to be used as a tool for the analysis of the algorithm performance. As an illustration of the above ideas, we apply this framework, and we propose a primal-dual algorithm for the

online scheduling problem of minimizing the total weighted flow time of jobs on unrelated machines when the preemption of jobs is not allowed.

We derived [40] a collection of results on on-line non-preemptive scheduling algorithms: Targeting the minimization of the weighted flow time on unrelated machines, we considered a version where the online algorithm can reject some $\varepsilon_r > 0$ fraction (by weight) of the jobs and have machines that are $1+\varepsilon_s$ as fast as the offline machines, for some $\varepsilon_s > 0$. We proved that this is already enough to achieve a competitive ratio of $O(1/(\varepsilon_s\varepsilon_r))$. More recently, we continued the study showing that it is sufficient to reject a 2ε fraction of the total number of jobs to achieve a competitive ratio of $2(\frac{1+\varepsilon}{\varepsilon})$. We also considered the speed scaling model, in which machines can be sped up if additional energy is invested, and the goal is to minimize the total weighted flow time plus energy usage. If the power function of machine i is given by $P(s_i(t)) = s_i(t)^\alpha$, where $s_i(t)$ is the current speed of machine i, there is an algorithm which is $O((1+1/\varepsilon)^{\alpha/(\alpha-1)})$-competitive that rejects jobs of total weight at most a fraction ε of the total weight of all the jobs. We also derived a positive result for jobs with hard deadlines, where the objective is to minimize the total energy usage and no job may be rejected. Finally, we closed the story in generalizing these results by showing that rejection alone is sufficient for an algorithm to be competitive even for weighted flow time. They presented an $O(1/\varepsilon^3)$-competitive algorithm that rejects at most $O(\varepsilon)$ of the total weight of the jobs.

8.1.3 Scheduling for Edge Infrastructures

Internet Of Things is a new step in IT intrusion in everyday life. Edge computing uses these components as a new scale of parallel computing platform. Federated learning (FL) is an approach that enables collaborative machine learning (ML) without sharing data over the network. We developed [22] a prototype that enables distributed ML model deployment, federated task orchestration, and monitoring of system state and model performance. The method was applied to predictive maintenance that aims to anticipate industrial equipment failures in order to allow early scheduling of corrective actions.

Another study [26] has considered the operational cost of edge computing platforms. It aims to provide decision aid support to a technical smart buildings manager to potentially reduce the emission of data produced by sensors inside a building and, more generally, to acquire knowledge on the data produced in the facility. The description and the construction of learning models over data sets are crucial in engineering studies to advance critical analysis and serve diverse researchers' communities, such as architects or data scientists.

8.1.4 Gradient algorithms for online decentralized optimization

A new research topic started in 2022 in DataMove, following the arrival of Kim Thang Nguyen. The question of decentralized online optimization problems that is frequent in scheduling environment was study in [11, 21]. The approach consists in a stochastic conditional gradient approach using the Frank-Wolfe algorithm. The resulting method permits a an asymptotically tight regret guarantee of $O(\sqrt{T})$ with T the time horizon. It also allows to address the uncertainty of settings, partial information for agents and low resources for communications and computations. An application was run on a smart building context [20] that permit to enhance the applicability of the method.

8.2 Empirical Studies of Large Scale Platforms

Participants: Olivier Richard, Pierre-François Dutot, Christophe Cerin.

8.2.1 A Methodology to Scale Containerized HPC Infrastructures in the Cloud

We developped a generic method to scale HPC clusters on top of the Kubernetes cloud orchestrator. Users define their targeted infrastructure with the usual Kubernetes syntax for recipes, and our approach automatically translates the description to a full-fledged containerized HPC cluster. Moreover, resource extensions or shrinks are handled, allowing a dynamic resize of the containerized HPC cluster without

disturbing its running. The Kubernetes orchestrator acts as a provisioner. We applied the generic method to three orthogonal architectural designs Open Source HPC schedulers: SLURM, OAR, and OpenPBS. Through a series of experiments, the paper demonstrates the potential of our approach regarding the scalability issues of HPC clusters and the simultaneous deployment of several job schedulers in the same physical infrastructure. It should be noticed that our plan does not require any modification either in the containers orchestrator or in the HPC schedulers. Our proposal is a step forward to reconciling the two ecosystems of HPC and cloud. It also calls for new research directions and concrete implementations for the dynamic consolidation of servers or sober placement policies at the orchestrator level. The works contribute a new approach to running HPC clusters in a cloud environment and test the technique on robustness by adding and removing nodes on the fly.

8.2.2 Model-free control for resource harvesting in computing grids

Cloud and High-Performance Computing (HPC) systems are increasingly facing issues of dynamic variability, in particular w.r.t. performance and power consumption. They are becoming less predictable, and therefore demand more runtime management by feedback loops. In this work, we describe results addressing autonomic administration in HPC systems through a control theoretical approach. We more specifically consider the need for controllers that can adapt to variations a long time in the behavior of controlled systems, but also to being reused on different systems and processors. We therefore explore the application of Model-Free Control (MFC) in the context of resource harvesting in a Computing Grid, by regulating the injection of flexible jobs while limiting perturbation of the prioritary applications.

8.2.3 Painless Transposition of Reproducible Distributed Environments with NixOS Compose

Development of environments for distributed systems is a tedious and time-consuming iterative process. The reproducibility of such environments is a crucial factor for rigorous scientific contributions. We think that being able to smoothly test environments both locally and on a target distributed platform makes development cycles faster and reduces the friction to adopt better experimental practices. To address this issue, this paper introduces the notion of environment transposition and implements it in NixOS Compose, a tool that generates reproducible distributed environments. It enables users to deploy their environments on virtualized (docker, QEMU) or physical (Grid'5000) platforms with the same unique description of the environment. We show that NixOS Compose enables to build reproducible environments without overhead by comparing it to state-of-the-art solutions for the generation of distributed environments (EnOSlib and Kameleon). NixOS Compose actually enables substantial performance improvements on image building time over Kameleon (up to 11x faster for initial builds and up to 19x faster when building a variation of an existing environment).

