This is a post-peer-review, pre-copyedit version of an article published in *Journal of computer science and technology*. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11390-020-9801-1.

Ad hoc file systems for HPC

André Brinkmann[1], Kathryn Mohror[2], Weikuan Yu[3], Philip Carns[4], Toni Cortes[5], Scott A. Klasky[6], Alberto Miranda[7], Franz-Josef Pfreundt[8], Robert B. Ross[4], and Marc-André Vef[1]

[1] *Zentrum für Datenverarbeitung, Johannes Gutenberg University Mainz, 55128 Mainz, Germany*

[2] *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA*

[3] *Department of Computer Science, Florida State University, Tallahassee, FL 32306, USA*

[4] *Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*

[5] *Department of Computer Architecture, Universitat Politecnica de Catalunya, Barcelona 08034, Spain*

[6] *Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

[7] *Computer Science Department, Barcelona Supercomputing Center, Barcelona 08034, Spain*

[8] *Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern 67663, Germany*

E-mail: brinkman@uni-mainz.de; mohror1@llnl.gov; yuw@cs.fsu.edu; carns@mcs.anl.gov; toni.cortes@bsc.es; klasky@ornl.gov; alberto.miranda@bsc.es; pfreundt@itwm.fhg.de; rross@mcs.anl.gov; vef@uni-mainz.de

Abstract
Storage backends of parallel compute clusters are still based mostly on magnetic disks, while newer and faster storage technologies such as flash-based SSDs or non-volatile random access memory (NVRAM) are deployed within compute nodes. Including these new storage technologies into scientific workflows is unfortunately today a mostly manual task, and most scientists therefore do not take advantage of the faster storage media. One approach to systematically include node-local SSDs or NVRAMs into scientific workflows is to deploy ad hoc file systems over a set of compute nodes, which serve as temporary storage systems for single applications or longer-running campaigns.
This paper presents results from the Dagstuhl Seminar 17202 "Challenges and Opportunities of User-Level File Systems for HPC" and discusses application scenarios as well as design strategies for ad hoc file systems using node-local storage media. The discussion includes open research questions, such as how to couple ad hoc file systems with the batch scheduling environment and how to schedule stage-in and stage-out processes of data between the storage backend and the ad hoc file systems. Also presented are strategies to build ad hoc file systems by using reusable components for networking and how to improve storage device compatibility. Various interfaces and semantics are presented, for example those used by the three ad hoc file systems BeeOND, GekkoFS, and BurstFS. Their presentation covers a range from file systems running in production to cutting-edge research focusing on reaching the performance limits of the underlying devices.

Keywords Parallel Architectures, Distributed File Systems, High-Performance Computing, Burst Buffers, POSIX

1 Introduction

For decades, magnetic disks have served as the storage backbone of high-performance computing (HPC) clusters, and their physical properties have significantly influenced the design of parallel file systems. Magnetic disks are composed of rotating platters and a mechanical arm positioning the read and write heads on the platter tracks. The bandwidth of magnetic disks is limited by the disk rotation speed and access latency (the seek time of the disk arm) [67]. Sequential accesses to magnetic disks can easily achieve sustained transfer rates of more than 200 MByte/s, and thus thousands of disks accessed in parallel can read or write at speeds of more than one TByte/s.

In contrast to the high throughput achievable with sequential accesses to magnetic disks, random accesses to a single disk reduce disk transfer rates to less than one percent of peak performance. Random access patterns executed by a single application on a storage sys-

tem can easily lead to a denial of service to the complete HPC system [62]. Applications that perform huge amounts of random accesses can be highly problematic for HPC clusters and parallel file systems that are optimized for sequential read and write access patterns, which occur, for example, during the stage-in of data at application start-up or during checkpointing [63].

The mechanical composition of magnetic disks also leads to a high failure rate compared with that of pure semiconductor components. As a result, magnetic disks are used mostly for maintenance-friendly, dedicated storage clusters that are physically separated from the compute nodes. In this configuration, failing disks can easily be replaced when compared with configurations with disks attached to compute nodes, where there is generally no easy manual disk access. Unfortunately, the dedicated storage cluster configuration hinders the scalability of the storage backend with respect to an increased compute node count.

The challenges and limitations associated with magnetic disks have led to the introduction of new storage technologies into HPC systems, including non-volatile random-access memory (NVRAM) devices such as flash-based solid state drives (SSDs), as new storage tiers in addition to the parallel file system (PFS). SSDs are fully semiconductor-based and can benefit from decreasing process technologies. Today, a single non-volatile memory express (NVMe) SSD can deliver an order of magnitude higher bandwidth than that of a magnetic disk with little difference in the performance between random and sequential access patterns. Additionally, the reliability of SSDs is higher than that of magnetic disks. The replacement rates of SSDs in the field are much smaller than of magnetic disks [68] and depend mostly on the amount of data written to the SSD [49]. Different bit error rate characteristics during the lifetime of an SSD also enable administrators to identify the likelihood that a device will fail and therefore to proactively exchange it [54].

	MOGON II	Summit	MareNostrum 4	Theta
Node capacity (in GB)	400	1,600	240	128
Node bandwidth (in MB/s)	500	6,000	320	$\geq$ 2,150
Node count	1,868	4,608	3,456	4,392
Cluster capacity (in TB)	$\approx$ 747	7,300	$\approx$ 829	562
Cluster bandwidth (in GB/s)	934	$\approx$ 27,648	1,106	$\geq$ 9,307

Table 1. SSD usage in different sized cluster environments.

The use of SSDs is already widespread in most new HPC systems and they can be used as metadata storage for parallel file systems [84], in-system burst buffers [46], and node-local storage. The bandwidth of node-local SSDs typically exceeds the peak bandwidth of the attached parallel file system, while the maximum number of I/O operations (IOPS) can even be more than 10,000× higher than that of the parallel file system (see Table 1).

Several efforts have been made to explore best methods for using node-local SSDs, including as an additional caching layer for parallel file systems [61] or as file systems run over node-local SSDs, with a larger number of these file systems being implemented at the user level [78] [81].

This paper focuses on *ad hoc file systems* that aim to efficiently use temporarily available storage on either node-local SSDs or global burst buffers. While existing parallel file systems can be in principle built on top of node-local SSDs or global burst buffers, it is important to adhere to the following definitions to be useful in the context of ad hoc file systems for HPC.

- Ad hoc file systems can be deployed on HPC clusters for lifetimes as small as the runtime of a sin-

gle job to use node-local SSDs or external burst buffers. It is therefore important that the deployment overhead of the ad hoc file systems is low compared with the application runtime.

- Ad hoc file systems provide a global namespace for all nodes being linked to the ad hoc file system, while the semantics of the file system can be optimized for the application scenario, enabling optimized parallel access schemes.
- Ad hoc file systems interact with the backend storage system using data staging, and most ad hoc file systems have been and will be implemented completely in user space, while they also may contain a kernel component.

User-level file systems are an attractive implementation option for ad hoc file systems because it is relatively easy to swap in new, specialized implementations for use by applications on a case-by-case basis, as opposed to the current mainstream approach of using general-purpose, system-level file systems, which may not be optimized for specific HPC workloads and must be installed by administrators. In contrast, user-level file systems can be tailored for specific HPC workloads for high performance and can be used by applications without administrator intervention.

Although the benefits of hierarchical storage have been adequately demonstrated, critical questions remain for supporting hierarchical storage systems including ad hoc file systems:

- How should we manage data movement through a storage hierarchy for best performance and resilience of data?
- Are user-level file systems fast enough to be used in HPC systems, and how should we present hierarchical storage systems to user applications, such that they are easy to use and that application code is portable across systems?
- How do the particular I/O use cases mandate the way we manage data?
- Is it possible to reuse components like building blocks when designing user-level file systems?

The Dagstuhl Seminar 17202 "Challenges and Opportunities of User-Level File Systems for HPC" brought together experts in I/O performance, file systems, and storage and collectively explored the space of current and future problems and solutions for I/O on hierarchical storage systems [8].

This paper provides a summary of the Dagstuhl seminar and puts the results into the context of the related work. The paper therefore starts in Section 2 with use cases and continues in Section 3 with a presentation of reusable components as basic building blocks of ad hoc file systems that can simplify the implememtation of ad hoc file systems. Section 4 analyzes different implementation choices to present the ad hoc file system to the client application. Section 5 discusses existing ad hoc file systems and their approaches. Ad hoc file systems also require changes to the HPC infrastructure. It is for example in many cases necessary to synchronize data between the backend file system and the ad hoc file system. This staging process is discussed in Section 6. Section 7 provides a summary of our conclusions.

2 Use cases for ad hoc file systems

While parallel file systems such as Lustre, GPFS, or BeeGFS have already been serving as reliable backbones for HPC clusters for more than two decades, a need for changes in the HPC storage architecture arose with the arrival of *data-intensive applications* in HPC. These applications shift the bottleneck from being com-

pute intensive, and thus being restricted by the performance of the CPUs, to being bound by the quantity of data, its complexity, and the speed at which it changes [38].

Node-local SSDs and burst buffers have been introduced into the HPC storage hierarchy to support the new application requirements. This hierarchy level can be used by ad hoc file systems to share data and to provide better performance than that provided by general-purpose storage backends. Ad hoc file systems can be tailored for specific application semantics and can be applied, for example, in the following use cases.

Big data workloads. Data processing and analysis have always been important applications for smaller and mid-sized HPC clusters. Researchers, for example from high-energy physics, astronomy, or bioinformatics, have developed community-specific workflow and processing environments that often have been adapted to the specific properties of HPC backend storage systems [2] [22] [39] [90]. Using HPC backend storage as primary file systems, however, unfortunately also restricted these big data applications to the drawbacks of centralized storage, for example that bandwidths and IOPS are shared between all concurrently running applications. Big data processing is nevertheless not restricted to HPC, and cloud-specific big data environments such as Hadoop or Spark [71] [89], as well as scalable NoSQL databases such as MongoDB or Cassandra [4] [13], attracted researchers to adapt their applications to new and more easily programmmed environments [34] [55]. Thus, unified environments are needed in order to process both HPC and big data workloads, where the converged environment should keep the promises and benefits of both approaches [15] [26] [83]. Ad hoc file systems can help couple locality with a global namespace, while additionally providing the random access rates of node-local SSDs. Nevertheless, in this case the ad hoc file systems must also support long-running campaigns, so that data staging between the backend storage and the backend parallel file system can be reduced to a minimum (see also Section 6).

Bulk-synchronous applications. Bulk-synchronous applications are the dominant workload seen on today's HPC systems. Here, applications run in a loosely synchronized fashion, generally synchronizing on major timestep boundaries. At these boundaries, the applications perform collective communication and I/O operations, for example output or visualization dumps and checkpoint/restart. In the general case, the I/O operations per process are independent and written either to per-process files or to process-isolated offsets in a shared file. Additionally, read and write operations occur in bulk phases, without concurrent interleaving of reads and writes.

These behaviors can easily be supported by ad hoc file systems that can provide higher performance than general-purpose file systems can. In the shared file case, the ad hoc file system can create a shared namespace across disjoint storage devices, enabling applications with this behavior to use fast storage tiers. Also, because we know that the processes will not read and write concurrently and that each process will write to its own isolated offsets, the ad hoc file system does not need to implement locking around write operations and hence can greatly improve performance.

Checkpoint/restart. HPC applications can survive failures by regularly saving their global state in checkpoints, which are often stored in the backend parallel file systems. The application can, in case of a failure, restart from the last checkpoint. Especially long-running applications benefit from the ability to restart failed simulation runs. However, the time to take a single application checkpoint increases linearly with the size of the application, and the overall check-

pointing overhead increases with the checkpoint frequency. Older studies have shown that up to 65% of applications' runtimes were spent in performing checkpoints [24] [59], and studies indicate that up to 80% of HPC I/O traffic are induced by checkpoints [58].

