A

Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations

MATTHIEU DORIER[1], Argonne National Laboratory, IL, USA
GABRIEL ANTONIU, Inria, Rennes - Bretagne Atlantique Research Centre, France
FRANCK CAPPELLO, Argonne National Laboratory, IL, USA
MARC SNIR, Argonne National Laboratory, IL, USA
ROBERT SISNEROS, University of Illinois at Urbana Champaign, IL, USA
ORCUN YILDIZ, Inria, Rennes - Bretagne Atlantique Research Centre, France
SHADI IBRAHIM, Inria, Rennes - Bretagne Atlantique Research Centre, France
TOM PETERKA, Argonne National Laboratory, IL, USA
LEIGH ORF, University of Wisconsin - Madison, WI, USA

With exascale computing on the horizon, reducing performance variability in data management tasks (storage, visualization, analysis, etc.) is becoming a key challenge in sustaining high performance. This variability significantly impacts the overall application performance at scale and its predictability over time.

In this paper, we present Damaris, a system that leverages *dedicated cores* in multicore nodes to offload data management tasks, including I/O, data compression, scheduling of data movements, in situ analysis and visualization. We evaluate Damaris with the CM1 atmospheric simulation and the Nek5000 computational fluid dynamic simulation on four platforms, including NICS's Kraken and NCSA's Blue Waters. Our results show in particular that (1) Damaris fully hides the I/O variability as well as all I/O-related costs, which makes simulation performance predictable; (2) it increases the sustained write throughput by a factor of up to 15 compared with standard I/O approaches; (3) it allows almost perfect scalability of the simulation up to over 9,000 cores, as opposed to state-of-the-art approaches that fail to scale; (4) it enables a seamless connection to the VisIt visualization software to perform in situ analysis and visualization in a way that does not impact the performance of the simulation, nor its variability.

In addition, we further extended our implementation of Damaris to also support the use of *dedicated nodes* and conducted a thorough comparison of the two approaches –dedicated cores and dedicated nodes– for I/O tasks with the aforementioned applications.

Categories and Subject Descriptors: D.1.3 [**Concurrent Programming**]: Parallel Programing; E.5 [**Files**]: Optimization; I.6 [**Simulation and Modeling**]: Simulation Output Analysis

General Terms: Design, Experimentation, Performance

Additional Key Words and Phrases: Exascale Computing, I/O, In Situ Visualization, Dedicated Cores, Dedicated Nodes, Damaris

ACM Reference Format:
Matthieu Dorier, Gabriel Antoniu, Franck Cappello, Marc Snir, Robert Sisneros, Orçun Yildiz, Shadi

[1]The work was done while the author was at ENS Rennes, IRISA, Rennes, France

Author's addresses: Matthieu Dorier, Franck Cappello, Marc Snir and Tom Peterka, Argonne National Laboratory, 9700 South Cass Avenue B109, Lemont, IL 60439, USA; Gabriel Antoniu, Orçun Yildiz and Shadi Ibrahim, Inria Rennes, campus de Beaulieu, 35042 Rennes, France; Robert Sisneros, NCSA, 1205 West Clark Street Urbana, IL 61801, USA; Leigh Orf, Central Michigan University, 1200 S Franklin St, Mt Pleasant, MI 48859, USA.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© YYYY ACM 1539-9087/YYYY/01-ARTA $15.00
DOI:http://dx.doi.org/10.1145/0000000.0000000

Ibrahim, Tom Peterka and Leigh Orf, 2015. Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations. *ACM Trans. Parallel Comput.* V, N, Article A (January YYYY), 44 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

As supercomputers become larger and more complex, one critical challenge is to efficiently handle the immense amounts of data generated by extreme-scale simulations. The traditional approach to data management consists of writing data to a parallel file system, using a high-level I/O library on top of a standardized interface such as MPI-I/O. This data is then read back for analysis and visualization purpose.

One major issue posed by this traditional approach to data management is that it induces a high performance variability. This variability can be observed at different levels. Within a single application, I/O contention across processes leads to large variations in the time each process takes to complete its I/O operations (I/O jitter). Such differences from a process to another in a massively parallel application makes all processes wait for the slowest one. These processes thus waste valuable computation time. The variability is even larger from one I/O phase to another, due to interference with other applications sharing the same parallel file system.

While scientists have found a potential solution to this problem by coupling their simulations with visualization software in order to bypass data storage and derive results early on, the current practices of coupling simulations with visualization tools also expose simulations to high performance variability, as their run time does not depend anymore on their own scalability only, but also on the scalability of visualization algorithms. This particular problem is further amplified in the context of interactive in situ visualization, where the user himself and his interactions with the simulation become the cause of run-time variability.

To make an efficient use of future exascale machines, it becomes important to provide data management solutions that do not solely focus on pure performance, but address performance variability as well. Addressing this variability is indeed the key to ensure that each and every component of these future platforms is optimally used.

To address these challenges, we have proposed a new system for I/O and data management called Damaris. Damaris leverages dedicated I/O cores on each multicore SMP (Symmetric multiprocessing) node, along with the use of shared memory, to efficiently perform asynchronous data processing I/O and in situ visualization. We picked this approach based on the intuition that the usage of dedicated cores for I/O-related tasks combined with the usage of intranode shared memory can help overlapping I/O with computation, but also lowering the pressure on the storage system by reducing the number of files to be stored and, at the same time, the amount of data. Such dedicated resources can indeed perform data aggregation, filtering or compression, all in an asynchronous manner. Moreover, such dedicated cores can further be leveraged to enable non-intrusive in situ data visualization with optimized resource usage. Damaris' only overhead results from removing computation resources from the simulation. Yet as we move towards post-petascale and exascale machines, the growing number of cores per node makes it computationally affordable to remove one or a few core from the computation for the purpose of managing data. Besides, our experiments show that this overhead is largely counterbalanced by the performance gain of overlapping I/O with computation.

Some of these aspects of the Damaris approach have been introduced in previous conference papers [Dorier et al. 2012a; Dorier et al. 2013]. This paper aims to provide a comprehensive, global presentation and discussion of the Damaris approach in its current state and of its evaluation and applications.

We evaluated Damaris on three different platforms including the Kraken Cray XT5 supercomputer [NICS 2015], with the CM1 atmospheric model [Bryan and Fritsch 2002] and the Nek5000 [P. F. Fischer and Kerkemeier 2008] computational fluid dynamics code. By overlapping I/O with computation and by gathering data into large files while avoiding synchronization between cores, our solution brings several benefits: (1) it fully hides the jitter as well as all I/O-related costs, which makes the simulations performance predictable; (2) it substantially increases the sustained write throughput (by a factor of 15 in CM1, 4.6 in Nek5000) compared with standard approaches; (3) it allows almost perfect scalability of the simulation (up to over 9,000 cores with CM1 on Kraken), as opposed to state-of-the-art approaches (file-per-process and collective I/O), which fail to scale in the CM1 application with as few as a few hundreds of cores (see Section 4; (4) it enables data compression without any additional overhead, leading to a major reduction of storage requirements.

Furthermore, we extended Damaris with Damaris/Viz, an in situ visualization framework based on the Damaris approach. By leveraging dedicated cores, external high-level structure descriptions and a simple API, our framework provides adaptable in situ visualization to existing simulations at a low instrumentation cost. Results obtained with the Nek5000 and CM1 simulations show that our framework can completely hide the performance impact of visualization tasks and the resulting run-time variability. In addition, the proposed API allows efficient memory usage through a shared-memory-based, zero-copy communication model.

Finally, in order to compare the Damaris, dedicated-core-based approach with other approaches such as dedicated nodes, forwarding nodes, and staging areas, we further extended Damaris to support the use of dedicated nodes as well. We leverage again the CM1 and Nek5000 simulations on Grid'5000, the national French grid testbed, to shed light on the conditions under which a dedicated-core-based approach to I/O is more suitable than a dedicated-node-based one, and vice versa.

To the best of our knowledge, Damaris is the first open-source middleware to enable the use of dedicated cores or/and dedicated nodes for data management tasks ranging from storage I/O to complex in situ visualization scenarios.

The rest of this paper is organized as follows: Section 2 presents the background and motivation for our work, discusses the limitations of current approaches to I/O and to in situ visualization. Our Damaris approach, including its design principles, implementation detail and use cases, is described in Section 3. We evaluate Damaris in Section 4, first in scenarios related to storage I/O, then in scenarios related to in situ visualization. Our experimental evaluation continues in Section 5 with a comparison between dedicated cores and dedicated nodes in various situations. Section 6 discusses our positioning with respect to related work and Section 7 summarizes our conclusions and discusses open further directions.

2. BACKGROUND AND MOTIVATION

HPC simulations create large amounts of data that are then read offline by analysis tools. In the following we present the traditional approaches to parallel I/O as well as the problems they pose in terms of performance variability. We then dive into the trend toward coupling simulations with analysis and visualization tools, going from offline to in situ analysis and visualization.

2.1. I/O and Storage for Large-Scale HPC Simulations

Two I/O approaches have been traditionally used for performing I/O in large-scale simulations.

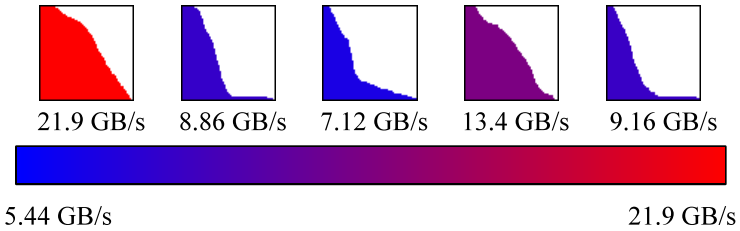


Fig. 1: Variability across processes and across I/O phases in the IOR benchmark using a file-per-process approach on Grid'5000's Rennes site [Grid'5000 2015], with a PVFS2 [Carns et al. 2000] file system. Each graph represents a write phase. The 576 processes are sorted by write time on the y axis and an horizontal line is draw with a length proportional to this write time. These graphs are normalized so that the longest write time spans the entire graph horizontally. Each graph is colored according to a scale that gives the aggregate throughput of the phase, that is, the total amount of data written divided by the write time of the slowest process.[2]

The File-per-process. approach consists of having each process access its own file. This reduces possible interference between the I/O of different processes, but increases the number of metadata operations. This is especially a problem for file systems with a single metadata server, such as Lustre [Donovan et al. 2003]. It is also hard to manage the large number of files thus created and have them read by analysis or visualization codes that use a different number of processes

Collective I/O. leverages communication phases between processes to aggregate access requests and reorganize them. These operations are typically used when several processes need to access different parts of a shared file, and benefit from tight interactions between the file system and the MPI-I/O layer in order to optimize the application's access pattern [Prost et al. 2001].

2.1.1. Variability in Traditional I/O Approaches. The periodic nature of scientific simulations, which alternate between computation and I/O phases, leads to burst of I/O activity. The overlap between computation and I/O is reduced, so that both the compute nodes and the I/O subsystem may be idle for periods of time.

With larger machines, the higher degree of I/O concurrency between processes of a single application or between concurrent applications pushes the I/O system to its limits. This leads to a substantial variability in I/O performance. Reducing or hiding this variability is critical, as it is an effective way to make a more efficient use of these new computing platforms through improved predictability of the behavior and of the execution time of applications.

Figure 1 illustrates this variability with the IOR application [Shan and Shalf 2007], a typical benchmark used to evaluate the performance of parallel file systems with predefined I/O patterns. It shows that even with very well optimized I/O (each process here writes the same amount of data contiguously in a separate file using large requests that match the file system's distribution policy) there is a large difference in the time taken by each process to complete its I/O operations within a single I/O phase and also across I/O phases. Since during these I/O phases all processes have to wait for the slowest one before resuming computation, this I/O variability leads to a waste of performance and to unpredictable overall run times. I/O variability is therefore a key issue that we aim to address in this paper.

2.1.2. Causes and Effects of the I/O Variability. Skinner at al. [Skinner and Kramer 2005] point out four causes of performance variability in supercomputers (here presented in a different order).

(1) Communication, causing synchronization between processes that run within the same node or on separate nodes. In particular, network access contention causes collective algorithms to suffer from variability in point-to-point communications.
(2) Kernel process scheduling, together with the jitter introduced by the operating system.
(3) Resource contention within multicore nodes, caused by several cores accessing shared caches, main memory and network devices.
(4) Cross-application contention, which constitutes a random variability coming from simultaneous accesses to shared components of the computing platform, such as the network or the storage system, by distinct applications.

Future systems will have additional sources of variability, such as power management, and fault masking activities. Issues 1 and 2, respectively, cause communication and computation jitter. Issue 1 can be addressed through more efficient network hardware and collective communication algorithms. The use of lightweight kernels with less support for process scheduling can alleviate issue 2. Issues 3 and 4, on the other hand, cause I/O performance variability.

At the level of a node, the increasing number of cores per node in recent machines makes it difficult for all cores to access the network all at once with an optimal throughput. Requests are serialized in network devices, leading to a different service time for each core. This problem is further amplified by the fact that an I/O phase consists of many requests that are thus serialized in an unpredictable manner.

Parallel file systems also represent a well-known bottleneck and a source of high variability [Uselton et al. 2010]. The time taken by a process to write some data can vary by several orders of magnitude from one process to another and from one I/O phase to another depending on many factors, including (1) network contention when several nodes send requests to the same I/O server [Dorier et al. 2014], (2) access contention at the level of the file system's metadata server(s) when many files are created simultaneously [Dorier et al. 2012b], (3) unpredictable parallelization of I/O requests across I/O servers due to different I/O patterns [Lofstead et al. 2010], (4) additional disk-head movements due to the interleaving of requests coming from different processes or applications [Gainaru et al. 2014]. Other source of I/O variability at disk level include the overheads of RAID group reconstruction, data scrubbing overheads, or various firmware activities.

Lofstead et al. [Lofstead et al. 2010] present I/O variability in terms of *interference*, with the distinction between *internal interference* caused by access contention between processes of the same application, and *external interference* that are due to sharing the access to the file system with other applications, possibly running on different clusters. While the sources of I/O performance variability are numerous and difficult to track, we can indeed observe that some of them originate from contentions within a single application, while other come from the contention between multiple applications concurrently running on the same platform. The following section describes how to tackle these two sources of contention.

2.1.3. Approaches to Mitigate the I/O Variability. While most efforts today address performance and scalability issues for specific types of workloads and software or hardware components, few efforts target the causes of performance variability. We highlight two practical ways of hiding or mitigating the I/O variability.

[2]Due to the use of colors, this figure may not be properly interpretable if this document was printed in black and white. Please refer to an electronic version.

Asynchronous I/O. The main solution to prevent an application from being impacted by its I/O consists of using asynchronous I/O operations, i.e., non-blocking operations that proceed in the background of the computation.

The MPI 2 standard proposes rudimentary asynchronous I/O functions that aim to overlap computation with I/O. Yet these functions are available only for independent I/O operations. Besides, popular implementations of the MPI-I/O standard such as ROMIO [Thakur et al. 1999b] actually implement most of these functions as synchronous. Only the small set of functions that handle contiguous accesses have been made asynchronous, provided that the backend file system supports it.

Released in 2012, the MPI 3 standard completes this interface with asynchronous collective I/O primitives. Again, their actual implementation is mostly synchronous. As of today, there is no way to leverage completely asynchronous I/O using only MPI-I/O. Higher-level libraries such as HDF5 [HDF5 2015; Folk et al. 1999] or NetCDF [Unidata 2015] have also no support yet for asynchronous I/O.

Dedicated I/O Resources. Over the past few years, dedicated I/O resources have been proposed to address the limitation of MPI implementations in terms of asynchronous I/O. These resources can take various forms. Explicit I/O threads [Fu et al. 2012] have been used to achieve fully asynchronous I/O at the potential price of additional OS jitter. Dedicated cores have been proposed to leverage a subset of cores in each multi-core node used by the application [Dorier et al. 2012a; Li et al. 2010], and have them perform I/O operations on behalf of the cores that run the application. Staging areas [Abbasi et al. 2009; Nisar et al. 2008; Prabhakar et al. 2011] is another approach that usually consists of dedicated nodes deployed along with an application. Forwarding nodes [Ali et al. 2009; Stone et al. 2006] and burst buffers [Liu et al. 2012; Ma et al. 2006] consist of a set of nodes, independent of the applications and interposed between the compute nodes and the storage system. These nodes may feature a larger memory capacity than compute nodes, in the form of SSDs or NVRAMs.

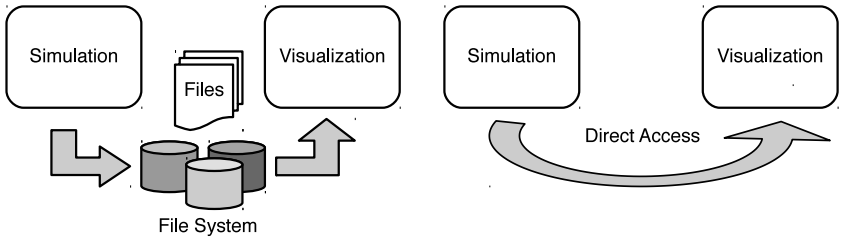
This trend toward using dedicated resources has benefited the field of data analysis and visualization as well, where dedicated cores or nodes are seen as new ways to efficiently get access to simulations' data as they are generated. The next section explores this trend in more details.

2.2. Analysis and Visualization: an Overlooked Process

Data produced by HPC simulations can serve several purposes. One of them is fault tolerance using a checkpoint/restart method. The other, and most important, is the analysis and visualization of the simulated phenomenon. Analysis and visualization are important components of the process that leads from running a simulation to actually *discovering knowledge*.

