

2025 Activity Report

RESEARCH CENTRE: Inria Lyon Centre

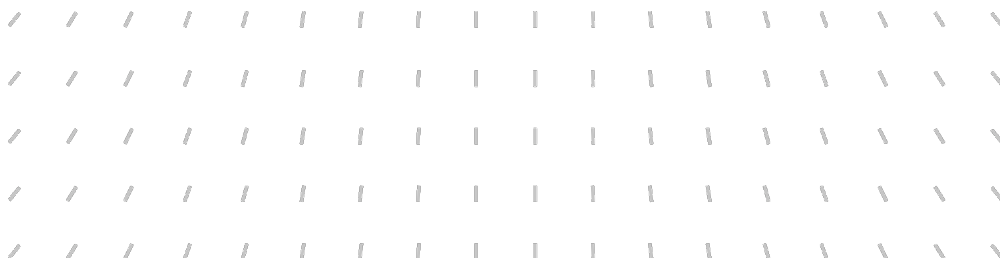
IN PARTNERSHIP WITH: Université Claude Bernard (Lyon 1), Ecole normale supérieure de Lyon, CNRS

Project-Team

AVALON

Algorithms and Software Architectures for Distributed
and HPC Platforms

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



Project-Team AVALON

Creation of the Project-Team: 2014 July 01

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

Keywords

Computer sciences and digital sciences

- A1.1.1. – Multicore, Manycore
- A1.1.2. – Hardware accelerators (GPGPU, FPGA, etc.)
- A1.1.4. – High performance computing
- A1.1.5. – Exascale
- A1.3.5. – Cloud
- A1.3.6. – Fog, Edge
- A1.6. – Green Computing
- A2.1.6. – Concurrent programming
- A2.1.7. – Distributed programming
- A2.1.10. – Domain-specific languages
- A2.2.8. – Code generation
- A2.5.2. – Component-based Design
- A2.6.2. – Middleware
- A2.6.4. – Ressource management
- A3.1.3. – Distributed data
- A4.4. – Security of equipment and software
- A7.1. – Algorithms
- A7.1.1. – Distributed algorithms
- A7.1.2. – Parallel algorithms
- A8.2.1. – Operations research
- A8.9. – Performance evaluation

Other research topics and application domains

- B1.1.7. – Bioinformatics
- B4.5. – Energy consumption
- B4.5.1. – Green computing
- B6.1.1. – Software engineering
- B9.5.1. – Computer science
- B9.7. – Knowledge dissemination
- B9.7.1. – Open access
- B9.7.2. – Open data
- B9.8. – Reproducibility

Contents

Project-Team AVALON	1
1 Team members, visitors, external collaborators	5
2 Overall objectives	6
2.1 Presentation	6
2.2 Objectives	6
3 Research program	7
3.1 Energy Application Profiling and Modeling	7
3.2 Data-intensive Application Profiling, Modeling, and Management	7
3.3 Resource-Agnostic Application Description Model	8
3.4 Application Mapping and Scheduling	8
3.4.1 Application Mapping and Software Deployment	8
3.4.2 Non-Deterministic Workflow Scheduling	9
4 Application domains	9
4.1 Overview	9
4.2 Climatology	9
4.3 Astrophysics	10
4.4 Bioinformatics	10
5 Social and environmental responsibility	10
5.1 Footprint of research activities	10
6 Highlights of the year	10
6.1 Awards	10
7 Latest software developments, platforms, open data	11
7.1 Latest software developments	11
7.1.1 Concerto	11
7.1.2 execo	11
7.1.3 Halley	12
7.1.4 SkyDSoft	12
7.1.5 XKBLAS	12
7.2 New platforms	13
7.2.1 Platform: Grid'5000	13
7.2.2 Platform: SLICES-FR	13
7.2.3 Platform: SLICES RI	13
8 New results	14
8.1 Energy Efficiency in Large Scale Distributed Systems	14
8.1.1 Estimating the power consumption of bare metal water-cooled servers	14
8.1.2 CPU Frequency Aware Power Modeling for IoT Edge Nodes	14
8.1.3 Revisiting virtual machine consolidation to save resources and energy in heterogeneous production cloud infrastructures	15
8.1.4 Estimating the environmental impact of Generative-AI services	15
8.1.5 Placing leverages on Clouds for footprint reduction	15
8.1.6 Analyzing the Full Life Cycle of IoT-Based 5G Solutions for Smart Agriculture	16
8.2 Edge, Cloud and Distributed Resource Management	16
8.2.1 SkyData: Autonomous Data paradigm	16
8.2.2 Numerics in the Cloud	16
8.2.3 A specialized model and implementation of an actuarial chatbot based on Federated Learning	17

8.3	HPC Applications and Runtimes	17
8.3.1	Measuring and interpreting performances of HPC applications with dependent tasks	17
8.3.2	Handling dynamicity of HPC applications designed by a task-based component model	18
8.3.3	Performance portable batched linear algebra kernels for transport sweeps using Kokkos	18
9	Bilateral contracts and grants with industry	18
9.1	Bilateral grants with industry	18
10	Partnerships and cooperations	19
10.1	International initiatives	19
10.1.1	Participation in other International Programs	19
10.2	European initiatives	20
10.2.1	Horizon Europe	20
10.3	National initiatives	22
11	Dissemination	26
11.1	Promoting scientific activities	26
11.1.1	Scientific events: organisation	26
11.1.2	Scientific events: selection	26
11.1.3	Journal	26
11.1.4	Invited talks	27
11.1.5	Scientific expertise	27
11.1.6	Research administration	27
11.2	Teaching - Supervision - Juries - Educational and pedagogical outreach	27
11.2.1	Supervision	28
11.2.2	Juries	29
11.2.3	Educational and pedagogical outreach	29
11.3	Popularization	29
11.3.1	Productions (articles, videos, podcasts, serious games, ...)	29
11.3.2	Participation in Live events	29
12	Scientific production	30
12.1	Major publications	30
12.2	Publications of the year	30
12.3	Cited publications	31

1 Team members, visitors, external collaborators

Research Scientists

- Christian Perez [Team leader, INRIA, Senior Researcher, HDR]
- Thierry Gautier [INRIA, Researcher, until Aug 2025, HDR]
- Laurent Lefevre [INRIA, Senior Researcher, from Oct 2025, HDR]
- Laurent Lefevre [INRIA, Researcher, until Sep 2025, HDR]

Faculty Members

- Yves Caniou [UNIV LYON I, Associate Professor]
- Eddy Caron [UNIV LYON I, Professor, HDR]
- Olivier Glück [UNIV LYON I, Associate Professor, HDR]
- Elise Jeanneau [UNIV LYON I, Associate Professor]

Post-Doctoral Fellows

- Quentin Guilloteau [INRIA, from Feb 2025]
- Mouna Safir [ENS DE LYON, Post-Doctoral Fellow, from Sep 2025]

PhD Students

- Maxime Agusti [OVH]
- Adrien Berthelot [OCTO TECHNOLOGY, CIFRE, until Feb 2025]
- Emile Egreteau–Druet [INRIA]
- Julien Gaupp [INRIA, from Dec 2025]
- Maxime Just [ENS DE LYON, from Sep 2025]
- Simon Lambert [CIRIL GROUP, CIFRE, until Oct 2025]
- Thomas Stavis [INRIA]

Technical Staff

- Hamza Aabirrouche [INRIA, Engineer, from Mar 2025 until Jun 2025]
- Annour Saad Allamine [INRIA, Engineer, until May 2025]
- Brice-Edine Bellon [INRIA, Engineer, from Feb 2025]
- Simon Delamare [CNRS, Engineer]
- Pierre Jacquot [INRIA, Engineer]
- Jean Christophe Mignot [CNRS, Engineer]
- Emeline Pegon [CNRS, Engineer]
- Pierre-Etienne Polet [INRIA, Engineer]
- Dominique Ponsard [CNRS, Engineer]

- Jean-Camille Seck [INRIA, Engineer, until Sep 2025]
- Cyril Seguin [ENS DE LYON, Engineer]
- Anass Serhani [INRIA, Engineer, until May 2025]

Interns and Apprentices

- Louann Coste [INRIA, Intern, from Mar 2025 until Jul 2025]
- Cyril Devaux [INRIA, Apprentice]
- Maxime Just [ENS DE LYON, Intern, until Mar 2025]
- Basile Leretaille [ENS DE LYON, Intern, from Feb 2025 until Jul 2025]
- Redhouane Messaoud [INRIA, Intern, from Feb 2025 until Jul 2025]
- Alix Peigue [INRIA, Intern, from Sep 2025]

Administrative Assistant

- Chrystelle Mouton [INRIA]

External Collaborator

- Doreid Ammar [AIVANCITY, Professor]

2 Overall objectives

2.1 Presentation

The fast evolution of hardware capabilities in terms of wide area communication, computation and machine virtualization leads to the requirement of another step in the abstraction of resources with respect to parallel and distributed applications. These large scale platforms based on the aggregation of large clusters (Grids), datacenters (Clouds) with IoT (Edge/Fog), or high performance machines (Supercomputers) are now available to researchers of different fields of science as well as to private companies. This variety of platforms and the way they are accessed also have an important impact on how applications are designed (*i.e.*, the programming model used) as well as how applications are executed (*i.e.*, the runtime/middleware system used). The access to these platforms is driven through the use of multiple services providing mandatory features such as security, resource discovery, load-balancing, monitoring, *etc.*

The goal of the AVALON team is to execute parallel and/or distributed applications on parallel and/or distributed resources while ensuring user and system objectives with respect to performance, cost, energy, security, *etc.* Users are generally not interested in the resources used during the execution. Instead, they are interested in how their application is going to be executed: the duration, its cost, the environmental footprint involved, *etc.* This vision of utility computing has been strengthened by the cloud concepts and by the short lifespan of supercomputers (around three years) compared to application lifespan (tens of years). Therefore a major issue is to design models, systems, and algorithms to execute applications on resources while ensuring user constraints (price, performance, *etc.*) as well as system administrator constraints (maximizing resource usage, minimizing energy consumption, *etc.*).

2.2 Objectives

To achieve the vision proposed in the previous section, the AVALON project aims at making progress on four complementary research axes: energy, data, programming models and runtimes, application scheduling.

