IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

Ecole normale supérieure de Cachan

Activity Report 2014

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

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]

Faculty Members

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

Engineers

Anirvan Basu [Inria, August–December 2014, granted by the EIT ICT Labs program]
Camelia-Elena Ciolac [Inria, August–December 2014, granted by the EIT ICT Labs program]
Loïc Cloatre [Inria, from February 2014, granted by the Inria ADT BlobSeer]
Rohit Saxena [Inria, until February 2014, granted by the ANR FP3C project]

PhD Students

Radu Tudoran [University Rennes 1, granted by ED Matisse/MESR until September 2014, and then granted by the ANR MapReduce project. PhD defended on December 10, 2014.]
Matthieu Dorier [ENS Rennes, granted by ENS Rennes/MESR. PhD defended on December 9, 2014.]
Álvaro García Recuero [Inria, granted by the Inria Large Wingspan Action Hemera]
Luis Eduardo Pineda Morales [Inria, granted by Microsoft Research Z-CloudFlow contract]
Lokman Rahmani [University Rennes 1, granted by ED Matisse/MESR]
Tien Dat Phan [University Rennes 1, granted by ED Matisse/MESR, from October 2014]
Orçun Yildiz [Inria, granted by the CORDI-S Program, from September 2014]

Visiting Scientists

Pierre Matri [Inria, predoctoral contract granted by the ANR MapReduce project, from October 2014]
Robert Ross [University Rennes 1, Invited professor, June 2014]

Administrative Assistant

Aurélie Patier [University Rennes 1]

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 long-term agenda of the Cloud Computing and Exascale HPC research communities.

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

Our research activities focus on 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., Map-Reduce-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).

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 [31] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [27]. 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 [30], EESI [28], ETP4HPC [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 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. Additionally, we are also looking at other kinds of infrastructures, 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, Barcelona Supercomputing Centre) and Industry (Microsoft, IBM).

The highly experimental nature of our research validation methodology should be stressed. Our approach relies on building prototypes and on validating them at a large scale on real testbeds and experimental platforms. We strongly rely on the Grid'5000 platform. Moreover, thanks to our projects and partnerships, we have access to reference software and physical infrastructures in the cloud area (Microsoft Azure, Amazon clouds, Nimbus clouds); in the post-Petascale HPC area we 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 advanced realistic platforms.

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

3.2. Our research agenda

Three typical application scenarios will be described in detail in the next section:

- Joint genetic and neuroimaging data analysis on Azure clouds;
- Structural protein analysis on Nimbus clouds;
- I/O intensive 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 address the aforementioned challenges efficiently, 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 two usage scenarios in more detail, which led to two more specific approaches: 1) Pyramid [35] (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 [34], 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 (JLPC) 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. 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. It has remained poorly understood so far. Both neuroimaging- and genetic-domain observations include a huge amount of variables (of the order of millions). Performing rigorous statistical analyses on such amounts of data is a major computational challenge that cannot be addressed with conventional computational techniques only. On the one hand, sophisticated regression techniques need to be used in order to perform significant analysis on these large datasets; on the other hand, the cost entailed by parameter optimization and statistical validation procedures (e.g. permutation tests) is very high.

The A-Brain (AzureBrain) Project was carried out within the Microsoft Research-Inria Joint Research Center. It was co-led by the KerData (Rennes) and Parietal (Saclay) Inria teams. They jointly address this computational problem using cloud related techniques on the Microsoft Azure cloud infrastructure. The two teams 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. This project is a typical multi-disciplinary Data Science project which serves as background for several on-going research activities.

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.

4.2. 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, using a cloud computing facility unlocks 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 update all derivative predictive models periodically with new experimental data. The applications in this complete chemo-proteomic vision address the identification of new druggable protein targets, and thereby the generation of new drug candidates.

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

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

5.1. Major Software

5.1.1. BlobSeer

Participants: Loïc Cloatre, 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*. This year we have mainly focused on the deployment in production of the BlobSeer software on IBM's cluster at Montpellier, in the context of the ANR MapReduce project. To this end, several bugs were solved, and several optimizations were brought to the communication layer of BlobSeer. To showcase the benefits of BlobSeer on this platform we focused on the Terasort benchmark. Currently, preliminary tests on Grid5000 with this benchmark show that BlobSeer performs better than HDFS for block sizes lower than 2 MB. We have also improved the continuous integration process of BlobSeer by deploying daily builds and automatic tests on Grid5000.

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

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

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

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

A *Technology Research Action* (ADT, *Action de recherche technologique*) 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.

5.1.2. Damaris

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

Contact: Gabriel Antoniu.

Presentation: Damaris is a middleware for multicore SMP nodes enabling them to handle data transfers for storage and visualization efficiently. The key idea is to dedicate one or a few cores of each SMP node to the application I/O. It is developed within the framework of a collaboration between KerData and the *Joint Laboratory for Petascale Computing* (JLPC). Damaris enables efficient asynchronous I/O, hiding all I/O related overheads such as data compression and post-processing, as well as direct (*in-situ*) interactive visualization of the generated data. Version 1.0 was released in November 2014 and enables other approaches such as the use of dedicated nodes instead of dedicated cores.

Users: Damaris has been preliminarily evaluated at NCSA/UIUC (Urbana-Champaign, IL, USA) with the CM1 tornado simulation code. CM1 is one of the target applications of the Blue Waters supercomputer in production at, in the framework of the Inria-UIUC-ANL Joint Lab (JLPC). Damaris now has external users, including (to our knowledge) visualization specialists from NCSA and researchers from the France/Brazil Associated research team on Parallel Computing (joint team between Inria/LIG Grenoble and the UFRGS in Brazil). Damaris has been successfully integrated into four large-scale simulations (CM1, OLAM, Nek5000, GTC).

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

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

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

5.2. New Software

5.2.1. Omnisc'IO

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

Contact: Matthieu Dorier

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

Users: Omnisc'IO is currently used only within the KerData team.

URL: http://omniscio.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.

5.2.2. Darshan-Web

Participants: Matthieu Dorier, Thomas Bouguet.

