RESEARCH CENTRE

Inria Lyon Centre

IN PARTNERSHIP WITH:

Université Claude Bernard (Lyon 1), Ecole normale supérieure de Lyon, CNRS

2023

ACTIVITY REPORT

Project-Team

AVALON

Algorithms and Software Architectures for Distributed and HPC Platforms

IN COLLABORATION WITH: Laboratoire de l'Informatique du Parallélisme (LIP)

DOMAIN

Networks, Systems and Services, Distributed Computing

THEME

Distributed and High Performance Computing

Inria

Contents

Project-Team AVALON

Creation of the Project-Team: 2014 July 01

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

1 Team members, visitors, external collaborators

Research Scientists

- Christian Perez [Team leader, INRIA, Senior Researcher, HDR]
- Thierry Gautier [INRIA, Researcher, HDR]
- Laurent Lefevre [INRIA, Researcher, HDR]

Faculty Members

- Yves Caniou [UNIV LYON I, Associate Professor]
- Eddy Caron [Université Lyon1, Professor, from Sep 2023, ISFA (Institut de Science Financière et d'Assurances), HDR]
- Eddy Caron [ENS DE LYON, Associate Professor, until Aug 2023, HDR]
- Olivier Glück [UNIV LYON I, Associate Professor]
- Elise Jeanneau [UNIV LYON I, Associate Professor]
- Etienne Mauffret [ENS DE LYON, ATER, from Feb 2023 until Jun 2023]

Post-Doctoral Fellows

- Yasmina Bouziem [INRIA]
- Jerry Lacmou Zeutouo [INRIA]
- Lucien Ndjie Ngale [ENS DE LYON, Post-Doctoral Fellow, from Dec 2023]

PhD Students

- Maxime Agusti [OVHcloud]
- Adrien Berthelot [OCTO TECHNOLOGY, CIFRE]
- Ghoshana Bista [ORANGE LABS, until Jan 2023]
- Hugo Hadjur [AIVANCITY, until Aug 2023]
- Mathilde Jay [Université Grenoble-Alpes]
- Simon Lambert [CIRIL GROUP, CIFRE]
- Lucien Ndjie Ngale [UPJV, until Oct 2023]
- Vladimir Ostapenco [INRIA]
- Romain Pereira [UVSQ, from Nov 2023]
- Romain Pereira [CEA, until Oct 2023]
- Pierre-Etienne Polet [INRIA, from Sep 2023]
- Pierre-Etienne Polet [THALES, until Jun 2023]

Technical Staff

- Brice-Edine Bellon [INRIA, Engineer, until Jan 2023]
- Arthur Chevalier [INRIA, Engineer, until Oct 2023]
- Simon Delamare [CNRS, Engineer]
- Matthieu Imbert [Inria, Engineer]
- Pierre Jacquot [INRIA, Engineer]
- Jean Christophe Mignot [CNRS, Engineer]
- Anass Serhani [INRIA, Engineer, from May 2023]

Interns and Apprentices

- Jules Bonhotal [INRIA, Intern, from May 2023 until Jul 2023]
- Charlène Broutier [ENS DE LYON, Intern, from May 2023 until Jul 2023]
- Sofiane Chogli [INRIA, Intern, from May 2023 until Jul 2023]

Administrative Assistant

- Chrystelle Mouton [INRIA]

External Collaborator

- Doreid Ammar [AIVANCITY]

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 SLIES 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 mantain. 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 used to valide 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 programing 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 Vera C. Rubin Observatory will produce 20 TB of data every night, with the goals of discovering thousands of exoplanets and of uncovering the nature of dark matter and dark energy in the universe. 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 Beagle 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 challenge (*Défi*) betwwen Inria and OVHcloud. Laurent Lefevre is also involved in the steering committe of the EcoInfo GDS CRNS group which deals with eco-responsibility of ICT.

6 Highlights of the year

6.1 Awards

- Best poster award at the conference CLOSER 2023 for the paper entitled *SkyData : Rise of the Data How can the Intelligent and Autonomous Data paradigm become real?* [10] .

6.2 Other highlights

- More than 1200 downloads in less than 10 months for the paper entitled *An experimental comparison of software-based power meters: focus on CPU and GPU* [9] .
- Eddy Caron was promoted *Professeur des Universités* at *Université Claude Bernard Lyon1* and attached to the LIP laboratory from September 2023.

7 New software, platforms, open data

7.1 New software

7.1.1 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 2023, the meta-model has been cleaned and the software has been updated accordingly so as to prepare its evolution. A first evolution that has started to be developed is the support of dynamic dataflow..

Publications: tel-01663718, hal-01518730, hal-01566288, hal-01901806

Contact: Christian Perez

Participants: Jérôme Richard, Christian Perez, Jerry Lacmou Zeutouo

7.1.2 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 versions: calls to BLAS kernels must be initiate by the same thread that initializes the XKBlas library. 0.2 versions: better support for libblas_wrapper and improved scheduling heuristic to take into account memory hierarchy between GPUs 0.4 versions: add support for AMD GPU

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

Contact: Thierry Gautier

Participants: Thierry Gautier, João Vicente Ferreira Lima

7.1.3 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.

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

Contact: Matthieu Imbert

Participants: Florent Chuffart, Laurent Pouilloux, Matthieu Imbert

7.1.4 Kwollect

Keywords: Monitoring, Power monitoring, Energy, Infrastructure software, Sensors

Functional Description: Kwollect is a monitoring framework for IT infrastructures. It focuses on collecting environmental metrics (energy, sensors, etc.) and make them available to users. It is used as the main monitoring service for Grid'5000 users.

