



## Data Management Services and Storage

A. Johansson<sup>a\*a</sup>, C. Piechurski<sup>b\*b</sup>, D. Pleiter<sup>c\*c</sup>, K. Wadówka<sup>d\*d</sup>

<sup>a</sup>Swedish National Infrastructure for Computing, <sup>b</sup>GENCI, <sup>c</sup>Forschungszentrum Jülich GmbH,

<sup>d</sup>Poznań Supercomputing and Networking Center

---

### Abstract

HPC storage systems have evolved from fairly simple systems attached to a single cluster to site-wide complex infrastructures supporting migration of data between tiers of different performance characteristics and I/O acceleration. Exascale systems and AI workloads will continue this trend by placing even greater demands on speedy access to data. This report summarises the history of this development and examines some of the technologies that are building blocks of near-future storage systems, both hardware and the software required to manage the large amounts of data.

---

---

<sup>a</sup> andjo@nsc.liu.se

<sup>b</sup> christelle.piechurski@genci.fr

<sup>c</sup> d.pleiter@fz-juelich.de

<sup>d</sup> kwadowka@man.poznan.pl

## Table of contents

Abstract.....	1
1. Introduction .....	3
2. Storage Infrastructure .....	3
2.1. Evolution of Data Storage Systems.....	3
2.2. High-Performance Storage.....	5
2.3. Storage Systems Solutions for the First Announced HPC Exascale Systems.....	7
2.4. Large Capacity Storage .....	7
3. File Systems.....	9
3.1. Lustre .....	9
3.2. Spectrum Scale (formerly GPFS) .....	10
3.3. BeeGFS.....	11
3.4. Ceph.....	12
3.5. File System Feature Comparison.....	13
4. Data Management Services .....	13
4.1. Cloud Storage Interfaces .....	13
4.2. Data Lake and ad-hoc Storage Systems .....	14
4.3. Tiered Storage Management.....	14
4.4. I/O Acceleration Solutions.....	15
4.5. Data Management Solutions .....	16
5. Trends.....	17
6. Conclusion.....	17
References.....	18
List of acronyms.....	20
Acknowledgements .....	21

## 1. Introduction

This technical report is part of a series of reports published in the Work Package “HPC Planning and Commissioning” (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups “Infrastructure Advisory Group” (INFRAG) and “Research & Innovation Advisory Group” (RIAG) and other European HPC sites. Further reports published so far cover “State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI” [1] and “Edge Computing: An Overview of Framework and Applications” [2]. The series will be continued in 2021 with further selected highly topical subjects.

HPC storage systems have evolved from fairly simple systems attached to a single cluster to site-wide complex infrastructures supporting migration of data between storage tiers with different performance characteristics and I/O acceleration. Exascale systems and AI workloads will continue this trend by placing even greater demands on speedy access to data.

Topics covered here are related to storage infrastructures and management of data stored on these infrastructures. Hardware and software are discussed due to the need for making cost trade-offs that influence both. Well managed data workflows are a concern if data is expected to be available over 20+ years and thus far beyond the lifetime of a single compute resource with attached storage system. Much data is only used during the computation and there the concerns are mainly high bandwidth and low latency, but results and data sets required to reproduce results require more long-term infrastructure.

## 2. Storage Infrastructure

This section explores a number of storage system options from a systems perspective, examining hardware interfaces and protocols.

### 2.1. Evolution of Data Storage Systems

The amount of data created in our society is growing rapidly, and even experts struggle to predict the growth rate. In response, there is an insatiable need for more advanced high-performance storage systems to store such amounts of data appropriately taking both cost and performance factors into account.

When discussing storage systems, the concept of storage tiers is often used. Different tiers have different trade-offs between volume cost and performance since the access pattern for most data varies during the lifetime of the dataset. In this report the following tiers will be used.

Level 1	Short-term data, also known as scratch or work storage, often referred to as “warm”
Level 2	Medium-term data, also known as project storage
Level 3	Long-term data, also known as archival storage, less frequently accessed data that is often referred to as “cold”

Ten years ago, the largest HPC data storage systems contained only a few petabytes (around 10 PB) of disk space based on traditional magnetic hard disk drives (HDD) for short-term data supported by an archive or Hierarchical Storage Management (HSM) system, both relying mainly on tape libraries to store cold data. Today, the equivalent systems are supporting an order of magnitude (around 20 times) of the capacity for short-term system storage, with additional storage tier levels to store data in an efficient and cost-optimised way depending on data access or criticality. See Figure 1 for an overview of the current storage hierarchy.

Generally, this ratio can also be applied to any common industrial and academic research computing systems, which can support up to several petabytes of storage today. Some research laboratories are able to create petabytes of data from their own scientific instruments. As an example, the large-scale scientific radio telescopes that form the Square Kilometre Array (SKA), built to explore the Universe, will generate Exabytes of scientific data per year. The raw SKA data will be filtered and post-processed using HPC resources, but the project still expects to

archive 600 Petabytes of data per year [3]. Also, the volume of generated computational data expands dramatically with increasing computing capabilities.

Flash memory technology has evolved to become part of mainstream high-performance storage devices. In the HPC context these are non-volatile disk drive devices, not to be confused with slow memory cards or small embedded memories for storing system configuration data. This technology is the basis for current low latency and high bandwidth devices such as a Solid-State Drive (SSD).

Today's storage needs will be further increased by the upcoming Exascale systems. The performance capabilities of these systems will also increase, entailing new storage requirements in term of capacity, throughput and I/O operations per second (IOPS) relative to the data profile of the workload. While bandwidth and storage capacity are still the dominant performance indicators when choosing a storage system, IOPS are now an important criterion to consider when analysing HPC storage needs. The data handling requirements of Exascale, Big Data and Artificial Intelligence systems are very high for both IOPS and volume. Level 1 storage systems based on SSDs are relatively small in capacity for cost reasons but capable of handling high throughput of IOPS intensive data. Protocol limitations on how the storage devices are interfaced to systems drove the creation of the Non-Volatile Memory Express (NVMe) protocol, a new communication standard between the processor and data storage. NVMe is a communication interface and driver specification that defines a set of commands and a set of functions for PCIe-based SSDs to increase IOPS performance and interoperability across a wide range of corporate and client systems. Due to its low latency and direct PCIe connection allowing tighter coupling, it can handle specific workloads more efficiently than large shared storage systems that support all types of workloads but in a more general manner.

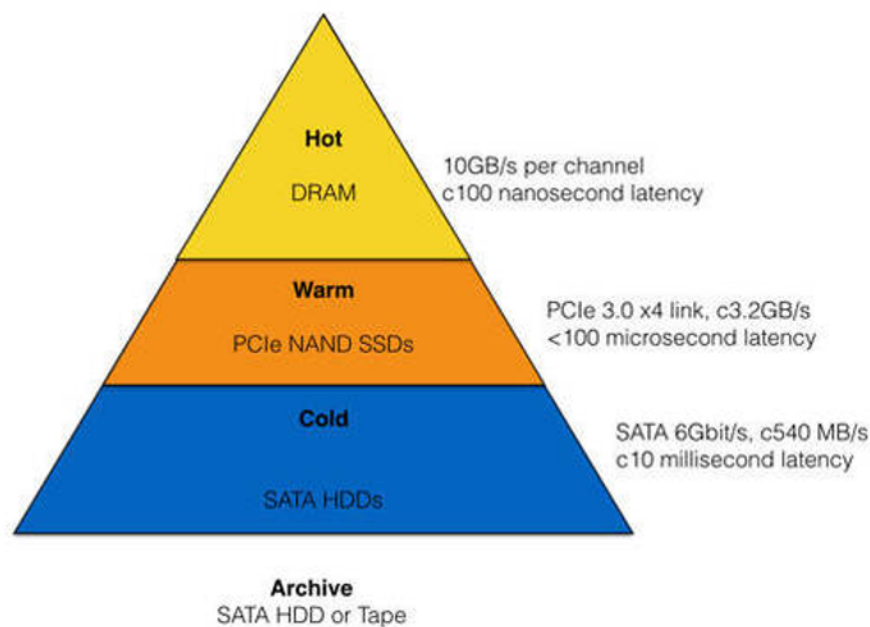


Figure 1: Storage and Memory Hierarchy today [4]

Compared to high-performance storage solutions for warm data, cold data is usually stored in less expensive storage environments with lower IOPS and bandwidth. Tape systems have been and are still a popular storage medium for cold data. Linear Tape-Open (LTO) was originally developed in the late 1990s as a low-cost vendor neutral storage option while some vendors (ex. IBM with their Jaguar tape drives) also market proprietary formats. See Section 2.4.2 for a more in-depth look at tape technologies.

