



**Hewlett Packard
Enterprise**

Technical white paper

Check if the document is available
in the language of your choice.



CRAY CLUSTERSTOR E1000 STORAGE SYSTEM

System overview



CONTENTS

Executive summary 3

Intended audience 3

Lustre file system 4

Cray ClusterStor E1000 Storage System overview 6

 Unique design philosophy 6

 Key technologies 8

 Subsystems 12

 Interconnect options and considerations 16

 Performance considerations 17

 File system configuration options and considerations 18

 Intra- and inter-file system data movement considerations 20

 Cray ClusterStor and the Lustre community 22

Conclusion 23



EXECUTIVE SUMMARY

"We're not doing our grandfather's HPC here," said Rick Stevens from Argonne National Laboratory in a recent expert panel on "Exascale Day."¹ Like Rick Stevens, we firmly believe that the confluence of AI/machine learning with traditional simulations running on one supercomputer or HPC clusters will transform the very nature of high-performance computing. We call this era of converged workloads the exascale era.

The workload convergence of classic modeling and simulation with AI/machine learning is happening right now. A recent study of the independent analyst firm Intersect360 found out that the majority (61%) of the HPC users already are running machine learning programs.² And an additional 10% of the respondents stated that they plan to do so by the end of the year 2020.

The workload convergence of classic modeling and simulation with AI has fundamental implications for HPC storage as the I/O profiles of the different workloads could not be more different:

- Modeling and simulation: Mainly **writing petabytes** of **large** files in **sequential** order
- AI/machine learning: Mainly **reading terabytes** of files of **all sizes** in **random** order

In the past modeling and simulation applications were running on CPU-nodes in supercomputers or HPC clusters while machine learning/deep learning applications were executed on GPU-systems like HPE Apollo 6500 or NVIDIA® DGX in a separate AI-dedicated infrastructure stack called **POD**.

The traditional storage systems that have served us well in the siloed past are breaking in the new era architecturally and economically:

- Classic mainly HDD-based HPC storage for modeling and simulation like Cray ClusterStor L300, DDN EXAScaler, or IBM ESS:
 - Strengths: Cost-effective and extremely scalable
 - Weaknesses: Not well suited to serve small, random I/O at high speeds
- Classic NFS-based All-Flash enterprise storage systems for AI PODs like NetApp AFF or Dell EMC Isilon F-Series:
 - Strengths: Well suited to serve small, random I/O at high speeds
 - Weaknesses: High cost per terabyte and limited scalability in a single namespace

The new era needs new HPC storage that combines the best of both worlds: As cost-effective and scalable as classic HPC storage and as good as serving small, random I/O as classic All-Flash enterprise NAS.

This new HPC storage is called the Cray ClusterStor E1000 storage system which is available as an HPE product starting in July 2020.

INTENDED AUDIENCE

This white paper is for everyone involved in the design of HPC storage solutions for large HPC clusters and supercomputers with hundreds or even thousands of CPU-based or GPU-accelerated compute nodes. This document is not intended for HPC users who are using CPU or GPU-accelerated compute nodes in low node count configurations for workgroup or departmental use. For those use cases HPE has a range of other storage solutions in the portfolio.

This document intends to give architects who are involved in the design of HPC storage solutions for large HPC clusters and supercomputers a first technical overview of the Cray ClusterStor E1000 storage system.

NOTE

This document is not a complete design guide and does not list all the available features or explain how to configure them. It is highly recommended to engage with HPE's storage architects when considering a Cray ClusterStor E1000 storage system to ensure that the specific configuration is tailored in the optimal way to the specific needs of the specific project.

¹ nextplatform.com/2019/10/22/exascale-is-not-your-grandfathers-hpc/

² Intersect360 HPC User Budget Map Survey: Machine Learning's Impact on HPC Environments, October 2019



LUSTRE FILE SYSTEM

The Cray ClusterStor E1000 storage system embeds the open-source parallel Lustre® file system for performance, scalability, and cost-effectiveness reasons. Around two thirds of the global top 100 supercomputers use Lustre in production.³ And a recent multiclient study of Hyperion Research regarding the file system landscape in the HPC ecosystem found that Lustre is the most widely used parallel file system and the only parallel file system that has shown consistent growth in the last years.

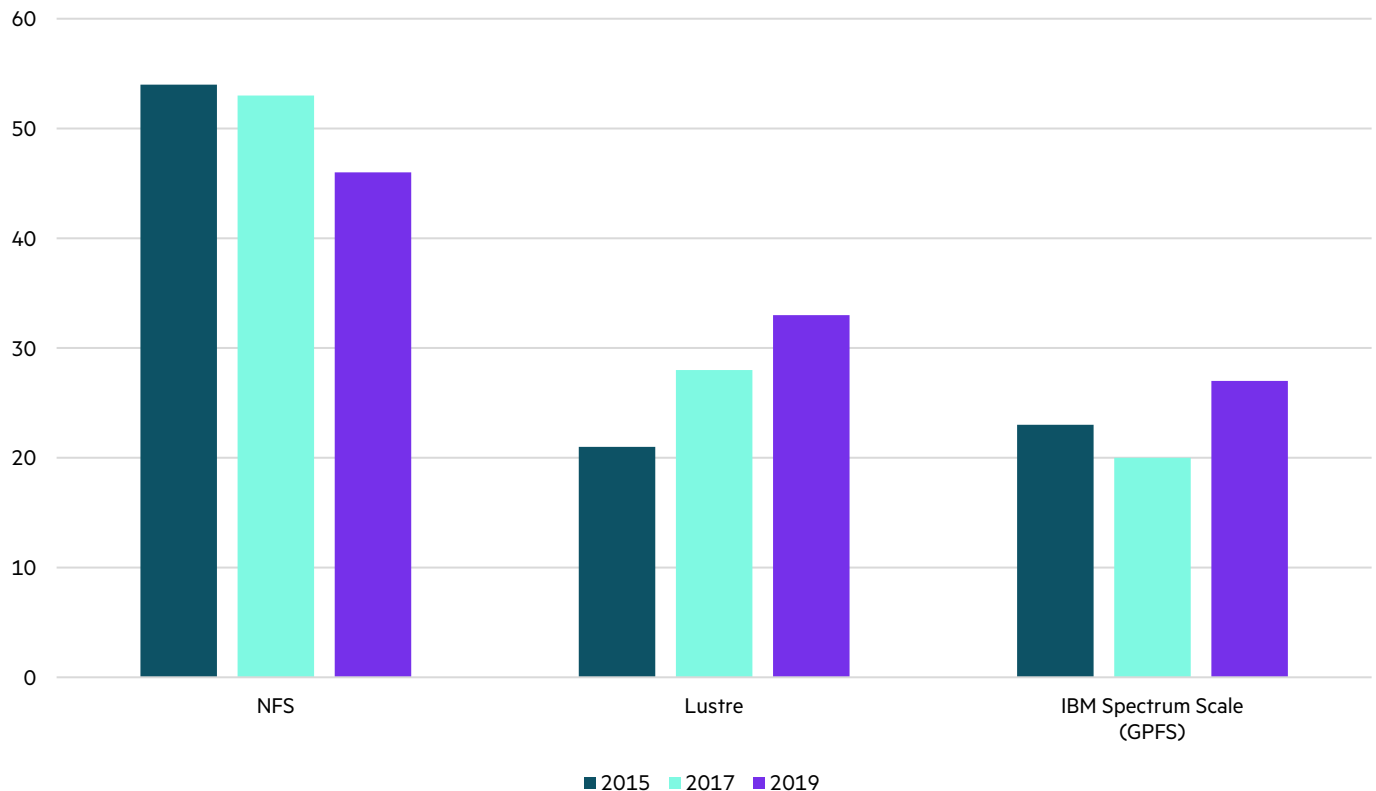


FIGURE 1. Top 3 file system adoption in HPC 2015–2019

Hyperion Research comments the finding of the study as follows:⁴

- While NFS remains the most widely adopted file system, it has dropped from being utilized at 54% of the sites down to 46%. NFS was one of the first file systems that could handle the initial scale of early HPC systems. Its first mover status, coupled with the fact that it's still adequate in smaller scale HPC systems today, is reflected in its continued, albeit shrinking, wide adoption.
- Lustre utilization has grown from 21% to 32.5%, and Lustre's open-source approach has enabled it to mature its feature set across a wide number of areas including performance, resiliency, reliability, and scalability. This maturity has also given it stability and the capability of scaling to meet the demands of petascale and emerging exascale configurations.
- GPFS/Spectrum Scale adoption has grown at a slightly lower rate, from 23% to 26.8%. This modest growth in adoption can be attributed to a slower feature advancement pace due to its proprietary nature along with a market shift away from IBM's dominance in the HPC sector and a change in the pricing model away from user-license-based to capacity-based.

This data confirms the selection of Lustre as the parallel file system for the new HPC storage for the new era: Cray ClusterStor E1000 storage system.

³ sc19.supercomputing.org/proceedings/bof/bof_pages/bof134.html

⁴ Hyperion Research, Special Study: Shifts Are Occurring in the File System Landscape, June 2020



In order to understand the system architecture of the Cray ClusterStor E1000 a basic understanding of the architecture of the Lustre parallel file system is required. For current users of NFS storage—who are outgrowing the scalability limitations of NFS—and for current users of IBM Spectrum Scale—who are struggling to afford the **software license tax** considering their constantly growing capacity requirements—we will review the fundamentals of the Lustre file system on the next section.

NOTE

Current users of Lustre are encouraged to go directly to the section [Cray ClusterStor E1000 storage system overview](#).

The following description of the basic architecture of Lustre is based on the introduction to Lustre on the home page of the Lustre community.

Lustre⁵ is an open-source, global single-namespace, POSIX-compliant, distributed parallel file system designed for scalability, high-performance, and high availability. Lustre runs on Linux®-based operating systems and employs a client-server network architecture. Storage is provided by a set of servers that can scale to populations measuring up to thousands hosts. Lustre servers for a single file system instance can, in aggregate, present up to multiple exabytes of storage to thousands of compute clients, with up to terabyte-per-second of combined throughput.

Redundant servers support storage failover, while metadata and data are stored on separate servers, allowing metadata and data to scale independently from each other which enables the extreme scalability Lustre is famous for. Lustre can deliver fast I/O to applications across high-speed network fabrics, such as HPE Slingshot, InfiniBand, and Ethernet.

Lustre's architecture uses distributed, object-based storage managed by servers and accessed by client compute nodes using the efficient network protocol, LNet. There are metadata servers, responsible for storage allocation, and managing the file system namespace, and object storage servers, responsible for the data content itself. A file in Lustre is comprised of a metadata inode object and one or more data objects.

