IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

Ecole normale supérieure de Cachan

Activity Report 2012

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 is associated with the IRISA CNRS joint laboratory, University Rennes 1, INSA Rennes and ENS Cachan/Rennes.

Creation of the Project-Team: July 01, 2012 .

1. Members

Research Scientist

Gabriel Antoniu [Team leader, Senior Researcher (DR2) Inria., HdR]

Faculty Members

Luc Bougé [Professor, ENS Cachan/Rennes., HdR]

Alexandru Costan [Associate Professor, INSA Rennes, since September 1, 2012. Before this date, he was an Inria Post-Doctoral Fellow until August 31, 2012, supported by the MapReduce ANR project.]

Engineer

Zhe Li [Research Engineer, fixed-term contract, funded through the BlobSeer Technology Development Action (ADT) since November 2012.]

PhD Students

Viet-Trung Tran [MESR Grant until September 2012, and the ACET Inria. PhD to be defended on January 21, 2013. PhD thesis started in October 2009.]

Houssem-Eddine Chihoub [SCALUS FP7 Marie-Curie Initial Training Network Grant. PhD thesis started in October 2010.]

Radu Tudoran [MESR Grant from University Rennes 1. PhD thesis started in October 2011]

Matthieu Dorier [MESR Grant from ENS Cachan. PhD thesis started in October 2011]

Post-Doctoral Fellows

Louis-Claude Canon [Post-Doctoral Fellow, funded by the Microsoft Research-Inria A-Brain project since October 2011.]

Shadi Ibrahim [Inria Post-Doctoral Fellow, since November 2012. Before this date, he was supported by the Hemera Large Wingspan Project since October 2011.]

Visiting Scientists

Bunjamin Memishi

Elena Apostol

Bharath Vissapragada

Administrative Assistants

Céline Gharsalli [Team Administrative Assistant, CNRS, until September 2012.]

Élodie Lequoc [Team Administrative Assistant, University Rennes 1, since September 2012.]

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

- The KerData project-team has been officially created on July 1st 2012 as a Joint Project-Team with ENS Cachan/Brittany and INSA Rennes, for a 4-year term.
- Alexandru Costan, a former Post-Doc fellow at the KerData project-team, has been hired on a permanent position at INSA. Alexandru got his PhD in Valentin Cristea's NCIT group at Polytechnic University of Bucharest (Romania), our partner in the *DataCloud@work* Inria Associate Team.
- The KerData project-team organized the 7th Workshop of the Inria-Illinois Joint Laboratory on Petascale Computing, June 13-15, 2012 http://jointlab.ncsa.illinois.edu/events/workshop7/.
- After successful experiments with up to 9000 cores on the Kraken Cray XT5 machine (NICS) in 2011, Damaris scaled up to 16000 cores on Oak Ridge's leadership supercomputer Titan (now first in the Top500), providing in-situ analysis to the CM1 tornado simulation.

3. Scientific Foundations

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 [32] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [28]. At the core of the Map-Reduce frameworks stays 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 highlights a major, more general research topic. It has been clearly identified by international panels of experts like IESP [30] and EESI [29], 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 infrastructure and methodology 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, hopefully soon, 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 this approach in preliminary experiments [37], we have initiated several collaborative projects led by KerData in the area of cloud data management, e.g., the MapReduce ANR project, the A-Brain 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 those projects, we have already started to enjoy a visible scientific positioning at the international level.

The particularly active DataCloud@work 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 6.5.

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 the 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. Thus, an 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 that 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 it off-site, reading it into a workflow, and analyzing it. It 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 when it comes online in 2012.

In this context, the KerData project-team started to explore ways to remove the limitations mentioned above through a 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 approaches, which typically consists in periodically writing a very large number of small files. This causes burst 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 those challenges through an approach based on dedicated I/O cores.

5. Software

5.1. BlobSeer

Participants: Viet-Trung Tran, Zhe Li, 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) in the OpenNebula IaaS cloud environment 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 new *Technology Research Action* (ADT, *Action de recherche technologique*) has been launched in Septembre 2012 for one year, with a possible 1-year renewal, to robustify the BlobSeer software and and make it a safeky distributable product. This project is funded by Inria *Technological Development Office* (D2T, *Direction du Développement Technologique*). Zhe Li has been hired as a senior (PhD) engineer for this task.

5.2. Damaris

Participants: Matthieu Dorier, 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. On-going work is targeting fast direct access to the data from running simulations, and efficient I/O scheduling.

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 developed by 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. Registration with APP is in progress.