Many HPC backend storage systems have been designed according to the peak demands of their checkpointing load. The node-internal SSDs often have a higher peak performance than the backend file system has, and the number of node-internal SSDs available for a checkpoint linearly scales with the job's size. Several approaches for using node-internal storage as the checkpoint target have been developed, for example by integrating them into the MPI-IO protocol [16] or by offering dedicated checkpointing libraries [51]. These can be combined with strategies to reduce the checkpoint size, for example by using compression or deduplication [33] [35].

These approaches are partly bound to the usage of MPI or the availability of libraries. An interesting alternative is the use of dedicated ad hoc checkpointing file systems, which can store the checkpoint either in main memory or on node-local SSDs and which then can asynchronously flush (some of the) checkpoints to persistent storage [63]. The asynchronous nature of flushing the checkpoints in a background process even allows first storing the checkpoint locally and later moving the data to storage that cannot be affected by a local failure [16]. One can even overcome network latencies and store a checkpoint at memory speed [63].

Machine and deep learning workloads. The desired input data sizes of machine and deep learning workloads are increasing rapidly. The reason is that the use of small dataset sizes can fail to produce adequately generalized models that can recognize real-world variations in input, such as poses, positions, and scales in images. The typical I/O workload for learning applications is that random samples from the full dataset are read repeatedly from the backend store during training. These random reads from a parallel file system can be a bottleneck, especially for learning frameworks being run on GPU clusters with very high computational throughput. For smaller datasets, the parallel file system cache, or perhaps node-local storage such as an SSD, can hold the entire dataset, and performance is not an issue. Larger datasets, however, may not fit in the file system cache or node-local storage, and the performance of the learning workload can suffer because of the I/O bottleneck [93].

Learning workloads can benefit from ad hoc file systems. One such ad hoc file system could distribute the very large datasets across the memory or storage on other compute nodes of a job and serve the randomly requested input to each process as needed. Another file system implementation could prefetch the randomly selected portions of the dataset from the parallel file system to the compute nodes if the dataset cannot fit in node-local storage. In both cases, the learning workload would see vast improvements in I/O performance resulting in better training throughput.

Producer-consumer workloads. Several variations of producer-consumer workloads on HPC systems exist. Perhaps the most canonical of these is found in climate model codes, for example E3SM [1], where different physical components are modeled in individual executables (e.g., land, ocean, atmosphere, or ice). The individual component executables consume data files as input and produce output files that can be in turn consumed by another component executable. A common workflow for these models is to run them concurrently and have the components produce and consume files to compute the overall simulation.

Another emerging example of producer-consumer workloads for HPC is data analytics [66]. Here, tradi-

tional HPC simulations produce simulation output that is read in and analyzed by processes in the same allocation, for example coupled simulation and machine learning tasks for climate analytics [41]. The analysis could be done in situ, as a library linked into the application or an executable running on the same compute node, or simply coscheduled in the allocation on separate compute nodes. The analysis processes can perform tasks such as feature extraction or machine learning.

In both these producer-consumer workloads, the workflow can benefit from ad hoc file systems that are able to facilitate the sharing of data between the components in the job without resorting to using the parallel file system. The ad hoc file system can keep the data on fast, local storage and manage moving the bytes as appropriate for best performance with respect to the producer and consumer components. For example, writing bytes to the local storage of the producer will result in best performance from the producer's perspective. When it is time for the consumer to read the bytes, however, the file system could move the file contents to the compute node where the consumer is running, for best read performance.

3 Reusable components

Ad hoc file system implementations can be specialized for a variety of use cases (Section 2) and interfaces (Section 4). Despite this specialization, however, each ad hoc file system will encounter a common set of underlying technical challenges posed by HPC platforms and HPC application workloads. These challenges include the following:

- HPC network fabric compatibility
- Storage device compatibility
- Maximizing concurrency
- Minimizing resource consumption

Reusable building block components can help address these challenges while improving file system developer productivity, reducing software maintenance cost, and enhancing portability. In this section, we highlight each of these challenges and survey the state of the art in reusable components to address them.

3.1 HPC network fabric compatibility

Ad hoc file systems that operate "in-system" must, by definition, interact with the system's HPC network fabric. HPC networks are characterized by low latency (to accommodate tightly coupled computations), RDMA access (to minimize CPU impact on data transfers), and low message loss (because the environment tends to be static and homogeneous). No single standard for HPC network hardware exists that meets these requirements, however. Many of the world's most powerful computers use different network technologies [3] [14] [23], each with its own distinct API and optimization strategy. Data services must therefore employ network abstraction layers to ensure portability.

TCP/IP sockets are the most widely used and most portable network abstraction model, but they lack the specialized features needed to accommodate the latency, RDMA, and message loss characteristics of a dedicated HPC network. The Message Passing Interface (MPI) provides a network abstraction and programming model that is directly tailored to HPC environments [28]. However, MPI was designed for application-level use and is not readily applicable to system services in practice [42]. More generalized HPC network fabric abstractions such as OFI/libfabric [29], UCX [70], and Portals [6] are not tied to MPI semantics or programming models and are thus more appropriate foundational building blocks for ad hoc file system implementations.

Remote procedure call framworks can be used in conjunction with network abstractions to further ease the task of constructing ad hoc file system services. Examples in general-purpose distributed computing include gRPC* and Apache Thrift†, while examples in HPC include Mercury [72] and Nessie [56]. RPC frameworks implement common client/server functionality such as request/response matching, protocol encoding, service handler invocation, and programmatic API bindings.

3.2 Storage device compatibility

Storage devices are the second crucial resource that ad hoc file systems must manage. Parallel and distributed file systems have long utilized local file systems such as EXT4 and XFS as the abstraction point between distributed service daemons and local storage resources. Local file system abstractions are still a valid and highly portable design choice, but the emergence of new storage device technologies has prompted renewed exploration of alternative interfaces (see also Section 4).

Persistent memory devices in particular have more in common with dynamic memory than they do with rotating magnetic media and can thus be accessed more efficiently through direct user-space load/store operations than through indirect kernel-space block device and page cache operations. This property has led to the creation of storage abstractions such as the Persistent Memory Developers Kit (PMDK)‡ that provide simple transactional storage primitives atop memory-mapped devices rather than block devices.

Faster block device interfaces can also benefit from lower-latency access paths. For example, the Storage Performance Development Kit (SPDK)§ provides an alternative interface to NVMe devices that relies on user-space poll-driven device drivers to minimize latency. PMDK and SPDK are intended for use with two different storage technologies, but they share the common goal of minimizing access cost for devices that do not conform to the design assumptions and performance characteristics of legacy hard drives.

3.3 Maximizing concurrency

The following are three major drivers of concurrency in ad hoc file system services:

- Request arrival rate from parallel applications
- Availability of multicore processors on service nodes
- Storage resources that require parallel request issue to maximize bandwidth

Balancing these factors while limiting implementation complexity is a daunting task. Several techniques and strategies are available to help, however. Services can leverage an asynchronous event-driven model using frameworks such as Seastar¶ or libraries such as libevent‖ or libev**, and storage device concurrency can be achieved by offloading operations to asynchronous APIs such as the POSIX asynchronous I/O interface or the SPDK framework described in the preceding section.

Multithreading can also be used to manage concurrency without adopting an event-driven programming model. Conventional POSIX threads are functionally effective but introduce excessive overhead when concurrency exceeds the number of compute cores available on a system. User-level threads seek to find a middle ground between the efficiency of event-driven frameworks and the programmability of POSIX threads. Ex-

*https://grpc.io/
†https://thrift.apache.org/
‡https://pmem.io/pmdk/
§https://spdk.io/
¶http://seastar.io/
‖https://libevent.org/
**http://software.schmorp.de/pkg/libev.html

amples of modern user-level threading packages include Qthreads [85], MassiveThreads [53], and Argobots [69]. Argobots is particularly amenable to distributed service development because it supports customized schedulers and flexible mappings of work units to compute resources.

3.4 Minimizing resource consumption

The term "ad hoc file system" implies that the file system does not necessarily have dedicated resources but is instead provisioned on-demand within an existing system. An ad hoc file system must therefore be able to coexist with other services or even application processes without dominating available resources or causing excessive jitter. Common resource consumption pitfalls include the following:

- Memory consumption
- Busy polling of network resources
- CPU or NUMA contention

No single component solves these challenges; they are more readily addressed by design patterns that account for the capabilities of HPC hardware resources. For example, non-volatile memory can be combined with RDMA-capable networks (either in user space or via kernel drivers) to limit memory consumption by transferring data directly between network and storage without intermediate buffering. RPC frameworks and network abstraction layers can use adaptive polling strategies to limit network and CPU use when services are idle. CPU affinity and NUMA control can prevent colocated services and applications from contending for CPU and memory resources. These design patterns can be used in any ad hoc file system implementation.

3.5 The Mochi project

Mochi [21] seeks to combine many of the components and best practices described in this section into a coherent environment to accelerate data service development. This type of environment also enables utility libraries to *span* components in order to implement complementary best practices. This is exemplified by the Margo [10] and abt-io [69] libraries that integrate reusable RPC and file I/O functionality into the Argobots threading framework.

Regardless of development environment and runtime system, however, the core principles of reusable components and design patterns can be applied to accelerate the development of ad hoc file systems and allow their creators to spend more time on data service innovation.

4 Interfacing ad hoc file systems

Since the arrival of parallel file systems over two decades ago, file systems have continued to become increasingly more complex, spanning intricate logic over million of lines of code [79] with the goal of offering a general-purpose solution for all applications (see Section 2). In order to allow for a familiar and portable interface that most applications can agree and rely on, such general-purpose file systems follow a set of rules in the POSIX family of standards that define the syntax and semantics of the I/O interface, also known as *POSIX I/O*. As a result, the POSIX I/O interface is deeply embedded within the operating system and common to users and applications, usually accessed via the GNU C library (glibc). Note that these application-oriented interfaces should not be confused with other interfaces such as block I/O, object-oriented I/O, or file-based I/O, which are defined for lower-level storage media access.

The I/O interface can be implemented at the user level or the kernel level. As shown in Figure 1, kernel-based file systems such as EXT4 may work entirely within the kernel, or they can utilize an additional user-space implementation where application I/O calls are redirected to, for example, FUSE. The kernel-based approach works by registering the file system to the *virtual file system* (VFS), which can be accessed by standard libraries. Exclusive user-space file systems, on the other hand, can also operate without a kernel component, providing their own implementation by *preloading* a library through the `LD_PRELOAD` environment variable, which intercepts defined I/O calls and redirects them to the user-space file system. These types of file systems are becoming increasingly popular (including in ad hoc file systems) because of their advantages for code development, porting, debugging, maintenance, and often performance.

An alternative to the standard I/O interface is proposed by so-called I/O libraries or I/O wrappers, providing a thinner set of API functions that may ease deployment and maintenance. Such libraries are not necessarily backed by a user-space file system and align access patterns with the capabilities of the underlying PFS (e.g., ADIOS [48]). They can also be used as an API to a separate user-space file system such as OrangeFS [52].

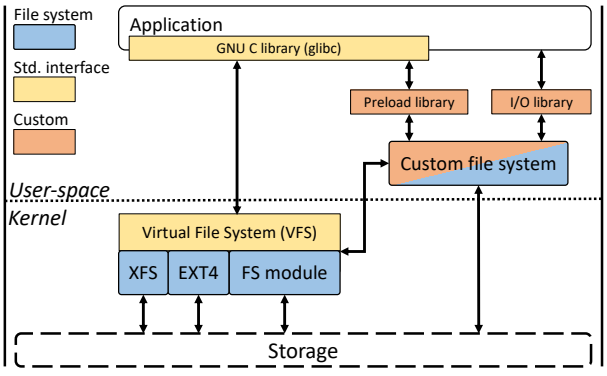


Fig.1. Techniques to interface file systems with their components.

