Technical White Paper

Everpure

FlashBlade//EXA MLPerf Storage v2.0

Benchmark test results

Everpure

Contents

Introduction

AI has shifted from experimental pilots to mission-critical production systems, and with that shift comes a new set of performance and economic pressures. Organizations are investing heavily in GPUs, yet many discover that their accelerators sit idle for a significant portion of training and inference cycles—not because of insufficient compute, but because their storage systems cannot deliver data fast enough.

Everpure™ FlashBlade//EXA™ is a scale-out, disaggregated file and object platform specifically engineered for very large-scale AI, high-performance computing (HPC), and unstructured data pipelines, combining extremely high throughput with exabyte-class scalability.

Executive summary: The new AI bottleneck

AI infrastructure has entered a new performance era where storage—not GPUs—determines the speed, scale, and economics of modern ML workloads. As models grow larger and data sets become more complex, the ability to deliver data to accelerators at sustained high throughput is now the primary factor influencing training time, operational efficiency, and total cost of ownership. The MLCommons MLPerf Storage benchmark suite provides an industry-standard method for evaluating this capability across diverse workloads, from 3D medical imaging (3D U-Net) and scientific simulation analysis (CosmoFlow) to high-volume image classification (ResNet-50) and frontier-scale LLMs (Llama 3). These benchmarks collectively model and replay demanding I/O patterns that make or break GPU utilization, offering a clear, objective view of how well a storage platform supports real-world AI pipelines.

Everpure FlashBlade//EXA consistently exceeds published MLPerf Storage 2.0 results, demonstrating superior throughput, parallelism, and latency under the exact conditions that typically stall GPU pipelines. By keeping GPU utilization above 90%, FlashBlade//EXA empowers organizations to achieve the same or greater model performance with fewer GPUs, fewer CPU servers, and a smaller supporting infrastructure footprint (networking, power, and cooling). This directly reduces capital expenditures on accelerators and networking while lowering ongoing power and cooling requirements. The result is a materially stronger return on investment: faster time to value for AI initiatives, reduced infrastructure sprawl, and a more predictable, scalable foundation for enterprise and research-class AI workloads.

Why the right storage matters for AI

In AI factories, storage is not an accessory; it is a core performance component that directly determines how quickly GPUs can transform data into competitive advantage. When storage is undersized, slow, or unreliable, even the largest GPU clusters sit idle, starved for data while timelines, budgets, and business expectations slip. The result is a widening gap between what the business paid for and what the AI team can actually deliver. In contrast, a storage architecture engineered for high throughput, low latency, and resilient checkpointing keeps GPUs continuously fed, minimizes the impact of inevitable hardware events, and allows AI teams to meet—rather than constantly revise—delivery commitments.

The following examples illustrate how storage architecture directly influences GPU efficiency and AI delivery outcomes.

Example 1: Underpowered storage

Example 1 illustrates the hidden cost of underpowered storage in a 2,000-GPU AI factory. After resolving upfront constraints around GPUs, space, power, and cooling, an organization invests over $25 million in capital with a recurring annual spend of roughly $5 million only to discover that GPUs are running at less than 50% utilization and the original 30-day results target must slip to 60 days. A subsequent hardware failure, combined with inadequate checkpointing on slow storage, corrupts data and forces the team to restart training. Once checkpointing is added, progress slows further.

The net effect: Usable results are now projected beyond 210 days from the initial start, turning a $30 million AI investment into a delayed, uncertain bet.

Example 2: Storage first

In Example 2, the same organization takes a storage-first approach, designing a 1,000-GPU AI factory around a high-performance Everpure FlashBlade//EXA platform with an approximate $2.5 million investment. With only half the GPUs, space, power, and cooling of Example 1, the environment is fully deployed in half the time at a significantly lower capital outlay of roughly $15 million and a lower recurring cost of about $2.5 million per year. Because the storage system delivers sustained bandwidth and resiliency, all GPUs operate above 90% utilization, keeping the original 30-day results target intact. When a hardware failure occurs in week two, checkpointed data on FlashBlade//EXA enables a rapid, one-day recovery with only a minor adjustment to the timeline—from 30 to 31 days.

The net effect: Just 61 days after the initial investment, the AI factory is already delivering actionable business insights, demonstrating that the right storage turns AI from an infrastructure experiment into a predictable, high-return business asset.