5.3. Derived software

Derived from BlobSeer, two additional platforms are currently being developed within KerData: 1) Pyramid, a software service for array-oriented active storage developed within the framework of the PhD thesis of Viet-Trung Tran; and 2) BlobSeer-WAN, a data management service specifically optimized for geographically distributed environments. It is also developed within the framework of the PhD thesis of Viet-Trung Tran in relation to the FP3C project. These platforms have not been publicly released yet.

6. New Results

6.1. Optimizing MapReduce processing

6.1.1. Hybrid infrastructures

Participants: Alexandru Costan, Bharath Vissapragada, 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 stays 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.

In 2012 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.) ([16]). 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 Bharath Vissapragada was dedicated to this topic.

This approach will be evaluated with real-life bio-informatics applications on existing Nimbus-powered cloud testbeds interconnected with desktop grids.

6.1.2. Scheduling: Maestro

Participants: Shadi Ibrahim, Gabriel Antoniu.

As data-intensive applications became popular in the cloud, data-intensive cloud systems call for empirical evaluations and technical innovations. We have investigated some performance limits in current MapReduce frameworks (Hadoop in particular). Our studies reveal that the current Hadoop's scheduler for map tasks is inadequate, as it disregards replicas distributions. It causes performance degradation due to a high number of non-local map tasks, which in turn causes too many needless speculative map tasks and leads to imbalanced execution of map tasks among data nodes. We addressed these problems by developing a new map task scheduler called Maestro.

In [19], we developed a scheduling algorithm (Maestro) to alleviate the nonlocal map tasks executions problem of MapReduce. Maestro is conducive to improving the locality of map tasks executions efficiency by virtue of the finer-grained replica aware execution of map tasks, thereby having one additional factor for the chunk hosting status: the expected number of map tasks executions to be launched. Maestro keeps track of the chunks' locations along with their replicas' locations and the number of other chunks hosted by each node. In doing so, Maestro can efficiently schedule the map task to the node with minimal impacts on other nodes' local map tasks executions. Maestro schedules the map tasks in two waves: first, it fills the empty slots of each data node based on the number of hosted map tasks and on the replication scheme for their input data; second, runtime scheduling takes into account the probability of scheduling a map task on a given machine depending on the replicas of the task's input data. These two waves lead to a higher locality in the execution of map tasks and to a more balanced intermediate data distribution for the shuffling phase.

We evaluated Maestro through a set of experiments on the Grid'5000 [35] testbed. Preliminary results [19] show the efficiency and scalability of our proposals, as well as additional benefits brought forward by our approach.

6.1.3. Fault tolerance

Participants: Bunjamin Memishi, Shadi Ibrahim, Gabriel Antoniu.

The simple philosophy of MapReduce has made huge community interest for its exploration, especially in environments where data-intensive applications are primary concern. Fault tolerance is one of the key features of the MapReduce system. MapReduce tasks are re-executed in case of failure, and a potential failure of a single master causes an additional bottleneck. It is observed that the detection of the failed worker tasks in Hadoop have a certain delay, yet not solved. Willing to improve the applications performance and optimal resource utilization, both of this concerns were more than a motivation so that we show in [36] that a little attention has been devoted to the failure detection in Hadoop's MapReduce which currently uses a timeout based mechanism for detecting failed tasks.

We have performed an in-depth analysis of MapReduce's failure detection, and these preliminary studies have revealed that the current static timeout value (600 seconds) is not adequate and demonstrate significant variations in the application's response time with different timeout value. Moreover, in the presence of single machine failure, the applications latencies vary not only in accordance to the occupancy time of the failure, similar to [33], but also vary with the job length (short or long).

Based on our aforementioned micro-analysis of failure detection in MapReduce, we are currently investigating an adaptive failure detection mechanism for Hadoop, which basically addresses the timeout adjustment in real-time for different jobs and applications, so that finally to adjust this model into a Shared Hadoop Cluster. Another work should discuss in details different failures types in MapReduce system and survey the different mechanisms used in MapReduce for detecting, handling and recovering from these failures and their inherited pros and cons; additionally, to a particular interest will be the analyzing of different execution environments including Cluster, Cloud and Desktop Grid on the efficiency of fault-tolerance in MapReduce. This work will soon be published.

6.2. A-Brain and TomusBlobs

6.2.1. TomusBlobs

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Enabling high-throughput massive data processing on cloud data becomes a critical issue, as it impacts the overall application performance. In the framework of the MSR-Inria A-Brain co-led by Gabriel Antoniu (KerData) and Bertrand Thirion (PARIETAL), the TomusBlobs[22] system was designed and implemented by KerData to address such challenges at the level of the cloud storage. The system we introduce is a concurrency-optimized data storage system which federates the virtual disks associated to VMs. As TomusBlobs does not require modifications to the cloud middleware, it can serve as a high-throughput globally-shared data storage for the cloud applications that require data passing among computation nodes.