Energy Application Profiling and Modeling AVALON will improve the profiling and modeling of scientific applications with respect to energy consumption. In particular, it will require to improve the tools that measure the energy consumption of applications, virtualized or not, at large scale, so as to build energy consumption models of applications.

Data-intensive Application Profiling, Modeling, and Management AVALON will improve the profiling, modeling, and management of scientific applications with respect to CPU and data intensive applications. Challenges are to improve the performance prediction of parallel regular applications, to model and simulate (complex) intermediate storage components, and data-intensive applications, and last to deal with data management for hybrid computing infrastructures.

Programming Models and Runtimes AVALON will design component-based models to capture the different facets of parallel and distributed applications while being resource agnostic, so that they can be optimized for a particular execution. In particular, the proposed component models will integrate energy and data modeling results. AVALON in particular targets OpenMP runtime as a specific use case and contributes to improve it for multi-GPU nodes.

Application Mapping and Scheduling AVALON will propose multi-criteria mapping and scheduling algorithms to meet the challenge of automating the efficient utilization of resources taking into consideration criteria such as performance (CPU, network, and storage), energy consumption, and security. AVALON will in particular focus on application deployment, workflow applications, and security management in clouds.

All our theoretical results will be validated with software prototypes using applications from different fields of science such as bioinformatics, physics, cosmology, *etc.* The experimental testbeds GRID'5000 and SLICES will be our platforms of choice for experiments.

3 Research program

3.1 Energy Application Profiling and Modeling

Despite recent improvements, there is still a long road to follow in order to obtain energy efficient, energy proportional and eco-responsible exascale systems. Energy efficiency is therefore a major challenge for building next generation large-scale platforms. The targeted platforms will gather hundreds of millions of cores, low power servers, or CPUs. Besides being very important, their power consumption will be dynamic and irregular.

Thus, to consume energy efficiently, we aim at investigating two research directions. First, we need to improve measurement, understanding, and analysis on how large-scale platforms consume energy. Unlike some approaches [21] that mix the usage of internal and external wattmeters on a small set of resources, we target high frequency and precise internal and external energy measurements of each physical and virtual resource on large-scale distributed systems.

Secondly, we need to find new mechanisms that consume less and better on such platforms. Combined with hardware optimizations, several works based on shutdown or slowdown approaches aim at reducing energy consumption of distributed platforms and applications. To consume less, we first plan to explore the provision of accurate estimation of the energy consumed by applications without pre-executing and knowing them while most of the works try to do it based on in-depth application knowledge (code instrumentation [24], phase detection for specific HPC applications [28], *etc.*). As a second step, we aim at designing a framework model that allows interaction, dialogue and decisions taken in cooperation among the user/application, the administrator, the resource manager, and the energy supplier. While smart grid is one of the last killer scenarios for networks, electrical provisioning of next generation large IT infrastructures remains a challenge.

3.2 Data-intensive Application Profiling, Modeling, and Management

The term “Big Data” has emerged to design data sets or collections so large that they become intractable for classical tools. This term is most of the time implicitly linked to “analytics” to refer to issues such as data curation, storage, search, sharing, analysis, and visualization. However, the Big Data challenge is not

limited to data-analytics, a field that is well covered by programming languages and run-time systems such as Map-Reduce. It also encompasses data-intensive applications. These applications can be sorted into two categories. In High Performance Computing (HPC), data-intensive applications leverage post-petascale infrastructures to perform highly parallel computations on large amount of data, while in High Throughput Computing (HTC), a large amount of independent and sequential computations are performed on huge data collections.

These two types of data-intensive applications (HTC and HPC) raise challenges related to profiling and modeling that the AVALON team proposes to address. While the characteristics of data-intensive applications are very different, our work will remain coherent and focused. Indeed, a common goal will be to acquire a better understanding of both the applications and the underlying infrastructures running them to propose the best match between application requirements and infrastructure capacities. To achieve this objective, we will extensively rely on logging and profiling in order to design sound, accurate, and validated models. Then, the proposed models will be integrated and consolidated within a single simulation framework (SIMGRID). This will allow us to explore various potential “what-if?” scenarios and offer objective indicators to select interesting infrastructure configurations that match application specificities.

Another challenge is the ability to mix several heterogeneous infrastructures that scientists have at their disposal (*e.g.*, Grids, Clouds, and Desktop Grids) to execute data-intensive applications. Leveraging the aforementioned results, we will design strategies for efficient data management service for hybrid computing infrastructures.

3.3 Resource-Agnostic Application Description Model

With parallel programming, users expect to obtain performance improvement, regardless its cost. For long, parallel machines have been simple enough to let a user program use them given a minimal abstraction of their hardware. For example, MPI [23] exposes the number of nodes but hides the complexity of network topology behind a set of collective operations; OpenMP [27] simplifies the management of threads on top of a shared memory machine while OpenACC [26] aims at simplifying the use of GPGPU.

However, machines and applications are getting more and more complex so that the cost of manually handling an application is becoming very high [22]. Hardware complexity also stems from the unclear path towards next generations of hardware coming from the frequency wall: multi-core CPU, many-core CPU, GPGPUs, deep memory hierarchy, *etc.* have a strong impact on parallel algorithms. Parallel languages (UPC, Fortress, X10, *etc.*) can be seen as a first piece of a solution. However, they will still face the challenge of supporting distinct codes corresponding to different algorithms corresponding to distinct hardware capacities.

Therefore, the challenge we aim to address is to define a model, for describing the structure of parallel and distributed applications that enables code variations but also efficient executions on parallel and distributed infrastructures. Indeed, this issue appears for HPC applications but also for cloud oriented applications. The challenge is to adapt an application to user constraints such as performance, energy, security, *etc.*

Our approach is to consider component based models [29] as they offer the ability to manipulate the software architecture of an application. To achieve our goal, we consider a “compilation” approach that transforms a resource-agnostic application description into a resource-specific description. The challenge is thus to determine a component based model that enables to efficiently compute application mapping while being tractable. In particular, it has to provide an efficient support with respect to application and resource elasticity, energy consumption and data management. OpenMP runtime is a specific use case that we target.

3.4 Application Mapping and Scheduling

This research axis is at the crossroad of the AVALON team. In particular, it gathers results of the other research axis. We plan to consider application mapping and scheduling addressing the following three issues.

3.4.1 Application Mapping and Software Deployment

Application mapping and software deployment consist in the process of assigning distributed pieces of software to a set of resources. Resources can be selected according to different criteria such as performance, cost, energy consumption, security management, *etc.* A first issue is to select resources at application launch time. With the wide adoption of elastic platforms, *i.e.*, platforms that let the number of resources allocated to

an application to be increased or decreased during its execution, the issue is also to handle resource selection at runtime.

The challenge in this context corresponds to the mapping of applications onto distributed resources. It will consist in designing algorithms that in particular take into consideration application profiling, modeling, and description.

A particular facet of this challenge is to propose scheduling algorithms for dynamic and elastic platforms. As the number of elements can vary, some kind of control of the platforms must be used accordingly to the scheduling.

3.4.2 Non-Deterministic Workflow Scheduling

Many scientific applications are described through workflow structures. Due to the increasing level of parallelism offered by modern computing infrastructures, workflow applications now have to be composed not only of sequential programs, but also of parallel ones. New applications are now built upon workflows with conditionals and loops (also called non-deterministic workflows).

These workflows cannot be scheduled beforehand. Moreover cloud platforms bring on-demand resource provisioning and pay-as-you-go billing models. Therefore, there is a problem of resource allocation for non-deterministic workflows under budget constraints and using such an elastic management of resources.

Another important issue is data management. We need to schedule the data movements and replications while taking job scheduling into account. If possible, data management and job scheduling should be done at the same time in a closely coupled interaction.

4 Application domains

4.1 Overview

The AVALON team targets applications with large computing and/or data storage needs, which are still difficult to program, deploy, and maintain. Those applications can be parallel and/or distributed applications, such as large scale simulation applications or code coupling applications. Applications can also be workflow-based as commonly found in distributed systems such as grids or clouds.

The team aims at not being restricted to a particular application field, thus avoiding any spotlight. The team targets different HPC and distributed application fields, which brings use cases with different issues. This will be eased with our participation to the Joint Laboratory for Extreme Scale Computing (JLESC), to BioSyL, a federative research structure about Systems Biology of the University of Lyon, or to the SKA project. Last but not least, the team has a privileged connection with CC-IN2P3 that opens up collaborations, in particular in the astrophysics field.

In the following, some examples of representative applications that we are targeting are presented. In addition to highlighting some application needs, they also constitute some of the use cases that will be used to validate our theoretical results.

4.2 Climatology

The world's climate is currently changing due to the increase of the greenhouse gases in the atmosphere. Climate fluctuations are forecasted for the years to come. For a proper study of the incoming changes, numerical simulations are needed, using general circulation models of a climate system. Simulations can be of different types: HPC applications (*e.g.*, the NEMO framework [25] for ocean modelization), code-coupling applications (*e.g.*, the OASIS coupler [30] for global climate modeling), or workflows (long term global climate modeling).

As for most applications the team is targeting, the challenge is to thoroughly analyze climate-forecasting applications to model their needs in terms of programming model, execution model, energy consumption, data access pattern, and computing needs. Once a proper model of an application has been set up, appropriate scheduling heuristics can be designed, tested, and compared. The team has a long tradition of working with CERFACS on this topic, since for example in the LEGO (2006-09) and SPADES (2009-12) French ANR projects.

4.3 Astrophysics

Astrophysics is a major field to produce large volumes of data. For instance, the **Square Kilometer Array** will produce 9 Tbits/s of raw data. One of the scientific projects related to this instrument called Evolutionary Map of the Universe is working on more than 100 TB of images. The **Euclid Imaging Consortium** will generate 1 PB data per year.

The **SKA project** is an international effort to build and operate the world's largest radiotelescopes covering all together the wide frequency range between 50 MHz and 15.4 GHz. The scale of the SKA project represents a huge leap forward in both engineering and research & development towards building and delivering a unique Observatory, whose construction has officially started on July 2021. The SKA Observatory is the second intergovernmental organisation for ground-based astronomy in the world, after the European Southern Observatory. AVALON participates to the activities of the SCOOP team in SKAO's SAFe framework that deals with platforms related issues such as application benchmarking and profiling, hardware-software co-design.

4.4 Bioinformatics

Large-scale data management is certainly one of the most important applications of distributed systems in the future. Bioinformatics is a field producing such kinds of applications. For example, DNA sequencing applications make use of MapReduce skeletons.