8.2.4 Reproducibility

The ability to reproduce an experiment is fundamental in this research axis. Existing approaches focus on repeatability, but this is only the first step to reproducibility: continuing a scientific work from a previous experiment requires to be able to modify it. We called this ability reproducibility with variation. We have shown that capturing the environment of execution is necessary but not sufficient ; we also need the environment of development. The variation also implies that those environments are subject to evolution, so the whole software development lifecycle needs to be considered. To take into account these evolutions, software environments need to be clearly defined, reconstructible with variation, and easy to share. We proposed to leverage functional package managers to achieve this goal.

8.3 Integration of High Performance Computing and Data Analytics

Participants: Bruno Raffin, Frederic Wagner, Yves Denneulin.

8.3.1 Data analysis for a single simulation run.

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. It give access to an already existing rich ecosystem to be used in situ such as the parallel versions of Numpy, Pandas and scikit-learn. Experiments in configurations up to 1024 cores show that DEISA can improve the simulation wallclock time (excluding analysis) by a factor up to 3 and the total experiment (including analysis) hour.core cost by a factor of up to 5 compared to parallel post hoc with plain Dask while requiring the modification of only two lines of python code, three of YAML, and none at all in a C simulation code already instrumented with PDI Data Interface. Index Terms—In situ processing, code coupling, task-based programming, MPI, Dask

8.3.2 Data analysis for ensemble simulation runs.

We put significant efforts in investigating in situ processing beyond a single large-scale simulation, considering use cases where the analysis needs to combine data from multiple simulation runs (also commonly called ensemble run). Such use-cases are becoming more common with the need to sample the simulation behavior within some parameter ranges for extracting knowledge using statistical or machine learning based methods, combined with the availability of large supercomputers capable today of running thousands of large simulation instances. Each simulation being potentially large, the amount of data generated by multiple runs is huge, leading to a pressing need for frugal I/O solutions. Initial work focused on sensibility analysis, where the data produced by the simulation are aggregated to compute statistics. We developed Melissa, a file avoiding, fault tolerant and elastic framework. Melissa is built around two key concepts: iterative (sometimes also called incremental) statistics algorithms and asynchronous client/server model for data transfer. Simulation outputs are never stored on disc. They are sent by the simulations to a parallel server, which aggregates them to the statistic fields in an iterative fashion, and then throw them away. This allows to compute oblivious statistics maps on every mesh element for every time step on a full scale study. 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.

Then we extended this work in two directions:

- Instead of computing statistics we are investigating how to train on-line surrogate models using deep learning. This join work with EDF Lab led to (PhD Lucas Meyer) led to revisite the classical epoch based training process to adapt it to the online setup enabled by Melissa. First investigation point is to add the necessary mechnisms to ensure that the no biais is intrudced due to the specificity of the online training: order of data generation that is dicted by the simulation its self, the resources available on the compute machine to run various simulations concurrently, and the limited amount of memory avaiable for storing data [32].

- The other direction focused on extending Melissa for ensemble based Data Assimilation (DA) (Sebastian Friedemann PhD). Data assimilation is more demanding than open-loop data processing, as the analysis results are used to steer the simulations progresses. The general approach consists in periodically correcting the simulations trajectories by minimizing the global error obtained from the combination of observation data, typically acquired by on-the-field sensors, and simulation data. Data assimilation is particularly used in domains like weather forecast and climate simulation where numerical models are highly sensible to parameter values. Our solution extends Melissa with simulation virtualization and dynamics load balancing, further improving the elasticity and efficiency of executions. Experiments run on up to 16240 cores, to propagate 16384 members using the ParFlow hydrology simulation code [9]

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.
- **Qarnot Computing (2019-2022)**. PhD grant (Angan Mitra). 175K euros
- **Ryax Technologies (2020-2023)**. PhD grant (Anderson Andrei Da Silva). 170K euros.
- **Berger-Levrault (2022-2025)**. PhD grant (Halmza Safri). 170K euros
- **ATOS (2022-2025)**. PhD grant (Abdessalam Benharii). 170K euros
- **ADEUNIS (2022-2025)**. PhD grant (Louis Closson). 170K euros

10 Partnerships and cooperations

10.1 International initiatives

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

- Associate Team UNIFY (Intelligent Unified Data Services for Hybrid Workflows Combining Compute-Intensive Simulations and Data-Intensive Analytics at Extreme Scales)
- Partners:
  - INRIA teams: KerData, DataMove
  - Argonne National Lab (Tom PETERKA)
- Duration: 2019-2022
- Project providing travel money, mainly untouched due to Covid.

