IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

École normale supérieure de Rennes

Activity Report 2015

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:
1.1.4. - High performance computing
1.1.6. - Cloud
1.3. - Distributed Systems
1.6. - Green Computing
2.6.2. - Middleware
3.1.3. - Distributed data
3.1.8. - Big data (production, storage, transfer)
3.3.3. - Big data analysis
6.2.7. - High performance computing
7.1. - Parallel and distributed algorithms

Other Research Topics and Application Domains:
1.1.2. - Molecular biology
2.6.1. - Brain imaging
3.2. - Climate and meteorology
4.4.1. - Green computing

1. Members

Research Scientists
Gabriel Antoniu [Team leader, Inria, Senior Researcher, HdR]
Shadi Ibrahim [Inria, Researcher]

Faculty Members
Luc Bougé [ENS Rennes, Professor, HdR]
Alexandru Costan [INSA Rennes, Associate Professor]

Engineer
Loïc Cloatre [Inria, until January 2015]

PhD Students
Luis Eduardo Pineda Morales [Microsoft Research Inria Joint Centre]
Lokman Rahmani [Univ. Rennes I]
Orçun Yildiz [Inria]
Tien Dat Phan [Univ. Rennes I]
Pierre Matri [Universidad Politécnica de Madrid, from March 2015]
Ovidiu-Cristian Marcu [Inria, from October 2015]
Mohammed-Yacine Taleb [Inria, from August 2015]

Visiting Scientist
Antonio Aguilera [Universidad de Granada, Invited PhD student, from April 2015 until June 2015]

Administrative Assistant
Aurélie Patier [Univ. Rennes I]

Others

Roxana-Ioana Roman [Inria, Master intern, from May 2015 until July 2015]
Nathanaël Cheriere [ENS Rennes, Master intern, from February 2015 until July 2015]

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 these fields, it becomes crucial to efficiently store and manipulate massive data, which are typically *shared* at a large scale and *concurrently accessed*. In all these examples, the overall application performance is highly dependent on the properties of the underlying data management service. With the emergence of recent infrastructures such as cloud computing platforms and post-petascale architectures, achieving highly scalable data management has become a critical challenge.

The KerData project-team is namely focusing on *scalable data storage and processing on clouds and post-petascale HPC 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 and Exascale HPC research communities.

2.2. Objective: efficient support for scalable data-intensive computing

Our research activities focus on the data storage and processing needs of data-intensive applications that that exhibit the need to handle:

- Massive data BLOBs (Binary Large OBjects), in the order of Terabytes, stored in a large number of nodes (thousands to tens of thousands), accessed under heavy concurrency by a large number of processes (thousands to tens of thousands at a time), with a relatively fine access grain, in the order of Megabytes;
- Very large sets (millions) of small objects potentially arriving in streams, stored and processed on geographically distributed infrastructures (e.g. multi-site clouds);
- Very large sets of scientific data processed on extreme-scale supercomputers.

Examples of such applications are:

- Massively parallel data analytics for Big Data applications (e.g., Map-Reduce-based data analysis as currently enabled by frameworks such as Ha doop, Spark or Flink);
- Advanced cloud services for data storage and transfer for geographically distributed workflows requiring efficient data sharing within and across multiple datacenters;
- Scalable solutions for I/O management and in situ visualization for data-intensive scientific simulations (e.g. atmospheric simulations, computational fluid dynamics, etc.) running on Extreme-Scale HPC systems.

3. Research Program

3.1. Our goals and methodology

Data-intensive applications demonstrate common requirements with respect to the need for data storage and I/O processing. These requirements lead to several core challenges discussed below.

Challenges related to cloud storage. In the area of cloud data management, a significant milestone is the emergence of the Map-Reduce [33] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [29]. At the core of Map-Reduce frameworks lies teh storage system, a key component which must meet a series of specific requirements that have not fully been 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, as data-intensive processing needs go beyond the frontiers of single datacenters, extra challenges related to the efficiency of metadata management concern the storage and efficient access to very large sets of small objects by Big Data processing workflows running on large-scale infrastructures.

Challenges related to data-intensive HPC applications. The requirements exhibited by climate simulations specifically highlight a major, more general research topic. They have been clearly identified by international panels of experts like IESP [32], EESI [30], ETP4HPC [31] in the context of HPC simulations running on post-petascale supercomputers. 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 such as climate modeling, solid earth sciences or astrophysics. In this context, the lack of data-intensive infrastructures and methodologies to analyze huge simulations is a growing limiting factor. The challenge is to find new ways to store, visualize and analyze massive outputs of data during and after the simulation without impacting the overall performance (i.e. while avoiding as much as possible the *jitter* generated by I/O interference). In this area, we specifically focus on *in situ processing* approaches and we explore approaches to *model and predict I/O* and to *reduce intra-application and cross-application I/O interference*.

The overall goal of the KerData project-team is to bring a substantial contribution to the effort of the research communities in the areas of cloud computing and HPC to address the above challenges. KerData's approach consists in designing and implementing distributed algorithms for scalable data storage and input/output management for efficient large-scale data processing. We target two main execution infrastructures: cloud platforms and post-petascale HPC supercomputers. Our collaboration porfolio includes international teams that are active in this areas both in Academia (e.g., Argonne National Lab, University of Illinois at Urbana-Champaign, Barcelona Supercomputing Centre) and Industry (Microsoft, IBM).

The highly experimental nature of our research validation methodology should be stressed. Our approach relies on building prototypes and on validating 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 (Microsoft Azure, Amazon clouds, Nimbus clouds); in the post-petascale HPC area we are running our experiments on 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.

Moreover, the consortiums of our current projects include application partners in the areas of Bio-Chemistry, Neurology and Genetics, and Climate Simulations. This is an additional asset, it enables us to take into account application requirements in the early design phase of our solutions, and to validate those solutions with real applications. We intend to continue increasing our collaborations with application communities, as we believe that this a key to perform effective research with a high impact.

3.2. Our research agenda

Three examples of motivating application scenarios will be described in detail in the next section:

- Joint genetic and neuroimaging data analysis on Azure clouds;
- Structural protein analysis on Nimbus clouds;
- I/O-intensive atmospheric simulations for the Blue Waters post-petascale machine.

They illustrate the above challenges in some specific ways. They all exhibit a common scheme: massively concurrent processes which access massive data at a fine granularity, where data is shared and distributed at a large scale. To address the aforementioned challenges efficiently, we have are exploring two main approaches:

- the BlobSeer approach, which stands at the center of some of our main research efforts in the area of cloud storage for Big Data processing. This approach relies on the design and implementation of *scalable* distributed algorithms for data storage and access. They combine advanced techniques for decentralized metadata and data management, with versioning-based concurrency control to optimize the performance of applications under heavy access concurrency.
- the Damaris approach (that is totally independent of BlobSeer), which exploits multicore parallelism in post-petascale supercomputers to enable jitter-free, low-overhead I/O management and non intrusive in situ visualization for large-scale simulations.

Our short- and medium-term research plan is devoted to storage challenges in two main contexts: clouds and post-petascale HPC architectures. Consequently, our research plan is split in two main themes, which correspond to their respective challenges. For each of those themes, we have initiated several actions through collaborative projects coordinated by KerData, which define our current research agenda.

