IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

Ecole normale supérieure de Cachan

Activity Report 2013

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

Keywords: High Performance Computing, Big Data, Cloud Computing, Middleware, Data Management, Data Storage

The KerData project-team was officially created on July 1, 2012. It is associated with the IRISA CNRS joint laboratory, University Rennes 1, INSA Rennes and ENS Cachan/Rennes.

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

1. Members

Research Scientists

Gabriel Antoniu [Team leader, Inria, Senior Researcher, HdR]
Shadi Ibrahim [Inria, Researcher, since Sep 2013]

Faculty Members

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

Engineers

Zhe Li [Inria, until Nov 2013]
Rohit Saxena [Inria, granted by the ANR FP3C Project]

PhD Students

Houssem-Eddine Chihoub [Inria, granted by the ANR MapReduce Project, until Nov 2013]
Matthieu Dorier [ENS Cachan]
Álvaro García Recuero [Inria, from Oct 2013]
Lokman Rahmani [Univ. Rennes I, from Sep 2013]
Viet-Trung Tran [Inria, granted by the ANR MapReduce Project, until Jan 2013]
Radu Tudoran [Univ. Rennes I]

Post-Doctoral Fellows

Elena Apostol [Inria, Post-Doctoral Fellow, funded by the Microsoft Research-Inria A-Brain project, from Jul 2013]
Shadi Ibrahim [Inria, Post-Doctoral Fellow, funded by Inria / Ministry of Research, until Aug 2013]

Visiting Scientists

Rui Wang [Inria, Master intern from Feb 2013 until Aug 2013]
Yue Li [Inria, Master intern from Feb 2013 until Jun 2013]
Ana-Ruxandra Ion [Inria, Master intern from May 2013 until Aug 2013]
Mihaela-Catalina Nita [Inria, Master intern from Apr 2013 until Aug 2013]

Administrative Assistant

Elodie Lequoc [Univ. Rennes I]

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 platforms*, 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 post-Petascale High-Performance Computing (HPC), which shape the longer-term agenda of the Cloud Computing and Exascale HPC research communities.

Our research activities focus on data-intensive high-performance applications 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.

Examples of such applications are:

- Massively parallel cloud data-mining applications (e.g., MapReduce-based data analysis).
- Advanced Platform-as-a-Service (PaaS) cloud data services requiring efficient data sharing under heavy concurrency.
- Advanced concurrency-optimized, versioning-oriented cloud services for virtual-machine-image storage and management at IaaS (Infrastructure-as-a-Service) level.
- Scalable storage solutions for I/O-intensive HPC simulations for post-Petascale architectures.
- Storage and I/O stacks for big data analysis in applications that manipulate structured scientific data (e.g. very large multi-dimensional arrays).

2.2. Highlights of the Year

Team: Shadi Ibrahim, a former Post-Doc fellow at the KerData project-team, has been hired as a permanent Junior Researcher at Inria (CR2) starting in October 2013.

A-Brain MSR-Inria project: The TomusBlobs data-storage layer developed in the framework of the A-Brain MSR-Inria project was demonstrated to scale up to 1000 cores on 3 Azure data centers; it exhibits improvements in execution time up to 50 % compared to standard solutions based on Azure BLOB storage. Based on this storage infrastructure, the A-Brain project has provided the first statistical evidence of the heritability of functional signals in a failed stop task in basal ganglia, using a ridge regression approach, while relying on the Azure cloud to address the computational burden.

Joint Lab for Petascale Computing: The Damaris middleware for I/O and in-situ visualization, initiated in 2010 in the framework of the Joint Laboratory for Petascale Computing, was ported to NCSA's Blue Waters supercomputer and provided in-situ visualization capabilities to the CM1 atmospheric simulation on up to 6400 cores.

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 [34] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [30]. At the core of Map-Reduce frameworks lies 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*. Additionally, as thousands of clients simultaneously access shared data, it is critical to preserve *fault-tolerance* and *security* requirements.

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] and EESI [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 and analyze massive outputs of data during and after the simulation without impacting the overall performance.

The overall goal of the KerData project-team is to bring a substantial contribution to the effort of the research community to address the above challenges. KerData aims to design and implement 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. We are also looking at other kinds of infrastructures (that we are considering as secondary), e.g. hybrid platforms combining enterprise desktop grids extended to cloud platforms. Our collaboration porfolio includes international teams that are active in this area both in Academia (e.g., Argonne National Lab, University of Illinois at Urbana-Champaign, University of Tsukuba) and Industry (Microsoft, IBM).

The highly experimental nature of our research validation methodology should be stressed. Our approach relies on building prototypes and on their large-scale experimental validation on real testbeds and experimental platforms. We strongly rely on the ALADDIN-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 have access to the Jaguar and Kraken supercomputers (ranked 3rd and 11th respectively in the Top 500 supercomputer list) and to the Blue Waters supercomputer. This provides us with excellent opportunities to validate our results on 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 potential impact.

3.2. Our research agenda

Three typical application scenarios are described in Section 4.1:

- Joint genetic and neuroimaging data analysis on Azure clouds;
- Structural protein analysis on Nimbus clouds;
- I/O intensive climate 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 efficiently address the aforementioned challenges we have started to work out an approach called BlobSeer, which stands today at the center of our research efforts. 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.

Preliminary experiments with our BlobSeer BLOB management system within today's cloud software infrastructures proved very promising. Recently, we used the BlobSeer approach as a starting point to address more in depth two usage scenarios, which led to two more specific approaches: 1) Pyramid (which borrows many concepts from BlobSeer), with a specific focus on array-oriented storage; and 2) Damaris (totally independent of BlobSeer), which exploits multicore parallelism in post-Petascale supercomputers. All these directions are described below.

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 agenda for the next 4 years.

Based on very promising results demonstrated by BlobSeer in preliminary experiments [36], we have initiated several collaborative projects in the area of cloud data management, e.g., the MapReduce ANR project, the A-Brain Microsoft-Inria project, the Z-CloudFlow Microsoft-Inria project. 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. Thanks to these projects, we have already started to enjoy a visible scientific positioning at the international level.

The particularly active Data@ExaScale Associate Team creates the framework for an enlarged research activity involving a large number of young researchers and students. It serves as a basis for extended research activities based on our approaches, carried out beyond the frontiers of our team. In the HPC area, our presence in the research activities of the Joint UIUC-Inria Lab for Petascale Computing at Urbana-Champaign is a very exciting opportunity that we have started to leverage. It 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, as described in Section 6.4.

4. Application Domains

4.1. Application Domains

Below are three examples which illustrate the needs of large-scale data-intensive applications with respect to storage, I/O and data analysis. They illustrate the classes of applications that can benefit from our research activities.

4.1.1. 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, and that has remained poorly understood so far. As both neuroimaging- and genetic-domain observations represent a huge amount of variables (of the order of millions), performing statistically rigorous 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.