Contact: Matthieu Dorier

Presentation: Darshan-Web is a web interface for Darshan-Ruby, based on Ruby on Rails and AJAX technologies. It allows to navigate through many Darshan log files and display graphs on demand, directly on a web brother. A demo of Darshan-Web is available at http://darshan-web.irisa.fr/, which includes 2 months of logs from ANL's Intrepid supercomputer. The code of this demo is available and can be installed and used by the community.

Users: The KerData team is currently seeking potential users, in particular from Argonne National Laboratory, and will push the development further according to potential users' feedback.

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

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

Status: Prototype and demo available on demand.

5.2.3. JetStream

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Contact: Alexandru Costan

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

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

License: Microsoft Public License.

Status: Prototype and demo available.

5.2.4. *OverFlow*

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Contact: Alexandru Costan.

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

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

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

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

5.2.5. *iHadoop*

Participants: Tien Dat Phan, Shadi Ibrahim.

Contact: Shadi Ibrahim

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

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

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

Status: Available on Inria's forge. Registration with APP is in progress.

6. New Results

6.1. Highlights of the Year

IEEE Cluster 2014. The KerData Team had a leading role the organization of the IEEE Cluster 2014 conference, held in Madrid (22–26 September 2014): Gabriel Antoniu as PC Chair, Luc Bougé as Student Mentoring Program Chair, Alexandru Costan as Submissions Chair.

6.2. Data Management for Geographically Distributed Workflows

6.2.1. *OverFlow: a multi-site-aware framework for Big Data management*

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The global deployment of cloud datacenters is enabling large-scale scientific workflows to improve performance and deliver fast responses. This unprecedented geographical distribution of the computation coincides with an increase in the scale of the data handled by such applications, bringing new challenges related to the efficient data management across sites. High throughput, low latencies or cost-related trade-offs are just a few concerns for both cloud providers and users when it comes to handling data across datacenters, as shown in earlier evaluations [21]. Existing solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemes. In turn, workflow engines need to find ad-hoc substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability.

We tackle these problems by trying to understand to what extent the intra- and inter-datacenter transfers can impact the total makespan of cloud workflows. We advocate storing data on the compute nodes and transferring files between them directly, in order to exploit data locality and to avoid the overhead of interacting with a shared file system. Under these circumstances, we propose a file management service that enables high throughput through self-adaptive selection among multiple transfer strategies (e.g. FTP-based, BitTorrent-based, etc.). Next, we focus on the more general case of large-scale data dissemination across geographically distributed sites. The key idea is to predict I/O and transfer performance accurately and robustly in a dynamic cloud environment in order to decide judiciously how to perform transfer optimizations over federated datacenters: predict the best combination of protocol and transfer parameters (e.g., multi-routes, flow count, multicast enhancement, replication degree) to maximize throughput or minimize costs, according to users policies. We have implemented these principles in OverFlow, as part of the Azure Cloud so that applications could use it using a Software-as-a-Service (SaaS) approach.

OverFlow [20] was validated on the Microsoft cloud across the 6 EU and US sites. The experiments were conducted on hundreds of nodes using synthetic benchmarks and real-life bio-informatics applications (A-Brain, BLAST). The results show that our system is able to model the cloud performance accurately and to leverage this for efficient data dissemination, being able to reduce the monetary costs and transfer time by up to 3 times.

6.2.2. Metadata management for geographically distributed workflows

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

Scientific workflow data can reach sizes that exceed single-site capabilities. It is needed to support fine-grain data stripping to handle either very large files or very large sets of small files across data centers. Therefore, metadata becomes a critical issue. Moreover, workflow metadata provides crucial information to optimize data management, particularly in the context of geographically distributed data centers. Many present-day distributed file systems, such as GoogleFS and HDFS, include a potential bottleneck as the number of files grows, because they use a centralized metadata management scheme. Thus, we argue for a new, *cloud-based, distributed metadata management* scheme.

We have designed four different approaches to a geographically distributed metadata registry, namely: a) baseline centralized version; b) distributed on each data center with centralized replication agent; c) decentralized non-replicated; and d) decentralized replicated with hierarchical access. A comparative analysis showed that the later strategy performs best in terms of metadata operations per time unit. We then evaluate each of our approaches against various workflow benchmarks, with the purpose of dynamically adapt the metadata handling scheme according to the underlying application and cloud contexts. In the next phase, we will provide a uniform metadata handling tool for scientific workflow engines across cloud datacenters, as well as derive a cost model to offer users the best trade-off (performance vs. cost) driven by their constraints.

6.2.3. Transfer-as-a-Service: a cost-effective model for multi-site cloud data management

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Existing cloud data management solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemas. Users are therefore forced to design and deploy custom solutions, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability. In [19] we have proposed a dedicated cloud data-transfer service that supports largescale data dissemination across geographically distributed sites, advocating for a Transfer-as-a-Service (TaaS) paradigm. The system aggregates the available bandwidth by enabling multi-route transfers across cloud sites, based on the approach previously described.

We argue that the adoption of such a TaaS approach brings several benefits for both users and the cloud providers who propose it. For users of multi-site or federated clouds, our proposal is able to decrease the variability of transfers and increase the throughput up to three times compared to baseline user options, while benefiting from the well-known high availability of cloud-provided services. For cloud providers, such a service can decrease the energy consumption within a datacenter down to half compared to user-based transfers. Finally, we propose a dynamic cost model schema for the service usage, which enables the cloud providers to regulate and encourage data exchanges via a data transfer market.

6.3. Optimizing Map-Reduce processing

6.3.1. Optimizing Map-Reduce 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.

In [32], we developed a novel disk I/O scheduling framework, named *Pregather*, to improve disk I/O efficiency through exposure and exploitation of the spatial locality in the virtualized environment (regional and sub-regional spatial locality corresponds to the virtual disk space and applications' access patterns, respectively). In [14], we extend *Pregather* to improve disk I/O utilization further while reducing the disk resource contention and ensuring the I/O performance of VMs with different degrees of spatial locality. To do so, we developed an adaptive time-slice allocation scheme based on the spatial locality of VMs, to adjust the lengths of I/O time slices of VMs dynamically. We evaluated *Pregather* through extensive experiments that involve multiple simultaneous applications of both synthetic benchmarks and a Map-Reduce application (e.g., distributed sort) on Xen-based platforms.

