RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

Institut national des sciences appliquées de Rennes, École normale supérieure de Rennes, Université Rennes 1

2021 ACTIVITY REPORT

Project-Team KERDATA

Scalable Storage for Clouds and Beyond

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

DOMAIN

Networks, Systems and Services, Distributed Computing

THEME

Distributed and High Performance Computing

Contents

Project-Team KERDATA 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 2

3 Research program 4

3.1 Research Axis 1: Convergence of HPC and Big Data Infrastructures 4

3.2 Research Axis 2: Advanced Data Processing on the Edge-to-Cloud Digital Continuum 5

3.3 Research Axis 3: Scalable I/O, communication, in situ visualization and analysis on HPC systems at extreme scales 7

4 Application domains 9

4.1 Climate and meteorology 9

4.2 Earth science 9

4.3 Sustainable development through precision agriculture 10

4.4 Smart cities 10

4.5 Botanical Science 11

5 Social and environmental responsibility 11

5.1 Footprint of research activities 11

5.2 Impact of research results 11

6 Highlights of the year 12

7 New software and platforms 12

7.1 New software 12

7.1.1 Damaris 12

7.1.2 KerA 13

7.1.3 E2Clab 13

8 New results 14

8.1 Convergence of HPC and Big Data Infrastructures 14

8.1.1 Provisioning storage resources for hybrid supercomputer/cloud infrastructures 14

8.1.2 Supporting seamless execution of HPC-enabled workflows across the Computing Continuum 14

8.1.3 Artificial intelligence-based data analysis in heterogeneous and volatile environments 15

8.1.4 Memory and data-awareness in hybrid workflows 16

8.2 Advanced Data Processing on the Edge-to-Cloud Digital Continuum 16

8.2.1 Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum 16

8.2.2 Enabling Provenance Capture in Edge-to-Cloud Workflows 17

8.2.3 Towards Data Parallel Rehearsal-based Continual Learning 17

8.2.4 Deploying Heterogeneity-aware Deep Learning Workloads on the Computing Continuum . . . 18
8.2.5 Modeling distributed stream processing across the Edge-to-Cloud Continuum 18
8.3 Scalable I/O, Communication, in-situ Visualization and Analysis on HPC Systems at Extreme Scales . . . 19
8.3.1 Exploring Alternative Strategies Leveraging Dedicated Resources for in-situ visualization for Large Scale Simulations . . . 19
8.3.2 Scalable asynchronous I/O and in-situ processing with Damaris for Carbon Sequestration . . . 20

9 Partnerships and cooperations **20**
9.1 International initiatives . . . 20
9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . 20
9.1.2 Inria associate team not involved in an IIL or an international program . . . 21
9.1.3 Participation in other International Programs . . . 22
9.2 European initiatives . . . 22
9.2.1 FP7 & H2020 projects . . . 22
9.2.2 Collaborations in European Programs, except FP7 and H2020 . . . 23
9.2.3 Other European programs/initiatives . . . 24
9.2.4 Collaborations with Major European Organizations . . . 24
9.3 National initiatives . . . 25
9.3.1 ANR . . . 25
9.3.2 Other National Projects . . . 25

10 Dissemination **26**
10.1 Promoting scientific activities . . . 27
10.1.1 Scientific events: organisation . . . 27
10.1.2 Scientific events: selection . . . 27
10.1.3 Journal . . . 28
10.1.4 Invited talks . . . 28
10.1.5 Leadership within the scientific community . . . 28
10.1.6 Scientific expertise . . . 29
10.2 Teaching - Supervision - Juries . . . 29
10.2.1 Teaching . . . 29
10.2.2 Supervision . . . 30
10.2.3 Juries . . . 31
10.3 Popularization . . . 31
10.3.1 Internal or external Inria responsibilities . . . 31
10.3.2 Interventions . . . 31

11 Scientific production **31**
11.1 Major publications . . . 31
11.2 Publications of the year . . . 33

11.3 Cited publications 34

Project-Team KERDATA

Creation of the Project-Team: 2012 July 01

Keywords

Computer sciences and digital sciences

A1.1.1. – Multicore, Manycore

A1.1.4. – High performance computing

A1.1.5. – Exascale

A1.1.9. – Fault tolerant systems

A1.3. – Distributed Systems

A1.3.5. – Cloud

A1.3.6. – Fog, Edge

A2.6.2. – Middleware

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.8. – Big data (production, storage, transfer)

A6.2.7. – High performance computing

A6.3. – Computation-data interaction

A7.1.1. – Distributed algorithms

A9.2. – Machine learning

A9.7. – AI algorithmics

Other research topics and application domains

B3.2. – Climate and meteorology

B3.3.1. – Earth and subsoil

B8.2. – Connected city

B9.5.6. – Data science

B9.8. – Reproducibility

B9.11.1. – Environmental risks

1 Team members, visitors, external collaborators

Research Scientists

- Gabriel Antoniu [Team leader, Inria, Senior Researcher, HDR]
- François Tessier [Inria, Starting Faculty Position]

Faculty Members

- Luc Bougé [École normale supérieure de Rennes, Professor, Emeritus since Sept 2021, HDR]
- Alexandru Costan [INSA Rennes, Associate Professor, HDR]

PhD Students

- Thomas Bouvier [Inria]
- Julien Monniot [Inria, from Oct 2021]
- Cedric Prigent [Inria, from Nov 2021]
- Daniel Rosendo [Inria]

Technical Staff

- Joshua Charles Bowden [Inria, Engineer]

Interns and Apprentices

- Hugo Chaugier [Inria, from Feb 2021 until Jul 2021]
- Juliette Fournis d'Albiat [Inria, from Oct 2021]
- Matthieu Robert [Univ de Rennes I, from Feb 2021 until Jul 2021]

Administrative Assistant

- Laurence Dinh [Inria, from Feb 2021]

2 Overall objectives

Context: the need for scalable data management. For several years now we have been witnessing a rapidly increasing number of application areas generating and processing very large volumes of data on a regular basis. Such applications, called *data-intensive*, range from traditional large-scale simulation-based scientific domains such as climate modeling, cosmology, and bioinformatics to more recent industrial applications triggered by the Big Data phenomenon: governmental and commercial data analytics, financial transaction analytics, etc. Recently, the data-intensive application spectrum further broadened by the emergence of IoT applications that need to process data coming from large numbers of distributed sensors.

Our objective. The KerData project-team is focusing on designing innovative architectures and systems for *scalable data storage and processing*. We target three types of infrastructures: *pre-Exascale high-performance supercomputers*, *cloud-based* and *edge-based* infrastructures, according to the current needs and requirements of data-intensive applications. In addition, as emphasized by the latest Strategic Research Agenda of ETP4HPC [6], new complex applications have started to emerge: they combine simulation, analytics and learning and require hybrid execution infrastructures combining supercomputers, cloud-based and edge-based systems. Our most recent research aims to address the data-related requirements (storage, processing) for such complex workflows. They are structured in three research axes summarized below.

Challenges and goals related to the HPC-Big Data-AI convergence. Traditionally, HPC, Big Data analytics and Artificial Intelligence have evolved separately, using different approaches for data storage and processing as well as for leveraging their respective underlying infrastructures. The KerData team has been tackling the convergence challenge from a data storage and processing perspective, trying to provide answers to questions like: what common storage abstractions and data management techniques could fuel storage convergence, to support seamless execution of hybrid simulation/analytics workflows on potentially hybrid supercomputer/cloud infrastructures? The team's activities in this area are grouped in Research Axis 1 (see below).

Challenges and goals related to cloud-based and edge-based storage and processing. The growth of the Internet of Things is resulting in an explosion of data volumes at the edge of the Internet. To reduce costs incurred due to data movement and centralized cloud-based processing, cloud workflows have evolved from single-datacenter deployment towards multiple-datacenter deployments, and further from cloud deployments towards distributed, edge-based infrastructures.

This allows applications to distribute analytics while preserving low latency, high availability, and privacy. Jointly exploiting edge and cloud computing capabilities for stream-based processing leads however to multiple challenges.

In particular, understanding the dependencies between the application workflows to best leverage the underlying infrastructure is crucial for the end-to-end performance. We are currently missing models enabling this adequate mapping of distributed analytics pipelines on the Edge-to-Cloud Continuum. The community needs tools that can facilitate the modeling of this complexity and can integrate the various components involved. This is the challenge we address in Research Axis 2 (described below).

Challenges and goals related to data-intensive HPC applications. Key research fields such as climate modeling, solid Earth sciences and astrophysics rely on very large-scale simulations running on post-Petascale supercomputers. Such applications exhibit requirements clearly identified by international panels of experts like IESP [38], EESI [36], ETP4HPC [37]. A jump of one order of magnitude in the size of numerical simulations is required to address some of the fundamental questions in several communities in this context. In particular, the lack of data-intensive infrastructures and methodologies to analyze the huge results of such simulations is a major limiting factor. The high-level challenge we have been addressing in Research Axis 3 (see below) is to find scalable ways to store, visualize and analyze massive outputs of data during and after the simulations through asynchronous I/O and *in-situ processing*.

Approach, methodology, platforms. KerData's global approach consists in studying, designing, implementing and evaluating distributed algorithms and software architectures for scalable data storage and I/O management for efficient, large-scale data processing. We target three main execution infrastructures: edge and cloud platforms and pre-Exascale HPC supercomputers.

The highly experimental nature of our research validation methodology should be emphasized. To validate our proposed algorithms and architectures, we build software prototypes, then validate them at large scale on real testbeds and experimental platforms.

We strongly rely on the Grid'5000 platform. Moreover, thanks to our projects and partnerships, we have access to reference software and physical infrastructures.

In the cloud area, we use the Microsoft Azure and Amazon cloud platforms, as well as the Chameleon [33] experimental cloud testbed. In the post-Petascale HPC area, we are running our experiments on systems including some top-ranked supercomputers, such as Titan, Jaguar, Kraken, Theta, Pangea and Hawk. This provides us with excellent opportunities to validate our results on advanced realistic platforms.

Collaboration strategy. Our collaboration portfolio includes international teams that are active in the areas of data management for edge, clouds and HPC systems, both in Academia and Industry. Our academic collaborating partners include Argonne National Laboratory, University of Illinois at Urbana-Champaign, Universidad Politécnica de Madrid, Barcelona Supercomputing Center. In industry, through bilateral or multilateral projects, we have been collaborating with Microsoft, IBM, Total, Huawei, ATOS/BULL.

Moreover, the consortia of our collaborative projects include application partners in multiple application domains from the areas of climate modeling, precision agriculture, earth sciences, smart cities or botanical science. This multidisciplinary approach is an additional asset, which enables us to take into account application requirements in the early design phase of our proposed approaches to data storage and processing, and to validate those solutions with real applications and real users.

Alignment with Inria's scientific strategy. Data-intensive applications exhibit several common requirements with respect to the need for data storage and I/O processing. We focus on some core challenges related to data management, resulting from these requirements. Our choice is perfectly in line with Inria's strategic objectives [39], which acknowledges in particular HPC-Big Data convergence as one of the Top 3 priorities of our institute.

In addition, we have recently engaged in collaborative projects with some of Inria's main strategic partners: DFKI (main German research center in artificial intelligence) though the ENGAGE Inria-DFKI project and ATOS (through the ACROSS and EUPEX H2020 EuroHPC projects). Gabriel Antoniu, Head of the KerData team, serves as the main scientific lead for Inria in these three projects.

3 Research program

3.1 Research Axis 1: Convergence of HPC and Big Data Infrastructures

Our research in the area of HPC-Big Data convergence addresses two main topics:

- Enabling storage-based HPC-Big Data convergence;
- Modeling and designing malleable storage systems.

Convergent storage for HPC systems and clouds through transactional blobs. We believe that HPC-Big Data convergence can efficiently be tackled from the angle of HPC/cloud storage convergence. Since 2016 we have been investigating how object-based storage could serve as a common abstraction for both HPC simulations (traditionally running on supercomputers) and for Big Data analytics (traditionally running on clouds). We argue that transactional blob storage is a strong candidate for future data platforms combining these infrastructures [26].