The A-Brain (AzureBrain) Project started in October 2010 within the Microsoft Research-Inria Joint Research Center. It is co-led by the KerData (Rennes) and Parietal (Saclay) Inria teams. They jointly address this computational problem using cloud related techniques on Microsoft Azure cloud infrastructure. The two teams bring together their complementary expertise: KerData in the area of scalable cloud data management, and Parietal in the field of neuroimaging and genetics data analysis.

In particular, KerData brings its expertise in designing solutions for optimized data storage and management for the Map-Reduce programming model. This model has recently arisen as a very effective approach to develop high-performance applications over very large distributed systems such as grids and now clouds. The computations involved in the statistical analysis designed by the Parietal team fit particularly well with this model.

4.1.2. Structural protein analysis on Nimbus clouds

Proteins are major components of the life. They are involved in lots of biochemical reactions and vital mechanisms for living organisms. The three-dimensional (3D) structure of a protein is essential for its function and for its participation to the whole metabolism of a living organism. However, due to experimental limitations, only few protein structures (roughly, 60,000) have been experimentally determined, compared to the millions of proteins sequences which are known. In the case of structural genomics, the knowledge of the 3D structure may be not sufficient to infer the function. A usual way to make a structural analysis of a protein or to infer its function is to compare its known, or potential, structure to the whole set of structures referenced in the *Protein Data Bank* (PDB).

In the framework of the MapReduce ANR project led by KerData, we focus on the SuMo application (*Surf the Molecules*) proposed by Institute for Biology and Chemistry of the Proteins from Lyon (IBCP, a partner in the MapReduce project). This application performs structural protein analysis by comparing a set of protein structures against a very large set of structures stored in a huge database. This is a typical data-intensive application that can leverage the Map-Reduce model for a scalable execution on large-scale distributed platforms. Our goal is to explore storage-level concurrency-oriented optimizations to make the SuMo application scalable for large-scale experiments of protein structures comparison on cloud infrastructures managed using the Nimbus IaaS toolkit developed at Argonne National Lab (USA).

If the results are convincing, then they can immediately be applied to the derived version of this application for drug design in an industrial context, called MED-SuMo, a software managed by the MEDIT SME (also a partner in this project). For pharmaceutical and biotech industries, such an implementation run over a cloud computing facility opens several new applications for drug design. Rather than searching for 3D similarity into biostructural data, it will become possible to classify the entire biostructural space and to periodically update all derivative predictive models with new experimental data. The applications in this complete chemo-proteomic vision concern the identification of new druggable protein targets and thereby the generation of new drug candidates.

4.1.3. 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 efficiently record and visualize data during the simulation 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 speeds. 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 (http://www.ncsa.illinois.edu/BlueWaters/), which aims to build one of the most powerful supercomputers in the world.

In this context, the KerData project-team started to explore ways to remove the limitations mentioned above through collaborative work in the framework of the Joint Inria-UIUC Lab for Petascale Computing (JLPC, Urbana-Champaign, Illinois, USA), whose research activity focuses on the Blue Waters project. As a starting point, we are focusing on a particular tornado simulation code called CM1 (Cloud Model 1), which is intended to be run on the Blue Waters machine. Preliminary investigation demonstrated the inefficiency of the current I/O approach, which typically consists in periodically writing a very large number of small files. This causes bursts of I/O in the parallel file system, leading to poor performance and extreme variability (jitter) compared to what could be expected from the underlying hardware. 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 JLPC, we started to address these challenges through an approach based on dedicated I/O cores.

5. Software and Platforms

5.1. BlobSeer

Participants: Zhe Li, Rohit Saxena, Alexandru Costan, Gabriel Antoniu, Luc Bougé.

Contact: Gabriel Antoniu.

Presentation: BlobSeer is the core software platform for most current 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*.

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*) started in November 2012 for two years, aiming at robustifying the BlobSeer software and making it a safely distributable product. This project is funded by Inria *Technological Development Office* (D2T, *Direction du Développement Technologique*). Loïc Cloatre, has been hired as a senior engineer for the second year of this project, as a successor of Zhe Li, starting in February 2014.

5.2. BlobSeer-WAN

Participants: Rohit Saxena, Alexandru Costan, Gabriel Antoniu.

Contact: Gabriel Antoniu.

Presentation: BlobSeer-WAN was initially designed as an extension of BlobSeer, targeting geographically distributed environments. With BlobSeer-WAN, the metadata is replicated asynchronously for low latency. There is a version manager on each site and vector clocks are used to allow collision detection and resolution under highly concurrent access. Several experiments have been conducted with this setup on the Grid'5000 testbed which have shown scalable metadata performance under geographically distributed environments. Currently, BlobSeer-WAN is integrated within BlobSeer, as a new release of the latter.

Users: BlobSeer-WAN has been preliminarily evaluated at University of Tsukuba (Japan) in the context of the FP3C project. BlobSeer-WAN is used as a storage backend for HGMDS, a multi master metadata server designed for a global distributed file system.

URL: http://blobseer.gforge.inria.fr./doku.php?id=ci:blobseer-wan

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

Status: This software is available on Inria's forge as part of BlobSeer. Registration with APP is in progress.

5.3. Damaris

Participants: Matthieu Dorier, Lokman Rahmani, Gabriel Antoniu.

Contact: Gabriel Antoniu.

Presentation: 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.

Users: 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).

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.

5.4. TomusBlobs

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Contact: Gabriel Antoniu.

Presentation: TomusBlobs is a software library for concurrency-optimized data storage for data-intensive applications running on Azure clouds, including MapReduce applications. It is being developed by the KerData Inria Project-Team in the framework of the A-Brain MSR-Inria project. It uses the BlobSeer library.

Users: TomusBlobs has been preliminarily evaluated within the A-Brain project where it was used to execute a real-life application aiming to search for significant associations between brain images and genetics data. The TomusBlobs data-storage layer developed in the framework of the A-Brain MSR-Inria project was demonstrated to scale up to 1000 cores on 3 Azure data centers; it exhibits improvements in execution time up to 50 % compared to standard solutions based on Azure BLOB storage. Based on this storage infrastructure, the A-Brain project consortium has provided the first statistical evidence of the heritability of functional signals in a failed stop task in basal ganglia, using a ridge regression approach, while relying on the Azure cloud to address the computational burden.

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

Status: This software is available on Inria's forge. Registration with APP is in progress.

5.5. Darshan-Ruby

Participant: Matthieu Dorier.

Contact: Matthieu Dorier.

Presentation: Darshan-Ruby is a Ruby extension to the Darshan scalable HPC I/O characterization tool (developed by the Mathematics and Computer Science division at Argonne National Lab). It simplifies the access to the contents of Darshan-generated log files, in an object-oriented manner through the Ruby scripting language.

Users: Darshan-Ruby is available as a Ruby gem package on the official Rubygems website (http://rubygems.org/), and is referenced on the Darshan website (http://www.mcs.anl.gov/research/projects/darshan/).

URL: http://darshan-ruby.gforge.inria.fr/

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

Status: This software is available on Inria's forge.

