Hewlett Packard Enterprise

Business white paper

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

THE NEW HPC ERA NEEDS NEW HPC STORAGE

The confluence of simulation and AI changes everything

TABLE OF CONTENTS

EXECUTIVE SUMMARY

As the market share leader in high-performance computing (HPC) servers,[1] Hewlett Packard Enterprise saw the convergence of classic modeling and simulation with artificial intelligence (AI) methods such as machine learning (ML) and deep learning (DL) coming and took decisive action. Through the acquisition of Cray, HPE has assembled additional intellectual property on the HPC software layer, the interconnect layer, and the storage layer for the new HPC era. This is characterized by modeling and simulation, along with AI running on the same supercomputer or HPC cluster in mission- or business-critical workflows. We call this the exascale era.

The impact of the workload convergence is most disruptive on the HPC storage layer as the input/output (I/O) profiles of both workloads could not be more different. The storage architectures that have served us well in the past when simulation and AI were deployed on separate infrastructure stacks are breaking architecturally and economically in the new era.

The costs of staying with the legacy storage architectures in the new era range from idling, unproductive compute nodes due to I/O bottlenecks to exploding HPC storage costs, and high coordination efforts between different vendors.

Do you want to learn how HPE can help you in the new era to feed your CPU and GPU-accelerated compute nodes without I/O bottlenecks? You can also learn how to contain storage spending growth and how to have unified accountability of HPE Pointnext Services for your complete HPC infrastructure.

WORKLOAD CONVERGENCE CHANGES EVERYTHING

HPC is transforming. Explosive data growth, that by far exceeds price/performance improvements in underlying storage media technologies, and the convergence of modeling and simulation, along with AI workloads have created new requirements for compute, software, networking, and storage. Called the exascale era, it's far more than a performance milestone. It's a technology inflection point characterized by a new set of capabilities for a new set of workloads running in mission- or business-critical workflows on one machine. And it affects every industry and field of inquiry.

These new realities put new demands on high-performance storage. And it is happening right now. For example, a recent study of the independent analyst firm Intersect360 found out that the majority (61%) of the HPC users today already are running ML programs. And an additional 10% of the respondents stated that they plan to do so until the end of the year 2020.[2]

Storage technology that worked for petascale era workloads cannot power the exascale era's converged workflows because the I/O patterns and the characteristics of the currently deployed storage technologies could not be more different. As a result, sticking with your current storage infrastructure will leave you unable to keep up—both in terms of performance and budget.

[1] 2019 market results, new forecasts, and HPC trends, Hyperion Research, April 2020

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

[3] nextplatform.com/2019/10/22/exascale-is-not-your-grandfathers-hpc/

"THE CONFLUENCE OF AI WITH TRADITIONAL SIMULATIONS IS GOING TO TRANSFORM THE VERY NATURE OF HIGH PERFORMANCE COMPUTING."

– Rick Stevens, Argonne National Laboratory[3]

TABLE 1. Comparison of traditional petascale era infrastructures

	Traditional HPC cluster	Traditional AI POD
Primary workload	Modeling and simulation	ML/DL
Compute node type	CPU nodes such as HPE Apollo 2000 or Dell PowerEdge servers	GPU-accelerated nodes such as HPE Apollo 6500 or NVIDIA® DGX
Number of compute nodes	Hundreds to thousands	A few
Typical interconnect	InfiniBand	Gigabit Ethernet
Typical I/O intensity	Write-intensive	Read-intensive
Typical storage	HDD-based parallel file system storage	All-flash enterprise file storage
Capacity measured in	Petabytes	Terabytes
Scalability in single namespace	Up to exabytes	Up to low double digits petabytes
Well suited for	• Mainly writing • Large files • In sequential order • At speeds of double or triple digits gigabytes per second	• Reading and writing • Files of all sizes • In both sequential and random order • At speeds of up to low double digits gigabytes per second
Price per terabyte	$	$$$$$$$$$$

It seems that in the new era, HPC users are in in a difficult situation:

- If they stick with the traditional HPC storage systems most likely they will experience I/O bottlenecks for their AI/ML workloads as traditional HPC storage is not suited well to serve the large number of files of all sizes that ML needs to read in the training phase. That can lead to job pipeline congestion, missed deadlines, unsatisfied data scientists, and constant escalations.
- If they try to scale their traditional AI storage to the multi-petabyte requirements of the converged workloads, they most likely will experience scalability issues and exploding storage costs.

The new HPC era really needs new HPC storage.