We proposed a pioneering approach to the convergence of HPC and Big Data from the perspective of storage, by introducing Týr — the first blob storage system to provide built-in, multiblob transactions, while retaining sequential consistency and high throughput under heavy access concurrency. It leverages novel version management and read protocols allowing applications to read a consistent version of a blob spanning multiple chunks in presence of concurrent, conflicting writers. Large-scale experiments on the Microsoft Azure cloud with a production application from CERN LHC show Týr throughput outperforming state-of-the-art solutions by more than 75 % [10] (***Best Student Paper Award Finalist at the ACM/IEEE SC16 conference*** — the reference international conference in the HPC area). We investigate this approach on various large-scale platforms with a diverse spectrum of applications.

Provisioning storage resources for hybrid supercomputer/cloud infrastructures. Another recent idea we are developing in the context of HPC and Big Data convergence concerns the provisioning of storage resources. The way these resources are accessed on supercomputers and clouds opposes a complex low-level vision that requires tight user control (on supercomputers) and a very abstract vision that implies uncertainty for performance modeling (on clouds). Nevertheless, taking full advantage of all available resources is critical in a context where storage is central for coupling workflow components. This line of work is still preliminary.

A first effort has begun in 2021 on how to make hybrid heterogeneous storage resources allocatable and elastic to meet the needs of I/O-intensive hybrid workloads. A publication is in preparation to share the first results. A PhD thesis started in October 2021 on this topic, with a focus on scheduling algorithms in the context of a storage-aware job scheduler.

Malleable storage systems. Using storage resources in an elastic, cloud-inspired manner on supercomputers or on hybrid supercomputer/cloud infrastructures requires adequate modeling. Naturally, energy and cost savings can be obtained by reducing idle resources. Malleability, which is the possibility for resource managers to *dynamically* increase or reduce the resources of jobs, appears as a promising means to progress towards this goal. However, state-of-the-art parallel and distributed file systems have not been designed with malleability in mind. This is mainly due to the supposedly high cost of storage decommission, which is considered to involve expensive data transfers. Nevertheless, as network and storage technologies evolve, old assumptions on potential bottlenecks can be revisited.

We investigate and evaluate the viability of malleability as a design principle for a distributed file system. We specifically model the minimal duration of the rescaling operations for storage resources. Furthermore, we introduce Pufferbench [24], a benchmark for evaluating how fast one can scale up and down a distributed storage system on a given infrastructure and, thereby, how viably can one implement storage malleability on it. We explore how it can serve to quickly prototype and evaluate mechanisms for malleability in existing distributed storage systems.

Towards unified data processing techniques for Extreme Computing and Big Data applications. As part of our contribution to HPC-Big Data convergence, another direction we explore regards how to efficiently leverage HPC infrastructures for data-intensive applications. We propose a solution based on burst buffers, which enables efficient analytics that can run concurrently to simulations on the same supercomputer. This is achieved thanks to a prefetching technique that fetches the input data for analytics applications to be stored close to computing nodes thus reducing the latency of reading data inputs [28].

3.2 Research Axis 2: Advanced Data Processing on the Edge-to-Cloud Digital Continuum

Distributed digital infrastructures for computation and analytics are now evolving towards an interconnected ecosystem allowing complex applications to be executed from IoT Edge devices to the HPC Cloud (aka the *Computing Continuum*, or the *Edge-to-Cloud Continuum*).

As a first step towards enabling the Computing Continuum vision, we aim for a holistic approach to data processing across the continuum: we thrive to design new algorithms, systems and methodologies able to seamlessly handle data atop these highly heterogeneous infrastructures. Our objective is to provide for the first time a uniform way to process data, able to hide from users the heterogeneity and volatility of Edge devices or the inherent complexity of hybrid Edge-Cloud deployments.

Scalable stream storage across the continuum. As myriads of frameworks emerge for for stream processing [8], we argue that a plausible path to follow towards scalable stream storage is the careful design and implementation of a unified architecture for stream ingestion and storage which can lead to the optimization of the processing of Big Data applications across the continuum. This approach minimizes data movement within the analytics architecture, finally leading to better utilized resources.

We propose a set of design principles for a scalable, unified architecture for data ingestion and storage: (1) dynamic partitioning based on semantic grouping and sub- partitioning, which enables more flexible and elastic management of stream partitions; (2) lightweight offset indexing (i.e., reduced stream offset management overhead) optimized for sequential record access; (3) adaptive and fine-grained replication to trade-off in-memory storage with performance (low-latency and high throughput with durability).

We implement and evaluate this architecture through the KerA prototype with the goal of efficiently handling diverse access patterns: low-latency access to streams and/or high throughput access to unbounded streams and/or objects [7]. KerA serves a playground for our experimental research in the area of cloud/edge storage.

Cost-efficient execution plans for the uniform placement of stream analytics on the Edge-to-Cloud continuum. In order to execute a request over hybrid Edge-Cloud deployments, one needs a specific execution plan for the Edge engines, another one for the Cloud engines and to ensure the right interconnection between them thanks to an ingestion system. Manually and empirically deploying this analytics pipeline (Edge-Ingestion-Cloud) can lead to sub-optimal computation placement with respect to the network cost (i.e., high latency, low throughput) between the Edge and the Cloud.

We argue that a uniform approach is needed to bridge the gap between Cloud SPEs (Stream Processing Engines) and Edge analytics frameworks in order to leverage a single, transparent execution plan for stream processing in both environments. We introduce Planner, a streaming middleware capable of finding cost-efficient cuts of execution plans between Edge and Cloud. Our goal is to find a distributed placement of operators on Edge and Cloud nodes to minimize the stream processing makespan.

Edge benchmarking. While a plethora of frameworks for Edge processing were recently proposed, the distributed systems community has no clear means today to discriminate between them. Some preliminary surveys exist, focusing on a feature-based comparison. We claim that a step further is needed, to enable a performance-based comparison.

To this purpose, the definition of a benchmark is a necessity. We make a step towards the definition of a methodology for benchmarking Edge processing frameworks [27]. Using two representative real-life stream processing applications and state-of-the-art processing engines, we perform an experimental evaluation based on the analysis of the execution of those applications in fully-Cloud computing and hybrid Cloud-Edge computing infrastructures.

Exploring the continuum through reproducible experiments. Understanding end-to-end performance in such a complex continuum is challenging. This breaks down to reconciling many, typically contradicting, application requirements and constraints with low-level infrastructure design choices. One important challenge is to accurately reproduce relevant behaviors of a given application workflow and representative settings of the physical infrastructure underlying this complex continuum.

We introduce a rigorous methodology for such a process and validate it through **E2C***lab* [11]. It is the first platform to support the complete analysis cycle of an application on the Computing Continuum: *(i)* the configuration of the experimental environment, libraries and frameworks; *(ii)* the mapping between the application parts and machines on the Edge, Fog and Cloud; *(iii)* the deployment of the application on the infrastructure; *(iv)* the automated execution; and *(v)* the gathering of experiment metrics.

We further propose a methodology to support the optimization of real-life applications on the Edge-to-Cloud Continuum. Our approach [16], implemented as an extension of **E2C***lab*, relies on a rigorous analysis of possible configurations in a controlled testbed environment to understand their behavior and related performance trade-offs. We illustrate our methodology by optimizing Pl@ntNet, a world-wide plant identification application. This methodology can be generalized to other applications in the Edge-to-Cloud Continuum.

Supporting AI across the continuum. An increasing number of AI applications turn to the Computing Continuum to speed-up decision making, by processing data close to their sources. One such example is the Earthquake Early Warning (EEW) which detects and characterizes medium and large earthquakes before their damaging effects reach a certain location.

Our research aims to improve the accuracy of EEW systems by means of machine learning across the continuum. We introduce the Distributed Multi-Sensor Earthquake Early Warning (DMSEEW) system, a novel machine learning-based approach that combines data from both types of sensors (GPS stations and seismometers) to detect medium and large earthquakes. The system builds on a geographically distributed infrastructure, which can be deployed on clouds and edge systems. This ensures an efficient computation in terms of response time and robustness to partial infrastructure failures.

Our experiments show that DMSEEW is more accurate than the traditional seismometer-only approach and the combined-sensors (GPS and seismometers) approach that adopts the rule of relative strength. These results have been accepted for publication at AAAI (Special Track on AI for Social Impact), a "A*" conference in the area of Artificial Intelligence [5]. This paper won the *Outstanding Paper Award - Special Track on AI for Social Impact*.

3.3 Research Axis 3: Scalable I/O, communication, in situ visualization and analysis on HPC systems at extreme scales

Over the past few years, the increasing amounts of data produced by large-scale simulations led to a shift from traditional offline data analysis to in-situ analysis and visualization. In-situ processing started by coupling a parallel simulation with an analysis or visualization library, to avoid the cost of writing data on storage and reading it back. Going beyond this simple pairwise tight coupling, complex analysis workflows today are graphs with one or more data sources and several interconnected analysis components.

The core direction of our research in the HPC area aimed to leverage our Damaris approach to support scalable I/O, in situ processing and analytics. Additional related directions concern efficient communication on HPC systems through: 1) scalable topology-aware communications; and 2) RDMA-based data replication.

Scalable I/O and in situ processing with Damaris: leveraging dedicated resources for HDF-based storage and for data processing for geoscience and climate applications. Supercomputers are expected to reach Exascale by 2022. The Fugaku supercomputer (now first in the Top500 supercomputer list) reached 442 PFlop/s (LINPACK performance, which serves to established the ranking) and was reported to have already overcome 1 EFlop/s as early as June 2020 on a specific benchmark (HPC-AI). Extreme-scale scientific simulations deployed on pre-Exascale HPC machines usually store the resulted datasets in standard formats such as HDF5 or NetCDF. In the data storage process, two different approaches are traditionally employed: 1) file-per-process; and 2) collective I/O.

In our research we explore an approach based on dedicated resources for data aggregation and storage. The computing resources are partitioned such that a subset of cores in each node or a subset of nodes of the underlying platform are dedicated to data management. The data generated by the simulation processes are transferred to these dedicated cores/nodes either through shared memory (in the case of dedicated cores) or through the MPI calls (in the case of dedicated nodes) and can be processed asynchronously, with data aggregated and stored in HDF5 format using a specific plug-in.

This approach is implemented in Damaris ([3] and [25]), a middleware software developed in our team. It originated in our collaboration with the JLESC, and was the first software resulted from this joint lab validated in 2011 for integration into the Blue Waters supercomputer project. It scaled up to 16,000 cores on the Titan supercomputer (first in the Top500 supercomputer list in 2013) before being validated on other major HPC systems.

We recently have focused on two main targets:

1. Validating the benefits of the dedicated-core approach for new types of storage backends, e.g., HDF5, where it enables performance improvements approaching the order of 300 % compared to the standard file-per-process approach.
2. Validating Damaris for in situ processing for new simulations codes from different scientific domains, i.e. geoscience (wave propagation) and ocean modeling (e.g., the CROCO coastal and ocean simulation).

Ongoing efforts are supporting the integration of Damaris with the Code_Saturne CFD application from EDF for in-situ visualization [21], within the PRACE6IP project; the integration with the OPM Flow simulator, developed at Sintef (Norway), used for reservoir simulations and carbon sequestration (within the ACROSS EuroHPC project). Future plans include the development of the Damaris as a key technology for asynchronous I/O and in-situ processing on future European Exascale machines.

Scalable communication: exploring topology-awareness on Dragonfly-based supercomputers. Recent network topologies in supercomputers have motivated new research on topology-aware collective communication algorithms for MPI. But such an endeavor requires betting on the fact that topology-awareness is the primary factor to accelerate these collective operations. Besides, showing the benefit of a new topology-aware algorithm requires not only access to a leadership-scale supercomputer with the desired topology, but also a large resource allocation on this supercomputer.