5.6. Derived software

Derived from BlobSeer, an additional platform is currently being developed within KerData: Pyramid, a software service for array-oriented active storage developed within the framework of Viet-Trung Tran's PhD thesis.

6. New Results

6.1. A-Brain and TomusBlobs

6.1.1. Experiments with TomusBlobs at large scale

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Joint genetic and neuro-imaging data analysis may help identifying risk factors in target populations. Performing such studies on a large number of subjects is challenging as genotyping DNA chips can record several hundred thousands values per subject, while the fMRI images may contain 100k–1M volumetric picture elements. Determining statistically significant links between the two sets of data entails a massive amount of computation as one needs not only to compare all pair-wise relations but also to correct for family-wise multiple comparisons. These false positives are controlled by generating permutations of the input data set. The A-Brain initiative is such a data analysis application involving large cohorts of subjects and used to study and understand the variability that exists between individuals. Supposing that such an application could be executed on a single machine, the computation would take years. Cloud infrastructures have the potential to decrease this computation time to days, by parallellizing and scaling out the application. In order to execute this computation in parallel at a large scale, we noticed that the A-Brain application can be easily described as a MapReduce process. The problem was further divided into 28,000 computation tasks, which were executed as map jobs.

The experiment timespan was 14 days and was performed across 4 cloud deployments in 2 different US Azure data centers — North and West. The processing time for a map job is approximatively 2 hours and there are no notable time differences between the map execution time with respect to the geographical location. This is achieved due to the load balancing of the workload, the data locality within the deployments and to the geographical partition. The global result was aggregated using a MapIterativeReduce technique which was composed of 563 reduce jobs. This reduction process eliminates the implicit barrier between mappers and reducers, the reduction process happens in parallel with the map computation. During the period of the experiment the Azure services became temporary inaccessible, due to a failure of a secured certificate. Despite this problem, the framework was capable to handle the failure due to a safety mechanism that we implemented which suspended the computation until all Azure services became available again. Regarding the lost map/reduce enqueued jobs, the monitor mechanism, which supervises the computation progress, was able to restore them. The cost of the experiment was approximatively 210,000 compute hours, where 1 compute hour is equivalent to 1 CPU running for one hour. The monetary cost of the experiment adds up to almost 20,000 $. The total amount combines the cost of the compute resources, for which a value of 0.09 $/h was considered, the persistent Azure storage cost and the outbound traffic from the data centers. As a result of this experiment, we have confirmed that brain activation signals are a heritable feature.

6.1.2. Using dedicated compute nodes for data management on public clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

A large spectrum of scientific applications, some generating data volumes exceeding petabytes, are currently being ported on clouds to build on their inherent elasticity and scalability. One of the critical needs in order to deal with this "data deluge" is an efficient, scalable and reliable storage. However, the storage services proposed by cloud providers suffer from high latencies, trading performance for availability. One alternative is to federate the local virtual disks on the compute nodes into a globally shared storage used for large intermediate or checkpoint data. This collocated storage supports a high throughput but it can be very intrusive and subject to failures that can stop the host node and degrade the application performance.

To deal with these limitations we proposed DataSteward [25], a data management system that provides a higher degree of reliability while remaining non-intrusive through the use of dedicated compute nodes. DataSteward harnesses the storage space of a set of dedicated VMs, selected using a topology-aware clustering algorithm, and has a lifetime dependent on the deployment lifetime. To capitalize on this separation, we introduced a set of scientific data processing services on top of the storage layer, that can overlap with the executing applications. We performed extensive experimentations on hundreds of cores in the Azure cloud: compared to state-of-the-art node selection algorithms, we show up to a 20 % higher throughput, which improves the overall performance of a real life scientific application by up to 45 %.

6.1.3. File transfers for workflows

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific workflows typically communicate data between tasks using files. Currently, on public clouds, this is achieved by using the cloud storage services, which are unable to exploit the workflow semantics and are subject to low throughput and high latencies. To overcome these limitations, we propose in [26] an alternative leveraging data locality through direct file transfers between the compute nodes. We rely on the observation that workflows generate a set of common data access patterns that our solution exploits in conjunction with context information to self-adapt, choose the most adequate transfer protocol and expose the data layout within the virtual machines to the workflow engines. This file management system was integrated within the Microsoft Generic Worker workflow engine and was validated using synthetic benchmarks and a real-life application on the Azure cloud. The results show it can bring significant performance gains: up to 5x file transfer speedup compared to solutions based on standard cloud storage and over 25 % application timespan reduction compared to Hadoop on Azure. This work was done in colaboration with Goetz Brasche and Ramin Rezai Rad from *Microsoft Advance Technology Lab Europe*.

6.2. Optimizing MapReduce Processing

6.2.1. Optimizing MapReduce in virtualized environments

Participant: Shadi Ibrahim.

As data-intensive applications become popular in the cloud, their performance on the virtualized platform calls for empirical evaluations and technical innovations. Virtualization has become a prominent tool in data centers and is extensively leveraged in cloud environments: it enables multiple virtual machines (VMs) — with multiple operating systems and applications — to run within a physical server. However, virtualization introduces the challenging issue of providing effective QoS to VMs and preserving the high disk utilization (i.e., reducing the seek delay and rotation overhead) when allocating disk resources to VMs. We addressed these challenges by developing two Disk I/O scheduling frameworks: *Flubber* and *Pregather*.

In [17], we developed a two-level scheduling framework that decouples throughput and latency allocation to provide QoS guarantees to VMs while maintaining high disk utilization. The high-level throughput control regulates the pending requests from the VMs with an adaptive credit-rate controller, in order to meet the throughput requirements of different VMs and ensure performance isolation. Meanwhile, the low-level latency control, by the virtue of the batch and delay earliest deadline first mechanism (BD-EDF), re-orders all pending requests from VMs based on their deadlines, and batches them to disk devices taking into account the locality of accesses across VMs.

In [24], we developed a novel disk I/O scheduling framework, named *Pregather*, to improve disk I/O efficiency through exposure and exploitation of the special spatial locality in the virtualized environment (regional and sub-regional spatial locality corresponds to the virtual disk space and applications' access patterns, respectively), thereby improving the performance of disk-intensive applications (e.g., MapReduce applications) without harming the transparency feature of virtualization (without a priori knowledge of the applications' access patterns). The key idea behind Pregather is to implement an intelligent model to predict the access regularity of sub-regional spatial locality for each VM.

We evaluated *Pregather* through extensive experiments that involve multiple simultaneous applications of both synthetic benchmarks and a MapReduce application (i.e., distributed sort) on Xen-based platforms. Our experiments indicate that *Pregather* results in high disk spatial locality, yields a significant improvement in disk throughput, and ends up with improved Hadoop performance. This work was done in collaboration with Hai Jin, Song Wu and Xiao Ling from Huazhong University of Science and Technology (HUST).

6.2.2. Investigating energy efficiency in MapReduce

Participants: Shadi Ibrahim, Houssem-Eddine Chihoub, Gabriel Antoniu, Luc Bougé.