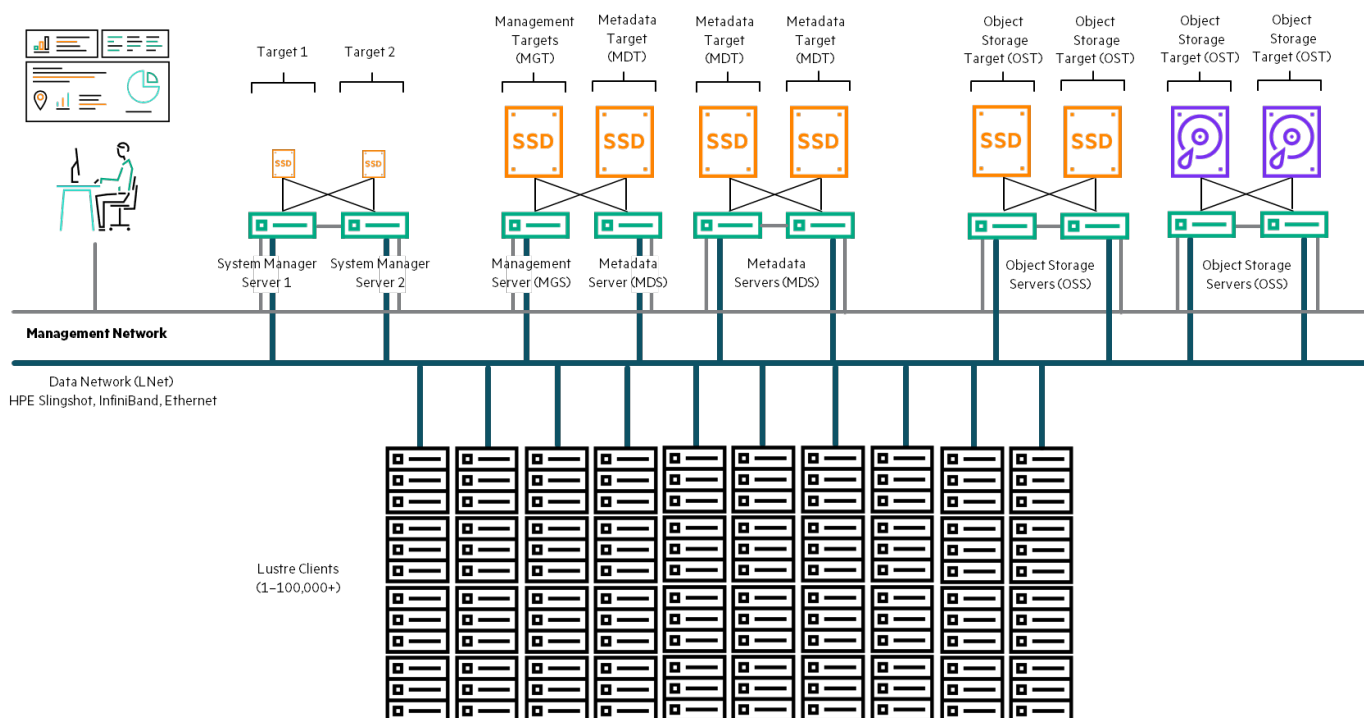


FIGURE 2. Lustre file system architecture

⁵ [wiki.lustre.org/Introduction to Lustre](http://wiki.lustre.org/Introduction%20to%20Lustre)

The major components of a Lustre storage system are:

- **Management service** with management servers (MGS) and management targets (MGT): Provides a registry of all active Lustre servers and clients, and stores Lustre configuration information. MGT is the management service storage target used to store configuration data.
- **Metadata service** with metadata servers (MDS) and metadata targets (MDTs): Provides file system namespace (the file system index), storing the inodes for a file system. MDT is the metadata storage target, the storage device used to hold metadata information persistently. Multiple MDS and MDTs can be added to provide metadata scaling.
- **Object storage services** with object storage servers (OSS) and object storage target (OST): Provides bulk storage of data. Files can be written in stripes across multiple object storage targets (OSTs). Striping delivers scalable performance and capacity for files. OSS are the primary scalable service unit that determines overall aggregate throughput and capacity of the file system.
- **Clients** (mainly compute nodes in supercomputers or HPC clusters): Lustre clients mount each Lustre file system instance using the Lustre Network protocol (LNet). Lustre represents a POSIX-compliant file system to the OS. Applications use standard POSIX system calls for Lustre I/O, and do not need to be written specifically for Lustre.
- **Network:** Lustre is a network-based file system, all I/O transactions are sent using network RPCs. Clients have no local persistent storage and are often diskless. Lustre network I/O is transmitted using the LNet protocol which has native support for TCP/IP networks, for the RDMA over Converged Ethernet (RoCE) protocol as well as RDMA networks such as InfiniBand.

A single Lustre file system can scale linearly based on the number of building blocks. The minimum high availability (HA) configuration for Lustre is a management building block (with redundant MGS and MGT), a metadata building block (with redundant MDS and MDT), plus at least one object storage building block (with redundant OSS and OST). Using these basic units, one can create file systems with hundreds of OSS/OST as well as several MDS/MDT to provide a reliable, high-performance shared storage platform.

To increase capacity and throughput in a single namespace, simply add more servers with the required storage. Lustre will automatically incorporate new servers and storage, and the clients will leverage the new capacity automatically. New capacity is automatically incorporated into the pool of available storage.

CRAY CLUSTERSTOR E1000 STORAGE SYSTEM OVERVIEW

The Cray ClusterStor E1000 storage system is a unique new HPC storage system for the new HPC era. This section gives the audience an overview over the design philosophy, the key technologies, and the modular subsystems. The Cray ClusterStor E1000 attaches directly to any supercomputer or any HPC cluster of any vendor as long as the compute supports modern high-speed networks like EDR/HDR InfiniBand, 100/200 Gigabit Ethernet or HPE Slingshot. Connectivity to legacy interconnects (for example, Intel® Omni-Path) can be realized via LNet routers.

Unique design philosophy

Designing, architecting, implementing, maintaining, and operating a Lustre file system DIY (do it yourself) based on a reference architecture can be hard, highly complex, and time consuming.

One of the key design points of Cray ClusterStor parallel storage systems is to completely shield the customer from that complexity by delivering an engineered system that:

- Ships from the factory after extensive soak-testing in the optimal configuration for the requirements of the customer
- Contains all the required, previously discussed components of a Lustre file system in a **single system image** system (see the following figure)
- Arrives at the customers' premises fully integrated with the file system already preinitialized



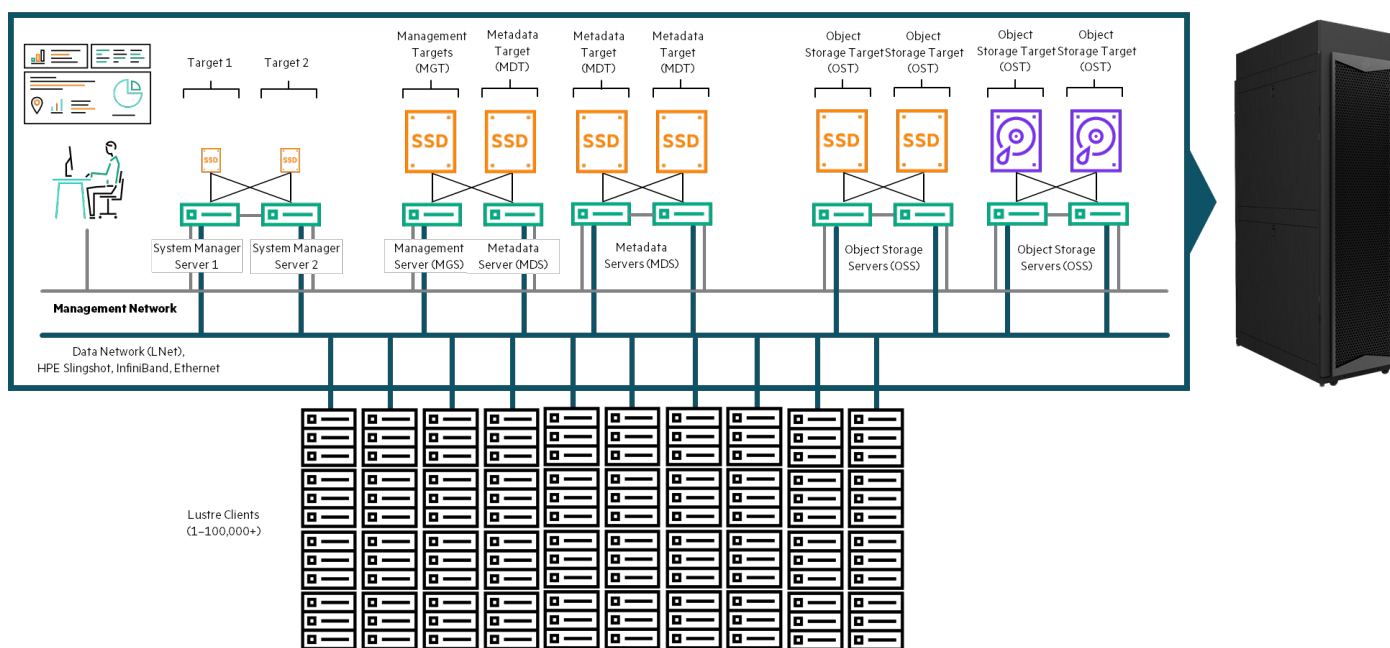


FIGURE 3. Cray ClusterStor E1000: An engineered parallel HPC storage system

The other critical questions in the design phase of every Cray ClusterStor storage system are:

- **How can we design HPC storage that delivers the highest performance through the file system to the compute nodes?**

The more data we can bring from/to the compute nodes per second the less time they spend idle waiting for their I/O to happen. This is critical for the return of investment as typically most of the investment in a supercomputer or HPC cluster resides in the CPU/GPU compute nodes.

- **How can we deliver that leading performance with the least amount of storage drives?**

More than 60% of the cost of any HPC storage system resides in the storage drives—SSDs or HDDs. Being able to deliver a given storage performance requirement with significantly fewer storage drives also means fewer storage enclosures, fewer servers, less power consumption, less floor space, and lower support cost over the lifetime of the storage system. We call that metric performance efficiency.

- **How can we do all of this using an open-source parallel file system that we can support with our own R&D team?**

The software licensing cost for proprietary parallel file systems that charge a software license tax per storage drive or per terabyte storage capacity easily can destroy the cost efficiency of the most efficient hardware storage architecture. For that reason Cray ClusterStor is embedding the open-source file system Lustre that is owned by a vibrant open-source community and not by a single company. Our own in-house Lustre development team supports the file system up to level 3 so that HPE can provide enterprise-grade support for customers in-house.

- **How can we deliver all of this as a high availability (HA) system without a single point of failure?**

The insights generated by HPC systems using methods like modeling and simulation, AI/machine learning or high-performance data analysis increasingly are becoming more and more mission or business critical for the continuous success of public sector organizations or enterprises.



Key technologies

End-to-end PCIe 4.0 storage controller

The design of the storage controller is one of the most important design decisions for any storage system as it fundamentally influences both the performance levels and the efficiency levels of how you deliver that performance with your storage system.