10.1.2 STIC/MATH/CLIMAT AmSud projects

SAQED

Title: Scalable Approximate Query Evaluation on Document Inverted Files for GPU based Big-Data Applications

Program: STIC-AmSud

Duration: January 2020 – December 2022

Local supervisor: Bruno Raffin

Partners:

- Marin (Chili)
- Senger (Brésil)

Inria contact: Bruno Raffin

Summary: Very large collections of documents have become frequent in several application areas such as in medicine, social sciences, natural language processing, e-commerce and many others. The web is an example of a large document collection. Such collections are continuously growing in terms of the amount of documents, and the number of search operations (e.g. queries, similarity evaluation, document ranking, etc.) executed on the datasets. These factors impact the scalability of search engines twofold. The increase in the number of documents directly impacts the size of document indexes and the complexity of individual text similarity evaluations. The increase in the rate of evaluations demanded also increase the need for throughput. CPUs are optimized for low latency execution of a moderate number of application threads. In contrast, GPU architectures are designed to deliver high throughput computing for massive numbers of threads. In this project, we intend to collaborate in the development of an enhanced parallel version of the WAND ranking algorithm using the heterogeneous power of GPUs and CPUs to execute document evaluation on massive collections of documents in scalable and efficient ways. A scheduling algorithm that uses the GPU as a static cache to process the most frequent and computationally expensive queries will be proposed and evaluated. Additionally, as a case of study we propose to evaluate our proposal in online e-commerce systems. As an outcome from the collaboration between the researchers in this proposal, we expect to publish the results in high ranking conferences and journals in the related research fields. We also expect to collaborate in thesis supervision of MSc and PhD students.

10.2 International research visitors

10.2.1 Visits of international scientists

Inria International Chair Ian Foster, ANL and University of Chicago. Inria International Chair 2021-2026. 2021 and 2022 visits canceled due to Covid traveling restrictions.

10.3 European initiatives

10.3.1 H2020 projects

EoCoE-II EoCoE-II project on cordis.europa.eu

Title: Energy Oriented Center of Excellence : toward exascale for energy

Duration: From January 1, 2019 to June 30, 2022

Partners:

- DATADIRECT NETWORKS FRANCE (DDN Storage), France
- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- UNIVERSITA DEGLI STUDI DI ROMA TOR VERGATA (UNITOV), Italy
- UNIVERSITY OF BATH (UBAH), United Kingdom
- FRIEDRICH-ALEXANDER-UNIVERSITAET ERLANGEN-NUERNBERG (FAU), Germany
- FORSCHUNGSZENTRUM JULICH GMBH (FZJ), Germany
- RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN (RWTH AACHEN), Germany
- FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV (FHG), Germany
- ECOLE NORMALE SUPERIEURE DE LYON (ENS DE LYON), France
- COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES (CEA), France
- UNIVERSITE DE STRASBOURG (UNISTRA), France
- CENTRO DE INVESTIGACIONES ENERGETICAS, MEDIOAMBIENTALES Y TECNOLOGICAS-CIEMAT (CIEMAT), Spain

- INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK, Poland
- UNIVERSITE LIBRE DE BRUXELLES (ULB), Belgium
- AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE (ENEA), Italy
- UNIVERSITE GRENOBLE ALPES (UGA), France
- CENTRE EUROPEEN DE RECHERCHE ET DEFORMATION AVANCEE EN CALCUL SCIENTIFIQUE (CERFACS), France
- CONSIGLIO NAZIONALE DELLE RICERCHE (CNR), Italy
- UNIVERSITA DEGLI STUDI DI TRENTO (UNITN), Italy
- IFP Energies nouvelles (IFPEN), France
- MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV (MPG), Germany
- INSTITUT NATIONAL POLYTECHNIQUE DE TOULOUSE (INPT), France
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS (CNRS), France
- BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS), Spain

Inria contact: Bruno RAFFIN

Coordinator:

Summary: The Energy-oriented Centre of Excellence (EoCoE) applies cutting-edge computational methods in its mission to accelerate the transition to the production, storage and management of clean, decarbonized energy. EoCoE is anchored in the High Performance Computing (HPC) community and targets research institutes, key commercial players and SMEs who develop and enable energy-relevant numerical models to be run on exascale supercomputers, demonstrating their benefits for low-carbon energy technology. The present project will draw on a successful proof-of-principle phase of EoCoE-I, where a large set of diverse computer applications from four such energy domains achieved significant efficiency gains thanks to its multidisciplinary expertise in applied mathematics and supercomputing. During this 2nd round, EoCoE-II will channel its efforts into 5 scientific Exascale challenges in the low-carbon sectors of Energy Meteorology, Materials, Water, Wind and Fusion. This multidisciplinary effort will harness innovations in computer science and mathematical algorithms within a tightly integrated co-design approach to overcome performance bottlenecks and to anticipate future HPC hardware developments. A world-class consortium of 18 complementary partners from 7 countries will form a unique network of expertise in energy science, scientific computing and HPC, including 3 leading European supercomputing centres. New modelling capabilities in selected energy sectors will be created at unprecedented scale, demonstrating the potential benefits to the energy industry, such as accelerated design of storage devices, high-resolution probabilistic wind and solar forecasting for the power grid and quantitative understanding of plasma core-edge interactions in ITER-scale tokamaks. These flagship applications will provide a high-visibility platform for high-performance computational energy science, cross-fertilized through close working connections to the EERA and EUROfusion consortia.