Based on very promising results demonstrated by BlobSeer in preliminary experiments [34], we have initiated several collaborative projects in the area of cloud data management, e.g., the MapReduce ANR project (aiming to improve both the performance and the fault-tolerance of the storage component of MapReduce processing frameworks to better support highly-concurrent data analytics applications); the A-Brain Microsoft-Inria project (that leverages these improvements on Microsoft Azure clouds to the benefit of joint neuroimaging and genetics analysis); the Z-CloudFlow Microsoft-Inria project (exploring how to efficiently manage metadata for geographically-distributed workflows). Such frameworks are for us concrete and efficient means to work in close connection with strong partners already well positioned in the area of cloud computing research.

Similarly, Damaris is the fruit of a very successful collaborative work within the Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing (JLESC, formerly called JLPC). It has become a reference framework illustrating the usage of a dedicated-core approach for scalable I/O and non-intrusive in situ visualization on post-petascale HPC systems. It led to the creation of the particularly active Data@Exascale Associate Team between Inria, ANL and UIUC, an excellent framework for an enlarged research activity involving a large number of young researchers and students of the KerData team and of its partners. This Associate Team serves as a basis for extended research activities based on our approaches (including Damaris and Omnisc'IO), carried out beyond the frontiers of our team. Our team is playing a leading role in the Big Data and I/O research activities in the JLESC lab. This joint lab facilitates high-quality collaborations and access to some of the most powerful supercomputers, an important asset which already helped us produce and transfer some results of our team (e.g. Damaris).

Thanks to these projects, we are now enjoying a visible scientific positioning at the international level.

4. Application Domains

4.1. Joint genetic and neuroimaging data analysis on Azure clouds

Joint acquisition of neuroimaging and genetic data on large cohorts of subjects is a new approach used to assess and understand the variability that exists between individuals. Both neuroimaging- and genetic-domain observations include a huge amount of variables (of the order of millions). Performing rigorous statistical analyses on such amounts of data is a major computational challenge that cannot be addressed with conventional computational techniques only. On the one hand, sophisticated regression techniques need to be used in order to perform significant analysis on these large datasets; on the other hand, the cost entailed by parameter optimization and statistical validation procedures (e.g. permutation tests) is very high.

To address the above challenges, the A-Brain (AzureBrain) Project was carried out within the Microsoft Research-Inria Joint Research Center. It was co-led by the KerData (Rennes) and Parietal (Saclay) Inria teams. They jointly address this computational problem using cloud related techniques on the Microsoft Azure cloud infrastructure. The two teams brought together their complementary expertise: KerData in the area of scalable cloud data management, and Parietal in the field of neuroimaging and genetics data analysis.

This application scenario is a typical multi-disciplinary Data Science project which serves as background for several on-going research activities, beyond the end of the A-Brain project.

4.2. Structural protein analysis on Nimbus and IBM clouds

In the framework of the MapReduce ANR project led by KerData (2010-2014), we have focused on the FastA bioinformatics application used for massive protein sequence similarity searching. This is a typical data-intensive application that can leverage the Map-Reduce model for a scalable execution on large-scale distributed platforms. FastA remains an interesting use case that we are considering beyond the end of the MapReduce project, for benchmarking our research results in the the area of optimized MapReduce processing.

4.3. I/O intensive climate simulations for the Blue Waters post-petascale machine

A major research topic in the context of HPC simulations running on post-petascale supercomputers is to explore how to record and visualize data during the simulation efficiently without impacting the performance of the computation generating that data. Conventional practice consists in storing data on disk, moving them off-site, reading them into a workflow, and analyzing them. This approach becomes increasingly harder to use because of the large data volumes generated at fast rates, in contrast to limited back-end performance. Scalable approaches to deal with these I/O limitations are thus of utmost importance. This is one of the main challenges explicitly stated in the roadmap of the Blue Waters Project, which aims to build one of the most powerful supercomputers in the world.

In this context, the KerData project-team is exploring innovative ways to remove the limitations mentioned above through collaborative work in the framework of the Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing (JLESC, formerly called JLPC), whose research activity focuses on the Blue Waters project. An example is the atmospheric simulation code CM1 (Cloud Model 1), one of the target applications of the Blue Waters machine. State-of-the-art I/O approaches, which typically consist in periodically writing a very large number of small files are inefficient: they cause bursts of I/O in the parallel file system, leading to poor performance and extreme variability (*jitter*). The challenge here is to investigate how to make an efficient use of the underlying file system, by avoiding synchronization and contention as much as possible. In collaboration with the JLESC, we are addressing these challenges through the Damaris approach.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

Gilles Kahn honorary award of the SIF and the Academy of Science: 2nd prize for Matthieu Dorier in 2015. The Gilles Kahn Honorary Award is given every year to at most the 3 best PhD theses in Computer Science in France and is jointly delivered by the *Société Informatique de France* (SIF) and the French Academy of Science. The candidates are judged on all aspects of their PhD work, including fundamental contributions to industrial transfers, publication impact, teaching, mentoring, and scientific dissemination activities. A Grand Prize and two *ex aequo* Accessit Prizes are given. Matthieu Dorier was given one of the latter.

PhD award of the Fondation Rennes 1: 2nd prize for Matthieu Dorier in the Matisse Doctoral School in 2015. The Rennes 1 Foundation PhD award from the Fondation Rennes 1 is given every year to 8 outstanding new doctors from the 4 doctoral schools associated with the University of Rennes 1 (2 awards per doctoral school). The candidates are judged on the innovative aspects of their PhD thesis, "innovative" being understood in the sense of impact on socioeconomic development and technology transfers.

5.1.2. 5 International Journals

This year the team published 5 papers in high-quality journals including IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Cloud Computing, Future Generation Computer Systems (2), World Wide Web.

6. New Software and Platforms

6.1. Major Software

6.1.1. BlobSeer

Participants: Alexandru Costan, Gabriel Antoniu, Luc Bougé, Loïc Cloatre.

Contact: Gabriel Antoniu.

Presentation: BlobSeer is the core software platform for many current cloud-oriented projects of the KerData team. It is a data storage service specifically designed to deal with the requirements of large-scale, data-intensive distributed applications that abstract data as huge sequences of bytes, called BLOBs (Binary Large OBjects). It provides a versatile versioning interface for manipulating BLOBs that enables reading, writing and appending to them.

BlobSeer offers both scalability and performance with respect to a series of issues typically associated with the data-intensive context: *scalable aggregation of storage space* from the participating nodes with minimal overhead, ability to store *huge data objects*, *efficient fine-grain access* to data subsets, *high throughput in spite of heavy access concurrency*, as well as *fault-tolerance*. This year we have mainly focused on the deployment in production of the BlobSeer software on IBM's cluster at Montpellier, in the context of the ANR MapReduce project. To this end, several bugs were solved, and several optimizations were brought to the communication layer of BlobSeer. To showcase the benefits of BlobSeer on this platform we focused on the Terasort benchmark. Currently, preliminary tests on Grid5000 with this benchmark show that BlobSeer performs better than HDFS for block sizes lower than 2 MB. We have also improved the continuous integration process of BlobSeer by deploying daily builds and automatic tests on Grid5000.

Users: Work is currently in progress in several formalized projects (see previous section) to integrate and leverage BlobSeer as a data storage back-end in the reference cloud environments: a) Microsoft Azure; b) the Nimbus cloud toolkit developed at Argonne National Lab (USA); and c) the OpenNebula IaaS cloud toolkit developed at UCM (Madrid).

