IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

École normale supérieure de Rennes

Activity Report 2019

Project-Team KERDATA

Scalable Storage for Clouds and Beyond

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Distributed and High Performance Computing

Table of contents

Project-Team KERDATA

Creation of the Team: 2009 July 01, updated into Project-Team: 2012 July 01

Keywords:

Computer Science and Digital Science:

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
A1.6. - Green Computing
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. - Algorithms
A7.1.1. - Distributed algorithms
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

1. Team, Visitors, External Collaborators

Research Scientist

Gabriel Antoniu [Team leader, Inria, Senior Researcher, HDR]

Faculty Members

Luc Bougé [École normale supérieure de Rennes, Professor, HDR]
Alexandru Costan [INSA Rennes, Associate Professor, HDR]

Post-Doctoral Fellow

Pedro de Souza Bento Da Silva [INSA Rennes]

PhD Students

Nathanaël Cheriere [École normale supérieure de Rennes, until Nov 2019]
Paul Le Noac'h [INSA Rennes, until Feb 2019]
Daniel Rosendo [Inria, from Oct 2019]

Technical staff

Ovidiu-Cristian Marcu [Inria, Engineer, until Oct 2019]

Interns and Apprentices

Tom Bordin [École Normale Supérieure de Rennes, from May 2019 until Jul 2019]

Juliette Fournis d Albiat [Inria, from Jun 2019 until Aug 2019]

Administrative Assistant

Gaëlle Tworkowski [Inria]

Visiting Scientists

José Aguilar Canepa [Instituto Politécnico Nacional, from Sep 2019 until Nov 2019]
Mario Rivero Angeles [Instituto Politécnico Nacional, from Jun 2019 until Jul 2019]
Edgar Romo Montiel [Instituto Politécnico Nacional, from Sep 2019 until Nov 2019]

2. Overall Objectives

2.1. Context: the need for scalable data management

We are witnessing a rapidly increasing number of application areas generating and processing very large volumes of data on a regular basis. Such applications are called *data-intensive*. Governmental and commercial statistics, climate modeling, cosmology, genetics, bio-informatics, high-energy physics are just a few examples in the scientific area. In addition, rapidly growing amounts of data from social networks and commercial applications are now routinely processed.

In all these examples, the overall application performance is highly dependent on the properties of the underlying data management service. It becomes crucial to store and manipulate massive data efficiently. However, these data are typically *shared* at a large scale and *concurrently accessed* at a high degree. With the emergence of recent infrastructures such as cloud computing platforms and post-Petascale high-performance computing (HPC) systems, achieving highly scalable data management under such conditions has become a major challenge.

2.1.1. Our objective

The KerData project-team is namely focusing on designing innovative architectures and systems for *scalable data storage and processing*. We target two types of infrastructures: *clouds* and *post-Petascale high-performance supercomputers*, according to the current needs and requirements of data-intensive applications.

We are especially concerned by the applications of major international and industrial players in cloud computing and extreme-scale high-performance computing (HPC), which shape the long-term agenda of the cloud computing [26], [23] and Exascale HPC [25] research communities. The Big Data area, emphasized the challenges related to Volume, Velocity and Variety. This is yet another element of context that further highlights the primary importance of designing data management systems that are efficient at a very large scale.

2.1.1.1. 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, resulted from these requirements. Our choice is perfectly in line with Inria's strategic plan [30], which acknowledges as critical the challenges of *storing, exchanging, organizing, utilizing, handling and analyzing* the huge volumes of data generated by an increasing number of sources. This topic is also stated as a scientific priority of Inria's research center of Rennes [29]: *Storage and utilization of distributed big data*.

2.1.1.2. Challenges and goals related to cloud data storage and processing

In the area of cloud data processing, a significant milestone is the emergence of the Map-Reduce [35] parallel programming paradigm. It is currently used on most cloud platforms, following the trend set up by Amazon [22]. At the core of Map-Reduce frameworks lies the storage system, a key component which must meet a series of specific requirements that are not fully met yet by existing solutions: the ability to provide efficient *fine-grain access* to the files, while sustaining a *high throughput* in spite of *heavy access concurrency*; the need to provide a high resilience to *failures*; the need to take *energy-efficiency* issues into account.

More recently, it becomes clear that data-intensive processing needs to go beyond the frontiers of single datacenters. In this perspective, extra challenges arise, related to the efficiency of metadata management. This efficiency has a major impact on the access to very large sets of small objects by Big Data processing workflows running on large-scale infrastructures.

2.1.1.3. Challenges and goals related to data-intensive HPC applications

Key research fields such as climate modeling, solid Earth sciences or 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 [28], EESI [24], ETP4HPC [25]. 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 challenge we have been addressing is to find new ways to store, visualize and analyze massive outputs of data during and after the simulations. Our main initial goal was to do it without impacting the overall performance, avoiding the *jitter* generated by I/O interference as much as possible. Recently, we started to focus specifically on *in situ processing* approaches and we explored approaches to *model and predict I/O phase occurrences* and to *reduce intra-application and cross-application I/O interference*.

2.1.2. Our approach

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 two main execution infrastructures: cloud platforms and post-Petascale HPC supercomputers.

2.1.2.1. Platforms and Methodology

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 a 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. In the post-Petascale HPC area, we are running our experiments on systems including some top-ranked supercomputers, such as Titan, Jaguar, Kraken or Blue Waters. This provides us with excellent opportunities to validate our results on advanced realistic platforms.

2.1.2.2. Collaboration strategy

Our collaboration portfolio includes international teams that are active in the areas of data management for clouds and HPC systems, both in Academia and Industry.

Our academic collaborating partners include Argonne National Lab, University of Illinois at Urbana-Champaign, Universidad Politécnica de Madrid, Barcelona Supercomputing Center, University Politehnica of Bucharest. In industry, we are currently collaborating with Huawei and Total.

Moreover, the consortiums of our collaborative projects include application partners in the area of climate simulations (e.g., the Department of Earth and Atmospheric Sciences of the University of Michigan, within our collaboration inside JLESC [31]). This is an additional asset, which enables us to take into account application requirements in the early design phase of our solutions, and to validate those solutions with real applications... and real users!

3. Research Program

3.1. Research axis 1: Convergence of HPC and Big Data

The tools and cultures of High Performance Computing and Big Data Analytics have evolved in divergent ways. This is to the detriment of both. However, big computations still generate and are needed to analyze Big Data. As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two ecosystems is crucial to the future.