A MapReduce system spans over a multitude of computing nodes that are frequency- and voltage-scalable. Furthermore, many MapReduce applications show significant variation in CPU load during their running time. Thus, there is a significant potential for energy saving by scaling down the CPU frequency. Some power-aware data-layout techniques have been proposed to save power, at the cost of a weaker performance. MapReduce applications range from CPU-Intensive to I/O-Intensive. More importantly, a typical MapReduce application comprises many subtasks, each subtask's workload being either a computation, a disk request or a bandwidth request. As a result, there is a high potential for power reduction by scaling down the CPU when the peak CPU performance is not used.

In this ongoing work, a series of experiments are conducted to explore the implications of *Dynamic Voltage Frequency scaling* (DVFS) settings on power consumption in Hadoop-clusters, by benefitting from the current maturity in DVFS research and of the introduction of governors (e.g., *performance*, *powersave*, *ondemand*, *conservative* and *userspace*). By adapting existing DVFS governors in Hadoop clusters, we observe significant variation in performance and power consumption of the cluster with different applications when applying these governors: the different DVFS settings are only sub-optimal for different MapReduce applications. Furthermore, our results reveal that current CPU governors do not exactly reflect their design goal and may even become ineffective to manage the power consumption. Based on this analysis, we are investigating a new approach to reduce the energy consumption in Hadoop through adaptively tuning the governors and/or the CPU frequencies during the execution of MapReduce jobs.

6.2.3. Hybrid infrastructures

Participants: Alexandru Costan, Ana-Ruxandra Ion, Gabriel Antoniu.

As Map-Reduce emerges as a leading programming paradigm for data-intensive computing, today's frameworks which support it still have substantial shortcomings that limit its potential scalability. At the core of Map-Reduce frameworks lies a key component with a huge impact on their performance: the storage layer. To enable scalable parallel data processing, this layer must meet a series of specific requirements. An important challenge regards the target execution infrastructures. While the Map-Reduce programming model has become very visible in the cloud computing area, it is also subject to active research efforts on other kinds of large-scale infrastructures, such as desktop grids. We claim that it is worth investigating how such efforts (currently done in parallel) could converge, in a context where large-scale distributed platforms become more and more connected together.

We investigated several directions where there is room for such progress: they concern storage efficiency under massive-data access concurrency, scheduling, volatility and fault-tolerance. We placed our discussion in the perspective of the current evolution towards an increasing integration of large-scale distributed platforms (clouds, cloud federations, enterprise desktop grids, etc.). We proposed an approach which aims to overcome the current limitations of existing Map-Reduce frameworks, in order to achieve scalable, concurrency-optimized, fault-tolerant Map-Reduce data processing on hybrid infrastructures. We are designing and implementing our approach through an original architecture for scalable data processing: it combines two approaches, BlobSeer and BitDew, which have shown their benefits separately (on clouds and desktop grids respectively) into a unified system. The global goal is to improve the behavior of Map-Reduce-based applications on the target large-scale infrastructures. The internship of Ana-Ruxandra Ion was dedicated to this topic and showed that for reliable hybrid Map-Reduce processing, one needs to first rely on public/private cloud resources, and then to scale them up using the local, yet volatile, desktop grid resources.

6.2.4. Key partitioning techniques

Participants: Shadi Ibrahim, Gabriel Antoniu.

Data locality is a key feature in MapReduce that is extensively leveraged in data-intensive cloud systems: it avoids network saturation when processing large amounts of data by co-allocating computation and data storage, particularly for the map phase. However, our studies with Hadoop, a widely used MapReduce implementation, demonstrate that the presence of partitioning skew (partitioning skew refers to the case when a variation in either the intermediate keys' frequencies or their distributions or both among different data nodes) causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes. As a result, the applications suffer from severe performance degradation due to the long data transfer during the shuffle phase along with the computation skew, particularly in reduce phase. We addressed these problems by developing a new key/value partitioning called *LEEN*.

In [16], we develop a novel algorithm named *LEEN* for locality-aware and fairness-aware key partitioning in MapReduce. *LEEN* aims at saving the network bandwidth dissipation during the shuffle phase of the MapReduce job along with balancing the reducers' inputs. *LEEN* is conducive to improve the data locality of the MapReduce execution efficiency by the virtue of the asynchronous map and reduce scheme, thereby having more control on the keys distribution in each data node. *LEEN* keeps track of the frequencies of buffered keys hosted by each data node. In doing so, *LEEN* efficiently moves buffered intermediate keys to the destination considering the location of the high frequencies along with fair distribution of reducers' inputs.

To quantify the locality, data distribution and performance of *LEEN*, we conducted a comprehensive performance evaluation study using *LEEN* in Hadoop. Our experimental results demonstrate that *LEEN* interestingly can efficiently achieve higher locality, and balance data distribution after the shuffle phase. This work was done in collaboration with Hai Jin, Song Wu and Lu Lu from Huazhong University of Science and Technology (HUST) and Bingsheng He from Nanyang Technological University (NTU).

6.3. Cloud Storage Trade-Offs: Consistency and Self-Adaptiveness

6.3.1. Cost-aware consistency management in the cloud

Participants: Houssem-Eddine Chihoub, Shadi Ibrahim, Gabriel Antoniu.

With the emergence of cloud computing, many organizations have moved their data to the cloud in order to provide scalable, reliable and highly available services. To meet ever growing user needs, these services mainly rely on geographically-distributed data replication to guarantee good performance and high availability. However, with replication, consistency comes into question. Service providers in the cloud have the freedom to select the level of consistency according to the access patterns exhibited by the applications. Most optimizations efforts then concentrate on how to provide adequate trade-offs between consistency guarantees and performance. However, as the monetary cost completely relies on the service providers, in [20] we argue that monetary cost should be taken into consideration when evaluating or selecting a consistency level in the cloud. Accordingly, we define a new metric called *consistency-cost efficiency*. Based on this metric, we present a simple, yet efficient economical consistency model, called *Bismar*, that adaptively tunes the consistency level at run-time in order to reduce the monetary cost while simultaneously maintaining a low fraction of stale reads. Experimental evaluations with the Cassandra cloud storage on a Grid'5000 testbed show the validity of the metric and demonstrate the effectiveness of the proposed consistency model.

6.3.2. Analysis of the impact of consistency mangement on energy consumption

Participants: Houssem-Eddine Chihoub, Shadi Ibrahim, Gabriel Antoniu.

Energy consumption within datacenters has grown exponentially in recent years. In the era of Big Data, storage and data-intensive applications are one of the main causes of the high power usage. However, few studies have been dedicated to the analysis of the energy consumption of storage systems. Moreover, the impact of consistency management has never been investigated in spite of its high importance. In this work, we address this particular issue. We investigate the energy consumption of application workloads with different consistency models. Thereafter, we leverage the observations about power and the resource usage with every consistency level in order to provide insight into energy-saving practices. In this context, we introduce adaptive configurations of the storage cluster according to the used consistency level. Our experimental evaluations on Cassandra deployed on Grid'5000 demonstrate the existence of the inevitable tradeoff between consistency and energy saving. Moreover, they show how reconfiguring the storage cluster can lead to energy saving, enhanced performance, and better consistency.