Given the increasing computation power in recent machines and the trend toward using dedicated resources, it will become more and more common to couple the simulation with the analysis and visualization tools. Simulation/Visualization coupling consists of making the simulation send its data directly to a visualization software instead of storing it and processing it offline. This approach, termed *in situ visualization* and illustrated in Figure 2 (b), has the advantage of bypassing the storage system and producing results faster. It also allows scientists to control their simulations as they run, efficiently overlapping simulation and knowledge discovery.

2.2.1. A Taxonomy of In Situ Visualization Methods. Several in situ visualization strategies exist that we separate into two main categories –tightly coupled and loosely coupled– depending on where visualization tasks run.



(a) Traditional Scientific Workflow (b) Coupling Simulation/Visualization

Fig. 2: Two approaches to retrieve insight from large-scale simulations: (a) the traditional approach of storing data in a parallel file system and reading it offline, (b) the new trend towards simulation/visualization coupling.

Tightly-Coupled In Situ Visualization. In a tightly-coupled scenario, the analysis and visualization codes run on the same node as the simulation and share its resources. The main advantage of this scenario is the proximity to the data, which can be retrieved directly from the memory of the simulation. Its drawback lies in the impact that such analysis and visualization tasks can have on the performance of the simulation and on the variability of its run time. Within this category, we make a distinction between *time partitioning* and *space partitioning*.

Time-partitioning visualization consists of periodically stopping the simulation to perform visualization tasks. This is the most commonly used method. For example, it is implemented in VisIt's *libsim* library [Whitlock et al. 2011] and ParaView's *Catalyst* library [Fabian et al. 2011; Johnston 2014].

In a space-partitioning mode, dedicated cores perform visualization in parallel with the simulation. This mode poses challenges in efficiently sharing data between the cores running the simulation and the cores running the visualization tasks, as these tasks progress in parallel. It also reduces the number of cores available to the simulation.

Loosely-Coupled In Situ Visualization. In a loosely coupled scenario, analysis and visualization codes run on a separate set of resources, that is, a separate set of nodes located either in the same supercomputer as the simulation [Zheng et al. 2010; Rasquin et al. 2011], or in a remote cluster [Malakar et al. 2010]. The data is sent from the simulation to the visualization nodes through the network.

Some in situ visualization frameworks such as GLEAN [Hereld et al. 2011] can be considered hybrid, placing some tasks close to the simulation in a time-partitioning manner while other tasks run on dedicated nodes.

2.2.2. From Offline to In Situ Visualization: Another Source of Variability. The increasing amounts of data generated by scientific simulations also leads to performance degradations when it comes to reading back data for analysis and visualization [Childs et al. 2010; Yu and Ma 2005]. While I/O introduces run time variability, in situ analysis and visualization can also negatively impact the performance of the simulation/visualization complete workflow. For instance, periodically stopping the simulation to perform in situ visualization in a time-partitioning manner leads to a loss of performance and an increase of run-time variability. Contrary to the performance of the simulation itself, the performance of visualization tasks may depend on the content of the data, which makes the rendering tasks imbalanced across processes and across iterations.

This variability is further amplified if the in situ visualization framework is interactive, in which case the user himself impacts the performance of his application.

In a loosely-coupled approach to in situ visualization, sending data through the network potentially impacts the performance of the simulation and forces a reduced number of nodes to sustain the input of a large amount of data. Transferring such large amounts of data through the network also have a potentially larger impact on the simulation than running visualization tasks in a tightly-coupled manner.

2.3. Our Vision: Using Dedicated Cores for I/O and In Situ Visualization

Despite the limitations of the traditional, offline approach to data analysis and visualization, users are still seldom moving to purely in situ visualization and analysis [Yu et al. 2010; Ma et al. 2007; Ma 2009]. The first reason is the development cost of such a step in large codes that were maintained for decades. The second reason is that storage I/O is still required for checkpoint-based fault tolerance, which makes offline analysis of checkpoints the natural candidate for scientific discovery.

To push further the adoption of in situ visualization and increase the productivity of the overall scientific workflow, *we postulate that a framework should be provided that deals with all aspects of Big Data management in HPC simulations*, including efficient I/O but also in situ processing, analysis and visualization of the produced data. Such a framework can at the same time provide efficient storage I/O for data that need to be stored, and efficient in situ visualization to speed up knowledge discovery and enable simulation monitoring.

Over the past 4 years we have been addressing this challenge by proposing, designing and implementing the Damaris system to data management. Damaris proposes to dedicate cores in multicore nodes for any type of data management task, including I/O and in situ visualization. We tried to make Damaris *simple to use*, *flexible*, *portable* and *efficient* in order to ease its adoption by the HPC community. The following section gives an overview of this approach and its implementation.

3. THE DAMARIS APPROACH: AN OVERVIEW

In order to address both I/O and in situ analysis/visualization issues, we propose to gather the I/O operations into a set of dedicated cores in each multicore node. These cores (typically one per node) are dedicated to data management tasks (i.e., they do not run the simulation code) in order to overlap writes and analysis tasks with computation and avoid contention for accesses to the file system. The cores running the simulation and the dedicated cores communicate data through shared memory. We call this approach Damaris. Its design, implementation and API are described below.

3.1. Design Principles

The Damaris approach is based on four main design principles.

3.1.1. Dedicated Cores. The Damaris approach is based on a set of processes running on dedicated cores in every multicore node. Each dedicated core performs in situ processing and I/O in response to user-defined events sent by the simulation. We call a process running the simulation a *client*, and a process running on a dedicated core a *server*. One important aspect of Damaris is that dedicated cores do not run the simulation. This gives dedicated core more freedom in scheduling there data management tasks or add additional processing tasks such as compressions. Such optimizations are discussed in Section 4.1.6.

With the current trend in hardware solutions, the number of cores per node increases. Thus dedicating one or a few cores has a diminishing impact on the performance of the simulation. Hence, our approach primarily targets SMP nodes featur-

ing a large number of cores per node: 12 to 24 in our experiments. This arrangement might be even more beneficial in future systems, for a variety of reasons: The number of cores increasing, neither memory bandwidth nor power constraints may allow all cores to run compute-intensive code; and reduced switching between different types of executions improves performance.

3.1.2. Data Transfers through Shared Memory. Damaris handles large data transfers from clients to servers through shared memory. This makes a write as fast as a `memcpy` and also enables direct allocation of variables within the shared memory. This option is especially useful to reduce the memory requirements of in situ visualization tasks, which can directly access the memory of the simulation without requiring a copy (see our previous work [Dorier et al. 2013]).

3.1.3. High-Level Data Abstraction. Clients write enriched datasets in a way similar to scientific I/O libraries such as HDF5 or NetCDF. That is, the data output by the simulation is organized into a hierarchy of groups and variables, with additional metadata such as the description of variables, their type, unit, and layout in memory. The dedicated cores thus have enough knowledge of incoming datasets to write them in existing high-level formats. This design principle differs from other approaches that capture I/O operations at a lower level [Li et al. 2010; Ma et al. 2006]. These approaches indeed lose the semantics of the data being written. While our design choice forces us to modify the simulation so that it writes its data using Damaris' API, it allows for implementing semantic-aware data processing functionalities in dedicated cores. In particular, keeping this level of semantics is mandatory in order for dedicated cores to be able to write data in a standard, high-level format such as HDF5 or NetCDF, or to feed an in situ visualization pipeline.

3.1.4. Extensibility through Plugins. Servers can perform data transformations prior to writing them, as well as analysis and visualization. One major design principle in the Damaris approach is the possibility for users to provide these transformations through a plugin system, thus adapting Damaris to the particular requirements of their application. Implementing such a plugin system at a lower level would not be possible, as it would not have access to the high-level information about the data (e.g., dimensions of arrays, data types, physical meaning of the variable within the simulation, etc.).

3.2. Architecture

Figure 3 presents the software architecture underlying the Damaris approach. While Damaris can dedicate several cores in large multicore nodes, only one client and one server are represented here.

Damaris has been designed in a highly modular way and features a number of decoupled, reusable software components. The *Shared Memory* component handles the shared buffer and ensures the safety of concurrent allocations/deallocations. The *Distributed Reactor* handles communications between clients and servers, and across servers. The *Metadata Manager* stores high-level information related to the data being transferred (type, size, layout, etc.). Finally the *Plugin Manager* on the server side loads and runs user-provided plugins.

This modular architecture greatly simplified the adaptation to several HPC platforms and simulations, as well as the development of extensions to support various scenarios such as storage, in situ visualization, data compression or I/O scheduling. The following sections describe each component in more detail.

3.2.1. Shared Memory. Data communications between the clients and the servers within a node are performed through the Shared Memory component. A large memory buffer is created on each node by the dedicated cores at start time, with a size set by

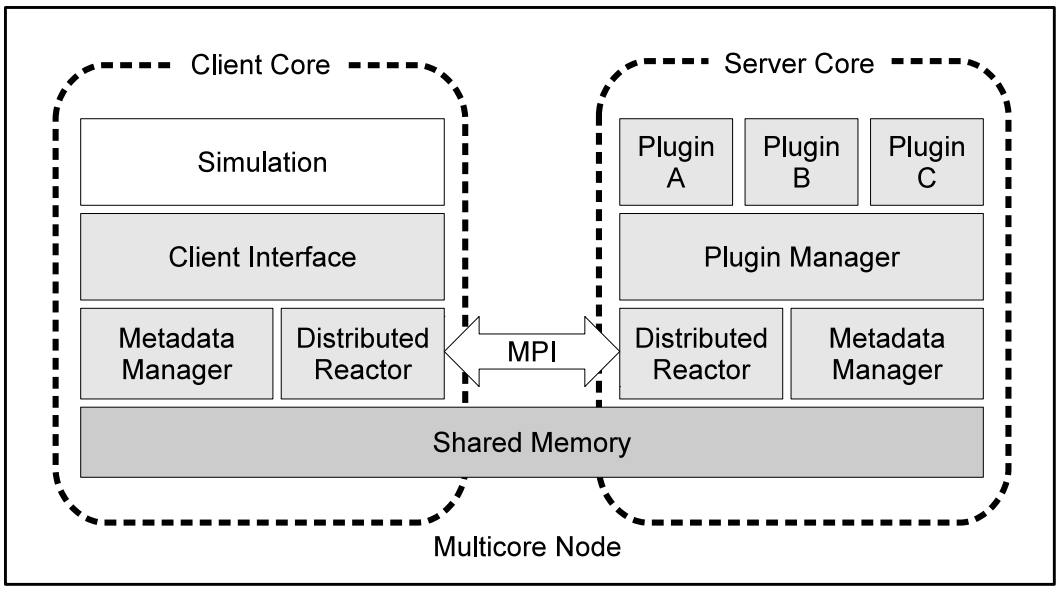


Fig. 3: Software architecture of the implementation of Damaris.

the user (typically several MB to several GB). Thus the user has full control over the resources allocated to Damaris. When a client submits new data, it reserves a segment of this shared-memory buffer. It then copies its data using the returned pointer so that the local memory can be reused.

3.2.2. Distributed Reactor. The Distributed Reactor is the most complex component of Damaris. It builds on the *Reactor* design pattern [Schmidt 1995] to provide the means by which different cores (clients and servers) communicate through MPI. Reactor is a behavioral pattern that handles requests concurrently sent to an application by one or more clients. The Reactor asynchronously listens to a set of *channels* connecting it to its clients. The clients send small events that are associated with event handlers (i.e., functions) in the Reactor. A synchronous event demultiplexer is in charge of queuing the events received by the Reactor and calling the appropriate event handlers. While clients communicate data through shared memory, they use the Distributed Reactor, based on MPI, to send short notifications that either new data is available in shared memory, or that a plugin should be triggered.

Contrary to a normal Reactor design pattern (as used in *Boost.ASIO*[3] for example), our Distributed Reactor also provides elaborate collective operations.

> *Asynchronous atomic multicast:.* A process can broadcast an event to a group of processes at once. This operation is asynchronous, that is, the sender does not wait for the event to be processed by all receivers to resume its activity. A receiver only processes the event when all other receivers are ready to process it as well. It is also atomic, that is, if two distinct processes broadcast a different event, the Distributed Reactor ensures that all receivers will handle the two events in the same order.
> *Asynchronous atomic labeled barrier:.* We call a "labeled" barrier across a set of processes a synchronization barrier associated with an event (its label). After all processes reach the barrier, they all invoke the event handler associated with the

[3]See http://www.boost.org/

event. This ensures that all processes agree to execute the same code at the same *logical* time. This primitive is asynchronous: it borrows its semantics from MPI 3's `MPI_Ibarrier` non-blocking barrier. It is atomic according to the same definition as the asynchronous atomic multicast.

These two distributed algorithms are important in the design of in situ processing tasks that include communications between servers. In particular, they ensure that plugins will be triggered in the same order in all servers, allowing collective communications to safely take place within these plugins.

3.2.3. Metadata Manager. The Metadata Manager component keeps information related to the data being written, including *variables*, *layouts* (describing the type and shape of blocks of data), *parameters*, etc. It is initialized using an XML configuration file.

This design principle is inspired by ADIOS [Lofstead et al. 2008] and other tools such as EPSN [Esnard et al. 2006]. In traditional data formats such as HDF5, several functions have to be called by the simulation to provide metadata information prior to actually writing data. The use of an XML file in Damaris presents several advantages. First, the description of data provided by the configuration file can be changed without changing the simulation itself, and the amount of code required to use Damaris in a simulation is reduced compared to existing data formats. Second, it prevents clients from transferring metadata to dedicated cores through shared memory. Clients communicate only data along with the minimum information required by dedicated cores to retrieve the full description in their own Metadata Manager.

Contrary to the XDMF format [KitWare 2015a], which leverages XML to store scientific datasets along with metadata (or points to data in external HDF5 files), our XML file only provides metadata related to data produced by the simulation. It is not intended to be an output format, or become part of one.

3.2.4. Plugin Manager. The Plugin Manager is the component that loads and stores plugins. Plugins are pieces of C++ or Python codes provided by the user. The Plugin Manager is capable of loading functions from dynamic libraries or scripts as well as from the simulation's code itself. It is initialized from the XML configuration file. Again, the use of a common configuration file between clients and servers allows different processes to refer to the same plugin through an identifier rather than its full name and attributes.

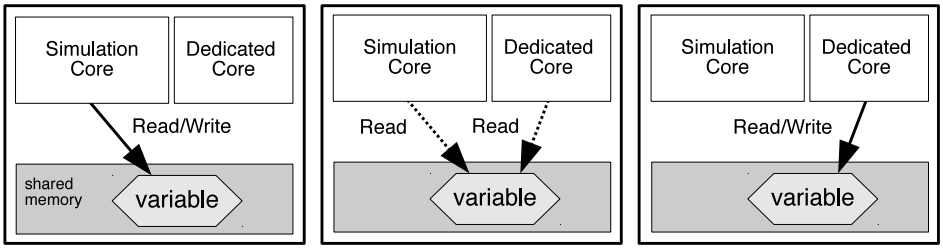
A server can call a plugin when it receives its corresponding event, or wait for all clients in a node or in the entire simulation to have sent the event. In these later cases, the collective algorithms provided by the Distributed Reactor ensure that all servers call the plugins in the same order.

3.3. Implementation

The Damaris approach is intended to be the basis for a generic, platform-independent, application-independent, easy-to-use tool. This section describes its main API and provides some technical details of its implementation.

3.3.1. Client API. Our implementation provides client-side interfaces for C, C++ and Fortran applications written with MPI. While the full API of Damaris can be found in its user guide,[4] we present some of its main functions hereafter.

[4]http://damaris.gforge.inria.fr/doc/DamarisUserManual-1.0.pdf



(a) After damaris_alloc (b) After damaris_commit (c) After damaris_clear

Fig. 4: Semantics of the three functions: (a) at iteration n, a segment is allocated for a given variable through `damaris_alloc`, the simulation holds it. (b) Eventually, a call to `damaris_commit` by the client notifies the dedicated core of the location of the data. From then on, the segment can be read by both processes (client and server), but should not be written or deleted by either of them. Finally, (c) a call to `damaris_clear` indicates that the simulation does not need the segment anymore, dedicated cores can modify it, delete it or move it to a persistent storage.

Initializing/finalizing. Initializing and finalizing Damaris is done through calls to `damaris_initialize("config.xml")` and `damaris_finalize()`, which have to be called respectively at the beginning and at the end of a simulation.

Writing data. `damaris_write("var_name",data)` copies the data in shared memory along with minimal information, and notifies the server on the same node that new data is available. All additional information such as the size of the data and its layout can be found by the servers in the configuration file.

Directly accessing the shared memory. Another way to transfer data from clients to dedicated cores is to directly allocate variables in shared memory and notify the dedicated cores when the data will not be subject to further modifications, at which point the server can start processing it. This is done using the `damaris_alloc("variable")`, `damaris_commit("variable")`, and `damaris_clear("variable")` functions. Figure 4 provides the semantics of these functions.

As shown in our previous work [Dorier et al. 2013], Damaris requires only limited code modifications in existing simulations, and is less intrusive than existing in situ visualization interface on this respect.

3.3.2. Memory management. As explained above, Damaris uses a fixed-size buffer to hold data transfered from clients to dedicated cores. If post-processing tasks in dedicated cores are too slow to cope with the rate at which data is produced by the simulation, it may happen that the buffer becomes full. In this situation, we considered two options. The first one consists of blocking the simulation until dedicated cores have freed enough memory. The second one consists of have future write calls fail without blocking the simulation. This later solution was preferred by the domain scientists with whom we discussed.

3.3.3. Technical Implementation Details. Damaris leverages the *Boost.Interprocess* library[5] to implement several versions of the Shared Memory component, suitable for different platforms.

Our implementation of the Distributed Reactor relies on MPI 2 communication primitives and, in particular, non-blocking *send* and *receive* operations. Events are simply implemented as 0-byte messages with the MPI *tag* carrying the type of the event. Since the MPI 3 standard provides new non-blocking collective functions such as `MPI_Ireduce`

[5]See http://www.boost.org/

or `MPI_Ibarrier`, our Distributed Reactor could be easily re-implemented with these MPI 3 functions without any impact on the rest of Damaris' implementation.

