DASH: Data-Aware Scheduling at Higher scale

Ordonnancement de données pour le calcul haute-performance

Défi 7 > Axe 7 > Infrastructures de HPC et traitement massif de données

Guillaume Aupy,
TADaaM project-team, INRIA Bordeaux

Appel à projets générique 2017

Jeunes chercheuses et jeunes chercheurs

Contents

Abstract

While computing power of supercomputers keeps on increasing at an exponential rate, their capacity to manage data movement experiences some limits. It is expected that this imbalance will be one of the key limitation to the development of future HPC applications. We propose to rethink how I/O is managed in supercomputers. More specifically, the novelty of this project is to account for known HPC application behaviors (periodicity, limited number of concurrent applications) to define static strategies. We expect that those strategies can be turned into more efficient dynamic strategies than current strategies.

In this study, we plan to include a dynamicity provision to cope with any uncertain behavior of applications. We also plan to research how to model and include emerging technologies. Finally we plan to explore the importance and impact of reliabilty and energy-efficiency into I/O management strategies.

Main changes with regard to the pre-proposition

The current proposition is in the line to the pre-preposition on the research directions considered. There are however some modification made after taking into account the evaluations. Specifically, we have extended the experimental portion of the project in order to increase dissemination and exploitation of our results:

- The postdoc position is now an engineer position to be able to develop a library based on the obtained results;
- The project is now planned on 48 months instead of 36 so that the library can contain all results obtained during the PhD;
- A collaboration has been initiated with Florin Isaila (UC3M) to provide expertise and a framework to integrate and evaluate the results (see Section 2.2, Task 4);
- A collaboration has been initiated with an industrial partner, Jean-Thomas Acquaviva (DDN) to integrate and evaluate the use of Burst-Buffers.

To accomodate this, we made some minor changes budget-wise: (i) the travel budget has been increased to take into account the collaboration with Florin Isaila; (ii) conference travel costs have also been increased to cover the additional year added to the project. Finally, we plan to organize a workshop dedicated to the problematics around I/O management at the beginning of the third year of the project in order to increase collaboration between the French I/O community (research teams in Rennes, and Grenoble, but also industrial partners) and some European partners (Madrid, Barcelona).

Project participants

Organization	Name	Professional Title	Contribution *research time*	Role and Responsibilities
Inria Bordeaux - Sud Ouest	Guillaume Aupy	CR INRIA	24PM (50%)	Project leader: he will ensure that everything goes according to plan. He will co-supervise the PhD student and supervise the interns and engineer
Inria Bordeaux - Sud Ouest	Emmanuel Jeannot	DR INRIA	6PM	He will help in supervising the PhD student
Inria Bordeaux - Sud Ouest	XXX	Research Intern funded by ANR	6PM	They will work in defining the problem and models
Inria Bordeaux - Sud Ouest	XXX	PhD Student funded by ANR	36PM	They will work in the algorithm design and implementaton
Inria Bordeaux - Sud Ouest	XXX	Engineer funded by ANR	12PM	They will help with the software development and dissemination of the results
Inria Bordeaux - Sud Ouest	XXX	Research Intern funded by ANR	6PM	They will work in exploring problematics related to reliability and energy-efficiency

1 Context, position, objectives

1.1 Context

In the race to larger supercomputers, the most commonly used metric is the peak computational power. However supercomputers are not simply computers with billions of processors. One of the reason why Sunway TaihuLight (the world fastest supercomputer as of June 2016), reaches 93 PetaFlops on HPL (performance benchmark based on dense linear algebra), but struggles to reach 0.37 PetaFlop on HPCG, a recent benchmark based on actual HPC applications [14] is data movement.

HPC applications and volume of data generated Nowadays, supercomputing applications create or have to deal with TeraBytes of data. This is true in all fields: as example LIGO (gravitational wave detection) generates 1500TB/year [31], the Large Hadron Collider generates 15PB/year, light sources project deal with 300TB of data per day and climate modelling are expected to have to deal with 100EB of data [23]. According to experts "Very few large scale applications of practical importance are not data intensive" (Alok Choudhary, Apr 2012).

Data management as a bottleneck for HPC applications Management of I/O operations is critical at scale. However, observations on the Intrepid machine at Argonne National Lab is that I/O transfer can be slowed down up to 70% due to congestion [21]. In 2013, Argonne upgraded its house supercomputer: moving from Intrepid (Peak performance: 0.56 PFlop/s; peak I/O throughput: 88 GB/s) to Mira (Peak performance: 10 PFlop/s; peak I/O throughput: 240 GB/s). In 2018, the new machine at Argonne, Aurora, is expected to have a Peak performance of 450 PFlops/s and a peak I/O throughput of 1 TB/s. While both criteria seem to continuously improve considerably, the reality behind is that for a given application, its I/O throughput scales linearly (or worse) with its performance, and hence, what should be noticed is a downgrade from 160 GB/PFlop (Intrepid) to 24 GB/PFlop (Mira) and finally 2.2 GB/PFlop (Aurora)!

With this in mind, to be able to scale, conception of new algorithms has to change paradigm: going from a compute-centric model to a data-centric model.

1.2 Positioning with respect to TADAAM + Inria Bordeaux

In December 2016, Guillaume Aupy has joined the TADaaM Inria project team.

The goal of the TADaaM project is to design and build a stateful system-wide service layer for HPC systems. This layer will be twofold. First, it will abstract low-level features of the system (e.g. topology, network, resource usage) and of the software stack (e.g. threads, data, runtime system). Second, applications will be able to register their needs and behavior thanks to a carefully designed API. With these two sets of information, the layer will optimize the execution of all the running applications in a coordinated fashion and at system-scale.

One of the key element of Guillaume Aupy's recruitment in TADaaM was to strengthen the I/O expertise of the team. This project will complement data-management performed at other level (memory, network) by other members in the team. Finally, this project also takes information from two levels, architecture and applications, hence in this regard, Guillaume will be able to use expertise already developed in the TADaaM project.

Finally, there is currently a consortium of different researchers from HPC, BigData, Applications being built at Inria Bordeaux in the context of the local LABEX. The goal is to make large scale computing meet Data Intensive Science. Guillaume Aupy is part of this consortium and the results of this ANR would fit perfectly into the proposal.