Systems based entirely on flash memories are ideal for warm data where the trade-off favours IOPS and bandwidth above capacity. Large volume storage and data retention policies create an environment more in favour of low-cost capacity for cold data. Operational costs are also lower if data can be stored on powered down storage when the latency of bringing it online is acceptable. For security reasons the use of totally offline media may also be mandated.

Until recently, the cold storage system was mostly provided by tape libraries which guarantee low power consumption, large capacity and average read/write speeds while incurring longer waiting times for data. Volume storage systems based on HDDs with SATA interfaces are starting to compete with tape libraries for cold storage due to them both becoming capable of spinning down drives to reduce power consumption and still retain their random-access nature. The time it takes to spin up a drive is much shorter than the mount and spool time for a tape cartridge, so for data that is intermittently used rather than archived this can justify the higher cost of HDD storage.

## 2.2. High-Performance Storage

In the Exascale computing era, data processing capabilities are becoming a key factor. Standard approaches used by traditional I/O solutions are increasingly becoming a bottleneck. While new technologies such as data caching can help solve performance issues at a higher cost per terabyte, it is important to consider implementing another layer of hierarchical storage management into HPC architectures using ultra-high-speed storage technologies to support the current and futures challenges. This section examines a few of these technologies.

### 2.2.1. 3D-NAND and V-NAND Memory

Typical NAND memory chips (SLC, MLC, TLC) are built in such a way that all cells that store data are in one plane (type (a) in Figure 2). To increase the capacity of the modules, the cells must be placed more densely. This can be done by reducing the space between them, but it cannot be done indefinitely. At some point the density of the cells will be so high that the electric charges stored in them will leak between them. The result will be data corruption or irretrievable loss. The solution to this problem is 3D-NAND and V-NAND modules where memory cells are stacked in layers. This technique not only allows to increase the capacity of the media, but also has a positive effect on their efficiency and is not associated with higher production costs. For this reason, 3D-NAND can be considered a major breakthrough for flash technologies. With the new layered style of memory more and more data can be stored within the same physical area.

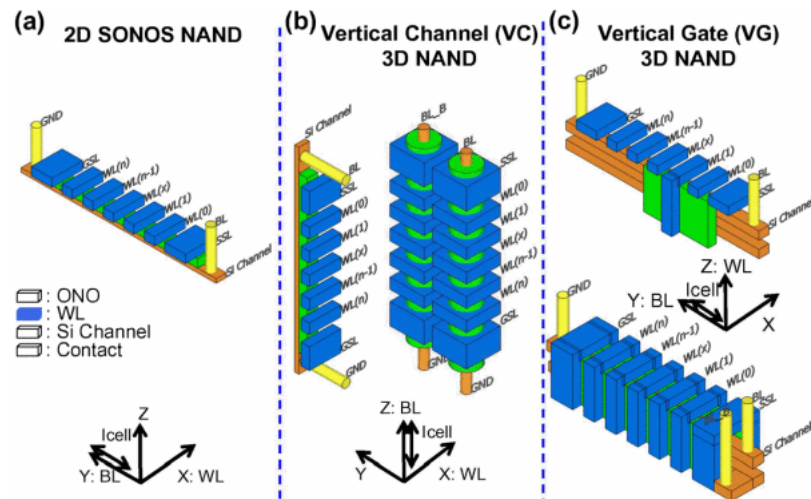


Figure 2: Schematic structure of some NAND flash types [5]

### 2.2.2. Intel Optane Memory

Intel Optane memory is not typical disk drive or DRAM computer memory (Figure 3), but a proprietary Intel standard. This is not a technology used for conventional storage. Instead, the M.2 form factor Optane module is a large cache bridge between the volatile DRAM and non-volatile storage, capable of storing a larger amount of data than traditional DRAM (but in a persistent way) and enabling faster data transfer between memory, storage and processor. Given proper OS support this additional layer speeds up most end user operations, using caching software that stores relevant data on an Optane drive for almost instant recall. Intel also uses the Optane name for smaller high-endurance low-latency SSDs with U.2 form factor.

The idea of using a small amount of super-fast flash memory to increase the performance of a basic memory disk is nothing new. In fact, Optane is essentially the next generation of Intel's Smart Response Technology (SRT) that

can use low-capacity, expensive SSDs (compared to HDDs) to cache data for slower, conventional high-capacity hard drives. The difference is that Optane uses memory produced and sold by Intel in conjunction with special hardware and software components on compatible motherboards. Optane memory works with all kinds of RAM modules, storage drives and graphics cards that match a compatible motherboard.

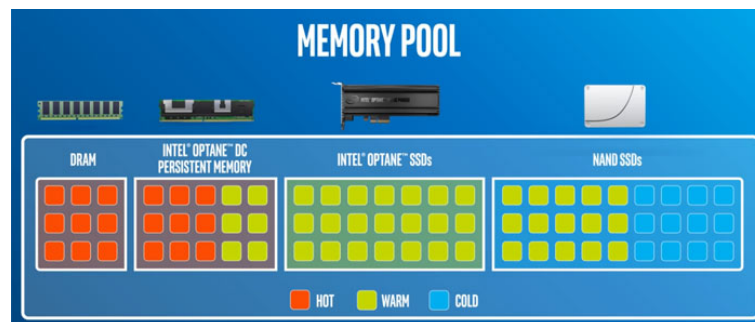


Figure 3: Optane Memory Pool Example [6]

### 2.2.3. NVMe Memory

NVMe memory eliminates much of the software overhead between applications and storage and is optimised for interfacing flash storage solutions to host systems, thus significantly reducing latency and increasing system and application performance. This is a vendor neutral standard for connecting storage to host systems, and while for example Optane described in Section 2.2.2 uses NVMe for its physical layer it uses proprietary software in addition to NVMe.

The NVMe specification has been designed for flash memory. I/O tasks performed with NVMe drivers have lower latency and more IOPS relative to older storage models using standards such as AHCI (Advanced Host Controller Interface) used with SATA SSD drives. Since the specification was designed specifically for flash storage, NVMe is becoming the new industry standard for data centre systems that require massive bandwidth and short access times. It is available in several form factors, with PCI-e and M.2 being the most popular. The M.2 form factor used by many NVMe devices allows high capacity in a small storage enclosure, such as a cache device, and are ideal for systems where the physical size of the device is a limiting factor. A disadvantage of NVMe is the use of flash memory which is more expensive than HDDs with respect to capacity, lacking legacy support that prevents it from upgrading older storage systems, and the generally lower capacity of flash drives compared to traditional disk drives.

NVMe memory is very well suited for IOPS intensive applications with very heavy workloads that require ultra-low latency.