PRACE-6IP PRACE-6IP project on cordis.europa.eu

Title: PRACE 6th Implementation Phase Project

Duration: From May 1, 2019 to December 31, 2022

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France

- CENTRUM SPOLOCNYCH CINNOSTI SLOVENSKEJ AKADEMIE VIED (CENTRE OF OPERATIONS OF THE SLOVAK ACADEMY OF SCIENCES), Slovakia
- GRAND EQUIPEMENT NATIONAL DE CALCUL INTENSIF (GENCI), France
- UNIVERSIDADE DO MINHO (UMINHO), Portugal
- LINKOPINGS UNIVERSITET (LIU), Sweden
- VSB - TECHNICAL UNIVERSITY OF OSTRAVA (VSB - TU Ostrava), Czechia
- MACHBA - INTERUNIVERSITY COMPUTATION CENTER (IUCC), Israel
- TECHNISCHE UNIVERSITAET WIEN (TU WIEN), Austria
- Gauss Centre for Supercomputing (GCS) e.V. (GCS), Germany
- FUNDACION PUBLICA GALLEGA CENTRO TECNOLOGICO DE SUPERCOMPUTACION DE GALICIA (CESGA), Spain
- UNIVERSITEIT ANTWERPEN (UANTWERPEN), Belgium
- NATIONAL UNIVERSITY OF IRELAND GALWAY (NUI GALWAY), Ireland
- AKADEMIA GORNICZO-HUTNICZA IM. STANISLAWA STASZICA W KRAKOWIE (AGH / AGH-UST), Poland
- KUNGLIGA TEKNISKA HOEGSKOLAN (KTH), Sweden
- FORSCHUNGSZENTRUM JULICH GMBH (FZJ), Germany
- EUDAT OY (EUDAT), Finland
- KORMANYZATI INFORMATIKAI FEJLESZTESI UGYNOKSEG (GOVERNMENTAL INFORMATION TECHNOLOGY DEVELOPMENT AGENCY), Hungary
- COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES (CEA), France
- NATIONAL INFRASTRUCTURES FOR RESEARCH AND TECHNOLOGY (GRNET S.A.), Greece
- GEANT VERENIGING (GEANT VERENIGING), Netherlands
- UNIVERSIDADE DE EVORA (UNIVERSIDADE DE EVORA), Portugal
- KOBENHAVNS UNIVERSITET (UCPH), Denmark
- UPPSALA UNIVERSITET (UU), Sweden
- INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK, Poland
- ISTANBUL TEKNIK UNIVERSITESI (ITU), Türkiye
- BAYERISCHE AKADEMIE DER WISSENSCHAFTEN (BADW), Germany
- SURF BV, Netherlands
- ASSOCIACAO DO INSTITUTO SUPERIOR TECNICO PARA A INVESTIGACAO E DESENVOLVIMENTO (IST ID), Portugal
- PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE AISBL (PRACE), Belgium
- UMEA UNIVERSITET, Sweden
- UNIVERSIDADE DE COIMBRA (UNIVERSIDADE DE COIMBRA), Portugal
- UNITED KINGDOM RESEARCH AND INNOVATION (UKRI), United Kingdom
- UNIVERSITE DU LUXEMBOURG (uni.lu), Luxembourg
- EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH (ETH Zürich), Switzerland
- SYDDANSK UNIVERSITET (SDU), Denmark
- "ASSOCIATION ""NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS" (NCSA), Bulgaria
- BILKENT UNIVERSITESI VAKIF (BILKENTUNIVERSITY BILIM KENTI), Türkiye
- UNIVERSITETET I OSLO (UNIVERSITY OF OSLO), Norway

- DANMARKS TEKNISKE UNIVERSITET (DTU), Denmark
- UNIVERSIDADE DO PORTO (U.PORTO), Portugal
- SIGMA2 AS (SIGMA2), Norway
- UNIVERSITY OF STUTTGART (USTUTT), Germany
- MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV (MPG), Germany
- UNIVERSITAET INNSBRUCK (UIBK), Austria
- CINECA CONSORZIO INTERUNIVERSITARIO (CINECA), Italy
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS (CNRS), France
- CENTRE INFORMATIQUE NATIONAL DE L'ENSEIGNEMENT SUPERIEUR (CINES), France
- POLITECHNIKA WROCLAWSKA (PWR), Poland
- POLITECHNIKA GDANSKA (GDANSK TECH), Poland
- NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET NTNU (NTNU), Norway
- THE CYPRUS INSTITUTE (THE CYPRUS INSTITUTE), Cyprus
- THE UNIVERSITY OF EDINBURGH (UEDIN), United Kingdom
- UNIVERZA V LJUBLJANI (UL), Slovenia
- BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS), Spain
- CSC-TIETEEN TIETOTEKNIIKAN KESKUS OY (CSC-IT CENTER FOR SCIENCE LTD), Finland

Inria contact: Jean ROMAM; Frédéric DESPREZ; Luc GIRAUD

Coordinator:

Summary: PRACE, the Partnership for Advanced Computing is the permanent pan-European High Performance Computing service providing world-class systems for world-class science. Systems at the highest performance level (Tier-0) are deployed by Germany, France, Italy, Spain and Switzerland, providing researchers with more than 17 billion core hours of compute time. HPC experts from 25 member states enabled users from academia and industry to ascertain leadership and remain competitive in the Global Race. Currently PRACE is finalizing the transition to PRACE 2, the successor of the initial five year period. The objectives of PRACE-6IP are to build on and seamlessly continue the successes of PRACE and start new innovative and collaborative activities proposed by the consortium. These include: assisting the development of PRACE 2; strengthening the internationally recognised PRACE brand; continuing and extend advanced training which so far provided more than 36 400 person·training days; preparing strategies and best practices towards Exascale computing, work on forward-looking SW solutions; coordinating and enhancing the operation of the multi-tier HPC systems and services; and supporting users to exploit massively parallel systems and novel architectures. A high level Service Catalogue is provided. The proven project structure will be used to achieve each of the objectives in 7 dedicated work packages. The activities are designed to increase Europe's research and innovation potential especially through: seamless and efficient Tier-0 services and a pan-European HPC ecosystem including national capabilities; promoting take-up by industry and new communities and special offers to SMEs; assistance to PRACE 2 development; proposing strategies for deployment of leadership systems; collaborating with the ETP4HPC, CoEs and other European and international organisations on future architectures, training, application support and policies. This will be monitored through a set of KPIs.