Event-driven network simulations can alleviate such constraints and speed up the search for appropriate algorithms by providing early answers on their relative merit. In our studies, we focused on the Scatter and AllGather operations in the context of the Theta supercomputer's Dragonfly topology [23]. We proposed a set of topology-aware versions of these operations as well as optimizations of the old, non-topology-aware ones. A trivial implementation of Scatter using non-blocking point-to-point communications can be faster than state-of-the art algorithms by up to a factor of 6. Traditional AllGather algorithms can also be improved by the same principle and exhibit a 4x speedup in some situations. These results highlight the need to rethink the collective operations under the light of non-blocking communications.

For this work, Nathanaël Cheriere received the Third Prize at the ACM Graduate Student Research Competition organized at the SC'16 conference.

Achieving fast atomic RDMA-based replication. Performance and fault tolerance are often contradictory goals. Replication is essential for fault-tolerance for in-memory storage systems, which are increasingly present on HPC systems as well as on cloud infrastructures. However, it is a source of high overhead. Remote direct memory access (RDMA) is an attractive technology to support the fast creation of copies of data, since it is low-latency and has no CPU overhead at the target. To ensure atomic data transfers, receivers check and apply only fully received messages. In contrast to such an approach, we introduced Tailwind [12], a zero-copy recovery-log replication protocol for scale-out in-memory databases.

Tailwind is the first replication protocol that eliminates *all* CPU-driven data copying and fully bypasses target server CPUs, thus leaving backups idle. Tailwind ensures all writes are atomic by leveraging a protocol that detects incomplete RDMA transfers. It substantially improves replication throughput and response latency compared with conventional RPC-based replication. Experiments show Tailwind improves RAMCloud's normal-case request processing throughput by 1.7x. It also cuts down writes median and 99th percentile latencies by 2x and 3x respectively. This work is carried out in collaboration with Ryan Stutsman, a coauthor of RAMCloud.

4 Application domains

The KerData team investigates the design and implementation of architectures for data storage and processing across clouds, HPC and edge-based systems, which address the needs of a large spectrum of applications. The use cases we target to validate our research results come from the following domains.

4.1 Climate and meteorology

The European Centre for Medium-Range Weather Forecasts (ECMWF) [35] is one of the largest weather forecasting centers in the world that provides data to national institutions and private clients. ECMWF's production workflow collects data at the edge through a large set of sensors (satellite devices, ground and ocean sensors, smart sensors). This data, approximately 80 million observations per day, is then moved to be assimilated, i.e. analyzed and sorted, before being sent to a supercomputer to feed the prediction models.

The compute and I/O intensive large-scale simulations built upon these models use ensemble forecasting methods for the refinement. To date, these simulations generate approximately 60 TB per hour, while the center predicts an annual increase of 40 % of this volume. Structured datasets called "products" are then generated from this output data and are disseminated to different clients, such as public institutions or private companies, at a rate of 1PB per month transmitted.

In the framework of the ACROSS EuroHPC Project started in 2020, our goal is to participate in the design of a hybrid software stack for the HPC, Big Data and AI domains. This software stack must be compatible with a wide range of heterogeneous hardware technologies and must meet the needs of the trans-continuum ECMWF workflow.

4.2 Earth science

Earthquakes cause substantial loss of life and damage to the built environment across areas spanning hundreds of kilometers from their origins. These large ground motions often lead to hazards such as tsunamis, fires and landslides. To mitigate the disastrous effects, a number of Earthquake Early Warning (EEW) systems have been built around the world. Those critical systems, operating 24/7, are expected to automatically detect and characterize earthquakes as they happen, and to deliver alerts before the ground motion actually reaches sensitive areas so that protective measures can be taken.

Our research aims to improve the accuracy of Earthquake Early Warning (EEW) systems. These systems are designed to detect and characterize medium and large earthquakes before their damaging effects reach a certain location. Traditional EEW methods based on seismometers fail to accurately identify large earthquakes due to their low sensitivity to ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to their propensity to produce noisy data. In addition, GPS stations and seismometers may be deployed in large numbers across different locations and may produce a significant volume of data consequently, affecting the response time and the robustness of EEW systems.

Integrating and processing in a timely manner high-frequency data streams from multiple sensors scattered over a large territory requires high-performance computing techniques and equipments. We therefore design distributed machine learning-based approaches [5] to earthquake detection, jointly with experts in machine learning and Earth data. Our expertise in swift processing of data on edge and cloud infrastructures allows us to learn from the data from the large number of sensors arriving at high sampling rate, without transferring all data to a single point and thus enables real-time alerts.

4.3 Sustainable development through precision agriculture

Feeding the growing world's population is a on-going challenge, especially in view of climate change, which adds a certain level of uncertainty in food production. Sustainable and precision agriculture is one of the answers that can be implemented to partly overcome this issue. Precision agriculture consists in using new technologies to improve crop management by considering environmental parameters such as temperature, soil moisture or weather conditions, for example. These techniques now need to scale up to improve their accuracy. Over recent years, we have seen the emergence of precision agriculture workflows running across the digital continuum, that is to say all the computing resources from the edge to High-Performance Computing (HPC) and Cloud-type infrastructures. This move to scale is accompanied by new problems, particularly with regard to data movements.

CybeleTech [34] is a French company that aims at developing the use of numerical technologies in agriculture. The core products of CybeleTech are based on numerical simulation of plant growth through dedicated biophysical models and machine learning methods extracting knowledge through large databases. To develop its models, CybeleTech collects data from sensors installed on open agricultural plots or in crop greenhouses. Plant growth models take weather variables as input and the accuracy of agronomic indices estimation heavily rely on the accuracy of these variables.

To this purpose, CybeleTech wishes to collect precise meteorological information from large forecasting centers such as the European Center for Medium-Range Weather Forecasting (ECMWF) [35]. This data gathering is not trivial since it involves large data movements between two distant sites under severe time constraints. Our objective in the context of the EUPEX EuroHPC project is to propose new data management techniques and data movement algorithms to accelerate the execution of these hybrid geo-distributed workflows running on large-scale systems in the area of precision agriculture.

4.4 Smart cities

The proliferation of small sensors and devices that are capable of generating valuable information in the context of the Internet of Things (IoT) has exacerbated the amount of data flowing from all connected objects to cloud infrastructures. In particular, this is true for Smart City applications. These applications raise specific challenges, as they typically have to handle small data (in the order of bytes and kilobytes), arriving at high rates, from many geographical distributed sources (sensors, citizens, public open data sources, etc.) and in heterogeneous formats, that need to be processed and acted upon with high reactivity in near real-time.

Our vision is that, by smartly and efficiently combining the data-driven analytics at the edge and in the cloud, it becomes possible to make a substantial step beyond state-of-the-art prescriptive analytics through a new, high-potential, faster approach to react to the sensed data of the smart cities. The goal is to build a data management platform that will enable comprehensive joint analytics of past (historical) and present (real-time) data, in the cloud and at the edge, respectively, allowing to quickly detect and react to special conditions and to predict how the targeted system would behave in critical situations. This vision is the driving objective of our SmartFastData associate team with Instituto Politécnico Nacional, Mexico.

4.5 Botanical Science

Pl@ntNet [32] is a large-scale participatory platform dedicated to the production of botanical data through AI-based plant identification. Pl@ntNet's main feature is a mobile app allowing smartphone owners to identify plants from photos and share their observations. It is used by around 10 million users all around the world (more than 180 countries) and it processes about 400K plant images per day. One of the challenges faced by Pl@ntNet engineers is to anticipate what should be the appropriate evolution of the infrastructure to pass the next spring peak without problems and also to know what should be done the following years.

Our research aims to improve the performance of Pl@ntNet. Reproducible evaluations of Pl@ntNet on large-scale testbed (e.g., deployed on Grid'5000 [31] by **E2C***lab* [11]), aim to optimize its software configurations in order to minimize the user response time.

5 Social and environmental responsibility

5.1 Footprint of research activities

HPC facilities are expensive in capital outlay (both monetary and human) and in energy use. Our work on Damaris supports the efficient use of high performance computing resources. Damaris [3] can help minimize power needed in running computationally demanding engineering applications and can reduce the amount of storage used for results, thus supporting environmental goals and improving the cost effectiveness of running HPC systems.

5.2 Impact of research results

Social impact. One of our target applications is Early Earthquake Warning. We proposed a solution that enables earthquakes classification with an outstandingly perfect accuracy. By enabling accurate identification of strong earthquakes, it becomes possible to trigger adequate measures and save lifes. For this reason, our work was distinguished with an Outstanding Paper Award — Special Track for Social Impact at AAAI-20, an A* conference in the area of Artificial Intelligence. This result has been highlighted by the *Le Monde* journal in its edition of December 28, 2020, in a section entitled: *Ces découvertes scientifiques que le Covid-19 a masquées en 2020*. This collaborative work continued in 2021.

Environmental impact. As presented in Section 4, we have started a collaboration with CybeleTech that we plan to materialize in the framework of the EUPEX EuroHPC project. CybeleTech is a French company specialized in precision agriculture. Within the framework of our collaboration, we propose to focus our efforts on a scale-oriented data management mechanism targeting two CybeleTech use-cases. They address irrigation scheduling for orchards and optimal harvest date for corn, and their models require the acquisition of large volumes of remote data. The results of our collaboration will have concrete applications as they will improve the accuracy of plant growth models and improve decision making for precision agriculture, which directly aims to contribute to sustainable development.

6 Highlights of the year

ETP4HPC White paper We published a vision paper [1] for the European HPC community addressing on the question: What are the main challenges when building Integrated Hardware/Software Ecosystems for the Edge-Cloud-HPC Continuum to address critical scientific, engineering and societal problems?

7 New software and platforms

7.1 New software

7.1.1 Damaris

Keywords: Visualization, I/O, HPC, Exascale, High performance computing

Scientific Description: Damaris is a middleware for I/O and data management targeting large-scale, MPI-based HPC simulations. It initially proposed to dedicate cores for asynchronous I/O in multicore nodes of recent HPC platforms, with an emphasis on ease of integration in existing simulations, efficient resource usage (with the use of shared memory) and simplicity of extension through plug-ins.

Over the years, Damaris has evolved into a more elaborate system, providing the possibility to use dedicated cores or dedicated nodes to in situ data processing and visualization. It proposes a seamless connection to the VisIt visualization framework to enable in situ visualization with minimum impact on run time. Damaris provides an extremely simple API and can be easily integrated into the existing large-scale simulations.

Damaris was at the core of the PhD thesis of Matthieu Dorier, who received an Accessit to the Gilles Kahn Ph.D. Thesis Award of the SIF and the Academy of Science in 2015. Developed in the framework of our collaboration with the JLESC – Joint Laboratory for Extreme-Scale Computing, Damaris was the first software resulted from this joint lab validated in 2011 for integration to the Blue Waters supercomputer project. It scaled up to 16,000 cores on Oak Ridge's leadership supercomputer Titan (first in the Top500 supercomputer list in 2013) before being validated on other top supercomputers. Active development is currently continuing within the KerData team at Inria, where it is at the center of several collaborations with industry as well as with national and international academic partners.

Functional Description: Damaris is a middleware for data management and in-situ visualization targeting large-scale HPC simulations. Damaris enables: In-situ data analysis by using selected dedicated cores/nodes of the simulation platform. Asynchronous and fast data transfer from HPC simulations to Damaris. Semantic-aware dataset processing through Damaris plug-ins, Writing aggregated data (by hdf5 format) or visualizing them either by VisIt or ParaView.

Release Contributions: This version includes the following changes: A Spack package.py file is available through the official Spack repository allowing automated source builds from source. Support for unstructured mesh model types are available for Paraview in-situ visualization.

URL: https://project.inria.fr/damaris/

Contact: Gabriel Antoniu

Participants: Gabriel Antoniu, Lokman Rahmani, Luc Bouge, Matthieu Dorier, Orçun Yildiz, Hadi Salimi, Joshua Charles Bowden

Partner: ENS Rennes

7.1.2 KerA

Name: KerAnalytics

Keywords: Big data, Distributed Storage Systems, Streaming, Real time

Scientific Description: Current state-of-the-art Big Data analytics architectures are built on top of a three-layer stack: data streams are first acquired by the ingestion layer (e.g., Kafka) and then they flow through the processing layer (e.g., Flink) which relies on the storage layer (e.g., HDFS) for storing aggregated data or for archiving streams for later processing. Unfortunately, in spite of potential benefits brought by specialized layers (e.g., simplified implementation), moving large quantities of data through specialized layers is not efficient: instead, data should be acquired, processed and stored while minimizing the number of movements. To that end, we design a unified architecture for stream ingestion and storage which leads to better utilized resources.