Finally we used Model-Driven Engineering (MDE) techniques to implement the Metadata Manager. Most of the source code of the Metadata Manager is indeed automatically generated from an XSD metamodel. This metamodel describes the concepts of *variables*, *layouts*, etc. as well as their relations to one another and how they are described in an XML format. The XSD file is used to synthesize C++ classes that correspond to the metamodel.

3.4. Managing Data with Damaris

Damaris is not a data format. It only provides a framework to dedicate cores for custom data processing and I/O tasks, to transfer data through shared memory and to call plugins. Thanks to its plugin system, Damaris can be adapted to many scenarios of in situ data processing. In this paper, we specifically use it to periodically write data and to perform in situ visualization.

3.4.1. Writing Data. We implemented a plugin that gathers data from client cores and writes them into HDF5 files. Each server running on a dedicated core produces a single file per iteration. Compared with the file-per-process approach, this way of writing produces fewer, bigger files, thus mitigating the bottleneck in metadata servers when files are created. Writing from a reduced number of writers also has the advantage of limiting network access contention across the cores of the same node. Finally, issuing bigger writes to the file system usually allows for better performance. Compared with the collective I/O approach, our writer plugin does not require synchronization between processes.

3.4.2. Visualizing and Analyzing. The high-level data description provided by Damaris enables a connection with existing visualization and analysis packages, including VisIt [LLNL 2015] or ParaView [KitWare 2015b], in order to build a full in situ visualization framework. Both VisIt and ParaView perform in situ visualization from in-memory data. Given that each of these software has strengths, a major advantage of our approach is the ability to switch between them with no code modification in the simulation.

We leveraged the XSD-based metadata management in Damaris to provide the necessary information to bridge simulations to existing visualization software. By investigating the in situ interfaces of different visualization packages including ParaView, VisIt, ezViz [ERDC DSRC 2015] and VTK [Schroeder et al. 2000], we devised a generic description of visualizable structures such as meshes, points or curves. Additionally, the Distributed Reactor enables synchronization between dedicated cores, which is necessary to run the parallel rendering algorithms implemented by the aforementioned visualization software.

4. EVALUATION

We evaluated Damaris with the CM1 atmospheric simulation [Bryan and Fritsch 2002] and ANL's Nek5000 CFD solver [P. F. Fischer and Kerkemeier 2008], on several platforms: NICS's Kraken [NICS 2015], three clusters of the French Grid'5000 platform [Grid'5000 2015], NCSA's BluePrint cluster and the Blue Waters supercomputer [NCSA 2015]. In the following, we first evaluate Damaris in the context of improving I/O performance by hiding the I/O variability. We then evaluate the use of Damaris for several other data management tasks, including data compression, I/O scheduling and in situ visualization.

4.1. Addressing the I/O Bottleneck with Damaris

In this first evaluation part, we show how Damaris is used to improve I/O performance.

4.1.1. Description of the Applications. The following applications were used in our experiments.

. **CM1 (Cloud Model 1)** is used for atmospheric research and is suitable for modeling small-scale atmospheric phenomena such as thunderstorms and tornadoes. It follows a typical behavior of scientific simulations, which alternate computation phases and I/O phases. The simulated domain is a regular 3D grid representing part of the atmosphere. Each point in this domain is characterized by a set of variables such as local *temperature* or *wind speed*. CM1 is written in Fortran 90. Parallelization is done using MPI, by distributing the 3D array along a 2D grid of equally-sized subdomains, each of which is handled by a process. The I/O phase leverages either HDF5 to write one file per process, or pHDF5 [Chilan et al. 2006] to write in a shared file in a collective manner. One of the advantages of using a file-per-process approach is that compression can be enabled, which cannot be done with pHDF5. However, at large process counts, the file-per-process approach generates a large number of files, making all subsequent analysis tasks intractable.

. **Nek5000** is a computational fluid dynamics solver based on the spectral element method. It is actively developed at ANL's Mathematics and Computer Science Division. It is written in Fortran 77 and solves its governing equations on an unstructured mesh. This mesh consists of multiple elements distributed across processes; each element is a small curvilinear mesh. Each point of the mesh carries the three components of the fluid's local velocity, as well as other variables. We chose Nek5000 for this particular meshing structure, different from CM1, and for the fact that it is substantially more memory-hungry than CM1. We modified Nek5000 in order to pass the mesh elements and fields data to Damaris.

Nek5000 takes as input the mesh on which to solve the equations, along with initial conditions. We call this set a *configuration*. In our experimental evaluation, we used the *MATiS* configuration, which was designed to run on 512 to 2048 cores. Another configuration, *turbChannel*, is used in Section 4.2 to evaluate in situ visualization. This configuration was designed to run on 32 to 64 cores.

4.1.2. Platforms and Configurations. With the CM1 application, our goal was to optimize CM1's I/O for future use on the upcoming Blue Waters Petascale supercomputer. Therefore we started with NCSA's IBM Power5 BluePrint platform as it was supposed to be representative of Blue Waters' hardware. On this platform, we evaluated the scalability of the CM1 application with respect to the size of its output, with the file-per-process and Damaris approaches. We then experimented on the *parapluie* cluster of Grid'5000's Rennes site. This cluster features 24-core nodes, which makes it very suitable to our approach based on dedicated cores. We then moved our experiments to NICS's Kraken supercomputer, which, in addition to allowing runs at much larger scales, has a hardware configuration very close to that of Blue Waters' final design.

With Nek5000, our goal was to confirm the usability of Damaris with a more memory-hungry application. We completed our experimentation on the *stremi* cluster of Grid'5000's Reims site, which provides the same type of hardware as the *parapluie* cluster, but a different network. All these platforms are detailed hereafter, along with the configuration of CM1 and Nek5000 we used.

. **BluePrint** is a test platform used at NCSA until 2011 when IBM was still in charge of delivering the Blue Waters supercomputer.[6] BluePrint features 120 Power5 nodes. Each node consists of 16 cores and includes 64 GB of memory. As for its file system, GPFS is deployed on 2 I/O servers. CM1 was run on 64 nodes (1024 cores), with a $960 \times 960 \times 300$-point domain. Each core handles a $30 \times 30 \times 300$-point subdomain with the standard approaches, that is, when no dedicated cores are used. When dedicating one core out of 16 on each node, computation cores handle a $24 \times 40 \times 300$-point subdomain. On this platform we vary the number of variables that CM1 writes, resulting in different sizes of the output. We enabled the compression feature of HDF5 for all the experiments done on this platform.

. **Grid'5000** is a French grid testbed. We use its *parapluie* cluster on the Rennes site and its *stremi* cluster on the Reims site. On the Rennes site, the parapluie cluster featured 40 nodes of 2 AMD 1.7 GHz CPUs, 12 cores/CPU, 48 GB RAM. We run CM1 on 28 nodes (672 cores) and 38 nodes (912 cores). We deployed a PVFS file system on 15 separate I/O servers (2 Intel 2.93 GHz CPUs, 4 cores/CPU, 24 GB RAM, 434 GB local disk). Each PVFS node was used both as I/O server and metadata server. All nodes (including the file system's) communicate through a 20G InfiniBand 4x QDR link connected to a common Voltaire switch. We use MPICH [ANL 2015] with ROMIO [Thakur et al. 1999a] compiled against the PVFS library, on a Debian Linux operating system. The total domain size in CM1 is $1104 \times 1120 \times 200$ points, so each core handles a $46 \times 40 \times 200$-point subdomain with a standard approach, and a $48 \times 40 \times 200$-point subdomain when one core out of 24 is used by Damaris.

On the Reims site the *stremi* cluster features the same type of node as the *parapluie* cluster. We run Nek5000 on 30 nodes (720 cores). We deploy PVFS on 4 nodes of the same cluster. Each PVFS node is used both as I/O server and metadata server. All nodes communicate through a 1G Ethernet network. We use the MATiS configuration of Nek5000, which contains 695454 elements (small $4 \times 2 \times 4$ curvilinear sub-meshes). These elements are distributed across available simulation processes. Thus the total number of elements (and thus the total amount of data output) does not vary whether we use dedicated cores or not. When no dedicated cores are used, each core handles 965 or 966 such elements. When dedicating one core out of 24, each simulation core handles 1007 or 1008 elements.

. **Kraken** was a supercomputer deployed at the National Institute for Computational Sciences (NICS). It was ranked 11th in the Top500 [Top500 2015] at the time of the experiments, with a peak Linpack performance of 919.1 Teraflops. It featured 9408 Cray XT5 compute nodes connected through a Cray SeaStar2+ interconnect and running Cray Linux Environment (CLE). Each node has 12 cores and 16 GB of local memory. Kraken provided a Lustre file system using 336 block storage devices managed by 48 I/O servers and one metadata server.

On this platform, we studied the weak scalability of the file-per-process, collective I/O and Damaris approaches in CM1, that is, we measured how the run time varies with a fixed amount of data per node. When all cores in each node are used by the simulation, each client process handles a $44 \times 44 \times 200$-point subdomain. Using Damaris, each client process (11 per node) handles a $48 \times 44 \times 200$-point subdomain, which makes the total problem size the same for a given total number of cores.

4.1.3. How Damaris Affects the I/O Variability.

[6]As IBM terminated its contract with NCSA in 2011 and Blue Waters was finally delivered by Cray, BluePrint was later decommissioned and replaced with a test platform, JYC, matching the new Blue Waters' design.

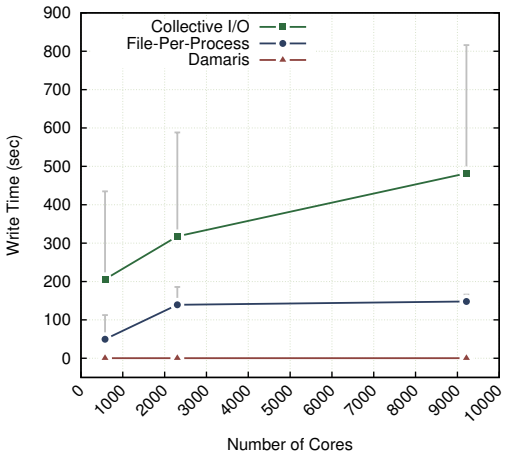


Fig. 5: Duration of a write phase of CM1 on Kraken (average and maximum) *from the point of view of the simulation*. For readability reasons we do not plot the minimum write time. Damaris shows to completely remove the I/O variability while file-per-process and collective-I/O have a big impact on the run-time predictability.

Impact of the Number of Cores on the I/O Variability. We studied the impact of the number of cores on the simulation's write time with the three I/O approaches: file-per-process, collective I/O, and Damaris. To do so, we ran CM1 on Kraken with 576, 2304, and 9216 cores.

Figure 5 shows the average and maximum duration of an I/O phase on Kraken from the point of view of the simulation. It corresponds to the time between the two barriers delimiting the I/O phase. This time is extremely high and variable with Collective I/O, achieving more than 800 seconds on 9216 cores. The average of 481 seconds still represents about 70% of the overall simulation's run time.

By setting the stripe size to 32 MB instead of 1 MB in Lustre, the write time went up to 1600 seconds with a collective I/O approach. This shows that bad choices of file system's configuration can lead to extremely poor I/O performance. Yet it is hard to know in advance the configuration of the file system and I/O libraries that will lead to a good performance.

The file-per-process approach appears to lead to a lower variability, especially at large process count, and better performance than collective I/O. Yet it still represents an unpredictability (difference between the fastest and the slowest phase) of about ± 17 seconds. For a one month run, writing every 2 minutes would lead to an uncertainty of several hours to several days of run time.

When using Damaris, we dedicate one core out of 12 on each node, thus potentially reducing the computation performance for the benefit of I/O efficiency (the impact on overall application performance is discussed in the next section). As a means to reduce the I/O variability, this approach is clearly effective: the time to write from the point of view of the simulation is cut down to the time required to perform a series of copies in shared memory. It leads to an apparent write time of 0.2 seconds (as opposed to the 481 seconds of collective I/O!) and does not depend anymore on the number of processes. The variability is in order of ± 0.1 seconds (too small to be seen on the figure).

Impact of the Amount of Data on the I/O Variability. On BluePrint, we vary the amount of data. We aim to compare the file-per-process approach with Damaris with respect to different output sizes. The results are reported in Figure 6. As we increase the amount of data, the variability of the I/O time increases with the file-per-process approach. With Damaris however, the write time remains in the order of 0.2 seconds for the largest amount of data and the variability in the order of ± 0.1 seconds again.

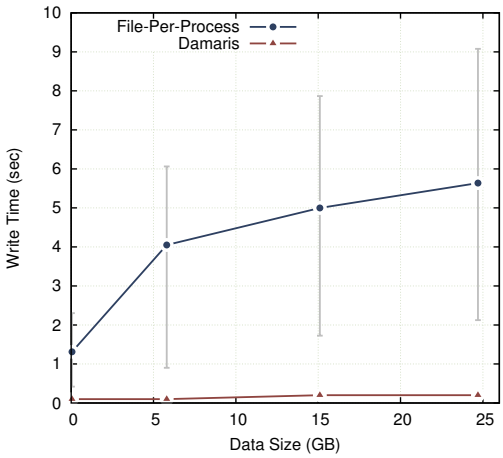


Fig. 6: Duration of a write phase of CM1 on 1024 cores on BluePrint (average, maximum and minimum) using file-per-process and Damaris. The amount of data is given in total per write phase.

Note that on this platform, data compression was enabled. Thus the observed variability comes not only from the bottleneck at the file system level, but also from the different amounts of data that are written across processes and across iterations. This illustrates the fact that I/O variability does not only comes from the variability of performance of data transfers and storage, but also on any pre-processing task occurring before the actual I/O. Damaris is therefore able to hide this pre-processing variability as well.

Impact of the Hardware. We studied the impact of the hardware on the I/O variability using Grid'5000's *parapluie* and *stremi* clusters. With the large number of cores per node (24) in these clusters as well as a network has substantially lower performance than that of Kraken and BluePrint, we aim to illustrate the large variation of write time across cores for a single write phase.

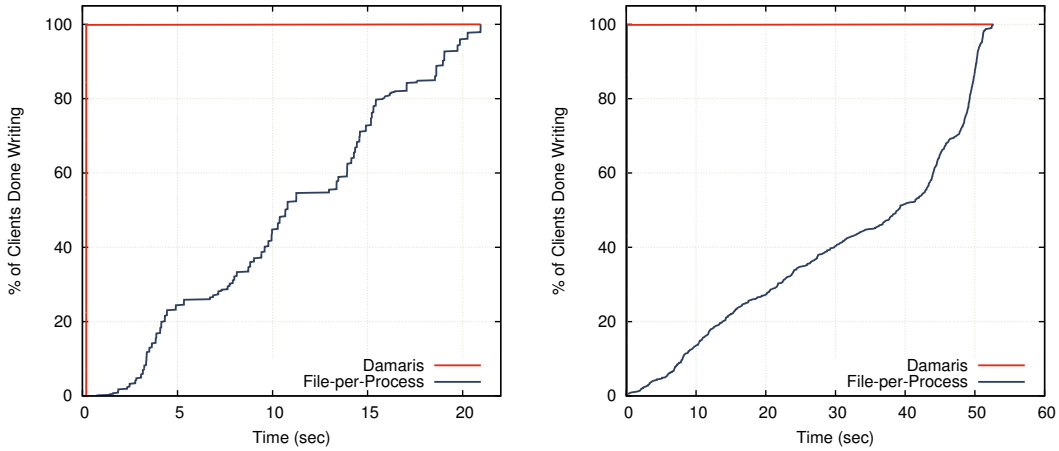
We ran CM1 using 672 cores on the *parapluie* cluster, writing a total of 15.8 GB uncompressed data (about 24 MB per process) every 20 iterations. With the file-per-process approach, CM1 reported spending 4.22% of its time in I/O phases. Yet the fastest processes usually terminate their I/O in less than 1 second, while the slowest take more than 25 seconds. Figure 7 (a) shows the CDF (cumulative distribution function) of write times for one of these write phases, with a file-per-process approach and with Damaris.

Finally we ran Nek5000 using 720 cores on the *stremi*, writing a total of 3.5 GB per iteration using a file-per-process approach. Figure 7 (b) shows the cumulative distribution function of write time for one of these write phases with the file-per-process approach and with Damaris.

In both simulations, we observe a large difference in write time between the fastest and the slowest process with a file-per-process approach, due to access contention either at the level of the network or within the file system. With Damaris however, all processes complete their write at the same time. This is due to the absence of contention when writing in shared memory.

Conclusion. Our experiments show that by replacing write phases with simple copies in shared memory and by leaving the task of performing actual I/O to dedicated cores, Damaris is able to completely hide the I/O variability from the point of view of the simulation, making the application run time more predictable.

4.1.4. Application's Scalability and I/O Overlap.



(a) CM1 on Grid'5000 Rennes cluster (b) Nek5000 on Grid'5000 Reims cluster

Fig. 7: Cumulative distribution function of the write time across processes when running CM1 on 672 cores of Grid'5000's Rennes cluster and Nek5000 on 720 cores of the Reims cluster.

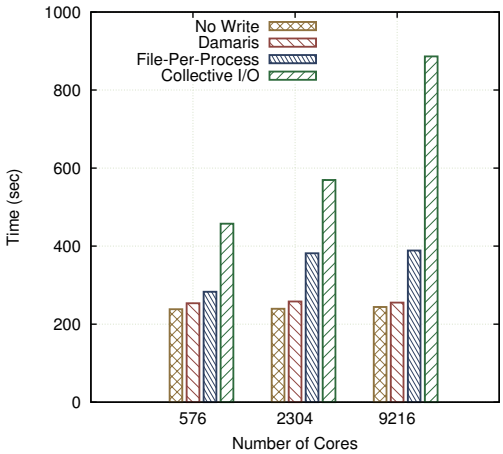


Fig. 8: Average overall run time of the CM1 simulation for 50 iterations and 1 write phase on Kraken.