URL: http://blobseer.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 3.

Status: This software is available on Inria's forge. Version 1.0 (released late 2010) registered with APP: IDDN.FR.001.310009.000.S.P.000.10700.

A *Technology Research Action* (ADT, *Action de recherche technologique*) was active for two years until January 2015, aiming to robustify the BlobSeer software and to make it a safely distributable product. This project is funded by Inria *Technological Development Office* (D2T, *Direction du Développement Technologique*).

6.1.2. Damaris

Participants: Matthieu Dorier, Gabriel Antoniu, Orçun Yildiz, Lokman Rahmani, Shadi Ibrahim.

Contact: Gabriel Antoniu.

Presentation: Damaris is a middleware for multicore SMP nodes enabling them to handle data transfers for storage and visualization efficiently. 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). Damaris 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. Version 1.0 was released in November 2014 and enables other approaches such as the use of dedicated nodes instead of dedicated cores.

Users: Damaris has been preliminarily evaluated at NCSA/UIUC (Urbana-Champaign, IL, USA) with the CM1 tornado simulation code. CM1 is one of the target applications of the Blue Waters supercomputer in production at, 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 four large-scale simulations (CM1, OLAM, Nek5000, GTC).

URL: http://damaris.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 3.

Status: This software is available on Inria's forge and registered with APP. Registration of the latest version with APP is in progress.

6.2. Other Software

6.2.1. Omnisc'IO

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

Contact: Matthieu Dorier

Presentation: Omnisc'IO is a middleware integrated in the POSIX and MPI-I/O stacks to observe, model and predict the I/O behavior of any HPC application transparently. It is based on formal grammars, implementing a modified version of the Sequitur algorithm. Omnisc'IO has been used on Grid'5000 with the CM1 atmospheric simulation, the LAMMPS molecular dynamics simulation, the GTC fusion simulation and the Nek5000 CFD simulation. Omnisc'IO was subject to a publication at SC14.

Users: Omnisc'IO is currently used only within the KerData team and at Argonne National Lab.

URL: http://omniscio.gforge.inria.fr/

License: GNU Lesser General Public License (LGPL) version 3.

Status: Currently unavailable for distribution (subject to major changes). Version 1.0 (released in November 2015) registered with APP: IDDN.FR.001.540003.000.S.P.2015.000.10000.

6.2.2. JetStream

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Contact: Alexandru Costan

Presentation: 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.

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

License: Microsoft Public License.

Status: Prototype and demo available.

6.2.3. OverFlow

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Contact: Alexandru Costan.

Presentation: 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. In 2015, OverFlow was extended with support for efficient metadata operations: the newly implemented strategies leverage workflow semantics in a 2-level metadata partitioning hierarchy that combines distribution and replication.

Users: 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.

License: GNU Lesser General Public License (LGPL) version 3.

Status: Registration of the latest version with APP is in progress

6.2.4. iHadoop

Participants: Tien Dat Phan, Shadi Ibrahim.

Contact: Shadi Ibrahim

Presentation: *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. In 2015, iHadoop was extended to simulate the execution and predict the performance of multiple Map-Reduce applications, sharing the same Hadoop cluster. Two schedulers (Fifo, Fair) are now available in iHadoop.

Users: 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. iHadoop was discussed with the SimGrid community during the SimGrid user days in Lyon (June 2015). We intend to add iHadoop to the contributions site of the SimGrid project and make it available to the SimGrid community.

License: GNU Lesser General Public License (LGPL) version 3.

Status: Available on Inria's forge. Registration of the latest version with APP is in progress.

7. New Results

7.1. Efficient data management for hybrid and multi-site clouds

7.1.1. JetStream: enabling high-throughput live event streaming on multi-site clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific and commercial applications operate nowadays on tens of cloud datacenters around the globe, following similar patterns: they aggregate monitoring or sensor data, assess the QoS or run global data mining queries based on inter-site event stream processing. Enabling fast data transfers across geographically distributed sites allows such applications to manage the continuous streams of events in real time and quickly react to changes. However, traditional event processing engines often consider data resources as second-class citizens and support access to data only as a side-effect of computation (i.e. they are not concerned by the transfer of events from their source to the processing site). This is an efficient approach as long as the processing is executed in a single cluster where nodes are interconnected by low latency networks. In a distributed environment, consisting of multiple datacenters, with orders of magnitude differences in capabilities and connected by a WAN, this will undoubtedly lead to significant latency and performance variations.

This is namely the challenge we addressed this year by proposing JetStream [15], a high performance batch-based streaming middleware for efficient transfers of events between cloud datacenters. JetStream is able to self-adapt to the streaming conditions by modeling and monitoring a set of context parameters. It further aggregates the available bandwidth by enabling multi-route streaming across cloud sites, while at the same time optimizing resource utilization and increasing cost efficiency. The prototype was validated on tens of nodes from US and Europe datacenters of the Windows Azure cloud with synthetic benchmarks and a real-life application monitoring the ALICE experiment at CERN. The results show a 3× increase of the transfer rate using the adaptive multi-route streaming, compared to state of the art solutions.

7.1.2. Multi-site metadata management for geographically distributed cloud workflows

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu.

With their globally distributed datacenters, clouds now provide an opportunity to run complex large-scale applications on dynamically provisioned, networked and federated infrastructures. However, there is a lack of tools supporting data-intensive applications (e.g. scientific workflows) on virtualized IaaS or PaaS systems across geographically distributed sites. As a relevant example, data-intensive scientific workflows struggle in leveraging such distributed cloud platforms. For instance, scientific workflows which handle many small files can easily saturate state-of-the-art distributed filesystems based on centralized metadata servers (e.g., HDFS, PVFS).

In [22], we explore several alternative design strategies to efficiently support the execution of existing workflow engines across multi-site clouds, by reducing the cost of metadata operations. These strategies leverage workflow semantics in a 2-level metadata partitioning hierarchy that combines distribution and replication. The system was validated on the Microsoft Azure cloud across 4 EU and US datacenters. The experiments were conducted on 128 nodes using synthetic benchmarks and real-life applications. We observe as much as 28% gain in execution time for a parallel, geo-distributed real-world application (Montage) and up to 50% for a metadata-intensive synthetic benchmark, compared to a baseline centralized configuration.

7.1.3. Understanding the performance of Big Data platforms in hybrid and multi-site clouds

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

Recently, hybrid multi-site big data analytics (that combines on-premise with off-premise resources) has gained increasing popularity as a tool to process large amounts of data on-demand, without additional capital investment to increase the size of a single datacenter. However, making the most out of hybrid setups for big data analytics is challenging because on-premise resources can communicate with off-premise resources at significantly lower throughput and higher latency. Understanding the impact of this aspect is not trivial, especially in the context of modern big data analytics frameworks that introduce complex communication patterns and are optimized to overlap communication with computation in order to hide data transfer latencies. This year we started to work on a study that aims to identify and explain this impact in relationship to the known behavior on a single cloud.

A first step towards this goal consisted of analysing a representative big data workload on a hybrid Spark setup [24]. Unlike previous experience that emphasized low end-impact of network communications in Spark, we found significant overhead in the shuffle phase when the bandwidth between the on-premise and off-premise resources is sufficiently small. We plan to continue this study by investigating additional parameters at a finer grain and adding new platforms, like Apache Flink.