In this context, the key-value interface has recently gained traction. Volos et al. [80] developed a flexible file system architecture called Aerie that can support user-level I/O accesses to storage-class memory. Nonetheless, a completely non-standardized custom interface usually requires the modification of applications and is not suitable in all cases.

In the following, we discuss some of the existing techniques to implement user-space file systems, and we highlight the most popular approach when developing an ad hoc file system in user space.

4.1 User-space file systems

User-space file systems that are interfaced via the traditional standard interface cannot be easily registered to an operating system compared with kernel-based file systems. Instead, exclusive user-space file systems are directly linked to an application at compile time or loaded at runtime, which on UNIX systems can be achieved by a preloading library. The latter technique is most popular and has been adopted by several ad hoc file systems, such as CRUISE [63], BurstFS [81], GekkoFS [78], and DeltaFS [92]. They all intercept the application I/O via a set of wrapper functions that are implemented in the form of a user-level preloading library, albeit not always being fully POSIX-compliant [78].

In addition to easing development or maintenance by using such a library, avoiding the kernel when calling I/O functions can yield significant performance benefits in terms of I/O throughput and latency. Volos et al. [80] argue that the existing kernel-based stack of components, although well suited for disks, unnecessarily limits the design and implementation of file systems for faster storage (e.g., storage-class memory). Their Aerie framework, a POSIX-like file system in user space, has been implemented with performance similar to or bet-

ter than a kernel implementation.

PMDK bypasses the kernel and builds on Linux's DAX features, which allow applications to use persistent memory as memory-mapped files. Therefore, techniques such as `LD_PRELOAD` are attractive choices for ad hoc file systems, which often utilize flash-based node-local storage. With more advanced future storage technologies (e.g., persistent memory), reducing the time spent within the operating system to increase I/O performance is going to become even more critical.

Because of the ability to intercept any function in preloading libraries, however, all I/O functions used by an application must be implemented and reinterpreted in the user-space file system. Hence, the more complex an application, the more the number of functions required in the file system for the application to work, potentially intercepting a large percentage of functions that are part of glibc, for instance. The *system call intercepting library*[††] (syscall_intercept), as part of the pmem project, aims to solve this challenge by providing a low-level interface for hooking Linux system calls in user space while still using the established `LD_PRELOAD` method. This can dramatically reduce the number of functions that need to be implemented in the user-space file system because the set of functions is limited to only system calls, such as `sys_mknod` or `sys_write`.

Other efforts have also explored ways to standardize the interception of POSIX-related I/O calls, including *libsysio* by Sandia National Laboratories and *Gotcha* by Lawrence Livermore National Laboratory. Libsysio provides a POSIX-like interface that redirects I/O function calls to file systems and supports the conventional VFS/vnode architecture for file-based accesses [37]. Gotcha is a wrapper library that works similar to `LD_PRELOAD` but operates via a programmable API [60].

[††]`https://github.com/pmem/syscall_intercept`

4.2 FUSE file systems

The FUSE (Filesystem in Userspace) [64] [77] framework is another popular approach when developing user-space file systems. It consists of two components, the FUSE kernel module and the *libfuse* user-space library, that support kernel-based and user-level redirection for I/O calls, respectively. The libfuse library is executed as part of a process that manages the entire file system at the user level and communicates through the FUSE kernel module with the kernel. Internally, all I/O calls are implemented as callbacks through the *libfuse* library, which supports the communication between the FUSE kernel module and the user-level file systems [64]. FUSE therefore provides a traditional file system interface, but it allows users to quickly implement a file system based on FUSE's API by avoiding much of the complexity within the kernel. As a result, a large number of FUSE-based file systems have been developed, including ChunkFS [30], SSHFS [31], FusionFS [91], and GlusterFS [18].

While convenient, FUSE-based user-level file systems also face several drawbacks inherent to its architecture. Since the libfuse library is typically executed as a separate user process, the communication round trip between an application and the FUSE process leads to nontrivial performance overhead [77]. To make the situation worse, an application leveraging the kernel module for I/O interception will experience even more overhead due to context switches across the user-kernel boundary. Furthermore, root permission is required to mount FUSE file systems, or system administrators have to give nonprivileged users the ability to mount FUSE file systems. In HPC environments, users do not typically own such superuser privileges and cannot easily mount the desired FUSE file systems without help from system administrators.

In ad hoc file systems, FUSE's disadvantages outweigh the advantage of convenient development, and it is therefore rarely used in these file systems.

5 Ad hoc file system implementations

The previous two sections have discussed basic building blocks of ad hoc file systems and how to interface them. This section now puts these components together and presents three different ad hoc file system implementations for different use cases.

- BeeGFS-On-Demand, or BeeOND for short, is a production-quality BeeGFS wrapper that simplifies the deployment of multiple independent BeeGFS instances on one cluster to aggregate the performance and capacity of internal SSDs or hard disks. Important aspects of BeeOND clients are developed as kernel modules.
- GekkoFS has been designed to overcome scalability limitations of the POSIX protocol [78]. It both slightly relaxes the POSIX semantics and reduces security guarantees, while still being able to ensure confidentiality in the setting of ad hoc file systems. Performance is achieved by spreading data and metadata as widely as possible.
- The Burst Buffer File System (BurstFS) uses techniques such as scalable metadata indexing, colocated I/O delegation, and server-side read clustering and pipelining to support scalable and efficient aggregation of I/O bandwidth from node-local storage [81]. Clients furthermore write to node-local logs to increase write performance, providing a different approach from that of GekkosFS and being suited for different applications.

All file systems have been developed to create temporary file systems that run for the duration of a compute job. Hence, the ad hoc file systems must be integrated with a workload manager such as Slurm or Torque, and the deployment time must be as short as possible so that the runtime of the compute job is not negatively affected by the start-up phase of the file system.

5.1 BeeGFS and BeeOND

BeeGFS, initially named *Fraunhofer Parallel File System* (FhGFS), was started in 2004 as a result of a project to integrate the Lustre PFS in a video streaming environment and as storage in an existing 64-node Linux cluster. The initial requirements for the development were to distribute metadata, to require no kernel patches, and to support zero-config clients with the goal of providing a scalable multithreaded architecture and dynamic failover for native InfiniBand and Ethernet.

The file system is built to run on any underlying POSIX-compliant local file system, allowing users the choice and flexibility if only a specific local file system is available. BeeGFS is designed to be easily deployed and maintained while focusing on performance instead of on features. The client is built as a Linux kernel module, while other software components were moved to user space, allowing increased usage flexibility.

Staying POSIX-compliant in a distributed world without sacrificing performance throughout the years is a challenging process, requiring detailed analyses and careful implementation. Today, BeeGFS has become a

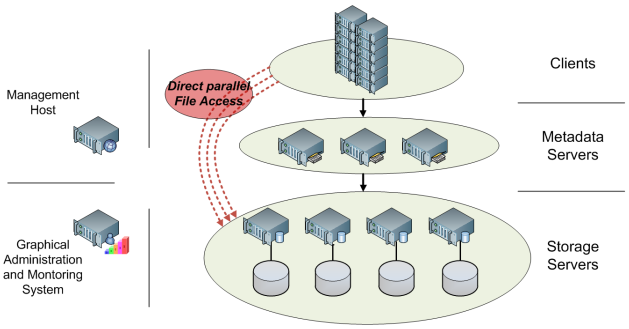


Fig.2. Architecture of the BeeGFS file system.

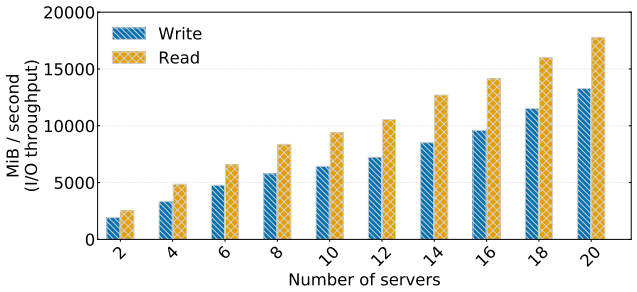


Fig.3. BeeGFS's I/O throughput for increasing server numbers for 192 client processes on one shared file.

true high-performance distributed file system that includes a reliable and fast distributed locking algorithm. After five years of heavy development, BeeGFS was ready for installation in larger environments, delivering the highest single-thread performance on the market up to this day and showing its strengths in N:1 shared file I/O cases (see Figure 3).

Figure 2 describes BeeGFS' overall architecture. The client works as a kernel module with its own caching strategy, which is kept updated and follows the latest kernel developments. The metadata server, the storage server, and the management server are independent processes that can be installed on a single server or in a distributed setting, depending on the needs of the user.

When a QDR InfiniBand system (with SSDs in each node) was installed at Fraunhofer in 2012, *BeeOND* was born out of the idea to use BeeGFS as an ad hoc file system on nodes that a user got assigned to by the cluster's batch system. BeeOND creates an empty, private distributed file system across all job nodes to separate challenging I/O patterns from the I/O load of the PFS. The file system is then started during the prolog process of the batch system. Using BeeGFS for this task was ideal because all server processes are already running in user-space and do not require further patches to be installed in advance. One of its first use cases was a data-sorting routine for seismic data that reads data from the PFS and uses the SSD storage as an intermediate data buffer before the sorted data is written back to the PFS. BeeGFS supports advanced features such as data mirroring, because of the high availability requirements in hyperconverged solutions to store large data sets economically, for instance.

Today, BeeOND is used around the world as an alternative to expensive burst buffers and allows users to manage the demanding requirements of deep learning applications. It inherits all the functionality of BeeGFS and offers a POSIX-compliant and scalable distributed file system. The installation at Tsubame 3.0 and the ABCI system in Japan are the most prominent BeeOND installations and can reach 1 TB/s streaming throughput. In the future, BeeGFS and BeeOND will be continuously developed to improve performance and to satisfy the needs of arising use cases.

5.2 GekkoFS

The basic design idea behind GekkoFS is to distribute data and metadata among cluster nodes as evenly as possible [78]. GekkoFS therefore aggressively uses hashing to distribute data and metadata. Each file inode is managed by one cluster node, which can be determined by simply calculating a hash function of the file path and name and then by taking the result modulo the number of nodes participating in the file system.

A client node, for example, creates a new file by first computing the managing cluster node and then by running the file create protocol between the client and the managing node. The serialization inside the managing node guarantees that an existing file cannot be created a second time. The protocol ensures that most metadata operations linearly scale in the number of participating cluster nodes.

Distributing inode metadata over all cluster nodes nevertheless also requires a different directory handling.

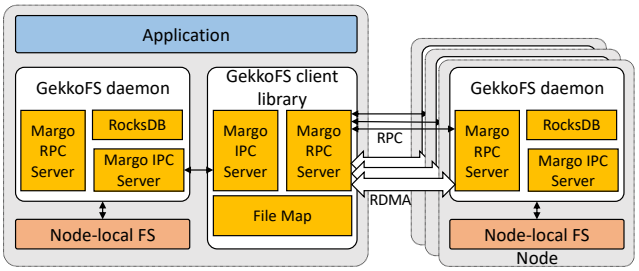


Fig.4. GekkoFS architecture.

GekkoFS therefore significantly relaxes the POSIX directory semantics. Directories are created similar to files, while the content of a directory is only implicitly available by collecting distributed inode entries using broadcast operations. GekkoFS is therefore not suited for applications that regularly require listing the content of directories using `ls`-operations or which rename directories, since this requires expensive one-to-all and all-to-one communication patterns and in the second case also requires updating all distributed inode entries. Fortunately, studies have shown that these operations are extremely rare while running a parallel job [44].