Release Contributions: Kwollect now provides job-scale metrics and allows users to insert their own metrics. It has also been updated to support the newest Omegawatt wattmeter devices, the most recent version of Postgresql, Grafana, and Debian, and includes a number of small improvements and fixes.

URL: https://gitlab.inria.fr/grid5000/kwollect

Publication: hal-03236421

Contact: Simon Delamare

7.1.5 IQ Orchestra

Keywords: -, Automatic deployment, Cybersecurity

Functional Description: IQ-orchestra (previously Qirinus-Orchestra) is a meta-modeling software dedicated to the securized deployment of virutalized infrastructures.

It is built around three main paradigmes:

1 - Modelization of a catalog of supported application 2 - A dynamic securized architecture 3 - An automatic virtualized environement Deployment (i.e. Cloud)

The software is strongly modular and uses advanced software engineering tools such as meta-modeling. It will be continuously improved along 3 axes:

* The catalog of supported applications (open source, legacy, internal). * The catalog of security devices (access control, network security, component reinforcement, etc.) * Intelligent functionalities (automatic firewalls configuration, detection of non-secure behaviors, dynamic adaptation, etc.)

Release Contributions: - Microservices Architecture - Multi-cloud support - Terraform export - Update of all old software embedded - Bugs fix

URL: http://www.qirinus.com

Publications: hal-00840734, hal-00840734, tel-01229874

Contact: Eddy Caron

Participants: Eddy Caron, Arthur Chevalier, Arnaud Lefray

Partner: ENS Lyon

7.2 New platforms

7.2.1 Platform: Grid'5000

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

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.

FUNCTIONAL DESCRIPTION The SLICES-FR infrastructure, that was known as SILECS, 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: Site under construction.

7.2.3 Platform: SLICES

Participants: Pierre Jacquot, Laurent Lefèvre, Christian Perez.

FUNCTIONAL DESCRIPTION SLICES is an European effort that aims at providing a flexible platform 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. SLICES-FR is the The French node of SLICES.

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

8 New results

8.1 Energy Efficiency in HPC and Large Scale Distributed Systems

8.1.1 Services Orchestration at the Edge and in the Cloud on Energy-Aware Precision Beekeeping Systems

Participants: Laurent Lefèvre, Doreid Ammar, Hugo Hadjur.

Honey bees have been domesticated by humans for several thousand years and mainly provide honey and pollination, which is fundamental for plant reproduction. Nowadays, the work of beekeepers is constrained by external factors that stress their production (parasites and pesticides, among others). Taking care of large numbers of beehives is time-consuming, so integrating sensors to track their status can drastically simplify the work of beekeepers.

Precision Beekeeping is a subfield of Precision Agriculture. It collects data on colonies of bees using sensors to assist beekeepers and preserve bees. The existing solutions include collecting voluminous data, such as images and sound, from up close or inside a beehive. The collection, transfer, and processing of such data are energy-intensive processes. However, most Precision Beekeeping systems work under a limited energy budget. Thus, such systems must take this constraint into account. This research focuses on the placement of energy-aware Precision Beekeeping services at the edge and in the cloud. We deploy a set of smart beehives in order to conduct experiments and collect real data. Then, we simulate large-scale systems with embedded services to add more realism to the experimental study. Our results show that parameters like the number of smart beehives, the frequency of data collection, and the characteristics of the cloud server impact the placement of these services to enhance energy efficiency. Some experimental parts of this work occur in the CPER LECO/GreenCube project and some parts are financially supported by aivancity School for Technology, Business & Society Paris-Cachan. This work, within the the Ph.D. of Hugo Hadjur is co-advised by Doreid Ammar (Academic Dean and Professor at aivancity School for Technology, Business & Society Paris-Cachan and external member of Avalon team) and Laurent Lefevre [8].

8.1.2 Ecodesign of large scale distributed applications

Participants: Laurent Lefèvre.

Creating energy aware application with limited environnemtal impacts needs a complete redesign. Laurent Lefevre with some colleagues from the GDS EcoInfo group have explored the various facets of ecodesign. This has resulted to a new version of a brochure available for software developers. This brochure [18] has been downloaded several thousands of times since the publication of the first version.

8.1.3 Environmental assessment of projects involving AI methods

Participants: Laurent Lefèvre.

With colleagues from the GDS EcoInfo group (Anne-Laure Ligozat, Dernis Trystram) we explore criteria for assessing the environmental impacts of responses to calls for projects involving Artificial Intelligence (AI) methods. When proposing these criteria, we take into account, in addition to the general impacts of digital services, the specificities of the AI field and in particular of machine learning: impacts of the learning and inference phases and data collection [17].

8.1.4 Memory over-allocation mechanisms for virtual machine consolidation

Participants: Eddy Caron, Laurent Lefèvre, Simon Lambert.

Efficient resource management is a challenge for Cloud service providers. They need to respond to growing demand by limiting the oversizing of their infrastructures, which can generate avoidable costs and environmental impact. Mechanisms derived from virtualization are used to optimize infrastructure sizing, such as elasticity and consolidation, but economic or physical constraints can hinder their adoption. In this work, we propose a consolidation method based on the observation of the resource consumption of virtual machines (VMs) in the infrastructure of a private Cloud host. With a suitable evaluation algorithm and using memory over-allocation mechanisms, we are able to generate sets of VMs positioned in such a way as to limit the number of physical machines used, while ensuring the quality of service required by users [14].

8.1.5 Environmental Impact of HTTP Requests

Participants: Mathilde Jay, Laurent Lefèvre.