1.3 Scientific and technical description

1.3.1 State of the art in I/O management for supercomputers

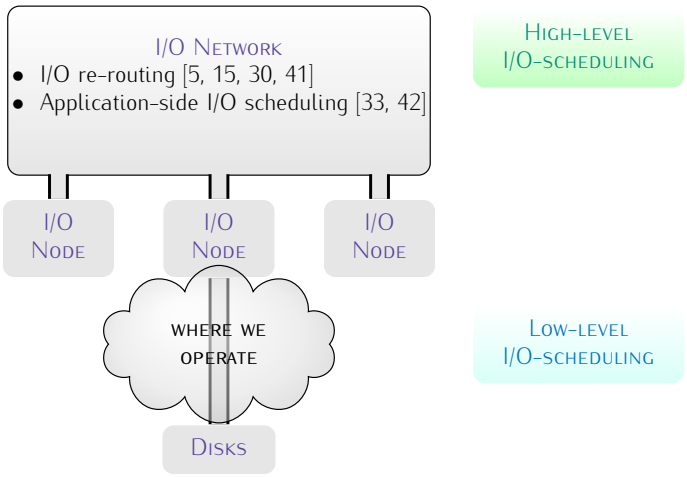
To deal with performance variability due to I/O management, many studies have analyzed the sources of performance degradation and several solutions have been proposed. In this section, we detail some of the existing work that copes with I/O congestion.

Data transformation To deal with the amount of data, recent work studied application-side I/O management and transformation. Amongst them, Lofstead et al. [33] propose adaptative strategies to deal with IO variability due to congestion by modifying at certain times both the number of process sending data, and the size of the data being sent. More recently Tessier et al. [38] study the impact of the locality of aggregate nodes. Those nodes are compute nodes dedicated to aggregating data sent by the other compute nodes during the I/O phase of an application. Those nodes also have the possibility to transform the data being sent (for instance by compressing it [16]). To reduce even more the amount of data being transferred, Di and Cappello [13] have discussed the advantage of compressing the data sent in a lossy way.

To deal with file systems reaching their limit, recent works [19] have started studying in-situ/in-transit analysis. In the past, some workflows created the data, stored it on disks before analyzing it as a second step. In-situ/in-transit analysis offers to dedicate some specific nodes to the analysis and to perform it as the data is created. Using this, the hope is to reduce the load on the file systems.

We consider all these solutions to be orthogonal to the problem studied here and hence can be used conjointly.

Software to deal with I/O movement There are many levels where one can help to reduce I/O congestion. Most of the work focus on high-level I/O scheduling, including I/O re-routing and application-side I/O scheduling. This proposal looks at the low-level of I/O scheduling.



Some papers consider application-side I/O scheduling [33, 42]. Recently, numerous works focus on using machine learning for auto tuning and performance studies [5, 30]. However these solution also work at the application level for IO-scheduling and do not have a global view of the I/O requirements of the system and they need to be supported by a platform level I/O management for better results. Cross-application contention has been studied recently [24, 37, 39]. The study in [24] evaluates the performance degradation in each application program when Virtual Machines (VMs) are executing two application programs concurrently in a physical computing server. The experimental results indicate

that the interference among VMs executing two HPC application programs with high memory usage and high network I/O in the physical computing server significantly degrades application performance. An earlier study in 2005 [37] cites application interference as one of the main problems facing the HPC community. While the authors propose ways of gaining performance by reducing variability, minimizing application interference is still left open. In [41], a more general study analyzes the behavior of the center wide shared Lustre parallel file system on the Jaguar supercomputer and its performance variability. One of the performance degradations seen on Jaguar was caused by concurrent applications sharing the filesystem. All these studies highlight the impact of having application interference on HPC systems, without, but they do not offer a solution.

Closer to this work, online schedulers for HPC systems were developed such as our previous work [21], the study by Zhou et al [43], and a solution proposed by Dorier et al [15]. In [15], the authors investigate the interference of two applications and analyze the benefits of interrupting or delaying either one in order to avoid congestion. Unfortunately their approach cannot be used for more than two applications. Another main difference with our previous work is the light-weight approach of this study where the computation is only done once.

Clarisse [26] proposes mechanisms for designing and implementing cross-layer optimizations of the I/O software stack. The specific implementation of the problem considered here is a naive First Come First Served approach. They however provide an excellent opportunity to study our results in a real framework. We discuss in Section 2.2 a collaboration with the group in charge of Clarisse to evaluate the algorithms and libraries developed in this proposal.

Hardware solutions Recently, architectural solutions such as burst buffers were introduced to reduce I/O congestion. They act by delaying accesses to the file storage, as they found that congestion occurs on a short period of time and the bandwidth to the storage system is often underutilized [32]. Note that because the computation power increases faster than the I/O bandwidth, this asumption may not hold in the future and the bandwidth may tend to be saturated more often and thus decreasing the efficiency of burst buffers. At Rikken, the team in charge of the Post-K computer (a supercomputer expected to replace the K-computer in 2022) is working on a 3-level hierarchical storage system [27].

As an example of the use of burst buffers, Kougkas et al [29] present a dynamic I/O scheduling at the application level using burst buffers to stage I/O and to allow computations to continue uninterrupted. They design different strategies to mitigate I/O interference, including partitioning the PFS, which reduces the effective bandwidth non-linearly. Note that for now, these strategies are designed for only two applications, furthermore they are not coupled with an efficient I/O bandwidth scheduling strategy and can only work because they considered an underutilized I/O bandwidth.

1.3.2 Taxonomy of HPC applications

Many recent HPC studies have observed independently patterns in the I/O behavior of HPC applications. The periodicity of HPC applications has been well observed and documented [8, 17, 21]: HPC applications alternate between computation and I/O transfer, this pattern being repeated over-time. Furthermore, fault-tolerance technique (such as periodic checkpointing [12]) also add to this periodic behavior. Carns et al. [8] observed with Darshan the periodicity of four different applications (MADBench2 [9], Chombo I/O benchmark [11], S3D IO [35] and HOMME [34]). Furthermore, in our previous work [21] we were able to verify the periodicity of gyrokinetic toroidal code (GTC) [20], Enzo [7], HACC application [22] and CM1 [6].

Recently, Hu et al. [25] sum up the four key characteristics of HPC applications observed in the literature:

1. *Periodicity*: Applications alternate between compute phases and I/O phases. Furthermore they do so in a periodic fashion: a regular pattern of compute-I/O is repeated over time.

2. *Burstiness*: In addition to the periodicity observed, sometimes, short I/O burst occur.
3. *Synchronization*: I/O accesses of an application are performed in a synchronized way between the different parallel processes.
4. *Repeatability*: The same jobs are often run many times with only different input, hence the compute-I/O pattern of an application can be predicted before it is executed.

The key idea in this project is to take into account those known structural behaviors of HPC applications and to include them in scheduling strategies.

