IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

École normale supérieure de Rennes

Activity Report 2017

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
A1.1.6. - Cloud
A1.1.9. - Fault tolerant systems
A1.3. - Distributed Systems
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

Other Research Topics and Application Domains:

B3.2. - Climate and meteorology
B3.3.1. - Earth and subsoil
B8.2. - Connected city
B9.4.5. - Data science

1. Personnel

Research Scientists

Gabriel Antoniu [Team leader, Inria, Senior Researcher, HDR]
Shadi Ibrahim [Inria, Researcher, until March 2017]

Faculty Members

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

Post-Doctoral Fellow

Chi Zhou [Inria, until March 2017 (also known as Amelie Chi Zhou)]

PhD Students

Lokman Rahmani [University of Rennes 1, until February 2017, then a visitor from April 2017 to June 2017]
Luis Eduardo Pineda Morales [Inria, until April 2017, currently R&D Engineer at ActiveEon]
Tien-Dat Phan [University of Rennes 1, until November 2017, currently R&D Engineer at Dassault Systèmes]
Orçun Yildiz [Inria, until November 2017]
Ovidiu-Cristian Marcu [Inria]
Mohammed-Yacine Taleb [Inria]
Nathanaël Cheriere [École normale supérieure de Rennes]
Paul Le Noac'h [INSA Rennes]
Pierre Matri [Universidad Politécnica de Madrid (UPM), Espagne]

Technical staff
Hadi Salimi [Research Engineer, ADT Damaris, Inria]

Intern
Mukrram Ur Rahman [Inria, from May 2017 to August 2017]

Administrative Assistant
Aurélie Patier [University of Rennes 1]

Visiting Scientist
Jose Aguilar Canepa [Instituto Politécnico Nacional, Mexico, Mexique, from November 2017 to December 2017]

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 [35], [32] and Exascale HPC [34] research communities. The Big Data area, which has recently captured a lot of attention, 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 [39], 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 centre of Rennes [38]: *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 [45] parallel programming paradigm. It is currently used on most cloud platforms, following the trend set up by Amazon [30]. 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 [37], EESI [33], ETP4HPC [34]. 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 [40]). 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 Extreme-Scale Computing and Big Data Infrastructures

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 eco-systems is crucial to the future. Our objective for the next years is premised on the idea that we must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change the plans that are now being laid for achieving Exascale computing.

Collaboration. *This axis is addressed in close collaboration with María Pérez (UPM), Rob Ross (ANL), Toni Cortes (BSC), Bogdan Nicolae (formerly at IBM Research, now at Huawei Research).*

Relevant groups with similar interests are the following ones.

- *The group of Jack Dongarra, Innovative Computing Laboratory at University of Tennessee/Oak Ridge National Laboratory, working on joint tools Exascale Computing and Big Data.*
- *The group of Satoshi Matsuoka, Tokyo Institute of Technology, working on system software for Clouds and HPC.*
- *The group of Franck Cappello at Argonne National Laboratory/NCSA 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

We argue that 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.

We argue 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, multi-blob 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. Big Data analytics on Exascale HPC machines

Big Data analytics is another interesting direction that we plan to explore, building on top of these converged storage architectures. Specifically, we will examine the ways in which Exascale infrastructures can be leveraged not only by HPC-centric, but also by scientific, cloud-centric applications. Many of the current state-of-the-art Big Data processing approaches, including Hadoop and Spark [41] are optimized to run on commodity machines. This impacts the mechanisms used to deal with failures and the limited network bandwidth.

A blind adoption of these systems on extreme-scale platforms would result in high overheads. It would therefore prevent users from fully benefiting from the high performance infrastructure. The objective that we set here is to explore design and implementation options for new data analytics systems that can exploit the features of extreme-scale HPC machines: multi-core nodes, multiple memory and storage technologies including a large memory, NVRAM, SSDs, etc.

3.2. Research axis 2: Advanced data processing on Clouds

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 management techniques to cope with the volatile requirements of data-intensive applications on large-scale dynamic clouds in a cost-efficient way.

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

Relevant groups with similar interests include the following ones.

- *The AMPLab, UC Berkeley, USA, working on scheduling stream data applications in heterogeneous clouds.*
- *The group of Ewa Deelman, USC Information Sciences Institute, working on resource management for workflows in Clouds.*
- *The XTRA group, Nanyang Technological University, Singapore, working on resource provisioning for workflows in the cloud.*

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 [36], Spark [41], Storm [42], and Google MillWheel [44]. 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. We will explore new mechanisms that expose resource heterogeneity (observed variability in resource utilization at runtime) when scheduling stream data applications. We will also investigate how to adapt to node failures automatically, and to adapt the failure handling techniques to the characteristics of the running application and to the root cause of failures.

3.2.2. Geographically distributed workflows on multi-site clouds

Many data processing jobs in data-intensive applications are modeled as workflows (i.e., as sets of tasks linked according to their data and computation dependencies) to facilitate the management and analysis of large volumes of data. With the fast growth of volumes of data to be handled at larger and larger scales, geographically distributed workflows are emerging as a natural data processing paradigm. This may bring several benefits: resilience to failures, distribution across partitions (e.g., moving computation close to data or vice versa), elastic scaling to support usage bursts, user proximity, etc.

In this context, sharing, disseminating and analyzing the data sets results in frequent large-scale data movements across widely distributed sites. Studies show that the inter-datacenter traffic is expected to triple in the following years. Our objective is to investigate approaches to data management enabling an efficient execution of such geographically distributed workflows running on multi-site clouds.

While in the past years we have addressed some data management issues in this area, mainly in support to efficient task scheduling of scientific workflows running on multisite clouds, we will now focus on an increasingly common scenario where workflows generate and process a huge number of small files, which is particularly challenging. As such workloads generate a deluge of small and independent I/O operations, efficient data and metadata handling is critical. We will explore specific means to better hide latency for data and metadata access in such scenarios, as a way to improve global performance.

3.3. Research axis 3: I/O management, 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 have motivated 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.

Collaboration. *This axis is worked out in close collaboration with Rob Ross (ANL), Tom Peterka (ANL), Matthieu Dorier (ANL), Toni Cortes (BSC), Bruno Raffin (Inria). Some additional collaborations are in discussion with other members of JLESC, and with CEA and Total.*

Relevant groups with similar interests include the following ones.