The AVALON team is a member of **BioSyL**, a Federative Research Structure attached to University of Lyon. It gathers about 50 local research teams working on systems biology. AVALON is in particular collaborating with the Inria **BioTiC** team on artificial evolution and computational biology as the challenges are around high performance computation and data management.

5 Social and environmental responsibility

5.1 Footprint of research activities

Through its research activities on energy efficiency and on energy and environmental impacts reductions, Avalon tries to reduce some impacts of distributed systems.

Avalon deals with frugality in clouds with the leadership of the FrugalCloud challenge (*Défi*) between Inria and OVHcloud. Laurent Lefevre is also involved in the steering committee of the **EcoInfo** GDS CRNS group which deals with eco-responsibility of ICT. Avalon is also involved in the sustainable management of large scale experimental infrastructures like Slices. Laurent Lefevre has proposed a Green Slices methodology which is under review in order to deal with the life cycle of such infrastructures. Laurent Lefevre and Emeline Pegon are strongly involved in the Alt-Impact programme on digital sufficiency, between Ademe, CNRS and Inria.

6 Highlights of the year

6.1 Awards

- Best Paper Award of MASCOT2025 conference for: Vladimir Ostapenco, Loïc Guégan, Salma Tofaily, Issam Raïs and Laurent Lefevre. "CPU Frequency Aware Power Modeling for IoT Edge Nodes", MASCOT2025: 33rd International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication System, Paris, France, October 21-23, 2025.
- Best Presentation Award during ICPADS2025 conference for: Maxime Agusti, Eddy Caron, Benjamin Fichel, Laurent Lefèvre, Olivier Nicol, Anne-Cécile Orgerie. "PEM-BM: Portable Power Estimation Methodology for Bare Metal Servers", ICPADS 2025: The 31st International Conference on Parallel and Distributed Systems, Hefei, China, December 14-18, 2025.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 Concerto

Name: Concerto

Keywords: Reconfiguration, Distributed Software, Component models, Dynamic software architecture

Functional Description: Concerto is an implementation of the formal model Concerto written in Python. Concerto allows:

1. the description of the life cycle and the dependencies of software components,
2. the description of a component assembly that forms the overall life cycle of a distributed software,
3. the automatic reconfiguration of a Concerto assembly of components by using a set of reconfiguration instructions as well as a formal operational semantics.

URL: <https://gitlab.inria.fr/VerDi-project/concerto>

Publications: [hal-03103714](#), [hal-02535077](#), [hal-01897803](#)

Contact: H el ene Coullon

Participants: Christian Perez, 3 anonymous participants

Partners: IMT Atlantique, LS2N, LIP

7.1.2 execo

Keywords: Toolbox, Deployment, Orchestration, Python

Functional Description: Execo offers a Python API for asynchronous control of local or remote, standalone or parallel, unix processes. It is especially well suited for quickly and easily scripting workflows of parallel/distributed operations on local or remote hosts: automate a scientific workflow, conduct computer science experiments, perform automated tests, etc. The core python package is execo. The execo_g5k package provides a set of tools and extensions for the Grid5000 testbed. The execo_engine package provides tools to ease the development of computer sciences experiments.

Release Contributions: Release 2.8.1 on October 21, 2024 (list of changes since version 2.7: adapt to changes in oarstat output format (compatibility with old and new output formats), g5k api cache stored as json instead of pickle, support clusters names ending with numbers (eg. abacus-X), canonical_host_name handles interface / kavlan / ipv6 and support cluster names ending with numbers + fix for ifname != ethX (eg. fpgaX), add get_host_interface, extend planning API to allow requests at node level additionally to cluster and site level, spawn process lifecycle handlers in separate threads to avoid blocking + refactoring, handle encoding (py3+) when writing to Process, full redesign of the Processes expect implementation, add get_cluster_queues and get_cluster_jobtypes, add KaconsoleProcess, add substitutions to filenames in stdout/stderr handlers, scp commands in Get / Put as list and shell=False to (securely) handle spaces in path, fix eating 100% of one core iterating through high number of fd to close them in conductor, fix various regexes with invalid escape sequences, add option in execo_engine for using pty to copy_outputs(), fix corner case in process args handling in Remote)

URL: <https://gitlab.inria.fr/mimbert/execo>

Contact: Matthieu Imbert

Participant: 3 anonymous participants

7.1.3 Halley

Name: Halley

Keywords: Software Components, HPC

Scientific Description: Halley is an implementation of the COMET component model that enables to efficiently compose independent parallel codes using both classical use/provide ports but also dataflow oriented ports that are used to generate tasks for multi-core shared-memory machines.

Functional Description: Halley transforms a COMET assembly into a L2C assembly that contains some special components that deal with the data flow section. In particular, a dataflow section of COMET generates a "scheduler" L2C component that contains the code that is in charged of creating its tasks.

Release Contributions: In 2025, support for StarPU, a unified execution system for heterogeneous multicore architectures, as the target execution environment for COMET began.

Publications: [tel-01663718](#), [hal-01518730](#), [hal-01566288](#), [hal-01901806](#)

Contact: Christian Perez

Participants: Christian Perez, Jerry Lacmou Zeutouo, an anonymous participant

7.1.4 SkyDSoft

Name: SkyData Prototype

Keywords: Distributed systems, Multi-Agents System, Data integration

Functional Description: SkyDSoft is build to conduct experiments for autonomous data (SKD). You can build your first class of autonomous SkyData agents. You can implement behaviours for Agents: add smart, useful behaviours to your agents. You can customize Your own Harbour. A graphical frontend is also given.

Publication: [hal-04040588](#)

Contact: Eddy Caron

Participant: 5 anonymous participants

7.1.5 XKBLAS

Name: XKBLAS

Keywords: BLAS, Dense linear algebra, GPU

Functional Description: XKBLAS is yet an other BLAS library (Basic Linear Algebra Subroutines) that targets multi-GPUs architecture thanks to the XKaapi runtime and with block algorithms from PLASMA library. XKBLAS is able to exploit large multi-GPUs node with sustained high level of performance. The library offers a wrapper library able to capture calls to BLAS (C or Fortran). The internal API is based on asynchronous invocations in order to enable overlapping between communication by computation and also to better composed sequences of calls to BLAS.

This current version of XKBlas is the first public version and contains only BLAS level 3 algorithms, including XGEMMT:

XGEMM XGEMMT: see MKL GEMMT interface XTRSM XTRMM XSYMM XSYRK XSYR2K XHEMM XHERK XHER2K

For classical precision Z, C, D, S.

Release Contributions: 0.1.x versions: calls to BLAS kernels must be initiated by the same thread that initializes the XKBlas library. 0.2.x versions: better support for libblas_wrapper and improved scheduling heuristic to take into account memory hierarchy between GPUs. 0.4.x versions: add support for AMD GPU. 0.5.x : better support for AMD GPU (MI250x). Add capacity to clustering GPUs and CPU threads. 0.6.x : support for APU GraceHopper and AMD MI300A

News of the Year: New development to support APU : NVidia GraceHopper and AMD MI300A to exploit their capabilities to share memory between CPU and GPU.

URL: <https://gitlab.inria.fr/xkblas/versions>

Contact: Thierry Gautier

Participant: 2 anonymous participants

7.2 New platforms

7.2.1 Platform: Grid'5000

Participants: Simon Delamare, Pierre Jacquot, Matthieu Imbert, Laurent Lefèvre, Christian Perez, Jean-Camille Seck, Cyril Devaux.

Functional Description: The Grid'5000 experimental platform is a scientific instrument to support computer science research related to distributed systems, including parallel processing, high performance computing, cloud computing, operating systems, peer-to-peer systems and networks. It is distributed on 10 sites in France and Luxembourg, including Lyon. Grid'5000 is a unique platform as it offers to researchers many and varied hardware resources and a complete software stack to conduct complex experiments, ensure reproducibility and ease understanding of results.

- Contact: Laurent Lefèvre
- URL: www.grid5000.fr/

7.2.2 Platform: SLICES-FR

Participants: Simon Delamare, Pierre Jacquot, Matthieu Imbert, Laurent Lefèvre, Christian Perez, Jean-Camille Seck, Cyril Devaux.

Functional Description: The SLICES-FR infrastructure aims at providing an experimental platform for experimental computer Science (Internet of things, clouds, HPC, big data, *etc.*). This new infrastructure will supersede two existing infrastructures, Grid'5000 and FIT.

- Contact: Christian Perez
- URL: www.slices-fr.eu/

7.2.3 Platform: SLICES RI

Participants: Simon Delamare, Pierre Jacquot, Laurent Lefèvre, Christian Perez, Brice-Edine Bellon.

Functional Description: SLICES RI is an European effort that aims at providing a flexible research infrastructure designed to support large-scale, experimental research focused on networking protocols, radio technologies, services, data collection, parallel and distributed computing and in particular cloud and edge-based computing architectures and services [20]. SLICES-FR is the french node of SLICES RI.

- Contact: Christian Perez
- URL: www.slices-ri.eu

8 New results

8.1 Energy Efficiency in Large Scale Distributed Systems

8.1.1 Estimating the power consumption of bare metal water-cooled servers

Participants: Maxime Agusti, Eddy Caron, Laurent Lefèvre.

In an effort to raise awareness on the increasing carbon emissions of Cloud computing, the European Corporate Sustainability Reporting Directive effectively requires providers to supply their customers with an assessment of the carbon impact associated with their use. This represents a challenge for bare metal servers, where the deployment of dedicated power meters is often unfeasible at scale. To address this, we design PPEM-BM, a novel sensor-driven modeling approach to estimate the power consumption of bare metal servers using CPU temperature data acquired via IPMI. PPEM-BM enhances and generalizes the existing POWERHEAT method, which correlates CPU temperature with power. Our methodology involves training individual power models, performing cross-evaluation to determine their portability, and then using a Learning to Rank (LTR) model to select the most appropriate pre-trained model for a target server based on its hardware configuration and CPU temperature statistics. An experiment conducted on 1,076 production servers at OVHcloud shows that PPEM-BM demonstrates a significant improvement compared to models based solely on hardware profiles. The approach offers a practical, scalable, and cost-effective solution for hosting providers to monitor energy consumption without widespread sensor deployment.

This result received the Best Presentation Award from the ICPADS2025 conference [8].