And as you can see in the following figure, HPC storage buyers value performance (#1 with 57%) and performance efficiency (tied on #2 and #3 with 37%) among all other things.

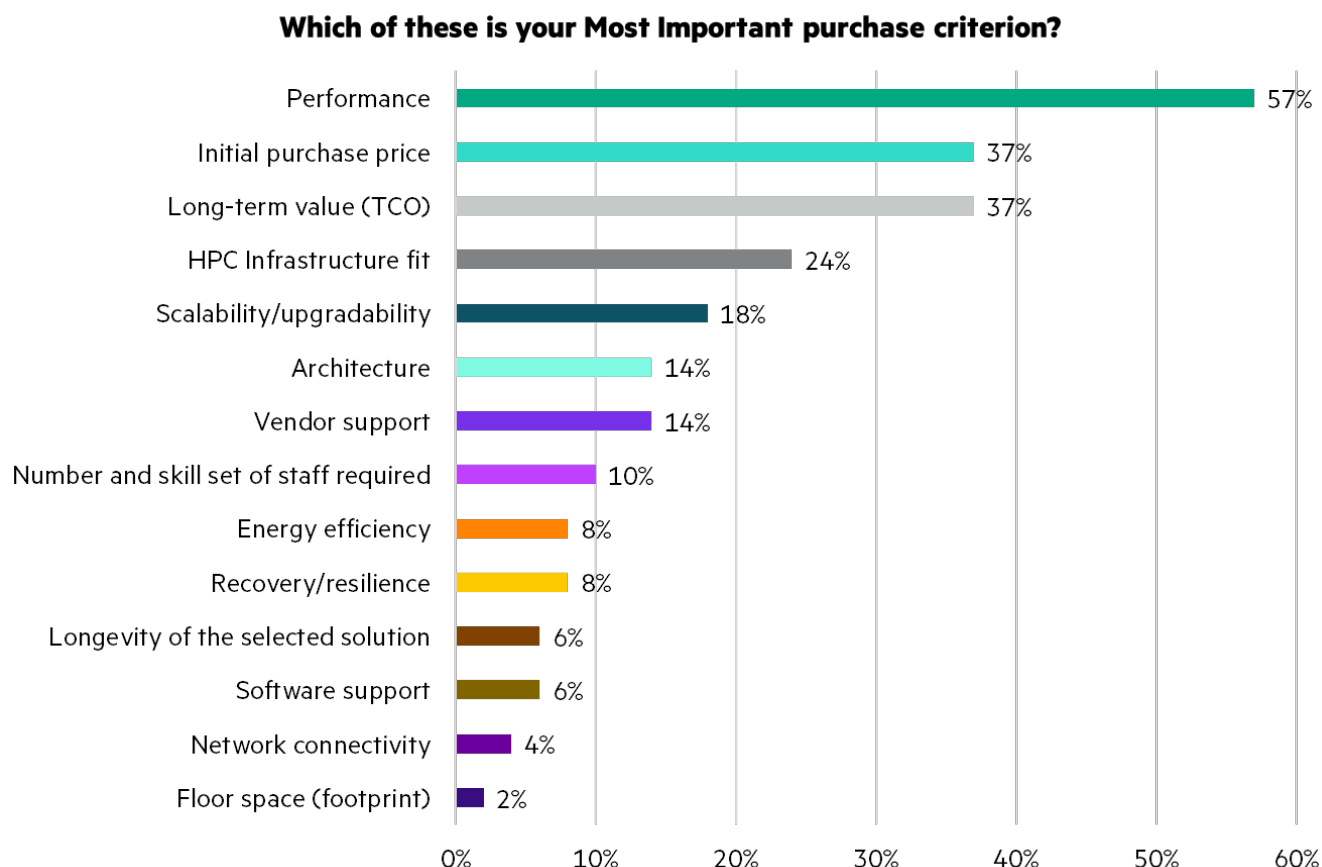


FIGURE 4. Most important purchase criteria of HPC storage buyers according to Hyperion Research⁶

⁶ Hyperion Research, The hidden costs of HPC storage, May 2020



In order to efficiently exploit both the performance of modern NVMe Gen4 SSDs and the throughput capabilities of modern 200 gigabit per second networks we knew that we need a balanced, zero bottleneck design like the one shown in the following figure.

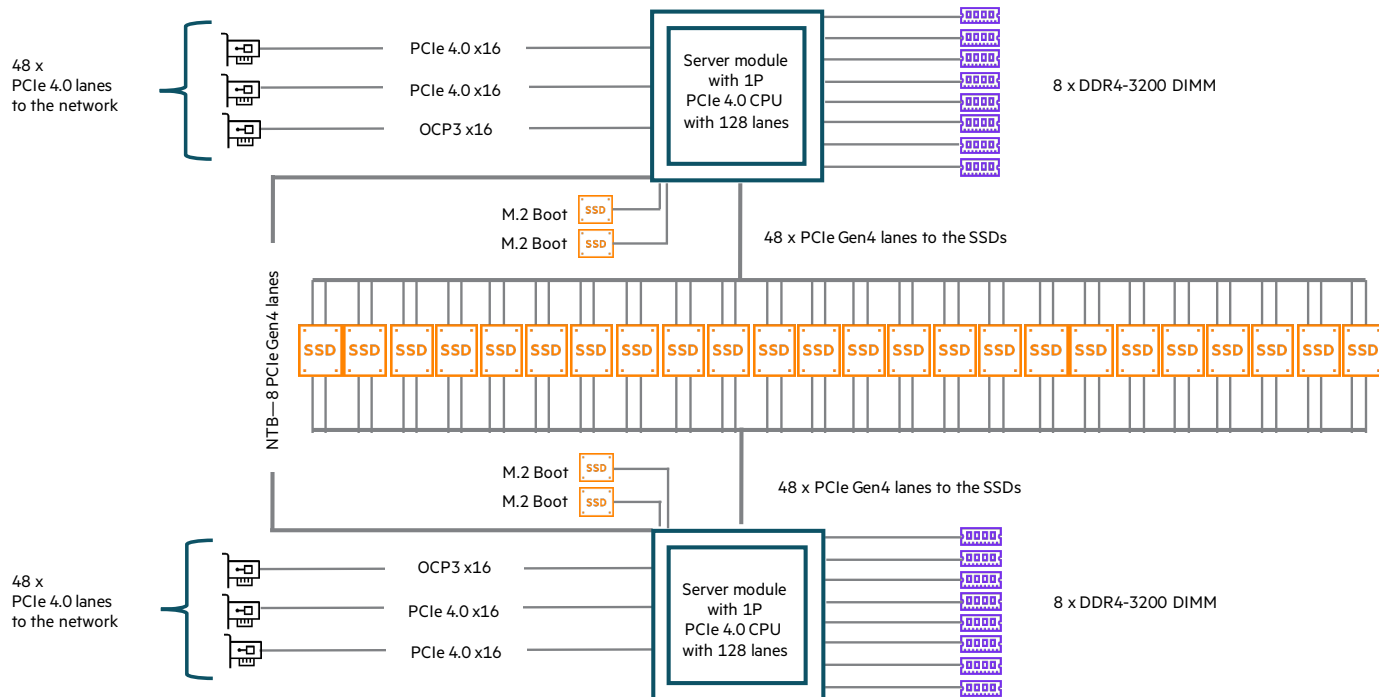


FIGURE 5. Next-generation storage controller: Balanced, zero-bottleneck design

In order to satisfy the demands of the next era, we partnered with an Original Design Manufacturer (ODM) to realize that ideal controller for the Cray ClusterStor E1000 storage system in a 2U form factor as shown in the following figures.

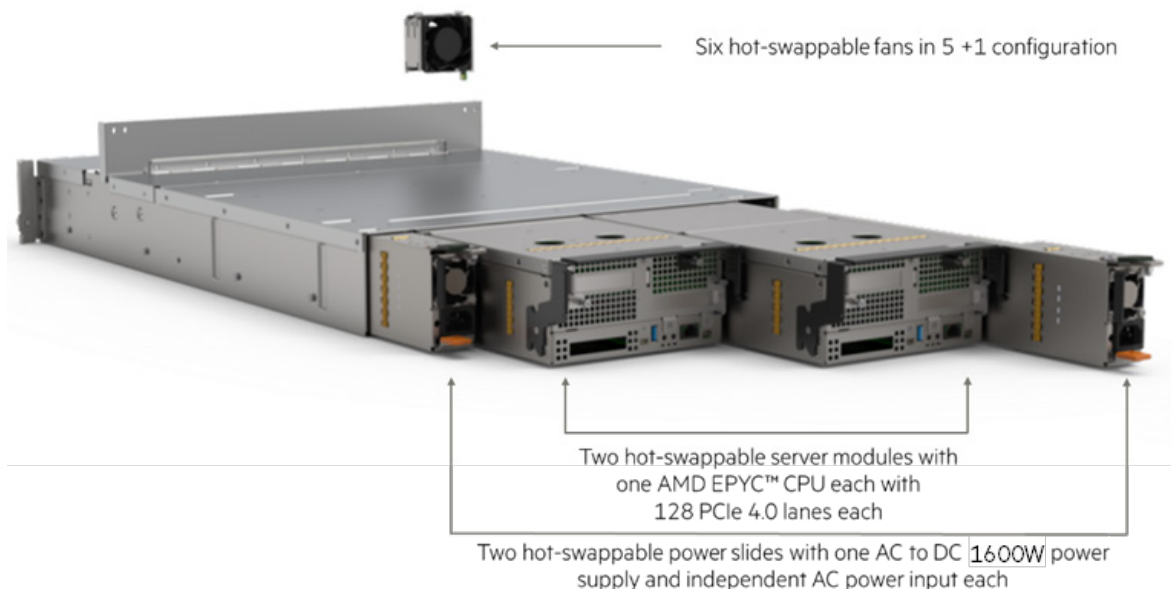


FIGURE 6. Cray ClusterStor E1000 storage controller: Rear view



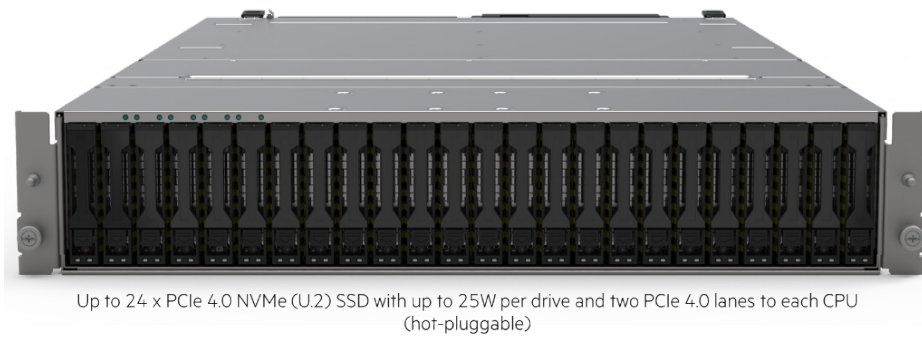


FIGURE 7. Cray ClusterStor E1000 storage controller: Front view

In different configurations this purpose-engineered storage controller serves as the common building block of the Cray ClusterStor E1000 storage system for deploying Cray ClusterStor Manager servers, Lustre management servers (MGS) and management targets (MGT), metadata servers (MDS) and metadata targets (MDTs) as well as object storage servers (OSS) and object storage targets (OSTs) based on NVMe SSDs. Object storage targets based on 7.2K RPM SAS HDDs are deployed by attaching high-density HDD enclosures to the storage controller via fast and redundant SAS connections.