6.3.3. Chameleon: customized consistency by means of application behavior modeling

Participants: Houssem-Eddine Chihoub, Gabriel Antoniu.

Multiple Big Data applications are being deployed worldwide to serve a very large number of clients nowadays. These applications vary in their performance and consistency requirements. Understanding such requirements at the storage system level is not possible. The high level semantics of an application is not exposed at the system level. In this context, the consequences of a stale read are not the same for all types of applications.

In [28] , we focus on managing consistency at the application level rather than at the system level. In order to achieve this goal, we propose an offline modeling approach of the application access behavior that considers its high–level consistency semantics. Furthermore, every application state is automatically associated with a consistency policy. At runtime, we introduce the *Chameleon* approach that leverages the application model to provide a customized consistency specific to that application. Experimental evaluations show the high accuracy of our modeling approach exceeding 96% of correct classification of the application states. Moreover, our experiments conducted on Grid'5000 show that *Chameleon* adapts, for every time period, according to the application behavior and requirements while providing best-effort performance.

6.4. Scalable I/O and Virtualization for Exascale Systems

6.4.1. Damaris/Viz

Participants: Matthieu Dorier, Gabriel Antoniu, Lokman Rahmani.

In the context of the Joint Inria/UIUC/ANL Laboratory for Petascale computing (JLCP), we are developing Damaris, which enables efficient I/O, data analysis and visualization at very large scale from SMP machines. The I/O bottlenecks already present on current petascale systems as well as the amount of data written by HPC applications force to consider new approaches to get insights from running simulations. Trying to bypass the need for storage or drastically reducing the amount of data generated will be of outmost importance for exascale. In-situ visualization has therefore been proposed to run analysis and visualization tasks closer to the simulation, as it runs.

We investigated the limitations of existing in-situ visualization software and proposed Damaris/Viz, a new version of Damaris that fills the gaps of these software by providing in-situ visualization support to Damaris. The use of Damaris/Viz on top of existing visualization packages allows us to:

- Reduce code instrumentation to a minimum in existing simulations,
- Gather the capabilities of several visualization tools to offer adaptability under a unified data management interface,
- Use dedicated cores to hide the run time impact of in-situ visualization and
- Efficiently use memory through a shared-memory-based communication model.

Experiments were conducted on Blue Waters (Cray XK6 at NCSA), Intrepid (BlueGene/P at ANL) and Grid'5000 with representative visualization scenarios for the CM1 [33] atmospheric simulation and the Nek5000 [35] CFD solver. Part of these experiments were carried by NCSA researcher Roberto Sisneros, who gave us important (and very positive) feedbacks on the usability of Damaris at scale (up to 6400 cores on Blue Waters) with real applications. The results of this work were presented as a poster in the PhD forum of IEEE IPDPS 2013 [22], published in a research report [29] and at the IEEE LDAV 2013 conference [23], and a demo of Damaris/Viz was presented at Inria's exhibition booth at the Supercomputing (SC 2013) conference.

This work enlightened the fact that interactive in-situ visualization, although greatly improved by Damaris/Viz, still lakes interactivity. Several meetings were organized with Tom Peterka (ANL) and Roberto Sisneros (NCSA) during the SC conference and during the 10th workshop of the JLPC. We started working on an approach that leverages information theory metrics to automatically find important features of the simulations' data and to reduce the visualization load accordingly.

6.4.2. CALCioM

Participants: Matthieu Dorier, Gabriel Antoniu.

Unmatched computation and storage performance in new HPC systems have led to a plethora of I/O optimizations ranging from application-side collective I/O to network and disk-level request scheduling on the file system side. As we deal with ever larger machines, the interference produced by multiple applications accessing a shared parallel file system in a concurrent manner becomes a major problem. Interference often breaks single-application I/O optimizations, dramatically degrading application I/O performance and, as a result, lowering machine wide efficiency.

Following discussions initiated in 2012 with ANL's Rob Ross and Dries Kimpe, a three month internship of Matthieu Dorier at Argonne National Lab during the summer 2013 led to the design and evaluation of CALCioM (Cross-Application Layer for Coordinated I/O Management), a framework that aims to mitigate I/O interference through the dynamic selection of appropriate scheduling policies. CALCioM allows several applications running on a supercomputer to communicate and coordinate their I/O strategy in order to avoid interfering with one another. Several I/O strategies were evaluated using this framework. Experiments on Argonne's BG/P Surveyor machine and on several clusters of Grid'5000 showed how CALCioM can be used to efficiently and transparently improve the scheduling strategy between several otherwise interfering applications, given specified metrics of machine wide efficiency.

Future work will explore approaches to automatically detect the temporal I/O patterns of simulations in order to further improve the scheduling decisions made by CALCioM.

6.4.3. Scalable metadata management for WAN

Participants: Rohit Saxena, Alexandru Costan, Gabriel Antoniu.

BlobSeer-WAN is a data management service specifically optimized for geographically distributed environments. It is an extension of BlobSeer, a large scale data management service. The metadata is replicated asynchronously for low latency. There is a version manager on each site and vector clocks are used to enable collision detection and resolution under highly concurrent access. It was developed within the framework of Viet-Trung Tran' s PhD thesis, in relation to the FP3C project.

BlobSeer-WAN is used as a storage backend for HGMDS, a multi master metadata server designed for a global distributed file system, developed at University of Tsukuba. Several experiments have been conducted with this setup on the Grid'5000 testbed which have shown scalable metadata performance under geographically distributed environments.

7. Bilateral Contracts and Grants with Industry

7.1. 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 (see below). 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 Gabriel Antoniu. The project will also provide funding for a PhD thesis to start in 2014. It is being conducted in collaboration with the Zenith team from Montpellier (led by Patrick Valduriez).

Microsoft: A-Brain (2010–2013). In the framework of the Joint Inria-Microsoft Research Center. See details in Section 4.1. To support this project, Microsoft provided 2 million computation hours on the Azure platform and 10 TB of storage per year. The project funded a complementary expertise mission for Radu Tudoran (*Mission complémentaire d'expertise*, 3 years, started in October 2011).

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

MapReduce (2010–2014). An ANR project (ARPEGE 2010) with international partners, which focuses on optimized Map-Reduce data processing on cloud platforms. This project started in October 2010 in collaboration with Argonne National Lab, the University of Illinois at Urbana Champaign, the UIUC/Inria Joint Lab on Petascale Computing, IBM, IBCP, MEDIT and the GRAAL Inria Project-Team. URL: http://mapreduce.inria.fr/.

8.1.2. Other National projects

HEMERA (2010–2014). An Inria Large Wingspan Project, started in 2010. Within Hemera, G. Antoniu (KerData Inria Team) and Gilles Fedak (GRAAL Inria Project-Team) co-lead the Map-Reduce scientific challenge.