We leveraged the performance of this solution to enable efficient data-intensive processing on commercial clouds by building an optimized prototype MapReduce framework for Azure. The system, deployed on 350 cores in Azure, was used to execute a real-life application, A-Brain with the goal of searching for significant associations between brain locations and genes.

The achieved throughput increased with an order of 2 for reading, respectively 3 for writing compared to the remote storage. With our approach for MapReduce data processing, the computation time is reduced to 50 % compared to the existing solutions, while the cost is reduced up to 30 %.

6.2.2. Iterative MapReduce

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Louis-Claude Canon.

While MapReduce has arisen as a major programming model for data analysis on clouds, there are many scientific applications that require processing patterns different from this paradigm. As such, reduce-intensive algorithms are becoming increasingly useful in applications such as data clustering, classification and mining. These algorithms have a common pattern: data are processed iteratively and aggregated into a single final result. While in the initial MapReduce proposal the reduce phase was a simple aggregation function, recently an increasing number of applications relying on MapReduce exhibit a reduce-intensive pattern, that is, an important part of the computations are done during the reduce phase. However, platforms like MapReduce or Dryad lack built-in support for reduce-intensive workloads.

To overcome these issues, we introduced MapIterativeReduce [23], a framework which: 1) extends the MapReduce programming model to better support reduce-intensive applications by exploiting the inherent parallelism of the reduce tasks which have an associative and/or commutative operation; and 2) substantially improves their efficiency by eliminating the implicit barrier between the Map and the Reduce phase. We showed how to leverage this architecture for scientific applications by enhancing the fault tolerance support in Azure and TomusBlobs, the underlying storage system, with a light checkpointing scheme and without any centralized control.

We evaluated MapIterativeReduce on the Microsoft Azure cloud with synthetic benchmarks and with a real-life application. Compared to state-of-art solutions, our approach enables faster data processing, by reducing the execution times by up to 75 %.

6.2.3. Adaptive file management for clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Recently, there is an increasing interest to execute general data processing schemas in clouds, as it would allow many scientific applications to migrate to this computing infrastructures. The natural way to do this is to designe and adopt Workflow Processing engines built for clouds. Such workflow processing in clouds would involve data propagation on the computation nodes based on well defined data access patterns. Having an efficient file management backend for a workflow engines is thus essential as we move to the world of BigData.

We proposed a new approach for a transfer-optimized file management in clouds On the one hand, our solution manages files within the deployment leveraging data locality. On the other hand, we envision an adaptive system that adopts the transfer method most suited based on the data transfer context.

The performance evaluation showed significant gains in terms of transfer throughput and computation time. File transfer times are reduced up to a factor of 5 with respect to the remote storage, while the timespan of running applications is reduced by more than 25% compared with other frameworks like Hadoop on Azure. This work was done in the context of a 3-month internship of Radu Tudoran hosted by the Advance Technology Lab from Microsoft Europe, Germany, Aachen.

6.3. Autonomic Cloud data storage management

Participants: Gabriel Antoniu, Alexandru Costan.

Providing the users with the possibility to store and process data on externalized, virtual resources from the cloud requires simultaneously investigating important aspects related to security, efficiency and quality of service. To this purpose, it clearly becomes necessary to create mechanisms able to provide feedback about the state of the storage system along with the underlying physical infrastructure. This information thus monitored, can further be fed back into the storage system and used by self-managing engines, in order to enable an autonomic behavior, possibly with several goals such as self-configuration, self-optimization, or self-healing. Within the DataCloud@work Associate Team in partnership with Politehnica University of Bucharest, our goal was to bring substantial contributions in this direction by leveraging previous efforts materialized through the BlobSeer data-sharing platform and several large-scale applications.

6.3.1. Evaluating BlobSeer for sharing application data on IaaS cloud infrastructures

. We showed how several types of large scale applications (e.g. scientific data aggregation, context-aware data management, video and image processing) rely on BlobSeer's support for high concurrency and increased data access throughput in order to achieve their goals. Several building blocks were implemented to address all the applications' requirements (new meta-data management, extended clients). An illustrative class of applications is represented by the context-aware ones. Our goal was to provide a cloud-based storage layer for sensitive context data, collected from a vast amount of sources: from smartphones to sensors located in the environment. We developed a layer on top of BlobSeer to allow two major things: efficient access to data based on meta-information (a catalogue of context data), and the support from mobility in the form of distributed caches able to support the movement of people and give support for fast access to real-time event of interest (dissemination of events of interest). The system as a whole was evaluated in extensive experiments, involving thousands of simulated clients, and the results proved its valuable contribution to advance the current state-of-the-art in the area of interested (middlewares to support context-aware apps).