1.3.3 Scientific and technical objectives

The overall scientific objectives are three-fold:

- To understand precisely the I/O behavior of HPC applications
- To design algorithms and policies for an adaptative schedule of I/O management.
- To provide a low-level I/O library for HPC systems that can be included in any I/O runtime

2 Project organization and means

We propose here a completely novel paradigm to deal with the I/O management problem. The originality of this project is to use known HPC application behaviors. Observations show that most HPC applications *periodically* alternate between (i) operations (computations, local data-accesses) executed on the *compute* nodes, and (ii) I/O transfers of data and this behavior can be predicted before-hand [3, 8].

Taking this structural argument, along with HPC-specific applications facts (there are in general very few applications running concurrently on a machine, and the applications run for many iterations with similar behavior) the goal is to design new algorithms for I/O scheduling. The novelty of this class of algorithms for I/O management is that we intend them to be computed statically. To deal with the complexity of designing static algorithms, they will be designed with a periodic behavior, the scheduling of the volume of I/O of the different applications is repeated over time. This is critical since often the number of instances of all applications is very high and the overall complexity of a non-periodic static schedule would make those algorithms non usable. We envision the implementation of this periodic scheduler to take place at two levels:

1. The job scheduler would know the applications profile. Using these profiles it would be in charge of computing a periodic schedule every time an application enters or leaves the system.

2. Application-side I/O management strategies then would be responsible to ensure the correct transfer of I/O at the right time by limiting the bandwidth used by nodes that transfer I/O.

To be able to manage the I/O at scale, extra care will be given to include solutions to the following challenges:

- Robustness and dynamicity to cope with uncertainties: while most applications are very structured, there might be some variability on the computational cost and I/O volumes transferred. The solutions designed will need to be robust to these variabilities. One solution might be to couple dynamicity to the offline solutions proposed.
- Integration of Emerging Technologies: with our industrial partner, DDN, we plan to model and integrate into the solutions developed Emerging Technologies such as Burst-Buffers.

Finally, as a last step, we will explore problematics related to these two challenges:

- Reliability: today's data management is considered to be highly reliable [1]. However with more stress put on the I/O system, failures and data corruption are expected to become the norm. At term the solution proposed will need to take fault-tolerance into account.
- Energy efficiency: finally, expectations are that energy cost of data-movement is going to be one of the key energy consumer in future systems [28]. Energy-efficiency will have to be incorporated in the scheduler we design.

Exploratory work by the project leader The main risk with this project is that such a scheduler proves inefficient. To mitigate this risk we not only plan to provide coupled dynamic/static strategies, but between the pre-proposition and the proposition, we have done an exploratory study: Guillaume Aupy, Ana Gainaru, and Valentin Le Fèvre, *Periodic I/O scheduling for supercomputers* [2]. This study was published as an Inria Research Report in February 2017. We plan to submit this work to the conference Cluster 2017. The key information from this is that with a simple periodic algorithm, we outperform all online strategies, both for performance and fairness simultaneously.

2.1 Project structure

This project is organized as follows:

Task 1 is used to complement existing information on HPC applications. While some information on applications exists, they have not been studied with an algorithmic perspective in mind. Outcome consists in reports, and a framework/software to automatize and repeat this study in order to keep track of future applications.

Tasks 2 and 3 are scientific tasks at the core of the project, the first one consists in developping scheduling algorithms for I/O management based on the input observed in Task 1. Task 3 consists in extending those algorithms to take into account emerging technologies such as Burst Buffers.

Task 4 has two goals: software integration (based on the prototypes developed in Tasks 2 and 3) and experimental evaluation using the different experimental platform to which we have access.

Task 5's goal is to take into account the future constraints such as energy and reliability on I/O management and to include them in our prototype. This task's main goal is to show a way towards more research in that direction.

2.2 Description of the tasks

2.2.1 Task 1: Full scale modeling of HPC applications and environments

leader: Guillaume Aupy				
Involved partners	Guillaume Aupy	Emmanuel Jeannot	Research Intern	PhD
Involvement	4	1	3	5
Participants: Jean-Thomas Acquaviva (DDN), Venkatram Vishwanath (ANL)				
Tools and environment: DIO-pro, Mira, Theta, Currie				

Goals: To obtain accurate data on I/O behavior in HPC systems.

Detailed work program: Currently many groups have started to caracterize HPC applications running on supercomputers. However these studies are done from a system perspective (e.g. what is the size of each data element moving on the I/O network), and not with a scheduling perspective (e.g. can we use tools from queueing theory to model the apparition of volumes of I/O).

To perform this study we will work with different HPC systems located at Argonne, CEA and Madrid. Jean-Thomas Acquaviva is developing a tool to study the behavior of I/O. His expertise will be key for this study. In this regard, we will study different classes of HPC applications such as Climate simulations (ICON), Deep-Learning (Caffe), Neuroscience (Nest).

Finally, we plan to use the output of this task as input to develop the I/O scheduling algorithms in the next tasks.

Deliverables:

T0+7 [D1.1]: An accurate model for I/O behaviors in HPC systems (report).
T0+21 [D1.2]: Pipeline prototype to study these behaviors in other systems (software).
T0+21 [D1.3]: An update on the model developed in D1.1.

Risks and backup solutions: The main risk of this task is not to find any other structural arguments that can be exploited from an algorithmic perspective. In that case the algorithms developed will solely be based on current structural arguments presented in Section 1.3.2. Furthermore, the pipeline derived to study I/O behavior will still be useful to keep track over time of the evolution of I/O behaviors.

2.2.2 Task 2: Design of a proof of concept, robustness etc

leader: Guillaume Aupy				
Involved partners	Guillaume Aupy	Emmanuel Jeannot	Research Intern	PhD
Involvement	9	3	3	18
Participants: Venkatram Vishwanath (ANL), Olivier Beaumont (Inria)				
Tools and environment: Mira, Theta, Currie				

Subtask 2.1: Algorithms for periodic applications

Goals: To develop static efficient algorithmic solutions based on structural arguments of HPC applications.

Detailed work program:

Based on the structural arguments observed in the literature and on our previous results (Task 1), we will develop and analyze algorithm that take those into account. As an example of natural extensions of our preliminary work [2]:

- The shape of a period: my preliminary work considers only one I/O transfer per instance, but we will have to consider I/O patterns within an instance;
- The multiplicity of entry-point into the job scheduler (multiple I/O nodes).

As in our previous work, all these algorithms will be evaluated as a first step using different parallel IO benchmarks such as IOR benchmarks [36] or HACC IO on real machines at Mellanox and Argonne, and via thorough simulations. The code for the algorithms will be made available under a free software licence.

Deliverables:

T0+13 [D2.1.1]: A description of the algorithms and thorough evaluation via simulation and benchmarking (report)
T0+13 [D2.1.2]: An implementation of the algorithm publicly available (software)

Risks and backup solutions: The main risk of this proposal is obviously that a static strategy based on periodicity does not perform as well as a dynamic strategy. To mitigate this risk, we have performed a preliminary study [2] which shows that a naive version of this strategy is already very efficient with respect to existing online strategies. If we realize that still these solutions do not perform better, then more time will be dedicated on Subtask 2.3 that is on the problematic of coupling dynamicity to the static solutions (using clairvoyance, see later).