7.2. Optimizing Map-Reduce

7.2.1. Chronos: failure-aware scheduling in shared Hadoop clusters

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop emerged as the de facto state-of-the-art system for MapReduce-based data analytics. The reliability of Hadoop systems depends in part on how well they handle failures. Currently, Hadoop handles machine failures by re-executing all the tasks of the failed machines (i.e., executing recovery tasks). Unfortunately, this elegant solution is entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. The unawareness of failures therefore may prevent Hadoop schedulers from operating correctly towards meeting their objectives (e.g., fairness, job priority) and can significantly impact the performance of MapReduce applications.

In [23], we propose Chronos, a failure-aware scheduling strategy that enables an early yet smart action for fast failure recovery while operating within a specific scheduler objective. Chronos takes an early action rather than waiting an uncertain amount of time to get a free slot (thanks to our preemption technique). Chronos embraces a smart selection algorithm that returns a list of tasks that need to be preempted in order to free the necessary slots to launch recovery tasks immediately. This selection considers three criteria: the progress scores of running tasks, the scheduling objectives, and the recovery tasks input data locations. In order to make room for recovery tasks rather than waiting an uncertain amount of time, a natural solution is to kill running tasks in order to create free slots. Although killing tasks can free the slots easily, it wastes the work performed by the killed tasks. Therefore, we present the design and implementation of a novel work-conserving preemption technique that allows pausing and resuming both map and reduce tasks without resource wasting and with little overhead.

We demonstrate the utility of Chronos by combining it with two state-of-the-art Hadoop schedulers: Fifo and Fair schedulers. The experimental results show that Chronos achieves almost optimal data locality for the recovery tasks and reduces the job completion times by up to 55% over state-of-the-art schedulers. Moreover, Chronos recovers to a correct scheduling behavior after failure detection within only a couple of seconds.

7.2.2. On the usability of shortest remaining time first policy in shared Hadoop clusters

Participants: Nathanaël Cheriere, Shadi Ibrahim.

A practical problem facing the Hadoop community is how to reduce job makespans by reducing job waiting times and execution times. Previous Hadoop schedulers have focused on improving job execution times, by improving data locality but not considering job waiting times. Even worse, enforcing data locality according to the job input sizes can be inefficient: it can lead to long waiting times for small yet short jobs when sharing the cluster with jobs with smaller input sizes but higher execution complexity.

We have introduced hSRTF [16], an adaption of the well-known Shortest Remaining Time First scheduler (i.e., SRTF) in shared Hadoop clusters. hSRTF embraces a simple model to estimate the remaining time of a job and a preemption primitive (i.e., kill) to free the resources when needed. We have implemented hSRTF and performed extensive evaluations with Hadoop on the Grid'5000 testbed. The results show that hSRTF can significantly reduce the waiting times of small jobs and therefore improves their make-spans, but at the cost of a relatively small increase in the make-spans of large jobs. For instance, a time-based proportional share mode of hSRTF (i.e., hSRTF-Pr) speeds up small jobs by (on average) 45% and 26% while introducing a performance degradation for large jobs by (on average) 10% and 0.2% compared to Fifo and Fair schedulers, respectively.

7.2.3. A Performance evaluation of Hadoop's schedulers under failures

Participants: Shadi Ibrahim, Gabriel Antoniu.

Recently, Hadoop has not only been used for running single batch jobs but it has also been optimized to simultaneously support the execution of multiple jobs belonging to multiple concurrent users. Several schedulers (i.e., Fifo, Fair, and Capacity schedulers) have been proposed to optimize locality executions of tasks but do not consider failures, although, evidence in the literature shows that faults do occur and can probably result in performance problems.

In [19], we have designed a set of experiments to evaluate the performance of Hadoop under failure when applying several schedulers (i.e., explore the conflict between job scheduling, exposing locality executions, and failures). Our results reveal several drawbacks of current Hadoop's mechanism in prioritizing failed tasks. By trying to launch failed tasks as soon as possible regardless of locality, it significantly increases the execution time of jobs with failed tasks, due to two reasons: 1) available resources might not be freed up as quickly as expected and 2) failed tasks might be re-executed on machines with no data on it, introducing extra cost for data transferring through network, which is normally the most scarce resource in today's datacenters.

Our preliminary study with Hadoop not only helps us to understand the interplay between fault-tolerance and job scheduling, but also offers useful insights into optimizing the current schedulers to be more efficient in case of failures.

7.2.4. Kvasir: empowering Hadoop with knowledge

Participants: Nathanaël Cheriere, Shadi Ibrahim.

Most of Hadoop schedulers are based on homogeneity hypotheses about the jobs and the nodes and therefore strongly rely on the location of the input data when scheduling tasks. However, our study revealed that Hadoop is a highly dynamic environment (e.g., variation in task duration within a job and across different jobs). Even worse, clouds are multi-tenant environments which in turn introduce more heterogeneity and dynamicity in Hadoop clusters. As a result, relying on static knowledge (i.e. data location) may lead to wrong scheduling decisions.

We have developed a new scheduling framework for Hadoop, named Kvasir. Kvasir aims to provide an up-to-date knowledge that reflects the dynamicity of the environment while being light-weight and performance-oriented. The utility of Kvasir is demonstrated by the implementation of several schedulers including Fifo, Fair, and SRTF schedulers.

7.3. Energy-aware data management in clouds and HPC

7.3.1. On understanding the energy impact of speculative execution in Hadoop

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé.

Hadoop emerged as an important system for large-scale data analysis. Speculative execution is a key feature in Hadoop that is extensively leveraged in clouds: it is used to mask slow tasks (i.e., stragglers) — resulted from resource contention and heterogeneity in clouds — by launching speculative task copies on other machines. However, speculative execution is not cost-free and may result in performance degradation and extra resource and energy consumption. While prior literature has been dedicated to improving stragglers detection to cope with the inevitable heterogeneity in clouds, little work is focusing on understanding the implications of speculative execution on the performance and energy consumption in Hadoop cluster.

In [21], we have designed a set of experiments to evaluate the impact of speculative execution on the performance and energy consumption of Hadoop in homogeneous and heterogeneous environments. Our studies reveal that speculative execution may sometimes reduce, sometimes increase the energy consumption of Hadoop clusters. This strongly depends on the reduction in the execution time of MapReduce applications and on the extra power consumption introduced by speculative execution. Moreover, we show that the extra power consumption varies among applications and is contributed to by three main factors: the duration of speculative tasks, the idle time, and the allocation of speculative tasks. To the best of our knowledge, our work provides the first deep look into the energy efficiency of speculative execution in Hadoop.

7.3.2. *On the energy footprint of I/O management in Exascale HPC systems*

Participants: Orçun Yildiz, Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The advent of unprecedentedly scalable yet energy hungry Exascale supercomputers poses a major challenge in sustaining a high performance-per-watt ratio. With I/O management acquiring a crucial role in supporting scientific simulations, various I/O management approaches have been proposed to achieve high performance and scalability. However, the details of how these approaches affect energy consumption have not been studied yet.

Therefore, we have explored how much energy a supercomputer consumes while running scientific simulations when adopting various I/O management approaches. In particular, we closely examined three radically different I/O schemes including time partitioning, dedicated cores, and dedicated nodes. To do so, we implemented the three approaches within the Damaris I/O middleware and performed extensive experiments with one of the target HPC applications of the Blue Waters sustained-petaflop supercomputer project: the CM1 atmospheric model.