Our objective is premised on the idea that we must explore the ways in which the major challenges associated with Big Data analytics intersect with, impact, and potentially change the directions now in progress for achieving Exascale computing.

In particular, a key milestone will be to achieve convergence through common abstractions and techniques for data storage and processing in support of complex workflows combining simulations and analytics. Such application workflows will need such a convergence to run on hybrid infrastructures combining HPC systems and clouds (potentially in extension to edge devices, in a complete digital continuum).

Collaboration. *This axis is addressed in close collaboration with María Pérez (UPM), Rob Ross (ANL), Toni Cortes (BSC), Several groups at Argonne National Laboratory and NCSA (Franck Cappello, Rob Ross, Bill Kramer, Tom Peterka).*

Relevant groups with similar interests are the following ones.

- *The group of Jack Dongarra, Innovative Computing Laboratory at University of Tennessee, who is leading international efforts for the convergence of Exascale Computing and Big Data.*
- *The group of Satoshi Matsuoka, RIKEN, working on system software for clouds and HPC.*
- *The group of Ian Foster, Argonne National Laboratory, working on on-demand data analytics and storage for extreme-scale simulations and experiments.*

3.1.1. High-performance storage for concurrent Big Data applications

Storage is a plausible pathway to convergence. In this context, we plan to focus on the needs of concurrent Big Data applications that require high-performance storage, as well as transaction support. Although blobs (binary large objects) are an increasingly popular storage model for such applications, state-of-the-art blob storage systems offer no transaction semantics. This demands users to coordinate data access carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior.

There is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications. In this context, one idea on which we plan to focus our efforts is exploring how blob storage systems could provide built-in, multiblob transactions, while retaining sequential consistency and high throughput under heavy access concurrency.

The early principles of this research direction have already raised interest from our partners at ANL (Rob Ross) and UPM (María Pérez) for potential collaborations. In this direction, the acceptance of our paper on the Týr transactional blob storage system as a Best Student Paper Award Finalist at the SC16 conference [10] is a very encouraging step.

3.1.2. Towards unified data processing techniques for Extreme Computing and Big Data applications

In the high-performance computing area (HPC), the need to get fast and relevant insights from massive amounts of data generated by extreme-scale computations led to the emergence of *in situ processing*. It allows data to be visualized and processed in real-time on the supercomputer generating them, in an interactive way, as they are produced, as opposed to the traditional approach consisting of transferring data off-site after the end of the computation, for offline analysis. As such processing runs on the same resources executing the simulation, if it consumes too many resources, there is a risk to "disturb" the simulation.

Consequently, an alternative approach was proposed (*in transit processing*), as a means to reduce this impact: data are are transferred to some temporary processing resources (with high memory and processing capacities). After this real-time processing, they are moved to persistent storage.

In the Big Data area, the search for real-time, fast analysis was materialized through a different approach: stream-based processing. Such an approach is based on a different abstraction for data, that are seen as a dynamic flow of items to be processed. Stream-based processing and in situ/in transit processing have been developed separately and implemented in different tools in the BDA and HPC areas respectively.

A major challenge from the perspective of the HPC-BDA convergence is their joint use in a unified data processing architecture. This is one of the future research challenges that I plan to address in the near future, by combining ongoing approaches currently active in my team: Damaris and KerA. We started preliminary work within the *"Frameworks"* work package of the HPC-Big Data IPL. Further exploring this convergence is a core direction of our current efforts to build collaborative European projects.

3.2. Research axis 2: Cloud and Edge processing

The recent evolutions in the area of Big Data processing have pointed out some limitations of the initial Map-Reduce model. It is well suited for batch data processing, but less suited for real-time processing of dynamic data streams. New types of data-intensive applications emerge, e.g., for enterprises who need to perform analysis on their stream data in ways that can give fast results (i.e., in real time) at scale (e.g., click-stream analysis and network-monitoring log analysis). Similarly, scientists require fast and accurate data processing techniques in order to analyze their experimental data correctly at scale (e.g., collectively analysis of large data sets distributed in multiple geographically distributed locations).

Our plan is to revisit current data storage and processing techniques to cope with the volatile requirements of data-intensive applications on large-scale dynamic clouds in a cost-efficient way, with a particular focus on streaming. More recently, the strong emergence of edge/fog-based infrastructures leads to to additional challenges for new scenarios involving hybrid cloud/fog/edge systems.

Collaboration. *This axis is addressed in close collaboration with María Pérez (UPM), Kate Keahey. (ANL)*

Relevant groups with similar interests include the following ones.

- *The group of Geoffrey Fox, Indiana University, working on data analytics, cloud data processing, stream processing.*
- *The group at RISE Lab, UC Berkeley, working on real-time stream-based processing and analytics.*
- *The group of Ewa Deelman, USC Information Sciences Institute, working on resource management for workflows in clouds.*

3.2.1. Stream-oriented, Big Data processing on clouds

The state-of-the-art Hadoop Map-Reduce framework cannot deal with stream data applications, as it requires the data to be initially stored in a distributed file system in order to process them. To better cope with the above-mentioned requirements, several systems have been introduced for stream data processing such as Flink [27], Spark [32], Storm [33], and Google MillWheel [34]. These systems keep computation in memory to decrease latency, and preserve scalability by using data-partitioning or dividing the streams into a set of deterministic batch computations.

However, they are designed to work in dedicated environments and they do not consider the performance variability (i.e., network, I/O, etc.) caused by resource contention in the cloud. This variability may in turn cause high and unpredictable latency when output streams are transmitted to further analysis. Moreover, they overlook the dynamic nature of data streams and the volatility in their computation requirements. Finally, they still address failures in a best-effort manner.

Our objective is to investigate new approaches for reliable, stream Big Data processing on clouds.

3.2.2. Efficient Edge, Cloud and hybrid Edge/Cloud data processing

Today, we are approaching an important technological milestone: applications are generating huge amounts of data and are demanding low-latency responses to their requests. Mobile computing and Internet of Things (IoT) applications are good illustrations of such scenarios. Using only Cloud computing for such scenarios is challenging. Firstly, Cloud resources are most of the time accessed through Internet, hence, data are sent across high-latency wide area networks, which may degrade the performance of applications. Secondly, it may be impossible to send data to the Cloud due to data regulations, national security laws or simply because an Internet connection is not available. Finally, data transmission costs (e.g., Cloud provider fees, carrier costs) could make a business solution impractical.