6.3.2. Fault-tolerant VM management in Clouds, using BlobSeer

. We were also concerned about the fault tolerance support for the aforementioned applications on the cloud. A first step towards this goal consisted in exploring ways to deploy, boot and terminate VMs very quickly, enabling cloud users to exploit elasticity to find the optimal trade-off between the computational needs (number of resources, usage time) and budget constraints. We built a VM management system based on the FUSE interface leveraging the high throughput under increased concurrency of BlobSeer. We integrated it within the Nimbus cloud to allow fast VM deployment / snapshotting/ live migration. An adaptive prefetching mechanism is used to reduce the time required to simultaneously boot a large number of VM instances on clouds from the same initial VM image (multi-deployment). This proposal does not require any foreknowledge of the exact access pattern. It dynamically adapts to it at run time, enabling the slower instances to learn from the experience of the faster ones. Since all booting instances typically access only a small part of the virtual image along almost the same pattern, the required data can be pre-fetched in the background. In parallel, we investigated ways to ensure the anonimity of the data management layer, a requirement for HPC applications deployed into the clouds.

6.4. Advanced techniques for scalable cloud storage

6.4.1. Adaptive consistency

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

In just a few years cloud computing has become a very popular paradigm and a business success story, with storage being one of the key features. To achieve high data availability, cloud storage services rely on replication. In this context, one major challenge is data consistency. In contrast to traditional approaches that are mostly based on strong consistency, many cloud storage services opt for weaker consistency models in order to achieve better availability and performance. This comes at the cost of a high probability of stale data being read, as the replicas involved in the reads may not always have the most recent write. In [17], we propose a novel approach, named Harmony, which adaptively tunes the consistency level at run-time according to the application requirements. The key idea behind Harmony is an intelligent estimation model of stale reads,

allowing to elastically scale up or down the number of replicas involved in read operations to maintain a low (possibly zero) tolerable fraction of stale reads. As a result, Harmony can meet the desired consistency of the applications while achieving good performance. We have implemented Harmony and performed extensive evaluations with the Cassandra cloud storage on Grid'5000 testbed and on Amazon EC2. The results show that Harmony can achieve good performance without exceeding the tolerated number of stale reads. For instance, in contrast to the static eventual consistency used in Cassandra, Harmony reduces the stale data being read by almost 80%. Meanwhile, it improves the throughput of the system by 45% while maintaining the desired consistency requirements of the applications when compared to the strong consistency model in Cassandra.

While most optimizations efforts for consistency management in the cloud focus on how to provide adequate trade-offs between consistency guarantees and performance, a little work has been investigating the impact of consistency on monetary cost. However, and since strict strong consistency is not always required for large class of applications, in [25] 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 allowing up to 31 % of money saving while tolerating a very small fraction of stale reads.

6.4.2. In-memory data management

Participants: Viet-Trung Tran, Gabriel Antoniu, Luc Bougé.

As a result of continuous innovation in hardware technology, computers are made more and more powerful than their prior models. Modern servers nowadays can possess large main memory capability that can size up to 1 Terabytes (TB) and more. As memory accesses are at least 100 times faster than disk, keeping data in main memory becomes an interesting design principle to increase the performance of data management systems. We design DStore [27], a document-oriented store residing in main memory to fully exploit high-speed memory accesses for high performance. DStore is able to scale up by increasing memory capability and the number of CPU-cores rather than scaling horizontally as in distributed data-management systems. This design decision favors DStore in supporting fast and atomic complex transactions, while maintaining high throughput for analytical processing (read-only accesses). This goal is (to our best knowledge) not easy to achieve with high performance in distributed environments.

To achieve its goals, DStore is built with several design principles. DStore follows a single threaded execution model to execute update transactions sequentially by one *master thread* while relying on a versioning concurrency control to enable multiple *reader threads* running simultaneously. DStore builds indexes for fast document lookups. Those indexes are built using the *delta-indexing* and *bulk updating* mechanisms for faster indexes maintenance and for atomicity guarantees of complex queries. Moreover, DStore is designed to favor stale reads that only need to access isolated snapshots of the indexes. Thus, it can eliminate interference between transactional processing and analytical processing.

We conducted multiple synthetic benchmarks on the Grid'5000 to evaluate the DStore prototype. Our preliminary results demonstrated that DStore achieved high performance even in scenarios where *Read*, *Insert* and *Delete* queries were performed simultaneously. In fact, the processing rate measured was about 600,000 operations per second for each concurrent process.