We implement and evaluate this architecture through the KerA prototype with the goal of efficiently handling diverse access patterns: low-latency access to streams and/or high-throughput access to objects. We further introduce a shared virtual log-structured storage approach for improving the cluster throughput when multiple producers and consumers write and read data streams in parallel. Stream partitions are associated with shared replicated virtual logs transparently to the user, effectively separating the implementation of stream partitioning (and data ordering) from data replication (and durability).

KerA was at the core of the PhD thesis of Ovidiu-Cristian Marcu. It was developed on top of RAMCloud, a low-latency key-value distributed system, and serves as a playground for our experimental research in the area of cloud/edge storage.

Functional Description: KerA is a software prototype which addresses the limitations of state-of-the-art storage solutions for stream processing. It illustrates the idea of a unified storage strategy for data ingestion and storage, while allowing low-latency access to streams and/or high throughput access to unbounded streams and objects.

News of the Year: Latest developments: in 2021, KerA was added support for containers allowing to facilitate further deployments. In particular, we aim at providing automatic scaling for stream processing applications.

URL: https://kerdata.gitlabpages.inria.fr/Kerdata-Codes/kera-website/

Publications: hal-01773799, tel-02127065, hal-01532070

Authors: Ovidiu-Cristian Marcu, Gabriel Antoniu, Alexandru Costan, Thomas Bouvier

Contact: Thomas Bouvier

7.1.3 E2Clab

Name: Edge-to-Cloud lab

Keywords: Distributed Applications, Distributed systems, Computing Continuum, Large scale, Experimentation, Evaluation, Reproducibility

Functional Description: E2Clab is a framework that implements a rigorous methodology that provides guidelines to move from real-life application workflows to representative settings of the physical infrastructure underlying this application in order to accurately reproduce its relevant behaviors and therefore understand end-to-end performance. Understanding end-to-end performance means rigorously mapping the scenario characteristics to the experimental environment, identifying and controlling the relevant configuration parameters of applications and system components, and defining the relevant performance metrics.

Furthermore, this methodology leverages research quality aspects such as the Repeatability, Replicability, and Reproducibility of experiments through a well-defined experimentation methodology and providing transparent access to the experiment artifacts and experiment results. This is an important aspect that allows that the scientific claims are verifiable by others in order to build upon them.

URL: https://gitlab.inria.fr/E2Clab/e2clab

Contact: Gabriel Antoniu

8 New results

8.1 Convergence of HPC and Big Data Infrastructures

8.1.1 Provisioning storage resources for hybrid supercomputer/cloud infrastructures

Participants: François Tessier, Julien Monniot, Gabriel Antoniu, Matthieu Robert.

Collaboration. *This work has been carried out in close co-operation with Rob Ross, Argonne National Lab.*

One of the recent axis we are developing in the context of HPC and Big Data convergence concerns the provisioning of storage resources. The way these resources are accessed on supercomputers and clouds opposes a complex low-level vision that requires tight user control (on supercomputers) and a very abstract vision that implies uncertainty for performance modeling (on clouds). Nevertheless, taking full advantage of all available resources is critical in a context where storage is central for coupling workflow components. Our goal is then to make heterogeneous storage resources distributed across HPC+Cloud infrastructures allocatable and elastic to meet the needs of I/O-intensive hybrid workloads.

A first effort has begun in 2021 through the MS internship of Matthieu Robert. Matthieu explored ways of extracting relevant data from I/O traces so they can be replayed in StorAlloc, our simulator of a storage-aware job scheduler. The PhD thesis of Julien Monniot also started in October 2021 on this topic, with a focus on scheduling algorithms for storage-aware job schedulers. It will build upon and extend the preliminary work that was started during the internship.

A publication is in preparation to share our first results. StorAlloc, still under very active development, serves as a testbed for our dynamic storage scheduling algorithms. This simulator is already able to replay and visualize the scheduling of I/O intensive jobs from Theta at Argonne National Laboratory, a 10 PFlops supercomputer.

8.1.2 Supporting seamless execution of HPC-enabled workflows across the Computing Continuum

Participants: Juliette Fournis d'Albiat, Alexandru Costan, François Tessier, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Rosa Badia, Barcelona Supercomputing Center*

Scientific applications have recently evolved to more complex workflows combining traditional HPC simulations with Data Analytics (DA) and ML algorithms. While these applications were traditionally executed on separate infrastructures (e.g., simulations on HPC supercomputers, DA/ML on cloud/edge), the new combined ones need to leverage myriads of resources from the edge to the HPC in order to promptly extract insights. Our goal is to enable seamless deployment, orchestration and optimization of HPC workflows across the Computing Continuum.

To this end, we have started to leverage the building blocks developed at BSC and Inria. In particular, the **E2C***lab* framework already enables the deployment and execution of DA/ML workflows on cloud and edge resources following a reproducibility-oriented methodology. We are exploring the possibility to extend this support workflows including simulations and other compute intensive applications on HPC infrastructures. The layers and services abstractions introduced by **E2C***lab* fit naturally to the approach based on containers for scientific codes pushed forward by BSC. Beyond the ease of deployment, this approach aims to enable the reproducibility of HPC experiments, one of the core design principles of **E2C***lab*.

In the framework of the internship of Juliette Fournis d'Albiat, one direction explored during this year towards such integration is supporting TOSCA-based HPC workflows in **E2C***lab*. Of particular interest are the links and interoperability issues with TOSCA, the Ystia orchestrator and the PyCOMPSs programming model proposed by BSC. A second step consists in investigating the semantics for workflow description that would be compatible with **E2C***lab* (which could possibly be extended) in order to fit the HPC Workflow as a Service paradigm. Finally, a proof-of-concept integrating the support for HPC with **E2C***lab* should be designed and evaluated with real-life applications.

8.1.3 Artificial intelligence-based data analysis in heterogeneous and volatile environments

Participants: Cédric Prigent, Alexandru Costan, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Loïc Cudennec (DGA) and with DFKI in the context of the ENGAGE Inria-DFKI project.*

One of the particularities of AI is to interfere – positively - in all domains: computer-aided design, the search for optimized solutions and of course all business applications with tedious, repetitive and sufficiently complex tasks that cannot be performed by a conventional computer program. Application domains benefiting greatly from these advances are the distributed heterogeneous systems, like the cooperative and autonomous vehicles. One challenge in this context lies in the ability to have relevant data at a given location and at a given time. For this, three mechanisms must be finely studied: data locality, task scheduling and orchestration. These issues call upon many well-studied research themes in homogeneous distributed systems such as supercomputers or large-scale file-sharing systems. But these themes run into new problems once the targeted system is made up of several heterogeneous systems, each bringing different paradigms for task and data management. The research context of the PhD thesis of Cédric Prigent started this year is related to the management of this heterogeneity.

The objective is to identify and adapt emerging approaches from the "computing continuum" in order to address the issues of distribution of calculations and processing, particularly in the case of workflows involving AI. This exploratory thesis has concrete applications such as the SCAF project or on civil warning systems and is situated upstream of these to help guide future technical choices.

During this year we have started to investigate and characterize different middleware for managing shared computations and data on HPC, Cloud, Fog, Edge systems, in centralized (client-server), decentralized (peer-to-peer), or hybrid configurations. We have also identified the communicating systems present in the targeted use cases and characterized them according to the reading grid established previously. We are currently surveying emerging scientific work on "continuum computing", and we are selecting and proposing adaptations for relevant contributions in the case of data intensive workflows.

8.1.4 Memory and data-awareness in hybrid workflows

Participant: François Tessier.

HPC and HPDA opens up the opportunity to solve a wide variety of questions and challenges. The number and complexity of challenges that these two domains can help with are limited by the performance of computer software and hardware. Increasingly, performance is now limited by how fast data can be moved within the memory and storage of the hardware. However, for years, the focus has been primarily on optimizing floating point operations. In the framework of the Maestro EU Project, we proposed a novel memory- and data-aware middleware called whose goal is to improve the performance of data movement in HPC and HPDA [29].

In particular, we have developed a data abstraction layer and an API designed for data orchestration. Our framework has been evaluated on synthetic benchmarks and on components of a production workflows for weather forecast.

8.2 Advanced Data Processing on the Edge-to-Cloud Digital Continuum

8.2.1 Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum

Participants: Daniel Rosendo, Alexandru Costan, Gabriel Antoniu.

In more and more application areas, we are witnessing the emergence of complex workflows that combine computing, analytics and learning. They often require a hybrid execution infrastructure with IoT devices interconnected to cloud/HPC systems (aka *Computing Continuum*). Such workflows are subject to complex constraints and requirements in terms of performance, resource usage, energy consumption and financial costs. This makes it challenging to optimize their configuration and deployment [22].

We propose a methodology [16] to support the optimization of real-life applications on the Edge-to-Cloud Continuum. We implement it as an extension of **E2C***lab* [11], a previously proposed framework supporting the complete experimental cycle across the Edge-to-Cloud Continuum. Our approach relies on a rigorous analysis of possible configurations in a controlled testbed environment (Grid'5000 [31]) to understand their behavior and related performance trade-offs. We illustrate our methodology by optimizing Pl@ntNet [32], a world-wide plant identification application. Our methodology can be generalized to other applications in the Edge-to-Cloud Continuum.

Since real-life Edge-to-Cloud deployments spans among IoT, Edge, Fog, and Cloud/HPC resources, currently we are extending **E2C***lab* to support multiple large-scale testbeds (e.g., Grid'5000 [31] + CHI@Edge + FIT IoT LAB [30]). The ultimate goal is to allow more realistic experimental evaluations through cross-testbed large-scale deployments combining small wireless sensors (FIT IoT LAB), Fog/Edge devices (CHI@Edge), and Cloud/HPC machines (Grid'5000).

8.2.2 Enabling Provenance Capture in Edge-to-Cloud Workflows

Participants: Daniel Rosendo, Alexandru Costan, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Marta Mattoso, Federal University of Rio de Janeiro.*

Supporting reproducibility of experiments carried out on large scale distributed and heterogeneous infrastructures is non-trivial. The experimental methodology, the artifacts used, and the data captured should provide additional context that more accurately explains the experiment execution and results.

In particular, one relevant challenge is the ***provenance capture*** (with a focus on debugging and data analysis support) on such highly heterogeneous and distributed infrastructures [18]. It requires the design and development of novel provenance systems to efficiently capture data from resource constrained hardware resources like IoT/Edge devices (*e.g.* requires smart data capture strategies to reduce the capture overhead).

In this context, meaningful questions that arise are: How do the existing provenance capture systems (e.g., Komadu, DfAnalyzer) perform in resource-constrained environments? How to minimize the provenance capture overhead in IoT/Edge environments? Regarding this research direction, we started to explore the performance (e.g., time required to capture provenance; amount of main memory used; amount of data transmitted) of reference provenance systems such as Komadu and DfAnalyzer to capture provenance in resource constrained IoT/Edge devices. Experimental evaluations will be carried out on real-life IoT/Edge hardware (e.g., CHI@Edge, FIT IoT LAB [30]).

The ultimate goal is that the novel approach could allow users to instrument their application code to capture provenance during the execution of a variety of tasks performed by IoT/Edge devices (e.g., model training, data processing, among others). The proposed approach will be integrated in **E2C***lab* to enable provenance capture in Edge-to-Cloud experiments and allow users to debug and analyze experiment data to better understand the results.