Our evaluations use synthetic benchmarks, a Map-Reduce application (distributed sort) and database workloads. They demonstrate that *Pregather* achieves high disk spatial locality, yields a significant improvement in disk throughput, ensures the performance guarantees of VMs, and enables 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).

6.3.2. A simulation approach to evaluate Map-Reduce performance under failure

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

Map-Reduce is emerging as a prominent tool for large-scale data analysis. It is often advocated as an easier-to-use, efficient and reliable replacement for the traditional programming model of moving the data to the computation. The popular open source implementation of Map-Reduce, Hadoop, is now widely used by major companies, including Facebook, Amazon, Last.fm, and the New York Times. Fault tolerance is one of the key features of the Map-Reduce system. Map-Reduce is designed to handle various kind of failures including stop-fail and time failures: Map-Reduce re-executes failed tasks and re-launches another copy of slow tasks. Although many studies have been dedicated to investigate and improve the performance of Map-Reduce, comparatively little attention has been devoted on investigating the performance of Map-Reduce under failures.

In this ongoing work, we investigate how Map-Reduce (i.e., Hadoop) behaves under failures. To do so, we developed *iHadoop*, a Hadoop simulator developed in Java on top of SimGrid. Experimental results demonstrated that *iHadoop* accurately simulates the behavior of Hadoop and therefore can accurately predict the performance of Hadoop when running on large-scale system using the Grid'5000 testbed. In particular, iHadoop can accurately predict the percentage of Map tasks locality, the number of speculative tasks and, more importantly, the overall execution time of Map-Reduce applications under failures.

6.3.3. Waste-Free Preemption Strategy for Hadoop

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop is widely used in the computer industry because of its scalability, reliability, ease of use, and low cost of implementation. Hadoop hides the complexity of discovery and handling failures from the schedulers, but the burden of failure recovery relies entirely on users, regardless of root causes. We systematically assess this burden through a set of experiments, and argue that more effort to reduce this cost to users is desirable. We also analyze the drawback of current Hadoop mechanism in prioritizing failed tasks. By trying to launch failed tasks as soon as possible regardless of locality, it significantly increases the execution time of jobs with failed tasks, due to two reasons: 1) available slots might not be free up as quickly as expected; and 2) the slots might belong to machines with no data on it, introducing extra cost for data transfer through network, which is normally the most scare resource in nowadays data centers.

In this ongoing work, we introduce a new algorithmic approach called the waste-free preemption. The waste-free preemption saves Hadoop scheduler from solely choosing between kill, which instantly releases the slots but is wasteful, and wait, which does not waste any previous effort but fails for the two above-mentioned reasons. With this new strategy, a preemptive version of Hadoop's default schedulers (FIFO and Fair) has been implemented. The evaluation demonstrates the effectiveness of the new feature by comparing its performance with the traditional Hadoop mechanism.

6.3.4. Optimizing incremental Map-Reduce computations for on-demand data upload

Participants: Stefan Ene, Alexandru Costan, Gabriel Antoniu.

Research on cloud-based Big Data analytics has focused so far on optimizing the performance and cost-effectiveness of the computations, while largely neglecting an important aspect: users need to upload massive datasets on clouds for their computations. In this context, we study the problem of running Map-Reduce applications by considering the simultaneous optimization of performance and cost of both the data upload and its corresponding computation taken together. We analyze the feasibility of incremental Map-Reduce approaches to let the computation progress as much as possible during the data upload by using already transferred data to compute intermediate results.

Current approaches that are either optimized for different purposes, or address the computational problem independent of the data upload. In contrast, to our best knowledge, this is the first approach which simultaneously focuses on both data upload and processing. In this context, we show in [17] that it is not always efficient to attempt to overlap the transfer time with as many incremental computations as possible: a better solution is to wait long enough to fill the computational capacity of the Map-Reduce cluster. Based on this idea, we developed and evaluated a preliminary prototype. To demonstrate the viability of our prototype in real-life, we run extensive experiments in a distributed setting that involves a 11-node large incremental Map-Reduce deployment based on Hourglass. The results show significant benefits for our approach compared with a simple incremental strategy that starts the next incremental job immediately after the previous has finished: the time-to-solution is improved by 1%, the compute time after the data transfer is finished is reduced by up to 40% and the cost is reduced 10 %-44 %. Compared with a serialized strategy that starts the computation only after all data is transferred, the time-to-solution is improved by up to 30 %, the compute time after the upload finished is reduced by up to 60 % and the cost is reduced between 4 % and 23 %.

6.4. Energy-Aware Data Management in the Cloud and Exascale HPC Systems

6.4.1. Energy-efficiency in Hadoop

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

With increasingly inexpensive cloud storage and increasingly powerful cloud processing, the cloud has rapidly become the environment to store and analyze data. Most of the large-scale data computations in the cloud heavily rely on the Map-Reduce paradigm and its Hadoop implementation. Nevertheless, this exponential growth in popularity has significantly impacted power consumption in cloud infrastructures.

In [18], we focus on Map-Reduce and we investigate the impact of dynamically scaling the frequency of compute nodes on the performance and energy consumption of a Hadoop cluster. To this end, a series of experiments are conducted to explore the implications of Dynamic Voltage Frequency scaling (DVFS) settings on power consumption in Hadoop-clusters. By adapting existing DVFS governors (i.e., *performance*, *powersave*, *on-demand*, *conservative* and *user-space*) in the Hadoop cluster, 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 Map-Reduce applications. Furthermore, our results reveal that the current CPU governors do not exactly reflect their design goal and may even become ineffective to manage power consumption in Hadoop clusters.