The high-density HDD enclosure is introduced in the next section.

High-density HDD enclosure

The high-density HDD enclosure can fit 106 large form factor HDDs in just 4 rack units. Currently, the enclosure supports 4, 6, 10, 12, 14, and 16 TB 7.2K RPM 12 Gb/s SAS HDDs.

The high-density HDD enclosure connects with 2X PCIe 12 Gb/s as HBAs per node connecting to the previously discussed 2U storage controller. The high-density 4U 106 HDD enclosure is designed as a fully redundant storage array with 2X SAS I/O modules and fully redundant internal SAS connectivity.

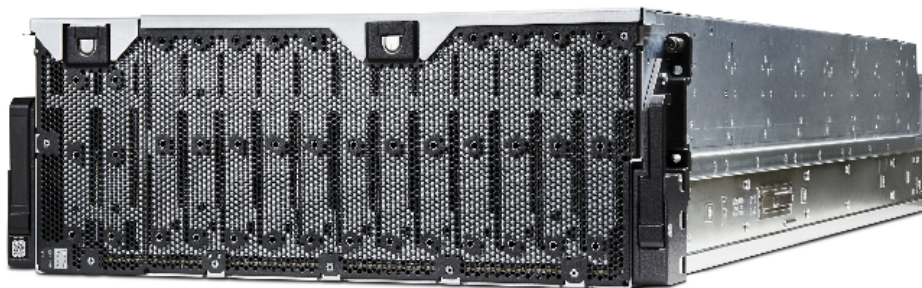


FIGURE 8. Front view



FIGURE 9. Rear view



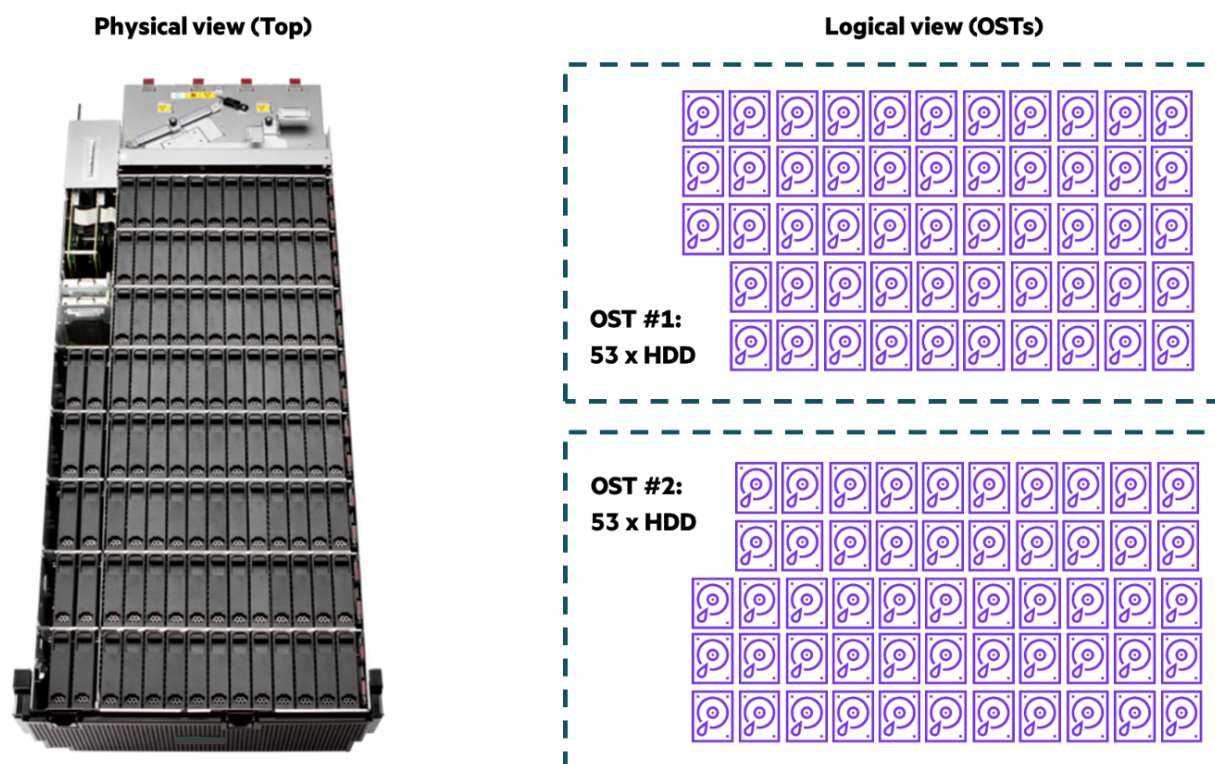


FIGURE 10. Top view and logical view

The high-density HDD enclosures always ship fully populated with 106 HDD with two OSTs consisting of 53 drives each that are organized differently depending on the selected back-end file system and data integrity concepts (see next section).

Back-end file system options and data integrity concepts

Lustre is a distributed, parallel file system that sits on top of a set of local file systems. To ensure harmony and consistency between the two file system layers, Lustre stores data in object format. The Cray ClusterStor E1000 storage system supports both LDISKFS and OpenZFS as back-end file system. Both back-end file systems support end-to-end data integrity (LDISKFS with T10-PI) to prevent silent data corruption.

Most of the Cray ClusterStor E1000 storage systems are shipping with LDISKFS while some very large systems are already contracted with OpenZFS as back-end file system. The choice between LDISKFS and OpenZFS depends on the specific customer requirements. LDISKFS is significantly faster than OpenZFS (more than 30% in aggregate bandwidth for OSTs and about 50% in metadata performance for MDTs) but rebuilding parity on OpenZFS is faster than LDISKFS (depending on the amount of data in the system). Also, OpenZFS offers some additional functionalities such as file system snapshots and compression.

IMPORTANT

The choice between LDISKFS and OpenZFS must be made carefully in the system design phase as once deployed a full reinstallation is required to change the underlying back-end file system.

Both back-end file systems need some form of RAID to protect from data loss in case of storage drive failures. Parity declustered RAID provides the best capacity option typically using a number of data blocks protected by two parity blocks. While this is similar to RAID 6, parity declustered RAID can use many more drives in case of a drive failure which dramatically improves rebuild times after a disk failure which is important with currently shipping 16 TB drive sizes. The ClusterStor E1000 storage system supports either GridRAID (with LDISKFS) or dRAID (with OpenZFS) for OSS/OSTs. However, one drawback with these RAID formats is that while they excel at throughput in gigabyte per second (GB/sec), input/output operations per second (IOPS) suffer. Therefore, Cray ClusterStor E1000 storage system uses RAID 10 (**mirroring**) for all IOPS-sensitive metadata services. The trade-off with RAID 10 capacity overhead and streaming performance penalty is that performance of small block I/O (IOPS) benefits tremendously from RAID 10. For SSD-based OSTs GridRAID is the standard, but for IOPS-intensive workloads RAID 10 is offered as an option.



As LDISKFS with GridRAID is the most widely deployed option for object storage services the following paragraphs describe GridRAID on a high level.

GridRAID is a proven parity declustered RAID concept that has been shipping in ClusterStor systems for more than seven years. It has proven to be very faster than a traditional RAID 6 solution with no compromise to reliability. As disk drives increase in capacity, the time it takes to rebuild a failed drive in an OST also increases significantly. Today, rebuilding a 10 TB drive configured as a RAID 6 volume, can easily take 24 to 36 hours. During rebuild of RAID 6 volumes, a significant performance impact is experienced and the risk of a second disk failure increases.

The high-density HDD enclosures of the Cray ClusterStor E1000 storage system feature two OSTs with 53 HDDs each. While GridRAID still uses eight data blocks and the corresponding two parity blocks of traditional RAID 6, the blocks are semi-randomly distributed over all drives to increase data reliability and distribute load as evenly as possible over the entire OST. GridRAID also uses distributed spare stripes for fast repair in case of a disk failure. This means that when a drive fails, the rebuilt data is read from 52 drives and written to 52 drives. Under normal use, a full rebuild can be performed without any significant performance degradation. As an example, a 10 TB drive can be rebuilt in less than five hours. Rebuild speed can be changed by the storage administrator during normal operation. Figure 11 illustrates the conceptual layout of the GridRAID with semi-randomly distributed data and parity blocks as well as distributed spare blocks (white blocks) for rapid rebuild.

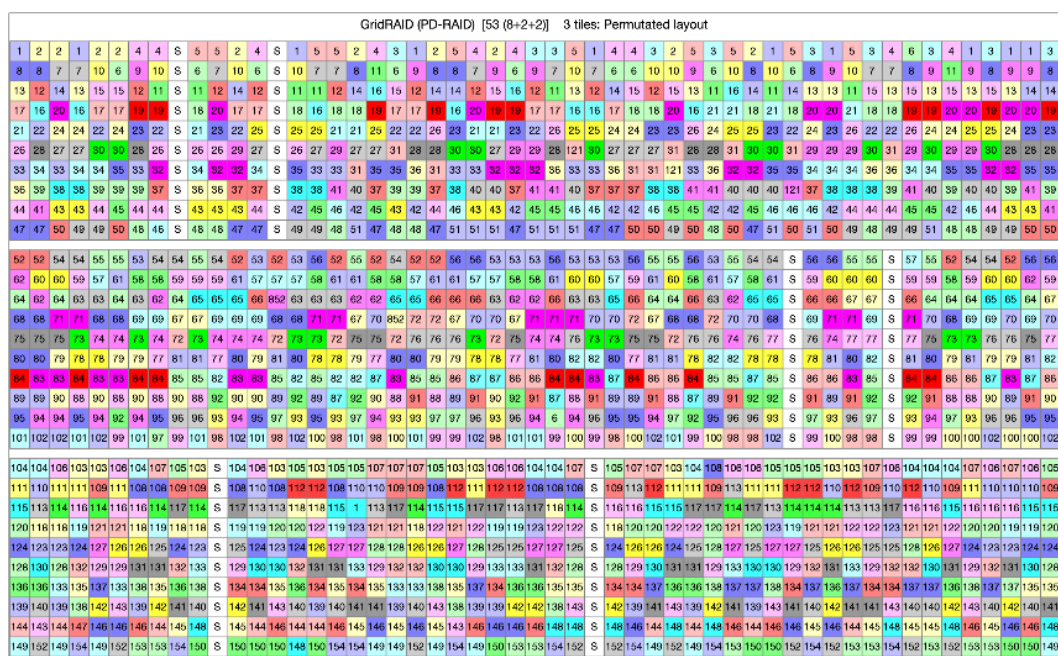


FIGURE 11. GridRAID layout for OST with 53 HDDs in high-density HDD enclosure

Subsystems

At a minimum a Cray ClusterStor E1000 storage system consists of three different subsystems (in addition to the needed network infrastructure):

- One Cray ClusterStor E1000 System Management Unit (SMU): Providing the management services for the storage system
- At least one Cray ClusterStor E1000 Metadata Unit (MDU): Providing the Lustre Metadata Servers (MDS) and the Lustre Metadata Targets (MDTs)
- At least one Cray ClusterStor E1000 Scalable Storage Unit (SSU): Providing the Lustre Object Storage Servers (OSS) and Object Storage Targets (OSTs)

Metadata performance of the file system can be scaled by adding up many MDUs per file system using Distributed Namespace Environment (DNE) functionality.