REGALE REGALE project on cordis.europa.eu

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

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
- ELECTRICITE DE FRANCE (EDF), France
- ETHNICON METSOVION POLYTECHNION (NATIONAL TECHNICAL UNIVERSITY OF ATHENS - NTUA), Greece
- TECHNISCHE UNIVERSITAET MUENCHEN (TUM), Germany
- GIOUMPITEK MELETI SCHEDIASMOS YLOPOIISI KAI POLISI ERGON PLIROFORIKIS ETAIREIA PERIORISMENIS EFTHYNIS (UBITECH DESIGN PLANNING IMPLEMENTATION & SALE OF INFORMATION WORKS), Greece
- BAYERISCHE AKADEMIE DER WISSENSCHAFTEN (BADW), Germany
- UNIVERSITE GRENOBLE ALPES (UGA), France
- E 4 COMPUTER ENGINEERING SPA (E4), Italy
- BULL SAS (BULL), France
- TWT GMBH SCIENCE & INNOVATION (TWT GMBH SCIENCE & INNOVATION), Germany
- ALMA MATER STUDIORUM - UNIVERSITA DI BOLOGNA (UNIBO), Italy
- SCIO IKE (SCIO P.C.), Greece
- CINECA CONSORZIO INTERUNIVERSITARIO (CINECA), Italy
- EREVNITIKO PANEPISTILIAKO INSTITOUTO SYSTIMATION EPIKOINONIAS KAI YPOLOGISTON-EMP (RESEARCH UNIVERSITY INSTITUTE COMMUNICATION AND COMPUTER SYSTEMS), Greece
- ANDRITZ HYDRO GMBH, Austria
- BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS), Spain

Inria contact: Bruno RAFFIN

Coordinator:

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.

10.4 National initiatives

Participants: Bruno Raffin, Olivier Richard, Denis Trystram, Fanny Dufossé, Gregory Mounié, Pierre-François Dutot.

10.4.1 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.

10.4.2 ANR

- **PPR Océan et Climat MEIDATION (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.

- **ANR grant Energumen (2018-2022).** Resource management: malleable jobs for a better use of the resources along with energy optimization. Coordinator: Denis Trystram. Partners: Grenoble-INP, IRIT, Sorbonne Université.

10.4.3 INRIA

- INRIA Challenge HPC-BigData (2018-2022). Convergence between HPC, Big Data and AI. Coordinator: Bruno Raffin. Partners: the INRIA teams Zenith, Kerdata, Datamove, Tadaam, SequeL, Parietal, Tau, and the external partners ATOS, ANL, IBPC, ESI-Group. See Web Site

10.4.4 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

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Journee sur la Recherche en Apprentissage Frugal. 24/25 Nov 2022, Grenoble

- General co-Chair of the 19th (Jinan) Annual IFIP International Conference on Network and Parallel Computing (IFIP NPC).

Member of the organizing committees

- ISAV'22 Workshop (In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization), November 2022, USA.

Member of the conference program committees

- ISPDC (21st International Symposium on Parallel and Distributed Computing) Basel, Switzerland, July 11-13.
- IPDPS Workshop: EduPar, Lyon, France, May.
- PPAM (14th International Conference on Parallel Processing and Applied Mathematics) Gdansk, Poland, sept. 11-14.
- IEEE Cluster Conference, Heidelberg, Germany, September 6-9.
- LATIN, Guanajuato, Mexico, Oct.
- Supercomputing Workshops, USA, Nov.
- ICPP (51st International Conference on Parallel Processing), Bordeaux, France, August 29th to Sept. 1st.
- IEEE Symposium on Large Data Analysis and Visualization 2022 (LDAV 2022), in conjunction with IEEE VIS in Oklahoma City, OK, USA, on October 15.
- Eurographics Symposium on Parallel Graphics and Visualization (EGPGV) , Italy, Nov.
- Compas (Conférence francophone d'informatique en Parallélisme, Architecture et Système), Amiens, France, July 5-8.

11.1.2 Journal

Member of the editorial boards

- Computational Methods in Science and Technology.
- ARIMA (revue africaine de recherche en informatique et maths appliquées).
- Theory of Computing System TOCS
- IEEE Transactions on Computers
- Springer Journal of Cloud Computing

11.1.3 Invited talks

- Keynote Speaker at the 19th Annual IFIP International Conference on Network and Parallel Computing (NPC 2022), Jinan, China on Sep. 2022.
- *Le numérique face au réchauffement climatique : opportunité ou handicap ?*. Keynote at ROADEF anual conference, Lyon,Fev. 2022.
- *Carbon cost of Euro-Par*. Invited talk Euro-Par 2022, Glasgow, Aug. 2022.

11.1.4 Leadership within the scientific community

- Member of the advisory committee of IEEE Cloud Computing ;

11.2 Teaching - Juries

11.2.1 Teaching

- Denis Trystram. 200 hours per year, ENSIMAG, Grenoble-INP, Master
- Fanny Dufossé. 20 hours per year, Algorithmic, Licence. Univ. Grenoble-Alpes and Licence Ensimag, Combinatorial scientific computing, Master, ENS Lyon.
- Pierre-François Dutot. 226 hours per year. Licence (first and second year) at IUT2/UPMF (Institut Universitaire Technologique de Univ. Grenoble-Alpes) and 9 hours Master M2R-ISC Informatique-Systèmes-Communication at Univ. Grenoble-Alpes.
- Grégory Mounié is responsible for the first year (M1) of the international Master of Science in Informatics at Grenoble (MOSIG-M1). 317 hours per year. Master (M1 and M2 year) at Engineering school ENSIMAG, Grenoble-INP, Univ Grenoble Alpes.
- Bruno Raffin is responsible of the Distributed Computing Track, Master international Science in Informatics at Grenoble (MoSIG), UGA. 35 hours per year (M2 - Parallel Systems, L3 - Big Data Processing).
- Olivier Richard is responsible for the third year of the computer science department of Grenoble INP. 222 hours per year. Master at Engineering school Polytech-Grenoble, Univ. Grenoble-Alpes.
- Frédéric Wagner. 220 hours per year. Engineering school ENSIMAG, Grenoble-INP, Master (M1/2nd year and M2/3rd year).
- Yves Denneulin. 192 hours per year. Engineering school ENSIMAG, Grenoble-INP, Master (M1/2nd year and M2/3rd year).