- *The group of Manish Parashar at Rutgers University, USA (I/O management for HPC systems, in situ processing).*
- *The group of Scott Klasky at Oak Ridge National Lab, USA (I/O management for HPC systems, in situ processing).*
- *The CNRS IPSL laboratory (Sébastien Denvil, Pôle de modélisation du climat) in Paris, France (in situ data analytics).*

3.3.1. Toward a joint optimized architecture for in situ visualization and advanced processing

From Inria and ANL, four tools at least have emerged to address some challenges of coupling simulations with visualization packages or analysis workflows. Each of them focused on some particular aspect:

Damaris (Inria, [5], [4]) exploits dedicated cores to enable jitter-free I/O and in situ visualization;

Decaf (ANL, [31]) implements a coupling service for workflows;

FlowVR (Inria, [43]) connects workflow components for in situ processing;

Swift (ANL, [46]) focuses on implicitly parallel data flows and was optimized for Big Data processing.

Our plan is to explore how these tools could best leverage their respective strengths in a *joint optimized architecture for in situ visualization and advanced processing* in the HPC area. We published a preliminary study describing the lessons learned from using these tools in production environments with real applications [7]. Such a joint architecture will contribute to address the data volume and velocity challenges raised by data-intensive workflows, including complex data-intensive analytics phases. It may also impact, in a subsequent step, future data analysis pipelines for converged Big Data and HPC architectures.

4. Highlights of the Year

4.1. Highlights of the Year

Euro-Par Steering Committee. Luc Bougé has been elected as the new Steering Commitee Chairman of the Euro-Par international conference on parallel and distributed processing. He is the successor of Prof. Christian Lengauer, University of Passau, Germany.

IEEE Cluster 2017 conference. Three years after the 2014 edition, the KerData team had again a leading role in the organization of the 2017 edition: Gabriel Antoniu served as Program Chair, Alexandru Costan served as Submissions Chair.

IEEE Big Data 2017 conference. Alexandru Costan served as Posters Chair.

5. New Software and Platforms

5.1. BlobSeer

BlobSeer : A Storage System For The Exascale Era
KEYWORDS: Versioning - HPC - Cloud storage - Distributed metadata - MapReduce
SCIENTIFIC DESCRIPTION: BlobSeer is a large-scale distributed storage service that addresses advanced data management requirements resulting from ever-increasing data sizes. It is centered around the idea of leveraging versioning for concurrent manipulation of binary large objects in order to efficiently exploit data-level parallelism and sustain a high throughput despite massively parallel data access.
FUNCTIONAL DESCRIPTION: BlobSeer is a large-scale distributed storage service for advanced management of massive data. Validated on Nimbus, OpenNebula and Microsoft Azure cloud platforms.

- Participants: Bogdan Nicolae, Gabriel Antoniu and Luc Bougé
- Partners: Université de Rennes 1 - ENS Cachan
- Contact: Gabriel Antoniu
- URL: http://blobseer.gforge.inria.fr/

5.2. Damaris

KEYWORDS: Big data - Visualization - I/O - HPC - Exascale
SCIENTIFIC DESCRIPTION: Damaris is a middleware for multicore SMP nodes enabling them to efficiently handle data transfers for storage and visualization. The key idea is to dedicate one or a few cores of each SMP node to the application I/O. It is developed within the framework of a collaboration between KerData and the Joint Laboratory for Petascale Computing (JLPC). The current version enables efficient asynchronous I/O, hiding all I/O related overheads such as data compression and post-processing, as well as direct (in situ) interactive visualization of the generated data.

Damaris has been preliminarily evaluated at NCSA (Urbana-Champaign) with the CM1 tornado simulation code. CM1 is one of the target applications of the Blue Waters supercomputer in production at NCSA/UIUC (USA), in the framework of the Inria-UIUC-ANL Joint Lab (JLPC). Damaris now has external users, including (to our knowledge) visualization specialists from NCSA and researchers from the France/Brazil Associated research team on Parallel Computing (joint team between Inria/LIG Grenoble and the UFRGS in Brazil). Damaris has been successfully integrated into three large-scale simulations (CM1, OLAM, Nek5000). Works are in progress to evaluate it in the context of several other simulations including HACC (cosmology code) and GTC (fusion).
FUNCTIONAL DESCRIPTION: Damaris is a middleware for data management targeting large-scale HPC simulations: • «In-situ» data analysis by some dedicated cores of the simulation platform • Asynchronous and fast data transfer from HPC simulations to Damaris • Semantic-aware dataset processing through Damaris plug-ins

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

5.3. iHadoop

FUNCTIONAL DESCRIPTION: iHadoop is a Hadoop simulator developed in Java on top of SimGrid to simulate the behavior of Hadoop and therefore accurately predict the performance of Hadoop in normal scenarios and under failures.

iHadoop is an internal software prototype, which was initially developed to validate our idea for exploring the behavior of Hadoop under failures. iHadoop has preliminarily evaluated within our group and it has shown very high accuracy when predicating the execution time of a Map-Reduce application. We intend to integrate iHadoop within the SimGrid distribution and make it available to the SimGrid community.

- Participants: Shadi Ibrahim and Tien Dat Phan
- Contact: Shadi Ibrahim

5.4. JetStream

FUNCTIONAL DESCRIPTION: JetStream is a middleware solution for batch-based, high-performance streaming across cloud data centers. JetStream implements a set of context-aware strategies for optimizing batch-based streaming, being able to self-adapt to changing conditions. Additionally, the system provides multi-route streaming across cloud data centers for aggregating bandwidth by leveraging the network parallelism. It enables easy deployment across .Net frameworks and seamless binding with event processing engines such as StreamInsight.

JetStream is currently used at Microsoft Research ATLE Munich for the management of the Azure cloud infrastructure.

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

5.5. 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. New Results

6.1. Convergence of HPC and Big Data

6.1.1. Týr: Blob-based storage convergence of HPC and Big Data

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

The increasingly growing data sets processed on HPC platforms raise major challenges for the underlying storage layer. A promising alternative to POSIX-I/O-compliant file systems are simpler blobs (binary large objects), or object storage systems. They offer lower overhead, better performance and horizontal scalability at the cost of largely unused features such as file hierarchies or permissions. Similarly, blobs are increasingly considered for replacing distributed file systems for big data analytics or as a base for storage abstractions like key-value stores or time-series databases.

This growing interest from both HPC and Big Data communities towards blob storage naturally fits with the current trend towards HPC and Big Data convergence. In this context, we seek to demonstrate that blob storage indeed constitutes a strong alternative to current storage infrastructures. Additionally, the data model of blob storage is close enough to that of distributed file systems so that this change is largely transparent for the applications running atop them.