### 2.2.4. Solid State Drives

While most storage systems based on HDD support large transactions (large file & large block transfer) well, an increasing number of those systems are facing I/O performance bottlenecks as they are less capable of absorbing low latency transactions in the same way. So, mixing both type of transactions, throughput and IOPS oriented is somewhat disturbing storage system performances. Therefore, operators are considering using SSD based storage system mainly for IOPS driven workloads on a low storage capacity, also being cost driven (min. 20 times higher per PB built in a high-performance oriented way) to increase overall storage system efficiency and reliability and reduce overall maintenance costs. On the other hand, it is unclear how operators can control the lifetime on an SSD based on DWPD (Drive Writes Per Day) set by device manufacturers. SSDs are manufactured in such a way that they can easily be implemented as replacements or additions to hard disks (HDDs) equipped with rotating magnetic plates. They are available in a variety of form factors, including standard 3.5- and 2.5-inch drive sizes, and support various communication protocols / interfaces. Devices can be directly attached using Serial ATA (SATA), Serial Attached SCSI (SAS) and recently PCIe (NVMe standard) to enable data transfer to and from server processors.

## 2.3. Storage Systems Solutions for the First Announced HPC Exascale Systems

Exascale storage requirements are no longer driven by traditional workloads with large streaming writes like checkpoint/restart but is increasingly driven by complex I/O patterns from new types of applications. High-performance data analytics workloads are generating vast quantities of random reads and writes. Artificial Intelligence workloads are reading far more than traditional high-performance computing workloads. Data streaming from instruments into an HPC cluster require better quality of service to avoid data loss. Data access time is now becoming as critical as write bandwidth. New storage semantics are required to query, analyse, filter, and transform datasets. A single storage platform in which next generation workflows combine HPC, Big Data, and AI to exchange data and communicate is essential.

The three planned DOE Exascale storage systems will rely on the Lustre file system. On top of Lustre, additional layers can be found including one based on the Intel Distributed Asynchronous Object Storage (DAOS) product which is optimised for non-volatile memory (NVM) technologies such as Optane persistent memory and Optane SSDs. DAOS is the foundation of the DOE Exascale storage stack and is an open-source software stack for scale out object storage that provides high bandwidth, low latency, and high I/O operations per second (IOPS) storage containers to HPC applications. It enables next generation data centric workflows that combine simulation, data analytics, and AI.

The DAOS instances of the Aurora system at Argonne National Lab, scheduled for delivery in 2022 and targeted as an Exascale system, is planned to feature a bandwidth of more than 25 TB/s and a capacity of 230 PB [7].

The EuroHPC pre-Exascale system LUMI will be using Lustre for both the all-flash warm data storage (7 PB) and the capacity storage (80 PB). In addition to this more traditional cluster storage LUMI will use Ceph storage for a data management service (30 PB). All these components will be connected to the cluster interconnect.

## 2.4. Large Capacity Storage

### 2.4.1. Hard Disk Drives and Hybrid Systems

While the popularity of SSDs has been growing for several years, the near future of the disk array market probably belongs to hybrid systems that support both HDDs and SSDs. While SSDs are characterised by higher performance and shorter access time than HDDs, their weakness is their lower capacity and corresponding higher price per TB compared to HDDs. For this reason, solid state drives are still not suitable for mid-price class storage solutions.

Automatic tiering is the most popular technology used in hybrid arrays. The device stores the most-used data on the fastest tier (ex. SSD) and migrates data to the slowest media tier (ex. HDDs) based on policy rules (age, access frequency, etc.) defined in the storage system. This method enables performance improvements on a global storage system level, where for example data accessed frequently are placed on fast storage to potentially benefit multiple hosts.

One example of a hybrid system that can be used as a building block for larger storage systems is the DDN SFA18KX hybrid disk array. It offers up to 3.2 million IOPS and 90GB/sec from a single 4U appliance. It uses NVMe devices and spinning drives. This high level of density makes the SFA18KX suited for data centres with limited space or any high-performance environment that aspires to expand capacity without adding the complexity of many appliances to manage and the cost of powering and cooling a large number of controllers.

Another interesting solution combining disk technologies is the NetApp E5700 series hybrid flash array where SAS attached HDDs and SSDs can be combined. It was designed specifically for heavy duty environments, including those for analysing large data sets, the E5700 provides over 1 million persistent IOPS and response times in microseconds. Bandwidth oriented loads can reach up to 21 GB/s. The E5700 is a 2U (24 drives) or 4U (60 drives) array supporting multiple high-speed host interfaces, including 32 Gb/s Fibre Channel, 25 Gb iSCSI, 100 Gb InfiniBand, 12 Gb SAS, 100 Gb NVMe via InfiniBand and 100 Gb NVMe via RoCE (RDMA – Remote Direct Memory Access – over Converged Ethernet).

### 2.4.2. Tape Libraries

Historically, tape drives were used for local system backups with a system administrator changing tapes in drives directly attached to servers. High speed networks, large disk drives and cloud storage have made the directly

attached drives for small scale backups a niche market. In most cases tape usage has been concentrated on large automated tape libraries used by many systems for backups and archive storage. This consolidation of storage and fewer drive sold have had consequences both for the business and the technical side.

Economies of scale mean that tape technology becomes a winner takes it all market with the need to amortise R&D costs over a low number of units. The StorageTek T10000 format was the main contender to IBM 3592 in the high-end tape market but could not compete on production volume, and development of the “E” format was cancelled a few years back. Thus, tape technology has become a mostly single vendor market on the drive side with IBM producing drive heads for both LTO and 3592 (aka Jaguar) drives. LTO media is produced by Sony and Fujifilm which in 2019 settled their lawsuit which had essentially halted production of LTO-8 media and slowed down the adaption of LTO-8 technology. In late 2020 LTO-9 is being introduced into the market, so LTO-8 may end up as one of the less widely deployed LTO generations.

IBM classifies the 3592 drives as “enterprise” technology, and new features are usually first introduced there. Main differences to LTO were historically the storage capacity, bandwidth and seek/access times. In recent generations the bandwidth gap has narrowed, and the pricing for tape media means that the price per TB is similar. Latency is the remaining large difference, with 3592 supporting recommended access order (RAO) and higher resolution directories for high-speed seeking. See Table 1 below for a summary.

<b>Tape Media</b>	<b>Uncompressed Capacity</b>	<b>Uncompressed Bandwidth</b>	<b>RAO Support</b>	<b>High Resolution Directory Size</b>
3592-JE	20 TB	400 MB/s	Yes	64
LTO-8	12 TB	360 MB/s	No	2
LTO-9	18 TB	400 MB/s	No	2

Table 1 Tape technology feature comparison

Mechanical constraints limit the possible performance of LTO drives due to pressure from system vendors to support the use case of tape drives inside a server chassis. This limits the physical form factor of the drive to what is known as half-height in the tape world. Combining this size with the fixed size of the tape cartridge leaves little room for the intricate mechanics needed for moving the tape with both high speed and precision. In the future this will probably require the LTO format to have divergent performance tiers for half- and full-height drives, for LTO-8 the difference is 300 vs 360 MB/s. Tape libraries usually use full-height drives since they are less constrained by space, but the lower price of half-height drives makes them an option for some libraries where price is more important than performance.

Competition remains on the tape library side with IBM, Oracle and Spectra Logic providing a range of library models targeting even the largest sites while Quantum are more focused on the low/mid-sized sites. All vendors are offering LTO technology in the libraries, with IBM and Spectra Logic also providing 3592 drives. Some notable differences between libraries are the different methods used for storing tapes inside the library. Optimising for mount latency leads to having libraries where all tapes are directly accessible and have short paths to drives (example StorageTek SL8500) and optimising for density leads to depth stacking of cartridges (for example, IBM TS4500). Trying to strike a balance is the drawer approach (for example Spectra Logic T950) where cartridges are placed in containers that can be pulled out and a single cartridge taken, or the entire container moved to/from the I/O station.