KerData also co-initiated a working group called *Efficient management of very large volumes of information for data-intensive applications*, co-led by G. Antoniu and Jean-Marc Pierson (IRIT, Toulouse).

Grid'5000. We are members of the Grid'5000 community: we make experiments on the Grid'5000 platform on a daily basis.

8.2. European Initiatives

8.2.1. FP7 Projects

The SCALUS FP7 Marie Curie Initial Training Network (2009–2013). Coordinator: André Brinkmann. Partners: Universidad Politécnica de Madrid (Spain), Barcelona Supercomputing Center (Spain), University of Paderborn (Germany), Ruprecht-Karls-Universität Heidelberg (Germany), Durham University (United Kingdom), FORTH (Greece), École des Mines de Nantes (France), XLAB (Slovenia), CERN (Switzerland), NEC (Germany), Microsoft Research (United Kingdom), Fujitsu (Germany), Sun Microsystems (Germany). Topic: scalable distributed storage. Abstract: The consortium of this Marie Curie Initial Training Network (MCITN) "SCALing by means of Ubiquitous Storage (SCALUS)" aims at elevating education, research, and development inside this exciting area with a focus on cluster, grid, and cloud storage. The vision of this MCITN is to deliver the foundation for ubiquitous storage systems, which can be scaled in arbitrary directions (capacity, performance, distance, security). We mainly collaborate with UPM (2 co-advised PhD theses).

8.3. International Initiatives

8.3.1. Inria Associate Teams

8.3.1.1. DATA@EXASCALE

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

Inria principal investigator: Gabriel Antoniu

International Partners (Institution - Laboratory - Researcher):

> Argonne National Laboratory (United States) - Mathematics and Computer Science Division - Rob Ross
>
> University of Illinois at Urbana Champaign (United States) - Marc Snir

Duration: 2013 - 2015

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

Description: 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 becomes clearly 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-UIUC Lab for Petascale Computing, recently extended to Argonne National Lab. 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.

8.3.2. Inria International Partners

8.3.2.1. Declared Inria International Partners

Politehnica University of Bucharest (since 1 January 2013, just after the end of our former Data-Cloud@work Associate Team).

8.3.3. Inria International Labs

Joint Inria-UIUC Lab for Petascale Computing (JLPC), since 2009. Collaboration on concurrency-optimized I/O for post-Petascale platforms (see details in Section 4.1). A joint project proposal with the team of Rob Ross (Argonne National Lab) has been completed in 2012. It served to prepare the creation of the Data@Exascale Associate Team with ANL and UIUC (2013-2015).

8.3.4. Participation In other International Programs

FP3C ANR-JST project (2010–2014). This project co-funded by ANR and by JST (Japan Science and Technology Agency) started in October 2010 for 42 months. It focuses on programming issues for Post-Petascale architectures. In this framework, KerData collaborates with the University of Tsukuba on data management issues.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

- Maria S. Perez (Universidad Politecnica de Madrid) and Toni Cortes (Universitat Politecnica de Catalunya) visited the KerData team for three days (December 2013) within the framework of the SCALUS project.

8.4.2. Internships

Participant: Mihaela Catalina Nita.

Subject: Smart Data Management for High-Performance Supercomputing

Date: from March 2013 until July 2013

Institution: Politehnica University of Bucharest (Romania)

Participant: Ana-Ruxandra Ion.

Subject: Enabling Map-Reduce-based Data-intensive Processing on Hybrid Cloud/Desktop Grid infrastructures

Date: from Mar 2013 until Jul 2013

Institution: Politehnica University of Bucharest (Romania)

Participant: Yue Li.

Subject: Energy Measurements for Cassandra Cloud Storage System: Exploring and improving Energy-Consistency Tradeoff

Date: from Feb 2013 until June 2013

Institution: Master student from Telecom Bretagne, Rennes (France)

Participant: Rui Wang.

Subject: Designing An Environment-Aware System for Geographically Distributed Data Transfers on Public Clouds

Date: from Feb 2013 until August 2013

Institution: Master student from Telecom Bretagne, Rennes (France)

8.4.3. Visits to International Teams

- Radu Tudoran visited ANL (Kate Keahey) for 3 months, funded by the Data@Exascale Associate Team.
- Matthieu Dorier visited ANL (Rob Ross, Tom Peterka, Phil Carns) for 2 months, funded by the Data@Exascale Associate Team.
- Radu Tudoran visited the ATL Lab at European Microsoft Innovation Center (Munich Germany) for 3 months, funded by Microsoft.

9. Dissemination

9.1. Scientific Animation

9.1.1. Gabriel Antoniu

- Program Co-Chair of the ScienceCloud 2013 International workshop held in conjunction with the ACM HPDC 2013 conference, New York, June 2013.
- Program Committee member (selection): ACM HPDC 2013 and 2014, ACM/IEEE SC 2013, IEEE/ACM CCGRID 2013, IEEE Big Data 2013, IEEE HPCC 2013, BDMC 2013 (workshop organized in conjunction with Euro-Par 2013).
- Coordinator for the MapReduce ANR project (see Section 8.1).

- G. Antoniu and B. Thirion (PARIETAL Project-Team, INRIA SACLAY – ÎLE-DE-FRANCE) co-led the A-Brain Microsoft-Inria Project (2010-2013).
- G. Antoniu and P. Valduriez (Zenith Project-Team, INRIA SOPHIA-ANTIPOLIS – MÉDITERRANÉE) co-lead the Z-CloudFlow Microsoft-Inria Project (2013-2016).
- Coordinator for the Data@Exascale Associate Team in collaboration with Argonne National Lab and the University of Illinois at Urbana - Champaign (2013-2015).
- Local coordinator for Inria Rennes – Bretagne Atlantique Research Center in the SCALUS Project of the Marie-Curie Initial Training Networks Programme (ITN), call FP7-PEOPLE-ITN-2008 (2009-2013).

9.1.2. Luc Bougé

- Vice-Chair of the Euro-Par conference steering committee (http://www.europar.org/). Workshop co-chair for the 2013 issue of the conference held in Aachen, Germany.
- PhD Forum Co-Chair for the 2013 IPDPS conference (http://www.ipdps.org/) held in Boston, MA, USA.

9.1.3. Shadi Ibrahim

- Program Committee member: ICPP 2013, ICA3PP 2013, CloudCom 2013, NPC 2013, CCGrid 2013 Doctoral Symposium, CLOUD COMPUTING 2013, BigData Book 2013
- Reviewer: IEEE Transactions on Cloud Computing, Future Generation Computer Systems, IEEE Systems Journal, Computing, Cluster Computing, Multimedia Tools and Applications Journal
- Subreviewer: SC 2013, HPDC 2013, CCGrid 2013, Cluster 2013, Euro-par 2013, HPCC 2013, ICPADS 2013, BigDataCloud 2013.