Subtask 2.2: Robustness to variability

Goals: To study the impact of system interference on static solutions.

Detailed work program: Before running on actual applications, a key question will be to see how robust those solutions are to unpredictable system interference (such as a slow compute node), or variability in the volume of I/O transfered. In this regard we plan:

1. As a first step, we will propose extensions to static solutions to deal with variability and evaluate them experimentally.
2. Then we will evaluate from a theoretical perspective the robustness that we can expect from these solutions and their limitations.
3. Finally we may have to develop a whole range of different type of solutions based on robust scheduling by budgeting uncertainties.

Deliverables:

T0+21 [D2.2.1]: A description of the algorithms and theoretical results. A thorough evaluation and comparison to previous results via simulation and benchmarking (report)
T0+21 [D2.2.2]: An implementation of the algorithms (software)

Risks and backup solutions: The main risk is for the solutions developed in the previous section not to be robust enough to variability. In this case, we plan to develop a new set of robust algorithms. The previous algorithms will then be useful to compare the loss due to the robustness of tbe new algorithms.

Subtask 2.3: Coupling Dynamicity and Staticity to cope with non periodic applications

Goals: Including not periodic applications in the solutions, as well as improving the solutions based on unpredictable events.

Detailed work program: Until now we have considered that we knew everything up to a variability factor on the applications running in the system.

This subtask will include in the scheduler those tasks that have not a predictable I/O pattern by coupling dynamicity and staticity. We will try at least two different strategies to do so:

- The first one will simply add an online scheduler on top of the static solutions, with the possibility to update a static schedule if the divergence to the expected solution is too high;
- The second one which is also one of the backup solution to the case where the static scheduler is not more efficient than the online scheduler, is to use known behavior of applications to inform online scheduler in a semi-clairvoyant fashion.

Deliverables:

T0+30 [D2.3.1]: A description of the algorithms and theoretical results. A thorough evaluation and comparison to previous results via simulation and benchmarking (report)
T0+30 [D2.3.2]: An implementation of the algorithms (software)

Risks and backup solutions: The main risk would be that including non periodic applications breaks entirely the static schedule. This is also the reason for the second strategy and should protect from this risk.

2.2.3 Task 3: Emerging technologies modeling and integration

leader: Guillaume Aupy			
Involved partners	Guillaume Aupy	Emmanuel Jeannot	PhD
Involvement	3	1	6
Participants: Jean-Thomas Acquaviva (DDN)			
Tools and environment: DDN's platform and Burst Buffers			

Goals: To integrate the modelisation of new technologies designed to assist I/O management.

Detailed work program: Recently, new architectural technologies are emerging (such as *Burst Buffers*) to help deal with the I/O data. Burst buffers behave as I/O buffers, where the data that should be sent to the disks can be stored if the bandwidth is saturated. Then when some bandwidth is available, the buffers are drained (and the data actually sent to the disks).

The first step of this proposal will be a precise modelisation of burst buffers. An idea to start this modelisation is to look at I/O transfers statistically, and use tools from queuing theory. To do this, we will have the assistance of Jean-Thomas Acquivava working at DDN, a french company specialized in the design of storage solutions. Their current research is focusing on burst buffers and they will make their architecture available to us during this collaboration.

Furthermore the use of these buffers is still subject to many questions, such as, should they be put close to the compute nodes and not shared between the different applications. This would mean more buffers at a higher cost and a more challenging architectural problem. Another solution would be to put them close to the I/O nodes and be shared within the different applications. We will then show how to integrate them in the algorithms that we designed in the previous tasks, where they could be used either in their primary buffer function or to assist robustness to face unpredicted bursty I/O. Independently, the solutions and that we develop for bust buffers may be extended to current I/O strategies, and included in DDN's software solutions for experimental evaluation.

Deliverables:

T0+37 [D3.1]: A modeling and experimental evaluation of emerging technologies. Integration of these constraints in the algorithms. (report)
T0+37 [D3.2]: An implementation of the algorithms (software)

Risks and backup solutions: With any emerging technologies, there is a risk of misunderstanding how they work. The collaboration with DDN, should be a great asset to protect from these risks. To reduce even more these risks, we have budgeted two one week visit of Thomas Acquaviva in Bordeaux.

2.2.4 Task 4: Integration and large-scale experimental validation

leader: Guillaume Aupy				
Involved partners	Guillaume Aupy	Emmanuel Jeannot	PhD	Engineer
Involvement	4	1	7	8
Participants: Florin Isaila (UC3M), Venkatram Vishwanath (ANL)				
Tools and environment: Clarisse, Mira, Theta, Currie				

Goals: To test the results from Tasks 2 to 3 on real machines. To provide multiple libraries that can be used by others.

Detailed work program:

This task will be to implement the strategies developed and integrate them into a software. We want to evaluate those solutions on real-life applications. The implementation will take place in different steps:

- Early evaluations will be done during Task 2.1 and 2.2 on synthetic applications. These evaluations will help us to confirm the models and results.
- An implementation in the I/O stack Clarisse [26] developed in UC3M.
- An evaluation of the algorithms developed for Tasks 2.1 to 2.3 in the HPC systems at UC3M on real applications.
- An implementation and evaluation of the algorithms developed for Tasks 2 and 3 on the supercomputers at Argonne.

In order to do this we plan multiple visits, including:

- an early one week visit for Florin Isaila to Bordeaux to present Clarisse and discuss with us locally in this project,
- a three-month internship in Madrid for the PhD student to implement the solutions into Clarisse. This visit will also include large scale experiments.
- During this visit Guillaume will also visit UC3M.

Then at a later step of the thesis, we plan for a one or two month visit for the PhD student at Argonne. This visit will be funded by external fundings (a joint call with Venkatram Vishwanath will be made to the Joint Laboratory for Exascale Computing (JLESC), see Section 2.5.2).

Finally, we plan to have an engineer work on the implementation. Initially this will be done with the help of the PhD student to understand the problem. Then the engineer's role then will be to create a well-documented library that can be integrated by other researchers in other systems.