Impact of Damaris on the Scalability of CM1. CM1 exhibits a very good weak scalability and very stable performance when it does not perform any I/O. Thus, as we increase the number of cores, the scalability becomes mainly driven by the scalability of the I/O phases.

Figure 8 shows the application run time for 50 iterations plus one write phase. The steady run time when no writes are performed illustrates this perfect scalability. Damaris enables a nearly perfect scalability where other approaches fail to scale. In particular, going from Collective I/O to Damaris leads to a $3.5\times$ speedup on 9216 cores.

I/O Overhead. Another way of analyzing the effect of dedicating cores to I/O is by looking at the CPU hours wasted in I/O tasks. With a time-partitioning approach, this overhead corresponds to the duration of a write phase (expressed in hours) multiplied

Table I: I/O Overhead in CPU-hours

Cores	*Simulation without I/O*	File-per-process	Collective-I/O	**Damaris**
576	*38.1*	7.9	32.9	**3.4**
2304	*152.5*	89.2	203.3	**13.8**
9216	*609.8*	378.8	1244.3	**54.5**

CPU hours wasted in I/O tasks (including processes remaining idle waiting for dependent tasks to complete), for 50 computation steps and 1 I/O phase of the CM1 application on Kraken. The "Simulation w/o I/O" column represents the CPU-hours required by the simulation to complete the 50 computation steps at this scale.

by the total number of cores. With dedicated cores, this overhead corresponds to the duration of the computation phase multiplied by the number of dedicated cores. Note that this metrics does not take into account the effect of dedicating cores on the duration of a computation phase, hence the need for the study of the impact on the application's scalability, conducted earlier.

Table I shows the CPU hours wasted in I/O tasks, when running CM1 for 50 computation steps and 1 I/O phase. To put these numbers in perspective, the "Simulation without I/O" column shows the CPU hours required by the simulation to complete the 50 iterations without any I/O and without any dedicated cores. It shows, for example, that using a Collective-I/O approach on 9216 wastes 1244.3 CPU-hours, twice as much as the CPU-hours required by the simulation at this scale. The CPU-hours wasted by Damaris at this scale, on the other hand are as low as 54.5.

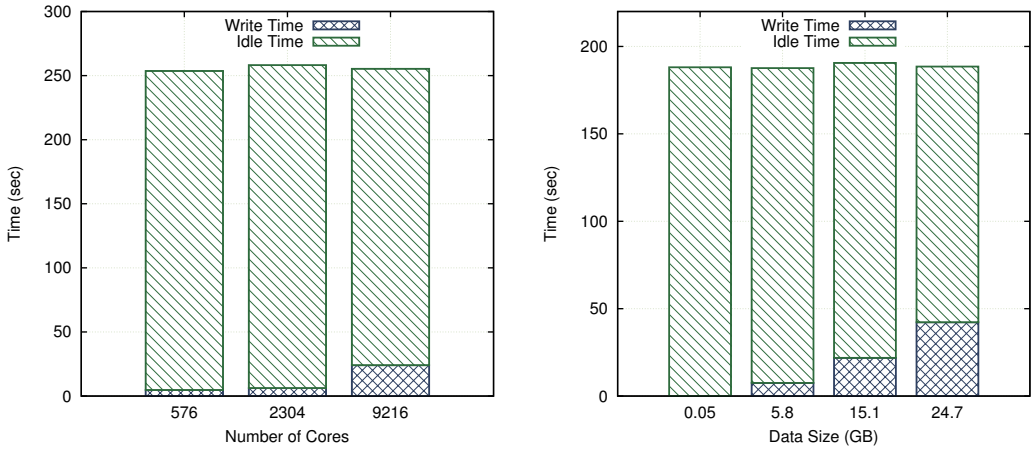
Idle Time in Damaris. Since the scalability of our approach comes from the fact that I/O overlaps with computation, we still need to show that the dedicated cores have enough time to perform the actual I/O while computation goes on.

Figure 9 shows the time used by the dedicated cores to perform the I/O on Kraken and BluePrint with CM1, as well as the time they remain idle, waiting for the next iteration to complete.

As the amount of data on each node is the same, the only explanation for the dedicated cores to take more time at larger process counts on Kraken is the access contention for the file system. On BluePrint the number of processes is constant for each experiment, thus the differences in write time come from the different amounts of data. In all configurations, our experiments show that Damaris has much spare time, during which dedicated cores remain idle. Similar results were obtained on Grid'5000. While the idle time of the dedicated cores may seem to be a waste (provided that no in situ data processing leverages it), it can reduce the energy consumption of the node; this saving will be significant in future systems that will have sophisticated dynamic power management.

With Nek5000, Figure 10 shows the cumulative distribution function (CDF) of the time spent by dedicated cores writing. This time averages to 9.41 seconds, which represents 10% of overall run time. Thus, dedicated cores remain idle 90% of the time. Additionally, this figure shows that the time spent by dedicated cores writing is stable across iterations and across processes, with a standard deviation of 1.08 seconds. This stability allows to add additional data processing tasks without worrying about the possibility that dedicated cores spend an unpredictable time writing.

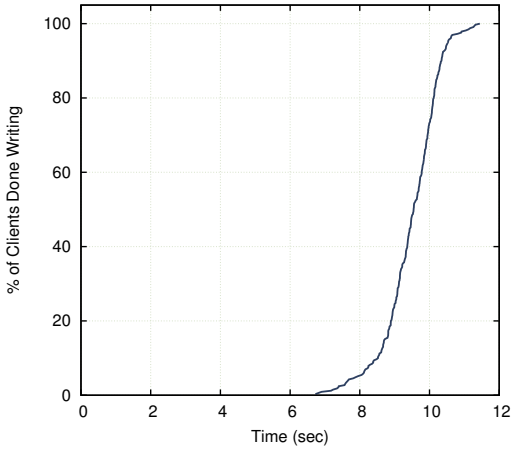
Conclusion. On all platforms, Damaris shows that it can fully overlap writes with computation and still remain idle 75% to 99% of time with CM1 (see Figure 9), and 90% with Nek5000 (see Figure 10). Thus, without impacting the application, it becomes possible to further increase the frequency of output, or to perform additional data processing operations such as in situ data analysis and visualization.



(a) Write / Idle Time on Kraken (b) Write / Idle Time on BluePrint

Fig. 9: Time spent by the dedicated cores writing data for each iteration. The spare time is the time dedicated cores are not performing any task.

Fig. 10: CDF of the time spent by dedicated cores writing (statistics across 11 iterations for 30 dedicated cores), with Nek5000 on the Reims cluster of Grid'5000.



4.1.5. Aggregate I/O Throughput. We then studied the effect of Damaris on the aggregate throughput observed from the computation nodes to the file system, that is, the total amount of data output by the simulation (whether it is transfered directly to the file system or goes through dedicated cores) divided by the amount of time it takes for this data to be stored.

Figure 11 presents the aggregate throughput obtained by CM1 with the three approaches on Kraken. At the largest scale (9216 cores) Damaris achieves an aggregate throughput about 6 times higher than the file-per-process approach, and 15 times higher than collective I/O. The results obtained on 672 cores of Grid'5000 are presented in Table II. The throughput achieved with Damaris here is more than 6 times higher than the other two approaches. Since compression was enabled on BluePrint, we do not provide the resulting throughputs, as it depends on the overhead of the compression algorithm used and the resulting size of the data.

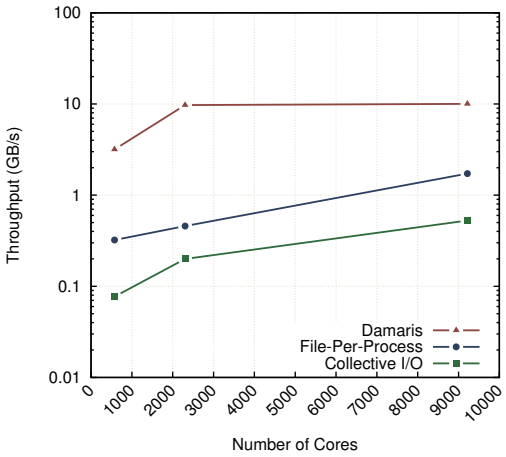


Fig. 11: Average aggregate throughput achieved by CM1 on Kraken with the different approaches. Damaris shows a 6 times improvement over the file-per-process approach and 15 times over Collective I/O on 9216 cores.

Table II: Average aggregate throughput (CM1)

	Aggregate throughput
File-per-process	695 MB/s
Collective-I/O	636 MB/s
Damaris	4.32 GB/s

Average aggregate throughput on Grid'5000's *parapluie* cluster, with CM1 running on 672 cores.

Table III: Avg. aggregate throughput (Nek5000)

	Aggregate throughput
File-per-process	73.5 MB/s
Damaris	337.6 MB/s

Average aggregate throughput on Grid'5000's *stremi* cluster, with Nek5000 running on 720 cores.

A higher aggregate throughput for the same amount of data represents a shorter utilization time of the network and file system. It reduces the probability that the simulation interfere with other applications concurrently accessing these shared resources, in addition to potentially reducing their energy consumption.

With Nek5000 on the *stremi* cluster of Grid'5000, Table III shows that Damaris enables a $4.6\times$ increase of throughput, going from 73.5 MB/s with the file-per-process approach, to 337.6 MB/s with one dedicated core per node.

Conclusion. By avoiding process synchronization and access contention at the level of a node and by gathering data into bigger files, Damaris reduces the I/O overhead, effectively hides the I/O variability and substantially increases the aggregate throughput, thus making a more efficient use of the file system.

4.1.6. Improvements: Leveraging the Spare Time. Section 4.1.4 showed that, with both applications, dedicated cores remain idle most of the time. In order to leverage the spare time in dedicated cores, we implemented two improvements: compression, and transfer delays. These improvements are evaluated hereafter in the context of CM1. Again here, Damaris aggregates data to write one file per dedicated core.

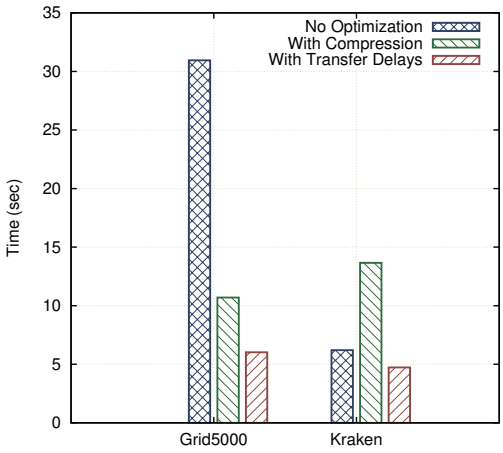


Fig. 12: Write time in the dedicated cores when enabling compression or transfer delays, with CM1 on Grid'5000.

Compression. We used dedicated cores to compress the output data prior to writing it. Using lossless gzip compression, we observed a compression ratio of $1.87:1$. When writing data for offline visualization, atmospheric scientists can afford to reduce the floating point precision to 16 bits, as it does not visually impact the resulting images. Doing so leads to nearly $6:1$ compression ratio when coupling with gzip. On Kraken, the time required by dedicated cores to compress and write data was twice longer than the time required to simply write uncompressed data. Yet contrary to enabling compression in the file-per-process approach, the overhead and jitter induced by the compression phase is completely hidden within the dedicated cores, and do not impact the running simulation. In other words, *compression is offered for free* by Damaris.

Data Transfer Delays. Additionally, we implemented in Damaris the capability to delay data movements. The algorithm is very simple and does not involve any communication between processes: each dedicated core computes an estimated duration of a simulation iteration by measuring the time between two consecutive calls to `damaris_end_iteration` (about 230 seconds on Kraken). This time is then divided into as many slots as there are dedicated cores. Each dedicated core waits for its slot before writing. This avoids access contention at the level of the file system. We evaluated this strategy on 2304 cores on Kraken, the aggregate throughput reaches 13.1 GB/s on average, instead of 9.7 GB/s when this algorithm is not used, which improves the file system utilization and makes dedicated cores spare more time that can be leveraged for other in situ processing tasks.

Summary. These two improvements have also been evaluated on 912 cores of Grid'5000. All results are synthesized in Figure 12, which shows the average write time in dedicated cores. The delay strategy reduces the write time in both platforms. Compression however introduces an overhead on Kraken, thus we are facing a trade-off between reducing the storage space used or reducing the spare time. A potential optimization would be to enable or disable compression at run time depending on the need to reduce write time or storage space.

4.2. Using Damaris for In Situ Visualization

Far from being restricted to performing I/O, Damaris can also leverage the high-level description of data provided in its configuration file to feed in situ visualization pipelines. In the following we evaluate such use of Damaris for in situ visualization.

We highlight two aspects: scalability of the visualization algorithms when using dedicated cores, and impact of in situ visualization on application run time.

4.2.1. Platforms and Configurations. We use again the CM1 and Nek5000 applications presented in the previous sections, respectively on Blue Waters and Grid'5000. The platforms and configurations of the experiments are described hereafter.

. **Blue Waters** Blue Waters [NCSA 2015] is a 13.3-petaflops supercomputer deployed at NCSA. It features 26,864 nodes in 237 Cray XE6 cabinets and 44 Cray XK7 cabinets, running Cray Linux Environment (CLE). We leveraged the XE6 nodes, each of which features 16 cores.

Methodology with CM1 on Blue Waters. CM1 requires a long run time before an interesting atmospheric phenomenon appears, and such a phenomenon may not appear at small scale. Yet contrary to the evaluation of I/O performance, we need visualizable phenomena to appear in order to evaluate the performance of in situ visualization tasks. Thus we first ran CM1 with the help of atmospheric scientists to produce relevant data. We generated a representative dataset of $3840 \times 3840 \times 400$ points spanning several iterations.

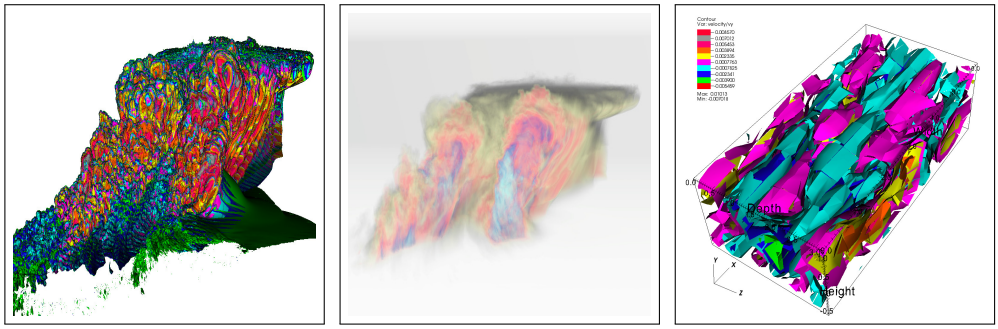
We then extracted the I/O kernel from the CM1 code and built a program that replays its behavior at a given scale and with a given resolution by reloading, redistributing and interpolating the precomputed data. The I/O kernel, identical to the I/O part of the simulation, calls Damaris functions to transfer the data to Damaris. Damaris then performs in situ visualization through a connection to VisIt's *libsim* library [Whitlock et al. 2011], either in a time-partitioning manner or using dedicated cores. Our goal with CM1 is to show the interplay between the scalability of the visualization tasks and the use of dedicated cores to run them.

Methodology with Nek5000 on Grid'5000. With Nek5000, we used the *stremi* cluster of Grid'5000 already presented in the previous section. In addition to the *MATiS* configuration, we also use the *turbChannel* configuration, which runs at smaller scales and is more appropriate for interactive in situ visualization. Our goal with Nek5000 is to show the impact of in situ visualization on the variability of the application's run time.

Using Damaris in Time-Partitioning Mode. In order to compare the traditional "time-partitioning" approach with the use of dedicated cores enabled by Damaris, we added a time-partitioning mode in Damaris. This mode, which can be enabled through the configuration file, prevents Damaris from dedicating cores, and runs all plugins in a synchronous manner on all cores running the simulation. This mode allows us to compare the traditional time-partitioning in situ visualization approach with the use of dedicated cores without having to modify the simulations twice.

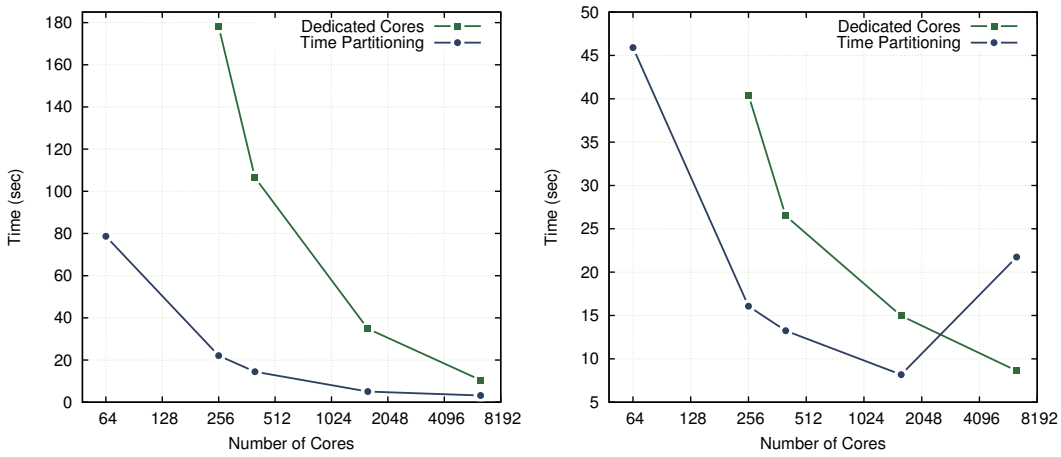
4.2.2. Impact of Dedicated Cores on the Scalability of Visualization Tasks. With CM1 on Blue Waters, we measured the time (average of 15 iterations) to complete either an isosurface rendering or a ray casting rendering using time partitioning and dedicated cores for each scenario. The comparative results are reported in Figure 14.