In [22] we provide a preliminary evaluation of blob storage in HPC and Big Data contexts. We leverage a series of real-world HPC applications as well as an industry-standard HPC benchmark. We analyze for each of these applications the storage requests sent to the underlying storage system. We discover that over 98% of these storage calls can be directly mapped to the data model offered by blobs. Interestingly, we also note that the remaining calls are using file systems features for convinience rather than by necessity. These calls may consequently be performed as offline pre- or post-processing, or avoided altogether without altering the application.

6.1.2. Modeling elastic storage

Participants: Nathanaël Cheriere, Gabriel Antoniu.

For efficient Big Data processing, efficient resource utilization becomes a major concern as large-scale computing infrastructures such as supercomputers or clouds keep growing in size. 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.

In [18], we evaluate the viability of malleability as a design principle for a distributed file system. We specifically model the duration of the decommission operation, for which we obtain a theoretical lower bound. Then we consider HDFS as a use case and we show that our model can explain the measured decommission times.

The existing decommission mechanism of HDFS is good when the network is the bottleneck, but could be accelerated by up to a factor 3 when the storage is the limiting factor. With the highlights provided by our model, we suggest improvements to speed up decommission in HDFS and we discuss open perspectives for the design of efficient malleable distributed file systems.

6.1.3. Eley: Leveraging burst-buffers for efficient Big Data processing on HPC systems

Participants: Orçun Yildiz, Chi Zhou, Shadi Ibrahim.

Burst Buffer is an effective solution for reducing the data transfer time and the I/O interference in HPC systems. Extending Burst Buffers (BBs) to handle Big Data applications is challenging because BBs must account for the large data inputs of Big Data applications and the performance guarantees of HPC applications – which are considered as first-class citizens in HPC systems. Existing BBs focus on only intermediate data of Big Data applications and incur a high performance degradation of both Big Data and HPC applications.In [26], we present *Eley*, a burst buffer solution that helps to accelerate the performance of Big Data applications while guaranteeing the performance of HPC applications. In order to improve the performance of Big Data applications, *Eley* employs a prefetching technique that fetches the input data of these applications to be stored

close to computing nodes thus reducing the latency of reading data inputs. Moreover, Eley is equipped with a full delay operator to guarantee the performance of HPC applications – as they are running independently on a HPC system. The experimental results show the effectiveness of *Eley* in obtaining shorter execution time of Big Data applications (shorter map phase) while guaranteeing the performance of HPC applications.

6.2. Scalable data processing on clouds

6.2.1. Low-latency storage for stream processing

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu, María Pérez, Radu Tudoran, Stefano Bortoli, Bogdan Nicolae.

We are now witnessing an unprecedented growth of data that needs to be processed at always increasing rates in order to extract valuable insights. Big Data applications are rapidly moving from a batch-oriented execution model to a streaming execution model in order to extract value from the data in real-time. Big Data streaming analytics tools have been developed to cope with the online dimension of data processing: they enable real-time handling of live data sources by means of stateful aggregations (window-based operators). In [21] we design a deduplication method specifically for window-based operators that rely on key-value stores to hold a shared state. Our key finding is that more fine-grained interactions between streaming engines and (key-value) stores (i.e., the data ingest, store, and process interfaces) need to be designed in order to better respond to scenarios that have to overcome memory scarcity.

Moreover, processing live data alone is often not enough: in many cases, such applications need to combine the live data with previously archived data to increase the quality of the extracted insights. Current streaming-oriented runtimes and middlewares are not flexible enough to deal with this trend, as they address ingestion (collection and pre-processing of data streams) and persistent storage (archival of intermediate results) using separate services. This separation often leads to I/O redundancy (e.g., write data twice to disk or transfer data twice over the network) and interference (e.g., I/O bottlenecks when collecting data streams and writing archival data simultaneously). In [20] and [27] we argue for a unified ingestion and storage architecture for streaming data that addresses the aforementioned challenge and we identify a set of constraints and benefits for such a unified model, while highlighting the important architectural aspects required to implement it in real life.

Based on these findings, we are currently developing a low-latency stream storage framework that addresses such critical real-time needs for efficient stream processing, exposing high-performance interfaces for stream ingestion, storage, and processing.

6.2.2. A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications

Participants: Paul Le Noac'h, Alexandru Costan.

Stream computing is becoming a more and more popular paradigm as it enables the real-time promise of data analytics. Apache Kafka is currently the most popular framework used to ingest the data streams into the processing platforms. However, how to tune Kafka and how much resources to allocate for it remains a challenge for most users, who now rely mainly on empirical approaches to determine the best parameter settings for their deployments. Our goal in [28] is to make a through evaluation of several configurations and performance metrics of Kafka in order to allow users avoid bottlenecks, reach its full potential and avoid bottlenecks and eventually leverage some good practice for efficient stream processing.

6.2.3. Hot metadata management for geographically distributed workflows

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu, Ji Liu, Esther Pacitti, Patrick Valduriez, Marta Mattoso.

Large-scale scientific applications are often expressed as scientific workflows (SWfs) that help defining data processing jobs and dependencies between jobs' activities. Several SWfs have huge storage and computation requirements, and so they need to be processed in multiple (cloud-federated) datacenters. It has been shown that efficient metadata handling plays a key role in the performance of computing systems. However, most of this evidence concern only single-site, HPC systems to date. In addition, the efficient scheduling of tasks among different datacenters is critical to the SWf execution. In [19], we present a hybrid distributed model and architecture, using hot metadata (frequently accessed metadata) for efficient SWf scheduling in a multisite cloud. We couple our model with a scientific workflow management system (SWfMS) to validate its applicability to real-life scientific workflows with different scheduling algorithms. We show that the combination of efficient management of hot metadata and scheduling algorithms improves the performance of SWfMS, reducing the execution time of highly parallel jobs up to 64.1 % and that of the whole scientific workflows up to 37.5 %, by avoiding unnecessary cold metadata operations. We also further discuss how to dynamically handle such hot metadata.