More recently, we extended our work to further illustrate the behavior of different governors, which influence the energy consumption in Hadoop Map-Reduce. We extend our experimental platform from 15 to 40 nodes and we employ two additional benchmarks: K-means and wordcount. Moreover, we investigate preliminary DVFS models that adjust to the various stages of Hadoop applications. We also demonstrate that achieving better energy efficiency in Hadoop cannot be done by tuning the governors parameters, nor through a naive coarse-grained tuning of the CPU frequencies or the governors according the running phase (i.e., map phase or reduce phase). In addition, we provide an extensive discussion of the sensitivity for different parameters employed in *ondemand* and *conservative* governors.

6.4.2. Exploring the impact of dedicated resources on energy consumption in Exascale systems

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

The advent of fast, unprecedentedly scalable, yet energy-hungry Exascale supercomputers poses a major challenge consisting in sustaining a high performance-per-Watt ratio. While much recent work has explored new approaches to I/O management, aiming to reduce the I/O performance bottleneck exhibited by HPC applications (and hence to improve application performance), there is comparatively little work investigating the impact of I/O management approaches on energy consumption.

In [23], we explore how much energy a supercomputer consumes while running scientific simulations when adopting various I/O management approaches. We closely examine three radically different I/O schemes including time partitioning, dedicated cores, and dedicated nodes. We implement the three approaches within the Damaris I/O middleware and perform extensive experiments with one of the target HPC applications of the Blue Waters sustained-Petaflops supercomputer project: the CM1 atmospheric model. The experimental results obtained on the French Grid'5000 platform highlight the differences between these three approaches and illustrate in which way various configurations of the application and of the system can impact performance and energy consumption.

Based on those experimental results, we are working on building a new energy model which can estimate the energy consumptions of various I/O management approaches and help users in selecting the optimal I/O approach to run their application.

6.4.3. Energy impact of data consistency management in the HBase distributed cloud data store

Participants: Álvaro García Recuero, Shadi Ibrahim, Gabriel Antoniu.

Cloud Computing has recently emerged as a key technology providing individuals and companies with access to remote computing and storage infrastructures. In order to achieve high-availability and fault-tolerance, cloud data storage relies on replication. That comes with the issue of consistency among distant replicas so one can always get the most up-to-date values from any of them (*e.g.*, fresh data).

In that context, being able to provide data consistency and continuous availability in the Cloud is yet a non-trivial problem, mainly due to the ever-increasing volume, variety and velocity of data in storage systems. Big data processing engines (e.g., Hadoop, Spark, etc.) as well as modern NoSQL storage back-ends (HBase, Cassandra) have to therefore deal with these high volumes of information at large scale while still providing applications with a consistent and on-time data delivery.

In this work, a set of synthetic workloads from YCSB (Yahoo! Cloud Service Benchmark) was configured to simulate random reads/writes and measure their impact into the overall energy consumption of a well-known distributed data store, HBase. The cluster is comprised of 40 servers and the results have been confirmed with several configurations and runs on the Grid5000 experimental platform. The results indicate that certain write-intensive workloads can be a bottleneck in terms of throughput, further deepening the problem of having an energy-efficient consistency management. Regarding read-intensive workloads, we observe similar patterns but with a very different impact on their energy footprint. We plan to further investigate how to leverage energy-aware mechanisms that overcome the energy-consistency trade-off, while taking into account the selected configuration.

6.5. Scalable I/O and Visualization for Exascale Systems

6.5.1. CALCioM: mitigating cross-application I/O interference

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

As larger supercomputers are used by an increasing number of applications in a concurrent manner, the interference produced by multiple applications accessing a shared parallel file system in contention becomes a major problem. Interference often breaks single-application I/O optimizations (such as access patterns preliminarily optimized to improve data locality on disks), thereby dramatically degrading application I/O performance, increasing run-time variability and, as a result, lowering machine-wide efficiency. We addressed this challenge by proposing CALCioM [15], 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. We examined four I/O strategies that can be accommodated in this framework: serializing, interrupting, interfering and coordinating. Experiments on Argonne's BG/P Surveyor machine and on several clusters of Grid'5000 showed that CALCioM can be used to improve the scheduling strategy efficiently and transparently between several otherwise interfering applications, given specified metrics of machine-wide efficiency. This work led to a publication at the IPDPS 2014 conference.

6.5.2. Omnisc'IO: Predicting the I/O patterns of HPC applications

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

Many I/O optimizations including prefetching, caching, and scheduling, have been proposed to improve the performance of the I/O stack. In order to optimize these techniques, modeling and predicting spatial and temporal I/O patterns of HPC applications as they run, have become crucial. In this direction we introduced Omnisc'IO [16], an original approach that aims to make a step forward toward an intelligent I/O management of HPC applications in next-generation, post-Petascale supercomputers. It builds a grammar-based model of the I/O behavior of any HPC application, and uses this model to predict when future I/O operations will occur, as well as where and how much data will be accessed. Omnisc'IO is transparently integrated into the POSIX and MPI-I/O stacks and does not require any modification to application sources or to high-level I/O libraries. It works without prior knowledge of the application, and converges to accurate predictions within a couple of iterations only. Its implementation is efficient both in computation time and in memory footprint. Omnisc'IO was evaluated with four real HPC applications — CM1, Nek5000, GTC, and LAMMPS — using a variety of I/O backends ranging from simple POSIX to Parallel HDF5 on top of MPI-I/O. Our experiments showed that Omnisc'IO achieves from 79 % to 100 % accuracy in spatial prediction and an average precision of temporal predictions ranging from 0.2 seconds to less than a millisecond. This work was published at the SC14 conference and initiated the development of the Omnisc'IO software.

6.5.3. Smart In-Situ Visualization

Participants: Lokman Rahmani, Matthieu Dorier, Gabriel Antoniu.