6.4.3. Scalable geographically distributed storage systems

Participants: Viet-Trung Tran, Gabriel Antoniu, Luc Bougé.

To build a globally scalable distributed file system that spreads over a wide area network (WAN), we propose an integrated architecture for a storage system relying on a distributed metadata-management system and BlobSeer, a large-scale data-management service. Since BlobSeer was initially designed to run on cluster environments, it is necessary to extend BlobSeer in order to take into account the latency hierarchy on geographically distributed environments.

We proposed BlobSeer-WAN, an extension of BlobSeer optimized for geographically distributed environments. First, in order to keep metadata I/O local to each site as much as possible, we proposed an asynchronous metadata replication scheme at the level of metadata providers. As metadata replication is asynchronous, we guarantee a minimal impact on the writing clients that generate metadata. Second, we introduced a distributed version management in BlobSeer-WAN by leveraging an implementation of multiple version managers and using vector clocks for detection and resolution of collision. This extension to BlobSeer keeps BLOBs consistent while they are globally shared among distributed sites under high concurrency.

Several experiments were performed on the Grid'5000 testbed demonstrated that BlobSeer-WAN can offer scalable aggregated throughput when concurrent clients append to one BLOB. The aggregated throughput reached to 1400 MB/s for 20 concurrent clients. We also compared BlobSeer-WAN and the original BlobSeer in local site accesses. The experiments shown that the overhead of the multiple version managers implementation and the metadata replication scheme in BlobSeer-WAN is minimal, thanks to our asynchronous replication scheme.

6.5. Scalable I/O for HPC

6.5.1. Damaris and HPC visualization

Participants: Matthieu Dorier, Gabriel Antoniu.

In the context of the Joint Inria/UIUC/ANL Laboratory for Petascale computing (JLCP), have proposed the Damaris approach to enable efficient I/O, data analysis and visualization at ver 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 storage or drastically reducing the amount of data generated will be of outmost importance for exascale. In-situ visualization has therefor been proposed to run analysis and visualization tasks closer to the simulation, as it runs.

The first results obtained with Damaris in achieving scalable, jitter-free I/O, were published this year [18]. In order to achieve efficient in-situ visualization at extreme scale, we investigated the limitations of existing in-situ visualization software and proposed to fill the gaps of these software by providing in-situ visualization support to Damaris. The use of Damaris 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 are now being conducted on BlueWaters (Cray XK6 at NCSA), Intrepid (BlueGene/P at ANL) and Grid5000 with representative visualization scenarios for the CM1 [31] atmospheric simulation and the Nek5000 [34] CFD solver.

Results will be submitted to a conference in early 2013. We plan to further investigate the role that Damaris can take in performing efficient and self-adaptive data analysis in HPC simulations.

6.5.2. Advanced I/O and Storage

Participants: Matthieu Dorier, Alexandru Costan, Gabriel Antoniu.

The recent extension of the JLPC to Argonne National Lab (ANL) has opened new research directions in the field of advanced I/O and storage for HPC, in collaboration with Robert Ross's team at ANL's Mathematics and Computer Science Division (MCS). A founding from the FACCTS program (France And Chicago CollaboraTing in Science) allowed multiple visits (see Section 8.4) of students and researchers from both sides to initiate this new collaboration and explore potential research directions.

One outcome of these visits has been the adaptation of Damaris to work on BlueGene/P and BlueGene/Q machines installed at ANL. Several exchanges led to the design of new I/O scheduling algorithms leveraging Damaris for efficient asynchronous I/O and storage. These algorithms are currently being evaluated, and expected to be published in early 2013.

During these exchanges we also investigated new storage architectures for Exascale systems leveraging BLOB-based large-scale storage able to cope with complex data models. We will explore how we can combine the benefits of the approaches to Big Data storage currently developed by the partners: the BlobSeer approach (KerData), which provides support for multi- versioning and efficient fine-grain access to huge data under heavy concurrency and the Triton approach (ANL), which introduces new object storage semantics. The final goal of the resulting architecture will be to propose efficient solutions to data-related bottlenecks in Exascale HPC systems.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

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 provides 2 million computation hours on the Azure platform and 10 TB of storage per year. The project is funding Louis-Claude Canon as a postdoc fellow (18 months since September 2011) and to complete the PhD MESR grant of Radu Tudoran (*Mission complémentaire d'expertise*, 3 years, started in October 2011).

IBM: MapReduce ANR Project (2010–2014). IBM is a partner of the MapReduce ANR Project: see Section 8.1.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

MapReduce (2010–2014). An ANR project (ARPEGE 2010) with international partners 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 an everyday basis.