The isosurface algorithm (resulting image presented in Figure 13 (a)) scales well with the number of cores using both approaches. A time-partitioning approach would thus be appropriate if the user does not need to hide the run time impact of in situ visualization. However, at the largest scale, the time to render from 400 dedicated cores is 10.5 seconds while the rendering time on all 6400 cores is 3.2 seconds. In terms of pure computational efficiency, an approach based on dedicated cores is thus 4.8 times more efficient.



(a) CM1 Isosurface (b) CM1 Ray Casting (c) Nek5000 Isosurface

Fig. 13: Example results obtained in situ with Damaris: (a) 10-level isosurface of the DBZ variable on 6400 cores (Blue Waters). (b) Ray-casting of the *dbz* variable on 6400 cores (Blue Waters). (c) Ten-level isosurface of the y velocity field in the TurbChannel configuration of Nek5000.



(a) In Situ Isosurface (b) In Situ Ray Casting

Fig. 14: Rendering time using ray-casting and isosurfaces, with time-partitioning and dedicated cores with CM1. Note that the number of cores represents the total number used for the experiments; using a dedicated-core approach, 1/16 of this total number is effectively used for in situ visualization, which explains the overall higher rendering time with dedicated cores.

The ray-casting algorithm (resulting image presented in Figure 13 (b)) on the other hand has a poorer scalability. After decreasing, the rendering time goes up again at a 6400-core scale, and it becomes about twice more efficient to use a reduced number of dedicated cores to complete this same rendering.

Conclusion. The choice of using dedicated cores versus a time-partitioning in situ visualization approach depends on (1) the intended visualization scenario, (2) the scale of the experiments and (3) the intended frequency of visual output. Our experiments show that at small scale, the performance of rendering algorithms are good enough to be executed in a time-partitioning manner, provided that the user is ready to increase the run time of his simulation. At large scale however, it becomes more efficient to

use dedicated cores, especially when using ray-casting, where the observed rendering performance is substantially better when using a reduced number of processes.

4.2.3. Impact of In Situ Visualization on Run Time Variability. Our goal in this series of experiments is to show the impact of in situ visualization tasks on the run-time variability of the simulation, and to show how dedicated cores help alleviate this variability. We show in particular the effect of interactivity on this variability. We use Nek5000 for this purpose.

Figure 13 (c) shows the result of a 10-level isosurface rendering of the fluid velocity along the y axis, with the TurbChannel case. We use the MATiS configuration to show the scalability of our approach based on Damaris against a standard, time-partitioning approach.

Results with the TurbChannel Configuration. To assess the impact of in situ visualization on the run time, we run TurbChannel on 48 cores using the two approaches: first we use a time-partitioning mode, in which all 48 cores are used by the simulation and synchronously perform in situ visualization. Then we switch on one dedicated core per node, leading to 46 cores being used by the simulation while 2 cores asynchronously run the in situ visualization tasks.

In each case, we consider four scenarios:

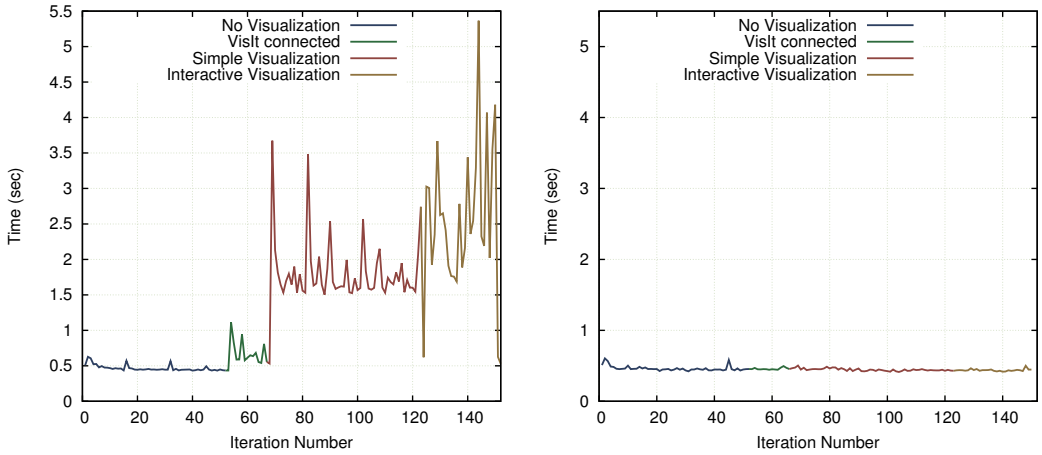
(1) The simulation runs without visualization;
(2) A user connects VisIt to the simulation but does not ask for any output;
(3) The user asks for isosurfaces of the velocity fields but does not interact with VisIt any further (letting the Damaris/Viz update the output after each iteration);
(4) The user has heavy interactions with the simulations (for example rendering different variables, using different algorithms, zooming on particular domains, changing the resolution).

Figure 15 presents a trace of the duration of each iteration during the four aforementioned scenarios using the two approaches. Figure 15 (a) shows that in situ visualization using a time-partitioning approach has a large impact on the simulation run time, even when no interaction is performed. The simple act of connecting VisIt without rendering anything forces the simulation to at least update metadata at each iteration, which takes time. When a visualization scenario is defined but the user does not interact with the running simulation, the run time still presents a large variability. This is due to load imbalance across processes and across iterations, as well as network performance variability when sending visual results to the user. Figure 15 (b) shows that in situ visualization based on dedicated cores, on the other hand, is completely transparent from the point of view of the simulation.

Results with the MATiS Configuration. We ran the MATiS configuration on 816 cores of the *stremi* cluster. Each iteration takes approximately one minute and due to the size of the mesh, it is difficult to perform interactive visualization. Therefore we connect VisIt and simply query for a 3D pseudo-color plot of the vx variable (x component of the fluid velocity) that is then updated at desired iterations.

For the following results, the time-partitioning approach initially outputs one image every time step, while dedicated cores adapted the output frequency to one image every 25 time steps in order to avoid blocking the simulation when the shared memory buffer becomes full. To conduct a fair comparison, we thus setup the time-partitioning mode such that it outputs one image every 25 iterations.

Figure 16 reports the behavior of the application with and without visualization performed, and with and without dedicated cores, for the configurations described above. Corresponding statistics are presented in Table IV.



(a) Time-Partitioning (b) Dedicated Cores

Fig. 15: Variability in run time induced by different scenarios of in situ interactive visualization.

Table IV: Average iteration time

Iteration Time		Average	Std. dev.
Time Partitioning	w/o vis.	75.07 sec	22,93
	with vis.	83.16 sec	43.67
Space Partitioning	w/o vis.	67.76 sec	20.09
	with vis.	64.79 sec	20.44

Average iteration time of the Nek5000 MATiS configuration with a time-partitioning approach and with dedicated cores, with and without visualization.

Conclusion. Time-partitioning visualization not only increases the average run time but also increases the standard deviation of this run time, making it more unpredictable. On the other hand, the approach based on dedicated cores yields more consistent results. One might expect dedicated cores to interfere with the simulation as it performs intensive communications while the simulation runs. However, in practice we observe very little such run time variation.

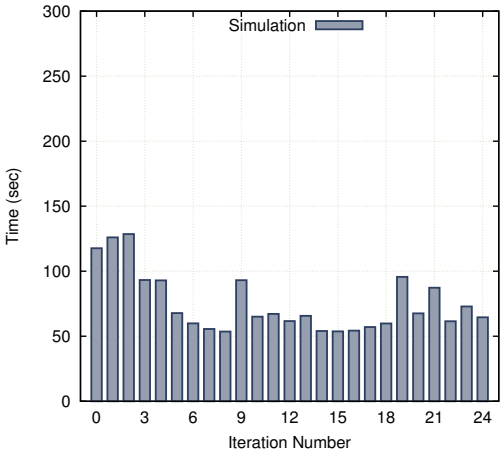
We also remark that decreasing the number of cores used by the simulation can actually decreases its run time. Nek5000 on Grid'5000, for instance, has to run with a number of nodes that is too large, in order to have enough memory.

5. DISCUSSION: DEDICATED CORES VS. DEDICATED NODES

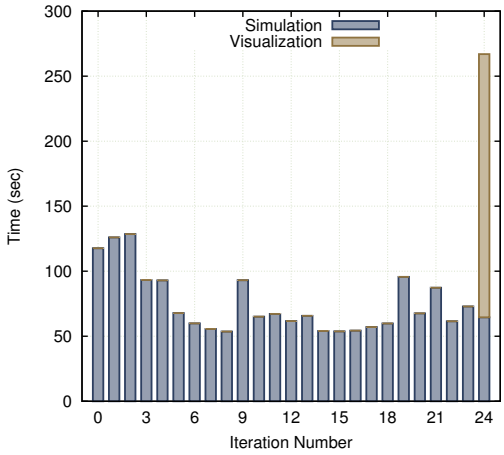
Two important questions can be asked about approaches like Damaris, which propose to dedicate cores for data processing and I/O.

— How many dedicated cores should be used?
— How does dedicating cores compares with dedicating nodes?

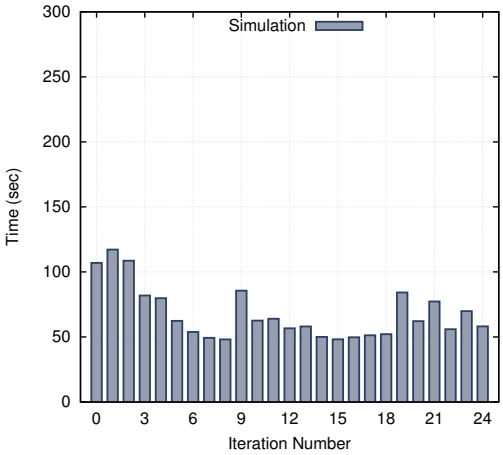
In this section we propose to answer these two questions through experiments with the CM1 and Nek5000 simulations on Grid'5000. We implemented in Damaris the option to use dedicated nodes instead of dedicated cores. Some details of this implementation are given hereafter, before diving into our experimental results.



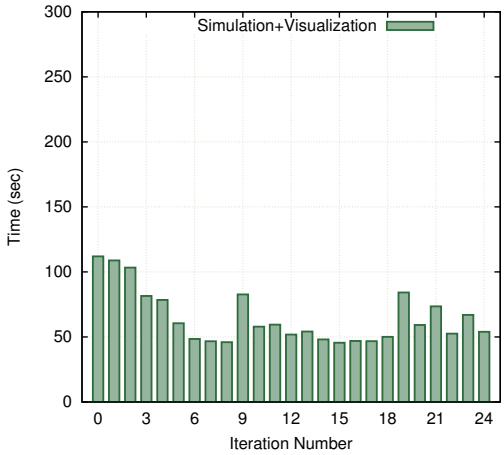
(a) Time Partitioning, w/o Visualization



(b) Time Partitioning, with Visualization performed every 25 time step



(c) Space Partitioning, w/o Visualization



(d) Space Partitioning, with Visualization performed every 25 time steps

Fig. 16: Iteration time of the MATiS configuration with a time partitioning approach (top) or a space partitioning approach (bottom), without visualization (left), with visualization (right).

We restrict our study to I/O. The choice of dedicating cores over dedicating nodes for in situ visualization indeed depends on too many parameters (including the amount of data involved, the simulation, the platform, and most importantly the visualization scenarios) and deserves an entire study that we reserve for a future work.

5.1. Dedicated Nodes in Damaris

In order to compare dedicated cores with dedicated nodes, we either needed a state-of-the-art framework that provides dedicated nodes, such as DataSpace [Rutgers 2015], or to implement dedicated nodes inside the Damaris framework. We chose the later because (1) our simulations are already instrumented with Damaris' API, allowing us to switch between each approach without having to modify the simulation with another

framework's API, and (2) comparing the use of dedicated cores in Damaris with the use of dedicated nodes in another framework would make it harder to distinguish performance benefits coming from the approach (dedicated cores vs. dedicated nodes) from performance benefits coming from specific optimizations of the framework itself. This following section gives an overview of our implementation of dedicated nodes in Damaris.

5.1.1. Implementation. The implementation of dedicated nodes in Damaris relies on asynchronous MPI communications through Damaris' Distributed Reactor. Each simulation core is associated with a server running in a dedicated node. A dedicated node hosts one server on each of its cores. Different simulation cores may thus interact with the same dedicated node, but with a different core (a different server) in this node.

When a client calls `damaris_write`, it first sends an event to its associated server. This event triggers a `RemoteWrite` callback in the server. When the server enters this callback, it starts a blocking `receive` to get the data sent by the client. The client sends its data to the server, along with metadata information such as the *id* of the variable to which the data belongs. A buffer is maintained in clients to allow these transfers to be non-blocking. When the client needs to send data to dedicated nodes, it copies the data into this buffer and issues a non-blocking `send` to the server using the copied data (note that this communication phase is non-blocking in clients, but blocking on servers). The status of this operation is checked in later calls to the Damaris API and the buffer is freed when the transfer is completed.

Other solutions exist in the literature, for example using RDMA [Docan et al. 2010] (remote direct memory access). We chose to use simple asynchronous communications for simplicity and portability. The flexibility of our design, along with the recent addition of dynamic RDMA windows in the MPI 3 standard, would ease such an RDMA-based implementation in Damaris in a near future.

5.1.2. "Switching Gears". Switching between dedicated cores and dedicated nodes, as well as changing the number of dedicated resources, can be done through the configuration, without recompiling the application.

— `<dedicated cores="n" nodes="0"/>` enables n dedicated cores per node. In our current implementation of Damaris, the number of cores per node must divide evenly into the number of dedicated cores.
— `<dedicated cores="0" nodes="n"/>` enables n dedicated nodes. The total number of nodes must divide evenly into the number of dedicated nodes.
— `<dedicated cores="0" nodes="0"/>` disables dedicated cores and nodes. It triggers the time-partitioning mode.

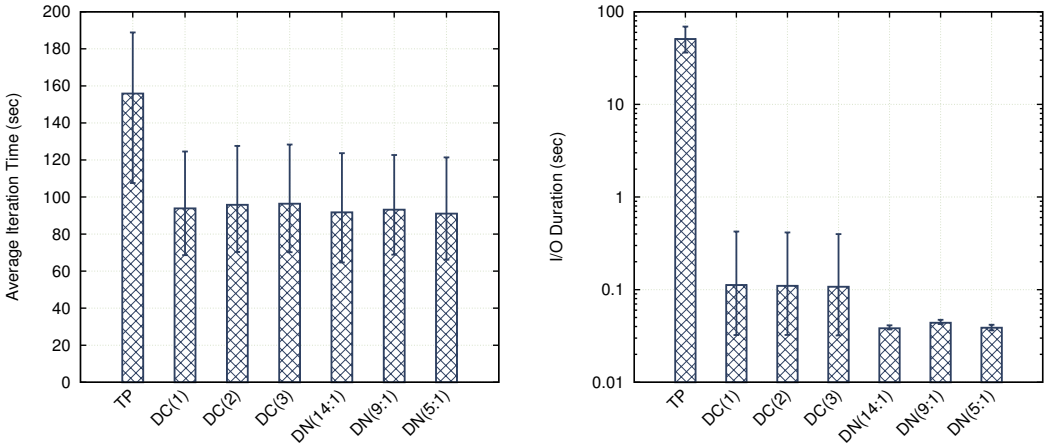
This configuration would allow for a hybrid approach that uses both dedicated cores and dedicated nodes. However this approach is not supported by Damaris yet, as we haven't found any real-life scenario that would benefit from it.

The implementation of all three approaches –time partitioning, dedicated cores, dedicated nodes– within the same framework allows us to evaluate their respective performance in the next sections.

5.2. Dedicated Core(s) vs. Dedicated Nodes: an Experimental Insight

In the following, we present the results obtained with the Nek5000 and CM1 simulations, using the different modes in which Damaris can now operate.

5.2.1. Results with the Nek5000 Application. We used the MATiS configuration of Nek5000 and ran it on 30 nodes (720 cores) of the Grid'5000's *stremi* cluster. We deploy PVFS



(a) Average iteration time

(b) Average I/O time

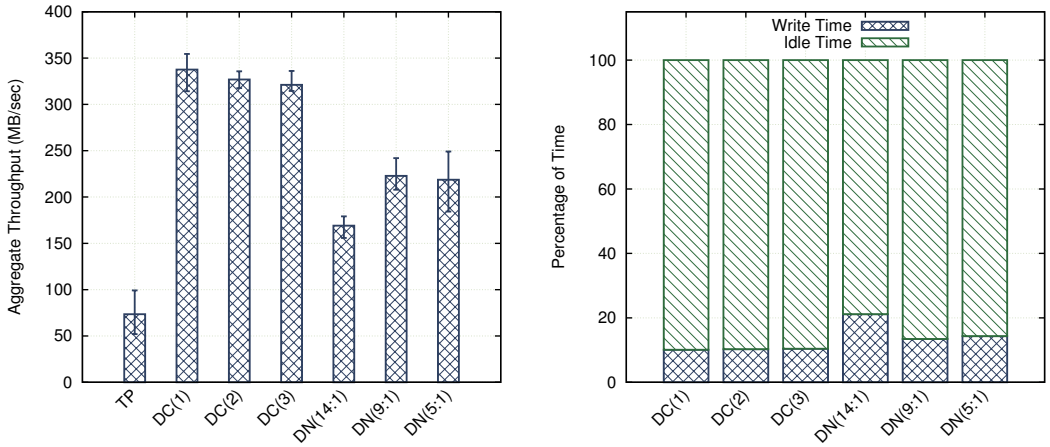
Fig. 17: Experiment with Nek5000 on 720 cores of Grid'5000 *stremi* cluster. Damaris is configured to use either no dedicated resources (TP), x =1, 2 or 3 dedicated cores (DC(x)), or a ratio of x computation nodes to y dedicated nodes (DN($x : y$)). We report (a) the average, maximum and minimum time of a single iteration (computation+I/O), and (b) the average, maximum and minimum time (logarithmic scale) of an I/O phase from the point of view of the simulation.

on 4 additional nodes of this cluster. All nodes (including the file system) communicate through a 1G Ethernet network.