INTRODUCING THE NEW HPC STORAGE FOR THE NEW ERA

We have developed a new kind of HPC storage system that combines the best of both worlds for the new HPC era of converged workloads running on one supercomputer or HPC cluster. The design goal was to create a new kind of HPC storage system that:

- Is as cost-effective and scalable as traditional HPC storage
- Is as well suited to serve huge numbers of files of all sizes such as all-flash enterprise file storage does today

Introducing, the Cray ClusterStor E1000 storage system, which is available for HPE Pointnext Services support in more than 140 countries from August 2020.

TABLE 2. Comparison of traditional petascale era infrastructures with converged exascale era infrastructure

	Traditional HPC cluster	Exascale era converged	Traditional AI POD
Primary workload	Modeling and simulation	Both	ML/DL
Compute node type	CPU nodes	Both	GPU-accelerated nodes
Number of compute nodes	Hundreds to thousands	Hundreds to thousands	A few
Typical interconnect	InfiniBand	Both (plus HPE Slingshot)	Gigabit Ethernet
Typical I/O intensity	Write-intensive	Both	Read-intensive
Typical storage	HDD-based parallel storage	Cray ClusterStor E1000	All-flash enterprise file storage
Capacity measured in	Petabytes	Petabytes	Terabytes
Scalability in single namespace	Up to exabytes	Up to exabytes	Up to low double digits petabytes
Well suited for	• Mainly writing • Large files • In sequential order • At speeds of up to double or triple digits gigabytes per second	• Reading and writing • Files of all sizes • In both sequential and random order • At speeds of up to double or triple digits gigabytes per second	• Reading and writing • Files of all sizes • In both sequential and random order • At speeds of up to low double digits gigabytes per second
Price per terabyte	$	$$$	$$$$$$$$$

We achieve this by combining new hardware and software technologies in an engineered HPC storage system that ships fully integrated after extensive soak-test from the HPE factory. Figure 1 illustrates how the system looks and gives a high-level overview of the key new technologies in the system.

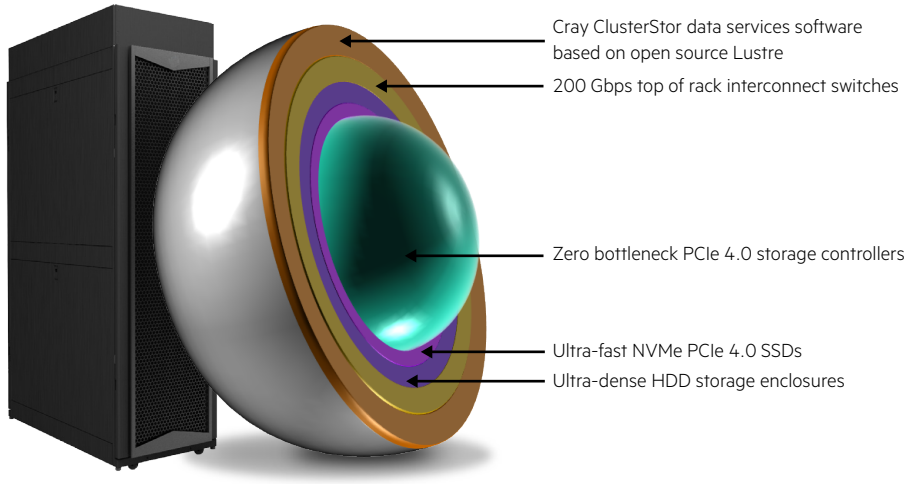

FIGURE 1. New technologies of the Cray ClusterStor E1000 storage system

Let's have a quick look at the new technologies:

- **End-to-end PCIe 4.0 storage controllers:** The zero bottleneck, end-to-end PCIe 4.0 design based on AMD EPYC™ CPUs enables us to bring significantly more performance per storage drive (SSD or HDD) through the parallel file system from and to the compute nodes than alternative storage systems. Throughput performance—measured in gigabits per second—is critical in the new era as a single modern network interface card in a CPU node or GPU-accelerated node has 64 gigabyte per second bidirectional data bandwidth. And modern nodes feature between 2 and 8 of those network interface cards to bring data from/to the compute nodes.
- **Ultra-fast NVMe PCIe 4.0 SSDs:** The new SSDs are so fast that you need modern 200 gigabit per second networks to bring the full performance of the SSDs to the CPU nodes and GPU-accelerated compute nodes. Those ultra-fast NVMe Gen4 SSDs can only unlock their inherent performance if they are embedded in an HPC storage system that was architected for this new technology.
- **Ultra-dense HDD storage enclosures:** The performance is mainly driven from SSDs in the new era but you still need HDD-enclosures to provide cost-effective storage capacity in the same file system/namespace. We deploy one of the densest HDD enclosures in the industry that can pack 1.25 petabyte usable storage capacity in just four rack units (with 16 TB HDDs).
- **200 gigabit per second interconnect switches:** With a choice of top of rack switches, we connect to any supercomputer or HPC cluster that uses either InfiniBand HDR, 200GbE, or HPE Slingshot interconnects. While 100 gigabit per second InfiniBand and Ethernet is supported too, it is highly recommended to use modern 200 gigabit per second high-speed interconnects to get the full storage performance out of Cray ClusterStor E1000.
- **Cray ClusterStor data services software:** ClusterStor data services accompanies the Cray ClusterStor E1000 storage system with two objectives. First objective is to make Lustre easier to use through automated and faster execution of currently manual tasks. The second objective is to provide the orchestration of the data flow with the workflow by enabling efficient data movement from HDD pools to SSD pool and vice versa in the same name space. The idea is to have the right data on the right storage medium at the right time to accelerate the workflow.

This unique combination of hardware and software innovation as a fully integrated package has enabled the Cray ClusterStor E1000 storage system to experience unprecedented customer adoption.

For example, three exascale sites in the U.S.—Argonne National Laboratory,[4] Oak Ridge National Laboratory,[5] and Lawrence Livermore National Laboratory[6]—have selected the Cray ClusterStor E1000 storage system as their external parallel file systems. As all three HPC leadership sites area are known in the industry for their very demanding and thorough vendor selection processes, this is a strong endorsement.

You don't have to worry if you are not planning to deploy exascale-sized systems in the foreseeable future. We have architected the Cray ClusterStor E1000 storage system in a way that enables you to start anywhere, to then go wherever you need to grow, without architectural limitations, helping you reach your goals. You can start with the Cray ClusterStor E1000 storage system with just 10 rack units.

[4] alcf.anl.gov/news/does-argonne-leadership-computing-facility-and-hpe-expand-high-performance-computing-hpc

[5] nextplatform.com/2019/06/20/exascale-system-at-oak-ridge-will-blaze-new-storage-path/

[6] hpe.com/us/en/newsroom/press-release/2020/03/hpe-and-amd-power-complex-scientific-discovery-in-worlds-fastest-supercomputer-for-us-department-of-energys-doe-national-nuclear-security-administration-nnsa.html

THE HPE ADVANTAGE

Compare the Cray ClusterStor E1000 storage system with other vendor's approaches to HPC storage and it's clear that we are doing something different. That is because HPE approaches next-generation HPC storage from a very different place than the rest of the marketplace. We bring:

- **Unique motivation:** At HPE, we believe HPC storage is a means to an end—unlocking new insights through simulations and AI. In order to achieve fast time-to-insights it is important to build balanced supercomputers or HPC clusters. An optimal balance of compute, interconnect, and storage typically results in fast time-to-insights. We are focused on engineering performant and efficient HPC storage systems that achieve the demanding storage requirements at an effective purchase price and TCO. Unlike others, we believe that over-proportional HPC storage spending growth should be contained in order to protect the customer budget allocated for compute including interconnect.
- **Distinct philosophy:** We are not only striving to engineer the fastest HPC storage systems but we also mandate to achieve that performance in the most cost-effective way. HPE achieves this by engineering for leading performance efficiency, which means that we design for bringing as much performance as possible from every single storage drive through the file system to the compute nodes. We call this metric performance efficiency. Performance efficiency by design is important to for brining cost-effective HPC storage systems to the market. But as the most-efficient system design can be neutralized by charging a software license tax per terabyte storage capacity or per storage drive, at HPE, we are strong believers in open source software.

  HPE embeds the open source parallel file system Lustre into our HPC storage systems. Lustre is the most widely used parallel file system in the HPC market.[7] It enables us to not charge a file system license for the capacity and the storage drives deployed in our HPC storage systems.

- **Matchless flexibility:** We engineer our HPC storage systems in a way that they can attach to any supercomputer or HPC cluster of any vendor as long as the compute system supports either InfiniBand EDR/HDR, or 100/200GbE, or HPE Slingshot interconnects.
- **Unified customer support experience:** HPC users should have the option to have their full HPC stack—software, compute, interconnect, and storage—under unified customer support from one single support provider. HPE provides that option with HPE Pointnext Services support in more than 140 countries worldwide. In addition, we do not only provide hardware support for our HPC storage systems but also software support including the parallel file system. Our in-house Lustre development and support team can fix and patch the parallel file system for fast problem resolution of individual customer support cases. After that, we push the patch/fix up into the master code stream of the vibrant Lustre community so that the whole open source community can benefit from it.
- **As-a-service consumption model:** At HPE, we are committed to offer our entire portfolio through a range of subscription, pay-per-use, and consumption-driven offerings by the year 2022.[8] But doing that we want to give customers the option to combine the agility and economics of public cloud with the security and performance of on-premises IT infrastructure. If they choose to do so, as classic purchasing and financing sourcing options will continue to be available.