Edge computing is a new paradigm which aims to address some of these issues. The key idea is to leverage computing and storage resources at the "edge" of the network, i.e., on processing units located close to the data sources. This allows applications to outsource task execution from the main (Cloud) processing data centers to the edge. The development of Edge computing was accelerated by the recent emergence of stream processing, a new model for handling continuous flows of data in real-time, as opposed to batch processing, which typically processes bounded datasets offline.

However, Edge computing is not a silver bullet. Besides being a new concept not fully established in the community, issues like node volatility, limited processing power, high latency between nodes, fault tolerance and data degradation may impact applications depending on the characteristics of the infrastructure.

Some relevant research questions are: How much can one improve (or degrade) the performance of an application by performing data processing closer to the data sources rather than performing it in the cloud? How to progress towards a seamless scheduling and execution of a data analytics workflow and break the limitation the current dual approaches used in preliminary efforts in this area, that rely on manual and empirical deployment of the corresponding dataflow operator graphs, using separate analytics engines for centralized clouds and for edge systems respectively?

Our objective is to try to answer precisely such questions. We are interested in understanding the conditions that enable the usage of Edge or Cloud computing to reduce the time to results and the associated costs. While some state-of-the-art approaches advocate either "100% Cloud" or "100% Edge" solutions, the relative efficiency of a method over the other may vary. Intuitively, it depends on many parameters, including network technology, hardware characteristics, volume of data or computing power, processing framework configuration and application requirements, to cite a few. We plan to study their impact on the overall application performance.

3.3. Research axis 3: Supporting AI across the digital continuum

Integrating and processing high-frequency data streams from multiple sensors scattered over a large territory in a timely manner requires high-performance computing techniques and equipments. For instance, a machine learning earthquake detection solution has to be designed jointly with experts in distributed computing and cyber-infrastructure to enable real-time alerts. Because of the large number of sensors and their high sampling rate, a traditional centralized approach which transfers all data to a single point may be impractical. Our goal is to investigate innovative solutions for the design of efficient data processing infrastructures for a distributed machine learning-based approach.

In particular, building on our previous results in the area of efficient stream processing systems, we aim to explore approaches for unified data storage, processing and machine-learning based analytics across the whole digital continuum (i.e., for highly distributed applications deployed on hybrid edge/cloud/HPC infrastructures). Our ZettaFlow project is targeting a startup creation precisely this area.

Collaboration. *This recently started axis is worked out in close collaboration with the group of Manish Parashar, Rutgers University, and with the LACODAM team at Inria, focused on large-scale collaborative data mining.*

4. Application Domains

4.1. 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.

- Climate and meteorology
- Earth science
- Energy and sustainable development
- Smart cities
- Data science

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Contributions to the ETP4HPC agenda

The KerData team contributed to the new ETP4HPC Strategic Agenda (to appear). It will serve as a reference for the future EU funding strategy for HPC. Gabriel Antoniu served as a co-leader of the Programming Environment working group. He also served as a co-leader of 2 transversal (*"cross-working group"*) research clusters: *"HPC and the Digital Continuum"* and *"Data Everywhere"*. Alexandru Costan served as a member of these groups.

5.1.2. Paper co-authored with the LACODAM team published in a major AI conference

In 2019, Pedro Silva initiated a multi-disciplinary collaboration with the LACODAM Inria team and the team of Manish Parashar at Rutgers University. It addresses Machine Learning in the context of Edge stream processing. The target application is early earthquake detection from motion sensors distributed on the ground.

This collaboration resulted in a co-authored paper titled *Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning* [21]. It will be presented at the 34th AAAI Conference on Artificial Intelligence (AAAI-20), a top conference for Machine Learning (CORE Rank: A*). It is the first paper published by the team in a major AI venue.

5.1.3. Awards

Pierre Matri, earned a PhD in May 2018 co-advised by Maria Pérez (Universidad Politécnica de Madrid, UPM), Alexandru Costan (INSA Rennes) and Gabriel Antoniu (Inria). This PhD was defended at UPM and it received the Outstanding PhD Award (*Premio Extraordinario*) of UPM.

6. New Software and Platforms

6.1. 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: - In situ data analysis by some 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 visualizating them either by VisIt or ParaView

- Participants: Gabriel Antoniu, Lokman Rahmani, Luc Bougé, Matthieu Dorier, Orçun Yildiz and Hadi Salimi
- Partner: ENS Rennes
- Contact: Matthieu Dorier
- URL: https://project.inria.fr/damaris/

6.2. OverFlow

FUNCTIONAL DESCRIPTION: OverFlow is a uniform data management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. The software is environment-aware, as it monitors and models the global cloud infrastructure, offering high and predictable data handling performance for transfer cost and time, within and across sites. OverFlow proposes a set of pluggable services, grouped in a data-scientist cloud kit. They provide the applications with the possibility to monitor the underlying infrastructure, to exploit smart data compression, deduplication and geo-replication, to evaluate data management costs, to set a tradeoff between money and time, and optimize the transfer strategy accordingly.

Currently, OverFlow is used for data transfers by the Microsoft Research ATLE Munich team as well as for synthetic benchmarks at the Politehnica University of Bucharest.

- Participants: Alexandru Costan, Gabriel Antoniu and Radu Marius Tudoran
- Contact: Alexandru Costan

6.3. Pufferbench

KEYWORDS: Distributed Storage Systems - Elasticity - Benchmarking

SCIENTIFIC DESCRIPTION: Pufferbench is 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. Besides, it can serve to quickly prototype and evaluate mechanisms for malleability in existing distributed storage systems.

FUNCTIONAL DESCRIPTION: Pufferbench is a benchmark to designed to evaluate whether to use malleable distributed storage systems on a given platform. - It measures the duration of commission and decommission operations. - Its modularity allows to quickly change and adapt each component to the needs of the user. - It can serve as a baseline when implementing commission and decommission mechanisms in a distributed storage system.

RELEASE FUNCTIONAL DESCRIPTION: This is the first release of Pufferbench.

It includes default components for each of the customisable components: - storage: in memory, on drive with file system cache, and on drive without file system cache - network: MPI network - IODispatcher: basic, and with acknowledgements - DataTransferScheduler: basic - DataDistributionGenerator: uniform, and random - MetadataGenerator: Files of same size The diversity of available components enables Pufferbench to fit to multiple use cases.