8.2. European Initiatives

8.2.1. FP7 Projects

The SCALUS FP7 Marie Curie Initial Training Network (2009–2013). Partners: Universidad Politécnica de Madrid (UPM), Barcelona Supercomputing Center, University of Paderborn, Ruprecht-Karls-Universität Heidelberg, Durham University, FORTH, École des Mines de Nantes, XLAB, CERN, NEC, Microsoft Research, Fujitsu, Sun Microsystems. Topic: scalable distributed storage. We mainly collaborate with UPM (2 co-advised PhD theses).

8.2.2. Collaborations in European Programs, except FP7

CoreGRID ERCIM Working Group, since 2009. The CoreGRID Symposium held in Las Palmas de Gran Canaria, Spain, 25-26 August 2008 marked the end of the ERCIM-managed CoreGRID Network of Excellence funded by the European Commission. There, it was decided to re-launch CoreGRID as a self-sustained ERCIM Working Group covering research activities on both Grid and Service Computing while maintaining the momentum of the European collaboration on Grid research.

8.3. International Initiatives

8.3.1. Inria Associate Teams

8.3.1.1. DATACLOUD

Title: Distributed data management for cloud services

Inria principal investigator: Gabriel Antoniu

International Partner (Institution - Laboratory - Researcher):

Politechnica University of Bucharest (Romania) - NCIT - Valentin Cristea

Duration: 2010 - 2012

See also: http://www.irisa.fr/kerdata/doku.php?id=cloud_at_work:start

Our research topics address the area of distributed data management for cloud services. We aim at investigating several open issues related to autonomic storage in the context of cloud services. 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, while leveraging the advantages of the grid operating system approach.

Our research activities involve the design and implementation of experimental prototypes based on the following software platforms:

The BlobSeer data-sharing platform (designed by the KerData Team)

The XtreemOS grid operation system (designed under the leadership of the Myriads Team)

The MonALISA monitoring framework (using the expertise of the PUB Team).

The main results obtained in 2012 are described in Section 6.4.

8.3.2. Inria International Partners

Politehnica University of Bucharest

8.3.3. Participation In International Programs

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

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

- Robert Ross and Dried Kimpe (Argonne National Lab) visited the KerData team for a week (June 2012) within the framework of our FACCTS project.
- Florin Pop and Ciprian Dobre (Politehnica University of Bucharest) visited the KerData team for a week (June 2012) within the framework of our DataCloud@work Associate Team.

8.4.2. Internships

Elena Burceanu (from February 2012 until June 2012)

Subject: Distributed data storage for context-aware applications

Institution: Politehnica University of Bucharest (Romania)

Vlad Nicolae Serbanescu (from February 2012 until June 2012)

Subject: Distributed data aggregation using the BlobSeer cloud storage service

Institution: Politehnica University of Bucharest (Romania)

Bharath Vissapragada (from February 2012 until June 2012)

Subject: MapReduce data processing on hybrid (cloud/desktop grid) infrastructures

Institution: University of Hyderabad (India)

Mauricio De Oliveira de Diana (June 2012)

Subject: Performance modeling for the BlobSeer storage system

Institution: Master student from Brazil

Sergiu Vicol (June–August 2012)

Subject: Optimizing memory management in Damaris

Institution: Bachelor student from Oxford University. Former awardee of the ENS-Inria Excellence Award for the Laureates of the Romanian Olympiad in Informatics.

Alexandru Farcasanu(June–August 2012)

Subject: Optimizing the DStore in-memory storage system

Institution: Bachelor students from Politehnica University of Bucharest. Former awardee of the ENS-Inria Excellence Award for the Laureates of the Romanian Olympiad in Informatics.

8.4.3. Visits to International Teams

- Viet-Trung Tran visited Microsoft Research Cambridge (Dushyanth Narayanan) for a 3-month internship, funded by MSR.
- Houssem-Eddine Chihoub visited the Polytechnical University of Madrid (Maria Perez) for 3 months, funded by the FP7 SCALUS MCITN project.
- Radu Tudoran visited the ATL Lab at European Microsoft Innovation Center (Aaachen Germany) for 3 months, funded by Microsoft.
- Matthieu Dorier visited ANL (Rob Ross, Tom Peterka, Phil Carns) and UIUC (Franck Cappello) for one month, funded by our FACCTS grant.