Deliverables:

T0+15 [D4.1]: Evaluation of task 2.1 on synthetic applications on a HPC system, integrated to report D2.1.1. (report)
T0+30 [D4.2]: Thorough evaluation on an HPC system with different sets of application, integrated to report D2.3.1. (report)
T0+42 [D4.3]: An open-source library (software)

Risks and backup solutions: The main risk is due to the lack of experience of the project leader in developing software. To mitigate this risk, we will use a strong collaboration (three visits, a three month internship for the PhD student) with Florin Isaila at UC3M. Furthermore we do

not plan to create the full I/O stack but to integrate our approach in the I/O software stack though the cross-layer approach developed in the Clarisse project (funded by the European Union). Furthermore, there is a risk that we cannot experiments on actual applications. In this case we will use mini-apps specifically developed for I/O studies such as the one here: `https://www.vi4io.org`.

2.2.5 Task 5: Long-term objectives, coping with energy-efficiency and reliability

leader: Guillaume Aupy			
Involved partners	Guillaume Aupy	Intern	Engineer
Involvement	4	6	4
Participants: Shadi Ibrahim (Inria)			
Tools and environment: Grid 5K			

Goals: To evaluate the impact of I/O management energy-wise, to integrate reliability into I/O management.

Detailed work program: This task's goal is more exploratory than the others.

Until now few studies have addressed the energetic impact of I/O management. On the contrary, it has been observed by Kogge and Schalf [28] that in future systems the energy cost of data movement is expected to be the main leading costs. To this end, we plan to study with Shadi Ibrahim how different I/O management techniques inccur different energy costs. Shadi Ibrahim has already started a study [18] and studied the cost of using dedicated cores and nodes to manage I/O (that is Application-side I/O scheduling). Together, one the direction would be to model all machine activities with regard to their power cost [4]. Similarly, we could try to evaluate all algorithmic-specific costs, such as the overhead of an online strategy, the overhead of interrupting a computation etc. We expect the engineer to expend the evaluation workflow developed by Shadi Ibrahim to evaluate the energy impact of I/O management.

The second problem will be to study the impact of different fault-tolerance techniques on I/O. Fault-tolerance nowaday has focused on computation. Many models even specifically consider no failures during the I/O management [12]. One of the reason is that one could consider additional techniques to make this transfer fault-tolerant (for instance HDFS). With regards to these problematics, we propose to evaluate the impact of HDFS and other similar fault-tolerance techniques, and offer other algorithmic based solutions and evaluate them with regard to the load that is added to the I/O system and energy costs incurred. Other direction include an I/O based optimization of reliability techniques. For example, in general optimization in periodic checkpointing assumes the optimal period with respect to time [12]. However, it has been observed that a small variations in the optimal period has little impact on the time wasted [3]. In this regard, techniques

Deliverables:

T0+47 [D5.1]: An experimental evaluation of both the performance and energy costs of different I/O management strategies (including reliable management). (report)

Risks and backup solutions: One of the risk and difficulty is to measure the power performance of the different items of a machine. This is why we are using Grid 5K instead of a usual supercomputer. On reliability, because Grid 5K is much smaller than a supercomputer, it is very unlikely that we experience faults in our experiments. This is why we are focusing on the fault-tolerant techniques and algorithms, we will extend it mathematically to cases where fault occur.

2.3 Tasks schedule, deliverable and milestones

We sum up the Task schedule graphically in Table 2.3 and a synthetic view of the deliverables is presented in Table 2.3.

Table 1: Tentative task schedule for the program's objective

		Timing diagram/critical path															
		Year 1				Year 2				Year 3				Year 4			
		3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48
Task 1				✍				♠									
Task 2	SubTask 2.1					✍											
	SubTask 2.2							✍									
	SubTask 2.3										✍						
Task 3													✍				
Task 4																♠	
Task 5																	✍
Meetings										❖						❖	
Progress report/Expenses							✪						✪				☆

Legend ✍ ♠ Deliverable is a report / Deliverable is a software system; ✪ ☆ Progress report + expenses / Final report + expenses summary; ❖ Workshop

Table 2: Deliverables and Milestones

Task	Title and substance of the deliverables and milestones	Delivery	Participants
T1: Full scale study of HPC applications			
	[D1.1] An accurate model for I/O behaviors in HPC systems	T0+7	Intern
	[D1.2] pipeline prototype to study these behaviors in other systems	T0+21	PhD
	[D1.3] An update on the model developed in D1.1	T0+21	PhD
T2: Design of a proof of concept, robustness			
2.1 Algorithms for periodic applications			
	[D2.1] Full description + Implementation	T0+13	PhD+Research Intern
2.2 Robustness to variability			
	[D2.2] Full description, Implementation and thorough evaluation	T0+21	PhD
2.3 Coupling Dynamicity and Staticity, inclusion of non-periodic applications			
	[D2.3] Full description + Implementation	T0+30	PhD
T3: Emerging technologies modeling and integration			
	[D3.1] An accurate model and experimental evaluation	T0+37	PhD
	[D3.2] Integration and implementation in current stack	T0+37	PhD
T4: Integration and large-scale experimental validation			
	[D4.1] Evaluation of T2.1 with synthetic applications	T0+15	PhD
	[D4.2] Thorough evaluation on an HPC system with application	T0+30	PhD
	[D4.3] Open-source library	T0+42	PhD+Engineer
T5: Long-term objectives, coping with energy-efficiency and reliability			
	[D5.1] Evaluation of energy cost and impact of reliability	T0+47	Intern+Engineer

2.4 Resource Management

2.4.1 Costs supported by ANR proposal

Staff Two research interns (master-level, 6 months) will be hired for the project (on the 1st and 44th months) as well as a PhD student (36 months) and an engineer (12 months).

The first research intern will work on studying application behavior in HPC systems and modeling them. The PhD student will then work on the design and evaluation of the algorithmic bricks of the

project. The engineer will be hired on the 36th month of the project and will work on (i) assisting with the experiments on real applications and machines; (ii) making the software available and usable to the HPC community. They will also work on exploratory work on the impact of reliability strategies, and on energy performance of the developed strategies. For this last task they will be assisted by a research intern (master-level).

Mission, travelling and workshop The traveling costs are evaluated as follows:

1. Bi-annual trip to the JLESC workshops where we will meet most of our collaborators for two members of the project during the 4 years. *Cost:* $2 \times 8 \times 1.3k€ = 20.8k€$.
2. Attending two international (non-Europe) conferences for two members of the project during 4 years. *Cost:* $2 \times 8 \times 3k€ = 48k€$.
3. Attending one european conferences for one member of the project during 4 years. *Cost:* $4 \times 1.6k€ = 6.4k€$.
4. A technical one-week visit for the industrial project partner (Task 3). *Cost:* $1 \times 1.2k€ = 1.2k€$.
5. Two technical one-week visit to France for the Spain project partner (Task 4). *Cost:* $2 \times 1.2k€ = 2.4k€$.
6. A three-month visit for the PhD student to the Spain partner (Task 4). *Cost:* $1 \times 6.8k€ = 6.8k€$.
7. One technical one-week visit to the Spain partner (Task 4). *Cost:* $1 \times 1.2k€ = 1.2k€$.