8.1.2 CPU Frequency Aware Power Modeling for IoT Edge Nodes

Participants: Laurent Lefèvre.

The Internet of Things (IoT) is used for various domains such as monitoring the environment, health care, and smart cities. Monitoring and measuring energy consumption of these systems is a crucial step in making them energy efficient. External Hardware-based power monitoring is not always available for IoT edge nodes. An alternative is to create an accurate power model that relates easy-to-monitor parameters (e.g., instructions count, cache misses, node temperature, etc) to externally monitored power. This relationship helps to estimate the power drawn by the nodes.

IoT edge nodes have several power optimization leverages like Dynamic Voltage and Frequency Scaling (DVFS). When models calibration does not consider these leverages, the gap between power estimation and actual power usage increases.

In related works, several power models and corresponding Software-defined power meters do not consider CPU frequency on IoT edge nodes. These Software-defined power meters provide regression-based power models for IoT edge nodes. This work compares predictions made by these state of the art power models to accurate external power monitoring. We show that not considering CPU frequency can result in incorrect estimations. We investigate and compare several methodologies for building power models, considering the CPU frequency, power, and energy leverage. Different performance metrics and regression methods are explored to estimate power usage. We demonstrate that linear and polynomial regression-based models are able to account for various CPU frequencies on IoT edge nodes. Using these models, we can predict the power consumed by IoT edge nodes running a specific workload, with a MAPE of 2% compared to accurate Hardware-based power meters.

This result is a joint work with the University of Tromsø (UiT) from Norway as part of the PHC Aurora Project with University of Tromsø (Norway) on "Exploring energy monitoring and leveraging energy

efficiency on end-to-end worst edge-fog-cloud continuum for extreme climate environments observatories". These result received the Best Award during the MASCOT2025 conference[10]

8.1.3 Revisiting virtual machine consolidation to save resources and energy in heterogeneous production cloud infrastructures

Participants: Eddy Caron, Laurent Lefèvre, Simon Lambert.

Due to some overprovisioning policies and variable usage, data centers in production can face low average resource utilization. This can result in a waste of underused servers and energy. In this context, virtual machine (VM) consolidation combined with shutdown policies can be a pertinent approach for improving resource utilization and reducing energy consumption of the entire cloud infrastructure. However, VM consolidation requires expensive migration techniques, which can potentially affect performance. Consolidation of workload has been proposed and studied as a core capability since the invention of the Cloud. But after two decades of deployment of Cloud infrastructures, VM consolidation is still rarely used in production for small and large-scale environments. In this article, we explore and revisit the potential of savings that can be achieved through a versatile and efficient Virtual Machine consolidation in small and large-scale production infrastructures through usage analysis of two Cloud providers infrastructures. We show that potential benefits in terms of saved cloud resources and energy usage reduction can occur for systems in production [18, 6, 5].

8.1.4 Estimating the environmental impact of Generative-AI services

Participants: Eddy Caron, Laurent Lefèvre.

Generative AI (Gen-AI) represents a major growth potential for the digital industry, a new stage in digital transformation through its many applications. Unfortunately, by accelerating the growth of digital technology, Gen-AI is contributing to the significant and multiple environmental damage caused by its sector. The question of the sustainability of IT must include this new technology and its applications, by measuring its environmental impact. To best respond to this challenge, we propose various ways of improving the measurement of Gen-AI's environmental impact. Whether using life-cycle analysis methods or direct measurement experiments, we illustrate our methods by studying Stable Diffusion a Gen-AI image generation available as a service. By calculating the full environmental costs of this Gen-AI service from end to end, we broaden our view of the impact of these technologies. We show that Gen-AI, as a service, generates an impact through the use of numerous user terminals and networks. We also show that decarbonizing the sources of electricity for these services will not be enough to solve the problem of their sustainability, due to their consumption of energy and rare metals. This consumption will inevitably raise the question of feasibility in a world of finite resources. We therefore propose our methodology as a means of measuring the impact of Gen-AI in advance. Such estimates will provide valuable data for discussing the sustainability or otherwise of Gen-AI solutions in a more transparent and comprehensive way [4]. This result is a joint work explored during the PhD of Adrien Berthelot co-advised by Laurent Lefevre and Eddy Caron and during the PhD of Mathilde Jay co-advised by Laurent Lefevre and Denis Trystram (UGA).

8.1.5 Placing leverages on Clouds for footprint reduction

Participants: Thomas Stavis, Laurent Lefèvre.

Data centers have significant environmental impacts, including resource depletion, carbon emissions, and high energy consumption. Because of their size and complexity, controlling these impacts is both challenging

and crucial to make them more sustainable. Diverse techniques called leverages are used to manage and change behaviors of data centers, but combining many leverages simultaneously remains difficult because of their high number and heterogeneity. In this work, we address a modeling of leverage placement and a methodology for strategically managing leverages towards impact reduction. To guarantee effectiveness and practical applicability, this approach takes into consideration data center architecture, the operational behavior of equipment and leverages, and sustainability goals defined by provider expertise. Preliminary results from well-established scenarios demonstrate the effectiveness of this method in reducing power consumption, enhancing management of leverages, and resolving scalability issues due to the size of data centers and the number of leverages[19, 12].

8.1.6 Analyzing the Full Life Cycle of IoT-Based 5G Solutions for Smart Agriculture

Participants: Egreteau-Druet Emile, Doreid Ammar, Laurent Lefèvre.

Agriculture, as a crucial production sector, has significant environmental impacts. In the context of an environmental crisis, those negative impacts must be mitigated. Therefore, Internet of Things (IoT) technologies combined with AI are often promoted to reduce those impacts and encourage agroecological practices. However, the positive effects of IoT technologies could be balanced by their own negative impacts, potentially leading to a rebound effect. Indeed, IoT environmental impacts are still not well understood. Three use cases were defined, and a life cycle analysis based on these use cases will be conducted to contribute to a better understanding of these impacts[17].

8.2 Edge, Cloud and Distributed Resource Management

8.2.1 SkyData: Autonomous Data paradigm

Participants: Eddy Caron, Elise Jeanneau, Laurent Lefèvre, Christian Perez.

With the rise of Data as a Service, companies understood that whoever controls the data has the power. The past few years have exposed some of the weaknesses of traditional data management systems. For example, application owner can collect and use data to their own advantage without the user's consent. We defined the SkyData concept, which revolves around autonomous data evolving in a distributed system. This new paradigm is a complete break from traditional data management systems. This paradigm is born from the many issues associated with traditional data management systems, such as resells or private information collected without consent, for example. Self managed data, or SKDs, are agents endowed with data, capabilities and goals to achieve. They are free to behave as they wish and try to accomplish their goals as efficiently as possible. They use learning algorithms to improve their decision making and learn new capabilities and services. We introduced how SKDs could be developed and provided some insight on useful capabilities.

In 2025, we explored algorithms that reach a trade-off between data autonomy and cohesion in a given subset of replicas. We propose a deterministic algorithm that ensures cohesion under a costly assumption on the number of failures. Alternatively, we investigate the use of an eventual leader election mechanism in a probabilistic algorithm where a designated leader manages coordination and acts as a communication relay. Compared to the naïve approach of broadcasting new positions to all replicas after each migration, we show experimentally that this approach reduces the loss of cohesion in most scenarios, even without assumption on the number of failures.

8.2.2 Numerics in the Cloud

We have defined some guidelines for critical applications to reduce the arithmetic numerical issue and we provide additional guidelines dedicated to Cloud platform.

Using a simple experiment we shown how the result of a floating-point computation can be affected when the program is compiled and executed in different environments (different processors, with different floating-point extensions and different compiler options), which is to be expected when running applications on a Cloud. Our example is simply based on floating-point summation, which is well known to be “not as easy to compute accurately as it seems” in the literature. However, this experiment is really meant to illustrate the difficulty to guarantee reproducible results, but not to exhibit real accuracy problems. With some care, porting numerical software from a micro-controller to the Cloud, or directly writing applications to the Cloud, can be achieved provided certain recommendations we have defined are followed.

We built an automated, flexible testing framework designed to evaluate the numerical stability of an application across diverse configurations and Cloud platforms. By leveraging advanced DevOps practices, this environment enables:

- The evaluation of numerical stability under varying hardware and software setups, and deployment methods (containerized or native).
- Among the many available Cloud providers, this report focuses on cross-Cloud consistency tests conducted on Grid’5000, AWS, and Azure as a proof of concept.
- The preliminary work for cost-performance analyses to identify optimal Cloud platforms.

We created a scalable pipeline was implemented using tools such as Jenkins, Terraform, Ansible, Docker, and GitLab. The framework supports detailed configuration tests and generates structured outputs for further numerical and cost-effectiveness evaluations. This testing environment forms the backbone of future analytical efforts, enabling accurate and reproducible results across multiple configurations and Cloud environments.

8.2.3 A specialized model and implementation of an actuarial chatbot based on Federated Learning

In this work we focused on developing a language model specialized in the actuarial domain using modern machine learning techniques, including federated learning (FL). The primary objective is to create a chatbot based on large language models (LLMs) capable of meeting actuaries’ specific needs in risk modeling, financial forecasting, and other technical tasks.

Training language models is a crucial process where parameters like batch size, learning rate, and optimizers are adjusted to improve performance. Specific fine-tuning techniques, like instruction-based adaptation and alignment with human preferences via Reinforcement Learning from Human Feedback (RLHF), are employed to tailor LLMs to the actuarial field.

In the context of this work, actuarial schools and other financial institutions become clients of the federated system, contributing to the training of a decentralized actuarial model. The central server aggregates updates from local models trained on each client’s private data, ensuring data remains secure throughout the process. To optimize performance and reduce communication costs between clients and the server, techniques for compressing updates are applied, including sketched and structured updates.

In this exploratory work, we have gained expertise in LLMs and federated learning.

8.3 HPC Applications and Runtimes

8.3.1 Measuring and interpreting performances of HPC applications with dependent tasks

Participants: Thierry Gautier, Romain Pereira.

Breaking down the parallel time into work, idleness, and overheads is crucial for assessing the performance of HPC applications but is challenging to measure in runtime systems with dependent tasks. No existing tools allow its measurement accurately. In [7, 11], we introduce POT: a tool-suite for parallel applications performance analysis with support for dependent tasks. We focus on its low-disturbance methodology consisting of parallel object modeling, discrete-event tracing, and post-mortem simulation-based analysis. The POT tool-suite allows the tracing and analysis of OMPT (OpenMP), PMPI (MPI) and pthreads events.