9. Dissemination

9.1. Scientific Animation

Gabriel Antoniu:

- General Co-Chair of the ScienceCloud 2012 International workshop held in conjunction with the ACM HPDC 2012 conference.
- Local Chair of the 7th International Workshop of the Joint Inria-UIUC-ANL Lab for Petascale Computing, Rennes, June 2012.
- Track chair at IEEE CloudCom 2012 international conference.
- Editor for a Special Issue of Concurrency and Computation: Practice and Experience Journal on Cloud Computing for Data-driven Science and Engineering, 2012.
- Program Committee member (selection): ACM HPDC 2012, ACM/IEEE SC 2013, IEEE/ACM CCGRID 2013, ICCCN 2012, IEEE HPCC 2012, IEEE AINA 2012, IEEE CloudCom 2012, ICPADS 2012.
- Coordinator for the MapReduce ANR project (see Section 8.1).
- G. Antoniu and B. Thirion (PARIETAL Project-Team, INRIA SACLAY – ÎLE-DE-FRANCE) co-lead the AzureBrain Microsoft-Inria Project (2010-2013).
- Coordinator for the DataCloud@work Associate Team, a project involving the KerData and MYRIADS Inria Teams in Rennes and the Distributed Systems Group from Politehnica University of Bucharest (2010–2012).
- 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).

Alexandru Costan:

- Organizer of the 1st Workshop on Big Data Management in Clouds BDMC2012 in conjunction with EuroPar 2012, see http://www.irisa.fr/kerdata/bdmc/
- Program Committee member: CloudCom 2012, ScienceClouds 2012, EIDWT 2012
- Reviewer: J. of Parallel and Distributed Computing, IEEE Internet Computing, Concurrency and Computation: Practice and Experience, Intl. J. of Grid and Utility Computing, Innovative Studies Intl. J., Scalable Computing: Practice and Experience, Simulation Modelling Practice and Theory, HPDC 2012, AINA 2012, RenPar 2012

Luc Bougé:

- Since September 2012: Scientific coordinator for the Information & Communication Science & Technology Department of the National Research Agency (ANR).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Gabriel Antoniu:

Master (Engineering Degree, 5th year): Grid and cloud computing, 18 hours (lectures), M2 level, Ecole Supérieure d'Informatique, Electronique, Automatique, Paris, France.

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

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

Alexandru Costan

- Object-oriented programming, 18h, L3, ENS Cachan - Antenne de Bretagne
- Databases, 28h, L2, INSA Rennes, France
- Object-oriented design, 28h, M1, INSA de Rennes
- Practical case studies, 16h, L3, INSA de Rennes

Matthieu Dorier:

- Java programming (lectures, seminars, practical sessions), L1 level, INSA de Rennes, 56h.
- Programming techniques (seminars, practical sessions), L3 level, ENS Cachan - Antenne de Bretagne 42h

9.2.2. Supervision

PhD & HdR :

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 in progress : Houssem Chihoub, Consistency issues in cloud storage systems, thesis started in October 2010 co-advised by Maria Pérez (UPM - Madrid) and Gabriel Antoniu.

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.

9.2.3. Juries

Gabriel Antoniu served as a member of Inria's national Jury for hiring researchers (junior positions, confirmed positions and starting research positions).

Gabriel Antoniu served as a Referee and as a Chair for a PhD Jury at the University of Bordeaux; as a Chair for a PhD Jury at the University of Nantes.

9.2.4. Miscellaneous

Gabriel Antoniu serves as a member of Inria's Evaluation Committee.

Luc Bougé serves as Head of the Computer Science Department of ENS Cachan - Brittany.

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, 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, p. 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, p. 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° 2, p. 169-184, Special issue on data intensive computing. To appear, 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° 1, p. 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

Articles in International Peer-Reviewed Journals

[11] 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, To appear, http://hal.inria.fr/hal-00767029.

[12] G. ANTONIU, A. COSTAN, B. DA MOTA, B. THIRION, R. TUDORAN. *A-Brain: Using the Cloud to Understand the Impact of Genetic Variability on the Brain*, in "ERCIM News", April 2012, p. 21-22, http://hal.inria.fr/hal-00684384.

[13] A. CARPEN-AMARIE, A. COSTAN, C. LEORDEANU, C. BASESCU, G. ANTONIU. *Towards a Generic Security Framework for Cloud Data Management Environments*, in "International Journal of Distributed Systems and Technologies (IJDST), Special Issue on Security, Privacy and Trust", 2012, http://hal.inria.fr/hal-00670923.

[14] 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.

[15] 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° 1, p. 19-25 [*DOI :* 10.1145/2146382.2146387], http://hal.inria.fr/hal-00640900.

International Conferences with Proceedings

[16] 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 "1st International IBM Cloud Academy Conference - ICA CON 2012", Research Triangle Park, North Carolina, États-Unis, 2012, http://hal.inria.fr/hal-00684866.