[7] Shifts Are Occurring in the File System Landscape, Hyperion Research (Special Study), June 2020

[8] hpe.com/us/en/newsroom/press-release/2019/06/hpe-announces-plans-to-offer-entire-portfolio-as-a-service-by-2022.html

Performance
- Up to 80 GB/sec from just two rack units
- More than 1,000 GB/sec from a standard 19" data center rack
- No more I/O bottlenecks

Efficiency
- Up to 3.3 GB/sec from just one storage drive
- No software license tax per storage drive
- HPC storage spending growth contained

Support
- HPE Pointnext Services support for the full HPC storage infrastructure
- No more hassles between HPC compute and storage vendors

Flexibility
- Connects to any cluster or supercomputer of any vendor
- Choice of InfiniBand, Gigabit Ethernet, or HPE Slingshot interconnects

As a service
- Sourcing option to combine the agility and economics of public cloud with the security and performance of on-premises IT

FIGURE 2. The HPE advantage

START YOUR HPC STORAGE TRANSFORMATION

The independent analyst firm Hyperion Research forecasts that—current course and speed—HPC storage spending will grow with a 40% faster compound annual growth rate than HPC compute spending throughout the year 2024.[9] That simply means that HPC storage is forecasted to consume a much larger part of your overall budget for your next supercomputer or HPC cluster going forward.

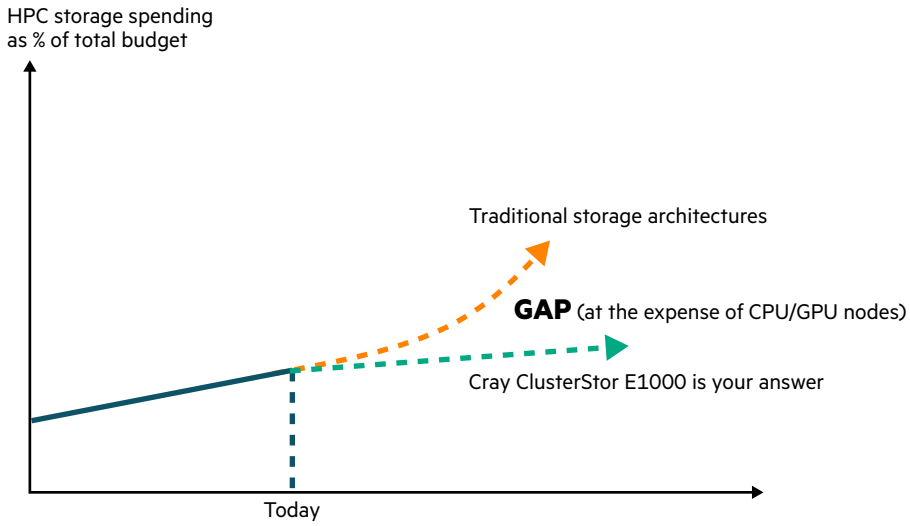

FIGURE 3. The not-so-desirable future as predicted by Hyperion Research

If you are interested to get onto the green curve (Figure 3), contact your HPE sales representative to discuss the new HPC storage from HPE.

Your architects can review a technical white paper on how the Cray ClusterStor E1000 storage system works. Read the technical white paper here.

If you worry about the risk and complexity of a potential data migration from your currently installed HPC storage system to Cray ClusterStor E1000, worry no more. HPE owns the proven HPE Data Management Framework (DMF) software that not only enables risk-free and fast data migration from legacy HPC storage systems but also can help you to manage and protect the permanent data that resides in your parallel file systems.

[9] 2019 market results, new forecasts, and HPC trends, Hyperion Research, April 2020

Resources

Cray ClusterStor E1000 Storage System Technical White Paper
hpe.com/h20195/v2/Getdocument.aspx?docname=a50001954enw

HPE Data Management Framework (DMF) Technical White Paper
hpe.com/h20195/v2/Getdocument.aspx?docname=a50001461enw

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

Hewlett Packard Enterprise

a50002042ENW, June 2020