Our experimental results obtained on the French Grid'5000 platform highlighted the differences among these three approaches and illustrate in which way various configurations of the application and of the system can impact performance and energy consumption. Considering that choosing the most energy-efficient approach for a particular simulation on a particular machine can be a daunting task, we provided a model to estimate the energy consumption of a simulation under different I/O approaches. Our proposed model gives hints to pre-select the most energy-efficient I/O approach for a particular simulation on a particular HPC system and therefore provides a step towards energy-efficient HPC simulations in Exascale systems.

We validated the accuracy of our proposed model using a real-life HPC application (CM1) and two different clusters provisioned on the Grid'5000 testbed. The estimated energy consumptions are within 5.7% of the measured ones for all I/O approaches.

7.3.3. *Exploring energy-consistency trade-offs in cloud storage systems and beyond*

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé.

Apache Cassandra is an open-source cloud storage system that offers multiple types of operation-level consistency including eventual consistency with multiple levels of guarantees and strong consistency. It is being used by many datacenter applications (e.g., Facebook and AppScale). Most existing research efforts have been dedicated to exploring trade-offs such as: consistency vs. performance, consistency vs. latency and consistency vs. monetary cost. In contrast, a little work is focusing on the consistency vs. energy trade-off. As power bills have become a substantial part of the monetary cost for operating a datacenter, we aim to provide a clearer understanding of the interplay between consistency and energy consumption.

In [17], a series of experiments have been conducted to explore the implication of different factors on the energy consumption in Cassandra. Our experiments have revealed a noticeable variation in the energy consumption depending on the consistency level. Furthermore, for a given consistency level, the energy consumption of Cassandra varies with the access pattern and the load exhibited by the application. This further analysis indicated that the uneven distribution of the load amongst different nodes also impacts the energy consumption in Cassandra. Finally, we experimentally compared the impact of four storage configuration and data partitioning policies on the energy consumption in Cassandra: interestingly, we achieve 23% energy saving when assigning 50% of the nodes to the hot pool for the applications with moderate ratio of reads and writes, while applying eventual (quorum) consistency.

This study points to opportunities for future research on consistency-energy trade-offs and offers useful insight into designing energy-efficient techniques for cloud storage systems. This work was done in collaboration with Houssem-Eddine Chihoub (LIG lab, Grenoble) and María Pérez (UPM, Madrid).

Recently, we have been looking at in-memory storage systems. In particular, we are investigating the current replication schemes, data placement strategies and consistency models which are used in in-memory storage systems. Next, an empirical study will be performed to analyze the potential impact of the aforementioned issues on energy consumption. At this point, we are working with RAMCloud.

7.3.4. *Governing energy consumption in Hadoop through CPU frequency scaling: an analysis*

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu.

In [12], we studied the impact of different existing DVFS (*Dynamic Voltage and Frequency Scaling*) governors (i.e., performance, powersave, on-demand, conservative and userspace) on Hadoop's performance and power efficiency. Interestingly, our experimental results reported not only a noticeable variation of the power consumption and performance with different applications and under different governors, but also demonstrate the opportunity to achieve a better tradeoff between performance and power consumption.

The primary contributions of this work are as follows: (1) it provides an overview of the state-of-the-art techniques for energy-efficiency in Hadoop; (2) it discusses and demonstrates the need for exploiting DVFS techniques for energy reduction in Hadoop; (3) it experimentally demonstrates that MapReduce applications experience variations in performance and power consumption under different CPU frequencies and also under different governors. A micro-analysis section is provided to explain this variation and its cause; (4) it illustrates in practice how the behavior of different governors influences the execution of MapReduce applications and how it shapes the performance of the entire cluster; (5) it also brings out the differences between these governors and CPU frequencies and shows that they are not only sub-optimal for different applications but also sub-optimal for different stages of MapReduce execution; (6) it demonstrates that achieving better energy efficiency in Hadoop cannot be done simply by tuning the governor parameters, nor through a naive coarse-grained tuning of the CPU frequencies or the governors according to the running phase (i.e., map phase or reduce phase).

7.4. Scalable I/Os: visualization and processing

7.4.1. *Modeling and predicting I/O patterns of large-scale simulations*

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The increasing gap between the computation performance of post-petascale machines and the performance of their I/O subsystem has motivated many I/O optimizations including prefetching, caching, and scheduling. In order to further improve these techniques, modeling and predicting spatial and temporal I/O patterns of HPC applications as they run has become crucial. Our work in this context focuses on Omnisc'IO, an approach that builds a grammar-based model of the I/O behavior of HPC applications and uses it to predict when future I/O operations will occur, and where and how much data will be accessed. To infer grammars, Omnisc'IO is based on StarSequitur, a novel algorithm extending Nevill-Manning's Sequitur algorithm [11]. Omnisc'IO is transparently integrated into the POSIX and MPI I/O stacks and does not require any modification in applications or higher-level I/O libraries. It works without any prior knowledge of the application and converges to accurate predictions of any N future I/O operations within a couple of iterations. Its implementation is efficient in both computation time and memory footprint.

7.4.2. *In situ analysis and visualization workflows*

Participants: Matthieu Dorier, Lokman Rahmani, Gabriel Antoniu.

In situ visualization has been proposed in the past few years to couple running simulations with parallel visualization and analysis tools. While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with what previously was large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation. In this work, we tackle the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to its content and filters out part of it in order to feed a visualization pipeline with only a reorganized subset of the data produced by the simulation. Our framework monitors its own performance and reconfigures itself dynamically to achieve the best possible visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a $5\times$ speedup and is able to meet performance constraints.

7.5. Scalable storage for data-intensive applications

7.5.1. OverFlow: multi-site aware Big Data management for scientific workflows on clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The global deployment of cloud datacenters is enabling large-scale scientific workflows to improve performance and deliver fast responses. This unprecedented geographical distribution of the computation is doubled by an increase in the scale of the data handled by such applications, bringing new challenges related to the efficient data management across sites. High throughput, low latencies or cost-related trade-offs are just a few concerns for both cloud providers and users when it comes to handling data across datacenters. Existing solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemes. In turn, workflow engines need to improvise substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability.

In [14], we introduced OverFlow, a uniform data-management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. Our solution 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. The system was validated on the Microsoft Azure cloud across its 6 EU and US datacenters. The experiments were conducted on hundreds of nodes using synthetic benchmarks and real-life bio-informatics applications (A-Brain, BLAST). The results show that our system is able to model the cloud performance accurately and to leverage this for efficient data dissemination, being able to reduce the monetary costs and transfer time by up to 3 times.

7.5.2. Efficient transactional storage for data-intensive applications

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

As the computational power used by large-scale applications increases, the amount of data they need to manipulate tends to increase as well. A wide range of such applications require robust and flexible storage support for atomic, durable and concurrent transactions. Historically, databases have provided the *de facto* solution to transactional data management, but they have forced applications to drop control over data layout and access mechanisms, while remaining unable to meet the scale requirements of Big Data. More recently, key-value stores have been introduced to address these issues. However, this solution does not provide transactions, or only restricted transaction support, constraining users to carefully coordinate access to data in order to avoid race conditions, partial writes, overwrites, and other hard problems that cause erratic behavior.