6.3. Scalable I/O, storage and in-situ processing in Exascale environments

6.3.1. Extreme-scale logging through application-defined storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Applications generating data as logs and seeking to store it as such face hard challenges on HPC platforms. In distributed systems this storage model is key to ensuring fault-tolerance, developing transactional systems or publish-subscribe models. In scientific applications, distributed logs can play many roles such as in-situ visualization of large data streams, centralized collection of telemetry or monitoring events computational steering, data aggregation from array of physical sensors or live data indexing. Distributed shared logs are very difficult to implement on common HPC platforms due to the lack of efficient append operation in the current file-based storage infrastructures. While part of the POSIX standard, this operation has not been the main focus during the development of parallel file systems. While application-specific, custom-built solutions are possible, they require a significant development effort and often fail to meet the performance requirements of data-intensive applications running at large scale.

In this work we go through the basic requirements of storing telemetry data streams for computational steering and visualization. For simple use cases where the telemetry data is only temporary, we prove that distributed logging can be performed at scale by leveraging state-of-the-art blob storage systems such as Týr or RADOS. This approach is supported by the growing availability of node-local storage on a new generation of supercomputers, giving application developers the freedom to deploy transient storage systems alongside the application directly on the compute nodes.

When long-term storage of the generated data is needed for offline visualization or analytics, we prove that distributed logs require a significantly lower number of output logs to achieve peak performance compared to Lustre or GPFS. We also prove that this low number of output files obviates the need for an explicit post-processing merge step in most cases for iterating the whole output log in generation order. We finally prove on up to 100,000 cores of the Theta supercomputer that our findings are applicable to run distributed logging at large scale, while improving write throughput by several orders of magnitude compared to Lustre of GPFS.

6.3.2. Leveraging Damaris for in-situ visualization in support of GeoScience and CFD Simulations

Participants: Hadi Salimi, Matthieu Dorier, Luc Bougé.

Damaris is a middleware for in situ data analysis and visualization targeting extreme-scale, MPI-based simulations. The main goal of Damaris is to provide a simple method to instrument a simulation in order to benefit from in situ analysis and visualization. To this aim, the computing resources are partitioned such that a subset of cores in a SMP node or a subset of nodes of the underlying platform are dedicated to in situ processing. The data generated by the simulation are passed to these dedicated processes 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 both in synchronous and asynchronous modes. Afterwards, the processed data can be analyzed or visualized. Damaris also supports a very simple API to instrument simulations developed in different domains. Moreover, using some XML configuration files for defining simulation data types (e.g. meshes) makes the instrumentation process easier with minimum code modifications. Active development is currently continuing within the KerData team, where it is at the center of several collaborations with industry (e.g Total) as well as with national and international academic partners.

In recent developments of Damaris, we have focused on two main targets that are: 1) Instrumenting new simulations codes from different scientific domains, i.e. geoscience and ocean modeling, 2) Implementing new storage backends, i.e. HDF5 for Damaris. In this regard, we report the results of some experiments we made to evaluate Damaris with respect to performance. These experiments were conducted on Grid'5000 test bed. In these experiments Damaris was employed to visualize the data generated by the Wave Propagation geoscience simulation and also the CROCO coastal and ocean simulation. During the experiments, the impact of Damaris was measured by comparing the simulations instrumented by Damaris (space partitioning approach) with a baseline where those simulations include data processing codes directly on their source code (time partitioning approach). The results of these simulations show that the incorporation of Damaris into a simulation decreases the total run time of the simulation due to its asynchronous data processing and visualization capabilities. In addition, using Damaris for data visualization has nearly no impact on the total run time of the mentioned simulation codes. We also have shown that the amount of code changes necessary for instrumenting the simulation codes is much less compared to the case that the simulation code is instrumented by native visualization or storage APIs. Moreover, we also have studied the impact of new HDF5 storage backend, on storing simulation results in HDF5 format in both file-per-dedicated-core and collective I/O scenarios.

6.3.3. Accelerating MPI collective operations on the Theta cupercomputer

Participants: Nathanaël Cheriere, Matthieu Dorier, Misbah Mubarak, Robert Ross, Gabriel Antoniu.

Recent network topologies in supercomputers have motivated new research on topology-aware collective communication algorithms for MPI. But such 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 focus on the Scatter and AllGather operations in the context of the Theta supercomputer's dragonfly topology. We propose a set of topology-aware versions of these operations as well as optimizations of the old, non-topology-aware ones. We conduct an extensive simulation campaign using the CODES network simulator. Our results show that, contrary to our expectations, topology-awareness does not help improving significantly the speed of these operations. Rather, the high radix and low diameter of the dragonfly topology, along with already good routing protocols, enable simple algorithms based on non-blocking communications to perform better than state-of-the-art algorithms. A trivial implementation of Scatter using nonblocking 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 nonblocking communications.

6.4. Energy-aware data storage and processing at large scale

6.4.1. Performance and energy-efficiency trade-offs in in-memory storage systems

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu, Toni Cortes.

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and offer a low response time. As the main memory capacity of clusters and clouds increases, it becomes possible to keep most of the data in the main memory. This motivates the introduction of in-memory storage systems. While prior work has focused on how to exploit the low-latency of in-memory access at scale, there is very little visibility into the energy-efficiency of in-memory storage systems. Even though it is known that main memory is a fundamental energy bottleneck in computing systems (i.e., DRAM consumes up to 40% of a server's power). During this project, by the means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud - a well-known in-memory storage system. We reveal that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also find that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly, we show that replication can also play a negative role in crash-recovery.

6.4.2. Energy-aware straggler mitigation in Map-Reduce

Participants: Tien-Dat Phan, Chi Zhou, Shadi Ibrahim, Guillaume Aupy, Gabriel Antoniu.

Energy consumption is an important concern for large-scale data-centers, which results in huge monetary cost for data-center operators. Due to the hardware heterogeneity and contentions between concurrent workloads, straggler mitigation is important to many Big Data applications running in large-scale data-centers and the speculative execution technique is widely-used to handle stragglers. Although a large number of studies have been proposed to improve the performance of Big Data applications using speculative execution, few of them have studied the energy efficiency of their solutions.

In [23], we propose two techniques to improve the energy efficiency of speculative executions while ensuring comparable performance. Specifically, we propose a hierarchical straggler detection mechanism which can greatly reduce the number of killed speculative copies and hence save the energy consumption. We also propose an energy-aware speculative copy allocation method which considers the trade-off between performance and energy when allocating speculative copies. We implement both techniques into Hadoop and evaluate them using representative Map-Reduce benchmarks. Results show that our solution can reduce the energy waste on killed speculative copies by up to 100% and improve the energy efficiency by 20% compared to state-of-the-art mechanisms.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

7.1.1. Huawei: HIRP Low-Latency Storage for Stream Data (2016–2017)

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

The goal of this project is to explore the plausible paths towards a dedicated storage solution for low-latency stream storage. Such a solution should provide on the one hand traditional storage functionality and on the other hand stream-like performance (i.e., low-latency I/O access to items and ranges of items).