Aggregate throughput performance and usable storage capacity of the namespace is scaled out linearly by adding SSUs. The ClusterStor E1000 storage system will automatically incorporate new SSUs. New capacity is automatically incorporated into the appropriate pool (disk or flash) of available storage in the single namespace.



The first Cray ClusterStor E1000 rack is called the base rack containing the management network switches, the interconnect switches, one SMU, at least one MDU and the initial SSU(s). Once the space in the base rack is filled up with SSUs the system continues to scale with additional SSUs in expansion racks.

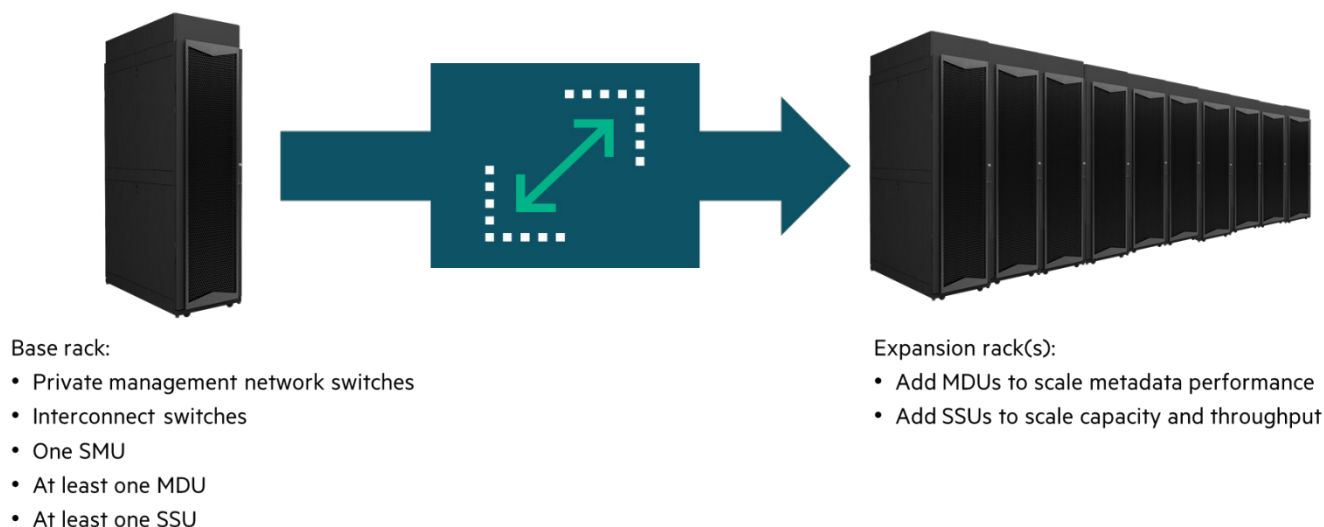


FIGURE 12. How to scale a single namespace with the Cray ClusterStor E1000 storage system

System Management Unit (SMU)

The System Management Unit (SMU) is responsible for the system management of the Cray ClusterStor E1000 storage system. All management and administration of the Cray ClusterStor E1000 storage system runs on the SMU, but the SMU does not play an active operational role in the Lustre file system or the datapath itself. The SMU consists of a high-availability management server pair with NVMe Gen4 RAID protected SSDs for storing system management data, logging, and boot services. The SMU is connected to all subsystems via a private 1GbE local management network implemented with Aruba CX 6300M switches as well as through the selected high-speed data network (InfiniBand EDR/HDR or 100/200 Gigabit Ethernet or HPE Slingshot).

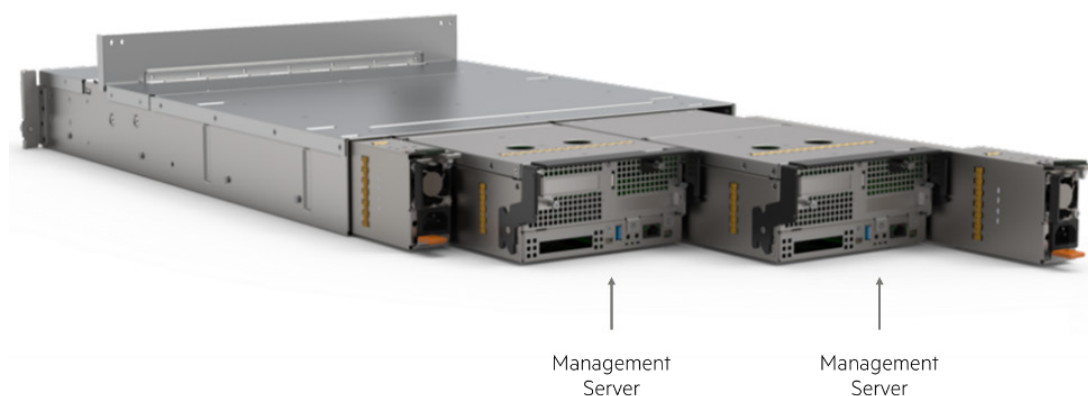


FIGURE 13. Cray ClusterStor E1000 storage controller in SMU configuration

The factory integrated application, ClusterStor Manager (CSM) runs on the SMU and provides the system management interface for the file system. All Cray ClusterStor systems come with an instantiation of CSM in high-availability mode to simplify the management of the Lustre file system. Whether using the GUI or the CLI, most standard functions are scripted and simplify everyday administration.

In addition to system statistics and inventory lists, CSM also provides:

- Simplified node control (for example, power status, manual failover/failback)
- Enhanced system performance details



- Management of support bundles
- Component health overview and drill downs
- Additional configuration details (for example, authentication settings, routing information, network, and RAID settings)

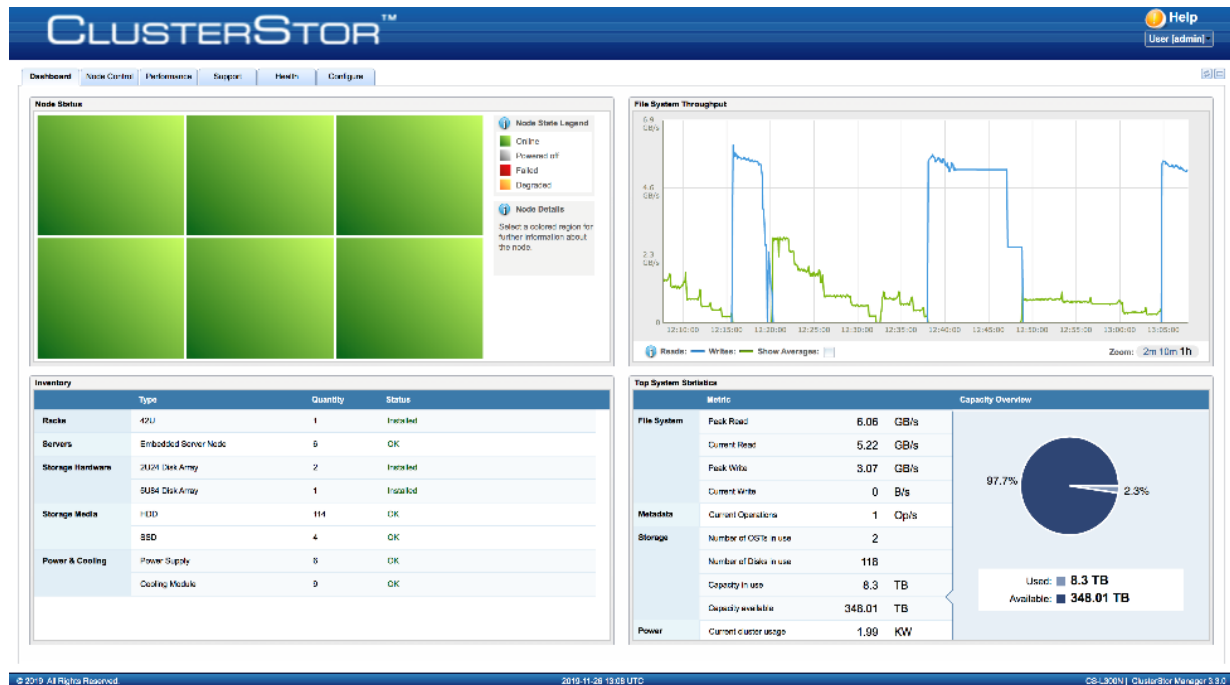


FIGURE 14. Graphical User Interface (GUI)

Metadata Unit (MDU)

The Metadata Unit (MDU) provides the Metadata Server (MDS) and Metadata Target (MDT) functionality for the Lustre file system.

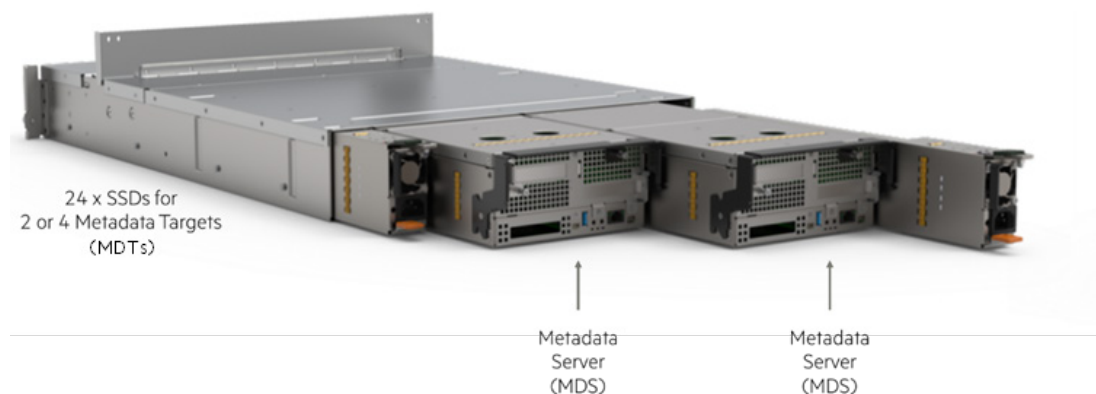


FIGURE 15. Cray ClusterStor E1000 storage controller in MDU configuration

One MDU supports up to 8 billion inodes per MDU in standard configuration (4 billion per MDS).