EcoIndex has been proposed to evaluate the absolute environmental performance of a given URL using a score ranging from 0 to 100 (the higher, the better). In this article, we make a critical analysis of the initial approach and propose alternatives that no longer calculate a plain score but allow the query to be situated among other queries. The generalized critiques come with statistics and rely on extensive experiments (first contribution). Then, we move on to low-cost Machine Learning (ML) approaches (second contribution) and a transition before obtaining our final results (third contribution). Our research aims to extend the initial idea of analytical computation, i.e., a relation between three variables, in the direction of algorithmic ML computations. The fourth contribution corresponds to a discussion on our implementation, available on a GitHub repository. Beyond computational questions, it is important for the scientific community to focus on this question in particular. We currently promote using well established ML techniques because of their potential. However, we also question techniques for their frugality or otherwise. Our data science project is still at the data exploration stage. We also want to encourage synergy between technical expertise and business knowledge because this is fundamental for advancing the data project [5].

8.1.6 Estimating the environmental impact of Generative-AI services

Participants: Adrien Berthelot, Eddy Caron, Mathilde Jay, 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 [3].

8.1.7 Comparing software-based power meters dedicated on CPU and GPU

Participants: Mathilde Jay, Laurent Lefèvre, Vladimir Ostapenco.

The global energy demand for digital activities is constantly growing. Computing nodes and cloud services are at the heart of these activities. Understanding their energy consumption is an important step towards reducing it. On one hand, physical power meters are very accurate in measuring energy but they are expensive, difficult to deploy on a large scale, and are not able to provide measurements at the service level. On the other hand, power models and vendor-specific internal interfaces are already available or can be implemented on existing systems. Plenty of tools, called software-based power meters, have been developed around the concepts of power models and internal interfaces, in order to report the power consumption at levels ranging from the whole computing node to applications and services. However, we have found that it can be difficult to choose the right tool for a specific need. In this work, we qualitatively and experimentally compare several software-based power meters able to deal with CPU or GPU-based infrastructures. For this purpose, we evaluate them against high-precision physical power meters while executing various intensive workloads. We extend this empirical study to highlight the strengths and limitations of each software-based power meter. This research is a joint work with Denis Trystram (LIG Laboratory) and Anne-Cécile Orgerie (IRISA, Rennes) [9].

8.1.8 Orchestrating heterogeneous environmental leverages for reducing the impacts of large-scale data centers infrastructures

Participants: Laurent Lefèvre, Vladimir Ostapenco.

Data centers are very energy-intensive facilities that can generate various environmental impacts. Numerous energy, power, and environmental leverages exist and can help cloud providers and data center managers to reduce some of these impacts. But dealing with such heterogeneous leverages can be a challenging task that requires some support from a dedicated framework. This research presents a new approach for modeling, evaluating, and orchestrating a large set of technological and logistical leverages. Our framework is based on leverages modeling and Gantt chart leverages mapping. Some experimental results based on selected scenarios show the pertinence of the proposed approach in terms of management facilities and potential impacts reduction [2]. The research is done inside the challenge (Défi) between Inria and OVHcloud company.

8.2 Edge, Cloud and Distributed Resource Management

8.2.1 Total cost modeling of software ownership in Virtual Network Functions

Participants: Ghoshana Bista, Eddy Caron.

Today, a massive shift is ongoing in telecommunication networks with the emergence of softwarization and cloudification. Among the technologies which are assisting these shifts, one of them is NFV (Network Function Virtualization). NFV is the network architecture that decouples network functions from hardware devices (middleboxes) with the help of a virtual component known as VNF (Virtual Network Function). VNF has shifted the network technological paradigm. Before: Network Function was performed by physical equipment, and service providers acquired its property for the lifetime of the relying hardware (instead counted in years). Today, Network functions are software that service providers develop or acquire purchasing licenses. A license defines software's Right to Use (RTU).

Therefore, if licensing in NFV is not appropriately managed, service providers might (1) be exposed to counterfeiting and risk heavy financial penalties due to non-compliance; (2) might overbuy licenses to cover poorly estimated usages. Thus, mastering network function license through implementing Software Asset Management and FinOps (Finance and DevOps) is essential to control costs. In this research, our primary problem is to minimize the TCO (Total Cost of Ownership) of software cost (VNF), providing Quality of Services (QoS) to a specific amount of users. Software costs include various costs, from development to maintenance, integration to release management, and professional services. Our research focuses on proprietary software (developed by a publisher and sold via a paid license).

We considered that TCO consists of the software license cost, the resources necessary to execute and operate SW, and the energy consumed by this execution. In this research, first, we have identified the need for a standardized VNF licensing model, which is highly dependent on the VNF provider's creativity; This lack of standards places CSPs (Communication Service Providers) at risk of having to delegate the management of rights to suppliers. Hence, we proposed a licensing model based on the metrics, which help to quantify the usage of the VNF. After estimating the license of VNF, we estimated the license cost. Afterward, we presented several ways to minimize the license cost depending upon the different use cases, which depend on the user's scenario and needs. Then after, with the help of industrial knowledge, we found that reducing resource consumption to minimize the TCO providing QoS affects the deployment of the VNF directly or indirectly, which impacts the licensing. Thus, the licenses and resources are interdependent. We used these costs to construct the software's total cost. After that, we proposed several ways to reduce the software's total cost by fulfilling the client's requirements. Then after, we considered the energy and its associated cost of VNF. The energy consumption of the VNF is dependent on resource consumption, and resources usages impact the license. Thus, we can see that these three costs are interdependent: license, resources, and energy cost of VNF. Hence, we consider these costs and constructed TCO. Minimizing TCO fulfilling the client's requirements is challenging since it is a multi-parameter. Therefore, we proposed several heuristical algorithms based on resource sharing and consolidation to reduce the TCO depending on the license, resource preference, and the client's scenarios.

This work [16] was obtained during the PhD of Ghoshana Bista co-advised by Anne-Lucie Vion (Orange) and Eddy Caron.