11.2.2 Juries

- Jury member of Yishu DU, Fault-tolerant algorithms for iterative applications and batch schedulers, Dec. 6th, Lyon.

11.2.3 Internal or external Inria responsibilities

- France Exascale Project:
  - Co-coordinator of the SUN (Sciences et Usages du Numérique) axis (Les applications françaises face à l'exascale).
  - INRIA representative at GENCI for the sub-project 3 (SP3).

12 Scientific production

12.1 Major publications

[1] 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.

[2] 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.

[3] 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.

[4] F. Zanon Boito, E. Camilo Inacio, J. Luca Bez, P. O. A. Navaux, M. A. R. Dantas and Y. Denneulin. 'A Checkpoint of Research on Parallel I/O for High Performance Computing'. In: *ACM Computing Surveys* 51.2 (Mar. 2018), 23:1–23:35. DOI: 10.1145/3152891. URL: https://hal.univ-grenoble-alpes.fr/hal-01591755.

12.2 Publications of the year

International journals

[5] J. Bertrand, F. Dufossé, S. Singh and B. Uçar. 'Algorithms and Data Structures for Hyperedge Queries'. In: *ACM Journal of Experimental Algorithmics* 27 (31st Dec. 2022), pp. 1–23. DOI: 10.1145/3568421. URL: https://inria.hal.science/hal-03905905.

[6] C. Cérin, K. Kimura and M. Sow. 'Data stream clustering for low-cost machines'. In: *Journal of Parallel and Distributed Computing* 166 (Aug. 2022), pp. 57–70. DOI: 10.1016/j.jpdc.2022.04.009. URL: https://hal.science/hal-03960663.

[7] F. Dufossé, C. Dürr, N. Nadal, D. Trystram and Ó. Vásquez. 'Scheduling with a processing time oracle'. In: *Applied Mathematical Modelling* 104 (Apr. 2022), pp. 701–720. DOI: 10.1016/j.apm.2021.12.020. URL: https://hal.science/hal-03523262.

[8] F. Dufossé, K. Kaya, I. Panagiotas and B. Uçar. 'Scaling matrices and counting the perfect matchings in graphs'. In: *Discrete Applied Mathematics* 308 (Feb. 2022), pp. 130–146. DOI: 10.1016/j.dam.2020.07.016. URL: https://inria.hal.science/hal-01743802.

[9] S. Friedemann and B. Raffin. 'An elastic framework for ensemble-based large-scale data assimilation'. In: *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.

[10] Y. Li, D. Ou, X. Zhou, C. Jiang and C. Cérin. 'Scalability and performance analysis of BDPS in clouds'. In: *Computing* 104.6 (June 2022), pp. 1425–1460. DOI: 10.1007/s00607-022-01056-7. URL: https://hal.science/hal-03960673.

[11] N. K. Thắng, A. Srivastav, D. Trystram and P. Youssef. 'A stochastic conditional gradient algorithm for decentralized online convex optimization'. In: *Journal of Parallel and Distributed Computing* 169 (Nov. 2022), pp. 334–351. DOI: 10.1016/j.jpdc.2022.07.010. URL: https://hal.science/hal-03955088.

[12] S. Zrigui, R. y. de Camargo, A. Legrand and D. Trystram. 'Improving the Performance of Batch Schedulers Using Online Job Runtime Classification'. In: *Journal of Parallel and Distributed Computing* 164 (2nd Feb. 2022), pp. 83–95. DOI: 10.1016/j.jpdc.2022.01.003. URL: https://hal.science/hal-03023222.

International peer-reviewed conferences

[13] 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.

[14] D. Carastan dos Santos and T. H. T. Pham. 'Understanding the Energy Consumption of HPC Scale Artificial Intelligence'. In: CARLA 2022 - Latin America High Performance Computing Conference. Porto Alegre, Brazil, 26th Sept. 2022, pp. 1–14. URL: https://hal.inria.fr/hal-03845090.

[15] O. Carrillo, C. J. B. Hernandez, F. Le Mouel, H. Barrera, Y. Denneulin, J. T. Hernandez, F. J. Vargas, L. X. B. Rozo, C. Roncancio and M. Riveill. 'Scalable Architectures to Support Sustainable Advanced Information Technologies'. In: *IEEE Xplore*. CLUSTER 2022 - 2022 IEEE International Conference on Cluster Computing. Heidelberg, Germany: IEEE, 5th Sept. 2022, pp. 512–515. DOI: 10.1109/CLUSTER51413.2022.00065. URL: https://hal.science/hal-03844189.

[16] E. Foussard, M.-L. Espinouse, G. Mounié and M. Nattaf. 'Ordonnancement de la production et de la maintenance sur une machine multicomposant: étude de complexité'. In: 23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Villeurbanne - Lyon, France, 23rd Feb. 2022. URL: https://hal.science/hal-03595366.

[17] Q. Guilloteau, J. Bleuzen, M. Poquet and O. Richard. 'Painless Transposition of Reproducible Distributed Environments with NixOS Compose'. In: CLUSTER 2022 - IEEE International Conference on Cluster Computing. Vol. CLUSTER 2022 - IEEE International Conference on Cluster Computing. Heidelberg, Germany, 6th Sept. 2022, pp. 1–12. URL: https://hal.science/hal-03723771.