Nek5000 initially wrote most of its checkpoint/restart data in the form of ASCII files, which appeared to be highly inefficient compared to using a high-level data format such as HDF5. We thus rewrote its I/O part as an HDF5-based plugin for Damaris, and used Damaris in 7 configurations: without dedicated resources (time partitioning, abbreviated TP), using 1, 2, or 3 dedicated cores per node (abbreviated DC(1), DC(2) and DC(3)), and using 2, 3 or 5 dedicated nodes (DN(14:1), DN(9:1), DN(5:1) respectively, where the notation $x : y$ represents the ratio of computation nodes to dedicated nodes). Despite the different number of simulation cores in each configuration, the same mesh is used as input for Nek5000 and, therefore, the same amount of data is produced (about 3.5 GB per iteration). We run Nek5000 for 10 such iterations in each configuration.

Overall run time. All configurations based on dedicated resources enable a 40% decrease of overall run time compared with the time-partitioning configuration. Note that because of the inherent variability of the duration of the computation phases within a single iteration (represented in Figure 17 (a) by the minimum and maximum iteration times), it is not possible to tell which of the configuration is actually the best one. Considering these results only, we can argue that using more dedicated cores or more dedicated nodes is potentially an advantageous choice (as long as the efficiency of running your simulation is not affected) because it offers more resources for post-processing and I/O tasks. The choice of using dedicated cores or dedicated nodes can then be based on the characteristics of these post-processing time (scalability, memory requirement, execution time, etc.).

I/O impact. Figure 17 (b) shows that the duration of the I/O phase as perceived by the simulation becomes negligible when using an approach based on dedicated re-



(a) Average aggregate throughput (b) Spare time in dedicated resources

Fig. 18: Experiment with Nek5000 on 720 cores of Grid'5000 *stremi* cluster. Damaris is configured to use either no dedicated resources (TP), x =1, 2 or 3 dedicated cores (DC(x)), or a ratio of x computation nodes to y dedicated nodes (DN($x : y$)). We report (a) the average, maximum and minimum aggregate throughput from writer processes, and (b) the spare time in dedicated processes for the approaches that leverage them.

sources. Dedicated cores reduce this time down to about 0.1 seconds, while dedicated nodes reduce it to about 0.04 seconds. This difference in communication time between dedicated cores and dedicated nodes can be easily explained. When using dedicated cores, the client competes with other clients for the access to a mutex-protected segment of shared memory. When using dedicated nodes on the other hand, this contention does not occur, as each client simply makes a local copy of its data and issues a non-blocking send that proceeds in parallel with the simulation. Therefore, while the I/O phase appears faster with dedicated nodes, our results do not show the potential impact that background communications with dedicated nodes may have on the performance of the simulation.

Aggregate throughput. From the point of view of writer processes (or from the point of view of the file system), the different configurations lead to different aggregate throughput. Figure 18 (a) shows that dedicated cores achieve the highest throughput. This throughput is slightly degraded as the number of dedicated cores per node increases, due to contention between dedicated cores on the same node. Dedicated nodes also increase the aggregate throughput compared with time partitioning, but do not achieve the throughput of dedicated cores. This is due to the fact that all cores in dedicated nodes are writing, and thus compete for the network access at the level of each single dedicated nodes. Additionally, the lower throughput observed when using only two dedicated nodes can be explained by the fact that the file system features four data servers. Therefore, dedicating only two nodes does not fully take advantage of parallelism across writers.

Spare time. Finally, Figure 18 (b) shows the spare time in dedicated resources. In all configurations based on dedicated cores, the dedicated cores spend 10% of their time writing, and remain idle 90% of the time. Dedicated nodes spend slightly more time writing (from 13 to 20% of their time). This is a direct consequence of the difference in aggregate throughput.

Conclusion. Overall, all the configurations based on dedicated resources improve the simulation run time in a similar way. These configurations however differ in other aspects. By avoiding contention at the level of a node, dedicated cores achieve a higher throughput and therefore spare more time that can be used for data processing. Yet, if we weight this spare time by the number of cores that can be used to leverage it (90 when dedicating 3 cores per node, 120 when dedicating 5 nodes), the configuration based on 5 dedicated nodes appears to spare more resources (core-seconds) in spite of sparing less time per core.

The choice of whether one should use an approach based on dedicated cores or dedicated nodes is of course not restricted to these considerations. Some memory-bound simulations may not be able to afford allocating shared memory to dedicated cores, and would rather benefit from dedicated nodes. Some I/O intensive simulations on the other hand may not be able to transfer large amounts of data to a reduced number of dedicated nodes and will prefer dedicated cores.

5.2.2. Results with the CM1 application. In this section, we leverage experiments with the CM1 simulation to show that the choice of one approach over another also depends on the platform considered.

We used CM1 on Grid'5000's Nancy and Rennes sites. On the Nancy site we use the *graphene* cluster. Each node of this cluster consists of a 4-core Intel Xeon 2.53 GHz CPU with 16 GB of RAM. Intra-cluster communication is done through a 1G Ethernet network. A 20G InfiniBand network is used between these nodes and the OrangeFS file system deployed on 6 I/O servers.

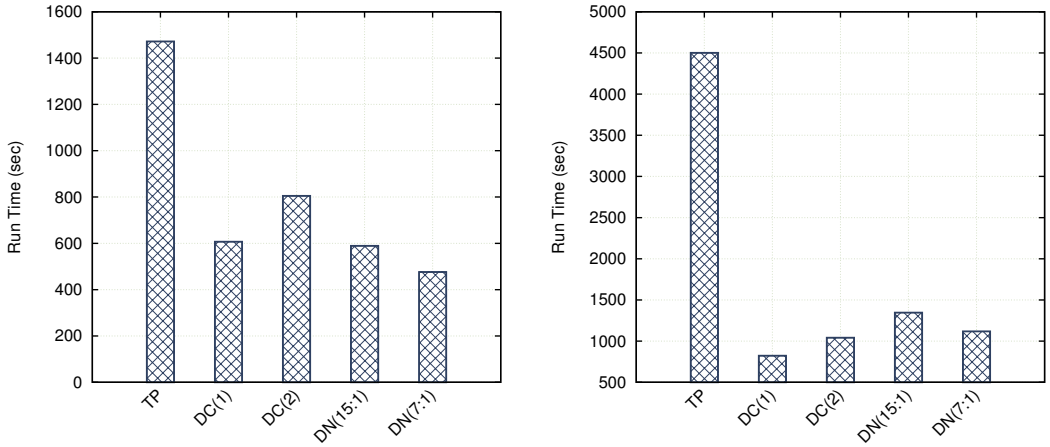
On the Rennes site we use the *parapluie* cluster, already presented in Section 4.1. The nodes communicate with one another through a 1G Ethernet network and with an OrangeFS file system deployed on 3 servers across a 20G InfiniBand network.

We deploy CM1 on 32 nodes (128 cores) on the Nancy site. On the Rennes site, we deploy it on 16 nodes (384 cores). In both cases, we configure CM1 to complete 2520 time steps. We vary its output frequency, using 10, 20 or 30 time steps between each output. Damaris is configured to run with CM1 in five different scenarios that cover the three I/O approaches considered: time partitioning, dedicated cores (one or two – DC(1) and DC(2)), and dedicated nodes using a ratio of 7:1 (DN(7:1), 7 compute nodes for one dedicated node) or 15:1 (DN(15:1), 15 compute nodes for one dedicated node). DN(7:1) thus uses four dedicated nodes on the Nancy site, two on the Rennes site. DN(15:1) dedicates two nodes on the Nancy site, one on the Rennes site.

Impact of the platform. Figure 19 shows that in both clusters, dedicating resources drastically improve the performance of CM1 compared with a time-partitioning approach. Dedicating four nodes on Nancy enables an almost $3\times$ overall speedup, while dedicating one core in each node on the Rennes cluster leads to more than $5\times$ speedup. Our results also show that the best approach in terms of overall run time depends on the platform. It consists of using dedicated nodes with a 7:1 ratio on the Nancy cluster, and using one dedicated core per node on the Rennes cluster. This conclusion is not surprising, since the Nancy cluster only provides 4 cores per node. Dedicating some of these cores thus has a large impact on the simulation. On the Rennes cluster, which provides 24 cores per node, dedicating some of these cores does not remove such an important fraction of computational power from the simulation and is thus more efficient than dedicating nodes.

5.3. Conclusion

Over the years, several research groups have proposed new approaches to I/O and data processing based on dedicated resources. These approaches can be divided into a group of approaches based on dedicated cores, and a group of approaches based on



(a) Run time of CM1 on Nancy cluster

(b) Run time of CM1 on Rennes cluster

Fig. 19: Experiment with CM1 on Grid'5000 Rennes (24 cores per node) and Nancy (4 cores per node) sites. Damaris is configured to use either no dedicated resources (TP), x =1 or 2 dedicated cores (DC(x)), or a ratio of 7:1 or 15:1 dedicated nodes (DN(7:1) and DN(15:1)). We report total run time for 2520 time steps.

dedicated nodes. While Damaris was initially part of the first one, we extended it to support a wider range of configurations. It now offers to dedicate either a subset of cores in each multicore node, or entire nodes. Additionally, it also offers to not dedicate any resource at all, performing all data processing and movement synchronously. This flexibility, made possible in particular through a configuration file that allows us to switch between modes very easily, let us to compare these approaches.

Our results show that dedicating resources for I/O is a highly efficient method to improve the I/O performance of a simulation, both in terms of overall run time, aggregate throughput and performance variability. They also highlighted the fact that there is no clear advantage of one approach over the other, at least for the considered applications: dedicating cores appears more efficient than dedicated nodes under certain conditions, and the opposite holds under different conditions. The choice of using dedicated cores or dedicated nodes, and how many of such resources, depends on the memory requirements of the simulation, the memory requirements of the data processing tasks running in plugins, the scalability of the simulation, the scalability of the processing tasks, and the amount of data involved. It is difficult to give rules of thumbs, although some factors such as a simulation being memory-bound and the post-processing tasks being memory-hungry, should direct the user to using dedicated nodes rather than dedicated cores, for example. The strength of Damaris lies in the fact that switching from dedicated cores to dedicated nodes and changing their numbers is only a matter of changing a line in a configuration file, making it possible for trial-and-error runs in order to find an appropriate configuration. The choice of one approach over the other may also depend on criteria other than the overall run time. Our experiments with Nek5000 showed that while this run time is very similar under the different approaches, the resulting aggregate throughput favors dedicating cores, while the resulting spared resources (spare time $\times$ number of cores in dedicated resources) advocates for using dedicated nodes. Our experiments with CM1 showed that the choice of one approach over the other also depends on the platform. While an approach based on dedicated cores is more suitable on a platform featuring a large

number of cores per node, it may be more efficient to use dedicated nodes on a platform with a reduced number of cores per node.

6. RELATED WORK

In this section, we position our work with respect to related work. We start by discussing approaches that attempt at improving I/O performance. We then examine approaches to in situ visualization.

6.1. Damaris in the "I/O Landscape"

Through its capability of gathering data into larger buffers and files, Damaris can be compared to the data aggregation feature in ROMIO [Thakur et al. 1999a]. This feature is an optimization of Collective I/O that leverages a subset of processes, called "aggregators", to actually perform the I/O on behalf of other processes. Yet, data aggregation is performed synchronously in ROMIO: all processes that do not perform actual writes in the file system must wait for the aggregator processes to complete their operations. Besides, aggregators are not dedicated processes, they run the simulation after completing their I/O. Through dedicated cores, Damaris can perform data aggregation and potential transformations in an asynchronous manner and still use the idle time remaining in the dedicated cores.

Other efforts focus on overlapping computation with I/O in order to reduce the impact of I/O latency on overall performance. Overlap techniques can be implemented directly within simulations [Patrick et al. 2008], using asynchronous communications. Non-blocking I/O primitives started to appear as part of the current MPI 3 standard, these primitives are still implemented as blocking in practice.

Other approaches leverage data-staging and caching mechanisms [Nisar et al. 2008; Isaila et al. 2010], or forwarding approaches [Ali et al. 2009] to achieve better I/O performance. Forwarding architectures run on top of dedicated resources in the platform, which are not configurable by the end-user, that is, the user cannot run custom data processing in forwarding resources. Similarly to the parallel file system, these dedicated resources are shared by all users. This leads to cross-application access contention and thus, to I/O variability. However, the trend toward I/O delegate systems underlines the need for new I/O approaches. Our approach relies on dedicated I/O cores at the application level, or dedicated nodes bound to the application, rather than relying on hardware I/O-dedicated or forwarding nodes, with the advantage of letting users configure their dedicated resources to best fit their needs.

The use of local memory to alleviate the load on the file system is not new. The Scalable Checkpoint/Restart (SRC) by Moody et al. [Moody et al. 2010] already makes use of node-level storage to avoid the heavy load caused by periodic global checkpoints. Yet their work does not use dedicated resources or threads to handle or process data, and the checkpoints are not asynchronous.

Dedicated-Core-Based Approaches. Closest to our work are the approaches by Li et al. [Li et al. 2010], and Ma et al. [Ma et al. 2006]. While the general goals of these approaches are similar (leveraging service-dedicated cores for non-computational tasks), their design is different, and so is the focus and the (much lower) scale of their evaluation. The first one mainly explores the idea of using dedicated cores in conjunction with SSDs to improve the overall I/O throughput. Architecturally, it relies on a FUSE interface, which introduces unnecessary copies through the kernel and reduces the degree of coupling between cores. Using small benchmarks we noticed that such a FUSE interface is about 10 times slower in transferring data between cores than using shared memory. In the second, active buffers are handled by dedicated processes that can run on any node and interact with cores running the simulation through the

network. In contrast to both approaches, Damaris makes a much more efficient design choice using the shared intra-node memory, thereby avoiding costly copies and buffering. The approach defended by Li et al. is demonstrated on a small 32-node cluster (160 cores), where the maximum scale used in the work by Ma et al. is 512 cores on a Power3 machine, for which the overall improvement achieved for the global run time is marginal. Our experimental analysis is much more extensive and more relevant for today's scales of HPC simulations: we demonstrated the excellent scalability of Damaris on a real supercomputer (Kraken, ranked 11th in the Top500 supercomputer list at the time of the experiments) with up to almost 10,000 cores, and with the CM1 tornado simulation, one of the target applications of the Blue Waters post-Petascale supercomputer project. We demonstrated not only a speedup in I/O throughput by a factor of 15 (never achieved by previous approaches), but we also showed that Damaris totally hides the I/O jitter and substantially cuts down the application run time at such high scales. With Damaris, the execution time for CM1 at this scale is even divided by 3.5 compared to approaches based on collective I/O! Moreover, we further explored how to leverage the spare time of the dedicated cores. We demonstrated for example that it can be used to compress data by a factor of 6.

Managing variability through QoS scheduling. While Damaris and the approaches presented above work at the application's side, another way of addressing I/O variability consists of enforcing quality of service levels in storage systems. Such techniques mitigate I/O variability by addressing one of its potential sources: the contention between distinct applications running concurrently on the same platform. While these techniques cannot *hide* the I/O variability, they attempt to maintain it within well-defined bounds.

QoS-based scheduling is used in particular in enterprise storage systems [Gulati et al. 2007; Wachs et al. 2007] and in the field of cloud computing [Pu et al. 2010], where pricing models require performance guaranties. It generally involves isolation techniques and bandwidth allocation to ensure that applications' I/O performance is guaranteed. In high-performance computing, however, the lack of a pricing model has not extensively motivated the implementation of such techniques. Additionally, as seen above, I/O variability already appears within the processes of a single application because of contention, communications, or metadata overhead. QoS-based scheduling is therefore only a secondary solution that mitigates I/O variability provided that the applications have already individually optimized their I/O.

Zhang *et al.* [Zhang et al. 2011] propose to meet QoS requirements set by each application in terms of application run time. The required application run time is converted into bandwidth and latency bounds through machine learning techniques, so that bandwidth can be allocated to each application individually.

6.2. Damaris in the "In Situ Visualization Landscape"

Loosely-Coupled Visualization Strategies. Ellsworth et al. [Ellsworth et al. 2006] propose to use distributed shared memory (DSM) to avoid writing files when performing concurrent visualization. Such an approach has the advantage of decoupling the simulation and visualization processes, but reading data from the memory of the simulation's processors can increase run time variability. The scalability of a distributed shared memory design is also a limiting factor.

Rivi et al. [Rivi et al. 2011] introduce the ICARUS plugin for ParaView together with a description of VisIt and ParaView's in situ visualization interfaces. ICARUS employs an HDF5 DSM file driver to ship data to a distributed shared memory buffer that is used as input to a ParaView pipeline. This DSM stores a view of the HDF5 files that can be concurrently accessed by the simulation and visualization tools. The HDF5 API

allows to bridge the simulation and ParaView with minimum code changes (provided that the simulation already uses HDF5), but it produces multiple copies of the data and a complete transformation of data into an intermediate HDF5 representation. Also, the visualization library on the remote resource requires the original data to conform to this HDF5 representation. Damaris, on the other hand, is not based on any data format and efficiently leverages shared-memory to avoid as much as possible unnecessary copies of data. Besides, its API is simpler than that of HDF5 for simulations that do not already use HDF5.