8.2.2 SkyData: Autonomous Data paradigm

Participants: Eddy Caron, Elise Jeanneau, Laurent Lefèvre, Etienne Mauffret, 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 weakenesses 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 [10] we investigated to show a way to define autonomous data as well as some challenges associated with their specificities. A first version of a SKYDATA Environment is being developed using JADE and JASON to implement SKDs using the Beliefs-Desires-Intentions model with AgentSpeak.

8.3 HPC Applications and Runtimes

8.3.1 Improving Simulations of Task-Based Applications on Complex NUMA Architectures

Participants: Thierry Gautier.

Modeling and simulation are crucial in high-performance computing (HPC), with numerous frameworks developed for distributed computing infrastructures and their applications. Despite node-level simulation of shared-memory systems and task-based parallel applications, existing works overlook non-uniform memory access (NUMA) effects, a critical characteristic of current HPC platforms. In this work, we introduce a modeling for complex NUMA architectures and enhance a simulator for dependency-based task-parallel applications. This facilitates experiments with varied data locality models: we refine a communication-oriented model leveraging topology information for data transfers, and devise a more intricate model incorporating a cache mechanism for last-level cache data storage. Dense linear algebra test cases are used to validate both models, demonstrating that our simulator reliably predicts execution time with minimal relative error. This work [6] was one of the result obtained during the PhD of Idriss Douadi co-advised by Samuel Thibault (Prof. Univ. Bordeaux, EPI STORM) and Thierry Gautier.

8.3.2 Suspending OpenMP Tasks on Asynchronous Events: Extending the Taskwait Construct

Participants: Romain Peirera, Thierry Gautier.

Many-core and heterogeneous architectures now require programmers to compose multiple asynchronous programming model to fully exploit hardware capabilities. As a shared-memory parallel programming model, OpenMP has the responsibility of orchestrating the suspension and progression of asynchronous operations occurring on a compute node, such as MPI communications or CUDA/HIP streams. Yet, specifications only come with the task detach(event) API to suspend tasks until an asynchronous operation is completed, which presents a few drawbacks. In this paper, we introduce the design and implementation of an extension on the taskwait construct to suspend a task until an asynchronous event completion. It aims to reduce runtime costs induced by the current solution, and to provide a standard API to automate portable task suspension solutions. The results show twice less overheads compared to the existing task detach clause. This work [11] was one of the result obtained during the PhD of Romain Peirera co-advised by Adrien Roussel (CEA), Patrick Carribaut (CEA) and Thierry Gautier.

8.3.3 Introducing Moldable Tasks in OpenMP

Participants: Pierre-Etienne Polet, Thierry Gautier.

The paper [13] introduces a new approach to handle implicit parallelism in library functions. If the library already utilizes a thirdparty programming model like OpenMP, it may run in parallel. Otherwise, if the library remains sequential, OpenMP directives in client code cannot be used for direct parallelization. To express implicit parallelism and, in the meanwhile, dynamically adjust the parallel degree of a task when its starts, we propose to use moldable tasks. We handle this by introducing a new construct called taskmoldable that generates multiples tasks from a single function call and an iteration space. For the Lapack Cholesky factorization algorithm, our taskmoldable directive allows simple code annotation to express parallelism between tiles and improves programmability. Performance results on a beamforming application indicates that our moldable implementation is slightly faster by 5% in mean, than a parallel execution achieved with Intel MKL. This work is one of the result of the PhD of Pierre-Etienne Polet co-advised by Ramy Fantar(Thales) and Thierry Gautier

8.3.4 Investigating Dependency Graph Discovery Impact on Task-based MPI+OpenMP Applications Performances

Participants: Romain Peirera, Thierry Gautier.

The architecture of supercomputers is evolving to expose massive parallelism. MPI and OpenMP are widely used in application codes on the largest supercomputers in the world. The community primarily focused on composing MPI with OpenMP before its version 3.0 introduced task-based programming. Recent advances in OpenMP task model and its interoperability with MPI enabled fine model composition and seamless support for asynchrony. Yet, OpenMP tasking overheads limit the gain of task-based applications over their historical loop parallelization (parallel for construct). This paper identifies the OpenMP task dependency graph discovery speed as a limiting factor in the performance of task-based applications. We study its impact on intra and inter-node performances over two benchmarks (Cholesky, HPCG) and a proxy-application (LULESH). We evaluate the performance impacts of several discovery optimizations, and introduce a persistent task dependency graph reducing overheads by a factor up to 15 at run-time. We measure 2x speedup over parallel for versions weak scaled to 16K cores, due to improved cache memory use and communication overlap, enabled by task refinement and depth-first scheduling. This work [12] was one of the result obtained during the PhD of Romain Peirera co-advised by Adrien Roussel (CEA), Patrick Carribaut (CEA) and Thierry Gautier.

8.3.5 Enhancing productivity on heterogeneous supercomputers with task-based programming model

Participants: Romain Peirera.

Heterogeneous supercomputers with GPUs are one of the best candidates to build Exascale machines. However, porting scientific applications with millions of lines of code lines is challenging. Data transfers/locality and exposing enough parallelism determine the maximum achievable performance on such systems. Thus porting efforts impose developers to rewrite parts of the application which is tedious and time-consuming and does not guarantee performances in all the cases. Being able to detect which parts can be expected to deliver performance gains on GPUs is therefore a major asset for developers. Moreover, task parallel programming model is a promising alternative to expose enough parallelism while allowing asynchronous execution between CPU and GPU. OpenMP 4.5 introduces the « target » directive to offload computation on GPU in a portable way. Target constructions are considered as explicit OpenMP task in the same way as for CPU but executed on GPU. In this work [15], we propose a methodology to detect the most profitable loops of an application that can be ported on GPU. While we have applied the detection part on several mini applications (LULESH, miniFE, XSBench and Quicksilver), we experimented the full methodology on LULESH through MPI+OpenMP task programming model with target directives. It relies on runtime modifications to enable overlapping of data transfers and kernel execution through

tasks. This work has been integrated into the MPC framework, and has been validated on distributed heterogeneous system.