GekkoFS provides the same consistency as POSIX for any file system operation that accesses a specific file. These include read and write operations as well as any metadata operation that targets a single file, for example, file creation. Nevertheless, similarly to PVFS [11], GekkoFS does not provide a global byte-range lock mechanism. In this sense, applications are responsible for ensuring that no conflicts occur, in particular w.r.t. overlapping file regions, in order to avoid complex locking within the file system.

The GekkoFS-architecture shown in Figure 4 consists of two main components: a client library and a server process. An application that uses GekkoFS must first preload the client interposition library that intercepts all file system operations and forwards them to a server (GekkoFS daemon), if necessary. The GekkoFS daemon, which runs on each file system node, receives forwarded file system operations from clients and processes them, sending a response when finished. The daemons operate independently and do not communicate with other server processes on remote nodes, therefore being effectively unaware of each other.

The client consists of an interception interface that catches relevant calls to GekkoFS and forwards unrelated calls to the node-local file system; a file map that manages the file descriptors of opened files and directories, independently of the kernel; and an RPC-based communication layer that forwards file system requests to local/remote GekkoFS daemons.

A GekkoFS daemon's purpose is to process forwarded file system operations of clients to store and retrieve data and metadata that hashes to a daemon. To achieve this goal, GekkoFS daemons use a RocksDB [20] key-value store (KV store) for handling metadata operations, an I/O persistence layer that reads/writes data from/to the underlying node-local storage system, and an RPC-based communication layer that accepts local and remote connections to handle file system operations.

The communication layer uses the Mercury RPC framework, which allows GekkoFS to be independent from the network implementation [73]. Mercury is interfaced indirectly through the Margo library, which provides wrappers to Mercury's API with the goal of providing a simple multithreaded execution model [12]. It uses the lightweight, low-level threading and tasking framework Argobots, which has been developed to support massive on-node concurrency.

The experiments for the results presented in Figures 5 and 6 have been performed on the MOGON II cluster at the Johannes Gutenberg University Mainz, Germany. The cluster consists of 1,876 nodes in total, with 822 nodes using Intel 2630v4 Intel Broadwell processors (two sockets each) and 1046 nodes using Xeon Gold

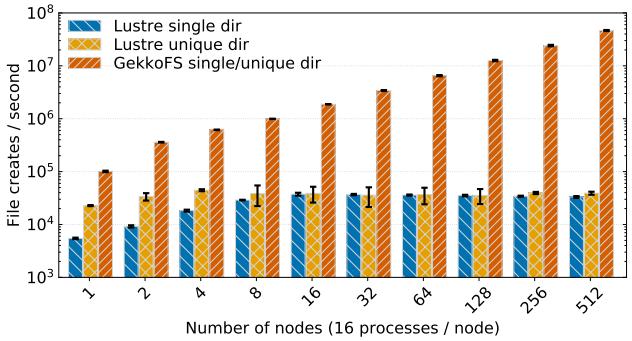


Fig.5. GekkoFS's file create performance compared with a Lustre file system for increasing node numbers.

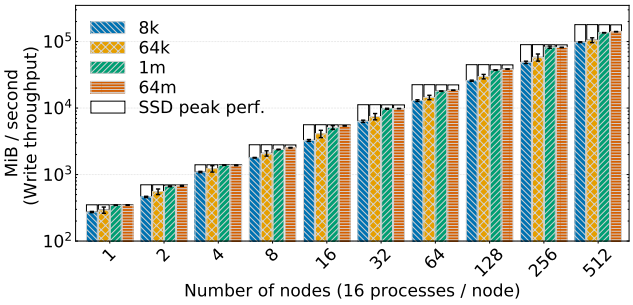


Fig.6. GekkoFS's write throughput for increasing node numbers.

6130 Intel Skylake processors (two sockets each). The Intel Broadwell processors have been used in all presented experiments. The main memory capacity inside the nodes ranges from 64 GiB to 512 GiB, and the nodes are connected by a 100 Gbit/s Intel Omni-Path interconnect. The cluster is attached to a 7.5 PiB Lustre storage backend.

In addition, each node includes an Intel SATA SSD DC S3700 Series with 200 GiB or 400 GiB, which have been used for storing data and metadata of GekkoFS. All Lustre experiments were performed on a Lustre scratch file system with 12 Object Storage Targets (OSTs), 2 Object Storage Servers (OSSs), and 1 Metadata Service (MDS) with a total of 1.2 PiB of storage. The experiments were run at least five times with each data point representing the mean of all iterations.

Figure 5 compares GekkoFS with Lustre for file creates for up to 512 nodes on a logarithmic scale. GekkoFS's workload was scaled with 100,000 files per process. Lustre's workload was fixed to 4 million files for all experiments. We fixed the number of files for Lustre's metadata experiments because Lustre was detecting hanging nodes when scaling to too many files. Lustre experiments were run in two configurations: all processes operated in a single directory (`single dir`), or each process worked in its own directory (`unique dir`). Moreover, Lustre's metadata performance was evaluated while the system was accessible by other applications as well.

GekkoFS outperforms Lustre by a large margin, regardless of whether Lustre processes operated in a single directory or in isolated directories. GekkoFS achieved around 46 million creates per second, while each operation was performed synchronously without any caching mechanisms in place, showing close to linear scaling. Lustre's create performance did not scale beyond approximately 32 nodes and has been ∼1,405x lower than the create-performance of GekkoFS, demonstrating the well-known metadata scalability challenges of general-purpose PFS.

GekkoFS's data performance is not compared with the Lustre scratch file system because Lustre's peak performance of 12 GByte/s is already reached for 10 nodes for sequential accesses. Moreover, Lustre has shown to scale linearly for sequential access patterns in larger deployments with more OSSs and OSTs being available [57]. Figure 6 shows GekkoFS's sequential I/O throughput in MiB/s for an increasing number of nodes for different transfer sizes. In addition, each data point is compared with the peak performance that all aggregated SSDs could deliver for a given node configuration, visualized as a white rectangle. The results demonstrate GekkoFS's close-to-linear scalability, achieving about 141 GiB/s (∼80% of the aggregated SSD peak bandwidth) and 204 GiB/s (∼70% of the aggregated SSD peak bandwidth) for write and read

operations with a transfer size of 64 MiB for 512 nodes. At 512 nodes, this translates to more than 13 million write IOPS and more than 22 million read IOPS, while the average latency can be bounded by at most 700 μs for file system operations with a transfer size of 8 KiB.

5.3 The Burst Buffer File System (BurstFS)

BurstFS shares a number of basic design considerations with GekkoFS. It has been designed to have the same temporary life cycle as a batch-submitted job, while it uses node-local burst buffers to improve applications' read and write performance [81]. The main distinction between the two ad hoc file systems is that BurstFS clients always write to local storage in a log-structured way. This can significantly improve write performance since no network latency is involved; but it also requires building a metadata directory to reconstruct writes coming from multiple clients to one file in case of N-1 access file patterns.

When a batch job is allocated over a set of compute nodes, an instance of BurstFS will be constructed on the fly across these nodes, using the locally attached burst buffers, which may consist of memory, SSDs, or other fast storage devices. These burst buffers enable very fast log-structured local writes; in other words, all processes can store their writes to the local logs. Next, one or more parallel programs launched on a portion of these nodes can leverage BurstFS to write data to or read data from the burst buffers.

BurstFS is mounted with a configurable prefix and transparently intercepts all POSIX functions under that prefix. Data sharing between different programs can be accomplished by mounting BurstFS using the same prefix. Upon the unmount operation from the last program, all BurstFS instances flush their data for data persistence (if requested), clean up their resources, and exit.

BurstFS uses MDHIM as distributed KV store for metadata, along with log-structured writes for data segments [27]. Figure 7 shows the organization of data and metadata for BurstFS. Each process stores its data to the local burst buffer as data logs, which are organized as data segments. New data are always appended to the data logs. With such log-structured writes, all segments from one process are stored together regardless of their global logical position with respect to data from other processes.

When the processes in a parallel program create a global shared file, a key-value pair (e.g., M1 or M2) is generated for each segment. A key consists of the file ID (8-byte hash value) and the logical offset of the segment in the shared file. The value describes the actual location of the segment, including the hosting burst buffer, the log containing the segment (there can be more than one log from multiple processes on the same node), the physical offset in the log, and the length. The key-value pairs for all the segments then provide the global layout for the shared file. All KV pairs are consistently hashed and distributed among the key-value servers (e.g., KVS0 and KVS1). With such an organization, the metadata storage and services are spread across multiple key-value servers.

Lazy synchronization provides efficient support for bursty writes. Each process holds a small memory pool for metadata KV pairs from write operations, and, at

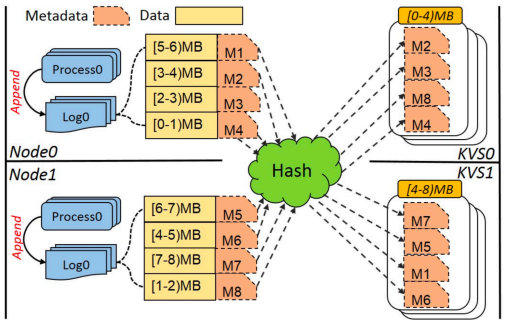


Fig.7. Diagram of the distributed key-value store for BurstFS.

the end of a configurable interval, KV pairs are periodically stored to the key-value stores. An `fsync` operation can force an explicit synchronization. BurstFS leverages the batch put operation from MDHIM to transfer these KV pairs together in a few round trips, minimizing the latency incurred by single put operations. During the synchronization interval, BurstFS searches for contiguous KV pairs in the memory pool to combine, which can span a bigger range and reduce the number of data segments. As shown in Fig. 7, segments [2-3) MB and [3-4) MB are contiguous and map to the same server, so their KV pairs are combined into one.

Read operations involve a metadata look-up for the distributed data segments. Thus, they search for all KV pairs whose offsets fall in the requested range. With batched read requests, BurstFS needs to search for all KV pairs that are targeted by the read requests in the batch. However, range queries are not directly supported by MDHIM and indirectly performing them by iterating over consecutive KV pairs induces additive round-trip latencies.

BurstFS therefore includes parallel extensions for both MDHIM clients and servers [81]. On the client side, incoming range requests are broken into multiple small range queries to be sent to each server based on consistent hashing [36]. Compared with sequential cursor operations, this extension allows a range query to be broken into many small range queries, one for each range server. On the server side, for the small range query within its scope, all KV pairs inside that range are retrieved through a single sequential scan in the key-value store.

Scalable read and write services are furthermore achieved through a mechanism called co-located I/O delegation. BurstFS launches an I/O proxy process on each node, a delegator. Delegators are decoupled from the applications in a batch job and are launched across

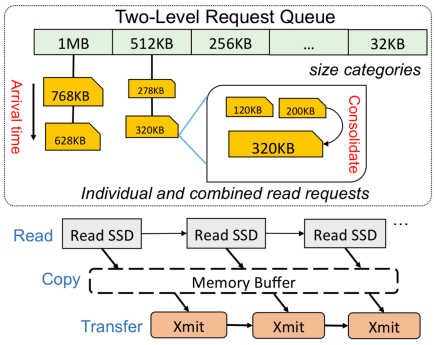


Fig.8. Server-side read clustering and pipelining.

all compute nodes. The delegators collectively provide data services for all applications in the job by offering a request manager and an I/O service manager. In this way, a conventional client-server model for I/O services is transformed into a peer-to-peer model among all delegators.