[18] Q. Guilloteau, B. Robu, C. Join, M. Fliess, É. Rutten and O. Richard. 'Model-free control for resource harvesting in computing grids'. In: CCTA 2022 - Conference on Control Technology and Applications, CCTA 2022. Trieste, Italy: IEEE, 22nd Aug. 2022. DOI: 10.1109/CCTA49430.2022.9966035. URL: https://hal.science/hal-03663273.

[19] 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. URL: https://inria.hal.science/hal-04030223.

[20] A. Mitra, N. K. Thang, T.-A. Nguyen, D. Trystram and P. Youssef. 'Online Decentralized Frank-Wolfe: From theoretical bound to applications in smart-building'. In: Global IoT Conference (GloTS 2022). Vol. 13533. Lecture Notes in Computer Science. Dublin, Ireland, 20th June 2022, pp. 43–54. DOI: 10.1007/978-3-031-20936-9_4. URL: https://hal.science/hal-03710138.

[21] T.-A. Nguyen, N. K. Thang and D. Trystram. 'One gradient frank-wolfe for decentralized online convex and submodular optimization'. In: ACML 2022 - 14th Asian Conference in Machine Learning. Hyderabad, India, 12th Dec. 2022, pp. 1–33. URL: https://hal.inria.fr/hal-03835713.

[22] H. Safri, M. M. Kandi, Y. Miloudi, C. Bortolaso, D. Trystram and F. Desprez. 'Towards Developing a Global Federated Learning Platform for IoT'. In: *Proceedings of IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS). Bologna, Italy: IEEE, 10th July 2022, pp. 1312–1315. DOI: 10.1109/ICDCS54860.2022.00145. URL: https://hal.science/hal-03955067.

[23] M. Vasconcelos, D. Cordeiro and F. Dufossé. 'Dimensioning of multi-clouds with follow-the-renewable approaches for environmental impact minimization'. In: 23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Villeurbanne - Lyon, France, 23rd Feb. 2022. URL: https://hal.science/hal-03595278.

National peer-reviewed Conferences

[24] V. Fagnon, G. Lucarelli, C. Mommessin and D. Trystram. 'Ordonnancement Deux-Agents avec Augmentation de Ressources'. In: 23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Villeurbanne - Lyon, France, 23rd Feb. 2022. URL: https://hal.science/hal-03595369.

[25] T. Zhang, D. Ou, Z. Ge, C. Jiang, C. Cérin and L. Yan. 'Interference-aware Workload Scheduling in Co-located Data Centers'. In: Network and Parallel Computing: 19th IFIP WG 10.3 International Conference, NPC 2022. Vol. 13615. Lecture Notes in Computer Science. Jinan, China: Springer Nature Switzerland, 1st Dec. 2022, pp. 69–82. DOI: 10.1007/978-3-031-21395-3_7. URL: https://hal.science/hal-03960635.

Conferences without proceedings

[26] L. Closson, C. Cérin, D. Donsez and D. Trystram. 'Towards a Methodology for the Characterization of IoT Data Sets of the Smart Building Sector'. In: 2022 IEEE International Smart Cities Conference (ISC2). Pafos, Cyprus: IEEE; IEEE, Oct. 2022, pp. 1–7. DOI: 10.1109/ISC255366.2022.9921984. URL: https://hal.science/hal-03951666.

[27] 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.

[28] V. Grandgirard, K. Obrejan, D. Midou, Y. Asahi, P.-E. Bernard, J. Bigot, E. Bourne, J. Dechard, G. Dif- Pradalier, P. Donnel, X. Garbet, A. Gueroudji, G. Hager, H. Murai, Y. Ould-Ruis, T. Padioleau, L. Nguyen, M. Peybernes, Y. Sarazin, M. Sato, M. Tsuji and P. Vezolle. 'New advances to prepare GYSELA-X code for exascale global gyrokinetic plasma turbulence simulations: porting on GPU and ARM architectures'. In: PASC22 - Conderence on The Platform for Advanced Scientific Computing. Bâle (virtual event), Switzerland, 27th June 2022, pp. 1–21. URL: https://hal-cea.archives-ouvertes.fr/cea-03740685.

[29] N. Greneche, T. Menouer, C. Cérin and O. Richard. 'A Methodology to Scale Containerized HPC Infrastructures in the Cloud'. In: Euro-Par 2022: Parallel Processing: 28th International Conference on Parallel and Distributed Computing. Vol. 13440. Lecture Notes in Computer Science. Glasgow, United Kingdom: Springer International Publishing, 1st Aug. 2022, pp. 203–217. DOI: 10.1007/978-3-031-12597-3_13. URL: https://hal.science/hal-03960678.

[30] Q. Guilloteau, J. Bleuzen, M. Poquet and O. Richard. 'Transposition d'environnements distribués reproductibles avec NixOS Compose'. In: COMPAS 2022 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Amiens, France, 5th July 2022, pp. 1–9. URL: https://hal.science/hal-03696485.

[31] Q. Guilloteau, O. Richard and É. Rutten. 'Étude des applications Bag-of-Tasks du méso-centre Gricad'. In: COMPAS 2022 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Amiens, France, 5th July 2022, pp. 1–7. URL: https://hal.science/hal-03702246.

[32] L. Meyer, A. Ribés and B. Raffin. 'Simulation-Based Parallel Training'. In: NeurIPS 2022 - AI for Science Workshop. New Orleans, United States, 28th Nov. 2022. URL: https://hal.science/hal-03842106.