8.2.3 Towards Data Parallel Rehearsal-based Continual Learning

Participants: Thomas Bouvier, Hugo Chaugier, Alexandru Costan, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Bogdan Nicolae (Argonne National Laboratory, USA), who co-advised the internship of Hugo Chaugier.*

Although not a new concept, the adoption of deep learning accelerated in recent years because of the ever-increasing amount of data being generated. Extracting valuable knowledge from it requires training large learning models using parallel and distributed strategies. However as more data accumulates in data centers over time, it becomes prohibitively expensive to re-train deep learning models from scratch to update them with new data. This concern is emphasized in the context of the Edge-to-Cloud Continuum, where 1) real-world sensing applications requires the need for agents that can adapt to continuously evolving environments and 2) compute capacities might be constrained across processing elements.

Besides, standard deep neural networks (DNNs) suffer from the well-known issue of catastrophic forgetting, making continual learning difficult. Specifically, incremental training introduces a bias towards newly learned patterns which makes the model forget about the previously learned ones.

These two challenges, distributing the training process and preventing catastrophic forgetting, have not been studied together yet are essential in the implementation of large scale deep learning. A first effort on that topic has begun in 2021 through the MS internship of Hugo Chaugier. Since then, we studied how different continual learning strategies behave at scale using Grid'5000. In particular, we showed that rehearsal-based methods did not suffer from data parallelization but on the contrary benefited from it, as increasing the total rehearsal memory size improves the ability of the DNN to remember previous tasks.

8.2.4 Deploying Heterogeneity-aware Deep Learning Workloads on the Computing Continuum

Participants: Thomas Bouvier, Alexandru Costan, Gabriel Antoniu.

The Computing Continuum led to the evolution of centralized data centers towards interconnected processing elements spanning from edge devices to cloud data centers. While this type of infrastructure benefits real-time analysis by reducing latency, the network and compute heterogeneity across and within clusters may negatively impact deep learning workloads.

We conducted preliminary experiments on Grid'5000 using the methodology implemented in **E2C***lab* to help reproduce both relevant behaviors of the given DL workloads and representative settings of the physical infrastructure underlying the continuum. The goal is to understand the end-to-end performance of deep learning training in such heterogeneous settings in order to identify the main factors leading to stragglers. We propose such an approach in [20] and [17]. The architectural differences of DNNs are compared to help predict inefficient gradients computation and propagation across nodes.

The outcomes will allow to devise novel intra- and inter-cluster strategies to address infrastructure bottlenecks in DL scenarios. This should lead to the design of a middleware helping to distribute DL training, accounting for deterministic and dynamic of both network and compute heterogeneity. Various policies should improve the following objectives while ensuring system scalability: makespan, cost and fairness. Eventually, it will be interesting to study the extent to which these strategies apply to continual learning.

8.2.5 Modeling distributed stream processing across the Edge-to-Cloud Continuum

Participants: Alexandru Costan, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Daniel Balouek, University of Utah, and Pedro De Souza Bento Da Silva, formerly a post-doc student in the team, and now at Hasso Plattner Institute, Berlin, Germany.*

The growth of the Internet of Things is resulting in an explosion of data volumes at the Edge of the Internet. To reduce costs incurred due to data movement and centralized cloud-based processing, it is becoming increasingly important to process and analyze such data closer to the data sources. Exploiting Edge computing capabilities for stream-based processing is however challenging. It requires addressing the complex characteristics and constraints imposed by all the resources along the data path, as well as the large set of heterogeneous data processing and management frameworks. Consequently, the community needs tools that can facilitate the modeling of this complexity and can integrate the various components involved.

To tap into the power of the Computing Continuum, we propose MDSC [13], an approach for modeling distributed stream-based applications running across the Edge-to-Cloud continuum. The objective of MDSC is to facilitate the definition and evaluation of alternative deployment options for such applications atop the Edge-to-Cloud Continuum, in a versatile way. To enable this vision, MDSC leverages a set of simple but expressive architectural concepts (e.g., layers, processing units, buses), which aim to describe the essential components of the applications and to make more explicit their inter-relations and functionalities. These abstractions further enable grouping for components of similar nature and allow to characterize the communication among those groups. Subsequently, this model facilitates the definition of alternative mappings onto the underlying edge/fog/cloud infrastructure.

We demonstrate how MDSC can be applied to a concrete real-life ML-based application - early earthquake warning - to help answer questions such as: when is it worth decentralizing the classification load from the Cloud to the Edge and how?

8.3 Scalable I/O, Communication, in-situ Visualization and Analysis on HPC Systems at Extreme Scales

8.3.1 Exploring Alternative Strategies Leveraging Dedicated Resources for in-situ visualization for Large Scale Simulations

Participants: Joshua Bowden, Gabriel Antoniu, François Tessier.

Collaboration. *This work has been carried out in close co-operation with collaborators within the PRACE 6IP project, Simone Bna (CINICA), Miroslav Puskaric, Martin Bidner (HLRS), and with support through the EDF Code_Saturne user group.*

As the exascale era approaches, maintaining scalable performance in data management tasks (storage, visualization, analysis, etc.) remains a key challenge in enabling high performance for the application execution. To address this challenge, the Damaris approach proposed by our team leverages dedicated computational resources in multicore nodes to offload data management tasks, including I/O, data compression, scheduling of data movements, in-situ analysis, and visualization. In this study, we evaluate the benefits of Damaris to improve the efficiency of in-situ visualization for Code_Saturne, a fluid dynamics modeling environment. The experiments show Damaris to adequately hide the I/O processing of various visualization processing pipelines in Code_Saturne using the Paraview visualization engine. In all cases the Damaris enabled version of Code_Saturne was found to be more efficient than the identical non-Damaris capable version when running the same Paraview pipeline.

This work investigates and evaluates alternative methods for allocation of resources for Damaris using either dedicated nodes or dedicated cores for visualization within multicore nodes. It shows what problems asynchronous processing can encounter due to under-allocation of resources. The paper also presents an analysis of the integration effort required to add the Damaris functionality to Code_Saturne, which is another important consideration when deciding to integrate a library. The Damaris Paraview interface has now implemented support for an unstructured-grid mesh type, thus opening up a large domain of simulation types relying on this mesh model. This mesh type was a requirement for the interface with Code_Saturne which uses these type extensively. Results have presented to the Code_Saturne development team at EDF and were published as a PRACE white paper [21] available from the PRACE website.

8.3.2 Scalable asynchronous I/O and in-situ processing with Damaris for Carbon Sequestration

Participants: Joshua Bowden, Alexandru Costan, François Tessier, Gabriel Antoniu.

Collaboration. *This work has been carried out in close co-operation with Atgeirr Rasmussen (SINTEF) within the framework of the EuroHPC H2020 ACROSS project.*

Carbon capture and storage (CCS) is one of the technologies that have a large potential for mitigating CO2 emissions, and can also enable carbon-negative processes. Before one can commit to large-scale carbon storage operations, it is essential to do simulation studies to assess the storage potential and safety of the operation and to optimize the placement and operation of injection wells. Such a simulation is done by computer programs that solve the equations that describe the motion and state of the fluids within the porous rocks. In the ACROSS project, we use OPM Flow, an open-source reservoir simulator program suitable for both industrial uses as well as research.

As such large-scale simulations can take a long time to run, and require significant high-performance computing resources, we investigate how asynchronous I/O and in-situ processing can help improve the performance, scaling, and efficiency of OPM Flow and of workflows using the program. The ACROSS project is using a co-development method, where software requirements inform the hardware design process for next generation HPC systems. The flow software is typical of MPI based simulation software where I/O inhibits the scaling of the simulation to larger machine sizes due to its serialized nature. We started to investigate how the Damaris approach could be leveraged by OPM Flow to provide asynchronous I/O.

In addition, we started to investigate methods for asynchronous analytics that could be integrated into the Damaris library. This includes exploratory work for integration of a Python interface to a distributed server component based on Ray or Dask. The result of this exploratory work has converged on the use of Python with Dask as Dask offers a suite of useful distributed analytic methods using familiar Pythonesque interfaces, similar to NumPy and Pandas. Pursuing this direction will enable new usage possibilities for Damaris that could be further studied in collaboration with CEA within the future NumPEx exploratory PEPR project, currently under discussion.

9 Partnerships and cooperations

9.1 International initiatives

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

UNIFY

Title: *Intelligent Unified Data Services for Hybrid Workflows Combining Compute-Intensive Simulations and Data-Intensive Analytics at Extreme Scales*

Duration: 2019 -

Coordinator: Gabriel Antoniu.

Partners: Argonne National Laboratory (United States)

Inria contact: Gabriel Antoniu

Partner coordinator: Tom PETERKA (tpeterka@mcs.anl.gov)

Summary: The landscape of scientific computing is being radically reshaped by the explosive growth in the number and power of digital data generators, ranging from major scientific instruments to the Internet of Things (IoT) and the unprecedented volume and diversity of the data they generate. This requires a rich, extended ecosystem including simulation, data analytics, and learning applications, each with distinct data management and analysis needs. Science activities are beginning to combine these techniques in new, large-scale workflows, in which scientific data is produced, consumed, and analyzed across multiple distinct steps that span computing resources, software frameworks, and time. This paradigm introduces new data-related challenges at several levels. The UNIFY Associate Team aims to address three such challenges. First, to allow scientists to obtain fast, real-time insights from complex workflows combining extreme-scale computations with data analytics, we will explore how recently emerged Big Data processing techniques (e.g., based on stream processing) can be leveraged with modern in situ/in transit processing approaches used in HPC environments. Second, we will investigate how to use transient storage systems to enable efficient, dynamic data management for hybrid workflows combining simulations and analytics. Finally, the explosion of learning and AI provides new tools that can enable much more adaptable resource management and data services than available today, which can further optimize such data processing workflows.

Activity in 2021:

- Alexandru Costan (Inria), Gabriel Antoniu (Inria) and Bogdan Nicolae (ANL) co-advised the internship of Hugo Chaugier (Inria) on data parallel rehearsal-based continual learning.

9.1.2 Inria associate team not involved in an IIL or an international program

SmartFastData

Title: Efficient Data Management in Support of Hybrid Edge/Cloud Analytics for Smart Cities

Duration: 2019 -

Coordinator: Alexandru Costan

Partners: Centro de Investigación en Computación, Instituto Politécnico Nacional (Mexico)

Inria contact: Alexandru Costan

Partner coordinator: Rolando Menchaca-Mendez (rolando.menchaca@gmail.com)

Website: team.inria.fr/smartfastdata

Summary: The proliferation of small sensors and devices that are capable of generating valuable information in the context of the Internet of Things (IoT) has exacerbated the amount of data flowing from all connected objects to private and public cloud infrastructures. In particular, this is true for Smart City applications, which cover a large spectrum of needs in public safety, water and energy management. Unfortunately, the lack of a scalable data management subsystem is becoming an important bottleneck for such applications, as it increases the gap between their I/O requirements and the storage performance.

The vision underlying the SmartFastData associated team is that, by smartly and efficiently combining the data-driven analytics at the edge and in the cloud, it becomes possible to make a substantial step beyond state-of-the-art prescriptive analytics through a new, high-potential, faster approach to react to the sensed data. The goal is to build a data management platform that will enable comprehensive joint analytics of past (historical) and present (real-time) data, in the cloud and at the edge, respectively, allowing to quickly detect and react to special conditions and to predict how the targeted system would behave in critical situations.

Activity in 2021:

- Alexandru Costan (Inria), Gabriel Antoniu (Inria), Rolando Menchaca (IPN), José Aguilar (IPN), Edgar Romo (IPN) have set the outline of a joint publication based on the main objective of the associate team: exploring analytical models for performance evaluation of stream storage and ingestion systems in Edge/Fog environments. While no internships or research visits were possible, instead, we organized a virtual collaborative working environment through regular video-conferences on specific tasks of Objectives 1 and 2 with each of the members involved.

9.1.3 Participation in other International Programs

FlexStream: Automatic Elasticity for Stream-based Applications

Program: PHC PROCOPE 2020

Project acronym: FlexStream

Project title: Automatic Elasticity for Stream-based Applications

Duration: January 2020–December 2021

Coordinator: Alexandru Costan

Other partners: University of Dusseldorf (UDUS)

Summary: This project aims at developing concepts providing automatic scaling for stream processing applications. In particular, FlexStream aims at developing and evaluating a prototype which will integrate a stream ingestion-system from IRISA and an in-memory storage from UDUS. For this approach a tight cooperation is mandatory in order to be successful which in turn requires visits on both sides and longer exchanges, especially for the involved PhD students, in order to allow an efficient integrated software design, development as well as joint experiments on large platforms and preparing joint publications.