- Participants: Nathanaël Cheriere, Matthieu Dorier and Gabriel Antoniu
- Partner: ENS Rennes
- Contact: Nathanaël Cheriere
- Publication: hal-01886351
- URL: https://gitlab.inria.fr/Puffertools/Pufferbench/wikis/home

6.4. Tyr

KEYWORDS: Cloud storage - Distributed Storage Systems - Big data

FUNCTIONAL DESCRIPTION: Tyr is the first blob storage system to provide built-in, multiblob transactions, while retaining sequential consistency and high throughput under heavy access concurrency. Tyr offers fine-grained random write access to data and in-place atomic operations.

- Partner: Universidad Politécnica de Madrid
- Contact: Gabriel Antoniu

6.5. Planner

KEYWORDS: Edge elements - Cloud computing - Scheduling

FUNCTIONAL DESCRIPTION: Planner is a middleware for uniform and transparent stream processing across Edge and Cloud. Planner automatically selects which parts of the executiongraph will be executed at the Edge in order to minimize thenetwork cost.

- Partner: ENS Cachan
- Contact: Gabriel Antoniu
- URL: https://team.inria.fr/kerdata/

6.6. KerA

KerAnalytics

KEYWORD: Distributed Storage Systems

FUNCTIONAL DESCRIPTION: A unified architecture for stream ingestion and storage which can lead to the optimization of the processing of Big Data applications. This approach minimizes data movement within the analytics architecture, finally leading to better utilized resources.

- Contact: Gabriel Antoniu

6.7. TailWind

KEYWORDS: Fault-tolerance - Data management. - Distributed Data Management

FUNCTIONAL DESCRIPTION: Replication is essential for fault-tolerance. However, in in-memory systems, it is a source of high overhead. Remote direct memory access (RDMA) is attractive to create redundant copies of data, since it is low-latency and has no CPU overhead at the target. However, existing approaches still result in redundant data copying and active receivers. To ensure atomic data transfers, receivers check and apply only fully received messages. Tailwind is 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. Tailwind substantially improves replication throughput and response latency compared with conventional RPC-based replication. In symmetric systems where servers both serve requests and act as replicas, Tailwind also improves normal-case throughput by freeing server CPU resources for request processing. We implemented and evaluated Tailwind on RAMCloud, a low-latency in-memory storage system. Experiments show Tailwind improves RAMCloud's normal-case request processing throughput by 1.7×. It also cuts down writes median and 99^{th} percentile latencies by 2x and 3x respectively.

- Contact: Gabriel Antoniu

7. New Results

7.1. Convergence HPC and Big Data

7.1.1. Convergence at the data-processing level

Participants: Gabriel Antoniu, Alexandru Costan, Daniel Rosendo.

Traditional data-driven analytics relies on Big Data processing techniques, consisting of batch processing and real-time (stream) processing, potentially combined in a so-called *Lambda architecture*. This architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data.

On the other side, simulation-driven analytics is based on computational (usually physics-based) simulations of complex phenomena, which often leverage HPC infrastructures. The need to get fast and relevant insights from massive amounts of data generated by extreme-scale simulations led to the emergence of in situ and in transit processing approaches: they allow data to be visualized and processed interactively in real-time as data are produced, while the simulation is running.

To support hybrid analytics and continuous model improvement, we propose to combine the above data processing techniques in what we will call the *Sigma architecture*, a HPC-inspired extension of the Lambda architecture for Big Data processing [17]. Its instantiation in specific application settings depends of course of the specific application requirements and of the constraints that may be induced by the underlying infrastructure. Its main conceptual strength consists in the ability to leverage in a unified, consistent framework, data processing techniques that became reference in HPC in the Big Data communities respectively, without however being combined so far for joint usage in converged environments.

The given framework will integrate previously-validated approaches developed in our team, such as Damaris, a middleware system for efficient I/O management and large-scale in situ data processing, and KerA, a unified system for data flow ingestion and storage. The overall objective is to enable the usage of a large spectrum of Big Data analytics and Intelligence techniques at extreme scales in the Cloud and Edge, to support continuous intelligence (from streaming and historical data) and precise insights/predictions in real-time and fast decision making.

7.1.2. Pufferscale: Elastic storage to support dynamic hybrid workflows systems

Participants: Nathanaël Cheriere, Gabriel Antoniu.

User-space HPC data services are emerging as an appealing alternative to traditional parallel file systems, because of their ability to be tailored to application needs while eliminating unnecessary overheads incurred by POSIX compliance. Such services may need to be rescaled up and down to adapt to changing workloads, in order to optimize resource usage. This can be useful, for instance, to better support complex workflows that mix on-demand simulations and data analytics.

We formalized the operation of rescaling a distributed storage system as a multi objective optimization problem considering three criteria: load balance, data balance, and duration of the rescaling operation. We proposed a heuristic for rapidly finding a good approximate solution, while allowing users to weight the criteria as needed. The heuristic is evaluated with Pufferscale, a new, generic rescaling manager for microservice-based distributed storage systems [18].

To validate our approach in a real-world ecosystem, we showcase the use of Pufferscale as a means to enable storage malleability in the HEPnOS storage system for high energy physics applications.

7.2. Cloud and Edge processing

7.2.1. Benchmarking Edge processing frameworks

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

With the spectacular growth of the Internet of Things, edge processing emerged as a relevant means to offload data processing and analytics from centralized Clouds to the devices that serve as data sources (often provided with some processing capabilities). While a large 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 [20].

7.2.2. Analytical models for performance evaluation of stream processing

Participants: José Aguilar Canepa, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

One of the challenges of enabling the Edge computing paradigm is to identify the situations and scenarios in which Edge processing is suitable to be applied. To this end, applications can be modeled as a graph consisting of tasks as nodes and data dependencies between them as edges. The problem comes down to deploying the application graph onto the network graph, that is, operators need to be put on machines, and finding the optimal cut in the graph between the Edge and Cloud resources (i.e., nodes in the network graph).

We have designed an algorithm that finds the optimal execution plan, with a rich cost model that lets users to optimize whichever goal they might be interested in, such as monetary costs, energetic consumption or network traffic, to name a few.

In order to validate the cost model and the effectiveness of the algorithm, a series of experiments were designed using two real-life stream processing applications: a closed-circuit television surveillance system, and an earthquake early warning system.