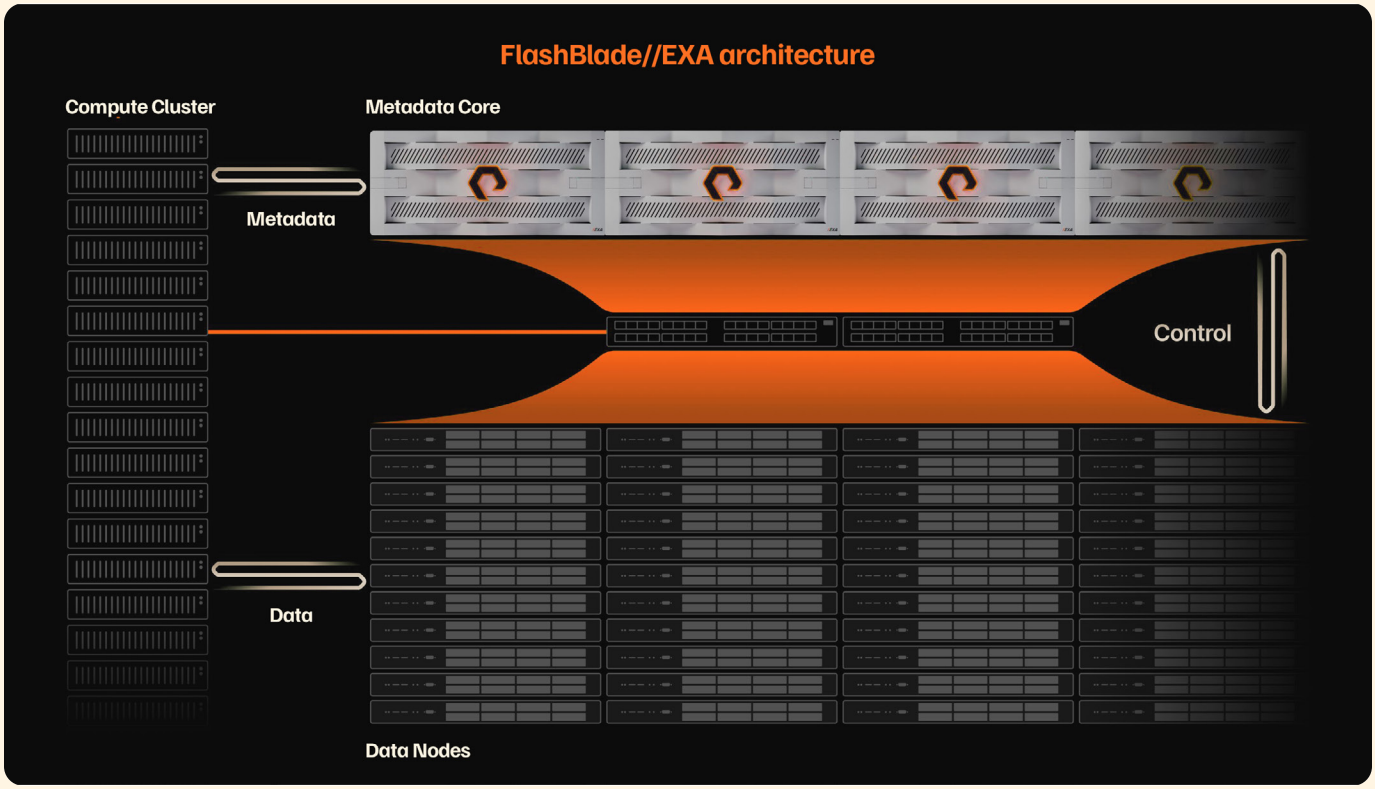
Understanding MLPerf Storage benchmarks

MLPerf Storage benchmarks provide a standardized, workload-realistic method for evaluating how well a storage system can sustain the data delivery rates required by modern AI pipelines. Unlike synthetic tests that measure isolated metrics such as raw IOPS or sequential throughput, MLPerf Storage focuses on end-to-end performance under real ML workloads. Each benchmark reproduces the I/O patterns, data set structures, and parallel access behaviors seen in production AI environments, making it possible to assess whether storage can keep GPUs fully utilized during training and inference. This approach allows organizations to compare systems using a common, vendor-neutral framework that reflects actual operational demands rather than idealized lab conditions.

The suite spans a diverse set of AI domains—medical imaging (3D U-Net), scientific simulation analysis (CosmoFlow), computer vision (ResNet-50), and large language model (LLM) training (Llama 3). Together, these benchmarks expose the full range of storage stressors: large 3D tensor reads, high-volume random access, multi-GPU parallelism, and multi-terabyte checkpointing. Because each workload exercises different aspects of the storage stack, performance across the suite provides a comprehensive view of how a system behaves under real-world pressure. For organizations scaling AI initiatives, MLPerf Storage results offer a reliable indicator of whether their infrastructure can deliver consistent, predictable performance as models grow larger and data pipelines become more complex.