We argue that there is a gap between existing storage solutions and application requirements that limits the design of transaction-oriented data-intensive applications. We have started working on a prototype of a massively parallel distributed transactional blob storage system, aiming to fill this gap.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Microsoft: Z-CloudFlow (2013–2016). In the framework of the Joint Inria-Microsoft Research Center, this project is a follow-up to the A-Brain project. The goal of this new project is to propose a framework for the efficient processing of scientific workflows in clouds. This approach will leverage the cloud infrastructure capabilities for handling and processing large data volumes. In order to support data-intensive workflows, the cloud-based solution will: adapt the workflows to the cloud environment and exploit its capabilities; optimize data transfers to provide reasonable times; manage data and tasks so that they can be efficiently placed and accessed during execution. The validation will be performed using real-life applications, first on the Grid5000 platform, then on the Azure cloud environment, access being granted by Microsoft through a *Azure for Research Award* received by G. Antoniu. The project also provides funding for the PhD thesis of Luis Pineda, started in 2014. The project is being conducted in collaboration with the Zenith team from Montpellier, led by Patrick Valduriez.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

OverFlow (2015–2019). 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. Ultimately, OverFlow will propose a new, pioneering paradigm: Workflow Data Management as a Service — a general and easy-to-use, cloud-provided service that bridges for the first time the gap between single- and multi-site workflow data management. It aims to reap economic benefits from the geo-diversity while accelerating the scientific discovery through a democratization of access to globally distributed data. Within this project, A. Costan is jointly working with Kate Keahey (University of Chicago and Argonne National Laboratory), Bogdan Nicolae (IBM Research) and Christophe Blanchet (Institut Français de Bioinformatique).

9.1.2. Other National Projects

DISCOVERY (2015–2019). An Inria Project Lab, led by Adrien Lebre (ASCOLA), that aims at exploring a new way of operating Utility Computing (UC) resources by leveraging any facilities available through the Internet in order to deliver widely distributed platforms that can better match the geographical dispersal of users as well as the unending demand. Project-teams: ASAP, ASCOLA, Avalon, Myriads, and KerData. Within DISCOVERY, S. Ibrahim (KerData Inria Team) is working with Gilles Fedak (Avalon Inria Project-Team) to address the VM images management challenge.

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

9.2. European Initiatives

9.2.1. FP7 and H2020 Projects

9.2.1.1. BigStorage

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

Program: H2020

Duration: January 2015–January 2019

Coordinator: Universidad politecnica de Madrid

Participants:

- Barcelona Supercomputing Center — Centro Nacional de Supercomputacion (Spain)
- CA Technologies Development Spain (Spain)
- CEA — Commissariat a 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)

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

Inria contact: Gabriel Antoniu and Adrien Lèbre

BigStorage is a European Training Network (ETN) whose main goal is to train future *data scientists* in order to enable them and us to apply holistic and interdisciplinary approaches for taking advantage of a data-overwhelmed world, which 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 will be hosting 2 Early Stage Researchers in this framework.

9.3. International Initiatives

9.3.1. Inria International Labs

9.3.1.1. 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. It has ben created in 2014 as a follow-up of the Inria-UIUC JLPC — *Joint Laboratory for Petascale Computing* to collaborate on concurrency-optimized I/O for Extreme-scale platforms (see details in Section 7.4). The KerData team is collaborating with teams from ANL and UIUC within this lab since 2009. This collaboration has now been formalized as the *Data@Exascale* Associate Team with ANL and UIUC (2013–2015).

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

Title: Ulta-scalable I/O and storage for Exascale systems

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

Start year: 2013

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

As the computational power used by large-scale scientific applications increases, the amount of data manipulated for subsequent analysis increases as well. Rapidly storing this data, protecting it from loss and analyzing it to understand the results are significant challenges, made more difficult by decades of improvements in computation capabilities that have been unmatched in storage. For many applications, the overall performance and scalability clearly become driven by the performance of the I/O subsystem. As we anticipate Exascale systems in 2020, there is a growing consensus in the scientific community that revolutionary new approaches are needed in computational science storage. These challenges are at the center of the activities of the Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing (JLESC, formerly called JLPC). This project gathers researchers from Inria, Argonne National Lab and the University of Illinois at Urbana Champaign to address 3 goals: 1) investigate new storage architectures for Exascale systems; 2) investigate new approaches to the design of I/O middleware for Exascale systems to optimize data processing and visualization, leveraging dedicated I/O cores and I/O forwarding techniques; 3) explore techniques enabling adaptive cloud data services for HPC.

9.3.2. Inria International Partners

9.3.2.1. DataCloud@work

Title: DataCloud@Work — Distributed data management for cloud services

International Partner: Politehnica University of Bucharest (Romania) — Computer Science and Engineering Department — Valentin Cristea and Nicolae Tapus

Start year: January 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

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.

9.4. International Research Visitors

9.4.1. Visits of International Scientists

9.4.1.1. Research stays abroad

Luis Eduardo Pineda Morales: Research visit at ANL, hosted by Kate Keahey and Balaji Subramaniam for 3 months (June–August), funded by the PUF NextGen peoject and by the Microsoft Research Inria Joint Centre project. This work is done in the context of the Joint Laboratory for Extreme-Scale Computing (JLESC).

Orçun Yildiz Research visit at ANL, hosted by Rob Rossa and Matthieu Dorier for 3 months, funded by the PUF NextGen project and by the Data@Exascale Associate Team. This work is done in the context of the Joint Laboratory for Extreme-Scale Computing (JLESC).

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

Gabriel Antoniu: Program Co-Chair of the EIT Digital Future Cloud Symposium (Rennes, October 2015).

10.1.1.2. Member of the organizing committees

Alexandru Costan: Organizer of the IRISA D1 Department Day Seminar.

10.1.2. Scientific events selection

10.1.2.1. Chair of conference program committees

Gabriel Antoniu: Track Chair for the following international conferences: IEEE Cluster 2015 (Data, Storage, and Visualization Track - Chicago, September 2015) and 3PGCIC 2015 (Distributed Algorithms Track - Krakow, November 2015).

Alexandru Costan: Program Co-Chair of the following international workshops: BigDataCloud 2015 held in conjunction with the Euro-Par 2015 conference (Vienna, August 2015) and ScienceCloud 2015 held in conjunction with HPDC 2015 (Portland, June 2015).

10.1.2.2. Member of the conference program committees

Gabriel Antoniu: ACM HPDC 2015, ACM/IEEE CCGrid'2015, ACM/IEEE SC'15, Euro-Par 2015, BigDataCloud 2015 workshop (held in conjunction with the Euro-Par 2015 conference).

Luc Bougé: BigDataCloud 2015, BigData 2015, ICA3PP 2015, CCGRID 2016.

Alexandru Costan: Member of the following Program Committees: IEEE Cluster 2015, CSCS 2015, ARMS-CC 2015, BigDataCloud 2015, ScienceCloud 2015, CSE

Other reviews: IEEE BigData 2015, SC 2015, HPDC 2015, CCGrid 2015, Euro-Par 2015.

Shadi Ibrahim: Member of the following Program Committees: IEEE Cluster 2015, IEEE Cloudcom 2015, ICPADS 2015, IEEE CSE 2015, IEEE FCST 2015, IEEE ICA3PP 2015, IFIP NPC 2015, MEDES 2015, BigDataCloud 2015 workshop (held in conjunction with the Euro-Par 2015 conference), SCRAMBL 2015 workshop (held in conjunction with the CCGrid 2015 conference).