Finally, we plan to organize a two-day workshop around I/O management in Bordeaux at the end of Year 2/beginning of Year 3, whose cost (food and breaks) will be partly supported by ANR (20 people): 1500€.

Inward billing A laptop will be bought for the PhD student to be recruited by this project: $3k€$.

Estimated cost of the Project				
Staff Cost	Missions and Traveling	Expenses of Inward Billing	Administrative fees	Total Cost
$171k€$ (1 PhD + 1 Post-Doc + 2 Research Interns)	$87.9k€$ (conferences, visits, workshop)	$3k€$ (1 laptop)	$20.9k€$	$282.8k€$

2.4.2 Costs not supported by ANR proposal

Experimental resources For experimental platforms we will rely mainly on our partners but also on project that we plan to submit in addition to this one:

- We will work with researchers at UC3M and Argonne to have access to different supercomputers.
- DDN will provide us enough machine time to experiments with their burst-buffers.
- We expect to have access to Currie the supercomputer at the CEA.
- Finally, we plan to ask compute-hours on GENCI (Grand Equipement de Calcul Intensif), which includes the supercomputers from CEA and CNRS.

Additional travels and workshop We will ask for dditional funding to the JLESC (see Section 2.5.2) for a one or two month internship of the PhD student at Argonne National Laboratory to assist in implementation of D.4.2.

We will also seek additional funding to organize another workshop at the end of Year 4.

2.5 Consortium around DASH

2.5.1 Scientific leader

Guillaume Aupy (CR Researcher in the TADaaM team, Inria Bordeaux Sud Ouest research center) will lead the project. Guillaume started his research career with a PhD in computer science at ENS Lyon in 2014 mostly focused on reliability and energy efficiency in HPC. After this, he became a Research Assistant Professor in Penn State University then in Vanderbilt University (USA) working on data movement in HPC algorithms. Finally, he was hired as a permanent research scientist at Inria in 2016. His research interests include Scheduling and optimization, data-movement, resilience, energy. He has published in the top conferences (SC, IPDPS, ICPP amongst others) and journals (Journal of Parallel and Distributed Computing, Siam Journal of Scientific Computing amongst others) of the field.

He is currently the Technical Program vice-chair of SC17, the main conference in High-Performance Computing that unites 12,000 researchers, practitioners and industrial partners. This position until today has never been held by any european researcher. As part of this role, he has been instrumental in pushing the SC reproducibility initiative[1], making it mandatory to be considered for Best Paper and Best Student Paper.

Finally, Guillaume is currently involved in the writing of the Strategic Research Agenda (SRA) for the European call ETP4HPC, specifically in the work-packages (i) *Balance Compute, I/O and Storage Performance*, (ii) *Big Data and HPC usage Models* and (iii) *Energy and Resiliency*.

More details can be found at: `http://gaupy.org`

2.5.2 Scientific Environment

JLESC: Joint Laboratory for Extreme Scale Computing Inria is part of the Joint Laboratory for Extreme Scale Computing (JLESC: `https://jlesc.github.io/`). This international organization also includes UIUC, ANL, BSC, JSC and RIKEN-AICS. Its goal is to facilitate collaboration between members of all these organizations by providing both a platform to communicate recent development (through bi-annual meetings) and funding for collaborations as well as access to top-end HPC machines (Theta, Mira, K-computer, etc.).

Tool: DIO-pro DIO-pro is an I/O profiling tool developed by DDN designed to capture I/O events emanating from an arbitrary application. DIO-pro observes I/O events from the point of view of the application, it is unaware of events in the file system or storage units. DIO-pro records I/O events in the POSIX and MPIIO interface. Events are recorded per file and process. For each I/O event, the time stamp, time span and some context variables (event-dependent) are stored. The time stamp have microsecond accuracy. Dio-Pro records traces system wide with a microsecond clock to capture the temporal locality of the I/O events. Such property is key in order to evaluate the impact of hierarchies in the I/O path.

Machines: Mira, Theta, Currie Mira is the current supercomputer at Argonne National Laboratory. Through both the JLESC and our collaboration with Venkatram Vishwanath, we will evaluate our different tasks there. Similarly, Theta is the new development platform for the future supercomputer at Argonne. Currie is the supercomputer at the CEA. TADaaM has a long term collaborationwith the CEA, and we hope to be able to also evaluate our algorithms and models on their machine. However, no discusssion as been started yet.

[1] `http://sc17.supercomputing.org/submitters/technical-papers/reproducibility-initiatives-for-technical-papers`

2.5.3 Scientific and industrial partners

Our three partners have offered a letter of intent to collaborate, attached to this proposal.

Pr. Florin Isaila: Associate professor at University Carlos III, Madrid (UC3M) Florin is a long-term collaborator of the TADaaM team [38] which opened the possibility of discussion. With his team in Madrid, they have started the CLARISSE project: a mechanism for designing and implementing cross-layer optimizations of the I/O software stack, including I/O scheduling. The collaboration with Florin will help ensure the implementation and dissemination of our results on a real system.

Dr. Jean-Thomas Acquaviva: Research Engineer at Data Direct Network (DDN) DDN is a tech company specialized in storage for HPC. I met Jean-Thomas when he came to Bordeaux to present their recent work on Burst-Buffers. He further presented DIO-pro, DDN's software to evaluate I/O behavior of HPC applications. His observations were similar to what I knew about the periodicity of I/O patterns, so we talked about my project and he was strongly enthusiastic (see his supporting letter). We discussed about our common interests in the modeling and integration of Burst-Buffers and decided to work together.

Dr. Venkatram Vishwanath: Research Scientist at Argonne National Laboratory, USA Venkatram, Emmanuel Jeannot and I have started collaborating on a project dedicated to I/O management with Argonne National Laboratory. We have written together a 2 year grant proposal which was accepted. The JLESC also provides us many opportunity to work together. Venkat has strongly accepted to support our project and contribute his expertise in I/O management.

3 Project Impact, Dissemination and exploitation of results

3.1 Positioning with respect to the call

This project aims at answering the *"Défi 7 > Axe 7> Infrastructures de HPC et traitement massif de données*. Indeed, the core of this project is managing data at the I/O level in the scope of High-Performance Computing. It has been well documented that the CPU/IO imbalance of current architecture is a key limitation to the development of Big Data [10].

We believe that this project will make a strong impact in many fields. With the help of DDN we will have access to key societal applications that are run on HPC machines such as Climate simulations (ICON), Deep-Learning (Caffe), Neuroscience (Nest).