In 2021, we focused on the evaluation of the KerA ingestion system (developed by KerData) with different network backends (including Infiniband) compared to other state of the art ingestion systems (e.g., Kafka). This work was done in the context of the Master internship of Jakob Scheumann at UDUS (co-advised with Alexandru Costan).

9.2 European initiatives

9.2.1 FP7 & H2020 projects

EuroHPC ACROSS

Title: HPC, Big Data, and Artificial Intelligence convergent platform

Duration: March 2021- Feb 2024

Coordinator: LINKS Foundation, Italy

Partners: CINECA, IT4I, Atos, Avio Aero, Morfo, Neuopublic, LINKS, CINI, SINTEF, MPI-M, DELTARES, ECMWF

Inria contact: Gabriel Antoniu

Summary: ACROSS will combine traditional HPC techniques with Artificial Intelligence, such as Machine Learning and Deep Learning, and Big Data analytic techniques to enhance the application test case outcomes. The performance of Machine Learning and Deep Learning will be accelerated by using dedicated hardware devices.

Web site: www.acrossproject.eu

9.2.2 Collaborations in European Programs, except FP7 and H2020

PRACE 6th Implementation Phase Project (PRACE6-IP)

Program: H2020 RIA European project, call H2020-INFRAEDI-2018-1

Project acronym: PRACE-6IP

Project title: PRACE 6th Implementation Phase Project

Duration: May 2019-Dec 2021

Coordinator: FZJ

Other partners: HLRS, LRZ, GENCI, CEA, CINES, CNRS, IDRIS, Inria, EPCC, BSC, CESGA, CSC, ETH-CSCS, SURFsara, KTH-SNIC, CINECA, PSNC, CYFRONET, WCNS, UiOsingma2, GRNET, UC-LCA, Univ MINHO, ICHEC, UHEM, CASTORCm NCSA, IT4I-VSB, KIFU, UL, CCSAS, CENAERO, Univ Lux, GEANT

Web site: cordis.europa.eu/project/id/823767

Participants: Joshua Bowden, Gabriel Antoniu.

PRACE, the Partnership for Advanced Computing is the permanent pan-European High Performance Computing service providing world-class systems for world-class science. Systems at the highest performance level (Tier-0) are deployed by Germany, France, Italy, Spain and Switzerland, providing researchers with more than 17 billion core hours of compute time. HPC experts from 25 member states enabled users from academia and industry to ascertain leadership and remain competitive in the Global Race. Currently PRACE is finalizing the transition to PRACE 2, the successor of the initial five year period.

The objectives of PRACE-6IP are to build on and seamlessly continue the successes of PRACE and start new innovative and collaborative activities proposed by the consortium. These include: assisting the development of PRACE 2; strengthening the internationally recognized PRACE brand; continuing and extend advanced training which so far provided more than 36 400 person-training days; preparing strategies and best practices towards Exascale computing, work on forward-looking SW solutions; coordinating and enhancing the operation of the multi-tier HPC systems and services; and supporting users to exploit massively parallel systems and novel architectures. A high level Service Catalog is provided. The proven project structure will be used to achieve each of the objectives in 7 dedicated work packages. The activities are designed to increase Europe's research and innovation potential especially through: seamless and efficient Tier-0 services and a pan-European HPC ecosystem including national capabilities; promoting take-up by industry and new communities and special offers to SMEs; assistance to PRACE 2 development; proposing strategies for deployment of leadership systems; collaborating with the ETP4HPC, Cows and other European and international organizations on future architectures, training, application support and policies. This will be monitored through a set of KPIs.

In PRACE-6IP, the Damaris framework developed by the KerData team is being experimented as a technology to provide a service for in situ visualization and processing to PRACE users. In this context, a demonstrator is currently being built using Damaris for Code_Saturne, a CFD application.

9.2.3 Other European programs/initiatives

The ENGAGE Inria-DFKI project proposal

Participants: Gabriel Antoniu, Alexandru Costan, Thomas Bouvier, Daniel Rosendo.

In the area of HPC-AI convergence, Gabriel Antoniu coordinated the ENGAGE Inria-DFKI project proposal (submitted, under evaluation). In addition to the KerData team, it involves two other Inria teams: HiePACS (Bordeaux) and DATAMOVE (Grenoble). It aims to create foundations for a new generation of high-performance computing (HPC) environments for Artificial Intelligence (AI) workloads. The basic premise for these workloads is that in the future, training data for Deep Neural Networks (DNN) will no longer only be stored and processed in epochs, but rather be generated on-the-fly using parametric models and simulations. This is particularly useful in situations where obtaining data by other means is expensive or difficult or where a phenomenon has been predicted in theory, but not yet observed. One key application of this approach is to validate and certify AI systems through targeted testing with synthetically generated data from simulations.

The project proposes contributions on three levels: On the application level, it will address the question how the adaptive sampling of parameter spaces will allow for better choices on what data to generate. On the middleware level, it will address the question how virtualization and scheduling need to be adapted to facilitate and optimize the execution of resulting mixed workloads consisting of training and simulation tasks, running on potentially hybrid (HPC/cloud/edge) infrastructures. On the resource management level, it will contribute to novel strategies to optimize memory management and the dynamic choice of parallel resources to run the training and inference phases.

In summary, the project will create a blueprint for a new generation of AI compute infrastructures that goes beyond the concept of epoch-based data management and considers model-based online-training of Neural Networks as the new paradigm for DNN applications.

9.2.4 Collaborations with Major European Organizations

Participants: Gabriel Antoniu, Alexandru Costan.

Appointments by Inria in relation to European bodies

Big Data Value Association (BDVA) and ETP4HPC: Since 2018, Gabriel Antoniu and Alexandru Costan have been serving as Inria representatives in the working groups dedicated to *HPC-Big Data* convergence .

Community service at European level in response to external invitations

ETP4HPC: Since 2019: Gabriel Antoniu has served as a co-leader of the working group on Programming Environments and co-leader of two research clusters, contributing to the Strategic Research Agenda of ETP4HPC, published in March 2020. Alexandru Costan served as a member of these working groups. Activity is now continuing through regular meetings to refine the ETP4HPC technical focus areas to prepare the future edition of the SRA that should be published in 2022.

Transcontinuum Initiative (TCI) : In 2020, as a follow-up action to the publication of its Strategic Research Agenda, ETP4HPC initiated a collaborative initiative called TCI (Transcontinuum Initiative). It gathers major European associations in the areas of HPC, Big Data, AI, 5G, Cybersecurity, including ETP4HPC, BDVA, CLAIRE, HIPEAC, 5G IA, ECO). It aims to strengthen research and industry in Europe to support the Digital Continuum - infrastructure (including HPC systems, clouds, edge infrastructures) by helping to define a set of research focus areas/topics requiring interdisciplinary action. The expected outcome of this effort is the co-editing of multidisciplinary calls for projects to be funded by the European Commission. Gabriel Antoniu is in charge of ensuring the BDVA-ETP4HPC coordination and of co-animating the working group dedicated to the definition of representative application use cases.

Big Data Value Association: Gabriel Antoniu was asked by BDVA to coordinate Diva's contribution to the TCI initiative recently started (see above). He also participated to the organization of a joint BDVA-ETP4HPC seminar on HPC, Big Data, IoT and AI future industry-driven collaborative strategic topics.

9.3 National initiatives

9.3.1 ANR

OverFlow (2015–2021)

Participants: Alexandru Costan, Daniel Rosendo, Gabriel Antoniu.

Project Acronym: OverFlow
Project Title: Workflow Data Management as a Service for Multisite Applications
Coordinator: Alexandru Costan
Duration: October 2015–March 2021
Other Partners: None (Young Researcher Project, JCJC)
External collaborators: Kate Keahey (University of Chicago and Argonne National Laboratory), Bogdan Nicolae (Argonne National Lab)
Web site: sites.google.com/view/anroverflow

This project investigates approaches to data management enabling an efficient execution of geographically distributed workflows running on multi-site clouds.

In 2021, we continued our work on the reproducibility aspects of the workflows deployed on hybrid edge-cloud infrastructures. We focused on capturing provenance data, which plays an important role in reproducibility. In particular, we studied how do the existing provenance systems (e.g., Komadu or DfAnalyzer) perform in resource-constrained environments and how to minimize the provenance capture overhead in IoT/Edge environments.

9.3.2 Other National Projects

HPC-Big Data Inria Inria Challenge (ex-IPL)

Participants: Daniel Rosendo, Gabriel Antoniu, Alexandru Costan.

Collaboration. *This work has been carried out in close co-operation with Pedro De Souza Bento Da Silva, formerly a post-doc student in the team, and now at Hasso Plattner Institute, Berlin, Germany.*

Project Acronym: HPC-BigData
Project Title: The HPC-BigData INRIA Challenge
Coordinator: Bruno Raffin
Duration: 2018–2022
Web site: project.inria.fr/hpcbigdata

The goal of this HPC-BigData IPL is to gather teams from the HPC, Big Data and Machine Learning (ML) areas to work at the intersection between these domains. Research is organized along three main axes: high performance analytics for scientific computing applications, high performance analytics for big data applications, infrastructure and resource management. Gabriel Antoniu is a member of the Advisory Board and leader of the Frameworks work package.

In 2021 we proposed a methodology [16] to support the optimization of real-life applications on the Edge-to-Cloud Continuum. We implement it as an extension of **E2C***lab* [11], a previously proposed framework supporting the complete experimental cycle across the Edge-to-Cloud Continuum.

ADT Damaris 2

Participants: Joshua Charles Bowden, Gabriel Antoniu.

Project Acronym: ADT Damaris 2
Project Title: Technology development action for the Damaris environment
Coordinator: Gabriel Antoniu
Duration: 2019–2022
Web site: project.inria.fr/damaris

This action aims to support the development of the Damaris software. Inria's *Technological Development Office* (D2T, *Direction du Développement Technologique*) provided 3 years of funding support for a senior engineer.

In April 2020, Joshua Bowden was hired on this position. He introduced a support for unstructured mesh model types to Damaris. This capability opens up the use of Damaris for a large number of simulation types that depend on this data structure, in the areas of Computational Fluid dynamics, which has applications in energy production and combustion modeling, electric modeling and atmospheric and flow.

The capability has been developed and tested using Code_Saturne, a finite volume computational fluid dynamics (CFD) simulation environment. Code_Saturne is an open source CFD modeling environment which supports both single phase and multi-phase flow and includes modules for atmospheric flow, combustion modeling, electric modeling and particle tracking. This work is being validated on PRACE Tier-0 computing infrastructure in the framework of the PRACE-6IP project.

Grid'5000 We are members of Grid'5000 community and run experiments on the Grid'5000 platform on a daily basis.

10 Dissemination

Participants: Gabriel Antoniu, Luc Bougé, Alexandru Costan, François Tessier, Daniel Rosendo, Thomas Bouvier, Joshua Bowden.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

Luc Bougé: Steering Committee Chair of the Euro-Par International Conference on Parallel and Distributed Computing. Euro-Par celebrated its 25th anniversary in Göttingen, Germany, in 2019.

Gabriel Antoniu: Steering Committee Chair of the HPS Workshop series on High-Performance Storage, held in conjunction with the IEEE IPDPS conference since 2020. General Chair of HPS'21.

François Tessier: Co-Chair of the 4th and 5th SuperCompCloud, the Workshop on Interoperability of Supercomputing and Cloud Technologies respectively held in conjunction with ISC'21 and Supercomputing'21.

Alexandru Costan: Co-Chair of the 11th ScienceCloud Workshop on Scientific Cloud Computing, held in conjunction with HPDC'21.

Member of the organizing committees

François Tessier: Member of the organizing committee of the 4th and 5th SuperCompCloud, the Workshop on Interoperability of Supercomputing and Cloud Technologies respectively held in conjunction with ISC'21 and Supercomputing'21.

