IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Université Rennes 1

École normale supérieure de Rennes

Activity Report 2016

Project-Team KERDATA

Scalable Storage for Clouds and Beyond

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Distributed and High Performance Computing

Table of contents

Project-Team KERDATA

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

Keywords:

Computer Science and Digital Science:

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

Other Research Topics and Application Domains:

1.1.2. - Molecular biology
2.6.1. - Brain imaging
3.2. - Climate and meteorology
4.5.1. - Green computing
9.4.5. - Data science

1. Members

Research Scientists

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

Faculty Members

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

Engineer

Hadi Salimi [Inria, from April 2016]

PhD Students

Nathanaël Cheriere [ENS Rennes, from September 2016]
Paul Le Noac'h [INSA Rennes, from November 2016]
Ovidiu-Cristian Marcu [Inria]
Pierre Matri [Inria and Universidad Politécnica de Madrid]
Tien-Dat Phan [Univ. Rennes I]
Luis Eduardo Pineda Morales [Inria]
Lokman Rahmani [Univ. Rennes I]
Mohammed-Yacine Taleb [Inria]
Orçun Yildiz [Inria]

Post-Doctoral Fellow

Chi Zhou [Inria, from May 2016]

Administrative Assistant

Aurélie Patier [Univ. Rennes I]

Others

Muhammad Najeeb Aslam [Univ. Rennes I, Master intern, from May 2016 until August 2016]
Rémi Hutin [Inria, Undergraduate intern, from May 2016 until July 2016]

2. Overall Objectives

2.1. Context: the need for scalable data management

We are witnessing a rapidly increasing number of application areas generating and processing very large volumes of data on a regular basis. Such applications are called *data-intensive*. Governmental and commercial statistics, climate modeling, cosmology, genetics, bio-informatics, high-energy physics are just a few examples in the scientific area. In addition, rapidly growing amounts of data from social networks and commercial applications are now routinely processed.

In all these examples, the overall application performance is highly dependent on the properties of the underlying data management service. It becomes crucial to store and manipulate massive data efficiently. However, these data are typically *shared* at a large scale and *concurrently accessed* at a high degree. With the emergence of recent infrastructures such as cloud computing platforms and post-Petascale high-performance computing (HPC) systems, achieving highly scalable data management under such conditions has become a major challenge.

2.1.1. Our objective

The KerData project-team is namely focusing on designing innovative architectures and systems for *scalable data storage and processing*. We target two types of infrastructures: *clouds* and *post-Petascale high-performance supercomputers*, according to the current needs and requirements of data-intensive applications.

We are especially concerned by the applications of major international and industrial players in cloud computing and extreme-scale high-performance computing (HPC), which shape the long-term agenda of the cloud computing [40], [37] and Exascale HPC [39] research communities. The Big Data area, which has recently captured a lot of attention, emphasized the challenges related to Volume, Velocity and Variety. This is yet another element of context that further highlights the primary importance of designing data management systems that are efficient at a very large scale.

2.1.1.1. Alignment with Inria's scientific strategy

Data-intensive applications exhibit several common requirements with respect to the need for data storage and I/O processing. We focus on some core challenges related to data management, resulted from these requirements. Our choice is perfectly in line with Inria's strategic plan [44], which acknowledges as critical the challenges of *storing, exchanging, organizing, utilizing, handling and analyzing* the huge volumes of data generated by an increasing number of sources. This topic is also stated as a scientific priority of Inria's research centre of Rennes [43]: *Storage and utilization of distributed big data*.

2.1.1.2. Challenges and goals related to cloud data storage and processing

In the area of cloud data processing, a significant milestone is the emergence of the Map-Reduce [50] parallel programming paradigm. It is currently used on most cloud platforms, following the trend set up by Amazon [35]. At the core of Map-Reduce frameworks lies the storage system, a key component which must meet a series of specific requirements that are not fully met yet by existing solutions: the ability to provide efficient *fine-grain access* to the files, while sustaining a *high throughput* in spite of *heavy access concurrency*; the need to provide a high resilience to *failures*; the need to take *energy-efficiency* issues into account.

More recently, it becomes clear that data-intensive processing needs to go beyond the frontiers of single datacenters. In this perspective, extra challenges arise, related to the efficiency of metadata management. This efficiency has a major impact on the access to very large sets of small objects by Big Data processing workflows running on large-scale infrastructures.

2.1.1.3. Challenges and goals related to data-intensive HPC applications

Key research fields such as climate modeling, solid Earth sciences or astrophysics rely on very large-scale simulations running on post-Petascale supercomputers. Such applications exhibit requirements clearly identified by international panels of experts like IESP [42], EESI [38], ETP4HPC [39]. A jump of one order of magnitude in the size of numerical simulations is required to address some of the fundamental questions in several communities in this context. In particular, the lack of data-intensive infrastructures and methodologies to analyze the huge results of such simulations is a major limiting factor.

The challenge we have been addressing is to find new ways to store, visualize and analyze massive outputs of data during and after the simulations. Our main initial goal was to do it without impacting the overall performance, avoiding the *jitter* generated by I/O interference as much as possible. Recently, we started to focus specifically on *in situ processing* approaches and we explored approaches to *model and predict I/O phase occurrences* and to *reduce intra-application and cross-application I/O interference*.

2.1.2. Our approach

KerData's global approach consists in studying, designing, implementing and evaluating distributed algorithms and software architectures for scalable data storage and I/O management for efficient, large-scale data processing. We target two main execution infrastructures: cloud platforms and post-Petascale HPC supercomputers.

2.1.2.1. Platforms and Methodology

The highly experimental nature of our research validation methodology should be emphasized. To validate our proposed algorithms and architectures, we build software prototypes, then validate them at a large scale on real testbeds and experimental platforms.

We strongly rely on the Grid'5000 platform. Moreover, thanks to our projects and partnerships, we have access to reference software and physical infrastructures. In the cloud area, we use the Microsoft Azure and Amazon cloud platforms. In the post-Petascale HPC area, we are running our experiments on systems including some top-ranked supercomputers, such as Titan, Jaguar, Kraken or Blue Waters. This provides us with excellent opportunities to validate our results on advanced realistic platforms.

2.1.2.2. Collaboration strategy

Our collaboration portfolio includes international teams that are active in the areas of data management for clouds and HPC systems, both in Academia and Industry.

Our academic collaborating partners include Argonne National Lab, University of Illinois at Urbana-Champaign, Universidad Politécnica de Madrid, Barcelona Supercomputing Center, University Politehnica of Bucharest. In industry, we are mainly collaborating with Microsoft and IBM.

Moreover, the consortiums of our collaborative projects include application partners in the areas of Bio-Chemistry (e.g., IBCP Lyon in the MapReduce ANR project), Neurology and Genetics (e.g., the Parietal team at Inria, the NeuroSpin centre in Saclay within the A-Brain Microsoft Research-Inria project), and Climate Simulations (e.g., the Department of Earth and Atmospheric Sciences of the University of Michigan, within our collaboration inside JLESC [45]). This is an additional asset, which enables us to take into account application requirements in the early design phase of our solutions, and to validate those solutions with real applications... and real users!

3. Research Program

3.1. Research axis 1: Convergence of Extreme-Scale Computing and Big Data Infrastructures

The tools and cultures of High Performance Computing and Big Data Analytics have evolved in divergent ways. This is to the detriment of both. However, big computations still generate and are needed to analyze Big Data. As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two eco-systems is crucial to the future. Our objective for the next years is premised on the idea that we must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change the plans that are now being laid for achieving Exascale computing.

3.1.1. High-performance storage for concurrent Big Data applications

We argue that storage is a plausible pathway to convergence. In this context, we plan to focus on the needs of concurrent Big Data applications that require high-performance storage, as well as transaction support. Although blobs (binary large objects) are an increasingly popular storage model for such applications, state-of-the-art blob storage systems offer no transaction semantics. This demands users to coordinate data access carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior.

We argue there is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications. In this context, one idea on which we plan to focus our efforts is exploring how blob storage systems could provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency.

The early principles of this research direction have already raised interest from our partners at ANL (Rob Ross) and UPM (María Pérez) for potential collaborations. In this direction, the acceptance of our paper on the Týr transactional blob storage system as a Best Student Paper Award Finalist at the SC16 conference [25] is a very encouraging step.