Many MDUs are supported in one file system using the [Distributed Namespace Environment Phase 2 \(DNE2\)](#) functionality of Lustre that enables to spread a single large directory across multiple MDT. The test results given in Figure 16 demonstrate the scalability by reaching close to linear performance.

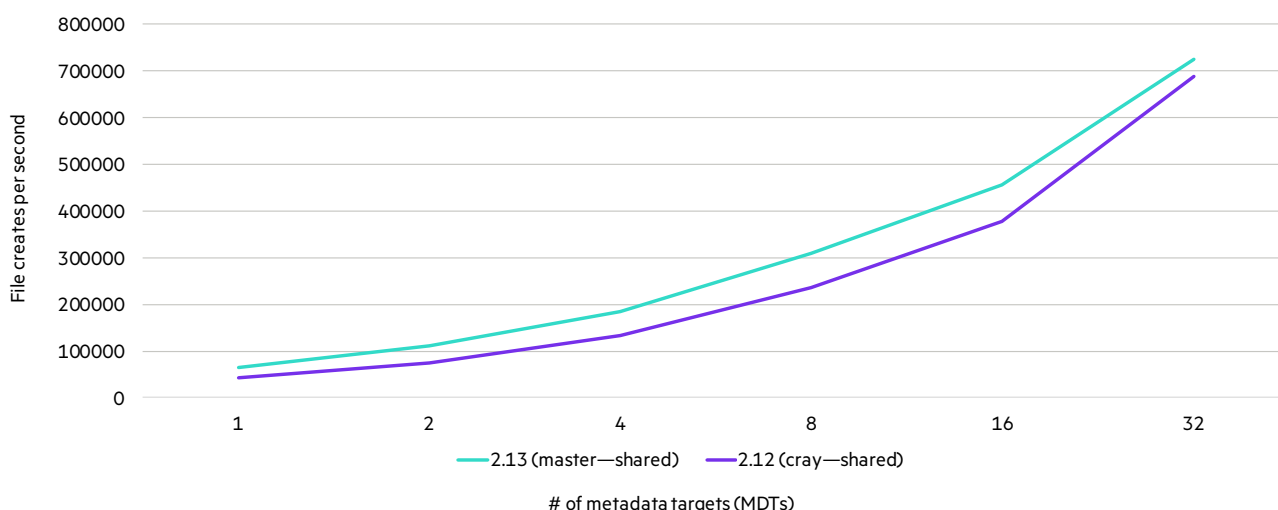


FIGURE 16. DNE2 Single Shared Directory—OK file creates/sec scaling up to 32 MDTs

In the past Lustre was optimized for large file performance. In the new era modern Lustre provides new functionality to significantly improve small file performance. The [Data on MDT \(DoM\)](#) feature allows the data for small files to be placed on the flash-based MDT so that additional RPCs to the OST can be eliminated and performance for small files correspondingly improved. System administrators can now set a layout policy that locates small files to be stored directly on an MDT.

The MDTs in the MDU always are organized in mirrored RAID configurations to optimize IOPS and small file performance. RAID 10 if LDISKFS is selected as back-end file system and dRAID 1 if OpenZFS is selected.

The standard configuration for the MDU is two uplinks to the selected high-speed network with an option to connect with four uplinks for very high-performance requirements.

NOTE

For most cases one MDU in standard configuration per file system will be enough to provide the required metadata performance. However it is highly recommended to engage with HPE's storage experts for correctly sizing both metadata and throughput performance for the specific workloads of the specific customer environment.

Scalable Storage Unit (SSU)





To increase storage capacity and throughput of the file system, simply add SSUs with two OSS and two or more OSTs. The ClusterStor E1000 storage system will automatically incorporate new SSUs. New capacity is automatically incorporated into the appropriate pool (disk or flash) of available storage in the single namespace.

IMPORTANT

Expanding the file system by design always happens by adding SSUs (OSS and OSTs) in order to provide linear performance scalability with predictable performance. Adding only OSTs (for example in an HDD enclosure) without adding the corresponding OSS is not supported by design.



TABLE 1. SSUs of the Cray ClusterStor E1000 storage system

Specifications	All-Flash SSU (SSU-F)	Disk SSU with 1 HDD enclosure (SSU-D1)	Disk SSU with 2 HDD enclosures (SSU-D2)	Disk SSU with 4 HDD enclosures (SSU-D4)
OSS/OST	<ul style="list-style-type: none"> • 2 OSS in 2U storage controller • 2 OSTs in same 2U storage controller 	<ul style="list-style-type: none"> • 2 OSS in 2U storage controller • 2 OSTs in 4U HDD storage enclosure (SAS-attached) 	<ul style="list-style-type: none"> • 2 OSS in 2U storage controller • 4 OSTs in 4U HDD storage enclosures (SAS-attached) 	<ul style="list-style-type: none"> • 2 OSS in 2U storage controller • 8 OSTs in 4U HDD storage enclosures (SAS-attached)
Form factor	2 rack units	6 rack units	10 rack units	18 rack units
Uplink ports to high-speed network	4 or 6	2 or 4	2 or 4	2 or 4
Physical view				
Storage drives	24 x NVMe Gen4 SSD (U.2)	106 x 3.5" 7.2K SAS HDD	212 x 3.5" 7.2K SAS HDD	424 x 3.5" 7.2K SAS HDD
Supported drive sizes	1.92/3.84/7.68/15.36 TB (1 DWPD*)	4/6/10/12/14/16 TB	4/6/10/12/14/16 TB	4/6/10/12/14/16 TB
Minimum usable capacity (with LDISKFS/GridRAID)	32 TB (1.92 TB SSD)	314 TB (4 TB HDD)	628 TB (4 TB HDD)	1,255 TB (4 TB HDD)
Maximum usable capacity (with LDISKFS/GridRAID)	258 TB (15.36 TB SSD)	1,255 TB (16 TB HDD)	2,510 TB (16 TB HDD)	5,020 TB (16 TB HDD)

* 3 DWPD SSD available as an option: 1.6/3.2/6.4/12.8 TB

The Cray ClusterStor E1000 storage system supports the combination of different SSU types in the same files system/namespace. That means that three types of file systems can be designed:

- All-Flash file system only deploying All-Flash SSU-F SSUs. Delivered in standard 19" 42U data center racks.
- HDD file system only deploying SSU-Dx SSUs. Delivered in strengthened Cray ClusterStor storage racks needed due to the weight of the high-density HDD enclosures.
- Tiered file system deploying SSU-F for the flash pool (delivering most of the performance) and SSU-Dx for the HDD pool (providing most of the storage capacity). Delivered in strengthened Cray ClusterStor storage racks needed due to the weight of the high-density HDD enclosures.

Interconnect options and considerations

The Cray ClusterStor E1000 storage system supports the direct attachment to any supercomputer or HPC cluster that uses one of the following interconnects:




- InfiniBand EDR or HDR
- 100 or 200 Gigabit Ethernet
- 200 Gbps HPE Slingshot

The Cray ClusterStor E1000 storage system connects to compute clusters using legacy interconnects like Intel Omni-Path with [LNet routers](#).

The previously described subsystems of the Cray ClusterStor E1000 storage system (SMU, MDU, SSU) connect with uplinks to a pair of redundant top of rack (TOR) switches in the base rack depending on the selected interconnect type as shown in the Table 2.



TABLE 2. TOR switches

Specifications	Cray ClusterStor E1000 attached to HPC cluster with InfiniBand EDR or HDR	Cray ClusterStor E1000 attached to HPC cluster with 100/200 Gigabit Ethernet	Cray ClusterStor E1000 attached to HPE Cray supercomputer with 200 Gbps HPE Slingshot
TOR switches	QM8790 Mellanox Quantum™ HDR Edge Switch	Arista 7060X4 Series 100/200/400G Data Center Switches	HPE Slingshot TOR switch
Ports	<ul style="list-style-type: none"> 40 x HDR 200 Gb/s ports 80 x HDR100 Gb/s ports (using splitter cables) 	<ul style="list-style-type: none"> 32 x 400GbE QSFP-DD ports All ports allow a choice of speeds including 400GbE, 200GbE, or 100GbE 	<ul style="list-style-type: none"> 64 x 200 Gbps ports (using splitter cables)
Form factor	1U	1U	1U
Physical view			

Once the number of required uplink ports from the subsystems exceeds the available ports additional TOR switches are added to the Cray ClusterStor E1000 storage racks.

NOTE

Cray ClusterStor E1000 storage systems attaching to the supercomputer or HPC cluster via either InfiniBand or HPE Slingshot are delivered from the factory with the TOR switches integrated in the storage racks. When customers want to provide the 100/200GbE switches of their choice, those Cray ClusterStor E1000 systems ship without the 100/200GbE TOR switches from the factory. Today the Arista 7060X4 Series is supported for customer-provided 100/200GbE TOR switches with certification of other 100/200GbE switch vendors (for example, Cisco) in progress.

IMPORTANT

The Cray ClusterStor E1000 storage system depends on modern 200 Gbps networks to deliver the full performance of the system to the compute nodes (see next section). While 100 Gbps networks are supported, it is highly recommended to deploy the Cray ClusterStor E1000 storage system with 200 Gbps networks.

Performance considerations





Performance in real-world production environments is highly dependent on the following factors:

- Specific I/O profiles of the individual workloads/applications in the environment
 - Most important factor that requires careful sizing together with our HPC storage architects and performance benchmarking experts to ensure that the reality of sustained performance in production meets the expectations
- Selected back-end file system (LDISKFS or OpenZFS)
 - Typically LDISKFS delivers about one third more throughput performance than OpenZFS
- Selected data integrity option (parity declustered RAID or mirroring)
 - Typically parity declustered RAID delivers significantly more throughput performance while mirrored RAID excels at IOPS performance
- Data rate of the selected high-speed network (100 Gbps or 200 Gbps networks)
 - Only 200 Gbps networks can deliver the full performance. While 100 Gbps networks are supported about one third of the inherent performance of the SSUs is not realized as the performance is bound by the network

For illustration purposes the following table shows maximum throughput performance numbers in IOR benchmarks observed in our performance testing labs with LDISKFS as back-end file system, GridRAID as data protection concept and InfiniBand HDR as high-speed interconnect.