9.1.4. Alexandru Costan

- Organizer of the 2nd Workshop on Big Data Management in Clouds BigDataCloud in conjunction with EuroPar 2013, see http://www.irisa.fr/kerdata/bigdatacloud/
- Program Committee member: CloudCom 2013, ISPDC 2013, CCGrid Doctoral Symposium 2013, BigData Book 2013
- Reviewer: Transactions on Parallel and Distributed Systems, J. of Parallel and Distributed Computing, IEEE Internet Computing, Concurrency and Computation: Practice and Experience, Intl. J. of Grid and Utility Computing, Intl. J. of Systems and Software, Intl. J. of Intelligent Systems Technologies and Application
- Subreviewer: HPDC 2013, CCGrid 2013, ICPADS 2013, HPCC 2013, Cluster 2013

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

9.2.1.1. 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: Grid, P2P and cloud data management, 18 hours (lectures), M2 level, ALMA Master, Distributed Architectures module, University of Nantes, France.

Master: Peer-to-Peer Applications and Systems, 10 hours (lectures), M2 level, PAP Module, M2RI Master Program, ENS Cachan/Rennes, France.

9.2.1.2. Luc Bougé

Licence: Introduction to programming concepts, 24 hours (lectures), L3 level, Informatics program, ENS Cachan/Rennes, France.

Master: Introduction to object-oriented high-performance programming, 24 hours (lectures), M1 level, Mathematics program, ENS Cachan/Rennes, France.

9.2.1.3. Shadi Ibrahim

Master : Cloud1 — Map/Reduce, 10 hours (lectures, lab sessions), M2 Level, École des Mines de Nantes, Nantes, France.

9.2.1.4. Alexandru Costan

License: Object-oriented programming, 18 hours, L3, ENS Cachan/Rennes

License: Databases, 28 hours, L2, INSA Rennes, France

License: Practical case studies, 16 hours, L3, INSA Rennes

Master: Object-oriented design, 28 hours, M1, INSA Rennes

9.2.1.5. Matthieu Dorier

Licence: Java programming 35 hours (lectures, lab sessions), L1 level, INSA de Rennes.

Licence: Ruby Programming, 15 hours (lectures, lab sessions), L3 level, ENS Cachan/Rennes.

Licence: Computer Architectures, 24 hours (lab sessions), L3 level, ENS Cachan/Rennes.

9.2.2. Supervision

PhD: Viet-Trung Tran, “Scalable data-management systems for Big Data”, thesis started in October 2009 co-advised by Gabriel Antoniu and Luc Bougé. Date of defense: 21 January 2013.

PhD: Houssem Chihoub, “Managing consistency for Big Data Applications: Tradeoffs and Self-Adaptiveness”, thesis started in October 2010 co-advised by Maria Pérez (UPM, Madrid) and Gabriel Antoniu. Date of defense: 10 December 2013.

PhD in progress : Matthieu Dorier, “Scalable I/O for postpetascale HPC systems”, thesis started in October 2011 co-advised by Gabriel Antoniu and Luc Bougé.

PhD in progress : Radu Tudoran, “Scalable data sharing for Azure clouds”, thesis started in October 2011 co-advised by Gabriel Antoniu and Luc Bougé.

PhD in progress : Álvaro García Recuero, “Scalable, Power-efficient Big Data Analysis on Geographically Distributed Clouds”, thesis started in October 2013 co-advised by Shadi Ibrahim and Gabriel Antoniu.

PhD in progress : 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.

9.2.3. Juries

Gabriel Antoniu served as a Referee for a PhD Jury at the University Pierre et Marie Curie, Paris; as a Member in a PhD Jury at Telecom Bretagne, Cesson-Sévigné.

Luc Bougé served as a jury member for several PhD and HDR defenses, in many cases as the jury chairman.

9.3. Miscellaneous

- Gabriel Antoniu serves as a member of Inria’s Evaluation Committee.
- Luc Bougé served as Head of the Computer Science Department of ENS Cachan/Rennes until September 2013.
- Luc Bougé is serving as a scientific co-ordinator for the IT Department (STIC) of the French *National Research Agency* (ANR), for 40 % of his working time. He is in charge of the logistic and the scientific management of the ANR call for proposals and of the various scientific ANR exhibitions and prizes in the IT domain.
- Luc Bougé is a member of the *National Academic Board* (CNU, *Conseil national des universités*) in the Informatics Committee (Section 27).

9.3.1. Popularization

9.3.1.1. Radu Tudoran's talks

Microsoft TechDays 2013, Paris, France. Invited Speaker together with Gabriel Antoniu to give a presentation about *Azure Brain: 4th paradigm, scientific discovery and (really) big data*.

Microsoft TechDays 2013, Paris, France. Invited Speaker to give a short talk about using the Azure cloud for science within Session *Introducing 10 Windows Azure projects in one hour*.

Microsoft ATLE, Munich, Germany. Internal Presentation about *Streaming and processing events across cloud data-centers*.

10. Bibliography

Major publications by the team in recent years

[1] G. Antoniu, L. Cudennec, M. Jan, M. Duigou. *Performance scalability of the JXTA P2P framework*, in "Proc. IEEE International Parallel and Distributed Processing Symposium (IPDPS 2007)", Long Beach, USA, 2007, 108 p. , http://hal.inria.fr/inria-00178653/en/

[2] G. Antoniu, J.-F. Deverge, S. Monnet. *How to bring together fault tolerance and data consistency to enable grid data sharing*, in "Concurrency and Computation: Practice and Experience", 2006, n^{o} 17, pp. 1-19, http://hal.inria.fr/inria-00000987/en/

[3] A. Costan, R. Tudoran, G. Antoniu, G. Brasche. *TomusBlobs: Scalable Data-intensive Processing on Azure Clouds*, in "Concurrency and Computation Practice and Experience", 2013, To appear, http://hal.inria.fr/hal-00767034

[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, http://hal.inria.fr/hal-00715252

[5] R. Morales, S. Monnet, I. Gupta, G. Antoniu. *MOve:Design and Evaluation of A Malleable Overlay for Group-Based Applications*, in "IEEE Transactions on Network and Service Management, Special Issue on Self-Management", 2007, vol. 4, pp. 107-116 [*DOI :* 10.1109/TNSM.2007.070903], http://hal.inria.fr/inria-00446067/en/

[6] B. Nicolae, G. Antoniu, L. Bougé, D. Moise, A. Carpen-Amarie. *BlobSeer: Next Generation Data Management for Large Scale Infrastructures*, in "Journal of Parallel and Distributed Computing", February 2011, vol. 71, n^{o} 2, pp. 169-184, Special issue on data intensive computing, http://hal.inria.fr/inria-00511414/en/

[7] B. Nicolae, J. Bresnahan, K. Keahey, G. Antoniu. *Going Back and Forth: Efficient Multi-Deployment and Multi-Snapshotting on Clouds*, in "The 20th International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2011)", San José, CA, United States, June 2011, Selection rate: 12.9%, http://hal.inria.fr/inria-00570682/en