3.1.2. Big Data analytics on Exascale HPC machines

Big Data analytics is another interesting direction that we plan to explore, building on top of these converged storage architectures. Specifically, we will examine the ways in which Exascale infrastructures can be leveraged not only by HPC-centric, but also by scientific, cloud-centric applications. Many of the current state-of-the-art Big Data processing approaches, including Hadoop and Spark [46] are optimized to run on commodity machines. This impacts the mechanisms used to deal with failures and the limited network bandwidth.

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

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

Relevant groups with similar interests are the following ones.

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

3.2. Research axis 2: Advanced data processing on Clouds

The recent evolutions in the area of Big Data processing have pointed out some limitations of the initial Map-Reduce model. It is well suited for batch data processing, but less suited for real-time processing of dynamic data streams. New types of data-intensive applications emerge, e.g., for enterprises who need to perform analysis on their stream data in ways that can give fast results (i.e., in real time) at scale (e.g., click-stream analysis and network-monitoring log analysis). Similarly, scientists require fast and accurate data processing techniques in order to analyze their experimental data correctly at scale (e.g., collectively analysis of large data sets distributed in multiple geographically distributed locations).

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

3.2.1. Stream-oriented, Big Data processing on clouds

The state-of-the-art Hadoop Map-Reduce framework cannot deal with stream data applications, as it requires the data to be initially stored in a distributed file system in order to process them. To better cope with the above-mentioned requirements, several systems have been introduced for stream data processing such as Flink [41], Spark [46], Storm [47], and Google MillWheel [49]. These systems keep computation in memory to decrease latency, and preserve scalability by using data-partitioning or dividing the streams into a set of deterministic batch computations.

However, they are designed to work in dedicated environments and they do not consider the performance variability (i.e., network, I/O, etc.) caused by resource contention in the cloud. This variability may in turn cause high and unpredictable latency when output streams are transmitted to further analysis. Moreover, they overlook the dynamic nature of data streams and the volatility in their computation requirements. Finally, they still address failures in a best-effort manner.

Our objective is to investigate new approaches for reliable, stream Big Data processing on clouds. We will explore new mechanisms that expose resource heterogeneity (observed variability in resource utilization at runtime) when scheduling stream data applications. We will also investigate how to adapt to node failures automatically, and to adapt the failure handling techniques to the characteristics of the running application and to the root cause of failures.

3.2.2. Geographically distributed workflows on multi-site clouds

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

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

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

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

Relevant groups with similar interests include the following ones.

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

3.3. Research axis 3: I/O management, in situ visualization and analysis on HPC systems at extreme scales

Over the past few years, the increasing amounts of data produced by large-scale simulations have motivated a shift from traditional offline data analysis to in situ analysis and visualization. In situ processing started by coupling a parallel simulation with an analysis or visualization library, to avoid the cost of writing data on storage and reading it back. Going beyond this simple pairwise tight coupling, complex analysis workflows today are graphs with one or more data sources and several interconnected analysis components.

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

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

Damaris (Inria, [12], [4]) exploits dedicated cores to enable jitter-free I/O and in situ visualization;

Decaf (ANL, [36]) implements a coupling service for workflows;

FlowVR (Inria, [48]) connects workflow components for in situ processing;

Swift (ANL, [51]) focuses on implicitly parallel data flows and was optimized for Big Data processing.

Our plan is to explore how these tools could best leverage their respective strengths in a *joint optimized architecture for in situ visualization and advanced processing* in the HPC area. We published a preliminary study describing the lessons learned from using these tools in production environments with real applications [6]. Such a joint architecture will contribute to address the data volume and velocity challenges raised by data-intensive workflows, including complex data-intensive analytics phases. It may also impact, in a subsequent step, future data analysis pipelines for converged Big Data and HPC architectures.

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

Relevant groups with similar interests include the following ones.

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

4. Application Domains

4.1. Application Domains

Our research work aims to improve large-scale, data-intensive applications running on clouds and extreme-scale HPC systems, with high requirements in terms of data storage and processing. Here are some classes of such applications.

Extreme-scale, data-intensive science simulations. A major research topic in the context of HPC simulations running on extreme-scale supercomputers is to explore how to record and visualize data during the simulation efficiently, without impacting the performance of the computation generating that data. In this area. We explore innovative approaches to I/O management and to in situ processing, in particular through our Damaris approach.

Map-Reduce-based data analytics. As Map-Reduce emerged as a dominant programming model for data analytics, we focus on several related challenges: how to enable fast failure recovery in shared Hadoop clusters; how to improve scheduling policies to favor resource allocation fairness; how to improve performance by detecting and mitigating stragglers.

Geographically-distributed cloud workflows. With fast-growing volumes of data to be handled at larger and larger scales, geographically distributed workflows are emerging as a natural data processing paradigm. They actually bring several benefits: resilience to failures, distribution across partitions, elastic scaling, user proximity etc. In this context, we investigate approaches to data management enabling an efficient execution of such geographically distributed workflows running on multi-site clouds. In projects like *ANR OverFlow* and *Z-CloudFlow* we explore means to better hide latency for data and metadata access and optimize transfers as a way of improving the global performance.

Stream data processing. The evolutions in the area of Big Data processing, the development of cloud computing and the success of the Map-Reduce model have fostered new types of data-intensive applications, in which obtaining fast and timely results is mandatory. Enterprises need to perform analysis on their stream data that can give fast results (i.e., in real time) at scale (e.g., click-stream analysis and network-monitoring log analysis). Similarly, scientists require fast and accurate data processing techniques in order to analyze their experimental data correctly at scale (e.g., analysis of data produced by massive-scale simulations and sensor deployments).

Besides processing, we are also focusing on efficient stream data storage. Unlike traditional storage, the main challenge of storing stream data is the large number of small items (arriving at rates easily reaching tens of millions per second). We explore the plausible paths towards a dedicated storage solution. We aim to provide on the one hand traditional storage functionality, and on the other hand stream-like performance (i.e., low-latency I/O access to items and ranges of items).

The team's projects and collaborations explicitly target concrete use cases belonging to the above application classes, in the following areas.

Smart Cities and Territories. In the framework on the *BigStorage project* where the KerData team is a major partner, we are focusing on several stream data applications in the context of Smart cities. The goal is to optimize current state-of-the-art processing engines to provide real-time analyzing of data collected from small sensors and devices. This will enable to make smart decisions in fields like healthcare, traffic management, water quality, air pollution and many more.

Climate and meteorology. An example is the atmospheric simulation code CM1 (Cloud Model 1), one of the target applications of the Blue Waters machine. We already used this code in collaborative research within *Data@Exascale* Associate Team, in the framework of the *Joint Laboratory for Extreme-Scale Computing* (JLESC), co-supported by Inria, UIUC, ANL, BSC, JSC and RIKEN/AICS.

Brain imaging. In the *A-Brain* MSR-Inria project (now completed), we applied Map-Reduce-based data analytics to neuro-imaging genetics.

Molecular biology. In the framework of the *MapReduce ANR project* led by KerData (now completed), we have focused on the *FastA* bioinformatics application used for massive protein sequence similarity searching. In the context of the *OverFlow ANR project* we are pursuing this analysis in collaboration with the Institut Français de Bioinformatique (IFB).@ We aim at using these results for drug design in an industrial context (i.e. the identification of new druggable protein targets and thereby the generation of new drug candidates).

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

SC16: Best Student Paper Finalist. The paper entitled *Týr: Blob Storage Meets Built-In Transactions* presented by Pierre Matri at the Supercomputing (SC16) Conference was one of the 7 papers selected for the Best Student Paper award.

This work was carried out in the context of the BigStorage project, under the supervision of Alexandru Costan, Gabriel Antoniu, María Pérez , and Jesús Montes.

There were 442 submissions, and 81 accepted papers.

ACM Graduate Student Research Competition SC16. Nathanaël Cheriere received the third prize in the SC16 ACM Student Research Competition for his work on optimizing the algorithms for the MPI collective *Scatter* and *AllGather* routines on the Dragonfly topology [1].

This work was carried out at the Argonne National Laboratory in the context of the JLESC, under the supervision of Matthieu Dorier, Rob Ross, Shadi Ibrahim, and Gabriel Antoniu.

As many as 62 posters were submitted for the Student Research Competition, out of which 14 have been selected in the Graduate category. After the presentation of their posters, 4 students have been invited to make a presentation of their work in front of a jury.