[17] H.-E. CHIHOUB, S. IBRAHIM, G. ANTONIU, M. PÉREZ. *Harmony: Towards Automated Self-Adaptive Consistency in Cloud Storage*, in "2012 IEEE International Conference on Cluster Computing", Beijing, Chine, IEEE, September 2012, http://hal.inria.fr/hal-00734050.

[18] 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, Chine, IEEE, September 2012, http://hal.inria.fr/hal-00715252.

[19] S. IBRAHIM, H. JIN, L. LU, B. HE, G. ANTONIU, S. WU. *Maestro: Replica-Aware Map Scheduling for MapReduce*, in "The 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID'2012)", Ottawa, Canada, 2012, http://hal.inria.fr/hal-00670813.

[20] D. MOISE, G. ANTONIU, L. BOUGÉ. *On-the-fly Task Execution for Speeding Up Pipelined MapReduce*, in "Euro-Par - 18th International European Conference on Parallel and Distributed Computing - 2012", Rhodes Island, Grèce, August 2012, http://hal.inria.fr/hal-00706844.

[21] R. TUDORAN, A. COSTAN, G. ANTONIU, L. BOUGÉ. *A Performance Evaluation of Azure and Nimbus Clouds for Scientific Applications*, in "CloudCP 2012 – 2nd International Workshop on Cloud Computing Platforms, Held in conjunction with the ACM SIGOPS Eurosys 12 conference", Bern, Suisse, 2012, To appear, http://hal.inria.fr/hal-00677842.

[22] 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, http://hal.inria.fr/hal-00670725.

[23] R. TUDORAN, A. COSTAN, G. ANTONIU. *MapIterativeReduce: A Framework for Reduction-Intensive Data Processing on Azure Clouds*, in "Third International Workshop on MapReduce and its Applications (MAPREDUCE'12), held in conjunction with ACM HPDC'12", Delft, Pays-Bas, 2012, To appear, http://hal.inria.fr/hal-00684814.

Research Reports

[24] L.-C. CANON, G. ANTONIU. *Scheduling Associative Reductions with Homogeneous Costs when Overlapping Communications and Computations*, Inria, March 2012, n° RR-7898, http://hal.inria.fr/hal-00675964.

[25] H.-E. CHIHOUB, S. IBRAHIM, G. ANTONIU, M. PÉREZ. *Consistency in the Cloud:When Money Does Matter!*, Inria, November 2012, http://hal.inria.fr/hal-00756314.

[26] M. DORIER, G. ANTONIU, F. CAPPELLO, M. SNIR, L. ORF. *Damaris: Leveraging Multicore Parallelism to Mask I/O Jitter*, Inria, April 2012, n° RR-7706, 36, http://hal.inria.fr/inria-00614597.

[27] V.-T. TRAN, D. NARAYANAN, G. ANTONIU, L. BOUGÉ. *DStore: An in-memory document-oriented store*, Inria, December 2012, n° RR-8188, 24, http://hal.inria.fr/hal-00766219.

References in notes

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

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

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

[31] G. H. BRYAN, J. M. FRITSCH. *A Benchmark Simulation for Moist Nonhydrostatic Numerical Models*, in "Monthly Weather Review", 2002, vol. 130, n° 12, p. 2917–2928 [*DOI :* 10.1175/1520-0493(2002)130<2917:ABSFMN>2.0.CO;2], http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%282002%29130%3C2917%3AABSFMN%3E2.0.CO%3B2.

[32] J. DEAN, S. GHEMAWAT. *MapReduce: simplified data processing on large clusters*, in "Communications of the ACM", 2008, vol. 51, n° 1, p. 107–113.

[33] F. DINU, T. E. NG. *Understanding the effects and implications of compute node related failures in hadoop*, in "Proceedings of the 21st international symposium on High-Performance Parallel and Distributed Computing", New York, NY, USA, HPDC '12, ACM, 2012, p. 187–198, http://doi.acm.org/10.1145/2287076.2287108.

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

[35] Y. JÉGOU, S. LANTÉRI, J. LEDUC, M. NOREDINE, G. MORNET, R. NAMYST, P. PRIMET, B. QUETIER, O. RICHARD, E.-G. TALBI, T. IRÉA. *Grid'5000: a large scale and highly reconfigurable experimental Grid testbed*, in "International Journal of High Performance Computing Applications", November 2006, vol. 20, n° 4, p. 481-494.

[36] B. MEMISHI, M. PEREZ, G. ANTONIU. *Enhanced failure detection mechanism in MapReduce*, in "High Performance Computing and Simulation (HPCS), 2012 International Conference on", july 2012, p. 690 -692, http://dx.doi.org/10.1109/HPCSim.2012.6266995.

[37] 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/.