The paper evaluates the accuracy of POT’s analysis on LLVM and MPC-OMP implementations. It shows that measurement bias may be neglected above workload per task, portably across two architectures and OpenMP runtime systems. We also illustrate the benefits unveiled by POT post-mortem simulation approach for analyzing mixed programming models with MPI+OpenMP.

8.3.2 Handling dynamicity of HPC applications designed by a task-based component model

Participants: Jerry Lacmou Zeutouo, Christian Perez, Thierry Gautier, Romain Pereira.

We extended the COMET component model with the support of dynamic dependencies in its data-flow model. From a meta-task based data-flow, the COMET compiler generates the OpenMP code that will submit the tasks as well as the associated dependencies. The limitation of COMET was that these dependencies were to be known when submitting the tasks of all the data-flow. Hence, a task could not depend on a value compute by another task. We have extended the COMET model with the support of dynamic dependencies and have modified accordingly its runtime. A major difficulty was to generate the code that handle those dynamically-known dependencies under HPC constraints. We evaluated the relevance and performance of three models of dependencies (flat, nested, and weak dependencies) provided by OpenMP related runtimes (LLVM, MPC, and OmpSs-2).

8.3.3 Performance portable batched linear algebra kernels for transport sweeps using Kokkos

Participants: Thierry Gautier, Gabriel Suau.

The paper [13] describes the development of performance portable batched linear algebra kernels for SN-DG neutron transport sweeps using Kokkos. We establish a new sweep algorithm for GPUs that relies on batched linear algebra kernels. We implement an optimized batched gesv solver for small linear systems that builds upon state-of-the-art algorithms. Our implementation achieves high performance by minimizing global memory traffic and maximizing the amount of computations done at compile-time. We assess the performance of the batched gesv kernel on NVIDIA and AMD GPUs. We show that our custom implementation outperforms state-of-the-art linear algebra libraries on these architectures. The performance of the new GPU sweep implementation is assessed on the H100 and MI300A GPUs. We demonstrate that it is able to achieve high performance on both architectures, and is competitive with an optimized multithreaded CPU implementation on a 128-core AMD Genoa CPU node.

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

Participants: Eddy Caron, Thierry Gautier, Laurent Lefevre, Adrien Berthelot, Simon Lambert.

Bosch We have a collaboration with Bosch and AriC (a research team of the LIP laboratory, jointly supported by CNRS, ENS de Lyon, Inria and Université Claude Bernard – Lyon 1). We conducted a study to provide guidelines for writing portable floating-point software in Cloud environments. With some care, porting numerical software from a micro-controller to the Cloud, or directly writing applications to the Cloud, can be eased with the help of some recommendations. It is in fact not more difficult than porting software from a micro-controller to any general-purpose processor.

CEA We have a collaboration with CEA INSTN/SFRES / Saclay. This collaboration is based on the co-advising of a CEA PhD. The research of the PhD student (Gabriel Suau) focuses on high performance codes for neutron transport. One of the goal of the PhD is to work on better integration of Kokkos with a task based model.

Octo technology We have a collaboration with Octo Technology (Part of Accenture). This collaboration is sealed through a CIFRE PhD grant. The research of the PhD student (Adrien Berthelot) focuses on accelerated and driven evaluation of the environmental impacts of an Information System with the full set of digital services

SynAApps We have a collaboration with SynAApps (part of Cyril Group). This collaboration is sealed through a CIFRE PhD grant. The research of the PhD student (Simon Lambert) focuses on forecast and dynamic resource provisioning on a virtualization infrastructure.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

JLESC

Participants: Thierry Gautier, Christian Perez.

Title: Joint Laboratory for Extreme Scale Computing

Partner Institutions: NCSA (US), ANL (US), Inria (FR), Jülich Supercomputing Centre (DE), BSC (SP), Riken (JP).

Date/Duration: 2014-

Summary: The purpose of the Joint Laboratory for Extreme Scale Computing (JLESC) is to be an international, virtual organization whose goal is to enhance the ability of member organizations and investigators to make the bridge between Petascale and Extreme computing. The founding partners of the JLESC are INRIA and UIUC. Further members are ANL, BSC, JSC and R-CCS. UTK is a research member. JLESC involves computer scientists, engineers and scientists from other disciplines as well as from industry, to ensure that the research facilitated by the Laboratory addresses science and engineering's most critical needs and takes advantage of the continuing evolution of computing technologies.

SKA

Participants: Anass Serhani, Laurent Lefevre, Christian Perez, Basile Leretaille.

Title: Square Kilometer Array Organization(SKA)

Summary: The AVALON team collaborates with SKA Organization (an IGO) whose mission is to build and operate cutting-edge radio telescopes to transform our understanding of the Universe, and deliver benefits to society through global collaboration and innovation.

10.2 European initiatives

10.2.1 Horizon Europe

SLICES-PP

Participants: Christian Perez, Laurent Lefevre, Pierre Jacquot.

Title: Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies - Preparatory Phase

Duration: From September 1, 2022 to December 31, 2025

Partners:

- Institut National de Recherche en Informatique et Automatique (INRIA), France
- Sorbonne Université (SU), France
- Universiteit van Amsterdam (UvA)
- Netherlands University of Thessaly (UTH), Greece
- Consiglio Nazionale delle Ricerche (CNR), Italy
- Instytut Chemii Bioorganicznej Polskiej Nauk (PSNC), Poland
- Mandat International (MI), Switzerland
- IoT Lab (IoTLAB), Switzerland
- Universidad Carlos III de Madrid (UC3M), Spain
- Interuniversitair Micro-Electronica Centrum (IMEC), Belgium
- UCLan Cyprus (UCLAN), Cyprus
- EURECOM, France
- Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI), Hungary
- Consorzio Interuniversitario Nazionale per l'Informatica (CINI), Italy
- Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy
- Université du Luxembourg (Uni.Lu), Luxembourg
- Technical Universitaet Muenchen (TUM), Germany
- Euskal Herriko Unibertsitatea (EHU), Spain
- Kungliga Tekniska Hoegskolan (KTH), Sweden
- Oulun Yliopisto (UOULU), Finland
- EBOS Technologies Ltd (EBOS), Cyprus
- Simula Research Laboratory AS (SIMULA), Norway
- Centre National de la Recherche Scientifique (CNRS), France
- Institut Mines-Télécom (IMT), France
- Université de Geneve (UniGe), Switzerland

Inria contact: Nathalie Mitton

Coordinator: Nathalie Mitton

Summary: The digital infrastructures research community continues to face numerous new challenges towards the design of the Next Generation Internet. This is an extremely complex ecosystem encompassing communication, networking, data-management and data-intelligence issues, supported by established and emerging technologies such as IoT, 5/6G, cloud-to-edge computing. Coupled with the enormous amount of data generated and exchanged over the network, this calls for incremental as well as radically new design paradigms. Experimentally-driven research is becoming worldwide a de-facto standard, which has to be supported by large-scale research infrastructures to make results trusted, repeatable and accessible to the research communities. SLICES-RI (Research Infrastructure), which was recently included in the 2021 ESFRI roadmap, aims to answer these problems by building a large infrastructure needed for the experimental research on various aspects of distributed computing, networking, IoT and 5/6G networks. It will provide the resources needed to continuously design, experiment, operate and automate the full lifecycle management of digital infrastructures, data, applications, and services. Based on the two preceding projects within SLICES-RI, SLICES-DS (Design Study) and SLICES-SC (Starting Community), the SLICES-PP (Preparatory Phase) project will validate the requirements to engage into the implementation phase of the RI lifecycle. It will set the policies and decision processes for the governance of SLICES-RI: i.e. the legal and financial frameworks, the business model, the required human resource capacities and training programme. It will also settle the final technical architecture design for implementation. It will engage member states and stakeholders to secure commitment and funding needed for the platform to operate. It will position SLICES as an impactful instrument to support European advanced research, industrial competitiveness and societal impact in the digital era.

ODISSEE

Participants: Christian Perez, Laurent Lefevre, Pierre Jacquot, Brice-Édine Bellon.

Title: Online Data Intensive Solutions for Science in the Exabytes Era

Duration: From January 1, 2025 to December 31, 2027

Partners:

- Institut National de Recherche en Informatique et Automatique (Inria), France
- Grand Equipement National de Calcul Intensif (GENCI), France
- Neovia Innovation (Neovia Innovation), France
- Simula Research Laboratory AS, Norway
- Sipearl, France
- Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
- Next Silicon LTD, Israel
- The Square Kilometre Array Observatory, United Kingdom
- Surf BV, Netherlands
- Eidgenoessische Technische Hochschule Zuerich (ETH Zürich), Switzerland
- Organisation Europeenne pour la Recherche Nucleaire (European Organization for Nuclear Research), Switzerland
- Observatoire de Paris (OBSPARIS), France
- Centre National de la Recherche Scientifique (CNRS), France
- Energy Aware Solutions SL, Spain
- Nextsilicon GMBH, Germany
- Stichting Nederlandse Wetenschappelijk Onderzoek Instituten (NWO-I), Netherlands

- Barcelona Supercomputing Center Centro Nacional de Supercomputacion (BSC CNS), Spain

Inria contact: Christian Perez

Coordinator: Damien Gratadour

Summary: This project federates efforts from 3 pan-European ESFRI infrastructures (HL-LHC, SKAO and SLICES-RI) in physical sciences, Big Data, and in the computing continuum supporting flagship instruments that will maintain and strengthen European leadership in high-energy physics and astronomy. The main goal is to enable key science projects, with the search for Dark Matter serving as a pilot program, combining the complementary capabilities of these three unique research infrastructures. ODISSEE will deliver evolutionary and revolutionary hardware and software platforms to address the corresponding digital challenges in a highly competitive international context. Developed through a joint and comprehensive R&D program with industry partners, as well as access to cutting edge experimental facilities from SLICES-RI, so as to enable HL-LHC and SKA to process and analyze the vast volumes of raw data they produce. Targeting such dataflow driven applications opens the way to a new range of technologies and services, feeding SLICES-RI with a unique yet representative set of specifications to progress their operational & experimental capacities at an unprecedented scale, increasing the dissemination potential. Bringing these 3 infrastructures to their full capacity, as well as operating and maintaining them, pose similar grand challenges across the digital continuum and require addressing the 3 dimensions of sustainability. Co-design and close partnership of academia with European companies will foster competitiveness of European industry and promote digital sovereignty. The project is deeply embedded into both regional and international R&I ecosystems, with strong connections to several major European initiatives and associated partnerships with main technology providers. Strong and lasting impact is built-in the two-fold exploitation strategy including the development of unique in-depth training for R.I. staff and extensive trans-sectoral dissemination.