We have investigated the main requirements and challenges, evaluated the different design choices (e.g., a standalone component vs. an extension of an existing Big Data solution like HDFS) and proposed a new converged architecture for smart storage.

7.1.2. Total: In situ Visualization with Damaris (2017-2018).

Participants: Hadi Salimi, Matthieu Dorier, Gabriel Antoniu, Luc Bougé.

The goal of this expertise contract is to 1) disseminate the usage of Damaris for engineers at Total; 2) to realize a feasibility study for the usage of Damaris for in situ analysis of data for Total's HPC simulations.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. OverFlow (2015–2019)

Participants: Alexandru Costan, Paul Le Noac'h.

- Project Acronym: OverFlow.
- Project Title: Workflow Data Management as a Service for Multisite Applications.
- Coordinator: Alexandru Costan.
- Duration: Octobre 2015–October 2019.
- Other Partners: None (Young Researcher Project).
- External collaborators: Kate Keahey (University of Chicago and Argonne National Laboratory), Bogdan Nicolae (Huawei Research) and Christophe Blanchet (Institut Français de Bioinformatique).
- Abstract: This JCJC project led by Alexandru Costan investigates approaches to data management enabling an efficient execution of geographically distributed workflows running on multi-site clouds.
- Progress: In 2017, we have reviewed in depth the technical and architectural needs of data storage for the use cases that drive OverFlow, in order to consolidate a set of requirements for its future architecture. Based on these workflow traces, in a second step, we have investigated the suitable benchmarks that reasonably represent them. In this direction, we have first focused on ingestion and storage optimisations for such complex deployments, in particular the novel support for concurrent writes. The project was successfully reviewed at T0+18.

8.1.1.2. KerStream (2017–2021)

Participant: Shadi Ibrahim.

- Project Acronym: KerStream.
- Project Title: Big Data Processing: Beyond Hadoop!
- Coordinator: Shadi Ibrahim .
- Duration: January 2017–January 2021.
- Other Partners: None (Young Researcher Project).
- Abstract: This JCJC project led by Shadi Ibrahim aims to address the limitations of Hadoop when running stream Big Data applications on large-scale clouds and to do a step beyond Hadoop by proposing a new approach, called KerStream, for scalable and resilient stream Big Data processing on clouds. The KerStream project can be seen as the first step towards developing the first French middleware that handles Stream Data processing at Scale.
- Note: Shadi Ibrahim left the KerData team in April 2017, so that this contract is no longer managed within the KerData team.

8.1.2. Other National Projects

8.1.2.1. ADT Damaris

Participants: Hadi Salimi, Alexandru Costan, Luc Bougé.

- Project Acronym: ADT Damaris
- Project Title: Technology development action for te Damaris environment.
- Coordinator: Alexandru Costan.
- Duration: 2016–2018.
- Abstract: 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.

Hadi Salimi is funded through this project to document, test and extend the Damaris software and make it a safely distributable product.

8.1.2.2. 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. FP7 & H2020 Projects

8.2.1.1. BigStorage

Title: BigStorage: Storage-based Convergence between HPC and Cloud to handle Big Data.

Programme: H2020.

Duration: January 2015–December 2018.

Coordinator: Universidad Politécnica de Madrid (UPM).

Partners:

- Barcelona Supercomputing Center — Centro Nacional de Supercomputacion (Spain)
- CA Technologies Development Spain (Spain)
- CEA — Commissariat à l'énergie atomique et aux énergies alternatives (France)
- Deutsches Klimarechenzentrum (Germany)
- Foundation for Research and Technology Hellas (Greece)
- Fujitsu Technology Solutions (Germany)
- Johannes Gutenberg Universitaet Mainz (Germany)
- Universidad Politecnica de Madrid (Spain)
- Seagate Systems UK (United Kingdom)

Inria contact: Gabriel Antoniu and Adrien Lèbre.

URL: http://www.bigstorage-project.eu/.

Description: BigStorage is a European Training Network (ETN) whose main goal is to train future *data scientists*. It aims at enabling them and us to apply holistic and interdisciplinary approaches to take advantage of a data-overwhelmed world. This world requires *HPC* and *Cloud* infrastructures with a redefinition of *storage* architectures underpinning them — focusing on meeting highly ambitious performance and *energy* usage objectives. The KerData team is hosting 2 *Early Stage Researchers* in this framework and co-advises an extra PhD student.

8.2.2. Collaborations with Major European Organizations

Gabriel Antoniu and Alexandru Costan are serving as Inria representatives in the working groups dedicated to *HPC-Big Data* convergence within the Big Data Value Association (BDVA) and the European Technology Platform in the area of High-Performance Computing (ETP4HPC). They are contributing to the respective Strategic Research Agendas of BDVA and ETP4HPC.

8.3. International Initiatives

8.3.1. Inria International Labs

8.3.1.1. JLESC: Joint Laboratory on Extreme-Scale Computing

The Joint Laboratory on Extreme-Scale Computing is jointly run by Inria, UIUC, ANL, BSC, JSC and RIKEN/AICS. It has been created in 2014 as a follow-up of the Inria-UIUC JLPC, the *Joint Laboratory for Petascale Computing*.

The KerData team is collaborating with teams from ANL and UIUC within this lab since 2009 on several topics in the areas of I/O, storage and in situ processing and cloud computing. This collaboration has been initially formalized as the *Data@Exascale* Associate Team with ANL and UIUC (2013–2015) followed by *Data@Exascale 2* Associate Team with ANL (2016–2018). Our activities in this framework are described here: http://www.irisa.fr/kerdata/data-at-exascale/

Since 2015, Gabriel Antoniu serves as a topic leader for Inria for the *I/O, Storage and In Situ Processing* topic. Ongoing lab research directions and projects he is co-supervising in this area are described here: https://jlesc.github.io/projects/ in the *I/O, Storage and In-Situ Processing* section.

Since 2017, Gabriel Antoniu is serving as *Vice-Executive Director of JLESC for Inria*.

8.3.1.1.1. Associate Team involved in the International Lab: Data@Exascale 2

Project Acronym: Data@Exascale 2.