Other reviews: IEEE BigData 2015, SC 2015, HPDC 2015, CCGrid 2015, Euro-Par 2015.

10.1.3. Journal

10.1.3.1. Member of the editorial boards

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

10.1.3.2. Reviewer

Shadi Ibrahim: IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Services Computing, IEEE Transactions on Cloud Computing, ACM Transactions on Architecture and Code Optimization, ACM Transactions on Internet Technology, Future Generation Computer Systems, IEEE Systems Journal, Cluster Computing.

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

10.1.4. Invited talks

Gabriel Antoniu:

3rd JLESC workshop: *To Overlap or Not to Overlap: Optimizing Incremental MapReduce Computations for On-Demand Data Upload*, 3rd workshop of the Joint Laboratory for Extreme Scale Computing, Barcelona, Spain, July 2015.

Huawei Workshop on New Directions in Algorithms and Software, *Scalable data-intensive processing for science on distributed clouds: A-Brain and Z-CloudFlow*, Paris, March 2015.

Inria-Mexico workshop: *Scalable data-intensive processing for science on distributed clouds: A-Brain and Z-CloudFlow*, First Inria-Mexico Workshop in Applied Mathematics and Computer ScienceMexico City, June 2015.

IRISA Data Science Symposium: *Damaris: Jitter-Free I/O Management and In Situ Visualization of HPC Simulations using Dedicated Cores*, Science Data Ecosystem workshop, Rennes, November 2015.

Luc Bougé

BigDataCloud 2016: *Data computing in distributed, very large-scale clouds: From execution models to programming models*, satellite workshop of the Euro-Par 2015 conference, Vienna, August 2015.

10.1.5. Leadership within the scientific community

Luc Bougé: Vice-Chair of the Steering Committee of the Euro-Par conference.

Gabriel Antoniu: Leader for the Big Data, I/O and visualization activity within the JLESC - Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing.

Gabriel Antoniu: Principal Investigator of the Z-CloudFlow Microsoft Research-Inria Project, for which he received an Azure Research Award.

10.1.6. Scientific expertise

Luc Bougé: Member of the jury for the *Agrégation de mathématiques*, the French national hiring committee to hire high-school mathematics teachers at the national level.

Gabriel Antoniu: Project evaluator for 12 ANR projects submitted to the ANR 2016 call (Phase 1).

10.1.7. Research administration

Luc Bougé: Nominated to seat at the CNU (*National University Council*) in the *Informatics* Section (27). His 4-year term has been renewed in November 2015.

Luc Bougé: Scientific Project Leader for Fundamental Informatics at the ANR - *Agence Nationale de la Recherche* until August 2015, for 40% of his time. It was the end of his 3-year delegation contract.

10.2. Teaching - Supervision - Juries

10.2.1. 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: Distributed Systems, 8 hours (lectures), M2 level, ALMA Master, Distributed Architectures module, University of Nantes, France.
- Master: Scalable Distributed Systems, 5 hours (lectures), M2 level, SDS Module, M2RI Master Program, ENS Rennes, France.
- Master: Scalable Distributed Systems, 12 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School, France.
- Master (Engineering Degree, 5th year, *Big Data* option), 10 hours (lectures), M2 level, INSA de Lyon, France.

Luc Bougé

- Bachelor: Introduction to programming concepts, 24 hours (lectures), L3 level, Informatics program, ENS Rennes, France.
- Master: Introduction to object-oriented high-performance programming, 24 hours (lectures), M1 level, Mathematics program, ENS Rennes, France.
- Master: Introduction to compilation, 12 hours (exercice classes), M1 level, Informatics program, Univ. Rennes 1, France.

Shadi Ibrahim

- Master (Engineering Degree, 5th year): Big Data, 24 hours (lectures and lab sessions), M2 level, ENSAI (*École Nationale Supérieure de la Statistique et de l'Analyse de l'Information*), Bruz, France.
- Master : Cloud Computing and Hadoop Technologies, 3 hours (lectures), M2 level, ENSAI (*École Nationale Supérieure de la Statistique et de l'Analyse de l'Information*), Bruz, France.
- Master: Distributed Systems (cloud data management), 4 hours (lectures), M2 level, ALMA Master, Distributed Architectures module, University of Nantes, 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 and Applications, 36h hours (lectures, lab sessions, project), M1, INSA Rennes.

10.2.2. Supervision

10.2.2.1. PhD in progress

Luis Eduardo Pineda Morales: *Efficient Big Data Management for Geographically Distributed Workflows*, thesis started in January 2014, co-advised by Alexandru Costan and Gabriel Antoniu.

Tien-Dat Phan: *Green Big Data Processing in Large-scale Clouds*, thesis started in October 2014, co-advised by Shadi Ibrahim and Luc Bougé.

Lokman Rahmani: *Big Data Management for Next-Generation High-Performance Computing Systems*, thesis started in October 2013 co-advised by Gabriel Antoniu and Luc Bougé.

Orçun Yildiz: *Energy-Efficient Big Data Management in Petasacle Supercomputers and Beyond*, thesis started in September 2014, co-advised by Shadi Ibrahim and Gabriel Antoniu.

Mohammed-Yacine Taleb: *Energy-impact of data consistency management in Clouds and Beyond*, thesis started in August 2015, co-advised by Shadi Ibrahim and Gabriel Antoniu.

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

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.

10.2.3. Juries

Gabriel Antoniu: Referee for the PhD thesis of Ms. Zhou Chi at the Nanyang Technological University (NTU), Singapore (to be defended in January 2016).

Gabriel Antoniu: Jury member of the PhD defense of Ms. Safae Dahmani at the University Bretagne Sud (December 14, 2015).

Shadi Ibrahim: Jury member of the PhD defense of Ms. Karine Pires at the University Pierre et Marie Curie, Paris (March 31, 2015).

10.2.4. Miscellaneous

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

Shadi Ibrahim: Project evaluator in the STIC-AMSUD Program 2015.

Shadi Ibrahim: Leader of the BigStorage project recruitment task force.

Gabriel Antoniu: Invited to give a tutorial on *Big Data Technologies* at the PUF Summer School co-organized with the 3rd JLESC workshop in Barcelona (July 2015).

Shadi Ibrahim: Invited to give a 2 days tutorial on *Hadoop* at IT4Innovations, Technical University Ostrava, Czech Republic (December 2015).

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 Seminar.

10.3. Popularization

Luc Bougé:

- Collège international, Valbonne. Invited presentation to the students of the preparatory classes on *Doing research in computer science* (January 2015).
- Lycée Chateaubriand, Rennes. Invited presentation to the students of the preparatory classes on *Science of numerics* (March 2015).
- IRISA Conf Lunch Program. Invited presentation about *Surviving the Data Deluge* (October 2015).
- Master Program, Rennes. Invited presentation to the M2 students about *Informatics as a scientific activity: Toward a responsible research* (December 2015).
- IRISA Open House Days, Rennes. Invited presentation about *Wikipedia back stage*, and management of an open booth to let visitors improve their skills about searching Wikipedia (December 2015).

Alexandru Costan:

- CumuloNumBio'15, Aussois. Invited presentation at the *Summer School of BioInformatics* about *Big Data Management on Clouds* (June 2015).

IRISA, Rennes. Invited presentation at the *Conf'Lunch* about *Clouds and MapReduce Programming* (October 2015).

EIT Digital, Rennes. Invited presentation at the Future Cloud Symposium about *Enhancing video gaming user experience with Big Data analytics based on Apache Flink - a use case* (October 2015).