8.3.6 Guidelines for writing portable floating-point software

We defined here some guidelines for critical applications to reduce the arithmetic numerical issue and we provide additional guidelines dedicated to Cloud platform. Many architecture was evaluted (OS native, Virtual Machine or container architecture). With some care, porting numerical software from a micro-controller to the Cloud, or directly writing applications to the Cloud, can be done provided that some recommendations listed below be followed. It is in fact not more difficult than porting software from a micro-controller to any general-purpose processor. The result of this research is an internal unpublished report dedicated to a private partner.

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

Participants: Eddy Caron, Thierry Gautier, Laurent Lefevre.

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 / DAM-Île de France. This collaboration is based on the co-advising of a CEA PhD. The research of the PhD student (Romain Pereira) focuses high performance OpenMP + MPI executions. MPC was developed for high performance MPI application. Recently a support for OpenMP was added. The goal of the PhD is to work on better cooperation of OpenMP and MPI thanks to the unique framework MPC.

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

Orange We have a collaboration with Orange. This collaboration is sealed through a CIFRE PhD grant. The research of the PhD student (Ghoshana Bista) focuses on the software asset management dedicated to the VNF (Virtual Network Function).

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.

Thales We have a collaboration with Thalès. This collaboration is sealed thanks to a CIFRE PhD grant. The research of the PhD student (Pierre-Etienne Polet) focuses on executing signal processing application on GPU for embedded architecture. The problem and its solutions are at the confluence of task scheduling with memory limitation, optimization, parallel algorithm and runtime system.

TotalLinux We have a collaboration with TotalLinux around the data center project Itrium. More specially we study the impact, the energy consumption, the behavior and the performances of new architectures based on immersion cooling.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

JLESC

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

Title: Joint Laboratory for Extreme Scale Computing

Partner Institution(s): 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, Thierry Gautier, Christian Perez.

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.

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

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 Bioorganiczenej 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
- Universite 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

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

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.

10.2.2 Other european programs/initiatives

PHC Aurora: Extrem Edge-Fog-Cloud continuum

Participants: Laurent Lefevre, Eddy Caron, Vladimir Ostapenco, Mathilde Jay.

Title: PHC Aurora

Date/Duration: 2023-2024

Partner Institution(s): University of Tromso, Norway

Summary: Laurent Lefevre co-eads a PHC Aurora project with University of Tromso (Norway) on the topic " Exploring Energy Monitoring and Leveraging Energy Efficiency on End-to-end Worst Edge-Fog- Cloud Continuum for Extreme Climate Environnements Observatories". The aim of this collaboration is to explore a best effort end-to-end energy monitoring system for a worst-case continuum, to discuss a long-term infrastructure continuum for the edge and to design an experimental validation of usable energy leverages at the edge.

10.3 National initiatives

Priority Research Programmes and Equipments (PEPR)

PEPR Cloud – Taranis

Participants: Christian Perez, Yves Caniou, Eddy Caron, Elise Jeanneau, Laurent Lefevre.

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

Partners: Inria, CNRS, IMT, U. Grenobles-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, Olivier Glück.

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.

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.

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, Etienne Mauffret, Lucien Ndjie Ngale, Laurent Lefèvre, Christian Perez.

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

Parners: 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 coonsumption. 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 laborary 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 animator of this project. Some Avalon PhD students are involved in this Inria Large Scale Initiative (Défi) : Maxime Agusti and Vladimir Ostanpenco.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Laurent Lefevre was co-general chair of the GreenDays2023@Lyon: "Energy efficiency, environmental impacts of digital technology, sobriety and digital frugality: a decompartmentalized vision !", Lyon, France, March 27-28, 2023

Member of the organizing committees

- Laurent Lefevre was co-organizer of the EcoInfo conference: "Screens: threats to health" , Lyon, France, May 9, 2023
- Laurent Lefevre and Christian Perez were co-organizers of the "20th anniversary of the Grid'5000 platform" , Lyon, France, May 10, 2023
- Christian Perez was member of the Organizing Committee of JCAD, the French Journées Calcul Données (Reims, 2-4 Oct 2023).

11.1.2 Scientific events: selection

Member of the conference program committees

- Yves Caniou was a program committee member for the International Conference on Computational Science and its Applications 2023.
- Eddy Caron was member of the PC for
  - CLOSER 2023. International Conference on Cloud Computing and Services Science.
  - Review Editor in Cloud Computing for Frontiers since 23 oct.
  - ICCS 2023. International Conference on Computational Science.
  - CCGRID 2023. International Symposium on Cluster, Cloud and Internet Computing.
  - CNRIA 2023. Conference on Research in Computer Science and its Applications.
- Laurent Lefevre was member of the PC of
  - ICPADS 2023: 29th IEEE International Conference on Parallel and Distributed Systems
  - CloudAM2023: 12th International Workshop on Cloud and Edge Computing, and Applications Management
  - ICA3PP 2023: The 23rd International Conference on Algorithms and Architectures for Parallel processing
  - SBAC-PAD 2023: IEEE International Symposium on Computer Architecture and High-Performance Computing
  - IC2E 2023: 11th IEEE International Conference on Cloud Engineering
  - ResIA: Résilience & IA, in conjunction with PFIA 2023
- Christian Perez was member of the PC of workshop "FutureG Experimental Test Platforms for Advanced Systems Implementation and Research", IEEE GLOBLECOM (Kuala Lumpur, Malaysia, December 2023), JCAD 2023 (Reims, 2-4 Oct 2023), of Compas (Annecy, July 5-7, 2023)
- Elise Jeanneau was a member of the PC of AlgoTel 2023: 25ème rencontres francophones sur les aspects algorithmiques des télécommunications.