5.1.2. 9 papers in international journals

This year the team published 9 papers in high-quality journals including *ACM Transactions on Parallel Computing*, *IEEE Transactions on Parallel and Distributed Systems*, *Future Generation Computer Systems*, *Concurrency and Computation: Practice and Experience* and *IEEE Transactions on Cloud Computing*.

BEST PAPER AWARD:

[25]
P. MATRI, A. COSTAN, G. ANTONIU, J. MONTES, M. S. PÉREZ. *Tyr: Blob Storage Meets Built-In Transactions*, in "IEEE ACM SC16 - The International Conference for High Performance Computing, Networking, Storage and Analysis 2016", Salt Lake City, United States, November 2016, https://hal.inria.fr/hal-01347652

6. New Software and Platforms

6.1. Týr

Title: Týr: Blob Storage Meets Built-In Transactions.

Keywords: Big Data; Transactions; Tyr; BlobSeer.

Scientific Description: Týr [25] is the first blob storage system to provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency.

Functional Description: Týr offers fine-grained random write access to data and in-place atomic operations. Large-scale experiments on Microsoft Azure with a production application from CERN LHC show Týr throughput outperforms state-of-the-art solutions by more than 75%. Týr leverages the approaches developed within BlobSeer, the reference data management system for large distributed blobs, developed over the past years in KerData.

Contact data:

Participants: Pierre Matri, Alexandru Costan and Gabriel Antoniu.

Partners: INSA Rennes, Universidad Politécnica de Madrid.

Contact: Gabriel Antoniu.

URL: http://tyr.io/.

6.2. Damaris

Title: Damaris: I/O and data management for large-scale, MPI-based HPC simulations.

Keywords: I/O; HPC; Data management; Visualization; Big Data; Exascale.

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

Damaris has been preliminarily evaluated at NCSA (Urbana-Champaign) with the CM1 tornado simulation code. CM1 is one of the target applications of the Blue Waters supercomputer in production at NCSA/UIUC (USA), in the framework of the JLESC. Damaris now has external users, including (to our knowledge) visualization specialists from NCSA, Big Data experts from the HDF group, 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, CROCO). Works are in progress to evaluate it in the context of several other simulation codes including HACC (cosmology) and GTC (fusion).

Damaris is the object of a *Technical Development Action* (ADT) supported by Inria.

Functional Description: Damaris targets large-scale HPC simulations: in situ data analysis by some dedicated cores of the simulation platform; asynchronous and fast data transfer from HPC simulations to Damaris; semantic-aware dataset processing through Damaris plug-ins.

Contact data:

Participants: Matthieu Dorier (ANL), Lokman Rahmani, Gabriel Antoniu, Orçun Yildiz, Hadi Salimi and Luc Bougé.

Partners: ENS Rennes, Argonne National Laboratory.

Contact: Gabriel Antoniu.

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

6.3. Other software

6.3.1. JetStream

Title: JetStream: Enabling High-Performance Event Streaming across Cloud Data-Centers.

Keywords: Big Data, streaming, data transfer, multisite cloud.

Scientific Description. JetStream is a middleware solution for batch-based, high-performance streaming across cloud data centers. JetStream implements a set of context-aware strategies to optimize batch-based streaming, being able to self-adapt to changing conditions.

Functional Description. The system provides multi-route streaming across cloud data centers for aggregating bandwidth by leveraging the network parallelism. It enables easy deployment across .Net frameworks and seamless binding with event processing engines such as StreamInsight. JetStream is currently used at Microsoft Research ATLE Munich for the management of the Azure cloud infrastructure.

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

Contact: Alexandru Costan.

6.3.2. Omnisc'IO

Title: Omnisc'IO: a Grammar-Based Approach to Spatial and Temporal I/O Patterns Prediction.

Keywords: HPC, Input-Output, Prediction, Grammar.

Scientific Description. Omnisc'IO is a library that aims to be integrated into I/O middleware.

Functional Description. It traces I/O operations, models the stream of such operations using grammar-inference techniques, and predicts when new I/O operations will be performed, as well as where and how much data will be written.

Participants: Matthieu Dorier (ANL), Gabriel Antoniu, Shadi Ibrahim.

Contact: Gabriel Antoniu.

6.3.3. OverFlow

Title: OverFlow: Workflow Data Management as a Service for Multi-Site Applications.

Keywords: Small data; workflow; multi-site cloud.

Scientific Description. OverFlow is a uniform data management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. The software is environment-aware, as it monitors and models the global cloud infrastructure, offering high and predictable performance for transfer cost and time, within and across sites.

Functional Description. OverFlow proposes a set of pluggable services, grouped in a data-scientist cloud kit. They provide the applications with the possibility to monitor the underlying infrastructure, to exploit smart data compression, deduplication and geo-replication, to evaluate data management costs, to set a tradeoff between money and time, and optimize the transfer strategy accordingly. Currently, OverFlow is used for data transfers by the Microsoft Research ATLE Munich team as well as for synthetic benchmarks at the Politehnica University of Bucharest.

Participants: Paul Le Noac'h, Ovidiu-Cristian Marcu, Alexandru Costan and Gabriel Antoniu.

Contact: Alexandru Costan.

6.3.4. iHadoop

Title: iHadoop: A Hadoop Simulator Developed In Java on Top of SimGrid.

Keywords: Simulation, Map-Reduce, Hadoop, SimGrid.

Scientific Description. iHadoop is a Hadoop simulator developed in Java on top of SimGrid. It simulates the behavior of Hadoop and therefore accurately predicts the performance of Hadoop in normal scenarios and under failures. iHadoop is extended to (1) simulate the execution and predict the performance of multiple Map-Reduce applications; (2) simulate the execution of Map-Reduce applications under various data distributions and data skew models.

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

Participants: Shadi Ibrahim and Tien-Dat Phan.

Contact: Shadi Ibrahim.

7. New Results

7.1. Convergence of HPC and Big Data

7.1.1. Transactional storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Concurrent Big Data applications often require high-performance storage, as well as ACID (Atomicity, Consistency, Isolation, Durability) transaction support. Although blobs (binary large objects) are an increasingly popular model for addressing the storage needs of such applications, state-of-the-art blob storage systems typically offer no transaction semantics. This demands users to coordinate access to data carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior. We argue there is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications.

Týr is the first blob storage system to provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency. Týr offers fine-grained random write access to data and in-place atomic operations.

Large-scale experiments on Microsoft Azure with a production application from CERN LHC show Týr throughput outperforms state-of-the-art solutions by more than 75 %.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.1.2. Big Data on HPC

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Over the last decade, Map-Reduce has stood as the most powerful Big Data processing model. Map-Reduce model is now used by many companies and research labs to facilitate large-scale data analysis. With the growing needs of users and size of data, commodity-based infrastructure (most commonly used as of now) will strain under the heavy weight of Big Data. On the other hand, HPC systems offer a rich set of opportunities for Big Data processing.

As first steps towards Big Data processing on HPC systems, several research efforts have been devoted to understand Map-Reduce performance on these systems. Yet, the impact of the specific features of HPC environments have not been fully investigated, yet.

We conducted an experimental campaign to provide a clearer understanding of Map-Reduce performance on HPC systems. We use Spark, a widely adopted Map-Reduce framework, and representative Big Data workloads on Grid'5000 testbed to evaluate how the latency, contention and file system's configuration can influence the application performance.

7.1.3. Energy vs. performance trade-offs

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

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and provide a low response time. As the memory capacity of clusters and clouds increases, it becomes possible to keep most of the data in the main memory.

This motivates the introduction of in-memory storage systems. While prior work has focused on how to exploit the low latency of in-memory access at scale, there is still little knowledge regarding the energy efficiency of in-memory storage systems. This is unfortunate, as it is known that main memory is a major energy bottleneck in many computing systems. For instance, DRAM consumes up to 40 % of a server's power.

By means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud — a well-known in-memory storage system. We demonstrated that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also found that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly enough, we also showed that replication can even play a negative role in crash-recovery.

Collaboration. *This work was carried out in collaboration with Toni Cortes (BSC, Spain).*

7.2. Efficient I/O and communication for Extreme-scale HPC systems

7.2.1. Adaptive performance-constrained in situ visualisation

Participant: Lokman Rahmani.

While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation.

We have been working on tackling the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to their contents and filters out part of them. Thereby, the visualization pipeline is only fed with a reorganized subset of the data produced by the simulation.