Project Title: Convergent Data Storage and Processing Approaches for Exascale Computing and Big Data Analytics.

International Partner: Argonne National Laboratory (United States) — Mathematics and Computer Science Division (MCS) — Rob Ross.

Start year: 2013.

URL: http://www.irisa.fr/kerdata/data-at-exascale/.

Description: In the past few years, countries including United States, the European Union, Japan and China have set up aggressive plans to get closer to what appears to be the next goal in terms of high-performance computing (HPC): Exaflop computing, a target which is now considered reachable by the next-generation supercomputers in 2020-2023. While these government-led initiatives have naturally focused on the big challenges of Exascale for the development of new hardware and software architectures, the quite recent emergence of the Big Data phenomenon introduces what could be called a tectonic shift that is impacting the entire research landscape for Exascale computing. As data generation capabilities in most science domains are now growing substantially faster than computational capabilities, causing these domains to become data-intensive, new challenges appeared in terms of volumes and velocity for data to be stored, processed and analyzed on the future Exascale machines.

To face the challenges generated by the exponential data growth (a general phenomenon in many fields), a certain progress has already been made in the recent years in the rapidly-developing, industry-led field of cloud-based Big Data analytics, where advanced tools emerged, relying on machine-learning techniques and predictive analytics.

Unfortunately, these advances cannot be immediately applied to Exascale computing: the tools and cultures of the two worlds, HPC (High-Performance Computing) and BDA (Big Data Analytics) have developed in a divergent fashion (in terms of major focus and technical approaches), to the detriment of both. The two worlds share however multiple similar challenges and unification now appears as essential in order to address the future challenges of major application domains that can benefit from both.

The scientific program we propose for the Data@Exascale 2 Associate Team is defined from this new, highly-strategic perspective and builds on the idea that the design of innovative approaches to data I/O, storage and processing allowing Big Data analytics techniques and the newest HPC architectures to leverage each other clearly appears as a key catalyst factor for the convergence process.

Activities in 2017 are described on the web site of the Associate Team.

8.3.2. Inria International Partners

8.3.2.1. Declared Inria International Partners

8.3.2.2. DataCloud@Work

Title: DataCloud@Work.

International Partner:

- Polytechnic University of Bucharest (Romania), Computer Science Department, Nicolae Tapus and Valentin Cristea.

Duration: 5 years.

Start year: 2013. The status of IIP was established right after the end of our former *DataCloud@work* Associate Team (2010–2012).

URL: https://www.irisa.fr/kerdata/doku.php?id=cloud_at_work:start.

Description: Our research topics address the area of distributed data management for cloud services, focusing on autonomic storage. The goal is explore how to build an efficient, secure and reliable storage IaaS for data-intensive distributed applications running in cloud environments by enabling an autonomic behavior.

8.3.2.3. Informal International Partners

Instituto Politécnico Nacional, IPN, Ciudad de México: We continued our informal collaboration in the area of stream processing. A PhD student from IPN (José Aguilar Canepa) was hosted by the KerData team for a 1-month internship, during which he identified optimization problems that can be subject to joint work (see Internships section below).

National University of Singapore (NUS): We collaborate on resource management for workflows in the cloud and optimizing graph processing in geo-distributed data-centers.

8.3.3. Participation in Other International Programs

8.3.3.1. International Initiatives

8.3.3.1.1. 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 an yearly series of invitation-based international workshops. In 2017 Gabriel Antoniu was solicited to co-lead the BDEC working group dedicated to exploring convergence-related challenges for hybrid architectures combining HPC systems, clouds and fog/edge computing infrastructures with Geoffrey Fox and Ewa Deelman. The contributions are reflected in the final report on convergence available on the BDEC web page.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

José Aguilar Canepa (Instituto Politécnico Nacional, IPN, Mexico) visited the KerData team for one month (November 2017) in order to setup a common topic of research for the future proposal of an Associate Team Kerdata-IPN.

8.4.1.1. Internships

Mukrram Rahman (M1, University of Rennes 1) has done a 3-month internship within the team, working with Ovidiu Marcu and Alexandru Costan on HDFS extensions for dedicated stream storage.

8.4.2. Visits to International Teams

8.4.2.1. Research Stays Abroad

Pierre Matri has done a 3-month internship at Argonne National Lab, to work on extreme-scale logging through application-defined storage under the supervision of Phil Carns and Rob Ross. See Section *New Results* for details.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Luc Bougé: Chair of the Steering Committee of the Euro-Par Series of conferences since August 2017.

9.1.2. Scientific Events Selection

9.1.2.1. Chair of Conference Program Committees

Gabriel Antoniu:

- Program Chair of the IEEE Cluster 2017 international conference.
- Vice-Chair of the Program Committee of the ACM/IEEE CCGrid 2017 international conference (Hybrid and Mobile Clouds Area), Madrid, May 2017.

Alexandru Costan:

- Program Co-Chair of the ScienceCloud 2017 international workshop held in conjunction with HPDC 2017, Washington, USA.
- Posters Chair of the IEEE Big Data 2017.
- Submissions Chair of the IEEE Cluster 2017.
- Track Chair of the IEEE ScalCom 2017 (Tools for Big Data track)

9.1.2.2. Member of Conference Program Committees

Gabriel Antoniu: IEEE/ACM SC'17 (Papers), ACM HPDC 2017.

Luc Bougé: ISC HPC 2017.

Alexandru Costan: IEEE/ACM SC'17 (Posters), ACM/IEEE CCGrid 2017, IEEE/ACM UCC 2017, ARMS-CC 2017 workshop (held in conjunction with PODC 2016), IEEE Big Data 2017, MLDS 2017, EBDMA 2017, ISPDC 2017, CSCS 2017.

9.1.2.3. Reviewer

Luc Bougé: ISC 2017, SC 2017, BigData 2017, IPDPS 2017, etc.

Alexandru Costan: HPDC 2017, SC 2017, IPDPS 2017

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

Gabriel Antoniu: Future Generation Computer Systems: Special Issue on Mobile, hybrid, and heterogeneous clouds for cyberinfrastructures (Guest Editor, 2017).

Luc Bougé: Concurrency and Computation: Practice and Experience, Special Issues on the Euro-Par conference.

Alexandru Costan: Soft Computing Journal, Special Issue on Autonomic Computing and Big Data Platforms

9.1.3.2. Reviewer - Reviewing Activities

Gabriel Antoniu: Concurency and Computation: Practice and Experience.