11.1.3 Journal

Reviewer - reviewing activities

- Eddy Caron was reviewer for
  - TCS (Theoretical Computer Science) journal
  - Engineering Applications of Artificial Intelligence journal.
- Christian Perez was reviewer for FGCS.

11.1.4 Invited talks

- Laurent Lefevre has given the following invited talks:
  - "Energy efficiency and environmental impacts of large scale parallel and distributed systems - Exploring frugality at large scale", Seminar in University of Tromso, Norway, November 13, 2023
  - "Le numérique : empilement, côté obscur, effets rebonds... Comment aborder et réduire ses impacts environnementaux ?", Alstom, Villeurbanne, October 2,2023
  - "Les impacts environnementaux du numérique : les prendre en compte dans les activités de recherche, les détecter, les mesurer, les éviter ?", Rencontres Huma-Num, Ecully, June 14, 2023

  - "Quizz EcoInfo", ICTS4S Conference, Rennes, France, June 6, 2023
  - "Impact environnemental du numérique : on en est où ?", Rencontres du GDR IG-RV, Lyon, May 30, 2023
  - "Les impacts (négatifs) du numérique : les détecter, les mesurer, les éviter ?", Ecole des Mines de Nancy, Nancy, May 12, 2023
  - "Grid'5000 : une plate-forme pour mesurer l'énergie !", Grid'5000 20th anniversary, May 10, 2023
  - "Numérique et impacts environnementaux", Cour des Comptes, April 4, 2023

- Christian Perez gave a keynote talk at the ComPas conference, entitled "Des plates-formes d'expérimenttation Grid'5000/FIT à SLICES-FR", Annecy, July 7, 2023.

11.1.5 Scientific expertise

- Yves Caniou evaluated a project for the ANR PRCE (Projets de recherche collaborative).
- Eddy Caron is member of an ANR committee (*for privacy the ID of the committe is not given here*)
- Christian Perez evaluated 3 projects for the French Direction générale de la Recherche et de l'Innovation.

11.1.6 Research administration

- 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 a member of the Inria Lyon Strategic Orientation Committee. He is in charge of organizing scientific collaborations between Inria and **SKA France**. He was a member of the jury for recruiting CRCN candidates in Inria Lyon center. He was president of a jury for hiring an Inria permanent research engineer.
- Since 2023, Elise Jeanneau is a member of the Inria Evaluation Committee.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Licence: Yves Caniou, Algorithmique programmation impérative initiation, 126h, niveau L1, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Algorithmique et programmation récursive, 42h, niveau L1, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Programmation Concurrente, 18h and Co-Responsible of UE, niveau L3, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Réseaux, 9h, niveau L3, Université Claude Bernard Lyon 1, France.
- Licence: Yves Caniou, Systèmes d'information documentaire, 20h, niveau L3, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Responsible of alternance students, 12h, niveau M1, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Sécurité, 30h and Responsible of UE, niveau M2, Université Claude Bernard Lyon 1, France.
- Master: Yves Caniou, Responsible of alternance students, 9h, niveau M2, Université Claude Bernard Lyon 1, France

- Licence: Eddy Caron, Programmation, 48h, L3, ENS de Lyon. France.
- Agreg Info (FEADéP): Eddy Caron, Operating System and Network, 15h, Agreg, ENS de Lyon. France.
- Agreg Info (FEADéP): Eddy Caron, TP Programmation, 11h, Agreg, ENS de Lyon. France.
- Master: Eddy Caron, Distributed System, 30h, M1, ENS de Lyon. France.
- Master: Eddy Caron, Langages, concepts et archi pour les données, 30h, M2, ISFA.
- Master: Laurent Lefèvre, Parallélisme, 12h, niveau M1, Université Lyon 1, France.
- CAPES Informatique : Laurent Lefèvre, Numérique responsable, 3h, Université Lyon1, France
- Agreg Info : Laurent Lefèvre, Impacts environnementaux du numérique 3h, Agreg, ENS de Lyon. France.
- Master : Laurent Lefèvre, Impacts environnementaux du numérique 3h, Master BioInfo, Université Lyon1. France.
- Licence: Laurent Lefèvre, TP Programmation Concurrente, 10h, niveau L3, Université Lyon1, France
- Master: Thierry Gautier, Introduction to HPC, 20h, niveau M2, INSA Lyon, France.
- Licence: Olivier Glück, Introduction Réseaux et Web, 54h, niveau L1, Université Lyon 1, France.
- Licence: Olivier Glück, Réseaux, 2x70h, niveau L3, Université Lyon 1, France.
- Master: Olivier Glück, Réseaux par la pratique, 10h, niveau M1, Université Lyon 1, France.
- Master: Olivier Glück, Responsible of Master SRS (Systèmes, Réseaux et Infrastructures Virtuelles) located at IGA Casablanca, 20h, niveau M2, IGA Casablanca, Maroc.
- Master: Olivier Glück, Administration systèmes et réseaux, 30h, niveau M2, Université Lyon 1, France.
- Master: Olivier Glück, Administration systèmes et réseaux, 24h, niveau M2, IGA Casablanca, Maroc.
- Licence : Frédéric Suter, Programmation Concurrente, 32.33, L3, Université Claude Bernard Lyon 1, France
- Licence: Elise Jeanneau, Introduction Réseaux et Web, 30h, niveau L1, Université Lyon 1, France.
- Licence: Elise Jeanneau, Réseaux, 52h, niveau L3, Université Lyon 1, France.
- Licence: Elise Jeanneau, Algorithmique, programmation et structures de données, 24h, niveau L2, Université Lyon 1, France
- Licence: Elise Jeanneau, Architecture des ordinateurs, 24h, niveau L2, Université Lyon 1, France
- Licence, Elise Jeanneau, Réseaux, systèmes et sécurité par la pratique, 24h, niveau L3, Université Lyon 1, France
- Master: Elise Jeanneau, Algorithmes distribués, 45h, niveau M1, Université Lyon 1, France.
- Master: Elise Jeanneau, Réseaux, 6h, niveau M1, Université Lyon 1, France.
- Master: Elise Jeanneau, Compilation et traduction de programmes, 22h, niveau M1, Université Lyon 1, France.