Two network infrastructures were designed to run the applications. The first one is a state-of-art infrastructure where all processing is done on the Cloud to serve as benchmark. The second one is an infrastructure produced by the algorithm. Both scenarios were executed on the Grid'5000. Several experiments are currently underway. The trade-offs of executing Cloud/Edge workloads with this model were published in [19].

7.2.3. Modeling smart cities applications

Participants: Edgar Romo Montiel, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Smart City applications have particular characteristics in terms of data processing and storage, which need to be taken into account by the underlying serving layers. The objective of this new activity is to devise clear models of the data handled by such applications. The data characteristics and the processing requirements does not have to match one-to-one. In some cases, some particular types of data might need one or more types of processing, depending on the use case. For example, small and fast data coming from sensors do not always have to be processed in real-time, but they could also be processed in a batch manner at a later stage.

This activity is the namely the topic of the SmartFastData associated team with the Instituto Politécnico Nacional of Mexico.

In a first phase, we focused on modeling the stream rates of data from sets of sensors in Smart Cities, specifically, from vehicles inside a closed coverage area. Those vehicles are connected in a V2I VANET, and they interact to applications in the Cloud such as traffic reports, navigation apps, multimedia downloading etc. This led to the design of a mathematical model to predict the time that a mobile sensor resides within a geographical designated area.

The proposed model uses Coxian distributions to estimate the time a vehicle requests Cloud services, so that the core challenge is to adjust their parameters. It was achieved by validating the model against real-life data traces from the City of Luxembourg, through extensive experiments on the Grid'5000.

Next, these models were used to estimate the resources needed in the Cloud (or at the Edge) in order to process the whole stream of data. We designed an auto-Scaling module able to adapt the resources with respect to the load. Using the Grid'5000, we evaluated the various possibility to place the prediction module: *(i)* at the Edge, close to data with less accuracy but faster results; or *(ii)* in the Cloud, with higher accuracy due to the global data, but higher latency as well.

7.3. AI across the digital continuum

7.3.1. Machine Learning in the context of Edge stream processing.

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Our research aims to improve the accuracy of Earthquake Early Warning (EEW) systems by means of machine learning. EEW 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 sensitivity to the ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to its 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.

In practice, EEW can be seen as a typical classification problem in the machine learning field: multi-sensor data are given in input, and earthquake severity is the classification result. 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.

DMSEEW is based on a new stacking ensemble method which has been evaluated on a real-world dataset validated with geoscientists. The system builds on a geographically distributed infrastructure (deployable on clouds and edge systems), ensuring 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, a "A*" conference in the area of Artificial Intelligence [21].

7.3.2. ZettaFlow: Unified Fast Data Storage and Analytics Platform for IoT

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

The ZettaFlow platform (system of systems) provides a high-performance multi-model analytics-oriented storage and processing system, while supporting publish-subscribe streams and streaming, key-value and in-memory columnar APIs [16].

The ZettaFlow project is funded by EIT Digital from October 2019 to December 2020. It includes three partners: Inria for the platform development, TU Berlin for edge to cloud IoT optimizations with microservices, and Systematic Paris Region for the go-to-market strategy.

Our goal is to create a startup that will commercialize the ZettaFlow platform: a dynamic, unified and auto-balanced real-time storage and analytics industrial IoT platform. ZettaFlow will provide real-time visibility into machines, assets and factory operations and will automate data driven decisions for high-performance industrial processes.

ZettaFlow will bring a threefold impact to the IoT market.

1. Enable novel real-time edge applications that truly automate manufacturing, transportation and utilities processes.
2. Reduce deployment efforts and time-to-decision of IoT edge-cloud applications by 75% through automation, unified dynamic data management and streaming analytics.
3. Reduce human costs for monitoring and engineering (through edge intelligence) and IoT hardware costs by 50% through unified data collection/storage/analytics.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. OverFlow (2015–2019)

Participants: Alexandru Costan, Pedro de Souza Bento Da Silva, Paul Le Noac'h.

Project Acronym: OverFlow

Project Title: Workflow Data Management as a Service for Multisite Applications

Coordinator: Alexandru Costan

Duration: October 2015–October 2019

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: https://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 2019, we focused on the challenges of stream processing at the Edge. In particular, Edge computing presents a significant opportunity to realize the potential of distributed ML models with regards to low latency, high availability and privacy. It allows for instance inferences on simple image, video or audio classification; as only the final result is transmitted, delays are minimized, while privacy and bandwidth are preserved in IoT applications. Also, neural networks could be partitioned such that some layers are evaluated at the Edge and the rest in the cloud.

In this context we proposed an architecture in which the initial layers can be used for feature-abstraction functions: as data travels through the neural network, they abstract into high-level features, which are more lightweight, helping reduce latency.

8.1.2. Other National Projects

8.1.2.1. HPC-Big Data Inria Project Lab (IPL)

Participants: Gabriel Antoniu, Alexandru Costan, Daniel Rosendo, Pedro de Souza Bento Da Silva.

Project Acronym: HPC-BigData

Project Title: The HPC-BigData Inria Project Lab

Coordinator: Bruno Raffin

Duration: 2018–2022

Web site: https://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 2019, Daniel Rosendo, who was hired in the context of this IPL project, focused on assessing the state of the art in high performance analytics on hybrid HPC/Big Data infrastructure. In particular, a new path for future work was identified: running Machine Learning algorithm at the Edge.

8.1.2.2. ADT Damaris 2

Participants: Ovidiu-Cristian Marcu, Gabriel Antoniu, Luc Bougé.

Project Acronym: ADT Damaris

Project Title: Technology development action for the Damaris environment

Coordinator: Gabriel Antoniu

Duration: 2019–2021

Web site: https://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 2 years of funding support for a senior engineer.

Ovidiu Marcu has been funded through this project to document, test and extend the Damaris software and make it a safely distributable product. In 2019, the main goal was to add Big Data analytics support in Damaris. We have extended Damaris with a streaming interface for writing and analyzing in real-time simulation data through KerA, a distributed streaming storage system.

KerA is further coupled with RAMCloud for in-memory key-value transactions and with Apache Flink for streaming analytics in an architecture that leverages Apache Arrow as in-memory columnar data representation for co-located streaming. This hybrid HPC-Big Data architecture is subject to further exploration within the ZettaFlow.io startup.