The ransomware attacks in the last few years have caused a resurgence of interest in tape technology for normal enterprise backups due to the possibilities of both keeping them separate from other systems and also physically removing the tapes for vault storage. HPC sites usually have tapes for archival storage and lower cost tiers in storage systems and keep all tapes accessible online. For most of these use cases the mount latency is not critical, and seek latency is more important.

Tape storage is viewed as low performance but is mostly high latency storage due to its sequential nature, not low bandwidth. When reading and writing data, tape drives prefer a steady stream of data to keep up with the movement of the tape and will do speed matching within a range but cannot go arbitrarily low. An LTO-8 drive, for example, has a lower limit of 112 MB/s when streaming. Technology projections [8] are becoming an increasingly important issue in the future. To be able to feed the tape libraries during writes they need to be matched with high-speed disk storage, so coupling flash and tape tiers directly will be attractive.

### 3. File Systems

Most HPC installations rely on distributed/parallel file systems, with Lustre and Spectrum Scale (formerly GPFS) being the most common. Thanks to their built-in scalability, they are able to manage huge amounts of data, support high bandwidth and provide high metadata performance. The increasingly demanding requirements of HPC systems are a good test case for new storage technologies. These file systems can perform hundreds of thousands of operations on metadata per second and stream multiple TBs of data per second. To meet the ever-increasing demands, cluster file systems are constantly evolving towards more universal, stable, and useful solutions. Current user expectations include high-performance access to small files, increased levels of security (encryption for example), support for data replication mechanisms and data retrieval via automatic tiering. Some of these are features that enterprise systems (NetApp FAS, EMC, etc.) have provided earlier without having the same level of performance.

In this section we will examine the evolution of the cluster file systems Lustre, Spectrum Scale and BeeGFS. All these have recently been developed with new features. Lustre is now implementing Distributed Namespace (DNE), Erasure Coding, Data on Metadata (DOM) and Persistent Client Cache (PCC). Spectrum Scale is now implementing Native Declustered RAID and BeeGFS has Storage-On Demand.

#### 3.1. Lustre

Lustre is open-source software whose development is supported and coordinated by the non-profit EOFS and OpenSFS organisations. Developed since 1999, it is used as a file system by many computing environments in the world. For more than 15 years, it has been used by at least half of the top 10 largest supercomputers and is known for supporting the largest high-performance computing clusters in the world, with tens of thousands of client systems, petabytes of storage deployed, and hundreds of gigabytes per second of I/O bandwidth. The central component of the Lustre architecture is the Lustre file system, which is supported on the Linux operating system and ensures compatibility with the POSIX standard.

Until recently, the Lustre file system performance has been optimised for large files. This results in many Remote Procedure Call (RPC) round trips to the Object Storage Targets (OSTs), which reduces small file performance. Therefore, a new functionality has been implemented to allow the placement of small files on Meta Data Targets (MDT) (Figure 4) so that these additional RPCs can be eliminated, and performance improved correspondingly. Used in conjunction with the Distributed Namespace (DNE), this will preserve efficiency without sacrificing horizontal scaling. Users or system administrators can set a layout policy that places small files on MDT. Files that grow beyond this size will use Progressive File Layouts to extend larger files onto OST objects and leave the small part of the file (defined by user or system admin) on the MDT.

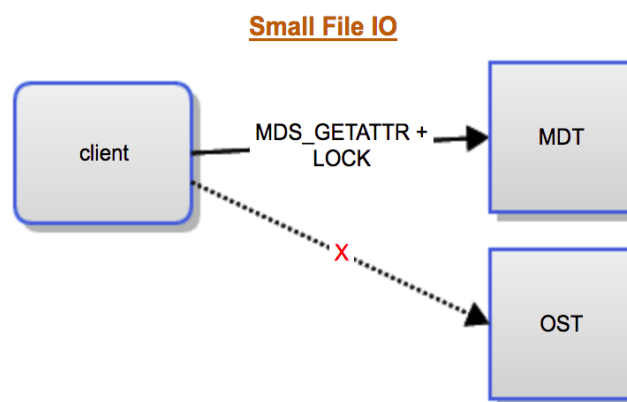


Figure 4: Small File IO (<http://wiki.lustre.org/> )

Persistent Client Cache (PCC) is another new feature that was implemented in Lustre version 2.13. Data is cached locally on SSDs or NVMe drives at the Lustre client side. These caches are not part of the global Lustre namespace,

instead each client uses its own SSD as a purely local cache. Cached data is managed using a local file system with I/O for cached files being fulfilled locally while other I/O is directed to the shared OSTs. PCC uses the previously existing HSM (Hierarchical Storage Management) support in Lustre for data synchronisation. It uses the HSM copy tool to move files from the local cache to the Lustre OSTs, acting as a HSM backend with a unique archive number. If another Lustre client accesses data cached in this manner it will trigger data synchronisation. Clients using PCC going offline are handled by making the data temporarily inaccessible for other clients. When the PCC client is online again the copy tool restarts, and the data is accessible again.

One feature on the Lustre 2.14 roadmap is client-side data encryption. This will increase security to a higher level, where a leak of data sent to the server is no longer a threat. Assumptions about this ability are as follows:

- encrypting file contents
- encrypting file name
- using the master key to encrypt data
- file data is no longer available after deleting the key
- ability to change the key without re-encrypting the files
- denying access to encrypted data after deleting the master key from the client's memory.

### 3.2. Spectrum Scale (formerly GPFS)

Another very well-known product on the clustered file systems market is IBM General Parallel File System (IBM GPFS) renamed IBM Spectrum Scale. The filesystem layout spreads data between multiple servers simultaneously, thus creating a global namespace. It supports both large scale HPC environments as well small-scale HPC systems. The word “parallel” in the former product name indicates the main feature of Spectrum Scale (Figure 5), namely the mechanism of data partitioning and their simultaneous distribution on many disks / disk arrays. This feature allows faster reading and writing of data. Additional mechanisms exist in Spectrum Scale to increase the reliability and performance of the entire system, such as automated management functions, high availability of resources, replication and mirroring.

Spectrum Scale is a clustered file system. This means that it provides simultaneous access to one file system from multiple nodes. All nodes can be connected to the SAN or network data storage systems. This enables high-performance access to the same resources through multiple access nodes simultaneously.

The Spectrum Scale file system provides the following mechanisms:

- Increasing the total throughput of a file system by spreading reads and writes across multiple disk resources.
- Simultaneous access to multiple processes or applications on all cluster nodes using standard file system calls.
- Load balancing by evenly distributing data across all drives. This increases the total system capacity and eliminates the bottlenecks in the transfer.
- Each physical disk device intended for use in the Spectrum Scale cluster should be defined as an NSD (Network Shared Disk). This allows you to create an additional layer of logs for input/output operations. This directly translates into increased system efficiency.
- Support for very large file sizes.
- Parallel, simultaneous reads and writes from multiple IBM Spectrum Scale cluster nodes.
- An extensive system of managing distributed tokens (locks) on files. Token management distribution reduces facility maintenance latency.
- Writing data to multiple disks via various disk controllers. Large files in IBM Spectrum Scale are split into blocks of equal size, and successive blocks are placed on different disks.
- Acceleration of reading by pre-fetching data into buffers.
- Native Declustered Raid: IBM Spectrum Scale RAID implements a sophisticated data and spare space disk layout scheme that allows for arbitrarily sized disk arrays while also reducing the overhead to clients when recovering from disk failures. To accomplish this, IBM Spectrum Scale RAID uniformly spreads or *declusters* user data, redundancy information, and spare space across all the disks of a declustered array.