11.2.2 Supervision

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. ENS de Lyon. Inria. Avalon), Benjamin Fichel (co-dir. OVHcloud), Laurent Lefevre (dir. Inria. Avalon) et Anne-Cécile Orgerie (co-dir. Inria. Myriads),.
- Adrien Berthelot. *Évaluation accélérée et assistée des impacts environnementaux d'un Système d'Information avec l'ensemble de ses services numériques*, Jan 2022, Eddy Caron (ENS de Lyon. Inria. Avalon), Christian Fauré (Octo Technology) and Laurent Lefevre (Inria. Avalon).
- Hugo Hadjur. *Designing sustainable autonomous connected systems with low energy consumption*, 2020-2023, PhD defended on July 13, 2023, Laurent Lefevre (dir. Inria. Avalon), Doreid Ammar (co-dir. Aivancity group)
- Mathilde Jay. *"Low-cost learning algorithm at the edge"*, 2021, Laurent Lefevre (co-dir. Inria. Avalon), Denis Trystram (dir. LIG, UGA)
- Simon Lambert. *Forecast and dynamic resource provisioning on a virtualization infrastructure*, 2022, Eddy Caron (dir. ENS de Lyon. Inria. Avalon), Laurent Lefevre (co-dir Inria. Avalon), Rémi Grivel (co-dir. Ciril Group).
- Vladimir Ostapenco. *Modeling and design of a framework and its environmental Gantt Chart in order to manage heterogeneous energy leverages*, FrugalCloud Inria-OVHCloud collaboration, Laurent Lefevre (dir. Inria. Avalon), Anne-Cécile Orgerie (co-dir. Inria. Myriads), Benjamin Fichel (co-dir. OVHcloud)

11.2.3 Juries

- Eddy Caron was PhD reviewer of Wilmer Garzon Alfonso: "Secure distributed workflows for biomedical data analytics", Universidad Escuela Colombiana de Ingeniería Julio Garavito (Université technologique Julio Garavito, Colombie) and IMT Atlantique Campus de Nantes, France.
- Eddy Caron PhD examiner of Nihel Kaboubi: "Système de partitionnement hybride pour une inférence distribuée et embarquée des CNNs sur les équipements en bordure de réseau.", Université Grenoble Alpes.
- Laurent Lefevre was PhD reviewer of Hergys Rexha : "Energy Aware Runtime Systems for Elastic Stream Processing Platforms", Abo Akademi University, Finland, April 10, 2023
- Christian Perez was president of the PhD defense committee of Marek Felsoci: "Solveurs rapides pour l'aéroacoustique haute-fréquence", Bordeaux, February 22, 2023.
- Christian Perez was PhD reviewer and member of the defense committe of Daavadorj Battulga: "Contributions to the Management of Stream Processing Pipelines in Fog Computing Environments", Rennes I, March 23, 2023.
- Christian Perez was PhD reviewer and member of the defense committe of Nicolas Greneche: "Élasticité des infrastructures HPC conteneurisées : Résonance entre le HPC et le Cloud", Paris, November 20, 2023.

11.3 Popularization

- "Transition numérique : Aujourd'hui, les grandes évolutions concernent la data", Emmanuelle Frenoux et Laurent Lefevre et Didier Mallarino, Archimag Magazine, February 20, 2023

11.3.1 Internal or external Inria responsibilities

- Eddy Caron was member of the "Jury Inria Concours CRCN et ISFP" at Nancy.

11.3.2 Education

- Yves Caniou co-organized the 6th Edition of Le Campus du Libre, on saturday Oct. 21 2023 at Université Lyon 2, Lyon.
- Eddy Caron was member of CSI (Comité de suivi de thèse) for
  - Mélanie Fontaine. Université de Picardie Jules Verne.
  - Sylvain Lejamble. Université Savoie Mont Blanc.
  - Cedric Prigent. (INSA Rennes) - 2ème année

12 Scientific production

12.1 Major publications

[1] 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. URL: https://hal.science/hal-04047008.

12.2 Publications of the year

International journals

[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. URL: https://hal.science/hal-04047008.

International peer-reviewed conferences

[3] A. Berthelot, E. Caron, M. Jay and L. Lefèvre. 'Estimating the environmental impact of Generative-AI services using an LCA-based methodology'. In: *Procedia CIRP*. 31ST CIRP CONFERENCE ON LIFE CYCLE ENGINEERING. Turin, Italy, 19th June 2024. URL: https://inria.hal.science/hal-04346102.

[4] E. Caron and N. Chappe. 'FicWebBoard: A Playful and Collaborative Learning Platform Built for All People and All Programming Languages'. In: 2023 IEEE Frontiers in Education Conference (FIE). College Station, TX, United States: IEEE, 18th Oct. 2023, pp. 1–8. DOI: 10.1109/FIE58773.2023.10343040. URL: https://inria.hal.science/hal-04380643.