10.3 National initiatives

Priority Research Programmes and Equipments (PEPR)

PEPR Cloud – Taranis

Participants: Christian Perez, Yves Caniou, Eddy Caron, Elise Jeanneau, Laurent Lefevre, Johanna Desprez, Quentin Quilloteau.

Title: Taranis : Model, Deploy, Orchestrate, and Optimize Cloud Applications and Infrastructure

Partners: Inria, CNRS, IMT, U. Grenoble-Alpes, CEA, U. Rennes, ENSL, U. Lyon I, U. Lille, INSA Lyon, INSA Rennes, Grenoble INP

Date: Sep 2023 – Aug 2030.

Summary: New infrastructures, such as Edge Computing or the Cloud-Edge-IoT computing continuum, make cloud issues more complex as they add new challenges related to resource diversity and heterogeneity (from small sensor to data center/HPC, from low power network to core networks), geographical distribution, as well as increased dynamicity and security needs, all under energy consumption and regulatory constraints.

In order to efficiently exploit new infrastructures, we propose a strategy based on a significant abstraction of the application structure description to further automate application and infrastructure management. Thus, it will be possible to globally optimize the resources used with respect to multi-criteria objectives (price, deadline, performance, energy, etc.) on both the user side (applications) and the provider side (infrastructures). This abstraction also includes the challenges related to the abstraction of application reconfiguration and to automatically adapt the use of resources.

The Taranis project addresses these issues through four scientific work packages, each focusing on a phase of the application lifecycle: application and infrastructure description models, deployment and reconfiguration, orchestration, and optimization.

PEPR Cloud – CareCloud

Participants: Laurent Lefevre, Eddy Caron, Thomas Stavis.

Title: Understanding, improving, reducing the environmental impacts of Cloud Computing

Partners: CNRS, Inria, Univ. Toulouse, IMT

Date: Sept 2023 - Aug 2030

Summary: The CARECloud project (understanding, improving, reducing the environmental impacts of Cloud Computing) aims to drastically reduce the environmental impacts of cloud infrastructures. Cloud infrastructures are becoming more and more complex: both in width, with more and more distributed infrastructures, whose resources are scattered as close as possible to the user (edge, fog, continuum computing) and in depth, with an increasing software stacking between the hardware and the user's application (operating system, virtual machines, containers, orchestrators, micro- services, etc.) The first objective of the project is to understand how these infrastructures consume energy in order to identify sources of waste and to design new models and metrics to qualify energy efficiency. The second objective focuses on the energy efficiency of cloud infrastructures, i.e., optimizing their consumption during the usage phase. In particular, this involves designing resource allocation and energy lever orchestration strategies: mechanisms that optimize energy consumption (sleep modes, dynamic adjustment of the size of virtual resources, optimization of processor frequency, etc.). Finally, the third objective targets digital sobriety in order to sustainably reduce the environmental impact of clouds. Indeed, current clouds offer high availability and very high fault tolerance, at the cost of significant energy expenditure, particularly due to redundancy and oversizing. This third objective aims to design infrastructures that are more energy and IT resource efficient, resilient to electrical intermittency, adaptable to the production of electricity from renewable energy sources and tolerant of the disconnection of a highly decentralized part of the infrastructure.

PEPR Cloud – SILECS

Participants: Simon Delamare, Pierre Jacquot, Laurent Lefevre, Christian Perez.

Title: Super Infrastructure for Large-Scale Experimental Computer Science for Cloud/Edge/IoT

Partners: Inria, CNRS, IMT, U. Lille, INSA Lyon, U. Strasbourg, U. Grenoble-Alpes, Sorbonne U., U. Toulouse, Nantes U., Renater.

Date: Sept 2023 - Aug 2030

Summary: The infrastructure component of the PEPR Cloud (SILECS) will structure the Cloud/Fog/Edge/IoT aspects of the SLICES-FR (Super Infrastructure for Large-Scale Experimental Computer Science) platform, the French node of the ESFRI SLICES-RI action. SILECS will enable the prototyping and conduct of reproducible experiments of any hardware and software element of current and future digital environments at all levels of the Cloud IoT continuum, addressing the experimental needs of the other PEPR components. SILECS will be complemented within SLICES-FR by funding from the PEPR Networks of the Future, which focuses on specific aspects of 5G and beyond technologies. There will therefore be continuous and coordinated strong interactions between the two PEPRs.

PEPR 5G Network of the Future – JEN

Participants: Laurent Lefevre, Doreid Ammar, Emile Egreteau-Druet.

Title: JEN: Network of the Future – Just Enough Networks

Partners: CEA, CNRS, ENSL, ESIEE, IMT, INPB, Inria, INSAL

Date: 2023-2028

Summary: In the NF-JEN project, partners propose to develop just enough networks: network whose dimension, performance, resource usage and energy consumption are just enough to satisfy users' needs. Along with designing energy-efficient and sober networks, we will provide multi-indicators models that could help policy-makers and inform the public debate.

PEPR NumPEX – Exa-Soft

Participants: Thierry Gautier, Christian Perez, Julien Gaupp, Pierre-Etienne Polet, Alix Paigue.

Title: Exa-Soft: HPC software and tools

Partners: Inria, CEA, CNRS, U. Paris-Saclay, Telcom SudParis, Bordeaux INP, ENSIIE, U. Bordeaux, U. Grenoble-Alpes, U. Rennes I, U. Strasbourg, U. Toulouse

Date: 2023-2029

Summary: Though significant efforts have been devoted to the implementation and optimization of several crucial parts of a typical HPC software stack, most HPC experts agree that exascale supercomputers will raise new challenges, mostly because the trend in exascale compute-node hardware is toward heterogeneity and scalability: Compute nodes of future systems will have a combination of regular CPUs and accelerators (typically GPUs), along with a diversity of GPU architectures.

Meeting the needs of complex parallel applications and the requirements of exascale architectures raises numerous challenges which are still left unaddressed.

As a result, several parts of the software stack must evolve to better support these architectures. More importantly, the links between these parts must be strengthened to form a coherent, tightly integrated software suite.

Our project aims at consolidating the exascale software ecosystem by providing a coherent, exascale-ready software stack featuring breakthrough research advances enabled by multidisciplinary collaborations between researchers.

ANR

SkyData

Participants: Eddy Caron, Elise Jeanneau, Laurent Lefèvre, Christian Perez, Maxime Just.

Title: SkyData: A new data paradigm: Intelligent and Autonomous Data

Partners: LIP, VERIMAHG, LIP6

Date: 01.2023-01.2027.

Summary: Nowadays, who controls the data controls the world, or at least the IT world. Usually data are managed through a middleware, but in this project, we propose a new data paradigm without any data manager. We want to endow the data with autonomous behaviors and thus create a new entity, so-called Self-managed data. We plan to develop a distributed and autonomous environment, that we call SKYDATA, where the data are regulated by themselves. This change of paradigm represents a

huge and truly innovative challenge! This goal must be built on the foundation of a strong theoretical study and knowledge on autonomic computing, since Self-managed data will now have to obtain and compute the services they need in autonomy. We also plan to actually develop a SKYDATA framework prototype and a green-IT use case that focuses data energy coconsumption. SKYDATA will be compliant with GDPR through the targeted datas and some internal process.

French Joint Laboratory

ECLAT

Participants: Anass Serhani, Laurent Lefèvre, Christian Perez.

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

Date/Duration: 2023-

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

Inria Large Scale Initiative

FrugalCloud: Défi Inria OVHCloud

Participants: Eddy Caron, Laurent Lefèvre, Christian Perez.

Summary A joint collaboration between Inria and OVH Cloud company on the topic challenge of frugal cloud has been launched in October 2021. It addresses several scientific challenge on the eco-design of cloud frameworks and services for large scale energy and environmental impact reduction. Laurent Lefèvre is the scientific leader of this project. Some AVALON PhD students are involved in this Inria Large Scale Initiative (Défi) : Maxime Agusti and Vladimir Ostanpenco.

Alt-Impact program

Participants: Laurent Lefèvre, Emeline Pegon.

Summary Alt Impact is a program supported by ADEME, CNRS and INRIA, designed to raise public awareness of the environmental impact of digital technology. Our mission is to provide information in a clear and accessible way, with verified, up-to-date and entertaining content. In addition to providing information, we offer practical solutions that are easy to implement on a day-to-day basis, so that everyone, whether an individual or an organization, and whatever their level of knowledge, can take concrete action to reduce their digital ecological footprint.

At the same time, we are pursuing our objective of accelerating and supporting the transition to digital sufficiency, with a focus on measuring and managing it, through the identification and sharing of reliable data and tools, as well as supporting actions aimed at integrating digital sufficiency into the strategies of local authorities and businesses.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Eddy Caron was co Publication Chair of CCGrid 2025 conference: The 25th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, Tromso, Norway, May 19-22, 2025
- Laurent Lefevre was General chair of the First Slices-FR school - Laurent Lefevre, Christian Perez, and Simon Delamare were member of the Organizing Committee of the 2025 SLICES-FR Summer School, Lyon, 7-11 Jul 2025
- Laurent Lefevre was co General Chair of CCGrid 2025 conference: The 25th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, Tromso, Norway, May 19-22, 2025
- Laurent Lefevre was co General Chair of the PAISE 2025: 7th Workshop on Parallel AI and Systems for the Edge , during the IPDPS2025 conference, Milan, Italy, June 3-7, 2025
- Laurent Lefevre was co workshop chair of CloudAM2025 : 14th International Workshop on Cloud and Edge Computing, and Applications Management, Nantes, December 1-4,2025
- Laurent Lefevre was co organizer of the GreenDays2025@Rennes : "Beyond efficiency, how can we imagine a more sufficient digital world?", Rennes, March 25-26, 2025
- Christian Perez was member of the Organizing Committee of 1st workshop on Research Infrastructures for experimenting across the HPC-Cloud-Edge continuum (ContinuumRI), colocated with CCGRID 2025, Tromsø, Norway, 19 May 2025,
- Christian Perez was member of the Organizing Committee of JCAD, the French Journées Calcul Données, Lille, 15-17 Sep 2025.