Our framework, as presented in [22], leverages fast, generic evaluation procedures to score blocks of data, using information theory, statistics, and linear algebra. It monitors its own performance and dynamically adapts to achieve appropriate visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a 5-time speedup with respect to the initial visualization pipeline, and is able to meet performance constraints.

Collaboration. *This was was carried out with the collaboration of Matthieu Dorier, ANL, USA.*

7.2.2. Dragonfly

Participants: Nathanaël Cheriere, Shadi Ibrahim, Gabriel Antoniu.

High-radix direct network topologies such as Dragonfly have been proposed for Petascale and Exascale supercomputers. It has been shown that they ensure fast interconnections and reduce the cost of the network compared to traditional network topologies. However, current algorithms for communication do not consider the topology and thus waste numerous opportunities of optimization for performance.

In our studies, we exploit the strength of the Dragonfly with topology-aware algorithms for AllGather and Scatter operations. We analyze existing algorithms, then propose derived algorithms, that we evaluate using CODES, an event-driven simulator.

As expected, making AllGather algorithms topology-aware does improve the performance and reduces the link utilization. However, simulations of various Scatter algorithms show surprising results, and point out the important role played by hardware for the efficiency of the algorithms. In particular, the knowledge of the number and size of input-output buffers in routers can be exploited to accelerate the Scatter operation by a factor up to 2 times.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.2.3. Interference between HPC jobs

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

As we move toward the Exascale era, performance variability in HPC systems remains a challenge. I/O interference, a major cause of this variability, is becoming more important every day with the growing number of concurrent applications that share larger machines. Earlier research efforts on mitigating I/O interference focus on a single potential cause of interference (e.g., the network). Yet the root causes of I/O interference can be diverse.

In [27], we conducted an extensive experimental campaign to explore the various root causes of I/O interference in HPC storage systems. We used micro-benchmarks on the Grid'5000 testbed to evaluate how I/O interference is influenced by the applications' access pattern, the network components, the file system's configuration, and the backend storage devices.

Our studies revealed that in many situations interference is a result of a bad flow control in the I/O path, rather than being caused by some single bottleneck in one of its components. We further show that interference-free behavior is not necessarily a sign of optimal performance. To the best of our knowledge, our work provides the first deep insight into the role of each of the potential root causes of interference and their interplay. Our findings can help developers and platform owners improve I/O performance and motivate further research addressing the problem across all components of the I/O stack.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.3. Workflow on clouds

7.3.1. Managing hot metadata for scientific workflows on multisite clouds

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

Large-scale scientific applications are often expressed as workflows that help defining data dependencies between their different components. Such workflows may incur huge storage and computation requirements, so that they need to be processed in multiple (cloud-federated) datacenters. A major challenge in such multisite clouds is the long latency of the network links between datacenters, that limits the performance of multisite applications. Moreover, it has been shown that poor metadata handling can further impact the efficiency of computing systems. Many efforts have been done to improve metadata management; however, most of them concern only single-site, HPC systems to date.

In [26], we assert that some workflow metadata are more frequently accessed than other, and thus should be handled with higher priority during the workflow's execution. We call them *hot metadata*. We present a hybrid decentralized/distributed model for handling hot metadata in *multisite* architectures. We couple our model with a scientific workflow management system (SWfMS) to validate and tune its applicability to various real-life scientific scenarios. We show that efficient management of hot metadata improves the performance of SWfMS, reducing the workflow execution time up to 50 % for highly parallel jobs by enabling timely data provisioning and avoiding unnecessary *cold* metadata operations.

7.3.2. Probabilistic optimizations for resource provisioning of cloud workflows

Participants: Chi Zhou, Shadi Ibrahim.

In many data-intensive applications, data management routines can be represented as workflows, where tasks are organized according to data and computation dependencies. Recently, the optimal provisioning of resources (e.g., VMs) for workflows running in the cloud has attracted a lot of attention. Most resource provisioning solutions overlook the important factor of cloud dynamics, e.g., the fluctuation of I/O, network performance, and system failures. In our experiments on the Amazon EC2 cloud, these issues significantly impact resource allocation quality. Therefore, we study how cloud dynamics should be incorporated into the resource provisioning process.

Our approach models cloud dynamics as time-dependent random variables (e.g., a probability distribution of workflow execution times) and performs probabilistic optimizations for resource provisioning problems using those random variables as optimization input. This solution yields more effective resource provisioning for cloud workflows. However, it involves heavy computation effort due to the complex structures of workflows and the large number of probability calculations.

To overcome this problem, we develop a three-stage pruning process, which simplifies workflow structure and reduces probability evaluation overhead. We have also implemented our techniques in a runtime library, which allows users to integrate our techniques into their existing resource provisioning methods. Experiments on two common resource provisioning problems show that probabilistic solutions can improve the performance by 51 % —70 % compared with state-of-the-art, static solutions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.3.3. A taxonomy and survey of scientific computing in the cloud

Participants: Chi Zhou, Shadi Ibrahim.

Cloud computing has evolved as a popular computing infrastructure for many applications. With (big) data acquiring a crucial role in eScience, efforts have been made recently to develop and deploy scientific applications efficiently on the unprecedentedly scalable cloud infrastructures.

In [29], we review recent efforts in developing and deploying scientific computing applications in the cloud. In particular, we introduce a taxonomy specifically designed for scientific computing in the cloud, and further review the taxonomy with four major kinds of science applications, including life sciences, physics sciences, social and humanities sciences, and climate and earth sciences.

Due to the large data size in most scientific applications, the performance of I/O operations can greatly affect the overall performance of the applications. As a consequence, the dynamic I/O performance of the cloud has made resource provisioning an important and complex problem for scientific applications in the cloud.

We present our efforts on improving the resource provisioning efficiency and effectiveness of scientific applications in the cloud. Finally, we present the open problems for developing the next-generation eScience applications and systems in the cloud and give our conclusions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.4. Fault tolerant data processing

7.4.1. Fast recovery

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop has emerged as a prominent tool for Big Data processing in large-scale clouds. Failures are inevitable in large-scale systems, especially in shared environments. Consequently, Hadoop was designed with hardware failures in mind. In particular, Hadoop handles machine failures by re-executing all the tasks of the failed machine. Unfortunately, the efforts to handle failures are entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. This may prevent Hadoop schedulers from meeting their objectives (e.g., fairness, job priority, performance) and can significantly impact the performance of the applications.

In our previous work, we addressed this issue through the design and implementation of a new scheduling strategy called Chronos. Chronos is conductive to improving the performance of Map-Reduce applications by enabling an early action upon failure detection. Chronos tries to launch recovery tasks immediately by preempting tasks belonging to low priority jobs, thus avoiding to wait until slots are freed.

In [20], we further investigated the potential benefit of launching local recovery tasks by implementing and evaluating Chronos*. To this end, we slightly changed the smart slot allocation strategy of Chronos into aggressive slot allocation strategy. With Chronos, recovery tasks with higher priority would preempt the selected tasks with less priority. With Chronos*, we also allow recovery tasks to preempt the selected tasks with the same priority (e.g., recovery tasks belonging to the same job with selected tasks). The experimental results indicate that Chronos* results in 100 % locality execution for recovery tasks thanks to its aggressive slot allocation strategy. Moreover, Chronos* improves the completion time of the jobs by up to 17 %.

7.4.2. Dynamic replica placement

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Large-scale applications are ever-increasingly geo-distributed. Maintaining the highest possible *data locality* is crucial to ensure high performance of such applications. Dynamic replication addresses this problem by dynamically creating replicas of frequently accessed data close to the clients. This data is often stored in decentralized storage systems such as Dynamo or Voldemort, which offer support for *mutable data*.

However, existing approaches to dynamic replication for such mutable data remain centralized, thus incompatible with these systems. We introduce a write-enabled dynamic replication scheme that leverages the decentralized architecture of such storage systems. We propose an algorithm enabling clients to locate tentatively the closest data replica without prior request to any metadata node. Large-scale experiments show a read latency decrease of up to 42% compared to other state-of-the-art, caching-based solutions.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5. Advanced data management on clouds

7.5.1. Benchmarking Spark and Flink

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

Spark and Flink are two Apache-hosted data analytics frameworks that represent the state of the art in modern in-memory Map-Reduce processing. They facilitate the development of multi-step data pipelines using directly acyclic graph (DAG) patterns. In the framework of our BigStorage project, we performed a comparative study [23] which evaluates the performance of Spark versus Flink. The objective is to identify and explain the impact of the different architectural choices and the parameter configurations on the perceived end-to-end performance.