10.1.2 Scientific events: selection

Gabriel Antoniu: Member of the Best Paper Award and Best Student Paper Award Committee for ACM/IEEE SC21.

Chair of conference program committees

Gabriel Antoniu: Track Chair of the HPC Asia conference

François Tessier:

- Track Chair of the SOS24 conference - Invitation-only series of interactive workshops on Distributed Supercomputing, digital event, March 2021.
- Minisymposium organizer at the PASC'21 conference - Platform for Advanced Scientific Computing, digital event, June 2021.

Member of the conference program committees

Alexandru Costan: IEEE/ACM SC'21 (Posters and ACM Student Research Competition, Best Poster Award Committee), IEEE/ACM UCC 2021, IEEE CloudCom 2021.

Gabriel Antoniu: ACM/IEEE SC21, IEEE Cluster 2021.

François Tessier: IEEE/ACM CCGrid 2021; ACM ICPP 2021; ACM HPC Asia 2021; COMPAS 2021;

Reviewer

Alexandru Costan: ACM HPDC 2021, IEEE IPDPS 2021, IEEE/ACM CCGrid 2021, IEEE Cluster 2021

François Tessier: IEEE Cluster 2021; IEEE IPDPS 2021; Supercomputing 2021;

Joshua Bowden: ICPP 2021; IEEE Cluster 2021; IEEE IPDPS 2021; Supercomputing 2021;

10.1.3 Journal

Member of the editorial boards

Gabriel Antoniu: Associate Editor of JPDC - the Elsevier Journal of Parallel and Distributed Computing.

François Tessier: IEEE Transactions on Parallel and Distributed Systems (TPDS) journal, Special Section on Innovative R&D toward the Exascale Era in 2021.

Reviewer - reviewing activities

Alexandru Costan: IEEE Transactions on Parallel and Distributed Systems, Future Generation Computer Systems, Concurrency and Computation Practice and Experience, IEEE Transactions on Cloud Computing, Journal of Parallel and Distributed Computing.

Gabriel Antoniu: IEEE Transactions on Parallel and Distributed Systems, SoftwareX.

Thomas Bouvier: IEEE International Parallel and Distributed Processing Symposium, Springer Computing.

Daniel Rosendo: IEEE International Parallel and Distributed Processing Symposium, Springer Computing.

Joshua Bowden: IEEE International Parallel and Distributed Processing Symposium, Springer Computing.

10.1.4 Invited talks

Gabriel Antoniu was invited to give a keynote talk on the 10-year collaboration with Argonne National Laboratory within the JLESC international Laboratory (virtual, February 2021).

Daniel Rosendo:

- **E2C***lab*: Optimizing Complex Workflow Deployments on the Edge-to-Cloud Continuum. 12th JLESC Workshop, February 24th, 2021.
- Enabling Reproducible Analysis of Complex Application Workflows on the Edge-to-Cloud Continuum. Experimentation Users-Group Meetings at Inria, December 7th, 2021.
- Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum. 13th JLESC Workshop, December 14th, 2021.

Thomas Bouvier:

- Deploying Heterogeneity-aware Deep Learning Workloads on the Computing Continuum. 13th JLESC Workshop, December 16th, 2021.

François Tessier:

- Storage allocation over hybrid HPC/Cloud Infrastructures. 12th JLESC Workshop, February 25th, 2021.

Joshua Bowden:

- Code_Saturne: Integration with Damaris - The potential of asynchronous I/O. Code_Saturne User Group Meeting, September 15th, 2021.
- Asynchronous I/O using Damaris - Results from integration with Code_Saturne CFD code. 13th JLESC Workshop, December 16th, 2021.

10.1.5 Leadership within the scientific community

Gabriel Antoniu:

TCI: Since 2020, co-leader of the Use-Case Analysis Working Group. TCI (The Transcontinuum Initiative) emerged as a collaborative initiative of ETP4HPC, BDVA, CLAIRE and other peer organizations, aiming to identify joint research challenges for leveraging the HPC-Cloud-Edge computing continuum and make recommendations to the European Commission about topics to be funded in upcoming calls for projects.

ETP4HPC: Since 2019, co-leader of the working group on Programming Environments and co-lead of two research clusters, contributing to the Strategic Research Agenda of ETP4HPC (next edition to appear in 2022).

International lab management: *Vice Executive Director of JLESC* for Inria. JLESC is the Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing. Within JLESC, he also serves as a *Topic Leader* for Data storage, I/O and in situ processing for Inria.

Team management: *Head of the KerData Project-Team* (INRIA-ENS Rennes-INSA Rennes).

International Associate Team management: Leader of the UNIFY Associate Team with Argonne National Lab (2019–2021).

Technology development project management: Coordinator of the Damaris 2 ADT project (2019–2022).

Luc Bougé:

SIF: Co-Vice-President of the *French Society for Informatics* (*Société informatique de France*, SIF), in charge of the Teaching Department.

CoSSAF: Member of the Foundation Committee for the *College of the French Scientific Societies* (CoSSAF, *Collège des sociétés savantes académiques de France*) gathering more than 40 French societies from all domains.

Alexandru Costan:

International Associate Team management: Leader of the SmartFastData Associate Team with Instituto Politécnico Nacional, Mexico City (2019–2021).

François Tessier:

JLESC: Member of the organizing committee of the bi-annual workshops of the Joint Laboratory for Extreme Scale Computing.

10.1.6 Scientific expertise

Alexandru Costan:

Estonian Research Council: Reviewer for the Postdoctoral and Young Researcher Grants.

GDR RSD: Member of the juries for the PhD award and Young Researcher award.

Thomas Bouvier: Participated in the hackAtech as a technical coach. The goal was to promote Inria technologies by conceiving a startup project within 72 hours, from September 30th to October 2nd. "Partner" challenge prize awarded by Ouest France.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

Alexandru Costan:

- Bachelor: Software Engineering and Java Programming, 28 hours (lab sessions), L3, INSA Rennes.
- Bachelor: Databases, 68 hours (lectures and lab sessions), L2, INSA Rennes.
- Bachelor: Practical case studies, 24 hours (project), L3, INSA Rennes.
- Master: Big Data Storage and Processing, 28h hours (lectures, lab sessions), M1, INSA Rennes.

- Master: Algorithms for Big Data, 28 hours (lectures, lab sessions), M2, INSA Rennes.
- Master: Big Data Project, 28 hours (project), M2, INSA Rennes.

Gabriel Antoniu:

- Master (Engineering Degree, 5th year): Big Data, 24 hours (lectures), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz.
- Master: Scalable Distributed Systems, 10 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School.
- Master: Infrastructures for Big Data, 10 hours (lectures), M2 level, IBD Module, SIF Master Program, University of Rennes.
- Master: Cloud Computing and Big Data, 14 hours (lectures), M2 level, Cloud Module, MIAGE Master Program, University of Rennes.
- Master (Engineering Degree, 5th year): Big Data, 16 hours (lectures), M2 level, IMT Atlantique, Nantes.

François Tessier:

- Bachelor: Computer science discovery, 18 hours (lab sessions), L1 level, DIE Module, ISTIC, University of Rennes.
- Bachelor: Introduction to operating systems, 12 hours (lab sessions), L3 level, SIN Module, MIAGE Licence Program, University of Rennes.
- Master (Engineering Degree, 4th year): Storage on Clouds, 5 hours (lecture and lab session), M2 level, IMT Atlantique, Rennes.

Daniel Rosendo:

- Master: Miage BDDA, 24 hours (lab sessions), M2, ISTIC Rennes.
- Bachelor: Algorithms for Big Data, 10 hours (lectures, lab sessions), INSA Rennes.

Thomas Bouvier:

- Bachelor: Databases, 36 hours (lectures and lab sessions), L2, INSA Rennes.

10.2.2 Supervision

PhD in progress

Daniel Rosendo: *Enabling HPC-Big Data Convergence for Intelligent Extreme-Scale Analytics*, PhD started in October 2019, co-advised by Gabriel Antoniu, Alexandru Costan and Patrick Valduriez (Inria).

Thomas Bouvier: *Reproducible deployment and scheduling strategies for AI workloads on the Computing Continuum*, PhD started in January 2021, co-advised by Alexandru Costan and Gabriel Antoniu.

Julien Monniot: *Bridging Supercomputers and Clouds at the Exascale Era Through Elastic Storage*, PhD started in October 2021, co-advised by François Tessier and Gabriel Antoniu.

Cédric Prigent: *Supporting Online Learning and Inference in Parallel across the Digital Continuum*, PhD started in November 2021, co-advised by Alexandru Costan, Gabriel Antoniu and Loïc Cudennec.

Internships

Juliette Fournis d'Albiat: *Supporting Seamless Execution of HPC-enabled Workflows Across the Computing Continuum*, 10-month internship started in October 2021 (6 months in France, 4 months in Spain), co-advised by Alexandru Costan, Gabriel Antoniu, François Tessier and Rosa Badia, Jorge Ejarque (Barcelona Supercomputing Center, Spain).

Hugo Chaugier: *Data Parallel Rehearsal-based Continual Learning*, 6-month internship from February to July 2021, co-advised by Alexandru Costan, Gabriel Antoniu and Bogdan Nicolae (Argonne National Laboratory).

Matthieu Robert: *Algorithms for Dynamic Provisioning of Storage Resources on Supercomputers*, 6-month internship from February to July 2021, co-advised by François Tessier and Gabriel Antoniu.

10.2.3 Juries

Gabriel Antoniu: member of the PhD jury of Xenia Human, ENS Lyon, 16 December 2021.
François Tessier: PhD mid-term evaluation of Amal Gueroudji, Maison de la simulation.

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

Gabriel Antoniu:

ETP4HPC and BDVA: Inria representative in the working groups of BDVA and ETP4HPC dedicated to HPC-Big Data convergence.

Luc Bougé:

ANR: Scientific officer at the *National Research Agency* (ANR, *Agence nationale de la recherche*) in charge of data management.

PNRIA: Programme National de Recherche en Intelligence Artificielle (*National Research Program for Artificial Intelligence*). Member of the management team for the IA for Humanity governmental program launched in March 2018. Bertrand Braunschweig, Inria, and now Isabelle Herlin, are the scientific directors of the program.

Alexandru Costan:

In charge of internships at the Computer Science Department of INSA Rennes.
In charge of the organization of the IRISA D1 Department Seminars.
In charge of the management of the KerData team access to Grid’5000.

10.3.2 Interventions

Luc Bougé:

ANR: Organization of various seminars and training about Large-Scale Data Management at *National Research Agency* (ANR, *Agence nationale de la recherche*). It focused on using Web API to access databases, for instance HAL and scanR, the search engine for French research and innovation.

François Tessier:

ENS Rennes Student Seminar: "Dynamic Provisioning of Storage Resources on Hybrid Infrastructures", March 2021.

11 Scientific production

11.1 Major publications

[1] G. Antoniu, P. Valduriez, H.-C. Hoppe and J. Krüger. *Towards Integrated Hardware/Software Ecosystems for the Edge-Cloud-HPC Continuum*. 2021. DOI: 10.5281/zenodo.5534464. URL: https://hal.archives-ouvertes.fr/hal-03358930.

[2] N. Cheriere, M. Dorier and G. Antoniu. 'How fast can one resize a distributed file system?' In: *Journal of Parallel and Distributed Computing* 140 (June 2020), pp. 80–98. DOI: 10.1016/j.jpdc.2020.02.001. URL: https://hal.archives-ouvertes.fr/hal-02961875.

[3] M. Dorier, G. Antoniu, F. Cappello, M. Snir, R. Sisneros, O. Yildiz, S. Ibrahim, T. Peterka and L. Orf. 'Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations'. In: *ACM Transactions on Parallel Computing* 3.3 (2016), p. 15. DOI: 10.1145/2987371. URL: https://hal.inria.fr/hal-01353890.

[4] M. Dorier, S. Ibrahim, G. Antoniu and R. Ross. 'Using Formal Grammars to Predict I/O Behaviors in HPC: the Omnisc'IO Approach'. In: *IEEE Transactions on Parallel and Distributed Systems* (2016). DOI: 10.1109/TPDS.2015.2485980. URL: https://hal.inria.fr/hal-01238103.