TABLE 3. Performance and performance efficiency

Specifications	All-Flash SSU (SSU-F)	Disk SSU with 1 HDD enclosure (SSU-D1)	Disk SSU with 2 HDD enclosures (SSU-D2)	Disk SSU with 4 HDD enclosures (SSU-D4)
Sequential read peak performance (IOR)	Up to 80 GB/sec	Up to 15 GB/sec	Up to 30 GB/sec	Up to 40 GB/sec
Sequential write peak performance (IOR)	Up to 50 GB/sec	Up to 15 GB/sec	Up to 30 GB/sec	Up to 40 GB/sec
Physical view				
Storage drives	24 x NVMe Gen4 SSD (U.2)	106 x 7.2K SAS HDD	212 x 7.2K SAS HDD	424 x 7.2K SAS HDD
Performance efficiency	<ul style="list-style-type: none"> Up to 3.3 GB/sec per SSD (Read) Up to 2.1 GB/sec per SSD (Write) 	Up to 141 MB/sec per HDD	Up to 141 MB/sec per HDD	Not applicable as SSU-D4 is a building block for cost-effective capacity and not for performance

As described in the design philosophy section at the beginning of the document a relevant gauge for HPC storage system cost efficiency is performance efficiency as shown in the last row of Table 3. We believe that the values in that table are a proof point for the fact that the Cray ClusterStor E1000 storage system not only delivers the performance that is needed in the new era but also delivers that performance in a very efficient way.

IMPORTANT

In order to ensure that sustained performance in actual production meets expectations it is highly recommended to engage with our HPC storage architects and performance benchmarking experts to custom-size Cray ClusterStor E1000 storage systems based on the specific workload mix of the specific HPC environment. The maximum performance numbers in Table 3 should not be used for actual sizing for sustained performance.

File system configuration options and considerations

The Cray ClusterStor E1000 storage system provides three fundamental ways to architect a file system/namespaces using fundamentally different storage media technologies with fundamentally different price points.

IDC forecasts⁷ that even by the year 2023 there still be a large difference (8X) in the price per terabyte between enterprise SSDs and nearline HDDs.

⁷ IDC: Worldwide 2019–2023 Enterprise SSD and HDD Combined Market Overview, June 2019



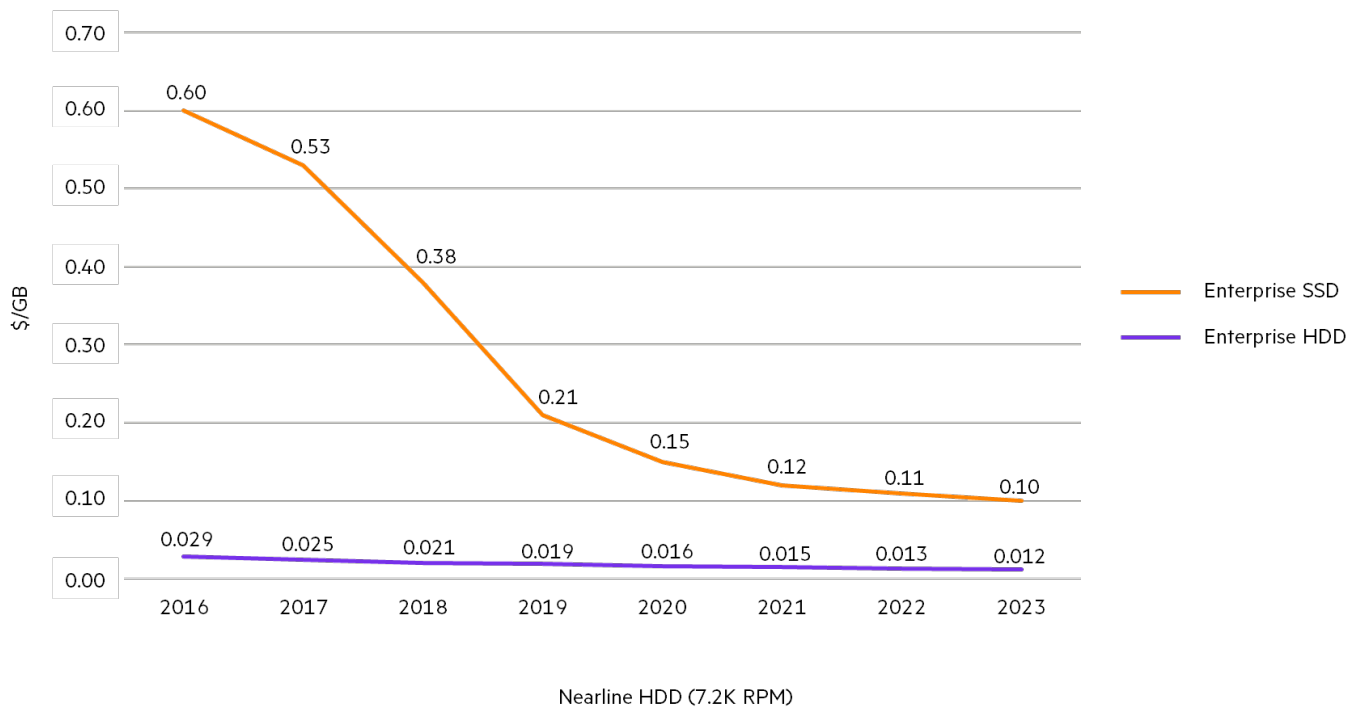


FIGURE 17. Enterprise blended HDD and SSD \$/GB trend comparison, 2016–2023

Based on those fundamental differences the design choices for a cost-effective Cray ClusterStor E1000 storage system can be summarized as follows:

1. Can the storage requirements be achieved with HDD-based SSUs without overprovisioning too much storage capacity?
If the answer is **Yes**, design HDD-based file system based on SSU-Dx.
2. If answer to 1 is **No**, then design a tiered file system with All-Flash SSU-F as flash pool providing most of the performance and SSU-Dx as disk pool providing most of the storage capacity.
3. Are the storage performance requirements very high with relatively low storage capacity requirements?
If the answer is **Yes**, design an All-Flash file system based on SSU-F.

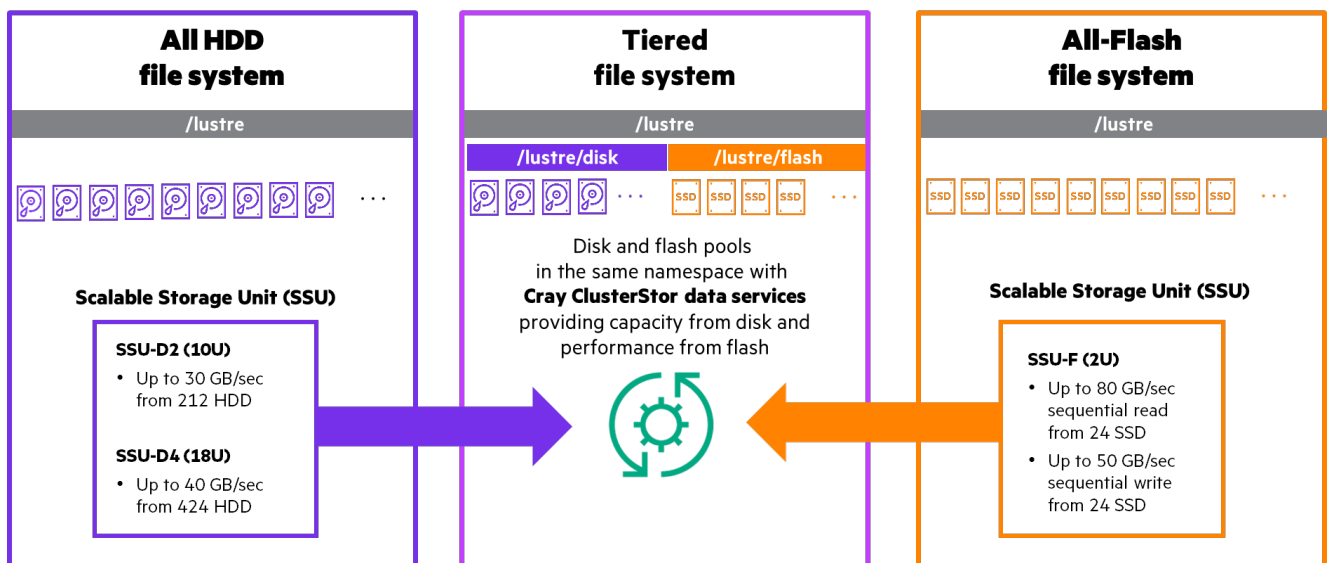


FIGURE 18. Fundamental file system design options with Cray ClusterStor E1000

Cray ClusterStor data services is a key capability for the tiered file system option which is briefly discussed in the next section.

Intra- and inter-file system data movement considerations

In the new era of coverage workloads it will be critical to have intelligent data management solutions in place that can move data within Lustre from fast flash pools to cost-effective HDD pools and vice versa (Intra-file system data movement). But it also will be key to have intelligent and automated data movement capabilities between different file systems (Inter-file system data movement) for the following use cases:

- Movement of data from one file system (for example, Lustre) to a different file system (for example, IBM Spectrum Scale or HPE XFS) as required by the workflow
- Automated data lifecycle management that creates space in the very fast parallel file system by moving files from completed application runs to even more resilient and cost-effective tiers in the storage hierarchy (for example, tape-based long-term archives, HDD-based active archives or public cloud long-term archives) and back. All automated based on defined policies
- Migration of large amounts of data from retiring parallel storage systems to the new parallel storage system

HPE offers innovative software solutions from both Intra-file system as well as for Inter-file system data movement.

Intra-file system movement: Cray ClusterStor data services

Cray ClusterStor data services is comprehensive set of new software tools for the Cray ClusterStor E1000 storage system increasing the cost-effectiveness of parallel storage by blending flash pools and disk pools in a single namespace while making Lustre easier to use.

Cray ClusterStor data services provides:

- A faster search engine based on new, more efficient indexing
- A new policy engine that uses common policy syntax for a fast learning curve
- A new tiering engine that directs the parallel data movers deployed on HPE ProLiant DL servers to take action

The following figure provides an overview over the five components of Cray ClusterStor data services, the users as well as the relationship with the Cray ClusterStor E1000 storage system.