Gabriel Antoniu:

EIT Digital. Invited presentation at the Future Cloud Symposium on *Scalable data-intensive processing for science on distributed clouds* (October 2015).

11. Bibliography

Major publications by the team in recent years

[1] 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

[2] 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

[3] 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

[4] 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

[5] M. Dorier, S. Ibrahim, G. Antoniu, R. Ross. *Omnisc'IO: A Grammar-Based Approach to Spatial and Temporal I/O Patterns Prediction*, in "SC'14 - International Conference for High Performance Computing, Networking, Storage and Analysis", New Orleans, United States, IEEE, ACM, November 2014, https://hal.inria.fr/hal-01025670

[6] 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

[7] 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/

[8] 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

[9] V.-T. TRAN, B. NICOLAE, G. ANTONIU. *Towards Scalable Array-Oriented Active Storage: the Pyramid Approach*, in "ACM Operating Systems Review", 2012, vol. 46, n^{o} 1, pp. 19–25, https://hal.inria.fr/hal-00640900

[10] 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

Articles in International Peer-Reviewed Journals

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

[12] S. IBRAHIM, T.-D. PHAN, A. CARPEN-AMARIE, H.-E. CHIHOUB, D. MOISE, G. ANTONIU. *Governing Energy Consumption in Hadoop through CPU Frequency Scaling: an Analysis*, in "Future Generation Computer Systems", February 2015, 14 p. [*DOI :* 10.1016/J.FUTURE.2015.01.005], https://hal.inria.fr/hal-01166252

[13] V. N. SERBANESCU, F. POP, V. CRISTEA, G. ANTONIU. *A formal method for rule analysis and validation in distributed data aggregation service*, in "World Wide Web", November 2015, vol. 18, n^{o} 6, pp. 1717–1736 [*DOI :* 10.1007/S11280-015-0334-4], https://hal.archives-ouvertes.fr/hal-01249152

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

[15] R. TUDORAN, A. COSTAN, O. NANO, I. SANTOS, H. SONCU, G. ANTONIU. *JetStream: Enabling high throughput live event streaming on multi-site clouds*, in "Future Generation Computer Systems", 2015, vol. 54 [*DOI :* 10.1016/J.FUTURE.2015.01.016], https://hal.inria.fr/hal-01239124

International Conferences with Proceedings

[16] N. CHERIERE, P. DONAT-BOUILLUD, S. IBRAHIM, M. SIMONIN. *On the Usability of Shortest Remaining Time First Policy in Shared Hadoop Clusters*, in "SAC 2016-The 31st ACM/SIGAPP Symposium on Applied Computing", Pisa, Italy, April 2016, https://hal.inria.fr/hal-01239341

[17] H.-E. CHIHOUB, S. IBRAHIM, Y. LI, G. ANTONIU, M. PÉREZ, L. BOUGÉ. *Exploring Energy-Consistency Trade-offs in Cassandra Cloud Storage System*, in "SBAC-PAD'15-The 27th International Symposium on Computer Architecture and High Performance Computing", Florianopolis, Santa Catarina, Brazil, October 2015, https://hal.inria.fr/hal-01184235

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

[19] S. Ibrahim, T. A. Phuong, G. Antoniu. *An Eye on the Elephant in the Wild: A Performance Evaluation of Hadoop's Schedulers Under Failures*, in "ARMS-CC'15-The second workshop on Adaptive Resource Management and Scheduling for Cloud Computing, held in conjunction with PODC 2015,", Donostia-San Sebastián, Spain, July 2015, https://hal.inria.fr/hal-01184236

[20] B. Memishi, M. S. Pérez-Hernández, G. Antoniu. *Diarchy: An Optimized Management Approach for MapReduce Masters*, in "ICCS 2015: Proceedings of the International Conference on Computational Science, Computational Science at the Gates of Nature", Reykjavík, Iceland, June 2015, pp. 9–18 [*DOI :* 10.1016/j.procs.2015.05.179], https://hal.archives-ouvertes.fr/hal-01249151

[21] T.-D. Phan, S. Ibrahim, G. Antoniu, L. Bougé. *On Understanding the Energy Impact of Speculative Execution in Hadoop*, in "GreenCom'15-The 2015 IEEE International Conference on Green Computing and Communications", Sydney, Australia, December 2015, https://hal.inria.fr/hal-01238055

[22] L. Pineda-Morales, A. Costan, G. Antoniu. *Towards Multi-site Metadata Management for Geographically Distributed Cloud Workflows*, in "CLUSTER 2015 - IEEE International Conference on Cluster Computing", Chicago, United States, September 2015 [*DOI :* 10.1109/CLUSTER.2015.49], https://hal.inria.fr/hal-01239150

[23] O. Yildiz, S. Ibrahim, T. A. Phuong, G. Antoniu. *Chronos: Failure-Aware Scheduling in Shared Hadoop Clusters*, in "BigData'15-The 2015 IEEE International Conference on Big Data", Santa Clara, CA, United States, October 2015, https://hal.inria.fr/hal-01203001

Conferences without Proceedings

[24] R.-I. Roman, B. Nicolae, A. Costan, G. Antoniu. *Understanding Spark Performance in Hybrid and Multi-Site Clouds*, in "6th International Workshop on Big Data Analytics: Challenges and Opportunities (BDAC-15)", Austin, TX, United States, November 2015, https://hal.inria.fr/hal-01239140

Research Reports

[25] M. Dorier, S. Ibrahim, G. Antoniu, R. Ross. *On the Use of Formal Grammars to Predict HPC I/O Behaviors*, ENS Rennes ; Inria Rennes Bretagne Atlantique ; Argonne National Laboratory ; Inria, August 2015, n^{o} RR-8725, https://hal.inria.fr/hal-01149941

[26] A. Lebre, J. Pastor, . The DISCOVERY Consortium. *The DISCOVERY Initiative - Overcoming Major Limitations of Traditional Server-Centric Clouds by Operating Massively Distributed IaaS Facilities*, Inria, September 2015, n^{o} RR-8779, 14 p. , https://hal.inria.fr/hal-01203648

[27] P. Matri, A. Costan, G. Antoniu, J. Montes, M. Pérez. *Týr: Efficient Transactional Storage for Data-Intensive Applications*, Inria Rennes Bretagne Atlantique ; Universidad Politécnica de Madrid, January 2016, n^{o} RT-0473, 25 p. , https://hal.inria.fr/hal-01256563

Other Publications

[28] L. Pineda-Morales, B. Subramaniam, K. Keahey, G. Antoniu, A. Costan, S. Wang, A. Padmanabhan, A. Soliman. *Scaling Smart Appliances for Spatial Data Synthesis*, November 2015, SC15 - ACM/IEEE International Conference in Supercomputing, Poster, https://hal.inria.fr/hal-01241718

References in notes

[29] *Amazon Elastic MapReduce*, 2010, http://aws.amazon.com/elasticmapreduce/

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

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

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

[33] 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

[34] B. NICOLAE, D. MOISE, G. ANTONIU, L. BOUGÉ, M. DORIER. *BlobSeer: Bringing High Throughput under Heavy Concurrency to Hadoop Map-Reduce Applications*, in "24th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2010)", Atlanta, GA, USA, IEEE and ACM, April 2010, A preliminary version of this paper has been published as Inria Research Report RR-7140