The delegators allow BurstFS to leverage the existing techniques of batched reads from the client side, where POSIX commands such as `lio_listio` allow read requests to be transferred in batches. BurstFS exploits the visibility of read requests at the server side for further performance improvements by introducing a mechanism called server-side read clustering and pipelining (SSCP) in the I/O service manager. In the two-level request queue, SSCP first creates several categories of request sizes, ranging from 32 KB to 1 MB (see Fig. 8). Incoming requests are inserted into the appropriate size category either individually or, if contiguous with other requests, combined with the existing contiguous requests and then inserted into the suitable size category. As shown in the figure, two contiguous requests of 120 KB and 200 KB are combined by the service manager. Within each size category, all requests are queued based on their arrival time. A combined request will use the arrival time from its oldest member. For best scheduling efficiency, the category with the largest request size is prioritized for service. Within

the same category, the oldest request will be served first. BurstFS enforces a threshold on the wait time of each category (default 5 ms). If any category has not been serviced longer than this threshold, BurstFS selects the oldest read request from this category for service.

Experiments comparing BurstFS with OrangeFS 2.8.8 [52] and the Parallel Log-Structured File System 2.5 (PLFS [7]) have been conducted on the Catalyst cluster at Lawrence Livermore National Laboratory, consisting of 384 nodes. Each node is equipped with two 12-core Intel Xeon Ivy Bridge E5-2695v2 processors, 128 GB of DRAM, and an 800-GB burst buffer comprising PCIe SSDs.

In the experiments, OrangeFS instantiated server instances across all the compute nodes allocated to a job to manage all node-local SSDs. PLFS is designed to accelerate N-1 writes by transforming random, dispersed N-1 writes into sequential N-N writes in a log-structured manner. Data written by each process is stored on the backend PFS as a log file. In the experiments, OrangeFS over node-local SSDs has been used as the PLFS backend. In PLFS with burst buffer support, processes store their metalinks on the backend PFS, which point to the real location of their log files in the burst buffers. Within the experiments, each process wrote to its node-local SSD, and the location was recorded on the center-wide Lustre parallel file system.

Figure 9 compares the write bandwidth with the PLFS burst buffer (PLFS-BB), PLFS, and OrangeFS. Sixteen processes are placed on each node, each writing 64 MB data following an N-1 strided pattern. Both BurstFS and PLFS-BB scale linearly with process count. The reason is that processes in both systems write locally and the write bandwidth of each node-local SSD is saturated. OrangeFS and PLFS also scale linearly, while their bandwidths increase at a much slower

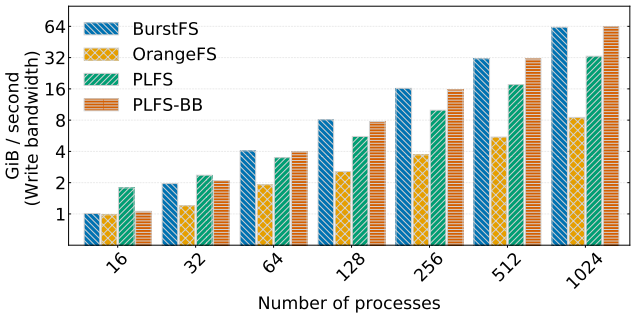


Fig.9. Comparison of BurstFS with PLFS and OrangeFS for N-1 segmented writes.

rate. The reason is that both PLFS and OrangeFS stripe their files across multiple nodes, which can cause degraded bandwidth due to contention when different processes write to the same node. On average, BurstFS delivers 3.5x the performance of OrangeFS and 1.6x the performance of PLFS.

PLFS initially delivers a higher bandwidth than does BurstFS at small process counts. PLFS internally transforms the N-1 writes into N-N writes. However, when `fsync` is called to force these N-N files to be written to the backend file system, OrangeFS does not completely flush the files to the SSDs before `fsync` returns.

In order to evaluate the support for data sharing among different programs in a batch job, read tests with IOR were conducted. A varying number of processes read a shared file written by another set of processes from a Tile-IO program. Processes in both MPI programs were launched in the same job. Each node hosted 16 Tile-IO processes and 16 IOR processes. Once Tile-IO processes completed writing on the shared file, this file was read back by IOR processes using the N-1 segmented read pattern. The same transfer sizes were used for IOR and Tile-IO. Since the read pattern did not match the initial write pattern of Tile-IO, each process needed to read from multiple logs on remote nodes. The size of each tile was fixed to 128 MB, and the number of tiles along the y axis to 4, while the number of tiles

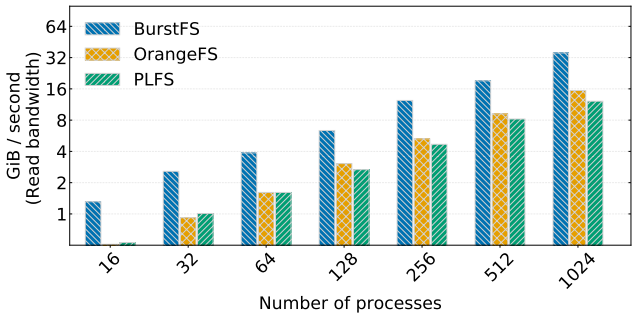


Fig.10. IOR read bandwidth on a shared file written by Tile-IO.

along the x axis was increased with the number of reading processes.

Figure 10 compares the read bandwidth of BurstFS with those of PLFS and OrangeFS. Both PLFS and OrangeFS are vulnerable to small transfer size (32 KB). BurstFS maintains high bandwidth because of locally combining small requests and server-side read clustering and pipelining. On average, when reading data produced by Tile-IO, BurstFS delivers 2.3x and 2.5x the performance of OrangeFS and PLFS, respectively.

6 Feeding ad hoc file systems: data staging

Ad hoc file systems can significantly speed up individual data accesses for an application when compared to a backend PFS, given that they allow exploiting faster node-local storage hardware such as NVRAM, and also since they provide application-specific data distributions and access semantics that can improve application I/O latency and/or bandwidth. Nonetheless, for ad hoc file systems to be useful, input data must be transferred (or "staged") into the file systems before running the targeted application and output data must be staged out after the application terminates. Since ad hoc file systems are ephemeral in nature, such *data staging* is typically done from/to the backend PFS. This section discusses requirements for coupling a batch scheduler, the backend PFS, and the ad hoc file system during stage-in and stage-out activities. Additionally, it discusses the potentially positive and negative interactions between concurrent jobs on a system while these stage-in and stage-out activities occur.

Data-intensive workloads are challenging for I/O subsystems because they present substantially larger I/O requirements than those of traditional compute-bound workloads [9] [81] [88]. Thus, even if modern HPC clusters can run multiple concurrent applications on top of millions of cores, severe I/O performance degradation is often observed because of cross-application interference [47] [87], a phenomenon that originates due to competing accesses to the supercomputer's shared resources. Nevertheless, the inclusion of fast node-local storage in compute nodes [5] [43] [86] enables the creation of a high-performance distributed staging layer where applications can efficiently store and retrieve data in isolation from each other which, if done correctly, can help reduce cross-application interference. Ad hoc file systems can become very useful in this scenario, both as enforcers of this isolation and as mediators for applications to transparently access this staging layer. Compared to traditional approaches, where applications directly access the PFS at their convenience, such a *staging architecture* has the advantage that arbitrary application I/O workloads would be substituted by system-controlled stage-in/stage-out I/O workloads, which could be scheduled to minimize interference between concurrent staging phases. Moreover, from the point of view of the PFS, global application I/O would be transformed from a stream of unrelated, random data accesses to a well-defined series of sequential read/write phases, which are better suited to be optimized.

Nonetheless, despite these benefits to HPC I/O performance, the presence of this node-local staging layer further increases the complexity of managing the storage hierarchy. When the hierarchy of the HPC storage

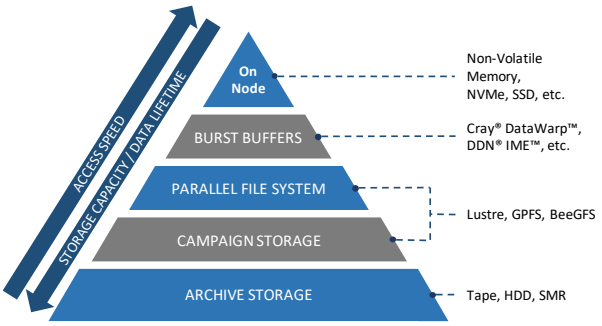
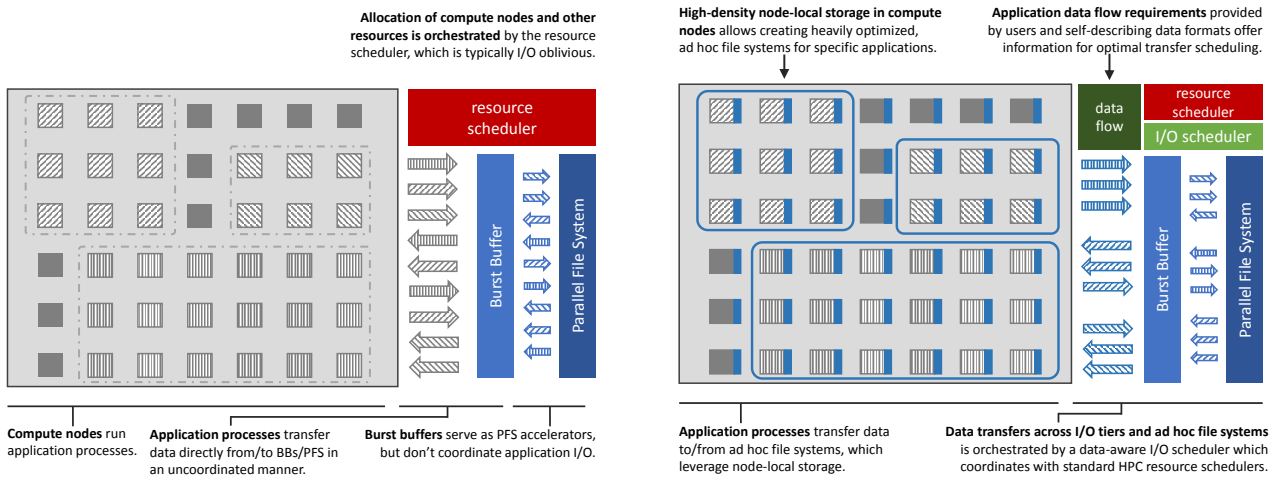


Fig.11. Growing complexity of the storage hierarchy of modern HPC systems, raising the need for research in order to understand how data and programming models expose and interact with this hierarchy.

system consists mainly of the parallel file system and archival storage, users of HPC clusters can easily manage the data movements required by their applications explicitly, either in the application code itself or in the scripts controlling their batch jobs. Explicitly managing data transfers between storage tiers, however, is not optimal since end users lack the required real-time information about the state of the cluster to decide the best moment to execute a transfer of data. Moreover, given the increasing complexity of current HPC storage architectures—which may include as many layers as node-local storage, storage on I/O nodes, parallel file systems, campaign storage, and archival storage (see Figure 11)—explicit transfer management forces users (i.e., scientists and researchers) to spend time learning the best way to use these technologies in their applications, an effort that would be better spent on their scientific problems. This means that opportunities for global I/O optimization are being missed by not communicating application data flows to the HPC services in charge of resource allocation and, as such, any architecture that does not expose information about the storage tiers to applications, and relies solely on the hardware and OS to transparently manage the I/O stack will lead to sub-optimal performance. Although with the advent of *shared burst buffers* [46], vendors are providing APIs for transferring data to/from the parallel file system into/out of burst buffers (e.g., Cray DataWarp API [17] and IBM BBAPI [32]), such APIs (1) do not yet schedule PFS I/Os by taking into account cross-application interference [40], and (2) do not yet concern themselves with node-local storage.