8.1.2.3. Grid'5000

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

8.2. European Initiatives

8.2.1. Collaborations in European Programs, Except FP7 & H2020

8.2.1.1. ZettaFlow: Unified Fast Data Storage and Analytics Platform for IoT

Program: EIT Digital Innovation Factory

Project acronym: ZettaFlow

Project title: ZettaFlow: Unified Fast Data Storage and Analytics Platform for IoT

Duration: October 2019–December 2020

Technical Coordinator: Ovidiu Marcu

Other partners: Technische Universität Berlin and System@tic

Web site: https://zettaflow.io/

The objective of this project is to create a startup in order to commercialize the ZettaFlow platform: a dynamic, unified and auto-balanced real-time storage and analytics industrial IoT platform. ZettaFlow is based on KerA, a streaming storage system prototype developed within the KerData team. ZettaFlow will provide real-time visibility into machines, assets and factory operations and will automate data driven decisions for high-performance industrial processes.

8.2.1.2. 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)

Elasticity is one of the key features of cloud computing providing virtual resources as needed according to dynamically changing workloads. This allows to minimize costs and reduce time-to-decision of IoT edge-cloud applications. However, while the underlying resources may easily be scaled many applications and services are not designed to support elastic scalability or require an administrator to manually control elastic scaling.

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.

8.2.2. Collaborations with Major European Organizations

8.2.2.1. BDVA and ETP4HPC

Gabriel Antoniu (as a working group leader) and Alexandru Costan (as a working group member) contributed to the new Strategic Research Agenda (version 4) of European Technology Platform in the area of High-Performance Computing (ETP4HPC).

Gabriel Antoniu and Alexandru Costan are serving as Inria representatives in the working group dedicated to *HPC-Big Data* convergence within the Big Data Value Association (BDVA).

8.2.2.2. International Initiatives

8.2.2.2.1. BDEC: Big Data and Extreme Computing

Since 2015, Gabriel Antoniu has been invited to participate to the yearly workshops of the international Big Data and Extreme-scale Computing (BDEC) working group focused on the convergence of Extreme Computing (the latest incarnation of High-Performance Computing - HPC) and Big Data. BDEC is organized as series of invitation-based international workshops.

In 2019 Gabriel Antoniu was invited again to contribute to the second and third workshops of the BDEC2 series, where he presented two white papers on HPC-Big Data convergence at the level of data processing.

8.3. International Initiatives

8.3.1. Inria International Labs

8.3.1.1. UNIFY: An associated team involved in the JLESC international lab

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

Inria International Lab: JLESC: Joint Laboratory for Extreme Scale Computing

International Partner: Argonne National Laboratory (USA) — Department of Mathematics, Symbolic Computation Group — Tom Peterka

Start year: 2019

See also: https://team.inria.fr/unify

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 insight 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.

8.3.2. Inria Associate Teams Not Involved in an Inria International Labs

8.3.2.1. SmartFastData

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

International Partner: Instituto Politécnico Nacional (Mexico) — Centro de Investigación en Computación — Rolando Menchaca-Mendez

- Start year: 2019
- See also: https://team.inria.fr/smartfastdata/

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.

In 2019, the first objective of the associated team (i.e., exploring analytical models for performance evaluation of stream storage and ingestion systems) was achieved by means of the two internships of José Canepa and Edgar Romo (described in the New Results section) as well as the visit of Mario Rivero as an Invited Professor, who set up the main research agenda for those internships.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Rosa Badia: Barcelona Supercomputing Center, Spain. Dates: 13-14 March 2019

Michael Schottner: University of Dusseldorf, Germany. Dates: 13-15 March 2019

Valentin Cristea: Politehnica University of Bucharest, Romania. Dates: 13-15 March 2019

Toni Cortés: Universitat Politècnica Catalunya, Spain. Dates: 4-5 November 2019

Kate Keahey: Argonne National Lab, USA. Dates: 4-5 November 2019

Matthieu Dorier: Argonne National Lab, USA. Dates: 4-5 November 2019

8.4.1.1. Invited Professors

Mario Rivero (Professor, Instituto Politécnico Nacional, Mexico) was an invited professor in the KerData team from June to July 2019, through the *Scientist Invitation Program* of IRISA and ISTIC. During his stay, he gave several talks at Inria/IRISA and worked on the modeling Smart City applications, laying the path for the work program of the upcoming internships of José Aguilar-Canepa and Edgar Romo.

8.4.1.2. Internships

Jose Aguilar-Canepa (PhD student, Instituto Politécnico Nacional, Mexico) has done a 3-month internship within the team, working with Alexandru Costan and Pedro Silva on hybrid Edge/Cloud stream processing. This work is validated through large scale experiments on Grid'5000 and is subject to a journal paper in submission, currently on the works, to be submitted by January 2020.

Edgar Romo (PhD student, Instituto Politécnico Nacional, Mexico) did a 3-month internship at KerData from September to November 2019. He worked on Objective 2 of the SmartFastData Associate Team, specifically on designing a complex model for predicting the stream arrival rates for vehicular networks in Smart Cities. To validate this proposal, he carried out several experiments on Grid'5000; this work is currently the topic of a workshop paper submission.

8.4.2. Visits to International Teams

Alexandru Costan and Gabriel Antoniu visited the NDS-Lab team at Instituto Politécnico Nacional from October 24 to November 3, 2019, in the context of the SmartFastData associate team. Working closely with Rolando Menchaca, they defined the work program for the upcoming year with respect to the team's objectives. They also presented KerData's vision on future hybrid analytics combining Edge, Cloud and HPC computing.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events: Organization

9.1.1.1. 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, this year.

Gabriel Antoniu: Co-Chair of Conv'2019 - the HPC-AI-Big Data Convergence Days.

9.1.1.2. Member of the Organizing Committees

Gabriel Antoniu: Member of the Organizing Committee of Conv'2019, the HPC-AI-Big Data Convergence Days.

9.1.2. Scientific Events: Selection

9.1.2.1. Chair of Conference Program Committees

Gabriel Antoniu: Track Co-Chair for the Clouds and Distributed Systems Track of the ACM/IEEE SC19 conference, Denver, USA.

Alexandru Costan: Program Co-Chair of the ScienceCloud 2019 international workshop held in conjunction with ACM HPDC 2019, Phoenix, AZ, USA.

9.1.2.2. Member of the Conference Program Committees

Alexandru Costan: IEEE/ACM SC'19 (Technical Program: Cloud Track, Posters and ACM Student Research Competition), ACM/IEEE CCGrid 2019, IEEE Cluster 2019, IEEE/ACM UCC 2019, IEEE Big Data 2019, SCRAMBL 2019, CSCS 2019, CEBDA 2019, CCIW 2019, IEEE CSE 2018, IEEE CloudCom 2019.