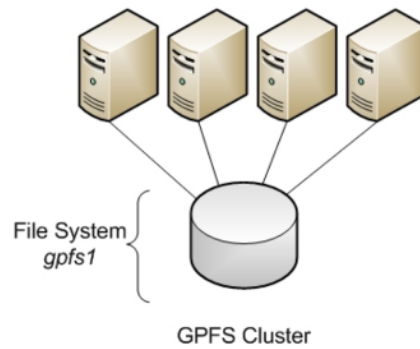


Figure 5: Simultaneous Access by Multiple Clients to GPFS Resources (IBM)

### 3.3. BeeGFS

The high-performance file system market is dominated by players like Lustre and Spectrum Scale. In recent years the European developed file system BeeGFS has emerged as a competitor with a growing number of sites implementing it. This file system was originally developed by Fraunhofer as an internal file system named FhGFS, but development was spun out by forming the company ThinkParQ.

BeeGFS boasts many features that are useful for users of high-performance file systems. It is software defined storage based on the POSIX file system standard. It means that applications can easily and efficiently use BeeGFS resources. System clients communicate with the cluster via a TCP/IP network or a high-performance Infiniband network.

The main features of a BeeGFS cluster are:

- Data is spread across multiple servers and increasing the number of servers and disks in the system translates directly into the capacity and performance of the file system represented as a single namespace.
- The BeeGFS network protocol is independent of the hardware platform. Hosts of different platforms can be mixed within the same file system instance.
- Management, metadata and storage services do not have direct access to the disks. Instead, they store data in any local POSIX file system (Ext4, XFS or ZFS). This gives you the flexibility to choose a basic file system that gives you maximum performance in the context of the hardware used.
- BeeGFS uses all available RAM in the server (which is not needed by other processes) automatically for buffering data. This gives a huge performance gain when handling small I/O requests and then aggregating them into larger blocks before saving to disk.
- BeeGFS has a feature that allows flash drives to become directly accessible to users. Users can request BeeGFS (via the `beegfs-ctl` command line tool) to transfer the current project to high-performance flash drives (e.g. NVMe) (Figure 6).
- BeeGFS supports all networks based on the TCP/IP protocol and the native InfiniBand or Omni-Path protocols. Servers and clients can handle requests from/to different networks at the same time.

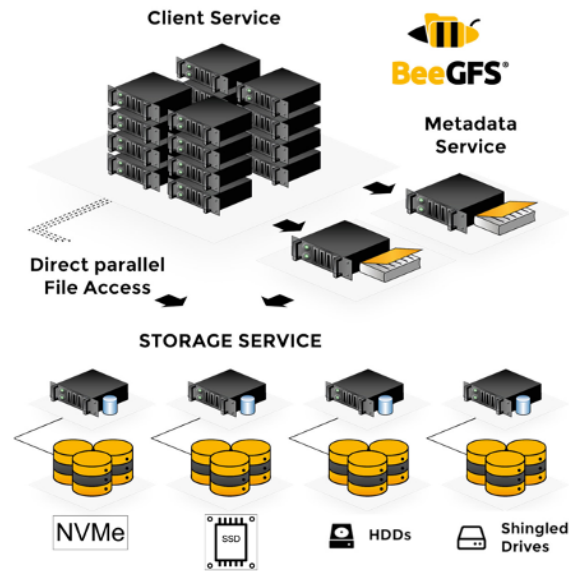


Figure 6: BeeGFS The Parallel Cluster File System [9]

### 3.4. Ceph

The Ceph storage system was introduced as a prototype in 2006 [10] and its file system client code been a part of the Linux kernel since 2010. Funding for the open-source project is provided by the Ceph Foundation, which in itself is hosted by the Linux Foundation. Ceph is designed as a distributed object storage cluster for commodity hardware with object, block and file system storage services layered on top (see Figure 7 for an overview of the architecture).

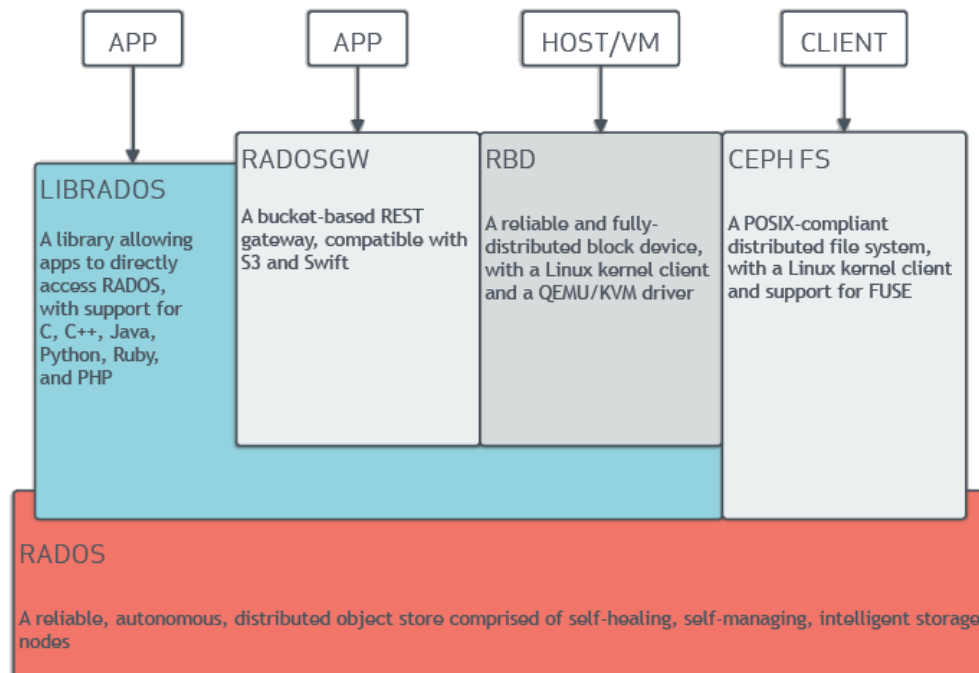


Figure 7: Ceph architecture [11]

Ceph is not designed to provide storage from single server, to take advantage of its features multiple Object Storage Devices (OSD) are needed to support for example replication and erasure coding. Most of the early deployments of Ceph have been in cloud environments where S3 or Swift access protocols and block storage are needed for the

virtual machines. As a file system Ceph is not yet a mainstream HPC choice but is starting to be used at sites that handle both traditional HPC clusters and cloud environments. One notable example is CERN where it is used as storage for their Openstack cloud. Upcoming deployments include the LUMI EuroHPC systems which also support a container cloud platform based on OpenShift and Kubernetes.

### 3.5. File System Feature Comparison

Table 2 compares the feature sets of some file systems popular in the HPC environment with regards to automatic data movement within the storage hierarchy.

Features mentioned below should be interpreted as

- **S3 Provider:** The file system can also act as object storage and provides an interface that is compatible with the S3 protocol popularised by AWS
- **Tier to Drives:** Data is migrated between different drive technologies transparently to optimise latency and cost trade-offs, with an example being SSD to/from HDD movement
- **Tier to Tape:** Data is migrated to a tape library, but still visible in the file system namespace and transparently migrated back if accessed
- **Tier to Cloud:** Data is migrated to object storage, either on premises or using an external provider

	<b>S3 Provider</b>	<b>Tier to Drives</b>	<b>Tier to Tape</b>	<b>Tier to Cloud</b>
<b>BeeGFS</b>	No	Yes	No	No
<b>Ceph</b>	Yes	Yes, cache	No	No
<b>Lustre</b>	No	Yes	Yes	Yes, unofficial
<b>Spectrum Scale</b>	Yes	Yes	Yes	Yes

Table 2: File System Features for Cloud and Tiering

## 4. Data Management Services

Today, parallel file systems, possibly in combination with a Hierarchical Storage Management (HSM) system, are still the most important approach to provision storage resources for HPC systems. To meet future needs, new technologies for data management and tiering are becoming increasingly important. For the future we expect not only growing needs in terms of storage capacity and performance, but also needs related to collaborative data management, realisation of workflows extending a single HPC data centre to support multiple groups sharing data as well as the provisioning of data according to the FAIR data principles [12] are becoming key drivers.