The increasing gap between computational power and I/O performance in new supercomputers has started to drive a shift from an offline approach to data analysis to an inline approach, termed *in-situ visualization* (ISV). While most visualization software now provides ISV, they typically visualize large dumps of unstructured data, by rendering everything at the highest possible resolution. This often negatively impacts the performance of simulations that support ISV, in particular when ISV is performed interactively, as in-situ visualization requires synchronization with the simulation. In this ongoing work, we investigate a smarter method of performing ISV. Our approach consists in adapting the resolution of regions of the visualization area based on how much their data are *relevant* with regards to the physical phenomena being simulated. In this direction, we first provide a generic definition of relevant data subsets based on *data variability*. Following this definition, we investigate various filtering algorithms to detect relevant data subsets automatically. The proposed filtering algorithms are derived from information theory, statistics and image processing. Our work is validated in the context of climate simulation, where we show an up to 40% improvement of time-to-solution without any significant loss regarding the quality of visualization (QoV). QoV loss is *quantified* using the structural similarity index metric (SSIM) that takes in consideration human visual system to compute visual errors.

6.6. Data Streaming and Small Data

6.6.1. JetStream: enabling high-performance event streaming across cloud data-centers

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The easily-accessible computation power offered by cloud infrastructures coupled with the revolution of Big Data are expanding the scale and speed at which data analysis is performed. In their quest for extracting value out of the 3 Vs of Big Data, applications process larger data sets, within and across clouds. Enabling fast data transfers across geographically distributed sites becomes particularly important for applications which manage continuous streams of events in real time. Scientific applications (e.g. the Ocean Observatory Initiative or the ATLAS experiment) as well as commercial ones (e.g. Microsoft's Bing and Office 365 large-scale services) operate on tens of data-centers around the globe and follow similar patterns: they aggregate monitoring data, assess the QoS or run global data mining queries based on inter-site event stream processing.

In [22] we propose a set of strategies for efficient transfers of events between cloud data-centers and we introduce JetStream: a prototype implementing these strategies as a high-performance, batch-based streaming middleware. JetStream is able to self-adapt to the streaming conditions by modeling and monitoring a set of context parameters. It further aggregates the available bandwidth by enabling multi-route streaming across cloud sites. The prototype was validated on tens of nodes from US and Europe data-centers of the Windows Azure cloud using synthetic benchmarks and with application code in the context of the Alice experiment at CERN. The results show an increase in transfer rate of 250 times over individual event streaming. Besides, introducing an adaptive transfer strategy brings an additional 25 % gain. Finally, the transfer rate can further be tripled thanks to the use of multi-route streaming.

6.6.2. Efficient management of many small data objects

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Large-scale intensive applications must often manage millions or even billions of small objects. Twitter, for example, has to record on average 5700 new tweets every second. Each of these objects are typically smaller than a kilobyte, and as a result, the database has to store billions of these objects. The sheer amount of objects and the small data sizes can also be found in many other applications, like sensor networks, or graph processing. Another important aspect are the access patterns of these applications where reads dominate over writes, which means the storage system has to be heavily optimized towards read performance.

To address these challenges, we are designing a novel storage system offering fast data access with minimal overhead. Learning from BlobSeer [33], we introduce a more efficient way to manage metadata. To this end, we propose to remove the centralised version manager and to distribute versions across the whole cluster using a distributed hash table. This greatly reduces the response times by allowing single-hop reads for most usage patterns. Additionally, this approach distributes the load over the whole cluster, thus providing a better horizontal scalability and fault tolerance.

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

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. ADT

ADT BlobSeer (2013–2014). To support the development of the BlobSeer software for ongoing cooperations, Inria provided support for a research engineer. Loïc Cloatre has been hired as a senior engineer for the second year of this project, starting in February 2014.

8.1.3. 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 and H2020 Projects

BigStorage (2015–2018)

Program: European Training Network (ETN).

Coordinator: María S. Pérez.

Partners: Universidad Politécnica de Madrid (UPM), Barcelona Supercomputing Center (PSC), Johannes Gutenberg Universität Mainz, Foundation for Research and Technology - Hellas (FORTH), Xyratex Technology Limited, Deutsches Klimarechenzentrum, CA Technologies, Fujitsu Technology Solutions GmbH, French Atomic Agency CEA, IBM Research Ireland, Bull SAS, and Informatica El Corte Ingles.

Abstract: The consortium of this Marie-Curie Innovative Training Networks (ITN) *BigStorage: Storage-based Convergence between HPC and Cloud to handle Big Data* aims at training future data scientists in order to enable them and us to apply holistic and interdisciplinary approaches for taking advantage of a data-overwhelmed world, which requires HPC and Cloud infrastructures with a redefinition of storage architectures underpinning them — focusing on meeting highly ambitious performance and energy usage objectives. KerData mainly collaborates with UPM and PSC 2 co-advised PhD theses).

8.2.2. Collaborations in European Programs, except FP7 and H2020

Program: EIT ICT Labs.

Project acronym: EUROPA Activity - Future Cloud Action Line.

Project title: Big Data Analytics with Apache Flink for Real Business Use-Cases.

Duration: May 2014–December 2014.

Coordinator: Gabriel Antoniu, Alexandru Costan.

Participants: Anirvan Basu, Camelia Ciolac.

Other partners: TU Berlin (Germany), VTT (Finland), F-Secure (Finland).

Abstract: In this project, we study the requirements with respect to Big Data analytics today, following several interviews with representative companies from various domains ranging from online mobile gaming to security and logistics. The goal is to identify those requirements that could be addressed by the Apache Flink (formerly known as Stratosphere) platform and apply them in some real-life business scenarios. We first present the state-of-the-art in the field of Big Data analytics, then validate the novel features of Flink. Finally we study how some of the requirements needed by the industry could be addressed by the latter, and illustrate them with 2 real use-cases. To this end, Camelia Ciolac and Anirvan Basu were hired and implemented two demos showing the use of Flink to solve Big Data problems from 2 companies: a mobile games developer (Tribeflame) and a security company (F-Secure), respectively.

8.3. International Initiatives

8.3.1. Inria International Labs

JLESC: Joint Laboratory on Extreme-Scale Computing. This laboratory is jointly run by Inria, UIUC, ANL and BSC. It has ben created in 2014 as a follow-up of the Inria-UIUC JLPC to collaborate on concurrency-optimized I/O for Extreme-scale platforms (see details in Section 4.3). This project is an extension of the Joint Inria-UIUC Laboratory for Petascale Computing (JLPC) which was used as the basis of the Data@Exascale Associate Team with ANL and UIUC (2013–2015).