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

[6] I. Daoudi, T. Gautier, S. Thibault and S. Perarnau. 'Improving Simulations of Task-Based Applications on Complex NUMA Architectures'. In: IWOMP 2023 - 19th International Workshop on OpenMP. Vol. 14114. Lecture Notes in Computer Science. Bristol, United Kingdom: Springer Nature Switzerland, 1st Sept. 2023, pp. 195–209. DOI: 10.1007/978-3-031-40744-4_13. URL: https://inria.hal.science/hal-04201317.

[7] A. Y. Guifo Fodjo, J. L. Lacmou Zeutouo and S. Bowong. 'Separation of Concerns in an Edge-Based Compartmental Modeling Framework'. In: 16th International Joint Conference on Biomedical Engineering Systems and Technologies. Lisbonne, Portugal, 16th Feb. 2023. URL: https://hal.sorbonne-universite.fr/hal-03953230.

[8] H. Hadjur, L. Lefèvre and D. Ammar. 'Services Orchestration at the Edge and in the Cloud on Energy-Aware Precision Beekeeping Systems'. In: PAISE 2023: 5th Workshop on Parallel AI and Systems for the Edge - co-conducted with IPDPS 2023. St. Petersburg, FL, United States, 15th May 2023. DOI: 10.1109/IPDPSW59300.2023.00129. URL: https://inria.hal.science/hal-04091575.

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

[10] ***Best Paper***
E. Mauffret, E. Jeanneau and E. Caron. 'SkyData : Rise of the Data How can the Intelligent and Autonomous Data paradigm become real?' In: The International Conference on Cloud Computing and Services Science (CLOSER). Prague (Czech Republic), Czech Republic, 26th Apr. 2023, p. 8. URL: https://hal.science/hal-04040588.

[11] R. Pereira, M. Martin, A. Roussel, P. Carribault and T. Gautier. 'Suspending OpenMP Tasks on Asynchronous Events: Extending the Taskwait Construct'. In: IWOMP 23 - International Workshop on OpenMP. Vol. 14114. Lecture Notes in Computer Science. Bristol, United Kingdom: Springer Nature Switzerland, 1st Sept. 2023, pp. 66–80. DOI: 10.1007/978-3-031-40744-4_5. URL: https://inria.hal.science/hal-04135481.

[12] R. Pereira, A. Roussel, P. Carribault and T. Gautier. 'Investigating Dependency Graph Discovery Impact on Task-based MPI+OpenMP Applications Performances'. In: 52nd International Conference on Parallel Processing (ICPP 2023). Vol. 52nd International Conference on Parallel Processing (ICPP 2023). Salt Lake City, United States: ACM, 7th Aug. 2023, pp. 163–172. DOI: 10.1145/3605573.3605602. URL: https://hal.science/hal-04136674.

[13] P.-É. Polet, R. Fantar and T. Gautier. 'Introducing Moldable Tasks in OpenMP'. In: IWOMP 23 - International Workshop on OpenMP. Vol. 14114. Lecture Notes in Computer Science. Bristol, UK, United Kingdom: Springer Nature Switzerland; Springer Nature Switzerland, 1st Sept. 2023, pp. 51–65. DOI: 10.1007/978-3-031-40744-4_4. URL: https://hal.science/hal-04409117.

National peer-reviewed Conferences

[14] S. Lambert, E. Caron, L. Lefèvre and R. Grivel. 'Étude et modélisation des mécanismes de surallocation de mémoire pour la consolidation de machines virtuelles'. In: Compas 2023 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Annecy, France, 4th July 2023. URL: https://hal.science/hal-04317788.

Conferences without proceedings

[15] A. Roussel, M. Boichot, R. Pereira and M. Ferat. 'Enhancing productivity on heterogeneous supercomputers with task-based programming model'. In: SIAM CSE 2023 - SIAM Conference on Computational Science and Engineering. Amsterdam, Netherlands, 26th Feb. 2023. URL: https://cea.hal.science/cea-03994014.

Doctoral dissertations and habilitation theses

[16] G. Bista. 'Total cost modeling of software ownership in Virtual Network Functions'. Ecole normale supérieure de lyon - ENS LYON, 24th Jan. 2023. URL: https://theses.hal.science/tel-04074989.

Reports & preprints

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

Other scientific publications

[18] C. Bonamy, C. Boudinet, L. Bourgès, K. Dassas, L. Lefèvre, B. Ninassi and F. Vivat. *Good practices in digital service ecodesign for software developers*. 7th Feb. 2023. URL: https://hal.science/hal-03977001.

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

12.3 Other

Scientific popularization

[20] C. Bonamy, C. Boudinet, L. Bourges, K. Dassas, L. Lefèvre, B. Ninassi and F. Vivat. 'Bonnes pratiques en écoconception de service numérique: Incontournables pour diminuer de façon drastique les GES, elles font l'objet d'un guide. On en découvre la genèse.' In: *Collection numérique de l'AMUE, Agence de mutualisation des universités et établissements d'enseignement supérieur* 29 (9th Oct. 2023), p. 2. URL: https://inria.hal.science/hal-04402624.

12.4 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. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4906989.

[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. URL: http://dx.doi.org/10.1177/1094342009347766.

[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.

[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.

[25] G. Madec. *NEMO ocean engine*. Note du Pole de modélisation 27. ISSN No 1288-1619. France: Institut Pierre-Simon Laplace (IPSL), 2008.

[26] OpenACC. *The OpenACC Application Programming Interface*. Version 1.0. Nov. 2011. URL: http://www.openacc-standard.org.

[27] OpenMP Architecture Review Board. *OpenMP Application Program Interface*. Version 3.1. July 2011. URL: http://www.openmp.org.

[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.

[29] C. Szyperski. *Component Software - Beyond Object-Oriented Programming*. 2nd ed. Addison-Wesley / ACM Press, 2002, p. 608.

[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.