[33] A. Ribés, R. Persicot, L. Meyer and J.-P. Ducreux. 'A hybrid Reduced Basis and Machine-Learning algorithm for building Surrogate Models: a first application to electromagnetism'. In: Workshop on Machine Learning and the Physical Sciences workshop at NeurIPS 2022. New Orleans, United States, 3rd Dec. 2022, pp. 1–6. URL: https://hal.science/hal-03850571.

[34] H. Safri, M. M. Kandi, Y. Miloudi, C. Bortolaso, D. Trystram and F. Desprez. 'A Federated Learning Framework for IoT: Application to Industry 4.0'. In: 2022 22nd International Symposium on Cluster, Cloud and Internet Computing (CCGrid). Taormina, Italy: IEEE, 16th May 2022, pp. 565–574. URL: https://hal.science/hal-03955062.

[35] M. F. Silva Vasconcelos, D. Cordeiro and F. Dufossé. 'Indirect Network Impact on the Energy Consumption in Multi-clouds for Follow-the-renewables Approaches'. In: *Proceedings of the 11th International Conference on Smart Cities and Green ICT Systems - SMARTGREENS*. 11th International Conference on Smart Cities and Green ICT Systems. Online Streaming, Brazil: SCITEPRESS - Science and Technology Publications, 27th Apr. 2022, pp. 44–55. DOI: 10.5220/0011047000003203. URL: https://hal.science/hal-03658289.

Scientific book chapters

[36] A. Ribés, T. Terraz, Y. Fournier, B. Iooss and B. Raffin. 'Unlocking Large Scale Uncertainty Quantification with In Transit Iterative Statistics'. In: *In Situ Visualization for Computational Science*. Mathematics and Visualization. Springer International Publishing, 5th May 2022, pp. 113–136. DOI: 10.1007/978-3-030-81627-8_6. URL: https://hal.inria.fr/hal-03950616.

Reports & preprints

[37] L. Angelelli, A. A. Da Silva, M. Mercier, G. Mounié, D. Trystram and Y. Georgiou. *Towards a Multi-objective Scheduling Policy for Serverless-based Edge-Cloud Continuum*. 4th May 2023. URL: https://inria.hal.science/hal-04027179.

[38] J. Bertrand, F. Dufossé, S. Singh and B. Uçar. *Algorithms and data structures for hyperedge queries*. RR-9390. Inria Grenoble Rhône-Alpes, 29th Apr. 2022, p. 28. URL: https://inria.hal.science/hal-03127673.

[39] D. Carastan dos Santos, K. Rzadca, L. Sousa and D. Trystram. *A community-guided discussion about social and environmental effects of post-COVID-19 Computer Science conferencing*. 2023. URL: https://hal.science/hal-03903632.

[40] V. Fagnon, G. Lucarelli, C. Mommessin and D. Trystram. *Two-Agent Scheduling with Resource Augmentation on Multiple Machines*. 12th May 2022. URL: https://hal.science/hal-03666851.

[41] S. Friedemann, K. Keller, Y.-S. Lu, B. Raffin and L. Bautista Gomez. *A Framework for Large Scale Particle Filters Validated with Data Assimilation for Weather Simulation*. 2023. URL: https://inria.hal.science/hal-03927612.

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

[43] Q. Guilloteau, O. Richard and É. Rutten. *Étude des applications Bag-of-Tasks du méso-centre Gricad*. 18th July 2022. URL: https://hal.science/hal-03726257.

[44] L. Lefèvre, A.-L. Ligozat, D. Trystram, S. Bouveret, A. Bugeau, J. Combaz, E. Frenoux, G. Guennebaud, J. Lefèvre and J.-P. Nicolaï. *Proposition de document de cadrage Évaluation environnementale de projets impliquant des méthodes d'IA*. EcoInfo, 2022, pp. 1–8. URL: https://hal.science/hal-03853135.

[45] 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.

[46] L. de Sousa Rosa, D. Carastan-Santos and A. Goldman. *A multiple linear regression approach for understanding the trade-offs in learning HPC job scheduling heuristics*. 2023. URL: https://hal.inria.fr/hal-03979237.

Other scientific publications

[47] M. Jay, L. Lefèvre and D. Trystram. 'Comparaison expérimentale de logiciels de mesure d'énergie : côté GPU'. In: Compas 2022. Amiens, France, 5th July 2022. URL: https://inria.hal.science/hal-03945744.

[48] A. Khedimi, T. Menouer, C. Cérin and M. Boudhar. 'A cloud weather forecasting service and its relationship with anomaly detection'. In: *Service Oriented Computing and Applications* 16.3 (Sept. 2022), pp. 191–208. DOI: 10.1007/s11761-022-00346-4. URL: https://hal.science/hal-03960646.

12.3 Other

Scientific popularization

[49] D. Trystram. *Calcul haute performance et ordinateurs superpuissants : la course à l'« exascale »*. 28th Dec. 2022. URL: https://hal.science/hal-03955098.

Softwares

[50] [SW] M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie and B. Fichel, *Artifact: An experimental comparison of software-based power meters*, 1st Dec. 2022. LIC: CC BY-SA 4.0. HAL: ⟨hal-03974900⟩, URL: https://inria.hal.science/hal-03974900, VCS: https://github.com/vladostp/an-experimental-comparison-of-software-based-power-meters, SWHID: ⟨swh:1:dir:6114d7d8e7bda182cd73ba519a7f4f2c27dbe211;origin=https://hal.archives-ouvertes.fr/hal-03974900;visit=swh:1:snp:4ceba8cdbbd0f71a6705518b9fb0f5943b3717b8;anchor=swh:1:rel:0c191544ad6eec42c116b0da34c5ce0c98ae4b45;path=/⟩.