8.3.2. Inria Associate Teams

Data@Exascale

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

Inria principal investigator: Gabriel Antoniu

International Partners:

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.

8.3.3. 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. Rohit Saxena was hired as an engineer until February 2014.

8.3.4. Inria International Partners

8.3.4.1. Declared Inria International Partners

Politehnica University of Bucharest. This status was established since January 2013, right after the end of our former DataCloud@work Associate Team.

8.3.4.2. Informal International Partners

Huazhong University of Science and Technology (HUST), China. We collaborate on optimizing Map-Reduce in virtualized environments.

Nanyang Technological University (NTU). We collaborate on optimizing Big Data applications in the Cloud and HPC systems.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Robert Ross (Argonne National Lab) visited the KerData team for one week (June 2014) within the framework of the Data@Exascale Associate Team, as an Invited Professor funded by the University of Rennes 1.

8.4.2. Internships

Stefan Ene

Subject: Overlapping cloud data transfers and computation for incremental Map-Reduce.

Date: April–September 2014.

Institution: Master student from Politehnica University of Bucharest (Romania). Co-funded by the Inria Internships Program.

Andreea Pintilie

Subject: Bio-informatics inspired algorithms for fast cloud data transfers.

Date:April–September 2014.

Institution: Master student from Politehnica University of Bucharest (Romania). Co-funded by the Inria Internships Program.

Anh-Phuong Tran

Subject: Failure-aware job scheduling in Hadoop cloud data centers.

Date: February–June 2014.

Institution: Master student enrolled in the European Master in Distributed Computing (EMDC) program, a joint program between KTH Royal Institute of Technology in Sweden and Instituto Superior Tecnico in Portugal.

Tien Dat Phan

Subject: A simulation approach to evaluate Map-Reduce performance under failure.

Date: February 2014–June 2014.

Institution: Master student from University Rennes 1, Rennes (France)

Orçun Yildiz

Subject: (In-)Efficiency in energy consumption of data management on Petascale supercomputers.

Date: February–July 2014.

Institution: Master student enrolled in the European Master in Distributed Computing (EMDC) program, a joint program between KTH Royal Institute of Technology in Sweden and Instituto Superior Tecnico in Portugal.

Thomas Bouguet

Subject: Development of a web platform for the analysis of Darshan I/O log files.

Date: May–July 2014.

Institution: Master student from University Rennes 1, Rennes (France).

8.4.3. Visits to International Teams

Lokman Rahmani visited ANL (Rob Ross, Tom Peterka) for 2 months, funded by the PUF NextGen project in the context of the Joint Laboratory for Extreme-Scale Computing (JLESC).

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

Gabriel Antoniu

- Program Chair of the IEEE Cluster 2014 conference, Madrid, September 2014.

Luc Bougé

- Vice-Chair of the Euro-Par Steering Committee. Chair of the Euro-Par Workshop Advisory Board.
- Co-Chair of the PhD Forum of the IPDPS 2014 Conference, Phoenix, May 2014.
- Chair of the PhD Student Mentoring Program of the IEEE Cluster 2014 conference, Madrid, September 2014.

Shadi Ibrahim

- PhD Consortium Co-Chair for the 2014 CloudCom conference, Singapore, December 2014.
- Workshop Co-Chair for the 2014 ScalCom conference, Indonesia, December 2014.

Alexandru Costan

- Program Co-Chair of the BigDataCloud 2014 International workshop held in conjunction with the Euro-Par 2104 conference, Porto, August 2014.
- Submission Chair of the IEEE Cluster 2014 conference, Madrid, September 2014.

9.1.2. Scientific events selection

Gabriel Antoniu

- Program Chair of the IEEE Cluster 2014 conference (Madrid, 22–26 September 2014).
- Member of the following Program Committees: ACM HPDC 2014, ACM/IEEE CC-Grid'2014, IEEE Big Data 2014, ACM/IEEE SC'14 (Technical Program Member - Posters Committee), BigDataCloud 2014 workshop (held in conjunction with the Euro-Par 2014 conference).

Shadi Ibrahim

- Program Committee Chair of the PhD Consortium for the 2014 CloudCom.
- Member of the following Program Committees: IEEE Cluster 2014, IEEE SCC 2014, IEEE CloudCom 2014, ICPADS 2014, HPCC 2014, ICA3PP 2014, NPC 2014, MEDES 2014, ISPDC 2014, PICom-2014, SCRAMBL workshop 2014, CLOUD COMPUTING 2014.
- Other reviews: HPDC 2014, Euro-par 2014, Big Data 2014, SC14 Posters and Student Research Competition.

Alexandru Costan

- Member of the following Program Committees: ICPP 2014, IEEE CloudCom 2014, IEEE CloudCom PhD Forum 2014, ISPDC 2014, ARMS-CC workshop 2014, IEEE Cluster 2014, BigDataCloud Workshop 2014
- Other reviews: ACM HPDC 2014, EuroPar 2014, ARMS-CC, IEEE Cluster 2014, IEEE/ACM CCGrid 2014, SC14.

9.1.3. Journal

Luc Bougé

- Member of the Editorial Board of Scientific Programming.

Shadi Ibrahim

- Guest editors for a Special Issue on *Advanced Techniques for Cloud Data Management* in the International Journal Transactions on Large-Scale Data and Knowledge Centered Systems (TLDKS), Springer.
- Other reviews: IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Computers, IEEE Transactions on Cloud Computing, ACM Transactions on Internet Technology, Future Generation Computer Systems, IEEE Systems Journal, Journal of Supercomputing, Cluster Computing, Springer Transactions on Large-Scale Data and Knowledge-Centered Systems.