Luc Bougé: IEEE Transactions on Distributed Parallel Systems.

Alexandru Costan: IEEE Transactions on Parallel and Distributed Systems, Future Generation Computer Systems, Concurrency and Computation Practice and Experience, IEEE Communications, IEEE Transactions on Storage, Information Sciences, IEEE Transactions on Big Data.

Chi Zhou: IEEE Transactions on Big Data, IEEE Transactions on Cloud Computing and ACM TAAS.

9.1.4. Keynote Talks and Invited Talks

Gabriel Antoniu:

BigStorage and WALL ITN Joint Meeting: *Týr: Storage-based Convergence Between HPC and Big Data*, Mainz, January 2017.

Huawei Seminar: *Convergence of HPC and Big Data*, Huawei European Research Center, Munich, October 2017.

Luc Bougé:

Dagstuhl School on Challenges and Opportunities of User-Level File Systems for HPC: *Are objects the right level of abstraction to enable the convergence between HPC and Big Data at storage level?*, joint talk with María Pérez, Universidad Politécnica de Madrid (slides).

ENS Rennes: *Map-Reduce: Very-Large Scale Programming for Big Data on Clouds*, November 2017.

Alexandru Costan:

Huawei Seminar: *Low-Latency Stream Storage*, Huawei Research Munich, October 2017.

BigStorage Summer School: *Making Cities Smarter: A Storage-based View on Stream Processing Engines*, Heraklion, July 2017.

UPB Scientific Days: *Science Driven, Scalable Data-Intensive Processing on Clouds*, University Politehnica of Bucharest, June 2017.

9.1.5. Leadership within the Scientific Community

Gabriel Antoniu:

- Scientific leader of the KerData project-team.
- Topic leader for Inria for the *Data storage, I/O and in situ processing* topic, supervising collaboration activities in this area within the JLESC, Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing.
- Co-leader the BDEC working group dedicated to exploring convergence-related challenges for hybrid architectures combining HPC systems, clouds and fog/edge computing infrastructures with Geoffrey Fox and Ewa Deelman.
- Work package leader within the BigStorage H2020 ETN project for the *Data Science* work package.

Luc Bougé: Vice-President of the *French Society for Informatics* (SIF), in charge of the Teaching department.

Alexandru Costan: Leader of the *Smart Cities* Working Group within the BigStorage H2020 ETN project.

9.1.6. Scientific Expertise

Gabriel Antoniu served as a project evaluator for the project proposals submitted within the Activity "Post-doctoral Research Support" provided from the European Regional Development Fund to the The State Education Development Agency (SEDA) of the Republic of Latvia.

Luc Bougé: Member of the jury for the *Agrégation de mathématiques* and the *CAPES of mathématiques*. These national committees select permanent mathematics teachers for secondary schools and high-schools, respectively.

9.1.7. Research Administration

Gabriel Antoniu: Vice Executive Director of JLESC, Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing for Inria.

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 : Cloud Computing, 15 hours (lectures and lab sessions), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz, France.
- Master: Scalable Distributed Systems, 12 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School, France.
- Master: Infrastructures for Big Data, 12 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.
- Master: Big Data Processing, 6 hours (lectures), M2 level, Health Big Data Master, University of Rennes, Faculty of Medecine, France.

Luc Bougé

- Bachelor: Introduction to programming concepts, 36 hours (lectures), L3 level, Informatics program, ENS Rennes, France.
- Master: Introduction to compilation, 24 hours (exercice and practical classes), M1 level, Informatics program, Univ. Rennes I, 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, 28h hours (lectures, lab sessions), M2, INSA Rennes.
- Master: Big Data Project, 28h hours (project), M2, INSA Rennes.

9.2.2. Supervision

Luis Eduardo Pineda Morales: *Efficient Big Data Management for Geographically Distributed Workflows*, INSA Rennes, defended on May 24, 2017. Co-advised by Alexandru Costan and Gabriel Antoniu. Manuscript: https://tel.archives-ouvertes.fr/tel-01645434.

Tien Dat Phan: *Energy-efficient Straggler Mitigation for Big Data Applications on the Clouds*, ENS Rennes, defended on November 30, 2017. Co-advised by Shadi Ibrahim and Luc Bougé.

Orçun Yildiz: *Efficient Big Data Processing on Large Scale Shared Platforms: Managing I/Os and Failures*, ENS Rennes, defended on December 8, 2017. Co-advised by Shadi Ibrahim and Gabriel Antoniu.

9.2.2.1. PhD in progress

Pierre Matri: *Predictive Models for Big Data*, thesis started in March 2015, co-advised by María Pérez (Universidad Politécnica de Madrid) and Gabriel Antoniu.

Mohammed-Yacine Taleb: *Energy-impact of data consistency management in Clouds and Beyond*, thesis started in August 2015, co-advised by Gabriel Antoniu and Toni Cortés (Barcelona Supercomputing Center).

Ovidiu-Cristian Marcu: *Efficient data transfer and streaming strategies for workflow-based Big Data processing*, thesis started in October 2015, co-advised by Alexandru Costan and Gabriel Antoniu.

Nathanaël Cheriere: *Resource Management and Scheduling for Big Data Applications in Large-scale Systems*, thesis started in September 2016, co-advised by Gabriel Antoniu and Matthieu Dorier.

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

9.2.3. Juries

Gabriel Antoniu: Referee for the PhD thesis of Jonathan Martí at the Barcelona Supercomputing Center (March 2017).

9.2.4. Miscellaneous

9.2.4.1. Responsibilities

Gabriel Antoniu: in charge of the IBD module of the SIF Master of Rennes and of the Cloud module of the MIAGE Master of the University of Rennes 1.

Luc Bougé: Co-ordinator between ENS Rennes and the Inria Research Center and the IRISA laboratory.

Luc Bougé: In charge of the Bachelor level (L3) and of the student seminar series at the Informatics Departement of ENS Rennes.

Alexandru Costan: In charge of communication at the Computer Science Department of INSA Rennes.

Alexandru Costan: In charge of the organization of the IRISA D1 Department Seminars.

9.2.4.2. Tutorials

Gabriel Antoniu gave a tutorial on in situ processing at the BigStorage Summer School in Heraklion (July 2017).

9.3. Popularization

Gabriel Antoniu:

ORAP Forum, Paris: ORAP is a French national forum intended for all stakeholders in the HPC area (industry and academia). Invited talk: *Convergence of HPC and Big Data: a storage-oriented perspective* in March 2017.