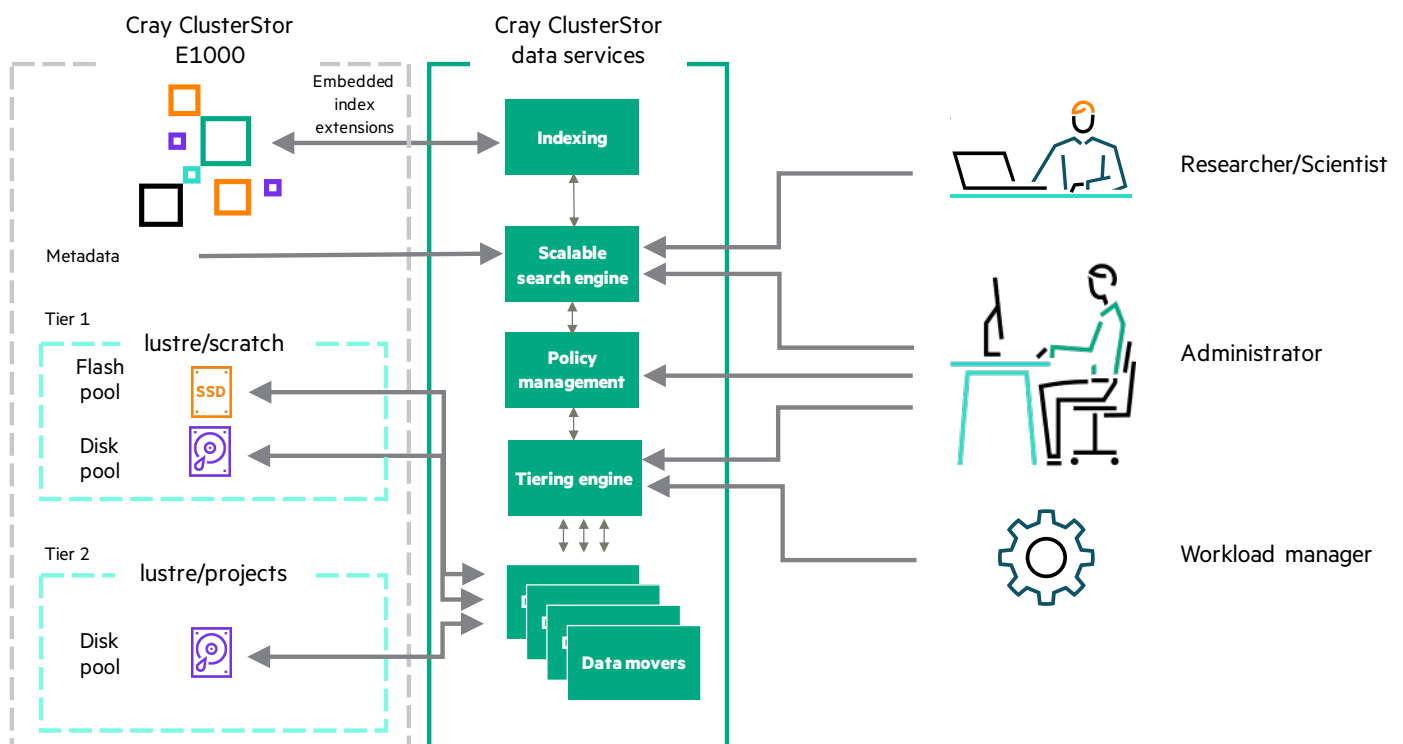


FIGURE 19. Cray ClusterStor data services overview

The two key innovations of Cray ClusterStor data services are:

- New HPE **indexing** technology that enhances performance of **search-at-scale**
 - Appended to the Lustre file system tree
 - Tracks file names, paths, and metadata
 - More effective than lfs find
 - Better performance vs. scan/copy or mirroring via changelog
 - Better storage utilization vs. an external storage pool for hosting a mirror
- New level of **automation** of previously manual **file system management** tasks through
 - New policy engine
 - New tiering engine
 - New parallel data movers

The following table shows some of the Lustre file system actions that ClusterStor data services automates. This means administrators do not have to execute commands manually, or manage scripts just to manage the file system.

TABLE 4. Tiering engine directives

Directive name	Description
Replication	Replicate files between OST pools (that is flash to disk)
Punch	Quickly free space from OST pools when capacity reaches a certain threshold
Mirror	Free up flash pool space more rapidly vs. an additional external storage system that only hosts a mirror of the metadata
Migrate	Migrate files from one OST pool to another
Purge	Delete files from namespace after reaching certain criteria
Archive	Send files to an external archive via Lustre HSM
Release	Release data from files that have been archived to free space in active file system

Inter-file system movement and data protection: HPE DMF v7

HPE Data Management Framework (HPE DMF) enables administrators to reduce the utilization of the parallel file system without limiting user access to their files, and the solution manages metadata and migrates data among storage assets based on workflow and administrator-defined criteria. This means that only the most critical or timely data resides on higher performance, premium storage systems that are co-located with HPC/AI compute clusters, while less frequently accessed data is automatically migrated to cost-effective, capacity-optimized storage media.

Highlights of HPE DMF v7 are:

- **Designed for HPC:** HPE DMF works in conjunction with HPC job schedulers and other workflow management systems to enable just-in-time data access for compute jobs and applications. With these capabilities, DMF radically improves storage utilization, saves on capital investment, and ensures that critical data assets are protected and secured.
- **Built for exascale but deploys at any scale:** Based on an incredibly rich feature set that has evolved over 20+ years of production customer deployment, the DMF v7 platform is based on a contemporary web scale architecture that is built to support exascale environments and the associated data volumes and object counts. However, any customer with over 1 petabyte of data to manage and protect can benefit from HPE DMF, and the solution can grow over time along with an organization's data requirements.
- **Integrated data protection:** HPE DMF is designed to integrate within administrator and user workflows as a tool that can simplify data-related operations while providing assurance that data is fully protected and won't be lost due to silent data corruption, system failure, or operator errors that can occur when working with extremely large data sets.
- **Namespace reflection:** The HPE DMF architecture supports deep integration with target filesystem environments—including Lustre, IBM Spectrum Scale (GPFS), and HPE XFS. This integration allows HPE DMF to be instantly notified of events such as file and directory creation, deletion, modification, and storage space utilization so that DMF's metadata repository can reflect these changes—and so the DMF policy engine can automatically take any needed data movement operations to protect, secure, and move data according to administrator-defined rules.
- **Modern open-source architecture:** Kafka for Changelog processing, Cassandra for Scalable Metadata, Mesos for Task Scheduling, and Spark for Query Engine are the foundation for a very flexible and scalable data management solution.



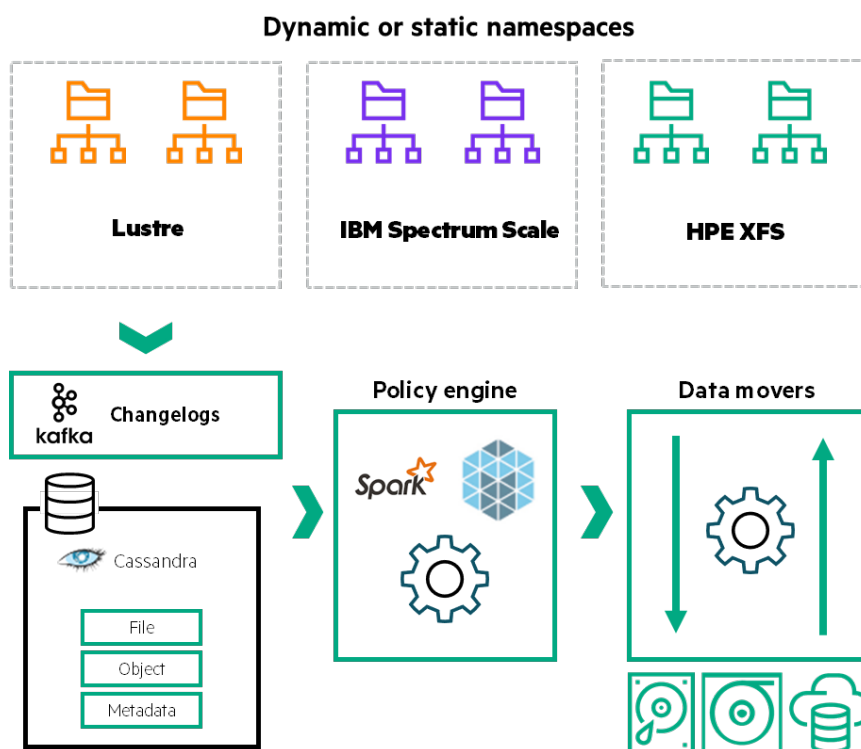


FIGURE 20. HPE DMF v7 architecture

Additional information on HPE DMF v7 is provided in the technical white paper [here](#).

Cray ClusterStor and the Lustre community

With the acquisition of Cray in the year 2019 HPE became one of the two largest Lustre development and support organizations.

Now, as a part of the Hewlett Packard Enterprise family, our Lustre development team continues to focus on hardening Lustre for reliable production at scale and adding enterprise features such as data resiliency with GridRAID, enabling fast rebuild times as well as balanced I/O and T10-PI phase 2 implementation to detect silent data corruption and bit rot. In addition, we continue to deliver user experience enhancing tools such as pool quotas, scalable search, and performance optimizations for flash technology in Lustre.

Another important contribution of HPE's Lustre development team to the Lustre community is the extensive testing of Lustre community features at scale in both HPE's labs and together with HPC leadership sites. The Lustre team reports identified defects to the community or develops patches. Developed patches are pushed upstream into the master code branch so that the whole community can benefit.

IMPORTANT

Occasional Lustre releases are designated as Long Term Stability (LTS) releases by the OpenSFS Lustre Working Group. LTS releases have a series of maintenance releases that include only small fixes and not major new functionality. The current LTS release is Lustre 2.12. The ClusterStor E1000 storage system uses Lustre 2.12 and will align with future LTS releases for stability in mission- or business-critical HPC environments.

Cray was a founding member of Open Scalable File Systems, Inc. (OpenSFS—opensfs.org/) which is a nonprofit organization dedicated to the success of the Lustre file system. OpenSFS was founded in 2010 to advance Lustre development, ensuring it remains vendor-neutral, open, and free.

The Lustre road map and development direction are not owned by a single company but are agreed upon in the OpenSFS Lustre Working Group (LWG—wiki.opensfs.org/Lustre_Working_Group). In the LWG the participants of OpenSFS come together to coordinate their software development efforts and the road map for the Lustre high-performance, open-source, parallel filesystem.



CONCLUSION

The Cray ClusterStor E1000 storage system is a unique new HPC storage system for the new HPC era purpose-engineered to:

- Provide enough storage performance to the CPU-based and GPU-accelerated compute nodes that are running simulations together with machine learning in mission- or business-critical workflows
- Provide that performance in the most efficient way in order to contain the forecasted over-proportional spending growth for HPC storage
- Provide all of that in an engineered system package that ships fully integrated after extensive soak-testing from the HPE factory with a comprehensive set of software tools to optimize all of the storage system operation, file system administration, and data curation

Additionally, for HPC organizations that already are using HPE's leading HPC compute platforms the Cray ClusterStor E1000 enables those users:

- To have, with HPE Pointnext Services, one support provider for the whole HPC infrastructure preventing frustrating **finger pointing** between HPC compute vendor and HPC storage vendor in case of end-to-end performance or system issues
- To have a path to a future consumption model **as a service** for the full HPC infrastructure with HPE GreenLake

This technical white paper only provides a first overview of the key hardware and software technologies of the Cray ClusterStor E1000 storage system. If you want to learn more about what the new storage system can do for you in your environment, please reach out to your HPE representative or HPE channel partner to arrange a dedicated deep dive briefing.

LEARN MORE AT

hpe.com/us/en/solutions/hpc-high-performance-computing/storage.html

Make the right purchase decision.
Contact our presales specialists.



Chat



Email



Call



Get updates