Malakar et al. [Malakar et al. 2010] present an adaptive framework for loosely-coupled visualization, in which data is sent over a network to a remote visualization cluster at a frequency that is dynamically adapted depending on resource availability. Our approach also adapts output frequency to resource usage.

The PreDatA [Zheng et al. 2010] middleware proposes to dedicate a set of nodes as a staging area to perform a first step of data processing prior to I/O for the purpose of subsequent visualization. The coupling between the simulation and the staging area is done through the ADIOS [Lofstead et al. 2008] I/O layer. The use of the ADIOS back-end allows to decouple the simulation and the visualization by simply integrating data analysis as part of an existing I/O stack [Zheng et al. 2011]. While Damaris borrows the use of an XML file from ADIOS in order to simplify its API, it makes the orthogonal choice of using dedicated cores rather than dedicated nodes. Thus it avoids potentially costly data movements across nodes.

GLEAN [Rasquin et al. 2011] provides in situ visualization capabilities with dedicated nodes. The authors use the PHASTA simulation on the Intrepid supercomputer and ParaView for analysis and visualization on the Eureka machine. Part of the analysis in GLEAN is done in a time-partitioning manner at the simulation side, which makes it a hybrid approach involving tightly- and loosely-coupled in situ analysis. Our approach shares some of the same goals, namely to couple a simulation with run-time visualization, but we run the visualization tool on one core of the same node instead of dedicated nodes. GLEAN is also used in conjunction with ADIOS [Moreland et al. 2011].

EPSN [Esnard et al. 2006] is an environment providing steering and visualization capabilities to existing parallel simulations. Simulations instrumented with EPSN ship their data to a visualization pipeline running on a remote cluster, thus EPSN is an hybrid approach including both code changes and the use of additional remote resources. In contrast to EPSN, all visualization tasks using Damaris can be performed on dedicated cores, closer to the simulation, thus reducing the network overhead.

Zheng et al. [Zheng et al. 2011] have provided a model to evaluate the tradeoff between in situ synchronous visualization and loosely-coupled visualization through staging areas. This model can be applied to compare in situ using dedicated cores instead of remote resources, with the difference being that approaches utilizing dedicated cores do not have network communication overhead.

Tightly-Coupled In Situ Visualization. SciRun [Johnson et al. 1999] is a complete computational-steering environment that includes visualization. Its in situ capabilities can be used with any simulation implemented with SciRun solvers and structures. SciRun is an example of the trend towards integrating visualization, data analysis and computational steering in the simulation process. Simulations are written specifically for use in SciRun in order to exchange data with zero data copy, but adapting an existing application to this framework can be a daunting task.

DIY [Peterka et al. 2011] offers a number of communication primitives allowing to easily build efficient parallel in situ analysis and visualization algorithms. However it does not aim to provide a way to dedicate resources on which to run these algo-

rithms. DIY could therefore very well be coupled with Damaris to implement powerful in situ analysis algorithms while Damaris provides the flexibility of running them on dedicated resources.

Tu et al. [Tu et al. 2006] propose an end-to-end approach for an earthquake simulation using the Hercule framework. All the components of the simulation, including visualization, run in parallel on the same machine, and the only output consists of a set of JPEG files. The data processing tasks in Hercule are still performed in a synchronous manner, and any operation initiated by a process to perform these tasks impacts the performance of the simulation.

In the context of ADIOS, CoDS (Co-located DataSpaces) [Zhang et al. 2012a] builds a distributed object-based data space abstraction and can use dedicated nodes (and recently dedicated cores with shared memory) with PreDatA, DataStager and DataSpace. ADIOS+CoDS has also been used for code coupling [Zhang et al. 2012b] and demonstrated with different simulation models. While the use of dedicated cores to accomplish two different tasks is a common theme in our approach, our objective in this chapter was to compare the performance impact on the simulation of a collocated visualization task with a directly embedded visualization. Besides, placement of data in shared memory in the aforementioned works is done through the ADIOS interface, which creates a copy of data from the simulation to the shared memory using a file-writing interface. We leverage the double-buffering technique usually implemented in simulations as an efficient alternative for sharing data.

Posteriorly to our work, Dreher and Rafin [Dreher et al. 2014b] built on the FlowVR framework (initially proposed for real-time interactive parallel visualization in the context of virtual reality) to provide a solution integrating both time partitioning, dedicated cores and dedicated nodes. They address usability by providing a simple *put/get* interface and a Python script that describes the various component of the visualization pipeline. They went one step further by providing in situ interactive simulation steering in a cave-like system with haptic devices [Dreher et al. 2014a], highlighting a case where the simulation process and research are part of the same workflow.

7. CONCLUSION AND FUTURE DIRECTIONS

As HPC resources exceeding millions of cores become a reality, science and engineering codes invariably must be modified in order to efficiently exploit these resources. An important challenge in maintaining high performance is data management, which nowadays does not only include writing and storing data efficiently, but also analyzing and visualizing these data in order to retrieve a scientific insight.

This paper provides a comprehensive overview of Damaris, an approach which proposes to offload data management tasks, including I/O, post-processing and visualization, into dedicated cores of multicore nodes. Damaris efficiently leverages shared-memory to improve memory usage when transferring data from cores running the simulation to cores running data-related tasks. Thanks to its plugin system and an external description of data, Damaris is highly adaptable to a wide range of simulations.

We first used Damaris to offload I/O tasks in dedicated cores, and compared the resulting performance with the two standard approaches to I/O in HPC simulations: the File-per-process and the Collective I/O approaches. By gathering I/O operations in a reduced number of cores and by avoiding synchronization between these cores, Damaris is able to completely hide all I/O-related costs, and in particular the I/O variability. Our experiments using the CM1 atmospheric simulation and the Nek5000 computation fluid dynamic, in particular on up to 9216 cores of the Kraken supercomputer, showed that Damaris can achieve a 15 times higher throughput compared with the collective I/O approach. Damaris also dramatically reduces the application run time, leading to a $3.5\times$ speedup in CM1, for example. Observing that dedicated cores still remain

idle a large fraction of the time, we implemented several improvements, including an overhead-free data compression that achieved up to 600% compression ratio.

We then leveraged the time spared by Damaris on dedicated cores by extending it to support in situ visualization through a connection with the VisIt visualization software. We evaluated our Damaris-based in situ visualization framework on the Grid'5000 and Blue Waters platforms. We showed that Damaris can fully hide the performance variability induced by in situ visualization tasks as well, even in scenarios involving interactions with a user. Besides, Damaris reduces visualization-related code modifications to a minimum in existing simulations.

Finally we further extended Damaris to support the use of dedicated nodes instead of dedicated cores. Based on our framework, we performed a thorough comparison of the dedicated cores, dedicated nodes and time-partitioning approaches for I/O on 3 different clusters of the Grid'5000 testbed, with the CM1 and Nek5000 simulations. Our evaluation shows that approaches based on dedicated resources always perform better than the time-partitioning approach for the selected simulations. They both manage to hide the I/O-related costs and, as a result, improve the overall simulation performance. While the choice of an approach based on dedicated cores over an approach based on dedicated nodes is primarily driven by the number of cores per node available in the platform, this choice also depends on the scalability of the application, its memory usage, and the potential use of spare time in dedicated resources.

To our knowledge, Damaris is the first middleware available to the community[7] that offers the use of dedicated cores or dedicated nodes to serve data management tasks ranging from I/O to in situ visualization. This work paves the way for a number of new research directions with high potential impact. Our study of in situ visualization using Damaris and CM1 revealed that in some simulations such as climate models, an important fraction of the data produced by the simulation does not actually contain any part of the phenomenon that are of interest to scientists. When visualizing this data in situ, it thus becomes possible to lower the resolution of non-interesting parts in order to increase the performance of the visualization process, an approach that we call "smart in situ visualization". Challenges to implement smart in situ visualization include automatically discriminating relevant and non-relevant data within the simulation while this data is being produced. This detection should be made without user intervention and be fast enough to not diminish the overall performance of the visualization process. The plugin system of Damaris together with its existing connection with the VisIt visualization software provide an excellent ground to implement and evaluate smart in situ visualization.

We also plan to investigate ways to reduce the energy consumption of simulations that use approaches like Damaris. We have already shown that the time spared by dedicated cores in Damaris can be leveraged to compress the data prior to storing it. An immediate question that can be asked is to which extent does compression in Damaris impacts this energy/performance tradeoff. On one hand, compression reduces the amount of data transferred and thus, the network traffic, which leads to lower energy consumption from data movements. On the other hand, compressing data requires more computation time and higher energy consumption as a result of data movement in the local memory hierarchy. Consequently, a promising direction will consist in investigating the tradeoff between energy, performance and compression level. We will also investigate how to use the Damaris approach in the context of out-of-core computation. This technique, usually meant for simulations whose data do not fit in memory, poses new challenges for Damaris to efficiently prefetch data from storage and monitor its memory usage.

[7]See http://damaris.gforge.inria.fr

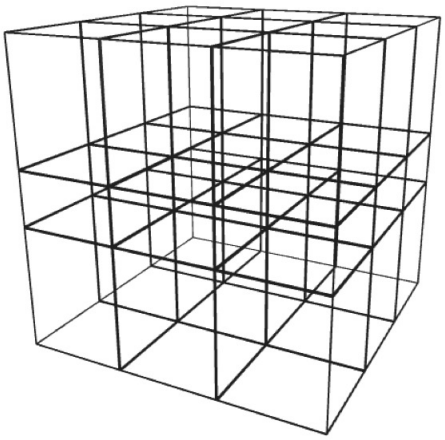


Fig. 20: Example of a $4 \times 4 \times 4$ rectilinear grid described by three arrays of coordinates. In this example there is a scalar value (such as *temperature* or *wind velocity*) at each node. The mesh itself is described through three coordinate arrays: `mesh_x = {0.0,1.0,2.0,3.0}`; `mesh_y = {0.0,1.0,2.0,3.0}`; `mesh_z = {0.0,1.2,1.8,3.0}`.

Acknowledgments

This work was done in the framework of a collaboration between the KerData (Inria Rennes Bretagne Atlantique, ENS Rennes, INSA Rennes, IRISA) team, the National Center for Supercomputing Applications (Urbana-Champaign, USA) and Argonne National Laboratory, within the Joint Inria-UIUC-ANL-BSC-JSC Laboratory for Extreme-Scale Computing (JLESC), formerly Joint Laboratory for Petascale Computing (JLPC). The work was supported by the U.S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357, by the National Center for Atmospheric Research (NCAR) and Central Michigan University. Some experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). The authors also thank Robert Wilhelmson for his insight on CM1, Dave Semeraro for the discussions regarding in situ visualization using VisIt, Paul Fischer and Aleksandr Obabko for helping understand Nek5000 and providing input datasets.

A. CODE SAMPLE USING DAMARIS

Listing 1 is an example of a Fortran program that makes use of Damaris. It writes three 1D arrays representing the coordinates of a rectilinear mesh. At every iteration it then writes a 3D array representing temperature values on the points of the mesh and sends an event to the dedicated core. Line 7 initializes Damaris using a configuration file. Line 8 starts the servers on dedicated resources. From line 10 to 29, the client code, that is, the simulation's main loop, is executed. This main loop include calls to `damaris_write` whenever data has to be transmitted to the servers, and calls to `damaris_signal` whenever a plugin should be called. The `damaris_end_iteration` function is used to notify the servers that an iteration of the simulation has completed, leading the servers to take appropriate decisions such as purging the memory from old data or updating in situ visualization backends. Line 28 is executed by all clients to stop the servers on dedicated resources after leaving the main loop. Damaris is finalized line 32, cleaning up resources such as shared memory and communication channels.

The associated configuration file, shown in Listing 2, describes the data that is expected to be received by the servers, and the action to perform upon reception of the

```
program example
  integer ierr, is_client
  real, dimension(64,16,2) :: temperature
  real, dimension(4) :: x3d, y3d, z3d

  ! initialization
  call damaris_initialize_f("config.xml", MPI_COMM_WORLD, ierr)
  call damaris_start_f(is_client, ierr)

  if(is_client.eq.1) then

        ! writing non-time-varying data
    call damaris_write_f("coordinates/x3d", x3d, ierr)
    call damaris_write_f("coordinates/y3d", y3d, ierr)
    call damaris_write_f("coordinates/z3d", z3d, ierr)

    do while(...) ! simulation main loop
      ...
          ! writing temperature data
      call damaris_write_f("temperature", temperature, ierr)
          ! sending signal
      call damaris_signal_f("my_event", ierr)
          ! end of iteration
      call damaris_end_iteration_f(ierr)
      ...
    enddo
    ! stopping the servers
    call damaris_stop_f(ierr)
  endif

  ! finalization
  call damaris_finalize_f(ierr)
end program example
```

Listing 1: Example of Fortran simulation using Damaris.

event. More specifically, lines 14, 15, 16 and 18 of this XML file define *layouts*, which describe the type and dimensions of a piece of data. Lines 26 to 33 define a group, and within this group a set of variables that use these layouts. The temperature variable is defined in line 35. Finally line 38 associates an event with a function (or *action*) to be called when the event is received. It also locates the function within a dynamically-loaded library.

The configuration file also contains information for visualization software. Lines 20 to 24 in the XML file correspond to mesh structure drawn in Figure 20, and built from the three coordinate variables. The temperature variable is mapped onto this mesh using its *mesh* attribute.

REFERENCES

Hasan Abbasi, Matthew Wolf, Greg Eisenhauer, Scott Klasky, Karsten Schwan, and Fang Zheng. 2009. DataStager: Scalable Data Staging Services for Petascale Applications. In *Proceedings of the 18th ACM International Symposium on High Performance Distributed Computing (HPDC '09)*. ACM, New York, NY, USA, 39–48. DOI:http://dx.doi.org/10.1145/1551609.1551618

Nawab Ali, Philip Carns, Kamil Iskra, Dries Kimpe, Samuel Lang, Robert Latham, Robert Ross, Lee Ward, and Ponnuswamy Sadayappan. 2009. Scalable I/O Forwarding Framework for High-Performance Computing Systems. In *Proceedings of the IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER '09*. DOI:http://dx.doi.org/10.1109/CLUSTR.2009.5289188

ANL. 2015. Mpich. http://www.mpich.org. (2015).

George H. Bryan and J. Michael Fritsch. 2002. A Benchmark Simulation for Moist Nonhydrostatic Numerical Models. *Monthly Weather Review* 130, 12 (2002), 2917–2928. DOI:http://dx.doi.org/10.1175/1520-0493(2002)130⟨2917:ABSFMN⟩2.0.CO;2

```
<simulation name="my_simulation" language="c"
    xmlns="http://damaris.gforge.inria.fr/damaris/model">
    <architecture>
      <domains count="1"/>
      <dedicated cores="1"/>
      <buffer name="the_buffer" size="67108864" />
      <queue name="the_queue" size="100" />
    </architecture>
    <data>
        <parameter name="w" type="int" value="4" />
        <parameter name="h" type="int" value="4" />
        <parameter name="d" type="int" value="4" />

        <layout name="mesh_x_layout" type="float" dimensions="w" />
        <layout name="mesh_y_layout" type="float" dimensions="h" />
        <layout name="mesh_z_layout" type="float" dimensions="d" />

        <layout name="data_layout" type="double" dimensions="w,h,d"/>

        <mesh name="mesh3d" type="rectilinear" topology="3">
            <coord name="coordinates/x3d" unit="m" label="Width"/>
            <coord name="coordinates/y3d" unit="m" label="Height"/>
            <coord name="coordinates/z3d" unit="m" label="Depth"/>
        </mesh>

        <group name="coordinates">
            <variable name="x3d" layout="mesh_x_layout"
                visualizable="false" time-varying="false" />
            <variable name="y3d" layout="mesh_y_layout"
                visualizable="false" time-varying="false" />
            <variable name="z3d" layout="mesh_z_layout"
                visualizable="false" time-varying="false" />
        </group>

        <variable name="temperature" layout="data_layout" mesh="mesh3d"/>
    </data>
    <actions>
      <event name="my_event" action="my_function" using="my_plugin.so" />
    </actions>
</simulation>
```

Listing 2: Configuration file associated with the Fortran example.

Philip H. Carns, Walter B. Ligon, III, Robert B. Ross, and Rajeev Thakur. 2000. PVFS: a Parallel File System for Linux Clusters. In *Proceedings of the 4th annual Linux Showcase & Conference - Volume 4*. USENIX Association, Berkeley, CA, USA, 1.

Christian M Chilan, M Yang, Albert Cheng, and Leon Arber. 2006. Parallel I/O Performance Study with HDF5, a Scientific Data Package. *TeraGrid 2006: Advancing Scientific Discovery* (2006).

H. Childs, D. Pugmire, S. Ahern, B. Whitlock, M. Howison, and others. 2010. Extreme Scaling of Production Visualization Software on Diverse Architectures. *IEEE Computer Graphics and Applications* (2010), 22–31.

Ciprian Docan, Manish Parashar, and Scott Klasky. 2010. Enabling High-Speed Asynchronous Data Extraction and Transfer Using DART. *Concurrency and Computation: Practice and Experience* (2010), 1181–1204. DOI:http://dx.doi.org/10.1002/cpe.1567

Stephanie Donovan, Gerrit Huizenga, Andrew J. Hutton, C. Craig Ross, Martin K. Petersen, and Philip Schwan. 2003. Lustre: Building a File System for 1000-node Clusters. In *Proceedings of the 2003 Linux Symposium*. Citeseer.