The result that we expect to obtain will impact every HPC application: from medical research (Brain initiatives), to astrophysics (HACC, Enzo, HOMME), including meteorology (CM1) and fusion plasma (GTC). I/O congestion has been called one of the top10 challenge for extreme-visual analytics [40].

3.2 Valorization strategies, dissemination and exploitation of results

First, a website including all documentation and information (progress, reports, deliverable, papers, slides, software) in relation to the project will be available to the public and kept up-to-date.

Publications, presentations All throughout this project, we expect to publish the scientific results obtained. These results will be dissiminated in major scientific conferences (SC, IPDPS, EuroPar, ICPP, etc.) and journals (TPDS, JPDC, J. of Scheduling, etc.). Results such as software prototypes of this research will be presented in poster sessions and demos co-located with major scientific

conferences. Furthermore as part of the JLESC, we will present bi-annual updates of those results. This should allow us to trigger more collaborations and fruitful discussions.

Software Instead of trying to recreate a full software stack from scratch, we made the realistic choice to integrate the resulting software of this proposal into an existing middleware: Clarisse [26] from UC3M, Madrid. We hope that this will give us more time to define efficient algorithms, while giving to our results and software more visibility and perennity. These software will be made available publicly under a free licence which will allow its integration into commercial services.

Industrial partners We will be closely collaborating with an industrial partner: DDN. The resulting algorithms and strategies are definitely of interest to them. This is a new collaboration for the TADaaM team and we hope that one outcome of this ANR would be new collaborations and contracts with DDN for the TADaaM team. Other than DDN, Guillaume Aupy has an active collaboration with the HPC architecture team at Mellanox [2], a company leading in extreme-scale architecture and particularly in I/O and networks. We expect that this collaboration will allow to keep a vision of the coming challenges with I/O architectures. It will also allow us to have a development platform to experiment the algorithms developed. Finally, we plan to work with ATOS/Bull and the CEA who will be interested in those new techniques and with whom TADaaM is already collaborating (ITEA2 Coloc Project, PIA ELCI project).

Gathering the French I/O community and H2020 discussion With this project, one of the core idea of this project is also to help consolidate a French community around I/O management. Currently only a few French research teams are working on I/O management. In this direction, we plan to organize a workshop at the end of Year 2 or in the beginning of Year 3 to present all tools and results obtained by the different teams working on I/O. This workshop would include amongst other, researcher from the newly formed DataMove and Polaris team in Grenoble (Bruno Raffin, Olivier Richard, Denis Trystram, Arnaud Legrand, ...), researchers from the KerData and Myriads team in Rennes (Gabriel Antoniu, Shadi Ibrahim, Anne-Cécile Orgerie, ...), but also industrial partners (DDN, ATOS/Bull, ...). We will also be looking for more fundings to organize another workshop at the end of Year 4.

Ultimately, the idea would be to co-write a multi-institution european project around I/O management. Specifically, a subject of interest would be around the convergence between Big Data and I/O management: what would be needed for HPC systems to use tools developed for Big Data (Spark, Hadoop etc).

Outreach Guillaume Aupy has been very involved in the dissemination of scientific results to the public. He has worked with two different organization whose goal is to introduce math (Plaisir-Maths) and math research (Maths-en-Jeans) through fun activities. He plans to continue working with Maths-en-Jeans, an organization whose goal is to get interest from middle-school and high-school student to mathematical research. In the future he plans to use material from this proposal in this context.

Furthermore, his recent presentation of problematics related to this proposal in front of the full Inria body (Unithé-ou-Café, `http://gaupy.org/ressources/files/unithe.pdf`) was very appreciated. He plans to extend it as a blog article eiher on `https://interstices.info/` or on `http://binaire.blog.lemonde.fr/` and find ways to develop more content.

3.3 Reproducibility Initiative

The ability to reproduce experiments is key towards trust in scientific results. Towards this goal, this project will be using strong openess fundations from the start. We will provide:

1. A public code repository where all code funded by this ANR will be available under a free licence;
2. To work as much as possible towards reproducibility of computational results, we will publish our data, the code for its analysis, description of artifacts and computational results following recommendation from the ACM.

Note that we have already done most of this for our exploratory work [2]: `https://github.com/vlefevre/IO-scheduling-simu`. All this information will be aggregated on the website of the project.

As a side note, a recent ADT at Inria Bordeaux Sud Ouest started to create reproducible software environments. We will discuss frequently with them to work on best practices.

References

[1] Advanced Scientific Computing Advisory Committee (ASCAC). *Ten technical approaches to address the challenges of Exascale computing*. `http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf`.

[2] Guillaume Aupy, Ana Gainaru, and Valentin Le Fèvre. *Periodic I/O scheduling for supercomputers*. Research Report 9037. Inria Bordeaux Sud-Ouest, Feb. 2017.

[3] Guillaume Aupy, Yves Robert, Frédéric Vivien, and Dounia Zaidouni. "Checkpointing algorithms and fault prediction." In: *Journal of Parallel and Distributed Computing* 74.2 (2014), pp. 2048–2064.

[4] Guillaume Aupy, Anne Benoit, Thomas Hérault, Yves Robert, and Jack Dongarra. "Optimal Checkpointing Period: Time vs. Energy." In: *Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*. LNCS. Springer-Verlag, 2013.

[5] Babak Behzad, L Huong Vu Thanh, Joseph Huchette, Surendra Byna, R Aydt Prabhat, Quincey Koziol, and Marc Snir. "Taming parallel I/O complexity with auto-tuning." In: *Proceedings of SC13*. 2013.

[6] George H Bryan and J Michael Fritsch. "A benchmark simulation for moist nonhydrostatic numerical models." In: *Monthly Weather Review* 130.12 (2002).

[7] Greg L Bryan et al. "Enzo: An Adaptive Mesh Refinement Code for Astrophysics." In: *arXiv:1307.2265* (2013).

[8] Philip Carns, Robert Latham, Robert Ross, Kamil Iskra, Samuel Lang, and Katherine Riley. "24/7 characterization of petascale I/O workloads." In: *Proceedings of CLUSTER09*. IEEE. 2009, pp. 1–10.

[9] Jonathan Carter, Julian Borrill, and Leonid Oliker. "Performance characteristics of a cosmology package on leading HPC architectures." In: *HiPC*. Springer, 2005, pp. 176–188.

[10] CL Philip Chen and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." In: *Information Sciences* 275 (2014), pp. 314–347.

[11] P Colella et al. *Chombo infrastructure for adaptive mesh refinement*. `https://seesar.lbl.gov/ANAG/chombo/`. 2005.

[12] John T Daly. "A higher order estimate of the optimum checkpoint interval for restart dumps." In: *Future generation computer systems* 22.3 (2006), pp. 303–312.