[5] K. Fauvel, D. Balouek-Thomert, D. Melgar, P. Silva, A. Simonet, G. Antoniu, A. Costan, V. Masson, M. Parashar, I. Rodero and A. Termier. 'A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning'. In: In Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, United States, 7th Feb. 2020, pp. 403–411. DOI: 10.1609/aaai.v34i01.5376. URL: https://hal.archives-ouvertes.fr/hal-02373429.

[6] M. Malms, M. Ostasz, M. Gilliot, P. Bernier-Bruna, L. Cargemel, E. Suarez, H. Cornelius, M. Duranton, B. Koren, P. Rosse-Laurent, M. S. Pérez-Hernández, M. Marazakis, G. Lonsdale, P. Carpenter, G. Antoniu, S. Narasimhamurthy, A. Brinkman, D. Pleiter, A. Tate, J. Krueger, H.-C. Hoppe, E. Laure and A. Wierse. *ETP4HPC's Strategic Research Agenda for High-Performance Computing in Europe 4*. ETP4HPC White Papers. 3rd Mar. 2020. DOI: 10.5281/zenodo.4605343. URL: https://hal.inria.fr/hal-03354396.

[7] O.-C. Marcu, A. Costan, G. Antoniu, M. S. Pérez-Hernández, B. Nicolae, R. Tudoran and S. Bortoli. 'KerA: Scalable Data Ingestion for Stream Processing'. In: ICDCS 2018 - 38th IEEE International Conference on Distributed Computing Systems. Vienna, Austria: IEEE, 2nd July 2018, pp. 1480–1485. DOI: 10.1109/ICDCS.2018.00152. URL: https://hal.inria.fr/hal-01773799.

[8] O.-C. Marcu, A. Costan, G. Antoniu and M. S. Pérez-Hernández. 'Spark versus Flink: Understanding Performance in Big Data Analytics Frameworks'. In: Cluster 2016 - The IEEE 2016 International Conference on Cluster Computing. Taipei, Taiwan, 12th Sept. 2016. DOI: 10.1109/cluster.2016.22. URL: https://hal.inria.fr/hal-01347638.

[9] P. Matri, Y. Alforov, A. Brandon, M. Pérez, A. Costan, G. Antoniu, M. Kuhn, P. Carns and T. Ludwig. 'Mission Possible: Unify HPC and Big Data Stacks Towards Application-Defined Blobs at the Storage Layer'. In: *Future Generation Computer Systems* 109 (Aug. 2020), pp. 668–677. DOI: 10.1016/j.future.2018.07.035. URL: https://hal.archives-ouvertes.fr/hal-01892682.

[10] P. Matri, A. Costan, G. Antoniu, J. Montes and M. S. Pérez. 'Tyr: Blob Storage Meets Built-In Transactions'. In: IEEE ACM SC16 - The International Conference for High Performance Computing, Networking, Storage and Analysis 2016. Salt Lake City, United States, 13th Nov. 2016. DOI: 10.1109/sc.2016.48. URL: https://hal.inria.fr/hal-01347652.

[11] D. Rosendo, P. Silva, M. Simonin, A. Costan and G. Antoniu. 'E2Clab: Exploring the Computing Continuum through Repeatable, Replicable and Reproducible Edge-to-Cloud Experiments'. In: Cluster 2020 - IEEE International Conference on Cluster Computing. Kobe, Japan, 14th Sept. 2020, pp. 1–11. DOI: 10.1109/CLUSTER49012.2020.00028. URL: https://hal.archives-ouvertes.fr/hal-02916032.

[12] Y. Taleb, R. Stutsman, G. Antoniu and T. Cortes. 'Tailwind: Fast and Atomic RDMA-based Replication'. In: ATC '18 - USENIX Annual Technical Conference. Boston, United States, 13th July 2018, pp. 850–863. URL: https://hal.inria.fr/hal-01676502.

11.2 Publications of the year

International peer-reviewed conferences

[13] D. Balouek-Thomert, P. Silva, K. Fauvel, A. Costan, G. Antoniu and M. Parashar. 'MDSC: Modelling Distributed Stream Processing across the Edge-to-Cloud Continuum'. In: DML-ICC 2021 workshop (held in conjunction with UCC 2021). Leicester, United Kingdom, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03510012.

[14] C. Haine, U.-U. Haus, M. Martinasso, D. Pleiter, F. Tessier, D. Sarmany, S. Smart, T. Quintino and A. Tate. 'A Middleware Supporting Data Movement in Complex and Software-Defined Storage and Memory Architectures'. In: ISC 2021 SuperCompCloud - 4th Workshop on Interoperability of Supercomputing and Cloud Technologies. Frankfurt, Germany, 2nd July 2021, pp. 1–12. URL: https://hal.inria.fr/hal-03313021.

[15] O.-C. Marcu, A. Costan, B. Nicolae and G. Antoniu. 'Virtual Log-Structured Storage for High-Performance Streaming'. In: Cluster 2021 - IEEE International Conference on Cluster Computing. Portland / Virtual, United States, 7th Sept. 2021, pp. 1–11. URL: https://hal.inria.fr/hal-03300796.

[16] D. Rosendo, A. Costan, G. Antoniu, M. Simonin, J.-C. Lombardo, A. Joly and P. Valduriez. 'Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum'. In: Cluster 2021 - IEEE International Conference on Cluster Computing. Portland, OR, United States, 7th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03310540.

National peer-reviewed Conferences

[17] T. Bouvier, A. Costan and G. Antoniu. 'Deploying Heterogeneity-aware Deep Learning Workloads on the Computing Continuum'. In: BDA 2021 - 37e Conférence sur la Gestion de Données - Principes, Technologies et Applications. Proceedings of the BDA 2021 conference. Paris, France, 28th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03338520.

[18] D. Rosendo, A. Costan, G. Antoniu and P. Valduriez. 'Enabling Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum'. In: BDA 2021 - 37e Conférence sur la Gestion de Données - Principes, Technologies et Applications. Proceedings of the BDA 2021 conference. Paris, France, 28th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03332524.

Other scientific publications

[19] G. Antoniu, P. Valduriez, H.-C. Hoppe and J. Krüger. *Towards Integrated Hardware/Software Ecosystems for the Edge-Cloud-HPC Continuum*. 2021. DOI: 10.5281/zenodo.5534464. URL: https://hal.archives-ouvertes.fr/hal-03358930.

[20] T. Bouvier, A. Costan and G. Antoniu. 'Heterogeneity-aware Deep Learning Workload Deployments on the Computing Continuum'. In: IPDPS 2021 - 35th IEEE International Parallel and Distributed Processing Symposium. Virtual / Portland, United States, 17th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03270129.

[21] J. C. Bowden, F. Tessier, C. Deltel, S. Bnà and G. Antoniu. *In-situ visualization using Damaris: the Code Saturne use case*. 17th Sept. 2021. URL: https://hal.inria.fr/hal-03354035.

[22] D. Rosendo, A. Costan, G. Antoniu and P. Valduriez. 'E2Clab: Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum'. In: IPDPS 2021 - 35th IEEE International Parallel and Distributed Processing Symposium. Virtual, France, 17th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03269852.

11.3 Cited publications

[23] N. Cheriere and M. Dorier. *Design and Evaluation of Topology-aware Scatter and AllGather Algorithms for Dragonfly Networks*. Salt Lake City, United States, 13th Nov. 2016. URL: https://hal.inria.fr/hal-01400271.

[24] N. Cheriere, M. Dorier and G. Antoniu. 'Pufferbench: Evaluating and Optimizing Malleability of Distributed Storage'. In: PDSW-DISCS 2018: 3rd Joint International workshop on Parallel Data Storage and Data Intensive Scalable computing Systems. Dallas, United States, 12th Nov. 2018, pp. 1–10. DOI: 10.1109/PDSW-DISCS.2018.00006. URL: https://hal.archives-ouvertes.fr/hal-01892713.

[25] M. Dorier, O. Yildiz, S. Ibrahim, A.-C. Orgerie and G. Antoniu. 'On the energy footprint of I/O management in Exascale HPC systems'. In: *Future Generation Computer Systems* 62 (21st Mar. 2016), pp. 17–28. DOI: 10.1016/j.future.2016.03.002. URL: https://hal.inria.fr/hal-01330735.

[26] P. Matri, Y. Alforov, A. Brandon, M. Kuhn, P. Carns and T. Ludwig. 'Could Blobs Fuel Storage-Based Convergence Between HPC and Big Data?' In: CLUSTER 2017 - IEEE International Conference on Cluster Computing. Honolulu, United States, 2017, pp. 81–86. DOI: 10.1109/CLUSTER.2017.63. URL: https://hal.inria.fr/hal-01617655.

[27] P. Silva, A. Costan and G. Antoniu. 'Towards a Methodology for Benchmarking Edge Processing Frameworks'. In: IPDPSW 2019 - IEEE International Parallel and Distributed Processing Symposium Workshops. Rio de Janeiro, Brazil: IEEE, 20th May 2019, pp. 904–907. DOI: 10.1109/IPDPSW.2019.00149. URL: https://hal.inria.fr/hal-02310154.

[28] O. Yildiz, A. C. Zhou and S. Ibrahim. 'Eley: On the Effectiveness of Burst Buffers for Big Data Processing in HPC systems'. In: Cluster'17-2017 IEEE International Conference on Cluster Computing. Honolulu, United States, 5th Sept. 2017. DOI: 10.1109/CLUSTER.2017.73. URL: https://hal.inria.fr/hal-01570737.

[29] C. Haine, U.-U. Haus, F. Tessier, D. Sármány, S. Smart, T. Quintino and A. Tate. 'A Middleware API for Data Movement in Scientific Workflows'. In: International Supercomputing Conference (ISC 2021). Paper submitted. Under review. Frankfurt, Germany, 23rd June 2021.

[30] C. Adjih, E. Baccelli, E. Fleury, G. Harter, N. Mitton, T. Noel, R. Pissard-Gibollet, F. Saint-Marcel, G. Schreiner, J. Vandaele and T. Watteyne. 'FIT IoT-LAB: A Large Scale Open Experimental IoT Testbed'. In: *IEEE World Forum on Internet of Things (IEEE WF-IoT)*. Milan, Italy, Dec. 2015. URL: https://hal.inria.fr/hal-01213938.

[31] R. Bolze, F. Cappello, E. Caron, M. Dayde, F. Desprez, E. Jeannot, Y. Jégou, S. Lanteri, J. Leduc, N. Melab, G. Mornet, R. Namyst, P. Primet, B. Quétier, O. Richard, E.-G. Talbi and I. Touche. 'Grid'5000: A Large Scale And Highly Reconfigurable Experimental Grid Testbed'. In: *International Journal of High Performance Computing Applications* 20.4 (2006), pp. 481–494. DOI: 10.1177/1094342006070078. URL: https://hal.inria.fr/hal-00684943.

[32] A. Joly, P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, A. Affouard, J. Carré, J.-F. o. Molino, N. Boujemaa and D. Barthélémy. 'A look inside the Pl@ntNet experience'. In: *Multimedia Systems* 22.6 (2016), pp. 751–766. DOI: 10.1007/s00530-015-0462-9. URL: https://hal.inria.fr/hal-01182775.

[33] *Chameleon Cloud*. 2021. URL: https://www.chameleoncloud.org/.

[34] *Cybeletech - Digital technologies for the plant world*. 2021. URL: https://www.cybeletech.com/en/home/.

[35] *ECMWF - European Centre for Medium-Range Weather Forecasts*. 2021. URL: https://www.ecmwf.int/.

[36] *European Exascale Software Initiative*. 2013. URL: http://www.eesi-project.eu/.

[37] *The European Technology Platform for High-Performance Computing*. 2012. URL: http://www.etp4hpc.eu/.

[38] *International Exascale Software Program*. 2011. URL: http://www.exascale.org/iesp/.

[39] *Inria's strategic plan "Towards Inria 2020"*. 2016. URL: https://www.inria.fr/fr/recherche-innovation-numerique.