Based on empirical evidences, the study points out that in Big Data processing there is not a single framework for all data types, sizes and job patterns and emphasize a set of design choices that play an important role in the behaviour of a Big Data framework: memory management, pipelined execution, optimizations and parameter configuration easiness. What raises our attention is that a streaming engine (i.e., Flink) delivers in many benchmarks better performance than a batch-based engine (i.e., Spark), showing that a more general Big Data architecture (treating batches as finite sets of streamed data) is plausible and may subsume both streaming and batching use cases.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5.2. Geo-distributed graph processing

Participants: Chi Zhou, Shadi Ibrahim.

Graph processing is an emerging model adopted by a wide range of applications to easily parallelize the computations over graph data. Partitioning graph processing workloads to multiple machines is an important task for reducing the communication cost and improving the performance of graph processing jobs. Recently, many real-world applications store their data on multiple geographically distributed datacenters (DCs) to ensure flexible and low-latency services. Due to the limited Wide Area Network (WAN) bandwidths and the network heterogeneity of the geo-distributed DCs, existing graph partitioning methods need to be redesigned to improve the performance of graph processing jobs in geo-distributed DCs.

To address the above challenges, we propose a heterogeneity-aware graph partitioning method named G-Cut, which aims at minimizing the runtime of graph processing jobs in geo-distributed DCs while satisfying the WAN usage budget. G-Cut is a two-stage graph partitioning method. In the traffic-aware graph partitioning stage, we adopt the one-pass edge assignment to place edges into different partitions while minimizing the inter-DC data traffic size. In the network-aware partition refinement stage, we map the partitions obtained in the first stage onto different DCs in order to minimize the inter-DC data transfer time. We evaluate the effectiveness and efficiency of G-Cut using real-world graphs and the evaluation results show that G-Cut can achieve both lower WAN usage and shorter inter-DC data transfer time compared to state-of-the-art graph partitioning methods.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.5.3. Fairness and scheduling

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Recently, Map-Reduce and its open-source implementation Hadoop have emerged as prevalent tools for big data analysis in the cloud. Fair resource allocation in-between jobs and users is an important issue, especially in multi-tenant environments such as clouds. Several scheduling policies have been developed to preserve fairness in multi-tenant Hadoop clusters. At the core of these schedulers, simple (non-) preemptive approaches are employed to free resources for tasks belonging to jobs with less share. For example, Hadoop Fair Scheduler is equipped with two approaches: wait and kill. While wait may introduce a serious violation in fairness, kill may result in a huge waste of resources. Yet, recently some work have introduced preemption approach in shared Hadoop clusters.

To this end, we closely examine three approaches including wait, kill and preemption when Hadoop Fair Scheduler is employed for ensuring fair execution between multiple concurrent jobs. We perform extensive experiments to assess the impact of these approaches on performance and resource utilization while ensuring fairness. Our experimental results bring out the differences between these approaches and illustrate that these approaches are only sub-optimal for different workloads and cluster configurations: the efficiency of achieving

fairness and the overall performance varies with the workload composition, resource availability and the cost of the adopted preemption technique.

7.5.4. Stragglers in Map-Reduce

Participants: Tien-Dat Phan, Shadi Ibrahim.

Big Data systems (e.g., Map-Reduce, Hadoop, Spark) rely increasingly on speculative execution to mask slow tasks also known as stragglers because a job's execution time is dominated by the slowest task instance. Big Data systems typically identify stragglers and speculatively run copies of those tasks with the expectation a copy may complete faster to shorten job execution times.

There is a rich body of recent results on straggler mitigation in Map-Reduce. However, the majority of these do not consider the problem of accurately detecting stragglers. Instead, they adopt a particular straggler detection approach and then study its effectiveness in terms of performance, e.g., reduction in job completion time, or its efficiency, e.g., extra resource usage.

In this work, we consider a complete framework for straggler detection and mitigation. We start with a set of metrics that can be used to characterizes and detect stragglers such as Precision, Recall, Detection Latency, Undetected Time and Fake Positive. We then develop an architectural model by which these metrics can be linked to measures of performance including execution time and system energy overheads.

We further conduct a series of experiments to demonstrate which metrics and approaches are more effective in detecting stragglers and are also predictive of effectiveness in terms of performance and energy efficiency. For example, our results indicate that the default Hadoop straggler detector could be made more effective. In certain cases, precision is low and only 65 % of those detected are actual stragglers and recall, i.e., the proportion of stragglers which are actually detected, is also relatively low at 56 %. For the same case, the hierarchical approach (i.e., a green-driven detector based on the default one) achieves a precision of 98 % and a recall of 33 %.

Further, these increases in precision can be used to achieve lower execution time and energy consumption, and thus higher performance and energy efficiency. Compared to the default Hadoop mechanism, energy consumption is reduced by almost 30 %. These results demonstrate how our framework can offer useful insights and be applied in practical settings to characterize and design new straggler detection mechanisms for Map-Reduce systems.

Collaboration. *This work was carried out in collaboration with Guillaume Aupy and Padma Raghavan whilst they were affiliated with Vanderbilt University, USA.*

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Microsoft: Z-CloudFlow (2013–2016). In the framework of the Joint Inria-Microsoft Research Center, this project is a follow-up to the A-Brain project. The goal of this new project is to propose a framework for the efficient processing of scientific workflows in clouds. This approach will leverage the cloud infrastructure capabilities for handling and processing large data volumes.

In order to support data-intensive workflows, the cloud-based solution will: adapt the workflows to the cloud environment and exploit its capabilities; optimize data transfers to provide reasonable times; manage data and tasks so that they can be efficiently placed and accessed during execution.

The validation will be performed using real-life applications, first on the Grid5000 platform, then on the Azure cloud environment, access being granted by Microsoft through a *Azure for Research Award* received by G. Antoniu. The project also provides funding for the PhD thesis of Luis Pineda-Morales, started in 2014.

Collaboration. *The project is being conducted in collaboration with the Zenith team from Montpellier, led by Patrick Valduriez.*

Huawei: HIRP Low-Latency Storage for Stream Data (2016–2017). The goal of this project is to explore the plausible paths towards a dedicated storage solution for low-latency stream storage. Such a solution should provide on the one hand traditional storage functionality and on the other hand stream-like performance (i.e., low-latency I/O access to items and ranges of items).

We plan to investigate the main requirements and challenges, evaluate the different design choices (e.g., a standalone component vs. an extension of an existing Big Data solution like HDFS) and then propose an architectural overview.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

9.1.1.1. OverFlow (2015–2019)

- Project Acronym: OverFlow.
- Project Title: Workflow Data Management as a Service for Multisite Applications.
- Coordinator: Alexandru Costan.
- Duration: Octobre 2015–October 2019.
- Other Partners: None (Young Researcher Project).
- External collaborators: Kate Keahey (University of Chicago and Argonne National Laboratory), Bogdan Nicolae (Huawei Research) and Christophe Blanchet (Institut Français de Bioinformatique).
- Abstract: This JCJC project led by Alexandru Costan investigates approaches to data management enabling an efficient execution of geographically distributed workflows running on multi-site clouds. Ultimately, OverFlow will propose a new, pioneering paradigm: Workflow Data Management as a Service — a general and easy-to-use, cloud-provided service that bridges for the first time the gap between single- and multi-site workflow data management. It aims to reap economic benefits from the geo-diversity while accelerating the scientific discovery through a democratization of access to globally distributed data.

9.1.2. Other National Projects

9.1.2.1. DISCOVERY (2015–2019)

- Project Acronym: DISCOVERY.
- Project Title: DIStributed and COoperative framework to manage Virtual EnviRonments autonomicallY.
- Coordinator: Adrien Lèbre.
- Duration: 2015–2019.
- Partners: Inria Project-Teams including ASAP, ASCOLA, Avalon, Myriads, and KerData.
- Abstract: An Inria Project Lab, led by Adrien Lèbre (ASCOLA), that aims at exploring a new way of operating Utility Computing (UC) resources by leveraging any facilities available through the Internet. The goal is to deliver widely distributed platforms that can better match the geographical dispersal of users, as well as the unending demand.

  Within DISCOVERY, S. Ibrahim (KerData Inria Team) is working with Gilles Fedak (Avalon Inria Project-Team) to address the VM images management challenge.