In this section we report on select developments on the market, based on conference presentations and the authors' personal experience, which are relevant in this context. In Section 4.1 we analyse primarily the evolution of cloud storage technologies which, among others, can help to facilitate data sharing. Large-scale cloud providers are pushing for a new approach to architecting storage. In Section 4.2 we report on the relatively new concept of data lakes. Unlike cloud storage architectures, HPC storage architectures have long relied on tiered architectures comprising tiers optimised for capacity and others optimised for performance. In Section 4.3 we summarise the status of more traditional solutions for managing tiered storage architectures, while in the following Section 4.4 we focus on emerging solutions for I/O acceleration architectures and technologies. Finally, in Section 4.5 we consider the existing and emerging data management frameworks that are attracting interest in the context of HPC infrastructures.

### 4.1. Cloud Storage Interfaces

Object store architectures are receiving an increasing interest in the context of HPC mainly as a possible option for addressing scalability issues related to POSIX compliant parallel file systems when going to Exascale. Object store technologies that are used for cloud infrastructures are, however, also of interest for HPC infrastructures as they are designed for geographically distributed storage infrastructures that are much more openly accessible than parallel file systems. Thus, they help to meet the need of data sharing and realising workflows that extend beyond a single data centre. This is an important aspect for realising the recently formulated vision of “Transcontinuum Extreme-Scale Infrastructures” [13].

The focus here is not on object store architectures but rather their interfaces, i.e. on S3 and Swift. S3 is an API introduced and controlled by Amazon [14], while Swift is an object store technology and API developed by the OpenStack community [15]. What they both have in common is that they are web-based interfaces implementing a REST API that supports a small set of operations like *get object* or *put object*.

The use of cloud storage interfaces to facilitate external access to HPC data centres is still at an early stage. The Fenix project has recently announced that they will use Swift to facilitate access to federated storage resources [16]. There are several commercial solutions on the market that will support this development. There are two approaches to provision access to storage through the aforementioned APIs: one can either use (1) native object store solutions like OpenStack Swift or (2) provision these interfaces on top of other storage solutions, e.g. parallel file system solutions.

The latter approach is realised by IBM's Cluster Export Services (CES) for Spectrum Scale [17]. A CES node acts as a protocol node providing non-Spectrum Scale clients with access to data managed by Spectrum Scale. This approach has multiple benefits in the context of HPC data centres. Parallel file systems are technologies that are well integrated and supported by these data centres. Additionally, various commercial solution providers, which are active in the HPC market, provide the necessary support for such a configuration. Currently, similar solutions are not yet available for Lustre.

Commercial support for deployment of native objects stores is improving. Atos announced its new BullSequana Xstor [18] with support of Ceph, which provides both an S3 and a Swift compatible API through its Rados gateway. The new object store architecture Mero, which is partially developed within the EU-funded Sage projects [19], supports different storage interfaces (like S3), based on a component called Lingua Franca. This component implements different meta-data formats and interfaces. A number of smaller suppliers started to provide proprietary object store solutions including S3 interface, including the French company OpenIO, which is positioning its solution also in the HPC market and show-cased their product at SC'19 [20]. Yet other suppliers bundle open-source software stacks like OpenStack Swift for enabling commercial offerings, e.g. SwiftStack [21], a company recently acquired by NVIDIA. The latter solution is notably positioned as a solution that allows to extend storage beyond the data centre towards the Edge Computing.

While commercially supported solutions are receiving an increasing interest by HPC data centres, there are also solutions developed by research organisations in the context of grid computing [22] moving towards support of web-based interfaces. One example is dCache that has been developed for high energy physics storage application and supports grid protocols, network file system access as well as WebDAV [23] using protocol "doors", similar to the CES nodes for Spectrum Scale mentioned above. It is designed to be distributed among sites and supports tertiary storage systems for tiering to tape, for example. Developed for grid applications, it has also been used in federated national storage.

## 4.2. Data Lake and ad-hoc Storage Systems

Driven by commercial cloud providers like Amazon, Azure and Google, the new concept of *data lakes* is gaining momentum and is expected to be adopted also in the context of HPC infrastructures. A data lake can logically be seen as a centralised repository that might be realised on distributed storage resources, which allows for the storage of structured and unstructured data at any scale. The data is imported from different sources and provisioned in its original format. The aim is to make data available for processing and analytics pipelines soon after it becomes available.

A data lake can typically be expected to be a storage tier for cold data objects that can be accessed with limited performance and in formats that are not suitable for further processing. To make data access suitable for high-performance computing and high-performance data analytics specialisation and locality must be improved [24]. This can be achieved through, for instance, dynamic provisioning of storage [25] [26]. One solution that realises this is BeeGFS on Demand (BeeOND) that is commercially supported through ThinkParQ [27].

## 4.3. Tiered Storage Management

Tiering is not new in HPC environments and solutions, which may be considered "classical", continue to be further enhanced. The classical tiered storage approach in HPC environments is extending the file system to support multiple tiers. Cluster file systems such as Spectrum Scale and Lustre support this natively or through addons that can be extensively configured to select which files are migrated between tiers. This creates the appearance of all

files being locally online while they may, for example, be stored in a faraway tape library with high access latency. Tiered storage management solutions may be largely invisible to the user.

The following part highlights different products for tiered storage management which are all actively developed and can be expected to continue being relevant for HPC infrastructures.

**IBM Spectrum Scale** is a parallel file system used at many HPC sites. Tiering support is built in for disk and cloud tiers with space management add-ons supporting tape. Spectrum Scale Information Lifecycle Management (ILM) is a set of tools that allows to define placement and migration policies (see [28] for a recent overview).

**Lustre HSM** support was added in release 2.5 of Lustre. The design is based on a coordinator and agents that are responsible for moving data between the Lustre and HSM worlds. Migration requests can be user-triggered or initiated by a policy engine like Robinhood, which was developed at CEA and is the most commonly used add-on for Lustre HSM. Lustre HSM continues to be actively developed (see [29] for a recent update).

**HPE Data Management Framework (DMF)** is a software-defined framework for managing multiple storage tiers [30]. It can connect high-performance file systems, e.g. Spectrum Scale or Lustre, and a back-end data store which could, for example, be based on tape or an object store with off-site data replication enabled. In the most recent version DMF7 support for extensible metadata was added, which enables new data management capabilities, e.g. handling of data sets, and better integration with HPC job schedulers.

**HPSS (High Performance Storage System)** is an HSM system built mainly by IBM and US DOE lab [31]. It has been developed for a long time and with support for tape usage. It is optimised for I/O bandwidth by supporting parallel I/O through software striping, e.g. through RAIT (Redundant Array of Independent Tapes), which allows for striping data on tape. The Spectrum Scale can also use HPSS as a space management backend.

**Versity Storage Manager (VSM)** is a software platform that automates the process of storing and retrieving archival data [32]. Versity's product VSM2 comprises an open-source archiving file system with a POSIX interface called ScoutFS [33] and the proprietary Scout Archive Manager ScoutAM. A design target of ScoutFS, which makes it particularly interesting for the future, is advanced indexing capabilities to allow for quick discovery of inode attribute and file content changes. Version 1 of the product was based on SAM-QFS and offers a migration path for installations using SAM-QFS/Oracle HSM.

#### 4.4. I/O Acceleration Solutions

While in the past typical HPC data centres realised storage infrastructures based on an online tier using HDDs and an offline tier using tapes, the increasing need for performance requires adding another shared storage tier or facilitating use of node-local storage. This storage is based on fast non-volatile memory technologies to realise high-performance both in terms of bandwidth and throughput of I/O operations. While the storage is persistent, the data is expected to remain there for short periods of time and is typically staged from or migrated to a slower but much larger storage system. The corresponding solutions can therefore be considered to be I/O accelerators.