[13] Sheng Di and Franck Cappello. "Fast error-bounded lossy HPC data compression with SZ." In: *Parallel and Distributed Processing Symposium, 2016 IEEE International*. IEEE. 2016, pp. 730–739.

[14] Jack Dongarra and Michael A Heroux. "Toward a new metric for ranking high performance computing systems." In: *Sandia Report, SAND2013-4744* 312 (2013).

[15] Matthieu Dorier, Gabriel Antoniu, Robert Ross, Dries Kimpe, and Shadi Ibrahim. "CALCioM: Mitigating I/O Interference in HPC Systems through Cross-Application Coordination." In: *Proceedings of IPDPS14*. 2014.

[16] Matthieu Dorier, Gabriel Antoniu, Franck Cappello, Marc Snir, and Leigh Orf. "Damaris: How to efficiently leverage multicore parallelism to achieve scalable, jitter-free I/O." In: *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*. IEEE. 2012, pp. 155–163.

[17] Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu, and Rob Ross. "Omnisc'IO: a grammar-based approach to spatial and temporal I/O patterns prediction." In: *SC*. IEEE Press. 2014, pp. 623–634.

[18] Matthieu Dorier, Orcun Yildiz, Shadi Ibrahim, Anne-Cécile Orgerie, and Gabriel Antoniu. "On the energy footprint of I/O management in Exascale HPC systems." In: *Future Generation Comp. Syst.* 62 (2016), pp. 17–28.

[19] Matthieu Dreher and Bruno Raffin. "A flexible framework for asynchronous in situ and in transit analytics for scientific simulations." In: *Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on*. IEEE. 2014, pp. 277–286.

[20] Stephane Ethier, Mark Adams, Jonathan Carter, and Leonid Oliker. "Petascale parallelization of the gyrokinetic toroidal code." In: *VECPAR: High Performance Computing for Computational Science* (2012).

[21] Ana Gainaru, Guillaume Aupy, Anne Benoit, Franck Cappello, Yves Robert, and Marc Snir. "Scheduling the I/O of HPC applications under congestion." In: *Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International*. IEEE. 2015, pp. 1013–1022.

[22] Salman Habib et al. "The universe at extreme scale: multi-petaflop sky simulation on the BG/Q." In: *Proceedings of SC12*. IEEE Computer Society. 2012, p. 4.

[23] Bill Harrod. *Big data and scientific discovery*. 2014.

[24] Y. Hashimoto and K. Aida. "Evaluation of Performance Degradation in HPC Applications with VM Consolidation." In: *IEEE International Conference on Networking and Computing (ICNC)* (2012), pp. 273–277.

[25] Wei Hu, Guang-ming Liu, Qiong Li, Yan-huang Jiang, and Gui-lin Cai. "Storage wall for exascale supercomputing." In: *Journal of Zhejiang University-SCIENCE* 2016 (2016), pp. 10–25.

[26] Florin Isaila, Jesus Carretero, and Rob Ross. "CLARISSE: A Middleware for Data-Staging Coordination and Control on Large-Scale HPC Platforms." In: *Cluster, Cloud and Grid Computing (CCGrid), 2016 16th IEEE/ACM International Symposium on*. IEEE. 2016, pp. 346–355.

[27] Yutuka Ishikawa. *Towards the Japanese next flagship supercomputer*. Presentation at 39th ORAP Forum. 2017.

[28] Peter Kogge and John Shalf. "Exascale Computing Trends: Adjusting to the "New Normal" in Computer Architecture." In: IEEE, 2013.

[29] Anthony Kougkas, Matthieu Dorier, Rob Latham, Rob Ross, and Xian-He Sun. "Leveraging Burst Buffer Coordination to Prevent I/O Interference." In: *IEEE International Conference on eScience*. IEEE. 2016.

[30] Sidharth Kumar et al. "Characterization and modeling of PIDX parallel I/O for performance optimization." In: *Proceedings of SC13*. ACM. 2013.

[31] Albert Lazzarini. *Advanced LIGO Data & Computing*. 2003.

[32] N. Liu et al. "On the Role of Burst Buffers in Leadership-Class Storage Systems." In: *MSST/SNAPI*. 2012.

[33] Jay Lofstead, Fang Zheng, Qing Liu, Scott Klasky, Ron Oldfield, Todd Kordenbrock, Karsten Schwan, and Matthew Wolf. "Managing variability in the IO performance of petascale storage systems." In: *Proceedings of SC10*. IEEE Computer Society. 2010.

[34] RD Nair and HM Tufo. "Petascale atmospheric general circulation models." In: *Journal of Physics: Conference Series*. Vol. 78. IOP Publishing. 2007, p. 012078.

[35] Sankaran et al. "Direct numerical simulations of turbulent lean premixed combustion." In: *Journal of Physics: conference series*. Vol. 46. IOP Publishing. 2006, p. 38.

[36] H. Shan and J. Shalf. "Using IOR to Analyze the I/O Performance for HPC Platforms." In: *Cray User Group* (2007).

[37] D. Skinner and W. Kramer. "Understanding the Causes of Performance Variability in HPC Workloads." In: *IEEE Workload Characterization Symposium* (2005), pp. 137–149.

[38] François Tessier, Preeti Malakar, Venkatram Vishwanath, Emmanuel Jeannot, and Florin Isaila. "Topology-aware data aggregation for intensive I/O on large-scale supercomputers." In: *Proceedings of the First Workshop on Optimization of Communication in HPC*. IEEE Press. 2016, pp. 73–81.

[39] A. Uselton, M. Howison, N. Wright, D. Skinner, N. Keen, J. Shalf, K. Karavanic, and L. Oliker. "Parallel I/O Performance: From Events to Ensembles." In: *Proceedings of IPDPS10* (2010), pp. 1–11.

[40] Pak Chung Wong, Han-Wei Shen, Christopher R Johnson, Chaomei Chen, and Robert B Ross. "The top 10 challenges in extreme-scale visual analytics." In: *IEEE computer graphics and applications* 32.4 (2012), p. 63.

[41] Bing Xie, J. Chase, D. Dillow, O. Drokin, S. Klasky, S. Oral, and N. Podhorszki. "Characterizing output bottlenecks in a supercomputer." In: *Proceedings of SC12* (2012), pp. 1–11.

[42] Xuechen Zhang, Kei Davis, and Song Jiang. "Opportunistic data-driven execution of parallel programs for efficient I/O services." In: *Proceedings of IPDPS12*. IEEE. 2012, pp. 330–341.

[43] Z. Zhou, X. Yang, D. Zhao, P. Rich, W. Tang, J. Wang, and Z. Lan. "I/O-Aware Batch Scheduling for Petascale Computing Systems." In: *2015 IEEE International Conference on Cluster Computing*. 2015, pp. 254–263.