[8] 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, IEEE and ACM, Apr 2010, A preliminary version of this paper has been published as Inria Research Report RR-7140, http://hal.inria.fr/inria-00456801

[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 [*DOI :* 10.1145/2146382.2146387], http://hal.inria.fr/hal-00640900

[10] R. TUDORAN, A. COSTAN, G. ANTONIU, H. SONCU. *TomusBlobs: Towards Communication-Efficient Storage for MapReduce Applications in Azure*, in "12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'2012)", Ottawa, Canada, 2012, A-Brain project, Inria-Microsoft Research Joint Centre, http://hal.inria.fr/hal-00670725

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] H.-E. CHIHOUB. , *Managing Consistency for Big Data Applications on Clouds: Tradeoffs and Self Adaptiveness*, École normale supérieure de Cachan - ENS Cachan, December 2013, http://hal.inria.fr/tel-00915091

[12] V.-T. TRAN. , *Scalable data-management systems for Big Data*, École normale supérieure de Cachan - ENS Cachan, January 2013, http://hal.inria.fr/tel-00920432

[13] V.-T. TRAN. , *Sur le passage à l'échelle des systèmes de gestion des grandes masses de données*, École normale supérieure de Cachan - ENS Cachan, January 2013, http://hal.inria.fr/tel-00783724

Articles in International Peer-Reviewed Journals

[14] G. ANTONIU, J. BIGOT, C. BLANCHET, L. BOUGÉ, F. BRIANT, F. CAPPELLO, A. COSTAN, F. DESPREZ, G. FEDAK, S. GAULT, K. KEAHEY, B. NICOLAE, C. PÉREZ, A. SIMONET, F. SUTER, B. TANG, R. TERREUX. *Towards Scalable Data Management for Map-Reduce-based Data-Intensive Applications on Cloud and Hybrid Infrastructures*, in "International Journal of Cloud Computing (IJCC)", 2013, vol. 2, n^{o} 2/3 [*DOI :* 10.1504/IJCC.2013.055265], http://hal.inria.fr/hal-00767029

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

[16] S. IBRAHIM, H. JIN, L. LU, B. HE, G. ANTONIU, S. WU. *Handling Partitioning Skew in MapReduce using LEEN*, in "Peer-to-Peer Networking and Applications", 2013, http://hal.inria.fr/hal-00822973

[17] H. JIN, X. LING, S. IBRAHIM, S. WU, W. CAO, G. ANTONIU. *Flubber: Two-level Disk Scheduling in Virtualized Environment*, in "Future Generation Computer Systems", 2013, http://hal.inria.fr/hal-00784889

International Conferences with Proceedings

[18] G. ANTONIU, T. BOKU, C. CALVIN, P. CODOGNET, M. DAYDE, N. EMAD, Y. ISHIKAWA, S. MATSUOKA, K. NAKAJIMA, H. NAKASHIMA, R. NAMYST, S. PETITON, T. SAKURAI, M. SATO. *Towards exascale with the ANR-JST japanese-french project FP3C (Framework and Programming for Post- Petascale Computing)*, in "9th International Conference on Computer Science and Information Technologies", Yerevan, Armenia, 2013, http://hal.inria.fr/hal-00922754

[19] H.-E. Chihoub. *Self-Adaptive Cost-Efficient Consistency Management in the Cloud*, in "27th IEEE International Parallel & Distributed Processing Symposium IPDPS 2013 PhD Forum", Boston, United States, May 2013, http://hal.inria.fr/hal-00823664

[20] H.-E. Chihoub, S. Ibrahim, G. Antoniu, M. Pérez. *Consistency in the Cloud:When Money Does Matter!*, in "CCGRID 2013- 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing", Delft, Netherlands, May 2013, http://hal.inria.fr/hal-00789013

[21] 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, http://hal.inria.fr/hal-00916091

[22] M. Dorier. *Efficient I/O using Dedicated Cores in Large-Scale HPC Simulations*, in "IPDPS PhD forum - International Parallel and Distributed Processing Symposium, PhD forum", Boston, MA, United States, IEEE, May 2013 [*DOI :* 10.1109/IPDPSW.2013.101], http://hal.inria.fr/hal-00831296

[23] M. Dorier, R. Sisneros, T. Peterka, G. Antoniu, D. Semeraro. *Damaris/Viz: a Nonintrusive, Adaptable and User-Friendly In Situ Visualization Framework*, in "LDAV - IEEE Symposium on Large-Scale Data Analysis and Visualization", Atlanta, United States, October 2013, http://hal.inria.fr/hal-00859603

[24] X. Ling, S. Ibrahim, H. Jin, S. Wu, T. Songqiao. *Exploiting Spatial Locality to Improve Disk Efficiency in Virtualized Environments*, in "IEEE 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (Mascots 2013)", San Francisco, United States, July 2013, http://hal.inria.fr/hal-00842076

[25] R. Tudoran, A. Costan, G. Antoniu. *DataSteward: Using Dedicated Compute Nodes for Scalable Data Management on Public Clouds*, in "Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications", Melbourne, Australia, IEEE, July 2013, pp. 1057–1064 [*DOI :* 10.1109/TrustCom.2013.129], http://hal.inria.fr/hal-00927283

[26] R. Tudoran, A. Costan, R. Ramin Rezai, G. Brasche, G. Antoniu. *Adaptive File Management for Scientific Workflows on the Azure Cloud*, in "IEEE Big Data", Santa Clara, United States, IEEE, October 2013, pp. 273 - 281, http://hal.inria.fr/hal-00926748

Scientific Books (or Scientific Book chapters)

[27] H.-E. Chihoub, S. Ibrahim, G. Antoniu, M. Pérez. *Consistency Management in Cloud Storage Systems*, in "Advances in data processing techniques in the era of Big Data", S. Sherif, G. Mohamed Medhat (editors), CRC Press, 2013, http://hal.inria.fr/hal-00784885

Research Reports

[28] H.-E. Chihoub, M. Pérez, G. Antoniu, L. Bougé. , *Chameleon: Customized Application-Specific Consistency by means of Behavior Modeling*, October 2013, http://hal.inria.fr/hal-00875947

[29] M. Dorier, R. Sisneros, T. Peterka, G. Antoniu, D. Semeraro. , *A Nonintrusive, Adaptable and User-Friendly In Situ Visualization Framework*, Inria, June 2013, n^{o} RR-8314, 26 p. , http://hal.inria.fr/hal-00831265

References in notes

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

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

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

[33] G. H. BRYAN, J. M. FRITSCH. *A Benchmark Simulation for Moist Nonhydrostatic Numerical Models*, in "Monthly Weather Review", 2002, vol. 130, n^{o} 12, pp. 2917–2928, http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(2002)130<2917:ABSFMN>2.0.CO;2

[34] 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

[35] P. F. FISCHER, J. W. LOTTES, S. G. KERKEMEIER. , *nek5000 Web page*, 2008, http://nek5000.mcs.anl.gov

[36] 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, http://hal.inria.fr/inria-00456801/en/