Matthieu Dorier, Gabriel Antoniu, Franck Cappello, Marc Snir, and Leigh Orf. 2012a. Damaris: How to Efficiently Leverage Multicore Parallelism to Achieve Scalable, Jitter-free I/O. In *Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER '12)*. IEEE, Beijing, China. http://hal.inria.fr/hal-00715252

Matthieu Dorier, Gabriel Antoniu, Franck Cappello, Marc Snir, and Leigh Orf. 2012b. *Damaris: Leveraging Multicore Parallelism to Mask I/O Jitter*. Research Report RR-7706. INRIA. 36 pages. http://hal.inria.fr/inria-00614597

Matthieu Dorier, Gabriel Antoniu, Robert Ross, Dries Kimpe, and Shadi Ibrahim. 2014. CALCioM: Mitigating I/O Interference in HPC Systems through Cross-Application Coordination. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS '14)*. Phoenix, Arizona, USA. http://hal.inria.fr/hal-00916091

Matthieu Dorier, R. Sisneros, Roberto, Tom Peterka, Gabriel Antoniu, and B. Semeraro, Dave. 2013. Damaris/Viz: a Nonintrusive, Adaptable and User-Friendly In Situ Visualization Framework. In *Proceedings of the IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV '13)*. Atlanta, Georgia, USA. http://hal.inria.fr/hal-00859603

Matthieu Dreher, Jessica Prevoteau-Jonquet, Mikael Trellet, Marc Piuzzi, Marc Baaden, Bruno Raffin, Nicolas Férey, Sophie Robert, and Sébastien Limet. 2014a. Exaviz: A Flexible Framework to Analyse, Steer and Interact with Molecular Dynamics Simulations. *Faraday Discussions* (2014).

Matthieu Dreher, Bruno Raffin, and others. 2014b. A Flexible Framework for Asynchronous In Situ and In Transit Analytics for Scientific Simulations. *ACM/IEEE International Symposium on Cluster, Cloud and Grid Computing (CCGrid '14)* (2014).

D. Ellsworth, B. Green, C. Henze, P. Moran, and T. Sandstrom. 2006. Concurrent Visualization in a Production Supercomputing Environment. *IEEE Transactions on Visualization and Computer Graphics (TVGC)* 12, 5 (Sept.-Oct. 2006), 997–1004. DOI:http://dx.doi.org/10.1109/TVCG.2006.128

ERDC DSRC. 2015. EzViz. http://daac.hpc.mil/software/ezViz/. (2015).

A. Esnard, N. Richart, and O. Coulaud. 2006. A Steering Environment for Online Parallel Visualization of Legacy Parallel Simulations. In *Proceedings of the IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications (DS-RT '06)*. IEEE, 7–14.

N. Fabian, K. Moreland, D. Thompson, A.C. Bauer, P. Marion, B. Geveci, M. Rasquin, and K.E. Jansen. 2011. The ParaView Coprocessing Library: A Scalable, General Purpose In Situ Visualization Library. In *Proceedings of the IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV '11)*.

Mike Folk, Albert Cheng, and Kim Yates. 1999. HDF5: A File Format and I/O Library for High Performance Computing Applications. In *Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC '99)*.

Jing Fu, Robert Latham, Misun Min, and Christopher D. Carothers. 2012. I/O Threads to Reduce Checkpoint Blocking for an Electromagnetics Solver on Blue Gene/P and Cray XK6. In *Proceedings of the International workshop on Runtime and Operating Systems for Supercomputers (ROSS '12)*.

Ana Gainaru, Guillaume Aupy, Anne Benoit, Franck Cappello, Yves Robert, and Marc Snir. 2014. *Scheduling the I/O of HPC Applications under Congestion*. Rapport de recherche RR-8519. INRIA. http://hal.inria.fr/hal-00983789

Grid'5000. 2015. Inria testbed. http://www.grid5000.fr. (2015).

Ajay Gulati, Arif Merchant, and Peter J. Varman. 2007. pClock: An Arrival Curve Based Approach for QoS Guarantees in Shared Storage Systems. In *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '07)*. ACM, New York, NY, USA, 13–24. DOI:http://dx.doi.org/10.1145/1254882.1254885

HDF5. 2015. Hierarchical Data Format. http://www.hdfgroup.org/HDF5/. (2015).

Mark Hereld, Michael E. Papka, and V. Vishwanath. 2011. Toward Simulation-Time Data Analysis and I/O Acceleration on Leadership-Class Systems. In *Proceeding of the IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV '11)*. Providence, RI.

Florin Isaila, Javier Garcia Blas, Jesus Carretero, Robert Latham, and Robert Ross. 2010. Design and Evaluation of Multiple Level Data Staging for Blue Gene Systems. *IEEE Transactions on Parallel and Distributed Systems (TPDS)* (2010). DOI:http://dx.doi.org/10.1109/TPDS.2010.127

C. Johnson, S.G. Parker, C. Hansen, G.L. Kindlmann, and Y. Livnat. 1999. Interactive Simulation and Visualization. *Computer* 32, 12 (1999), 59–65.

Donald B Johnston. 2014. First-of-a-kind supercomputer at Lawrence Livermore available for collaborative research. https://www.llnl.gov/news/newsreleases/2014/May/NR-14-05-02.html. (2014).

KitWare. 2015a. eXtensible Data Model and Format (XDMF). http://www.xdmf.org/. (2015).

KitWare. 2015b. ParaView. http://www.paraview.org/. (2015).

M. Li, S.S. Vazhkudai, A.R. Butt, F. Meng, X. Ma, Y. Kim, C. Engelmann, and G. Shipman. 2010. Functional Partitioning to Optimize End-to-End Performance on Many-Core Architectures. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC '10)*. IEEE Computer Society.

Ning Liu, Jason Cope, Philip Carns, Christopher Carothers, Robert Ross, Gary Grider, Adam Crume, and Carlos Maltzahn. 2012. On the Role of Burst Buffers in Leadership-Class Storage Systems. In *Proceedings of the 28th IEEE Symposium on Mass Storage Systems and Technologies (MSST '12)*. IEEE.

LLNL. 2015. VisIt, Lawrence Livermore National Laboratory. https://wci.llnl.gov/simulation/computer-codes/visit. (2015).

Jay Lofstead, Fang Zheng, Qing Liu, Scott Klasky, Ron Oldfield, Todd Kordenbrock, Karsten Schwan, and Matthew Wolf. 2010. Managing Variability in the IO Performance of Petascale Storage Systems. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC '10)*. IEEE Computer Society, Washington, DC, USA, 12. DOI:http://dx.doi.org/10.1109/SC.2010.32

Jay F. Lofstead, Scott Klasky, Karsten Schwan, Norbert Podhorszki, and Chen Jin. 2008. Flexible IO and integration for scientific codes through the adaptable IO system (ADIOS). In *Proceedings of the 6th international workshop on Challenges of large applications in distributed environments (CLADE '08)*. ACM, New York, NY, USA. DOI:http://dx.doi.org/10.1145/1383529.1383533

Kwan-Liu Ma. 2009. In Situ Visualization at Extreme Scale: Challenges and Opportunities. *IEEE Computer Graphics and Applications* 29, 6 (Nov.-Dec. 2009), 14–19. DOI:http://dx.doi.org/10.1109/MCG.2009.120

Kwan-Liu Ma, Chaoli Wang, Hongfeng Yu, and Anna Tikhonova. 2007. In-situ Processing and Visualization for Ultrascale Simulations. *Journal of Physics: Conference Series* 78, 1 (2007). http://stacks.iop.org/1742-6596/78/i=1/a=012043

Xiaosong Ma, Jonghyun Lee, and Marianne Winslett. 2006. High-Level Buffering for Hiding Periodic Output Cost in Scientific Simulations. *IEEE Transactions on Parallel and Distributed Systems (TPDS)* 17 (2006), 193–204. DOI:http://dx.doi.org/10.1109/TPDS.2006.36

Preeti Malakar, Vijay Natarajan, and Sathish S. Vadhiyar. 2010. An Adaptive Framework for Simulation and Online Remote Visualization of Critical Climate Applications in Resource-constrained Environments. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC '10)*. IEEE Computer Society, Washington, DC, USA, 11. DOI:http://dx.doi.org/10.1109/SC.2010.10

Adam Moody, Greg Bronevetsky, Kathryn Mohror, and Bronis R. de Supinski. 2010. Design, Modeling, and Evaluation of a Scalable Multi-Level Checkpointing System. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC '10)*. IEEE Computer Society, Los Alamitos, CA, USA. DOI:http://dx.doi.org/10.1109/SC.2010.18

K. Moreland, R. Oldfield, P. Marion, S. Jourdain, N. Podhorszki, V. Vishwanath, N. Fabian, C. Docan, M. Parashar, M. Hereld, and others. 2011. Examples of In Transit Visualization. In *Proceedings of the 2nd International Workshop on Petascale Data Analytics: Challenges and Opportunities (PDAC '11)*. ACM.

NCSA. 2015. Blue Waters supercomputer, National Center for Supercomputing Applications. http://www.ncsa.illinois.edu/BlueWaters/. (2015).

NICS. 2015. Kraken supercomputer, National Institute for Computational Sciences. http://www.nics.tennessee.edu/computing-resources/kraken. (2015).

A. Nisar, Wei keng Liao, and A. Choudhary. 2008. Scaling Parallel I/O Performance Through I/O Delegate and Caching System. In *Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC '08)*. DOI:http://dx.doi.org/10.1109/SC.2008.5214358

James W. Lottes P. F. Fischer and Stefan G. Kerkemeier. 2008. Nek5000 Web page http://nek5000.mcs.anl.gov. (2008).

Christina M. Patrick, Seung Woo Son, and Mahmut Kandemir. 2008. Comparative Evaluation of Overlap Strategies with Study of I/O Overlap in MPI-IO. *Operating Systems Review (SIGOPS)* 42 (Oct. 2008), 43–49. Issue 6. DOI:http://dx.doi.org/10.1145/1453775.1453784

Tom Peterka, Robert Ross, Wesley Kendall, Attila Gyulassy, Valerio Pascucci, Han-Wei Shen, Teng-Yok Lee, and Abon Chaudhuri. 2011. Scalable Parallel Building Blocks for Custom Data Analysis. In *Proceedings of Large Data Analysis and Visualization Symposium LDAV'11*. Providence, RI.

R. Prabhakar, S.S. Vazhkudai, Y. Kim, A.R. Butt, M. Li, and M. Kandemir. 2011. Provisioning a Multi-tiered Data Staging Area for Extreme-Scale Machines. In *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS '11)*. DOI:http://dx.doi.org/10.1109/ICDCS.2011.33

Jean-Pierre Prost, Richard Treumann, Richard Hedges, Bin Jia, and Alice Koniges. 2001. MPI-IO/GPFS an Optimized Implementation of MPI-IO on Top of GPFS. In *Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC '01)*. IEEE Computer Society, Los Alamitos, CA, USA. DOI:http://dx.doi.org/10.1145/582034.582051

Xing Pu, Ling Liu, Yiduo Mei, S. Sivathanu, Younggyun Koh, and C. Pu. 2010. Understanding Performance Interference of I/O Workload in Virtualized Cloud Environments. In *Proceedings of the IEEE International Conference on Cloud Computing (Cloud '10)*. 51–58. DOI:http://dx.doi.org/10.1109/CLOUD.2010.65

Michel Rasquin, Patrick Marion, Venkatram Vishwanath, Benjamin Matthews, Mark Hereld, Kenneth Jansen, Raymond Loy, Andrew Bauer, Min Zhou, Onkar Sahni, and others. 2011. Electronic Poster:

Co-Visualization of Full Data and In Situ Data Extracts from Unstructured Grid CFD at 160k Cores. In *ACM/IEEE SC Companion*. ACM, 103–104.

Marzia Rivi, Luigi Calori, Giuseppa Muscianisi, and Vladimir Slavnic. 2011. In-Situ Visualization: State-of-the-Art and Some Use Cases. *PRACE White Paper (2012), http://www.prace-ri.eu/Visualisation* (2011).

Rutgers. 2015. DataSpace. www.dataspaces.org/. (2015).

Douglas C. Schmidt. 1995. Reactor - An Object Behavioral Pattern for Demultiplexing and Dispatching Handles for Synchronous Events. (1995).

W.J. Schroeder, L.S. Avila, and W. Hoffman. 2000. Visualizing with VTK: a Tutorial. *IEEE Computer Graphics and Applications* 20, 5 (Sep.,Oct. 2000), 20–27. DOI:http://dx.doi.org/10.1109/38.865875

H. Shan and J. Shalf. 2007. Using IOR to Analyze the I/O Performance for HPC Platforms. In *Proceedings of the Cray User Group Conference (CUG '07)*. Seattle, Washington, USA.

D. Skinner and W. Kramer. 2005. Understanding the Causes of Performance Variability in HPC Workloads. In *Proceedings of the IEEE Workload Characterization Symposium (IISWC '05)*. IEEE Computer Society, 137–149. DOI:http://dx.doi.org/10.1109/IISWC.2005.1526010

N. T. B. Stone, D. Balog, B. Gill, B. Johanson, J. Marsteller, P. Nowoczynski, D. Porter, R. Reddy, J. R. Scott, D. Simmel, J. Sommerfield, K. Vargo, and C. Vizino. 2006. PDIO: High-Performance Remote File I/O for Portals Enabled Compute Nodes. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA '06)*. http://www.scientificcommons.org/43489982

Rajeev Thakur, William Gropp, and Ewing Lusk. 1999a. Data Sieving and Collective I/O in ROMIO. *Symposium on the Frontiers of Massively Parallel Processing* (1999), 182. DOI:http://dx.doi.org/10.1109/FMPC.1999.750599

Rajeev Thakur, William Gropp, and Ewing Lusk. 1999b. On Implementing MPI-IO Portably and with High Performance. In *Proceedings of the sixth Workshop on I/O in Parallel and Distributed Systems (IOPADS '99)*. ACM, 23–32.

Top500. 2015. Top500 List of Supercomputers. http://www.top500.org/. (2015).

Tiankai Tu, Hongfeng Yu, Leonardo Ramirez-Guzman, Jacobo Bielak, Omar Ghattas, Kwan-Liu Ma, and David R. O'Hallaron. 2006. From Mesh Generation to Scientific Visualization: an End-to-End Approach to Parallel Supercomputing. In *Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC '06)*. ACM, New York, NY, USA, Article 91. DOI:http://dx.doi.org/10.1145/1188455.1188551

Unidata. 2015. NekCDF. http://www.unidata.ucar.edu/software/netcdf/. (2015).

A. Uselton, M. Howison, N.J. Wright, D. Skinner, N. Keen, J. Shalf, K.L. Karavanic, and L. Oliker. 2010. Parallel I/O Performance: From Events to Ensembles. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS '10)*. DOI:http://dx.doi.org/10.1109/IPDPS.2010.5470424

Matthew Wachs, Michael Abd-El-Malek, Eno Thereska, and Gregory R. Ganger. 2007. Argon: Performance Insulation for Shared Storage Servers. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07)*. USENIX Association, Berkeley, CA, USA, 5–5. http://dl.acm.org/citation.cfm?id=1267903.1267908

Brad Whitlock, Jean M. Favre, and Jeremy S. Meredith. 2011. Parallel In Situ Coupling of Simulation with a Fully Featured Visualization System. In *Proceedings of the Eurographics Symposium on Parallel Graphics and Visualization (EGPGV '10)*. Eurographics Association.

H. Yu and K.L. Ma. 2005. A Study of I/O Techniques for Parallel Visualization. *Journal of Parallel Computing* 31, 2 (2005).

Hongfeng Yu, Chaoli Wang, R.W. Grout, J.H. Chen, and Kwan-Liu Ma. 2010. In Situ Visualization for Large-Scale Combustion Simulations. *IEEE Computer Graphics and Applications* 30, 3 (May-June 2010), 45–57. DOI:http://dx.doi.org/10.1109/MCG.2010.55

Fan Zhang, Solomon Lasluisa, Tong Jin, Ivan Rodero, Hoang Bui, and Manish Parashar. 2012a. In-situ Feature-Based Objects Tracking for Large-Scale Scientific Simulations. In *ACM/IEEE SC Companion*. IEEE.

Fan Zhang, Manish Parashar, Ciprian Docan, Scott Klasky, Norbert Podhorszki, and Hasan Abbasi. 2012b. Enabling In-situ Execution of Coupled Scientific Workflow on Multi-core Platform. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS '12)*. IEEE.

Xuechen Zhang, Kei Davis, and Song Jiang. 2011. QoS Support for End Users of I/O-Intensive Applications using Shared Storage Systems. In *Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC '11)*.

Fang Zheng, Hasan Abbasi, Jianting Cao, Jai Dayal, Karsten Schwan, Matthew Wolf, Scott Klasky, and Norbert Podhorszki. 2011. In-situ I/O Processing: A Case for Location Flexibility. In *Proceed-*

ings of the Sixth Workshop on Parallel Data Storage (PDSW '11). ACM, New York, NY, USA, 37–42. DOI:http://dx.doi.org/10.1145/2159352.2159362

Fang Zheng, H. Abbasi, C. Docan, J. Lofstead, Qing Liu, S. Klasky, M. Parashar, N. Podhorszki, K. Schwan, and M. Wolf. 2010. PreDatA – Preparatory Data Analytics on Peta-Scale Machines. In *Proceedings of the IEEE International Symposium on Parallel Distributed Processing (IPDPS '10)*. DOI:http://dx.doi.org/10.1109/IPDPS.2010.5470454

Fang Zheng, Jianting Cao, Jai Dayal, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, Hasan Abbasi, Scott Klasky, and Norbert Podhorszki. 2011. High End Scientific Codes with Computational I/O Pipelines: Improving their End-to-End Performance. In *Proceedings of the 2nd International Workshop on Petascal Data Analytics: Challenges and Opportunities (PDAC '11)*. ACM, New York, NY, USA, 23–28. DOI:http://dx.doi.org/10.1145/2110205.2110210