Gabriel Antoniu: IEEE Cluster 2019, IEEE IPDPS 2019, STREAM-ML 2019.

9.1.2.3. Reviewer

Alexandru Costan: ACM HPDC 2019, IEEE IPDPS 2019.

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

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

9.1.3.2. 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: SoftwareX, Philosophical Transactions A.

9.1.4. Invited Talks

Gabriel Antoniu

- Fourth BDEC2 Workshop on Big Data and Extreme Computing, San Diego, in October 2019. *Towards a demonstrator of the Sigma Data Processing Architecture for BDEC 2*. URL: https://www.exascale.org/bdec/agenda/poznan.
- Third BDEC2 Workshop on Big Data and Extreme Computing, Poznan, in May 2019. *ZettaFlow: Towards High-Performance ML-based Analytics across the Digital Continuum*. URL: https://www.exascale.org/bdec/agenda/sandiego.
- Ninth Workshop of the Joint Laboratory for Extreme-Scale Computing (JLESC) in April 2019. *Scalable Data Ingestion for Stream Processing*. URL: https://jlesc.github.io/.

Alexandru Costan

- Inria/ IPN Joint Workshop, Mexico City, November 2019. *From Big Data to Fast Data: Efficient Stream Data Managements*.

9.1.5. Leadership within the Scientific Community

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

Gabriel Antoniu

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

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 ADT project (2016–2018), to be continued with the Damaris 2 ADT project (2019–2021).

9.1.6. Scientific Expertise

Luc Bougé

HCERES: Chair of the evaluation committee for the CRIStAL joint laboratory, Lille (report)

HCERES: Member of the evaluation committee for the IRCICA research institute, Lille

9.1.7. Research Administration

Luc Bougé

ANR: Member of the management team for the IA for Humanity governmental program launched in March 2018. Bertrand Braunschweig, Inria, is the scientific director of the program.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

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, France.
- Master: Scalable Distributed Systems, 10 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School, France.
- Master: Infrastructures for Big Data, 10 hours (lectures), M2 level, IBD Module, SIF Master Program, University of Rennes, France.
- Master: Cloud Computing and Big Data, 10 hours (lectures), M2 level, Cloud Module, MIAGE Master Program, University of Rennes, France.

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, France.
- 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.

Luc Bougé

- Bachelor: Introduction to programming concepts, 36 hours (lectures), L3 level, Informatics program, ENS Rennes, France.
- Bachelor: Introduction to scientific research, 24 hours. Research center visits, individual research project supervised by local researchers, student seminars, summer internships, etc.
- Master Program, Rennes: Invited presentation to the M2 students about *Preparing your applications after your PhD* (October 2019); *Informatics as a scientific activity: Towards a responsible research* (November 2019).

Pedro Silva

- Master: Algorithms for Big Data, 4 hours (lectures), M2, INSA Rennes.
- Master: Algorithms for Big Data, 6 hours (lab sessions), M2, INSA Rennes.

9.2.2. Supervision

9.2.2.1. HdR completed this year

Alexandru Costan: *From Big Data to Fast Data: Efficient Stream Data Management*, ENS Rennes, March 2019 [15].

9.2.2.2. PhD completed this year

Nathanaël Cheriere: *Towards Malleable Distributed Storage Systems? From Models to Practice*, ENS Rennes, thesis defended in November 2019, co-advised by Gabriel Antoniu and Matthieu Dorier [14].

9.2.2.3. PhD in progress

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

Paul Le Noac'h: *Workflow Data Management as a Service for Multi-Site Applications*, INSA Rennes, thesis started in November 2016, co-advised by Luc Bougé and Alexandru Costan. Thesis stopped in February 2019.

9.2.3. Juries

Luc Bougé: Member of the jury the *CAPES of mathématiques*, *Informatics track*. This national committee selects more than 1000 mathematics teachers per year for French secondary schools and high-schools.

9.3. Popularization

9.3.1. Internal or external Inria responsibilities

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.

10. Bibliography

Major publications by the team in recent years

[1] N. CHERIERE, M. DORIER. *Design and Evaluation of Topology-aware Scatter and AllGather Algorithms for Dragonfly Networks*, November 2016, Supercomputing 2016, Poster, https://hal.inria.fr/hal-01400271

[2] A. COSTAN, R. TUDORAN, G. ANTONIU, G. BRASCHE. *TomusBlobs: Scalable Data-intensive Processing on Azure Clouds*, in "CCPE - Concurrency and Computation: Practice and Experience", May 2013, https://hal.inria.fr/hal-00767034

[3] B. DA MOTA, R. TUDORAN, A. COSTAN, G. VAROQUAUX, G. BRASCHE, P. J. CONROD, H. LEMAITRE, T. PAUS, M. RIETSCHEL, V. FROUIN, J.-B. POLINE, G. ANTONIU, B. THIRION. *Machine Learning Patterns for Neuroimaging-Genetic Studies in the Cloud*, in "Frontiers in Neuroinformatics", April 2014, vol. 8, https://hal.inria.fr/hal-01057325

[4] M. DORIER, G. ANTONIU, F. CAPPELLO, M. SNIR, L. ORF. *Damaris: How to Efficiently Leverage Multicore Parallelism to Achieve Scalable, Jitter-free I/O*, in "CLUSTER - IEEE International Conference on Cluster Computing", Beijing, China, IEEE, September 2012, https://hal.inria.fr/hal-00715252

[5] M. DORIER, G. ANTONIU, F. CAPPELLO, M. SNIR, R. SISNEROS, O. YILDIZ, S. IBRAHIM, T. PETERKA, L. ORF. *Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations*, in "ACM Transactions on Parallel Computing", 2016, https://hal.inria.fr/hal-01353890

[6] M. DORIER, G. ANTONIU, R. ROSS, D. KIMPE, S. IBRAHIM. *CALCioM: Mitigating I/O Interference in HPC Systems through Cross-Application Coordination*, in "IPDPS - International Parallel and Distributed Processing Symposium", Phoenix, United States, May 2014, https://hal.inria.fr/hal-00916091