Addressing these challenges and utilizing this new data-driven staging architecture effectively requires developing new interfaces and APIs that allow end users to convey application *data flow requirements* (e.g., expected data lifetime, type of access, or visibility to related applications) to the services responsible for managing the HPC infrastructure. For instance, conveying data flow dependencies between jobs can help utilize the storage stack more effectively: if `Job A` generates output data that is going to be fed as input to `Job B`, a data-aware job scheduler could reuse `Job A`'s compute nodes for `Job B` and keep data in NVRAM. Unfortunately, today's users have no way to either express these dependencies or to influence the job-scheduling process so that `Job A`'s output data is kept in local storage until `Job B` starts. Worse yet, given that the I/O stack remains essentially a black box for today's job schedulers, `Job A`'s output data could be staged out to the cluster's parallel file system and, at some point in the near future, staged back into a new set of compute nodes for `Job B`, which might end up including some of the original nodes allocated to `Job A`.

Thus, we argue that integrating application data flows with scheduling and resource managers is critical for effectively using and managing a HPC hierarchical storage stack powered by ad hoc file systems. The development and deployment of data-aware scheduling services that ingest application data flow requirements and facilitate data movement across storage tiers, can improve the coordination between HPC resource managers and the storage stack, resulting in reduced PFS



(a) Cluster with uncoordinated backend access.

(b) Data-driven cluster with coordinated staging.

Fig.12. Designing a data-driven HPC cluster is possible by coupling application-aware data staging with the I/O isolation provided by ad hoc file systems. On traditional HPC clusters applications access backend storage directly causing contention due to uncoordinated I/Os. On a data-driven HPC cluster, application I/O is absorbed by node-local storage and transfers of application input and output artifacts are coordinated to maximize the performance of backend storage.

I/O contention and, in turn, improved job run times and system efficiency (see Figure 12). While several services have been proposed with similar goals [19] [50] [75] [82], to the best of our knowledge no resource scheduling algorithms have yet been proposed that take into account a job's dynamic requirements in terms of I/O (e.g. capacity, latency, and bandwidth) in addition to the more static computing requirements (e.g. compute nodes). There are thus plenty of research opportunities to investigate and develop APIs, infrastructure services and scheduling algorithms that can include application data flow needs into resource allocation decisions for I/O optimization. Provided with this information, and since many HPC applications exhibit relatively regular I/O patterns that would run in isolation in an architecture based on ad hoc file systems, these scheduling algorithms should be able to coordinate/interleave the staging phases from/to node-local storage in order to minimize the I/O contention of the parallel file system. This kind of algorithms have proved successful in avoiding contention in shared burst buffers and we believe that they could be extended to node-local storage [76].

Note, however, that major challenges still remain that are associated with scheduling storage resources on HPC systems. For example, is it possible to predict the best moment to start staging data into a compute node [74]? Would the data scheduler background transfers affect the cluster's networking subsystem so much that they impaired normal application execution? How should elastic workflows be addressed? What should the scheduling algorithms do when a job that is staging in data crashes? Could these algorithms increase the energy efficiency of the supercomputer?

Besides exploiting user-provided information conveyed through APIs and services, a step further consists of taking advantage of self-describing I/O [25] [45] [48] [65], which has become a key aspect in managing large-scale datasets. By enriching the self-describing nature of large datasets, and integrating it with data-aware infrastructure services, we are not simply moving and storing large numbers of bytes but rather creating a vehicle to extract the most possible information as efficiently as possible. The idea is to promote intelligent I/O, a mechanism to allow information to be published

and later subscribed to at all scales, for all types of data. The key to this is the ability to have self-describing data in streams and to think of data in motion in the same context as data at rest. We envision an extension to the publish/subscribe metaphor to include a clerk that will sit between the publisher and subscriber and mediate or orchestrate data streams in a dynamic fashion.

In order to support the emerging analytics, processing, and storage use cases, data cannot be considered passive, hence directly falling through a chute connecting publishers and subscribers. Instead, a service-oriented architecture must connect them; actors must be involved to touch, manage, maintain, and abstract the data and to support inspection tasks such as in code coupling, in situ analysis, or visualization. These sets of actions must be managed and orchestrated across the wide array of resources in a way that enables not just imperative connections ("Output A must go to Input B") but also new models of learning and intelligence in the system ("Make this data persistent, but watch what I've been doing to other data sets and preprocess this data based on that").

While one can build systems that can be tuned dynamically by a human in the loop, intelligent systems with the capability to automatically tune workflows and drive them according to data events observed at runtime will lead the way in the design of modern computing infrastructure. In this case, storage for applications, which can be in terms of memory, individual burst buffers, burst buffers put together in an ad hoc file system, and parallel file systems need data placed and retrieved with enough contextual meaning that different codes for these workflows can publish and subscribe to this data. Thus, self-describing data streams can be at the core of all these storage systems both for on-line processing and for eventual postprocessing during the scientific campaign.

7 Conclusion

Ad hoc file systems enable the usage of node-local SSDs in many application scenarios. The three presented file systems BeeOND, GekkoFS, and BurstFS can show only a small fraction of the possibilities of such ad hoc file systems. Nevertheless, they start from a production file system that transfers BeeGFS's very high client performance to BeeOND's dynamic setting, including distributed SSDs, and range to research file systems showing possible performance capabilities either when using local writes or when spreading metadata and data.

The research file systems also show that applications and usage scenarios have to be partially adapted to ad hoc file systems. Many commands such as `ls -a` or `mv` are not widespread within parallel applications, but completely abstaining from them would simplify the development of ad hoc file systems while at the same time significantly increasing performance. Codesigning applications with storage systems can therefore benefit both sides. The benefit of this codesign can even be improved if reusable components can be applied as building blocks to shorten the time to develop production-ready file systems.

This paper has shown that ad hoc file systems help use the additional storage layer of node-local SSDs and NVRAMs. Nevertheless, simply using these file systems without considering data stage-in and stage-out within the batch environment makes using them a manual and error-prone task. It is therefore necessary to extend the overall HPC framework to make the usage of ad hoc file systems automatic, still leaving plenty of room for new research directions in the scheduling domain, on codesign, component isolation and performance tuning.

Acknowledgments

We thank the team managing the Dagstuhl seminar series and all participants of the Dagstuhl Seminar 17202.

We also gratefully acknowledge the computing time granted on the supercomputer Mogon II at Johannes Gutenberg University Mainz.

This work has also been partially funded by the German Research Foundation (DFG) through the German Priority Programme 1648 "Software for Exascale Computing" (SPPEXA) and the ADA-FS project, which is gratefully acknowledged. We also gratefully acknowledge the funding by the European Union's Horizon 2020 research and innovation program under the NEXTGenIO project (grant 671591), the Spanish Ministry of Science and Innovation (contract TIN2015-65316), and the Generalitat de Catalunya (contract 2014-SGR-1051). This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was also supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357. LLNL-JRNL-779789.

References

[1] "Energy Exascale Earth System Model," https://e3sm.org.

[2] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic *et al.*, "ROOT - A C++ framework for petabyte data storage, statistical analysis and visualization," *Computer Physics Communications*, vol. 182, no. 6, pp. 1384–1385, 2011.

[3] I. T. Association *et al.*, "Infiniband architecture specification, volume 1, release 1.0," 2003.

[4] K. Banker, *MongoDB in Action*, 2nd ed. Manning, 2016.

[5] Barcelona Supercomputing Center, "MareNostrum IV – Technical Information." [Online]. Available: https://www.bsc.es/marenostrum/marenostrum/technical-information

[6] B. W. Barrett, R. Brightwell, S. Hemmert, K. Pedretti, K. Wheeler, K. Underwood, R. Riesen, A. B. Maccabe, and T. Hudson, "The portals 4.0 network programming interface," *Sandia National Laboratories, November 2012, Technical Report SAND2012-10087*, 2012.

[7] J. Bent, G. A. Gibson, G. Grider, B. McClelland, P. Nowoczynski, J. Nunez, M. Polte, and M. Wingate, "PLFS: a checkpoint filesystem for parallel applications," in *Proceedings of the ACM/IEEE Conference on High Performance Computing (SC), November 14-20, 2009, Portland, Oregon, USA*, 2009.

[8] A. Brinkmann, K. Mohror, and W. Yu, "Challenges and opportunities of user-level file systems for HPC (Dagstuhl Seminar 17202)," *Dagstuhl Reports*, vol. 7, no. 5, pp. 97–139, 2017.

[9] P. Carns, K. Harms, W. Allcock, C. Bacon, S. Lang, R. Latham, and R. Ross, "Understanding and improving computational science storage access through continuous characterization," *ACM Transactions on Storage (TOS)*, vol. 7, no. 3, p. 8, 2011.

[10] P. Carns, J. Jenkins, C. D. Cranor, S. Atchley, S. Seo, S. Snyder, and R. B. Ross, "Enabling NVM for data-intensive scientific services," in *4th Workshop on Interactions of NVM/Flash with Operating Systems and Workloads (INFLOW 16)*. Savannah, GA: USENIX Association, 2016. [Online].

Available: https://www.usenix.org/conference/inflow16/workshop-program/presentation/carns

[11] P. H. Carns, W. B. L. III, R. B. Ross, and R. Thakur, "PVFS: A parallel file system for linux clusters," in *4th Annual Linux Showcase & Conference 2000, Atlanta, Georgia, USA, October 10-14, 2000*, 2000.

[12] P. H. Carns, J. Jenkins, C. D. Cranor, S. Atchley, S. Seo, S. Snyder, and R. B. Ross, "Enabling NVM for data-intensive scientific services," in *4th Workshop on Interactions of NVM/Flash with Operating Systems and Workloads, INFLOW@OSDI 2016, Savannah, GA, USA, November 1, 2016.*, 2016.

[13] J. Carpenter and E. Hewitt, *Cassandra: The Definitive Guide*, 2nd ed. O'Reilly UK Ltd., 2016.

[14] D. Chen, N. A. Eisley, P. Heidelberger, R. M. Senger, Y. Sugawara, S. Kumar, V. Salapura, D. L. Satterfield, B. Steinmacher-Burow, and J. J. Parker, "The ibm blue gene/q interconnection network and message unit," in *SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2011, pp. 1–10.

[15] J. Conejero, S. Corella, R. M. Badia, and J. Labarta, "Task-based programming in compss to converge from HPC to big data," *International Journal of High Performance Computing Applications (IJHPCA)*, vol. 32, no. 1, pp. 45–60, 2018.

[16] G. Congiu, S. Narasimhamurthy, T. Süß, and A. Brinkmann, "Improving collective I/O performance using non-volatile memory devices," in *IEEE International Conference on Cluster Computing (CLUSTER), Taipei, Taiwan, September 12-16*, 2016, pp. 120–129.

[17] CRAY Inc., "libdatawarp - the DataWarp API." [Online]. Available: https://pubs.cray.com/content/S-2558/CLE%206.0.UP06/xctm-series-datawarptm-user-guide/libdatawarp---the-datawarp-api

[18] A. Davies and A. Orsaria, "Scale out with glusterfs," *Linux Journal*, vol. 2013, no. 235, Nov. 2013.

[19] B. Dong, S. Byna, K. Wu, H. Johansen, J. N. Johnson, N. Keen *et al.*, "Data elevator: Low-contention Data Movement in Hierarchical Storage Systems," in *2016 IEEE 23rd International Conference on High Performance Computing (HiPC)*. IEEE, 2016, pp. 152–161.

[20] S. Dong, M. Callaghan, L. Galanis, D. Borthakur, T. Savor, and M. Strum, "Optimizing space amplification in rocksdb," in *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*, 2017.