In the following part we provide different I/O accelerator solutions that are being used for HPC infrastructures and are expected to continue being relevant in the future. Section 2.3 has further coverage on this subject.

**IME (Infinite Memory Engine)** from DDN is meanwhile used at various leading supercomputing centres [34]. It is designed as an intermediate storage layer between an HPC system and an external storage system and is implemented by servers with a larger number of NVMe SSDs. Data stored in IME can be accessed either through the IME native interface or via a POSIX client. It uses the namespace of the backing file system, which could be Spectrum Scale or Lustre.

**DAOS** is a software-defined object store solution optimised for distributed non-volatile memory [35]. It was developed mainly by Intel and has been open-sourced. Applications can access datasets stored in DAOS either directly through the native DAOS API or by using I/O libraries (e.g. POSIX emulation, MPI-IO, HDF5) or frameworks (e.g., Spark, TensorFlow). DAOS will be used for the upcoming Aurora system at ANL, which is planned to be the first US Exascale system. DAOS is supported by multiple HPC system vendors including HPE and Lenovo.

**Excelero NVMesh** is a software-defined storage solution that allows to dynamically create a block storage volume on top of distributed NVMe SSDs. It can be implemented using any of the high-speed network technologies commonly used in HPC systems. Different storage solutions can be deployed dynamically on top of block store volumes. STFC in the UK uses, for instance, BeeGFS on top of NVMesh [36]. Excelero is an SME in the US, which offers its solution also through HPC systems vendors like Lenovo.

**Atos Smart Data Management Suite** comprises two solutions for I/O acceleration [37]. At hardware level they are both realised by servers that host a set of high-performance SSDs and are integrated in the HPC system's high-performance network. When using the Smart Burst Buffer solution (SBB), the SSDs are used as an intermediate cache transparent to the user. It relies on I/O calls interception through the scheme implemented in the Bull IO Instrumentation library. With the Smart Bunch of Flash (SBF) applications can be enabled to explicitly request a static allocation of NVMe storage.

## 4.5. Data Management Solutions

In this section we consider different solutions that provide user interfaces and tools for managing data. As of today, none of these solutions can claim a wide uptake in HPC infrastructures. However, with the growing importance of collaboratively managing data, the increasing need for enabling data analytics on structured and unstructured data as well as the support of the FAIR principles for data access, these solutions are expected to become more relevant.

**iRODS (Integrated Rule-Oriented Data System)** is an open-source data grid middleware. It is based on an abstraction for data management processes and policies. It provides users with a uniform interface to heterogeneous storage systems (both POSIX and non-POSIX) [38]. It allows federating a distributed storage infrastructure under a unified namespace. It also includes, to give a few examples, a workflow engine where rules that trigger actions can be added when defined conditions apply and the possibility to define microservices that run inside the iRODS system. A key focus for iRODS in the future is improved support of metadata for managing data [39]. One example of a large European project using iRODS is EUDAT. iRODS is developed by a consortium of private and publicly funded organisations. Limited commercial support is available through a partner program.

**Rucio** is a framework for scientific data management developed in the high-energy physics community [40]. The impetus for the original work was the ATLAS experiment at CERN and its storage requirements. It was designed to integrate easily with other already existing components and to provide high level integration. Workflow and physical storage are handled by other systems, but for example rules on how many replicas a dataset should have and where to find them are in Rucio. Several access protocols (including WebDAV and S3) as well as different types of authentications (including username and password, SSH-RSA public key exchange) are supported. Therefore, Rucio could be a candidate for HPC infrastructures. Rucio is developed by a scientific community and no commercial support is available.

**Starfish** is a solution for managing data in the context of very large-scale storage systems possibly based on multiple file systems (including Spectrum Scale or Lustre), object stores or tape libraries [41]. It is designed to scale to billions of files or objects. Starfish allows users and applications to assign tags and key-value pairs to files and directories to classify content, drive batch processes, and enforce policies. Starfish can be used to migrate files between multiple storage devices, synchronise and replicate data between locations or different file systems, and archive stale or old data manually or automatically based on metadata attributes. Starfish Storage is a US company founded in 2013 and positions its product also in the HPC market.

**Nodeum** is a storage management framework that can work on top of multiple data stores based on different storage types, including POSIX file systems or object stores with S3 compliant interfaces [42]. A global view is implemented through a virtual file system layer. Nodeum is a small company in Belgium, which starts to explore the use of its product in the context of HPC infrastructures.

Except for more general-purpose solutions, domain specific solutions continue to play a critical role for realising important HPC workflows. Two specific examples are listed below.

**MARS** is the Meteorological Archival and Retrieval System developed at ECMWF (European Centre for Medium-Range Weather Forecasts). It stores GRIB and NetCDF files [43]. While the stored files can be retrieved as-is the main usage is to query MARS for certain parameters over time ranges, where the output files are synthesised from data in a stored file.

The **Earth System Grid Federation** has developed a publication system for climate research data, partly supported by the Horizon 2020 IS-ENES projects [44]. Data is shared by research groups across the world with QA processes before publication and checks for not publishing data sets with errors. Storage and global search indexes are distributed among federation sites.

## 5. Trends

With the exponential growth of data, distributed/parallel storage systems have become not only an essential part, but also one of the bottlenecks of large-scale supercomputing centres. High latency data access, poor scalability, difficulty managing large datasets, and lack of query capabilities are just a few examples of common hurdles. Traditional storage systems have been designed for HDD media and for POSIX I/O. These storage systems represent a key performance bottleneck, and they cannot evolve to support new data models and next generation workflows. A strong trend is observed leading towards very high-performance media, based on NVMe solutions. By designing new hardware interfaces and creating new software solutions such as I/O accelerators higher performance than previously can be achieved.

Storage requirements in terms of both capacity and performance will continue to increase, and the storage stack will be expanded to include more levels in the hierarchy as Exascale systems appear. Data is becoming more important in itself and not only an adjunct to the computation. Moving data around is becoming more costly and creating multiple copies for different access methods does not scale. Storage systems are starting to support multiple access methods (such as file system I/O and S3 protocol) to the same data.

The importance of long-term handling of data will be greater in the future with the increased move towards making data more publicly available. FAIR data principles increase the importance of handling data in a structured way during its entire lifetime. Finding data and making it reusable requires extensive meta data, and to access the data publicly documented protocols are needed. In Section 4 we have looked at a number of data management technologies that provide basic storage, handling of meta data and multiple access protocols.

For the foreseeable future, storage systems based on HDD technology and/or tape libraries will provide space for storing data with suitable cost/latency trade-offs.

## 6. Conclusion

Here the general conclusions for the areas investigated in this report are summarised.

### Storage infrastructure:

1. Data infrastructures needs to strike a balance between capacity and performance, with the right balance depending on their tiering level. The growing necessity for efficient operation with large datasets leads to purely flash-based solutions for warm data due to latency and bandwidth considerations.
2. The latest and fastest technologies, starting from NVMe disks, through cluster file systems such as DAOS, will become the basis for building new, ultra-fast Exascale systems.
3. Both traditional HDD based storage systems and tape libraries remain competitive for large volume storage, less IOPS intensive use cases and storage of cold data.

### Data management and access:

4. Data must be both findable and accessible, and software support for managing meta data is required. Some scientific workflows may benefit greatly from domain specific solutions, but unless resources for maintaining such tools are provided, a more general solution is recommended.
5. HPC systems traditionally use parallel file systems, which can be extended with I/O accelerators for faster access during computations and with tiering support for automatically moving data to lower cost media. Enabling access to this data through object storage interfaces allows more software workflows to use the data directly.