[7] M. DORIER, M. DREHER, T. PETERKA, G. ANTONIU, B. RAFFIN, J. M. WOZNIAK. *Lessons Learned from Building In Situ Coupling Frameworks*, in "ISAV 2015 - First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization (held in conjunction with SC15)", Austin, United States, November 2015 [*DOI :* 10.1145/2828612.2828622], https://hal.inria.fr/hal-01224846

[8] M. DORIER, S. IBRAHIM, G. ANTONIU, R. ROSS. *Omnisc'IO: A Grammar-Based Approach to Spatial and Temporal I/O Patterns Prediction*, in "SC14 - International Conference for High Performance Computing, Networking, Storage and Analysis", New Orleans, United States, IEEE, ACM, November 2014, https://hal.inria.fr/hal-01025670

[9] M. DORIER, S. IBRAHIM, G. ANTONIU, R. ROSS. *Using Formal Grammars to Predict I/O Behaviors in HPC: the Omnisc'IO Approach*, in "TPDS - IEEE Transactions on Parallel and Distributed Systems", October 2015 [*DOI :* 10.1109/TPDS.2015.2485980], https://hal.inria.fr/hal-01238103

[10] P. MATRI, A. COSTAN, G. ANTONIU, J. MONTES, M. S. PÉREZ. *Týr: 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, November 2016, https://hal.inria.fr/hal-01347652

[11] B. NICOLAE, G. ANTONIU, L. BOUGÉ, D. MOISE, A. CARPEN-AMARIE. *BlobSeer: Next-Generation Data Management for Large-Scale Infrastructures*, in "JPDC - Journal of Parallel and Distributed Computing", February 2011, vol. 71, n^{o} 2, pp. 169–184, http://hal.inria.fr/inria-00511414/en/

[12] B. NICOLAE, J. BRESNAHAN, K. KEAHEY, G. ANTONIU. *Going Back and Forth: Efficient Multi-Deployment and Multi-Snapshotting on Clouds*, in "HPDC 2011 - The 20th International ACM Symposium on High-Performance Parallel and Distributed Computing", San José, CA, United States, June 2011, http://hal.inria.fr/inria-00570682/en

[13] R. TUDORAN, A. COSTAN, G. ANTONIU. *OverFlow: Multi-Site Aware Big Data Management for Scientific Workflows on Clouds*, in "IEEE Transactions on Cloud Computing", June 2015 [*DOI :* 10.1109/TCC.2015.2440254], https://hal.inria.fr/hal-01239128

Publications of the year

Doctoral Dissertations and Habilitation Theses

[14] N. CHERIERE. *Towards Malleable Distributed Storage Systems: From Models to Practice*, École normale supérieure de Rennes, November 2019, https://tel.archives-ouvertes.fr/tel-02376032

[15] A. COSTAN. *From Big Data to Fast Data: Efficient Stream Data Management*, ENS Rennes, March 2019, Habilitation à diriger des recherches, https://hal.archives-ouvertes.fr/tel-02059437

Invited Conferences

[16] G. ANTONIU, A. COSTAN, O.-C. MARCU. *ZettaFlow: Towards High-Performance ML-based Analytics across the Digital Continuum*, in "BDEC2 2019 - Workshop on Big Data and Extreme-scale Computing", San Diego, United States, San Diego Supercomputing Center, October 2019, 4 p. , https://hal.archives-ouvertes.fr/hal-02428382

[17] G. ANTONIU, A. COSTAN, O.-C. MARCU, M. HERNÁNDEZ-PÉREZ, N. STOJANOVIC. *Towards a demonstrator of the Sigma Data Processing Architecture for BDEC 2*, in "BDEC2 2019 - Workshop on Big Data and Extreme-scale Computing", Poznan, Poland, Poznan Supercomputing and Networking Center, May 2019, 4 p. , https://hal.archives-ouvertes.fr/hal-02428391

International Conferences with Proceedings

[18] N. CHERIERE, M. DORIER, G. ANTONIU. *Is it Worth Relaxing Fault Tolerance to Speed Up Decommission in Distributed Storage Systems?*, in "CCGrid 2019 - IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing", Larnaca, Cyprus, IEEE, May 2019, pp. 1-10 [*DOI :* 10.1109/CCGRID.2019.00024], https://hal.archives-ouvertes.fr/hal-02116727

[19] P. SILVA, A. COSTAN, G. ANTONIU. *Investigating Edge vs. Cloud Computing Trade-offs for Stream Processing*, in "BigData 2019 - IEEE International Conference on Big Data", Los Angeles, United States, IEEE, December 2019, https://hal.archives-ouvertes.fr/hal-02415684

[20] P. SILVA, A. COSTAN, 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, May 2019, pp. 904-907 [*DOI :* 10.1109/IPDPSW.2019.00149], https://hal.inria.fr/hal-02310154

Other Publications

[21] K. FAUVEL, D. BALOUEK-THOMERT, D. MELGAR, P. SILVA, A. SIMONET, G. ANTONIU, A. COSTAN, V. MASSON, M. PARASHAR, I. RODERO, A. TERMIER. *A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02373429

References in notes

[22] *Amazon Elastic Map-Reduce (EMR)*, 2017, https://aws.amazon.com/emr/

[23] *Digital Single Market*, 2015, https://ec.europa.eu/digital-single-market/en/digital-single-market/

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

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

[26] *European Cloud Strategy*, 2012, https://ec.europa.eu/digital-single-market/en/european-cloud-computing-strategy/

[27] *Apache Flink*, 2016, http://flink.apache.org/

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

[29] *Scientific challenges of the Inria Rennes-Bretagne Atlantique research centre*, 2016, https://www.inria.fr/centre-inria-rennes-bretagne-atlantique/

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

[31] *Joint Laboratory for Extreme Scale Computing (JLESC)*, 2017, https://jlesc.github.io/

[32] *Apache Spark*, 2017, http://spark.apache.org/

[33] *Storm*, 2014, http://storm.apache.org/

[34] T. AKIDAU, A. BALIKOV, K. BEKIROĞLU, S. CHERNYAK, J. HABERMAN, R. LAX, S. MCVEETY, D. MILLS, P. NORDSTROM, S. WHITTLE. *MillWheel: fault-tolerant stream processing at internet scale*, in "Proceedings of the VLDB Endowment", 2013, vol. 6, n^{o} 11, pp. 1033–1044

[35] J. DEAN, S. GHEMAWAT. *MapReduce: simplified data processing on large clusters*, in "Communications of the ACM", 2008, vol. 51, n^{o} 1, pp. 107–113