11.1.2 Scientific events: selection

Member of the conference program committees

- Yves Caniou was member of the Programme Committee of the 25th International Conference on Computational Science and Its Applications.
- Eddy Caron was member of Programme Committee of CloudAM 2025, QUICK'25 and UCC 2025.
- Christian Perez was member of the Programme Committee of CCGRID 2025, ContinuumRI 2025, UCC/BDCAT 2025, and JCAD 2025.

Reviewer

- Eddy Caron was reviewer for CloudAM 2025, ICCS 2025, ICCS 2025, and UCC 2025

11.1.3 Journal

Reviewer - reviewing activities

- Eddy Caron was reviewer for journal of Engineering Applications of Artificial Intelligence.
- Eddy Caron was reviewer for journal of Cloud Computing (Springer Nature)

11.1.4 Invited talks

- Laurent Lefevre gave the opening keynote of the fifth edition of the Complex Days of the Academy "Complex Systems" on "The environmental and human impact of digital technology: when AI contributes to the runaway", Nice, France, February 6, 2025
- Laurent Lefevre gave the invited talk "Le numérique : entre fantastique et coté obscur - Les impacts environnementaux du numérique", (online) Invited Talk, TCHADIA – TCHAD Intelligence Artificielle, Tchad, May 7, 2025
- Christian Perez a keynote talk about SLICES at 25h IEEE International Symposium on Cluster, Cloud, and Internet Computing (CCGRID2025), 19-22 May 2025, Tromsø, Norway.

11.1.5 Scientific expertise

- Yves Caniou was in the selection committee to recruit a new associate professor at Université Côte d'Azur.
- Christian Perez evaluated 8 projects for the French Direction générale de la Recherche et de l'Innovation.

11.1.6 Research administration

- Eddy Caron is a member of the ANR CE25 committee («Sciences et génie du logiciel - Réseaux de communication multi-usages, infrastructures de hautes performances»)
- Eddy Caron is a member of the ASTRID 2025 committee («Accompagnement Spécifique des Travaux de Recherches d'intérêt Défense»)
- Élise Jeanneau is a member of the **Inria Evaluation Committee**.
- Christian Perez represents Inria in the overview board of the France Grilles Scientific Interest Group. He is a member of the executive board and the sites committee of the Grid'5000 Scientific Interest Group and member of the executive board of the SLICES-FR testbed. He is in charge of organizing scientific collaborations between Inria and SKA France.

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

- Licence: Yves Caniou, Algorithmique programmation impérative initiation, 150h, niveau L1, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Algorithmique et programmation récursive, 53h, niveau L1, Université Claude Bernard Lyon 1, France.
- IUT ASPE: Yves Caniou, Initiation Unix, 12h, niveau L1, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Réseaux, 12h, niveau L3, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Programmation Concurrente, 43h and Responsible of UE, niveau L3, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Responsible of alternance students, 21h, niveau M1, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Responsible of alternance students, 7h, niveau M2, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Sécurité Système, 30h and Responsible of UE, niveau M2, Université Claude Bernard Lyon 1, France.
- Master: Eddy Caron, Distributed System, 20h, M1, École Normale Supérieure de Lyon. France.

- Master: Eddy Caron, Langages, concepts et archi pour les données, 30h, M2, ISFA. Université Claude Bernard Lyon 1
- Master: Eddy Caron, Risques dans les Systèmes et Réseaux - Cloud, 15h, M2, ISFA. Université Claude Bernard Lyon 1.
- Master: Eddy Caron, Service Web et Sécurité, 15h, M2, ISFA. Université Claude Bernard Lyon 1.
- Master: Eddy Caron, Data Mining Avancé: Environnements parallèles et distribués, 12h, M2, ISFA. Université Claude Bernard Lyon 1.
- Licence: Élise Jeanneau, Introduction Réseaux et Web, 36h, niveau L1, Université Lyon 1, France.
- Licence: Élise Jeanneau, Réseaux, 53h, niveau L3, Université Lyon 1, France.
- Licence: Élise Jeanneau, Algorithmique, programmation et structures de données, 24h, niveau L2, Université Lyon 1, France
- Licence: Élise Jeanneau, Architecture des ordinateurs, 24h, niveau L2, Université Lyon 1, France
- Licence: Élise Jeanneau, Réseaux, systèmes et sécurité par la pratique, 23h, niveau L3, Université Lyon 1, France
- Master: Élise Jeanneau, Algorithmes distribués, 45h, niveau M1, Université Lyon 1, France.
- Master: Élise Jeanneau, Réseaux, 6h, niveau M1, Université Lyon 1, France.

11.2.1 Supervision

PhD defended:

- Simon Lambert. "Forecast and dynamic resource provisioning on a virtualization infrastructure", 2022, Eddy Caron (PhD advisor, ENS de Lyon, Inria Avalon), Laurent Lefevre (PhD advisor, Inria Avalon), Rémi Grivel (PhD advisor, Ciril Group), defended 10 Oct 2025.

Phd in progress:

- Maxime Agusti. "Observation de plate-formes de co-localisation baremetal, modèles de réduction énergétique et proposition de catalogues", FrugalCloud Inria-OVHCloud collaboration, Feb 2022, Eddy Caron (co-dir.), Benjamin Fichel (co-dir. OVHcloud), Laurent Lefevre (dir.) et Anne-Cécile Orgerie (co-dir. Magellan),
- Émile Egreteau-bruet. "Analyzing full life cycle of IoT based 5G solutions for smart agriculture", 2024, Laurent Lefevre (dir.), Nathalie Mitton (co-dir. FUN) and Doreid Ammar (co-dir.)
- Julien Gaupp, "Composabilité des algorithmes numériques au modèle de programmation", Dec 2025, Christian Perez (co-dir.) and Emmanuel Agullo (co-dir. Concace)
- Maxime Just, "Mise en oeuvre d'une solution distribuée de contrôle de données autonomes et sécurisées avec respect de la vie privée", 2025, Eddy Caron (co-dir.), Olivier Barais (co-dir DiverSE)
- Thomas Stavis. "Replay of environmental leverages in cloud infrastructures and continuums", 2024, Laurent Lefevre (dir.), Anne-Cécile Orgerie (co-dir. Magellan)
- Gabriel Suau. Résolution de l'équation de transport des neutrons sur des architectures massivement parallèles et hétérogènes : application aux géométries hexagonales. Thierry Gautier (dir.), Ansar CALLOO (co-dir. CEA), Romain LE TELLIER (co-dir. CEA), Remi LE BARON (co-dir. CEA).
- Yifei Sun, "Taming Experimentation for Distributed Systems from Testbed to Conduct Tools with Reproducible Guarantee", Jul 2025, Christian Perez (dir.) and Olivier Richard (co-supervisor DataMove)

11.2.2 Juries

- Eddy Caron was PhD reviewer and member of the defense committee of
 - Divi De Lacour. IMT Nantes. "Architecture and security of cooperative autonomous systems", IMT Nantes / Orange. June 16, 2025.
 - Youssouf Faye. "Distributed edge cloud architecture for executing AI based applications". Université Savoie Mont Blanc. December 18, 2025.
- Eddy Caron was HDR member of the defense committee of
 - Carlos Jaime Barrios Hernández. "MultiScale-HPC Hybrid Architectures: Developing Computing Continuum Towards Sustainable Advanced Computing", INSA Lyon. June 6, 2025.
- Laurent Lefevre was PhD reviewer and member of the defense committee of
 - Roblex Nana Tchakouté: "Energy-Aware High Performance Artificial Intelligence: From Measurement and Modeling to Multi-Objective Scheduling", Université Paris Sciences et Lettres, Mines Paris, Paris, December 5, 2025
 - Tristan Coignon: "Empirical Evaluation of the Energy Impact of Large Language Models for Code Generation and Optimization", University of Lille, Lille, November 13, 2025
 - Jorge Andrés Larracochea González : "RADIANCE: A Methodology for Green Software Design", Universidad Zaragoza and University of Pau and Pays de l'Addour, Anglet, February 8, 2025
- Christian Perez was PhD reviewer and member of the defense committee of
 - Marta Bertran FERRER "New approaches for resource management and job scheduling for HEP Grid computing", 25 Jun 2025, Barcelona, Spain,
 - Hugo MONFLEUR "Concern-Oriented MicroService Architecture: Language, Library, Toolbox, and Evaluation", 28 Nov 2025, Lille,
 - Lise Jolicoeur "Towards secure cluster architectures for HPC workflows", 10 Dec 2025, Bordeaux,
 - Khaled ARSALANE "Scalable Data Stream Processing in Heterogenous Environments", 15 Dec 2025, Rennes.

11.2.3 Educational and pedagogical outreach

Yves Caniou has coorganized the Campus du Libre event.

11.3 Popularization

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

- Laurent Lefevre has been interviewed for "IA : le mur de l'énergie", Epsilon Journal, #45, March 2025

11.3.2 Participation in Live events

Laurent Lefevre has performed :

- Laurent Lefevre was invited to the round table "Prendre nos loisirs à la légère ?", Semaine Climat, Marie 1er Arrondissement, Lyon, October 3, 2025
- "Standup on Digital sufficiency", Feu au Lac event, Imhotep Bar, Lyon, Paris, February 13, 2025
- Interview for "The FrugalCloud challenge between Inria and OVHCloud", during AI Action Summit, Laurent Lefevre and Gregory Lebourg, BFM TV, Paris, February 4, 2025

12 Scientific production

12.1 Major publications

- [1] Y. Caniou, E. Caron, A. Kong Win Chang and Y. Robert. ‘Budget-aware scheduling algorithms for scientific workflows with stochastic task weights on IaaS Cloud platforms’. In: *Concurrency and Computation: Practice and Experience* 33.17 (2021), pp. 1–25. URL: <https://hal.inria.fr/hal-03508925>.
- [2] V. Ostapenco, L. Lefèvre, A.-C. Orgerie and B. Fichel. ‘Modeling, evaluating, and orchestrating heterogeneous environmental leverages for large-scale data center management’. In: *International Journal of High Performance Computing Applications* 37.3-4 (2023). DOI: [10.1177/10943420231172978](https://doi.org/10.1177/10943420231172978). URL: <https://hal.science/hal-04047008>.
- [3] M. Rzepka, P. Boryło, M. Assunção, A. Lasoń and L. Lefèvre. ‘SDN-based fog and cloud interplay for stream processing’. In: *Future Generation Computer Systems* 131 (June 2022), pp. 1–17. DOI: [10.1016/j.future.2022.01.006](https://doi.org/10.1016/j.future.2022.01.006). URL: <https://hal.inria.fr/hal-03559874>.