[21] M. Dorier, P. Carns, K. Harms, R. Latham, R. Ross, S. Snyder, J. Wozniak, S. Gutiérrez, B. Robey, B. Settlemyer, G. Shipman, J. Soumagne, J. Kowalkowski, M. Paterno, and S. Sehrish, "Methodology for the rapid development of scalable hpc data services," in *2018 IEEE/ACM 3rd International Workshop on Parallel Data Storage Data Intensive Scalable Computing Systems (PDSW-DISCS)*, Nov 2018, pp. 76–87.

[22] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[23] G. Faanes, A. Bataineh, D. Roweth, T. Court, E. Froese, B. Alverson, T. Johnson, J. Kopnick,

M. Higgins, and J. Reinhard, "Cray cascade: A scalable hpc system based on a dragonfly network," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12. Los Alamitos, CA, USA: IEEE Computer Society Press, 2012, pp. 103:1–103:9. [Online]. Available: http://dl.acm.org/citation.cfm?id=2388996.2389136

[24] K. Ferreira, R. Riesen, R. Oldfield, J. Stearley, J. Laros, K. Pedretti, R. Brightwell, and T. Kordenbrock, "Increasing fault resiliency in a message-passing environment," Sandia National Laboratories, Tech. Rep. SAND2009-6753, 2009.

[25] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson, "An overview of the hdf5 technology suite and its applications," in *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. ACM, 2011, pp. 36–47.

[26] G. C. Fox, J. Qiu, S. Jha, S. Ekanayake, and S. Kamburugamuve, "Big data, simulations and HPC convergence," in *Big Data Benchmarking - 6th International Workshop (WBDB) 2015, Toronto, ON, Canada, June 16-17, 2015 and 7th International Workshop (WBDB), New Delhi, India, December 14-15, 2015*, 2015, pp. 3–17.

[27] H. Greenberg, J. Bent, and G. Grider, "MDHIM: A parallel key/value framework for HPC," in *7th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage), Santa Clara, CA, USA, July 6-7, 2015*, 2015.

[28] W. D. Gropp, W. Gropp, E. Lusk, and A. Skjellum, *Using MPI: portable parallel programming with the message-passing interface*. MIT press, 1999, vol. 1.

[29] P. Grun, S. Hefty, S. Sur, D. Goodell, R. D. Russell, H. Pritchard, and J. M. Squyres, "A brief introduction to the openfabrics interfaces - a new network api for maximizing high performance application efficiency," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, Aug 2015, pp. 34–39.

[30] V. Henson, A. van de Ven, A. Gud, and Z. Brown, "Chunkfs: Using divide-and-conquer to improve file system reliability and repair," in *Proceedings of the Second Workshop on Hot Topics in System Dependability (HotDep), Seattle, WA, USA, November 8*, 2006.

[31] M. E. Hoskins, "Sshfs: Super easy file access over ssh," *Linux Journal*, vol. 2006, no. 146, Jun. 2006.

[32] IBM, "IBM/CAST - Cluster Administration and Storage Tools." [Online]. Available: https://github.com/IBM/CAST

[33] T. Z. Islam, K. Mohror, S. Bagchi, A. Moody, B. R. de Supinski, and R. Eigenmann, "Mcrengine: a scalable checkpointing system using data-aware aggregation and compression," in *Conference on High Performance Computing Networking, Storage and Analysis (SC), Salt Lake City, UT, USA - November 11 – 15*, 2012.

[34] J. C. Jacob, D. S. Katz, G. B. Berriman, J. Good, A. C. Laity, E. Deelman, C. Kesselman, G. Singh, M. Su, T. A. Prince, and R. Williams, "Montage: a grid portal and software toolkit for science-grade astronomical image mosaicking," *International Journal of Computational Science and Engineering (IJCSE)*, vol. 4, no. 2, pp. 73–87, 2009.

[35] J. Kaiser, R. Gad, T. Süß, F. Padua, L. Nagel, and A. Brinkmann, "Deduplication potential of HPC

applications' checkpoints," in *IEEE International Conference on Cluster Computing (CLUSTER), Taipei, Taiwan, September 12-16*, 2016, pp. 413–422.

[36] D. R. Karger, E. Lehman, F. T. Leighton, R. Panigrahy, M. S. Levine, and D. Lewin, "Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web," in *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing (STOC), El Paso, Texas, USA, May 4-6, 1997*, 1997, pp. 654–663.

[37] S. M. Kelly and R. Brightwell, "Software architecture of the light weight kernel, catamount," in *Proceedings of the 2005 Cray User Group Annual Technical Conference*, 2005, pp. 16–19.

[38] M. Kleppmann, *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly UK Ltd., 2017.

[39] J. Köster and S. Rahmann, "Snakemake - a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 34, no. 20, p. 3600, 2018.

[40] A. Kougkas, H. Devarajan, X.-H. Sun, and J. Lofstead, "Harmonia: An interference-aware dynamic i/o scheduler for shared non-volatile burst buffers," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, 2018, pp. 290–301.

[41] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, Prabhat, and M. Houston, "Exascale Deep Learning for Climate Analytics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, ser. SC '18, 2018.

[42] R. Latham, R. Ross, and R. Thakur, "Can MPI be used for persistent parallel services?" *Lecture Notes in Computer Science*, vol. 4192, pp. 275–284, September 2006. [Online]. Available: http://www.springerlink.com/content/u5768x256v818u1p/

[43] Lawrence Livermoore National Lab, "Sierra." [Online]. Available: https://hpc.llnl.gov/hardware/platforms/sierra

[44] P. H. Lensing, T. Cortes, and A. Brinkmann, "Direct lookup and hash-based metadata placement for local file systems," in *6th Annual International Systems and Storage Conference (SYSTOR), Haifa, Israel - June 30 - July 02*, 2013.

[45] J. Li, W.-k. Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, and M. Zingale, "Parallel netcdf: A high-performance scientific i/o interface," in *SC'03: Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. IEEE, 2003, pp. 39–39.

[46] N. Liu, J. Cope, P. H. Carns, C. D. Carothers, R. B. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the role of burst buffers in leadership-class storage systems," in *IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST), April 16-20, 2012, Asilomar Conference Grounds, Pacific Grove, CA, USA*, 2012, pp. 1–11.

[47] J. Lofstead, F. Zheng, Q. Liu, S. Klasky, R. Oldfield, T. Kordenbrock, K. Schwan, and M. Wolf, "Managing variability in the IO performance of petascale storage systems," in *SC'10: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, 2010, pp. 1–12.

[48] J. F. Lofstead, S. Klasky, K. Schwan, N. Podhorszki, and C. Jin, "Flexible IO and integration for scientific codes through the adaptable IO system (ADIOS)," in *6th International Workshop on Challenges of Large Applications in Distributed Environments, CLADE@HPDC 2008, Boston, MA, USA, June 23, 2008*, 2008, pp. 15–24.

[49] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "A large-scale study of flash memory failures in the field," in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Portland, OR, USA, June 15-19*, 2015, pp. 177–190.

[50] A. Miranda, A. Jackson, T. Tocci, I. Panourgias, and R. Nou, "NORNS: Extending Slurm to Support Data-Driven Workflows through Asynchronous Data Staging," in *2019 IEEE International Conference on Cluster Computing*. IEEE, 2019.

[51] A. Moody, G. Bronevetsky, K. Mohror, and B. R. de Supinski, "Design, modeling, and evaluation of a scalable multi-level checkpointing system," in *Conference on High Performance Computing Networking, Storage and Analysis (SC), New Orleans, LA, USA, November 13-19, 2010*, 2010.

[52] M. Moore, D. Bonnie, B. Ligon, M. Marshall, W. Ligon, N. Mills, E. Quarles, S. Sampson, S. Yang, and B. Wilson, "Orangefs: Advancing pvfs," in *9th USENIX Conference on File and Storage Technologies (FAST), Poster Session, San Jose, CA, USA, February 15-17*, 2011.

[53] J. Nakashima and K. Taura, "Massivethreads: A thread library for high productivity languages," in *Concurrent Objects and Beyond*. Springer, 2014, pp. 222–238.

[54] I. Narayanan, D. Wang, M. Jeon, B. Sharma, L. Caulfield, A. Sivasubramaniam, B. Cutler, J. Liu, B. M. Khessib, and K. Vaid, "SSD failures in datacenters: What? when? and why?" in *Proceedings of the 9th ACM International on Systems and Storage Conference (SYSTOR), Haifa, Israel, June 6-8*, 2016, pp. 7:1–7:11.

[55] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'big data', hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[56] R. A. Oldfield, P. Widener, A. B. Maccabe, L. Ward, and T. Kordenbrock, "Efficient data-movement for lightweight i/o," in *2006 IEEE International Conference on Cluster Computing*, Sep. 2006, pp. 1–9.

[57] S. Oral, D. A. Dillow, D. Fuller, J. Hill, D. Leverman, S. S. Vazhkudai, F. Wang, Y. K. , J. Rogers, J. James Simmons, and R. Miller, "Olcf's 1 tb/s, next-generation lustre file system," in *Proceedings of Cray User Group Conference (CUG 2013)*, 2013.

[58] F. Petrini, "Scaling to thousands of processors with buffer coscheduling," in *Scaling to New Height Workshop*, 2002.

[59] I. R. Philp, "Software failures and the road to a petaflop machine," in *Proceedings of the 1st Workshop on High Performance Computing Reliability Issues (HPCRI)*, 2005.

[60] D. Poliakoff and M. Legendre, "GotCha," https://github.com/LLNL/GOTCHA.

[61] Y. Qian, X. Li, S. Ihara, A. Dilger, C. Thomaz, S. Wang, W. Cheng, C. Li, L. Zeng, F. Wang, D. Feng, T. Süß, and A. Brinkmann, "Lpcc: Hierarchical persistent client caching for lustre," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Denver, CO, USA, November 17 - 22*, 2019.

[62] Y. Qian, X. Li, S. Ihara, L. Zeng, J. Kaiser, T. Süß, and A. Brinkmann, "A configurable rule based classful token bucket filter network request scheduler for the lustre file system," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Denver, CO, USA, November 12 - 17*, 2017, pp. 6:1–6:12.

[63] R. Rajachandrasekar, A. Moody, K. Mohror, and D. K. Panda, "A 1 pb/s file system to checkpoint three million MPI tasks," in *The 22nd International Symposium on High-Performance Parallel and Distributed Computing (HPDC), New York, NY, USA - June 17 - 21*, 2013, pp. 143–154.

[64] A. Rajgarhia and A. Gehani, "Performance and extension of user space file systems," in *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre, Switzerland, March 22-26, 2010*, 2010, pp. 206–213.

[65] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE computer graphics and applications*, vol. 10, no. 4, pp. 76–82, 1990.

[66] R. Ross, L. Ward, P. Carns, G. Grider, S. Klasky, Q. Koziol, G. K. Lockwood, K. Mohror, B. Settlemyer, and M. Wolf, "Storage Systems and I/O: Organizing, Storing, and Accessing Data for Scientific Discovery," https://www.osti.gov/biblio/1491994/, 5 2019.

[67] C. Ruemmler and J. Wilkes, "An introduction to disk drive modeling," *IEEE Computer*, vol. 27, no. 3, pp. 17–28, 1994.

[68] B. Schroeder, R. Lagisetty, and A. Merchant, "Flash reliability in production: The expected and the unexpected," in *14th USENIX Conference on File and Storage Technologies (FAST), Santa Clara, CA, USA, February 22-25*, 2016, pp. 67–80.

[69] S. Seo, A. Amer, P. Balaji, C. Bordage, G. Bosilca, A. Brooks, P. Carns, A. Castelló, D. Genet, T. Herault *et al.*, "Argobots: A lightweight low-level threading and tasking framework," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 3, pp. 512–526, 2017.