Alexandru Costan

- Other reviews: IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Cloud Computing, Future Generation Computer Systems, Concurrency and Computation - Practice and Experience, Computers and Geosciences

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Gabriel Antoniu

Master (Engineering Degree, 5th year): Big Data, 24 hours (lectures), M2 level, ENSAI (*École Nationale Supérieure de la Statistique et de l'Analyse de l'Information*), Bruz, France.

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

Master: Scalable Distributed Systems, 10 hours (lectures), M2 level, SDS Module, M2RI Master Program, ENS Rennes, France.

Master: Scalable Distributed Systems, 17 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School, France.

Luc Bougé

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

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

Shadi Ibrahim

Master: Hadoop, 48 hours (Project), M1 Level, ENS Rennes, France

Master: Cloud1, Map-Reduce, 16 hours (lectures, lab sessions), M2 Level, École des Mines de Nantes, Nantes, France.

Master (Engineering Degree, 5th year): Big Data, 12 hours (lab sessions), M2 level, ENSAI (*École Nationale Supérieure de la Statistique et de l'Analyse de l'Information*), Bruz, France.

Alexandru Costan

Bachelor: Object-oriented programming, 18 hours (lectures), L3, ENS Rennes

Bachelor: Java programming, 28 hours (lab sessions), L2, INSA Rennes

Bachelor: Databases, 68 hours (lectures and lab sessions), L2, INSA Rennes, France

Bachelor: Practical case studies, 24 hours (project), L3, INSA Rennes

Master: Big Data and Applications, 36h hours (lectures, lab sessions, project), M1, INSA Rennes

Matthieu Dorier

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

Master: Initiation to Unix Systems, 2 hours (lab sessions), M1 level, ENS Rennes.

Lokman Rahmani

Bachelor: Java Programming, 18 hours (lab sessions), L3 level, MIAGE program, University Rennes 1, France.

Master: Distributed Programming, 48 hours (lab sessions), M1 level, GL program, University Rennes 1, France.

9.2.2. Supervision

PhD defended: Matthieu Dorier, *Addressing the Challenges of I/O Variability in Post-Petascale HPC Simulations*, thesis started in October 2011 co-advised by Gabriel Antoniu and Luc Bougé. Defended on December 9, 2014.

PhD defended: Radu Tudoran, *High-Performance Big Data Management Across Cloud Data Centers*, thesis started in October 2011, co-advised by Gabriel Antoniu and Luc Bougé. Defended on December 10, 2014.

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.

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

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

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

9.2.3. Juries

Gabriel Antoniu served as a member of Inria's Junior Researcher (CR2) Admissibility Jury for the Rennes and Grenoble Research Centers and of Inria's Junior Researcher Admission Jury (CR1 and CR2, all centers).

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

9.2.4. Miscellaneous

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

Shadi Ibrahim served as a member of Inria's Post-Doc Evaluation Committee since 2014.

Shadi Ibrahim served as a member of jury for L3 internship at ENS Rennes, September 2014.

Shadi Ibrahim is giving a Tutorial on *Green Big Data Processing using Hadoop: An introductory tutorial* at the Middleware 2014 conference, Bordeaux, France, December 2014 (with Anne-Cécile Orgerie).

Matthieu Dorier served as a member of jury for L3 internship at ENS Rennes, September 2014.

9.3. Popularization

- Gabriel Antoniu

  Microsoft Research, Redmond. Invited presentation at the Workshop on *E-Science in the Cloud*. Subject: Big Data management in the Cloud (A-Brain and Z-CloudFlow projects). Audience: engineers, students and researchers, from academia and industry (April 2014).

  ORAP Forum. Invited Speaker. *Big Data: Concepts-clés et enjeux* (April 2014).

  Radio France. Invited interview about Big Data at the *Labo des savoirs*. Available on line at http://labodessavoirs.fr/chroniques-et-reportages/un-nouveau-paradigme/.

  Institut Français de Bio-informatique, Grand-Ouest, Rennes. Invited presentation at the *12e Rencontres des plates-formes de Bioinformatique du Grand Ouest* about the team's experience with the Microsoft Azure cloud platform (November 2014).

- Shadi Ibrahim

  Parallel and Distributed Computing Centre (PDCC), Nanyang Technological University (NTU), Singapore. Invited seminar. Subject: *Consistency Management for Big Data Applications in the Clouds* (January 2014).
- Alexandru Costan

  EIT ICT Labs, Rennes. Invited presentation at the Workshop on *Trusted Clouds* about *KerData Team: Scalable Data Management on Clouds and Beyond* (March 2014).
- Matthieu Dorier

  Inria, Grenoble. Invited seminar about *Damaris: data management for scientific simulations on post-Petascale supercomputers* (June 2014).

10. Bibliography

Major publications by the team in recent years

[1] H.-E. Chihoub, S. Ibrahim, G. Antoniu, M. Pérez. *Consistency Management in Cloud Storage Systems*, in "Large Scale and Big Data - Processing and Management", S. Sakr, M. M. Gaber (editors), CRC Press, 2014, https://hal.inria.fr/hal-00784885

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

[3] B. Da Mota, R. Tudoran, A. Costan, G. Varoquaux, G. Brasche, P. J. Conrod, H. Lemaitre, T. Paus, M. Rietschel, V. Frouin, J.-B. Poline, G. Antoniu, B. Thirion. *Machine Learning Patterns for Neuroimaging-Genetic Studies in the Cloud*, in "Frontiers in Neuroinformatics", April 2014, vol. 8, https://hal.inria.fr/hal-01057325

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

[5] M. Dorier, G. Antoniu, R. Ross, D. Kimpe, S. Ibrahim. *CALCioM: Mitigating I/O Interference in HPC Systems through Cross-Application Coordination*, in "IPDPS - International Parallel and Distributed Processing Symposium", Phoenix, United States, May 2014, https://hal.inria.fr/hal-00916091

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

[7] 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, http://hal.inria.fr/inria-00511414/en/

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

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

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] M. DORIER. *Addressing the Challenges of I/O Variability in Post-Petascale HPC Simulations*, Ecole Normale Supérieure de Rennes, December 2014, https://tel.archives-ouvertes.fr/tel-01099105