12.2 Publications of the year

International journals

- [4] A. Berthelot, E. Caron, M. Jay and L. Lefèvre. ‘Understanding the environmental impact of generative AI services’. In: *Communications of the ACM Special Issue on Sustainability and Computing* 68.7 (2025), pp. 46–53. DOI: [10.1145/3725984](https://doi.org/10.1145/3725984). URL: <https://hal.science/hal-04920612> (cit. on p. 15).
- [5] S. Lambert, E. Caron, L. Lefèvre and R. Grivel. ‘Consolidation of virtual machines to reduce energy consumption of data centers by using ballooning, sharing and swapping mechanisms’. In: *Future Generation Computer Systems* 174 (Jan. 2026), p. 107968. DOI: [10.1016/j.future.2025.107968](https://doi.org/10.1016/j.future.2025.107968). URL: <https://hal.science/hal-05143704> (cit. on p. 15).
- [6] S. Lambert, V. Ostapenco, L. Lefèvre, E. Caron, A.-C. Orgerie, B. Fichel and R. Grivel. ‘Revisiting virtual machine consolidation to save resources and energy in heterogeneous production cloud infrastructures’. In: *International Journal of High Performance Computing Applications* (18th Dec. 2025). DOI: [10.1177/10943420251399694](https://doi.org/10.1177/10943420251399694). URL: <https://hal.science/hal-05457159> (cit. on p. 15).
- [7] R. Pereira, T. Gautier, A. Roussel and P. Carribault. ‘Measuring and interpreting performances of HPC applications with dependent tasks’. In: *Future Generation Computer Systems* (June 2025), p. 107933. DOI: [10.1016/j.future.2025.107933](https://doi.org/10.1016/j.future.2025.107933). URL: <https://hal.science/hal-05112187> (cit. on p. 17).

International peer-reviewed conferences

- [8] M. Agusti, E. Caron, B. Fichel, L. Lefèvre, O. Nicol and A.-C. Orgerie. ‘PPEM-BM: Portable Power Estimation Methodology for Bare Metal Servers’. In: *ICPADS 2025 - 31st IEEE International Conference on Parallel and Distributed Systems*. ICPADS 2025 - 31st IEEE International Conference on Parallel and Distributed Systems. Hefei, China: IEEE, 2025, pp. 1–8. URL: <https://hal.science/hal-05365793> (cit. on p. 14).
- [9] P. Jacquet, M. Agusti, E. Caron, C. Coti, M. Dias de Assuncao, L. Lefèvre and A.-C. Orgerie. ‘Untangling GPU Power Consumption: Job-Level Inference in Cloud Shared Settings’. In: *EUROSYS 2026 - European Conference on Computer Systems*. Edinbourg, Ecosse, United Kingdom, 2026. DOI: [10.1145/3767295.3769333](https://doi.org/10.1145/3767295.3769333). URL: <https://hal.science/hal-05291033>.
- [10] V. Ostapenco, L. Guegan, S. Tofaily, I. Raïs and L. Lefèvre. ‘CPU Frequency Aware Power Modeling for IoT Edge Nodes’. In: *MASCOT2025: 33rd International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication System*. MASCOT2025: 33rd International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication System. Paris, France, 21st Oct. 2025. URL: <https://inria.hal.science/hal-05345559> (cit. on p. 15).

- [11] R. Pereira, T. Gautier, A. Roussel and P. Carribault. ‘Measuring and Interpreting Dependent Task-based Applications Performances’. In: *Parallel Processing and Applied Mathematics 2024*. 15th International Conference on Parallel Processing & Applied Mathematics - PPAM 2024. Ostrava, Czech Republic, 4th Apr. 2025, pp. 227–241. URL: <https://hal.science/hal-04767262> (cit. on p. 17).

Conferences without proceedings

- [12] T. Stavis, L. Lefèvre and A.-C. Orgerie. ‘Placing leverages in Cloud for footprint reduction’. In: COMPAS 2025 - Conférence francophone d’informatique en Parallélisme, Architecture et Système. Bordeaux, France, 2025. URL: <https://hal.science/hal-05491835> (cit. on p. 16).
- [13] G. Suau, T. Gautier, A. Calloo, R. Baron and R. Le Tellier. ‘Performance portable batched linear algebra kernels for transport sweeps using Kokkos’. In: SC Workshops ’25: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. St Louis, United States: ACM, 15th Nov. 2025, pp. 1147–1158. DOI: [10.1145/3731599.3767491](https://doi.org/10.1145/3731599.3767491). URL: <https://hal.science/hal-05468442> (cit. on p. 18).

Reports & preprints

- [14] A. Berthelot, T. da Silva Barros, L. Lefèvre, A.-L. Ligozat and E. Pegon. *Multi-criteria and multi-stage environmental study of Pl@ntnet service for the year 2024*. Inria Lyon, 8th Jan. 2026. URL: <https://inria.hal.science/hal-05448455>.
- [15] S. Bouveret, A. Bugeau, E. Frenoux, J. Lefevre, L. Lefèvre, A.-L. Ligozat, P. Marquet, A.-C. Orgerie and D. Trystram. *Quiz sur les impacts environnementaux du numérique*. EcoInfo, Feb. 2025, pp. 1–5. URL: <https://hal.science/hal-04960328>.

Other scientific publications

- [16] A. Berthelot, E. Caron, R. de Laage, L. Lefèvre and A. Nicolas. ‘Des serveurs aux services, évaluer l’empreinte environnementale d’un système d’information et de l’ensemble de ses services’. In: Green Days 2025. Rennes, France, 25th Mar. 2025. URL: <https://hal.science/hal-05008615>.
- [17] E. Egreteau-Druet, D. Ammar, L. Lefèvre and N. Mitton. ‘Study scenarios to analyze the Full Life Cycle of IoT-Based Solutions for Smart Agriculture’. In: ICT4S 2025 - 11th International Conference on ICT for Sustainability. Dublin, Ireland, 9th June 2025. URL: <https://hal.science/hal-05491463> (cit. on p. 16).
- [18] S. Lambert, V. Ostapenco, L. Lefèvre, E. Caron, A.-C. Orgerie, B. Fichel and R. Grivel. ‘Revisiting virtual machine consolidation to save resources and energy in heterogeneous production cloud infrastructures’. In: ICT4S 2025 - International Conference on Information and Communications Technology for Sustainability. Dublin, Ireland, 9th June 2025, pp. 1–1. URL: <https://inria.hal.science/hal-05100569> (cit. on p. 15).
- [19] T. Stavis, L. Lefèvre and A.-C. Orgerie. ‘Placing leverages on Clouds for footprint reduction’. In: ICT4S - International Conference on Information and Communications Technology for Sustainability. Dublin, Ireland, 2025. URL: <https://hal.science/hal-05491814> (cit. on p. 16).

Scientific popularization

- [20] S. Fortun, N. Mitton and C. Pérez. ‘Experiment. Innovate. Transform. The future of digital infrastructure starts with SLICES.’ In: *Innovation Platform 24* (Dec. 2025), pp. 16–24. URL: <https://hal.science/hal-05409390> (cit. on p. 13).

12.3 Cited publications

- [21] R. Ge, X. Feng, S. Song, H.-C. Chang, D. Li and K. W. Cameron. ‘PowerPack: Energy Profiling and Analysis of High-Performance Systems and Applications’. In: *IEEE Trans. Parallel Distrib. Syst.* 21.5 (May 2010), pp. 658–671. DOI: [10.1109/TPDS.2009.76](https://doi.org/10.1109/TPDS.2009.76). URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4906989 (cit. on p. 7).

- [22] A. Geist and S. Dosanjh. ‘IESP Exascale Challenge: Co-Design of Architectures and Algorithms’. In: *Int. J. High Perform. Comput. Appl.* 23.4 (Nov. 2009), pp. 401–402. doi: [10.1177/1094342009347766](https://doi.org/10.1177/1094342009347766). URL: <http://dx.doi.org/10.1177/1094342009347766> (cit. on p. 8).
- [23] W. Gropp, S. Huss-Lederman, A. Lumsdaine, E. Lusk, B. Nitzberg, W. Saphir and M. Snir. *MPI: The Complete Reference – The MPI-2 Extensions*. 2nd ed. Vol. 2. ISBN 0-262-57123-4. The MIT Press, Sept. 1998 (cit. on p. 8).
- [24] H. Kimura, T. Imada and M. Sato. ‘Runtime Energy Adaptation with Low-Impact Instrumented Code in a Power-Scalable Cluster System’. In: *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. CCGRID ’10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 378–387 (cit. on p. 7).
- [25] G. Madec. *NEMO ocean engine*. Note du Pole de modélisation 27. ISSN No 1288-1619. France: Institut Pierre-Simon Laplace (IPSL), 2008 (cit. on p. 9).
- [26] OpenACC. *The OpenACC Application Programming Interface*. Version 1.0. Nov. 2011. URL: <http://www.openacc-standard.org> (cit. on p. 8).
- [27] OpenMP Architecture Review Board. *OpenMP Application Program Interface*. Version 3.1. July 2011. URL: <http://www.openmp.org> (cit. on p. 8).
- [28] B. Rountree, D. K. Lownenthal, B. R. de Supinski, M. Schulz, V. W. Freeh and T. Bletsch. ‘Adagio: Making DVS Practical for Complex HPC Applications’. In: *Proceedings of the 23rd international conference on Supercomputing*. ICS ’09. New York, NY, USA: ACM, 2009, pp. 460–469 (cit. on p. 7).
- [29] C. Szyperski. *Component Software - Beyond Object-Oriented Programming*. 2nd ed. Addison-Wesley / ACM Press, 2002, p. 608 (cit. on p. 8).
- [30] S. Valcke. ‘The OASIS3 coupler: a European climate modelling community software’. In: *Geoscientific Model Development* 6 (2013). doi:10.5194/gmd-6-373-2013, pp. 373–388 (cit. on p. 9).