9.1.2.2. ADT Damaris

- Project Acronym: ADT Damaris
- Project Title: Technology development action for te Damaris environment.

- Coordinator: Alexandru Costan.
- Duration: 2016–2018.
- Abstract: This action aims to support the development of the Damaris software. Inria's *Technological Development Office* (D2T, *Direction du Développement Technologique*) provided 2 years of funding support for a senior engineer.

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

9.1.2.3. Grid'5000.

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

9.2. European Initiatives

9.2.1. FP7 and H2020 Projects

9.2.1.1. BigStorage

- Title: BigStorage: Storage-based Convergence between HPC and Cloud to handle Big Data.
- Programme: H2020.
- Duration: January 2015–December 2018.
- Coordinator: Universidad Politécnica de Madrid (UPM).
- Partners:
  - Barcelona Supercomputing Center — Centro Nacional de Supercomputacion (Spain)
  - CA Technologies Development Spain (Spain)
  - CEA — Commissariat à l'énergie atomique et aux énergies alternatives (France)
  - Deutsches Klimarechenzentrum (Germany)
  - Foundation for Research and Technology Hellas (Greece)
  - Fujitsu Technology Solutions (Germany)
  - Johannes Gutenberg Universitaet Mainz (Germany)
  - Universidad Politecnica de Madrid (Spain)
  - Seagate Systems UK (United Kingdom)
- Inria contact: G. Antoniu and Adrien Lèbre.
- URL: http://www.bigstorage-project.eu/.
- Description: BigStorage is a European Training Network (ETN) whose main goal is to train future *data scientists*. It aims at enabling them and us to apply holistic and interdisciplinary approaches to take advantage of a data-overwhelmed world. This world requires *HPC* and *Cloud* infrastructures with a redefinition of *storage* architectures underpinning them — focusing on meeting highly ambitious performance and *energy* usage objectives. The KerData team will be hosting 2 *Early Stage Researchers* in this framework.

9.3. International Initiatives

9.3.1. Inria International Labs

9.3.1.1. JLESC: Joint Laboratory on Extreme-Scale Computing

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

The KerData team is collaborating with teams from ANL and UIUC within this lab since 2009 on several topics in the areas of I/O, storage and in situ processing and cloud computing. This collaboration has been initially formalized as the *Data@Exascale* Associate Team with ANL and UIUC (2013–2015) followed by *Data@Exascale 2* Associate Team with ANL (2016–2018).

Since 2015, Gabriel Antoniu serves as a topic leader for Inria for the *I/O, Storage and In Situ Processing* topic.

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

Project Acronym: Data@Exascale 2.

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

International Partner:

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

Start year: 2013.

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

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

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

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

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

9.3.2. Inria International Partners

9.3.2.1. DataCloud@Work

Title: DataCloud@Work.

International Partner:

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

Duration: 4 years.

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

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

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

9.3.3. Informal International Partners

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

9.4. International Research Visitors

9.4.1. Visits of International Scientists

Guillaume Aupy (Vanderbilt University) visited the KerData team for one week (February 2016).

9.4.2. Visits to International Teams

9.4.2.1. Research Stays Abroad

CIC-IPN, Mexico:
Participants: Gabriel Antoniu, Alexandru Costan, Luis Eduardo Pineda Morales, Pierre Matri.

From October 31 to November 4, four members of our team visited the Informatics Research Centre of the National Polytechnic Institute (CIC-IPN for its acronym in Spanish) in Mexico City, Mexico.

The visit was a follow up to previous discussions held with the Network and Data Science Laboratory. The goal is to create a scientific collaboration on the grounds of cloud-based big data for smart cities, for which a proposal has been submitted in August to the program ECOS-NORD (Mexico-France). The visit included scientific presentations from both teams, a plenary talk from KerData to the IPN community, as well as discussions on future common research lines. Additionally, we held meetings with the partnering coordinator to talk about possible funding sources for students exchanges.

ANL, USA:
Participant: Nathanaël Cheriere.

Nathanaël Cheriere visited Matthieu Dorier and Rob Ross at ANL for 5.5 months, co-funded by the PUF NextGen project in the context of the Joint Laboratory for Extreme-Scale Computing (JLESC).

Vanderbilt University, USA:
Participant: Tien-Dat Phan.

Tien-Dat Phan visited(Guillaume Aupy, Padma Raghavan at Vanderbilt University for 2 months, funded by Vanderbilt University.

Technische Universitat Munchen and Huawei Research Center in Munich:
Participant: Ovidiu-Cristian Marcu.

Ovidiu-Cristian Marcu is doing an internship at Huawei in Munich, Germany for 4 months, starting October 2016. The goal is to create a framework to improve memory management for streaming systems.

National University of Singapore, Singapore:
Participant: Tien-Dat Phan.

Tien-Dat Phan is visiting NUS (Bingsheng He) for 3 months, co-funded by a Mobility grant from University Bretagne Loire (UBL) and NUS.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

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

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

- Gabriel Antoniu: Vice-Chair of the Program Committee of the ACM/IEEE CCGrid 2016 international conference (Hybrid and Mobile Clouds Tracks), Cartagena, May 2016.
- Alexandru Costan: Program Co-Chair of the ScienceCloud 2016 international workshop held in conjunction with HPDC 2016, Kyoto, June 2016.

10.1.2.2. Member of the Conference Program Committees

- Gabriel Antoniu: ACM HPDC 2016, IEEE Cluster 2016, PDSW-DISCS workshop (held in conjunction with ACM/IEEE SC16 conference), ARMS-CC 2016 workshop (held in conjunction with the PODC 2016 conference).
- Luc Bougé: Euro-Par 2016, IPDPS 2017, ICDE 2017, Euro-Par 2017, ISPDC 2017.
- Alexandru Costan: ACM/IEEE SC'16 BoF Applications Track, ACM/IEEE CCGrid 2016, IEEE BigData 2016, ICPP 2016, ARMS-CC 2016 workshop (held in conjunction with PODC 2016), FiCLOUD 2016, ScienceCloud 2016 workshop (held in conjunction with HPDC 2016).
- Shadi Ibrahim: IEEE Cluster 2016, IEEE/ACM CCGrid 2016, IEEE ICPADS 2016, IEEE CloudCom 2016, IEEE ICA3PP 2016, SCRAMBL 2016 (held in conjunction with Euro-Par 2016).

10.1.2.3. Reviewer

- Alexandru Costan: ACM/IEEE SC16, ACM HPDC 2016, IEEE Cluster 2016.
- Shadi Ibrahim: ACM HPDC 2016, Euro-Par 2016.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Gabriel Antoniu: Future Generation Computer Systems, Special Issue on Resource Management for Big Data Platforms.
- Luc Bougé: Concurrency and Computation: Practice and Experience, Special Issues on the Euro-Par conference.
- Alexandru Costan: Soft Computing Journal, Special Issue on Autonomic Computing and Big Data Platforms

10.1.3.2. Reviewer, Reviewing Activities

- Alexandru Costan: IEEE Transactions on Parallel and Distributed Systems, Future Generation Computer Systems, Concurrency and Computation Practice and Experience, IEEE Communications, IEEE Transactions on Storage, Information Sciences
- Shadi Ibrahim: IEEE Transactions on Parallel and Distributed Systems, Future Generation Computer Systems, IEEE Transactions on Big Data, IEEE Transactions on Cloud Computing, Springer Parallel Computing, Computers and Electrical Engineering, Journal of Healthcare Engineering

10.1.4. Invited Talks

- Gabriel Antoniu:

  - BDEC 2016: Invited keynote talk at the *4th Big Data and Exascale Computing (BDEC) workshop*, Frankfurt, June 2016.
  - First Chinese-French Workshop on Extreme Computing: *Damaris: Jitter-Free I/O Management and In Situ Visualization of HPC Simulations using Dedicated Cores*, Guangzhou, May 2016.
  - 5th JLESC workshop: *Spark versus Flink: Understanding Performance in Big Data Analytics Frameworks*, Lyon, June 2016.
  - Inria/CIC-IPN workshop: *Scalable Big Data Processing on Clouds: A-Brain and Z-CloudFlow*, Mexico City, November 2016.
  - Inria/Technicolor workshop: *Spark versus Flink: Understanding Performance in Big Data Analytics Frameworks*, Rennes, November 2016.
  - 6th JLESC workshop: *Storage-Based Convergence Between HPC and Big Data*, Kobe, Japan, December 2016.
- Luc Bougé:
  - Comin Labs-DGA-ENSAI BigData day: *Support logiciel pour la gestion de données distribuées à très grande échelle*, IRISA, January 2016.
  - Société des agrégés: *Teaching informatics as a first-class subject*, annual meeting of the Regional Section, April 2016.
  - Luminy Algorithmics and Programming School: *Big Data: Tremendous challenges, great solutions*, Preparatory school teachers in Mathematics and Informatics, May 2016.
- Alexandru Costan:
  - UPB Scientific Days: *Big Data and Extreme Computing: A Storage-Based Pathway to Convergence*, The UPB Research Workshop on Distributed Systems, University Politehnica of Bucharest, June 2016.
  - Inria/CIC-IPN workshop: *Science Driven, Scalable Data-Intensive Processing on Clouds*, Mexico City, November 2016.
- Shadi Ibrahim:
  - Inria Scientific Days: *Big Data management at scale*, Rennes, June 2016.
- Chi Zhou:
  - 5th JLESC workshop: *Incorporating Probabilistic Optimizations for Resource Provisioning of Cloud Workflow Processing*, Lyon, June 2016.
- Nathanaël Cheriere:
  - 6th JLESC Workshop: *Accelerating the Scatter Operation on Dragonfly Networks*, Kobe, Japan, December 2016.
- Orçun Yildiz:
  - Grid'5000 Winter School: *Investigating the Root Causes of I/O Interference on Grid'5000*, Grenoble, February 2016.
  - 6th JLESC Workshop: *Towards Efficient Big Data Processing in HPC Systems*, Kobe, Japan, December 2016.
- Luis Eduardo Pineda Morales:
  - 5th JLESC workshop: *Exploring Elastic Scaling on Chameleon Cloud*, Lyon, June 2016.
  - Inria / CIC-IPN workshop: *Metadata Management for Geo-distributed Cloud Workflows*, Mexico City, November 2016.

10.1.5. Leadership within the Scientific Community

- Gabriel Antoniu: Scientific leader of the KerData project-team.

- Gabriel Antoniu: Topic leader for Inria for the *Data storage, I/O and in situ processing* topic, supervising collaboration activities in this area within the JLESC, Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing.
- Luc Bougé: serves as a Vice-President of the *French Society for Informatics* (SIF), in charge of the teaching department.
- Gabriel Antoniu: Work package leader within the BigStorage H2020 ETN project for the *Data Science* work package.
- Alexandru Costan: Leader of the *Smart Cities* Working Group within the BigStorage H2020 ETN project.
- Shadi Ibrahim: Leader for the *Resource Management and Scheduling for Data-Intensive HPC Workflows* activity within the JLESC, Joint Inria-Illinois-ANL-BSC-JSC-RIKEN/AICS Laboratory for Extreme-Scale Computing.

10.1.6. Scientific Expertise

- Gabriel Antoniu served as a project evaluator for the ANR 2016 call (Phase 1).
- Luc Bougé: Member of the jury for the *Agrégation de mathématiques* and the *CAPES of mathématiques*. These national committees select high-school mathematics teachers in secondary schools and high-schools, respectively.
- Luc Bougé has been solicited by the Ministry of Education to participate to the committee in charge of designing and preparing the new *Informatics track* in the CAPES of mathematics. It will be offered for the 2017 session.
- Shadi Ibrahim served as a project evaluator in the DOE-ECP Program 2016: The research and development in Software Technology of the US Department of Energy's (DOE's) *Exascale Computing Project* (ECP).

10.1.7. Research Administration

- Luc Bougé: Nominated to seat at the CNU (*National University Council*) in the *Informatics* Section (27). His term ended in November 2016.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Gabriel Antoniu

- Master (Engineering Degree, 5th year): Big Data, 24 hours (lectures), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz, France.
- Master : Cloud Computing, 15 hours (lectures and lab sessions), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz, France.
- Master: Distributed Systems, 8 hours (lectures), M2 level, ALMA Master, Distributed Architectures module, University of Nantes, France.
- Master: Scalable Distributed Systems, 12 hours (lectures), M1 level, SDS Module, EIT ICT Labs Master School, France.

Luc Bougé

- Bachelor: Introduction to programming concepts, 36 hours (lectures), L3 level, Informatics program, ENS Rennes, France.
- Master: Introduction to compilation, 24 hours (exercice and practical classes), M1 level, Informatics program, Univ. Rennes I, France.

Alexandru Costan

- Bachelor: Software Engineering and Java Programming, 28 hours (lab sessions), L3, INSA Rennes.
- Bachelor: Databases, 68 hours (lectures and lab sessions), L2, INSA Rennes, France.
- Bachelor: Practical case studies, 24 hours (project), L3, INSA Rennes.
- Master: Big Data and Applications, 36h hours (lectures, lab sessions, project), M1, INSA Rennes.

Shadi Ibrahim

- Master (Engineering Degree, 5th year): Big Data, 24 hours (lectures and lab sessions), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz, France.
- Master : Cloud Computing and Hadoop Technologies, 16.5 hours (lectures and lab sessions), M2 level, ENSAI (*École nationale supérieure de la statistique et de l'analyse de l'information*), Bruz, France.
- Master: Cloud and Big data, 24 hours (lectures and lab sessions), M1 level, ENS Rennes, France.
- Master: Cloud1, Map-Reduce, (lectures, lab sessions), 15 hours (lectures and lab sessions), M2 level, Ecole des Mines de Nantes (EMN Nantes), Nantes, France.

10.2.2. Supervision

10.2.2.1. 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.

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. Defense planned in Spring 2017.

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

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

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

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

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

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

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

10.2.3. Juries

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

Luc Bougé: Referee for the PhD thesis of Matthieu Perrin, LINA, Univ. Nantes (June 2016). Member of several PhD and HDR thesis juries in France.

10.2.4. Miscellaneous

10.2.4.1. Responsibilities

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

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

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

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

Shadi Ibrahim: Member of Grid'5000 Sites Committee: Responsible for the Rennes site.

10.2.4.2. Tutorials

Gabriel Antoniu and Shadi Ibrahim gave tutorials on *Big Data technologies and Hadoop* at the BigStorage Winter School in Barcelona (March 2016).

Shadi Ibrahim gave a Tutorial on *Green Big Data Processing using Hadoop* at the at the Euro-Par 2016 conference, Grenoble, France (with Anne-Cécile Orgerie).

10.3. Popularization

Luc Bougé:

Master Program, Rennes. Invited presentation to the M2 students about *Informatics as a scientific activity: Toward a responsible research* (December 2016).

Alexandru Costan:

Master Program, Rennes. Invited presentation to the M2 students about *Big Data Analytics* (November 2016).

11. Bibliography

Major publications by the team in recent years

[1] N. CHERIERE, M. DORIER. *Design and Evaluation of Topology-aware Scatter and AllGather Algorithms for Dragonfly Networks*, November 2016, Supercomputing 2016, Poster, https://hal.inria.fr/hal-01400271

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

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

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

[5] M. DORIER, G. ANTONIU, 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, M. DREHER, T. PETERKA, G. ANTONIU, B. RAFFIN, J. M. WOZNIAK. *Lessons Learned from Building In Situ Coupling Frameworks*, in "ISAV 2015 - First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization (held in conjunction with SC15)", Austin, United States, November 2015 [*DOI :* 10.1145/2828612.2828622], https://hal.inria.fr/hal-01224846

[7] M. DORIER, S. IBRAHIM, G. ANTONIU, R. ROSS. *Omnisc'IO: A Grammar-Based Approach to Spatial and Temporal I/O Patterns Prediction*, in "SC14 - International Conference for High Performance Computing, Networking, Storage and Analysis", New Orleans, United States, IEEE, ACM, November 2014, https://hal.inria.fr/hal-01025670

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

[9] B. NICOLAE, G. ANTONIU, L. BOUGÉ, D. MOISE, A. CARPEN-AMARIE. *BlobSeer: Next-Generation Data Management for Large-Scale Infrastructures*, in "JPDC - Journal of Parallel and Distributed Computing", February 2011, vol. 71, n° 2, pp. 169–184, http://hal.inria.fr/inria-00511414/en/

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

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

Publications of the year

Articles in International Peer-Reviewed Journals

[12] M. DORIER, G. ANTONIU, F. CAPPELLO, M. SNIR, R. R. SISNEROS, O. YILDIZ, S. IBRAHIM, T. PETERKA, L. G. ORF. *Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations*, in "ACM Transactions on Parallel Computing", 2016, https://hal.inria.fr/hal-01353890