[12] R. TUDORAN. *High-Performance Big Data Management Across Cloud Data Centers*, ENS Rennes, December 2014, https://tel.archives-ouvertes.fr/tel-01093767

Articles in International Peer-Reviewed Journals

[13] B. DA MOTA, R. TUDORAN, A. COSTAN, G. VAROQUAUX, G. BRASCHE, P. J. CONROD, H. LEMAITRE, T. PAUS, M. RIETSCHEL, V. FROUIN, J.-B. POLINE, G. ANTONIU, B. THIRION. *Machine Learning Patterns for Neuroimaging-Genetic Studies in the Cloud*, in "Frontiers in Neuroinformatics", April 2014, vol. 8 [*DOI :* 10.3389/FNINF.2014.00031], https://hal.inria.fr/hal-01057325

[14] X. LING, S. IBRAHIM, S. WU, H. JIN. *Spatial Locality Aware Disk Scheduling in Virtualized Environment*, in "IEEE Transactions on Parallel and Distributed Systems", September 2014, 14 p. [*DOI :* 10.1109/TPDS.2014.2355210], https://hal.inria.fr/hal-01087602

International Conferences with Proceedings

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

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

[17] S. ENE, B. NICOLAE, A. COSTAN, G. ANTONIU. *To Overlap or Not to Overlap: Optimizing Incremental MapReduce Computations for On-Demand Data Upload*, in "DataCloud '14: The 5th International Workshop on Data-Intensive Computing in the Clouds", New Orleans, United States, November 2014, pp. 9-16 [*DOI :* 10.1109/DATACLOUD.2014.7], https://hal.inria.fr/hal-01094609

[18] S. IBRAHIM, D. MOISE, H.-E. CHIHOUB, A. CARPEN-AMARIE, L. BOUGÉ, G. ANTONIU. *Towards Efficient Power Management in MapReduce: Investigation of CPU-Frequencies Scaling on Power Efficiency in Hadoop* , in "Workshop on Adaptive Resource Management and Scheduling for Cloud Computing, Held in conjunction with PODC", Paris, France, July 2014, https://hal.inria.fr/hal-01077285

[19] R. Tudoran, A. Costan, G. Antoniu. *Transfer as a Service: Towards a Cost-Effective Model for Multi-Site Cloud Data Management*, in "Proceedings of the 33rd IEEE Symposium on Reliable Distributed Systems (SRDS 2014)", Nara, Japan, IEEE, October 2014, https://hal.inria.fr/hal-01023282

[20] R. Tudoran, A. Costan, R. Wang, L. Bougé, G. Antoniu. *Bridging Data in the Clouds: An Environment-Aware System for Geographically Distributed Data Transfers*, in "14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing", Chicago, United States, May 2014, https://hal.inria.fr/hal-00978153

[21] R. Tudoran, K. Keahey, P. Riteau, S. Panitkin, G. Antoniu. *Evaluating Streaming Strategies for Event Processing across Infrastructure Clouds*, in "14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)", Chicago, United States, IEEE, May 2014, pp. 151 - 159 [*DOI :* 10.1109/CCGrid.2014.89], https://hal.inria.fr/hal-01089371

[22] R. Tudoran, O. Nano, I. Santos, A. Costan, H. Soncu, L. Bougé, G. Antoniu. *JetStream: Enabling High Performance Event Streaming across Cloud Data-Centers*, in "Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems DEBS'14", Mumbai, India, ACM, May 2014, pp. 23 - 34 [*DOI :* 10.1145/2611286.2611298], https://hal.archives-ouvertes.fr/hal-01090281

[23] O. Yildiz, M. Dorier, S. Ibrahim, G. Antoniu. *A Performance and Energy Analysis of I/O Management Approaches for Exascale Systems*, in "DIDC '14 Proceedings of the sixth international workshop on Data Intensive Distributed Computing", Vancouver, Canada, June 2014, pp. 35-40 [*DOI :* 10.1145/2608020.2608026], https://hal.inria.fr/hal-01076522

Scientific Books (or Scientific Book chapters)

[24] H.-E. Chihoub, S. Ibrahim, G. Antoniu, M. Pérez. *Consistency Management in Cloud Storage Systems*, in "Large Scale and Big Data - Processing and Management", S. Sakr, M. M. Gaber (editors), CRC Press, 2014, https://hal.inria.fr/hal-00784885

[25] L. Lopez, J. Zilinskas, A. Costan, R. G. Cascella, G. Kecskemeti, E. Jeannot, M. Cannataro, L. Ricci, S. Benkner, S. Petit, V. Scarano, J. Gracia, S. Hunold, S. L. Scott, S. Lankes, C. Lengauer, J. Carretero, J. Breitbart, M. Alexander. *Euro-Par 2014: Parallel Processing Workshops, Part I*, Lecture Note In Computer Science, Springer, December 2014, vol. 8805, https://hal.inria.fr/hal-01110069

[26] L. Lopez, J. Zilinskas, A. Costan, R. G. Cascella, G. Kecskemeti, E. Jeannot, M. Cannataro, L. Ricci, S. Benkner, S. Petit, V. Scarano, J. Gracia, S. Hunold, S. L. Scott, S. Lankes, C. Lengauer, J. Carretero, J. Breitbart, M. Alexander. *Euro-Par 2014: Parallel Processing Workshops, Part II*, Lecture Note In Computer Science, Springer, December 2014, vol. 8806, https://hal.inria.fr/hal-01110071

References in notes

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

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

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

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

[31] 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

[32] 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, August 2013, https://hal.inria.fr/hal-00842076

[33] B. Nicolae, G. Antoniu, L. Bougé. *BlobSeer: Efficient Data Management for Data-Intensive Applications Distributed at Large-Scale*, in "IPDPS '10: Proceedings of the 24th IEEE International Symposium on Parallel and Distributed Processing: Workshops and Phd Forum", Atlanta, United States, April 2010, pp. 1-4 [*DOI :* 10.1109/IPDPSW.2010.5470802], https://hal.inria.fr/inria-00457809

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

[35] 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], https://hal.inria.fr/hal-00640900