## References

- [1] A. Tekin, A. T. Durak, C. Piechurski, D. Kaliszan, F. A. Sungur, F. Robertsen and P. Gschwandtner, “State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI, PRACE Technical Report,” 2020.
- [2] E. Krishnasamy, S. Varrette and M. Mucciardi, “Edge Computing: An Overview of Framework and Applications,” PRACE Technical Report, 2020.
- [3] Square Kilometer Array, “Software and Computing,” [Online]. Available: <https://www.skatelescope.org/software-and-computing/>. [Accessed 16 11 2020].
- [4] The Register, [Online]. Available: [https://www.theregister.com/2015/11/03/intels\\_allflash\\_data\\_center/](https://www.theregister.com/2015/11/03/intels_allflash_data_center/).
- [5] P.-Y. Du, H.-T. Lue, Y.-H. Shih, K.-Y. Hsieh and C.-Y. Lu, “Overview of 3D NAND Flash and progress of split-page 3D vertical gate (3DVG) NAND architecture,” in *12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Guilin, China, 2014.
- [6] Intel, “Reimagining the Data Center Memory and Storage Hierarchy,” [Online]. Available: <https://newsroom.intel.com/editorials/re-architecting-data-center-memory-storage-hierarchy/>. [Accessed 11 2020].
- [7] “DAOS For Applications,” 6 Feb 2020. [Online]. Available: [https://ecpannualmeeting.com/assets/overview/sessions/DAOS\\_ECP.pdf](https://ecpannualmeeting.com/assets/overview/sessions/DAOS_ECP.pdf).
- [8] “INSIC Technology Roadmap 2019,” [Online]. Available: <http://www.insic.org/wp-content/uploads/2019/07/INSIC-Technology-Roadmap-2019.pdf>.
- [9] Thinkparq, “BeeGFS,” [Online]. Available: <https://www.beegfs.io/c/>. [Accessed 16 11 2020].
- [10] S. Weil, S. Brandt, E. Miller, D. Long and C. Maltzahn, “CRUSH: Controlled, scalable, decentralized placement of replicated data,” Tampa, FL, 2006.
- [11] Ceph Project, “Ceph Architecture,” [Online]. Available: <https://docs.ceph.com/en/latest/architecture/>. [Accessed 16 11 2020].
- [12] “FAIR Principles,” [Online]. Available: <https://www.go-fair.org/fair-principles/>.
- [13] M. Malms, “ETH4HPC’s SRA 4. Strategic Research Agenda for High-Performance Computing in Europe,” 2020.
- [14] Amazon, “Amazon Simple Storage Service. Developer Guide”.
- [15] “Swift,” [Online]. Available: <https://docs.openstack.org/swift/latest>.
- [16] “Fenix,” [Online]. Available: <https://www.fenix-ri.eu>.
- [17] D. Hildebrand, “A Deployment Guide for IBM Spectrum Scale Unified File and Object Storage,” 2017.
- [18] “XSTOR,” [Online]. Available: <https://atos.net/en/solutions/high-performance-computing-hpc/bullsequana-xstor>.
- [19] S. Narasimhamurthy, “The SAGE project: a storage centric approach for exascale computing,” in *Proceedings of the 15th ACM International Conference on Computing Frontiers 2018*, 2018.
- [20] “OpenIO,” [Online]. Available: <https://www.openio.io>.

- [21] “Swiftstack,” [Online]. Available: <https://www.swiftstack.com>.
- [22] I. Foster and C. Kesselman, *The Grid 2*, Morgan Kaufmann, 2003.
- [23] “dCache,” [Online]. Available: <https://www.dcache.org>.
- [24] P. Carns, “BYOFS: The opportunities and dangers of specialisation in the age of exascale data storage”.
- [25] F. Tessier, “Dynamically Provisioning Cray DataWarp Storage,” p. arXiv:1911.12162, 2019.
- [26] A. Brinkmann, “Ad Hoc File Systems for High-Performance Computing,” vol. 35, no. 1, 2020.
- [27] “BEEOND,” [Online]. Available: <https://thinkparq.com/products/beeond/>.
- [28] N. Haustein, “IBM Spectrum Scale Information Lifecycle Management,” London, 2019.
- [29] B. Evans, “HSM, Data Movement, Tiering and More”.
- [30] HPE, “HPE Data Management Framework 7,” November 2019. [Online]. Available: [https://support.hpe.com/hpesc/public/docDisplay?docLocale=en\\_US&docId=a00056652enw](https://support.hpe.com/hpesc/public/docDisplay?docLocale=en_US&docId=a00056652enw).
- [31] “HPSS,” [Online]. Available: <http://www.hpss-collaboration.org>.
- [32] “VERSITY,” [Online]. Available: <https://www.versity.com>.
- [33] “ScoutFS,” [Online]. Available: <https://www.scoutfs.org>.
- [34] DDN, “IME Datasheet,” 19 08 2020. [Online]. Available: <https://www.ddn.com/download/ime-datasheet/>.
- [35] “DAOS,” [Online]. Available: <http://daos.io>.
- [36] “NVMeshSTFC,” [Online]. Available: <https://www.excelero.com/wp-content/uploads/2019/11/GPU-Servers-for-Machine-Learning-and-AI.pdf>.
- [37] “SDMS,” [Online]. Available: <https://atos.net/wp-content/uploads/2019/06/Smart-Data-Management-Suite.pdf>.
- [38] “iRODS,” [Online]. Available: <https://irods.org>.
- [39] T. Russell, “Beyond Discoverability: Metadata to drive your data management,” 2020.
- [40] M. Barisits, “Rucio: Scientific Data Management,” *Computing and Software for Big Science*, p. 3:11, 2019.
- [41] “Starfish,” [Online]. Available: <https://starfishstorage.com>.
- [42] “Nodeum,” [Online]. Available: <https://www.nodeum.io>.
- [43] “MARS,” [Online]. Available: <https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation>.
- [44] “ESGF,” [Online]. Available: <https://esgf.llnl.gov>.

## List of acronyms

AHCI	Advanced Host Controller Interface
AI	Artificial Intelligence
ANL	Argonne National Laboratories
BeeOND	BeeGFS on Demand
CES	Cluster Export Services
DAOS	Distributed Asynchronous Object Storage
DMF	Data Management Framework
DNE	Distributed Namespace Environment
DoE	Department of Energy
DWDP	Disk Write Per Day
ECMWF	European Centre for Medium-Range Weather Forecasts
EOFS	European Open File System
EU	European Union
GDPR	General Data Protection Regulation
HBM	High Bandwidth Memory
HDD	Hard Disk Drive
HDR	High Data Rate
HPSS	High Performance Storage System
HSM	Hierarchical Storage Management
ILM	Information Lifecycle Management
INFRAG	Infrastructure Advisory Group
IOPS	I/O operations per second
iRODS	Integrated Rule-Oriented Data System
LTO	Linear Tape Open
MARS	Meteorological Archival and Retrieval System
MDT	Metadata Target
MLC	Multi Level Cell
NVMe	Non-Volatile Memory Express
OS	Operating System
OST	Object Storage Target
OU	Organisational Unit
PCC	Persistent Client Cache
PRACE	Partnership for Advanced Computing in Europe
QoS	Quality of Service
RAIT	Redundant Array of Independent Tapes
RDMA	Remote Direct Memory Access
REST	REpresentational State Transfer
RIAG	Research and Innovation Advisory Group
RoCE	RDMA over Converged Ethernet
RPC	Remote Procedure Call
SAS	Serial Attached SCSI
SATA	Serial ATA
SKA	Square Kilometer Array
SLC	Single Level Cell
SMT	Simultaneous Multithreading
SoC	System on Chip
SR-IOV	Single Root Input/Output Virtualization
SRA	Strategic Research Agenda
SRT	Intel Smart Response
TCO	Total Cost of Ownership
TCP/IP	Transmission Control Protocol/Internet Protocol
TLC	Triple Level Cell
UI	User Interface
VM	Virtual Machine
VSM	Versity Storage Manager

## **Acknowledgements**

This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.