[70] P. Shamis, M. G. Venkata, M. G. Lopez, M. B. Baker, O. Hernandez, Y. Itigin, M. Dubman, G. Shainer, R. L. Graham, L. Liss, Y. Shahar, S. Potluri, D. Rossetti, D. Becker, D. Poole, C. Lamb, S. Kumar, C. Stunkel, G. Bosilca, and A. Bouteiller, "Ucx: An open source framework for hpc network apis and beyond," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, Aug 2015, pp. 40–43.

[71] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Lake Tahoe, Nevada, USA, May 3-7*, 2010, pp. 1–10.

[72] J. Soumagne, D. Kimpe, J. Zounmevo, M. Chaarawi, Q. Koziol, A. Afsahi, and R. Ross, "Mercury: Enabling remote procedure call for

high-performance computing," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2013, pp. 1–8.

[73] J. Soumagne, D. Kimpe, J. A. Zounmevo, M. Chaarawi, Q. Koziol, A. Afsahi, and R. B. Ross, "Mercury: Enabling remote procedure call for high-performance computing," in *2013 IEEE International Conference on Cluster Computing (CLUSTER), Indianapolis, IN, USA, September 23-27, 2013*, 2013, pp. 1–8.

[74] M. Soysal, M. Berghoff, D. Klusácek, and A. Streit, "On the quality of wall time estimates for resource allocation prediction," in *48th International Conference on Parallel Processing (ICPP) 2019 Workshop Proceedings, Kyoto, Japan, August 05-08, 2019.*, 2019, pp. 23:1–23:8.

[75] P. Subedi, P. Davis, S. Duan, S. Klasky, H. Kolla, and M. Parashar, "Stacker: An Autonomic Data Movement Engine for Extreme-scale Data Staging-based in-situ Workflows," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. IEEE Press, 2018, p. 73.

[76] S. Thapaliya, P. Bangalore, J. Lofstead, K. Mohror, and A. Moody, "Managing i/o interference in a shared burst buffer system," in *2016 45th International Conference on Parallel Processing (ICPP)*, 2016, pp. 416–425.

[77] B. K. R. Vangoor, V. Tarasov, and E. Zadok, "To FUSE or not to FUSE: performance of user-space file systems," in *15th USENIX Conference on File and Storage Technologies (FAST), Santa Clara, CA, USA, February 27 - March 2, 2017*, 2017, pp. 59–72.

[78] M.-A. Vef, N. Moti, T. Süß, T. Tocci, R. Nou, A. Miranda, T. Cortes, and A. Brinkmann, "Gekkofs - a temporary distributed file system for hpc applications," in *Proceedings of the 2018 IEEE International Conference on Cluster Computing (CLUSTER), Belfast, UK, September 10-13*, 2018.

[79] M.-A. Vef, V. Tarasov, D. Hildebrand, and A. Brinkmann, "Challenges and solutions for tracing storage systems: A case study with spectrum scale," *ACM Trans. Storage*, vol. 14, no. 2, pp. 18:1–18:24, 2018.

[80] H. Volos, S. N. and Sankaralingam Panneerselvam and Venkatanathan Varadarajan and Prashant Saxena, and M. M. Swift, "Aerie: flexible file-system interfaces to storage-class memory," in *Ninth Eurosys Conference 2014 (EuroSys), Amsterdam, The Netherlands, April 13-16, 2014*, 2014, pp. 14:1–14:14.

[81] T. Wang, K. Mohror, A. Moody, K. Sato, and W. Yu, "An ephemeral burst-buffer file system for scientific applications," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Salt Lake City, UT, USA, November 13-18*, 2016, pp. 807–818.

[82] T. Wang, S. Oral, M. Pritchard, B. Wang, and W. Yu, "Trio: Burst Buffer Based I/O Orchestration," in *2015 IEEE International Conference on Cluster Computing*. IEEE, 2015, pp. 194–203.

[83] M. Wasi-ur-Rahman, X. Lu, N. S. Islam, R. Rajachandrasekar, and D. K. Panda, "High-performance design of YARN mapreduce on modern HPC clusters with lustre and RDMA," in *2015*

IEEE International Parallel and Distributed Processing Symposium (IPDPS), Hyderabad, India, May 25-29, 2015, pp. 291–300.

[84] B. Welch and G. Noer, "Optimizing a hybrid SSD/HDD HPC storage system based on file size distributions," in *IEEE 29th Symposium on Mass Storage Systems and Technologies (MSST), May 6-10, 2013, Long Beach, CA, USA*, 2013, pp. 1–12.

[85] K. B. Wheeler, R. C. Murphy, and D. Thain, "Qthreads: An api for programming with millions of lightweight threads," in *2008 IEEE International Symposium on Parallel and Distributed Processing*. IEEE, 2008, pp. 1–8.

[86] J. L. Whitt, "Oak Ridge Leadership Computing Facility: Summit and Beyond," 3 2017. [Online]. Available: https://indico.cern.ch/event/618513/contributions/2527318/attachments/1437236/2210560/SummitProjectOverview_jlw.pdf

[87] B. Xie, J. Chase, D. Dillow, O. Drokin, S. Klasky, S. Oral, and N. Podhorszki, "Characterizing output bottlenecks in a supercomputer," in *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2012, pp. 1–11.

[88] O. Yildiz, M. Dorier, S. Ibrahim, R. Ross, and G. Antoniu, "On the root causes of cross-application I/O interference in HPC storage systems," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2016, pp. 750–759.

[89] M. Zaharia, R. S. Xin, P. Wendell, T. Das *et al.*, "Apache spark: a unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[90] Z. Zhang, K. Barbary, F. A. Nothaft, E. R. Sparks, O. Z. and Michael J. Franklin, D. A. Patterson, and S. Perlmutter, "Scientific computing meets big data technology: An astronomy use case," in *2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, October 29 - November 1*, 2015, pp. 918–927.

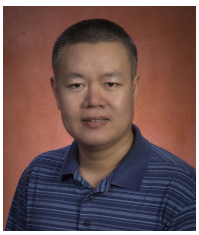
[91] D. Zhao, Z. Zhang, X. Zhou, T. Li, K. Wang, D. Kimpe, P. H. Carns, R. B. Ross, and I. Raicu, "Fusionfs: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems," in *2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, October 27-30, 2014*, 2014, pp. 61–70.

[92] Q. Zheng, C. D. Cranor, D. Guo, G. R. Ganger, G. Amvrosiadis, G. A. Gibson, B. W. Settlemyer, G. Grider, and F. Guo, "Scaling embedded in-situ indexing with deltafs," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) , Dallas, TX, USA, November 11-16*, 2018.

[93] Y. Zhu, F. Chowdhury, H. Fu, A. Moody, K. Mohror, K. Sato, and W. Yu, "Entropy-aware I/O pipelining for large-scale deep learning on HPC systems," in *26th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Milwaukee, WI, USA, September 25-28, 2018*, 2018, pp. 145–156.

André Brinkmann is a full professor at the computer science department of JGU and head of the university's data center ZDV (since 2011). He received his Ph.D. in electrical engineering in 2004 from the Paderborn University and has been an assistant professor in the computer science department of the Paderborn University from 2008 to 2011. Furthermore, he has been the managing director of the Paderborn Centre for Parallel Computing PC^2 during this time frame. His research interests focus on the application of algorithm engineering techniques in the area of data centre management, cloud computing, and storage systems.

Kathryn Mohror is the group leader for the Data Analysis Group at the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory (LLNL). Kathryn's research on high-end computing systems is currently focused on scalable fault tolerant computing and I/O for extreme scale systems. Her other research interests include scalable performance analysis and tuning, and parallel programming paradigms. Kathryn has been working at LLNL since 2010.



Weikuan Yu is a Professor in the Department of Computer Science at Florida State University. He received his PhD degree in Computer Science and master's degree in Neurobiology from the Ohio State University. He also holds a Bachelor degree in Genetics from Wuhan University, China. Yu's main research interests include big data management and analytics frameworks, parallel I/O and storage, GPU memory architecture, and high performance networking. Yu's research has won the 2012 Alabama Innovation Award and the First Prize of 2012 ACM Student Research Competition Grand Finals. He is a senior member of IEEE and life member of ACM.

Philip Carns is a principal software development specialist in the Mathematics and Computer Science Division of Argonne National Laboratory. He is also an adjunct associate professor of electrical and computer engineering at Clemson University and a fellow of the Northwestern-Argonne Institute for Science and Engineering. His research interests include characterization, modeling, and development of storage systems for data-intensive scientific computing.

Toni Cortes is an associate professor at Universitat Politècnica de Catalunya (since 1998) and researcher at the Barcelona Supercomputing Center. He received his M.S. in computer science in 1992 and his Ph.D. also in computer science in 1997 (both at Universitat Politècnica de Catalunya). Currently he develops his research at the Barcelona Supercomputing Center, where he acted as the leader of the Storage Systems Research Group from 2006 until 2019. His research concentrates in storage systems, programming models for scalable distributed systems and operating systems. He is also editor of the Cluster Computing Journal and served as the coordinator of the SSI task in the IEEE TCSS. He has also served in many international conference program committees and/or organizing committees and was general chair for the Cluster 2006 and 2021 conference, LaSCo 2008, XtreemOS summit 2009, and SNAPI 2010. He is also served as the chair of the steering committee for the Cluster conference series (2011-2014). His involvement in IEEE CS has been awarded by the "Certificate of appreciation" in 2007.

Scott A. Klasky is a distinguished scientist and the group leader for Scientific Data in the Computer Science and Mathematics Division at the Oak Ridge National Laboratory. He holds an appointment at the University of Tennessee, and Georgia Tech University. He obtained his Ph.D. in Physics from the University of Texas at Austin (1994). Dr. Klasky is a world expert in scientific computing and scientific data management, co- authoring over 200 paper.

Alberto Miranda is a Senior Researcher in advanced storage systems in the Computer Science Department of the Barcelona Supercomputing Center (BSC) and co-leader of the Storage Systems for Extreme Computing research group since January 2019. His research interests include scalable storage technologies, architectures for distributed systems, operating system internals, and high performance networking. He received a Ph.D. Cum Laude in Computer Science from the Technical University of Catalonia (UPC) in 2014, and has been working at BSC since 2007.

Franz-Josef Pfreundt is the Director of the Competence Center for HPC & Visualization at Fraunhofer ITWM since 1999. He studied Mathematics, Physics and Computer Science resulting in a Diploma in Mathematics and a Ph.D degree in Mathematical Physics (1986). In 2001 the prestigious Fraunhofer Research Prize was awarded to Franz-Josef Pfreundt, Konrad Steiner and their research group for their work on microstructure simulation. The developments in the area of visualization and implementations on IBM Cell Processor gained the Fraunhofer Research Price in 2005 and the IBM faculty award in 2006. His main research focus are parallel file systems and new parallel programming appraoches.

Robert B. Ross is a Senior Computer Scientist at Argonne National Laboratory and the Director of the DOE SciDAC RAPIDS Institute for Computer Science and Data. Rob's research interests are in system software for high performance computing systems, in particular distributed storage systems and libraries for I/O and message passing. Rob received his Ph.D. in Computer Engineering from Clemson University in 2000. Rob was a recipient of the 2004 Presidential Early Career Award for Scientists and Engineers.

Marc-André Vef is a third-year Ph.D. candidate in André Brinkmann's research team at the Johannes Gutenberg University Mainz. He started his Ph.D. in 2016 after receiving his B.Sc. and M.Sc. degrees in computer science from the Johannes Gutenberg University Mainz. His master thesis was in cooperation with IBM Research about analyzing file create performance in the IBM Spectrum Scale parallel file system (formerly GPFS). Marc's research interests focus on parallel and ad-hoc file systems and system analytics.