Available resources: Video and Slides.

Plenary Meeting of Inria's Department for International and European Relations: *Convergence of HPC and Big Data: Where Are We Going?*, Inria, Rocquencourt, October 2017.

Luc Bougé:

Doctoral Program, Rennes. Invited presentation to the PhD students about *Preparing your applications after your PhD* (November 2017).

Master Program, Rennes. Invited presentation to the M2 students about *Informatics as a scientific activity: Toward a responsible research* (November 2017).

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] T.-D. PHAN. *Energy-efficient Straggler Mitigation for Big Data Applications on the Clouds*, ENS Rennes, November 2017, https://tel.archives-ouvertes.fr/tel-01669469

[15] L. E. PINEDA MORALES. *Efficient support for data-intensive scientific workflows on geo-distributed clouds*, INSA de Rennes, May 2017, https://tel.archives-ouvertes.fr/tel-01645434

[16] O. YILDIZ. *Efficient Big Data Processing on Large-Scale Shared Platforms: Managing I/Os and Failures*, ENS Rennes, December 2017, https://tel.archives-ouvertes.fr/tel-01671413

Articles in International Peer-Reviewed Journals

[17] P. MATRI, M. S. PÉREZ, A. COSTAN, L. BOUGÉ, G. ANTONIU. *Keeping up with storage: Decentralized, write-enabled dynamic geo-replication*, in "Future Generation Computer Systems", June 2017, pp. 1-19 [*DOI :* 10.1016/J.FUTURE.2017.06.009], https://hal.inria.fr/hal-01617658

International Conferences with Proceedings

[18] N. CHERIERE, G. ANTONIU. *How Fast Can One Scale Down a Distributed File System?*, in "BigData 2017 - IEEE International Conference on Big Data", Boston, United States, December 2017, https://hal.archives-ouvertes.fr/hal-01644928

[19] J. LIU, L. PINEDA-MORALES, E. PACITTI, A. COSTAN, P. VALDURIEZ, G. ANTONIU, M. MATTOSO. *Efficient Scheduling of Scientific Workflows using Hot Metadata in a Multisite Cloud*, in "BDA: Conférence sur la Gestion de Données — Principes, Technologies et Applications", Nancy, France, November 2017, 13 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620231

[20] O.-C. MARCU, A. COSTAN, G. ANTONIU, M. S. PÉREZ-HERNÁNDEZ, R. TUDORAN, S. BORTOLI, B. NICOLAE. *Towards a Unified Storage and Ingestion Architecture for Stream Processing*, in "Second Workshop on Real-time & Stream Analytics in Big Data Colocates with the 2017 IEEE International Conference on Big Data", Boston, United States, IEEE, December 2017, pp. 1-6, https://hal.inria.fr/hal-01649207

[21] O.-C. MARCU, R. TUDORAN, B. NICOLAE, A. COSTAN, G. ANTONIU, M. S. PÉREZ-HERNÁNDEZ. *Exploring Shared State in Key-Value Store for Window-Based Multi-Pattern Streaming Analytics*, in "Workshop on the Integration of Extreme Scale Computing and Big Data Management and Analytics in conjunction with IEEE/ACM CCGrid 2017", Madrid, Spain, May 2017, https://hal.inria.fr/hal-01530744

[22] P. MATRI, Y. ALFOROV, A. BRANDON, M. KUHN, P. CARNS, 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, September 2017, pp. 81 - 86 [*DOI :* 10.1109/CLUSTER.2017.63], https://hal.inria.fr/hal-01617655

[23] T.-D. PHAN, S. IBRAHIM, A. C. ZHOU, G. AUPY, G. ANTONIU. *Energy-Driven Straggler Mitigation in MapReduce*, in "Euro-Par'17 - 23rd International European Conference on Parallel and Distributed Computing", Santiago de Compostela, Spain, August 2017, https://hal.inria.fr/hal-01560044

[24] Y. TALEB, S. IBRAHIM, G. ANTONIU, T. CORTES. *An Empirical Evaluation of How The Network Impacts The Performance and Energy Efficiency in RAMCloud*, in "Workshop on the Integration of Extreme Scale Computing and Big Data Management and Analytics in conjunction with IEEE/ACM CCGrid 2017", Madrid, Spain, May 2017, https://hal.inria.fr/hal-01376923

[25] Y. TALEB, S. IBRAHIM, G. ANTONIU, T. CORTES. *Characterizing Performance and Energy-Efficiency of The RAMCloud Storage System*, in "The 37th IEEE International Conference on Distributed Computing Systems (ICDCS 2017)", Atlanta, United States, June 2017, https://hal.inria.fr/hal-01496959

[26] O. YILDIZ, A. C. ZHOU, 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", Hawaii, United States, September 2017, https://hal.inria.fr/hal-01570737

Research Reports

[27] O.-C. MARCU, A. COSTAN, G. ANTONIU, M. S. PÉREZ-HERNÁNDEZ. *Kera: A Unified Storage and Ingestion Architecture for Efficient Stream Processing*, Inria Rennes - Bretagne Atlantique, June 2017, n° RR-9074, https://hal.inria.fr/hal-01532070

Other Publications

[28] P. LE NOAC'H, A. COSTAN, L. BOUGÉ. *A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications*, December 2017, IEEE Big Data 2017, Poster, https://hal.archives-ouvertes.fr/hal-01647229

[29] Y. TALEB, R. STUTSMAN, G. ANTONIU, T. CORTES. *Tailwind: fast and atomic RDMA-based replication*, January 2018, working paper or preprint, https://hal.inria.fr/hal-01676502

References in notes

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

[31] *The Decaf Project*, 2017, https://bitbucket.org/tpeterka1/decaf

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

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

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

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

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

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

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

[39] *Inria's strategic plan "Towards Inria 2020"*, 2016, https://www.inria.fr/en/institute/strategy/strategic-plan

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

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

[42] *Storm*, 2014, http://storm-project.net/

[43] *The FlowVR Project*, 2014, http://flowvr.sourceforge.net/

[44] 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

[45] 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

[46] S. WILDE, M. HATEGAN, J. M. WOZNIAK, B. CLIFFORD, D. KATZ, I. T. FOSTER. *Swift: A language for distributed parallel scripting*, in "Parallel Computing", 2011, vol. 37, n^{o} 9, pp. 633–652, http://dx.doi.org/10.1016/j.parco.2011.05.005