[13] M. DORIER, S. IBRAHIM, G. ANTONIU, R. ROSS. *Using Formal Grammars to Predict I/O Behaviors in HPC: the Omnisc'IO Approach*, in "IEEE Transactions on Parallel and Distributed Systems", 2016 [*DOI :* 10.1109/TPDS.2015.2485980], https://hal.inria.fr/hal-01238103

[14] M. DORIER, O. YILDIZ, S. IBRAHIM, A.-C. ORGERIE, G. ANTONIU. *On the energy footprint of I/O management in Exascale HPC systems*, in "Future Generation Computer Systems", March 2016, vol. 62, pp. 17–28 [*DOI :* 10.1016/J.FUTURE.2016.03.002], https://hal.inria.fr/hal-01330735

[15] S. IBRAHIM, T.-D. PHAN, A. CARPEN-AMARIE, H.-E. CHIHOUB, D. MOISE, G. ANTONIU. *Governing Energy Consumption in Hadoop through CPU Frequency Scaling: an Analysis*, in "Future Generation Computer Systems", January 2016, 14 p. [*DOI :* 10.1016/J.FUTURE.2015.01.005], https://hal.inria.fr/hal-01166252

[16] Y. SIMMHAN, L. RAMAKRISHNAN, G. ANTONIU, C. GOBLE. *Cloud computing for data-driven science and engineering: Special issue on the Cloud computing for data-driven science and engineering workshop (ScienceCloud 2012)*, in "Concurrency and Computation: Practice and Experience", 2016, vol. 28, n° 4, pp. 947–949 [*DOI :* 10.1002/CPE.3668], https://hal.inria.fr/hal-01351218

[17] R. TUDORAN, A. COSTAN, G. ANTONIU. *OverFlow: Multi-Site Aware Big Data Management for Scientific Workflows on Clouds*, in "IEEE Transactions on Cloud Computing", 2016 [*DOI :* 10.1109/TCC.2015.2440254], https://hal.inria.fr/hal-01239128

[18] R. TUDORAN, A. COSTAN, O. NANO, I. SANTOS, H. SONCU, G. ANTONIU. *JetStream: Enabling high throughput live event streaming on multi-site clouds*, in "Future Generation Computer Systems", January 2016, vol. 54 [*DOI :* 10.1016/J.FUTURE.2015.01.016], https://hal.inria.fr/hal-01239124

[19] S. WU, S. TAO, X. LING, H. FAN, H. JIN, S. IBRAHIM. *iShare: Balancing I/O performance isolation and disk I/O efficiency in virtualized environments*, in "Concurrency and Computation: Practice and Experience", 2016, https://hal.inria.fr/hal-01338404

[20] O. YILDIZ, S. IBRAHIM, G. ANTONIU. *Enabling Fast Failure Recovery in Shared Hadoop Clusters: Towards Failure-Aware Scheduling*, in "Future Generation Computer Systems", March 2016 [*DOI :* 10.1016/J.FUTURE.2016.02.015], https://hal.inria.fr/hal-01338336

International Conferences with Proceedings

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

[22] M. DORIER, R. R. SISNEROS, L. BAUTISTA-GOMEZ, T. PETERKA, L. ORF, L. RAHMANI, G. ANTONIU, L. BOUGÉ. *Adaptive Performance-Constrained In Situ Visualization of Atmospheric Simulations*, in "Cluster 2016 - The IEEE 2016 International Conference on Cluster Computing", Taipei, Taiwan, 2016, https://hal.inria.fr/hal-01351919

[23] O.-C. MARCU, A. COSTAN, G. ANTONIU, M. S. PÉREZ. *Spark versus Flink: Understanding Performance in Big Data Analytics Frameworks*, in "Cluster 2016 - The IEEE 2016 International Conference on Cluster Computing", Taipei, Taiwan, September 2016, https://hal.inria.fr/hal-01347638

[24] P. MATRI, A. COSTAN, G. ANTONIU, J. MONTES, M. S. PÉREZ. *Towards Efficient Location and Placement of Dynamic Replicas for Geo-Distributed Data Stores*, in "ScienceCloud'16 - 7th Workshop on Scientific Cloud Computing (in conjunction with ACM HPDC 2016)", Kyoto, Japan, June 2016 [*DOI :* 10.1145/2913712.2913715], https://hal.inria.fr/hal-01304328

[25] *Best Paper*
P. MATRI, A. COSTAN, G. ANTONIU, J. MONTES, M. S. PÉREZ. *Tyr: Blob Storage Meets Built-In Transactions*, in "IEEE ACM SC16 - The International Conference for High Performance Computing, Networking, Storage and Analysis 2016", Salt Lake City, United States, November 2016, https://hal.inria.fr/hal-01347652.

[26] L. PINEDA-MORALES, J. LIU, A. COSTAN, E. PACITTI, G. ANTONIU, P. VALDURIEZ, M. MATTOSO. *Managing Hot Metadata for Scientific Workflows on Multisite Clouds*, in "BIGDATA 2016 - 2016 IEEE International Conference on Big Data", Washington, United States, December 2016, https://hal.inria.fr/hal-01395715

[27] O. YILDIZ, M. DORIER, S. IBRAHIM, R. ROSS, G. ANTONIU. *On the Root Causes of Cross-Application I/O Interference in HPC Storage Systems*, in "IPDPS 2016 - The 30th IEEE International Parallel and Distributed Processing Symposium", Chicago, United States, May 2016, https://hal.inria.fr/hal-01270630

Scientific Books (or Scientific Book chapters)

[28] B. MEMISHI, S. IBRAHIM, M. S. PÉREZ-HERNÁNDEZ, G. ANTONIU. *On the Dynamic Shifting of the MapReduce Timeout*, in "Managing and Processing Big Data in Cloud Computing", R. KANNAN, R. U. RASOOL, H. JIN, S. BALASUNDARAM (editors), IGI Global, 2016, https://hal.inria.fr/hal-01338393

[29] A. C. ZHOU, B. HE, S. IBRAHIM. *A Taxonomy and Survey of Scientific Computing in the Cloud*, in "Big Data: Principles and Paradigms", eScience and Big Data Workflows in Clouds: A Taxonomy and Survey, Morgan Kaufmann, June 2016, https://hal.inria.fr/hal-01346745

Research Reports

[30] M. DORIER, R. R. SISNEROS, L. BAUTISTA-GOMEZ, T. PETERKA, L. G. ORF, R. ROSS, L. RAHMANI, G. ANTONIU, L. BOUGÉ. *Performance-Constrained In Situ Visualization of Atmospheric Simulations*, Inria Rennes - Bretagne Atlantique, February 2016, n^{o} RR-8855, 27 p. , https://hal.inria.fr/hal-01273718

[31] P. MATRI, A. COSTAN, G. ANTONIU, J. MONTES, M. S. PÉREZ. *Týr: Efficient Transactional Storage for Data-Intensive Applications*, Inria Rennes Bretagne Atlantique ; Universidad Politécnica de Madrid, January 2016, n^{o} RT-0473, 25 p. , https://hal.inria.fr/hal-01256563

Other Publications

[32] N. CHERIERE, M. DORIER. *Design and Evaluation of Topology-aware Scatter and AllGather Algorithms for Dragonfly Networks*, November 2016, Supercomputing 2016, Poster, https://hal.inria.fr/hal-01400271

[33] L. RAHMANI, M. DORIER, L. BOUGÉ, G. ANTONIU, R. R. SISNEROS, T. PETERKA. *Towards Smart Visualization Framework for Climate Simulations*, March 2016, working paper or preprint, https://hal.inria.fr/hal-01290268

[34] Y. TALEB, S. IBRAHIM, G. ANTONIU, T. CORTES. *Understanding how the network impacts performance and energy-efficiency in the RAMCloud storage system*, October 2016, working paper or preprint, https://hal.inria.fr/hal-01376923

References in notes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[49] T. AKIDAU, A. BALIKOV, K. BEKIROĞLU, S. CHERNYAK, J. HABERMAN, R. LAX, S. MCVEETY, D. MILLS, P. NORDSTROM, S. WHITTLE. *MillWheel: fault-tolerant stream processing at internet scale*, in "Proceedings of the VLDB Endowment", 2013, vol. 6, n^{o} 11, pp. 1033–1044

[50] 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

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