Everpure scalable architecture advantages

Everpure FlashBlade//EXA is a scale-out, disaggregated file and object platform specifically engineered for very large-scale AI, HPC, and unstructured data pipelines, combining extremely high throughput with exabyte-class scalability. Architecturally, FlashBlade//EXA separates metadata and data services so they can scale independently, eliminating the overprovisioning and contention that typically occur in monolithic, tightly coupled designs. This separation, combined with the best-in-class capacity of DirectFlash® Modules, allows the system to deliver more than 10TB/s of read bandwidth in a single namespace, while also optimizing power and cooling at high densities.



MLPerf Storage details

This section outlines the MLPerf Storage workloads used to assess FlashBlade//EXA performance.

3D U-Net

Description: 3D U-Net evaluates performance on volumetric medical image segmentation, measuring how efficiently systems process large 3D scans for precise voxel-level labeling.

Use case: It is used to validate infrastructure for clinical imaging workloads such as tumor segmentation, organ boundary detection, and automated radiology pipelines.

Relevant workflow: The benchmark mirrors high-throughput, 3D medical imaging inference and training, where large volumetric tensors must be loaded, preprocessed, and segmented in real time.

CosmoFlow

Description: CosmoFlow trains a 3D convolutional model on cosmology simulation volumes to predict physical parameters, stressing random-read I/O and large-scale scientific ML throughput.

Use case: It is ideal for assessing systems that support HPC-class scientific modeling, including astrophysics, climate simulations, and large-volume tensor analysis.

Relevant workflow: The benchmark reflects distributed scientific ML training, where massive 3D data sets are streamed in parallel to many workers.

ResNet-50

Description: ResNet-50 measures image classification performance using a deep residual network widely adopted as a baseline for computer vision workloads.

Use case: It is commonly used to evaluate general-purpose vision pipelines, edge-to-cloud inference, and high-volume image recognition systems.

Relevant workflow: The benchmark aligns with batch-oriented 2D image classification pipelines, emphasizing rapid preprocessing and GPU-efficient training loops.

Llama 3 8B

Description: Llama 3 8B benchmarks latency and throughput for compact, transformer-based LLMs optimized for cost-efficient inference.

Use case: It is ideal for testing lightweight enterprise chatbots, on-device assistants, and low-latency generative AI services.

Relevant workflow: The benchmark represents token-efficient LLM inference, where small models must respond quickly with minimal memory overhead.

Llama 3 70B

Description: Llama 3 70B evaluates performance on a large-capacity LLM requiring substantial memory bandwidth and parallel compute for high-quality reasoning.

Use case: It is suited for benchmarking enterprise-grade AI assistants, multilingual generation, and high-accuracy text-reasoning systems.

Relevant workflow: The benchmark reflects distributed LLM inference and fine-tuning, where multi-GPU coordination and high-bandwidth memory access dominate performance.

Llama 3 405B

Description: Llama 3 405B measures performance on an ultra-large-scale LLM that stresses extreme model parallelism, checkpoint bandwidth, and long-context reasoning.

Use case: It is used to validate infrastructure for frontier-scale AI training, research-class supercomputing, and massive-context generative workloads.

Relevant workflow: The benchmark mirrors frontier-scale LLM training, where trillion-parameter models require synchronized compute, high-speed interconnects, and sustained checkpoint I/O.

MLPerf Storage test results

This section summarizes the MLPerf Storage v2.0 test results for each benchmarked workload. Refer to the Appendix for details on the test setup.

3D U-Net

Description: 3D U-Net benchmarks volumetric medical image segmentation performance, measuring how efficiently systems process large 3D scans for pixel-accurate organ and tumor delineation.

Use case: It is ideal for evaluating infrastructure used in clinical imaging pipelines, such as kidney-tumor segmentation, radiology automation, and high-resolution 3D inference workloads. It emulates H100 GPUs running a medical imaging inference pipeline (for example, volumetric CT/MRI segmentation) and drives a mixed read-heavy, metadata-intensive storage workload characterized by large sequential data set reads, repeated tensor access, and frequent small file operations during preprocessing and result staging.

Test results: FlashBlade//EXA achieved 472 H100s in half a rack of storage with just 15 data nodes and a single chassis utilizing FlashBlade//EXA for metadata. This is about two times the performance of the closest competitor's published and audited results.

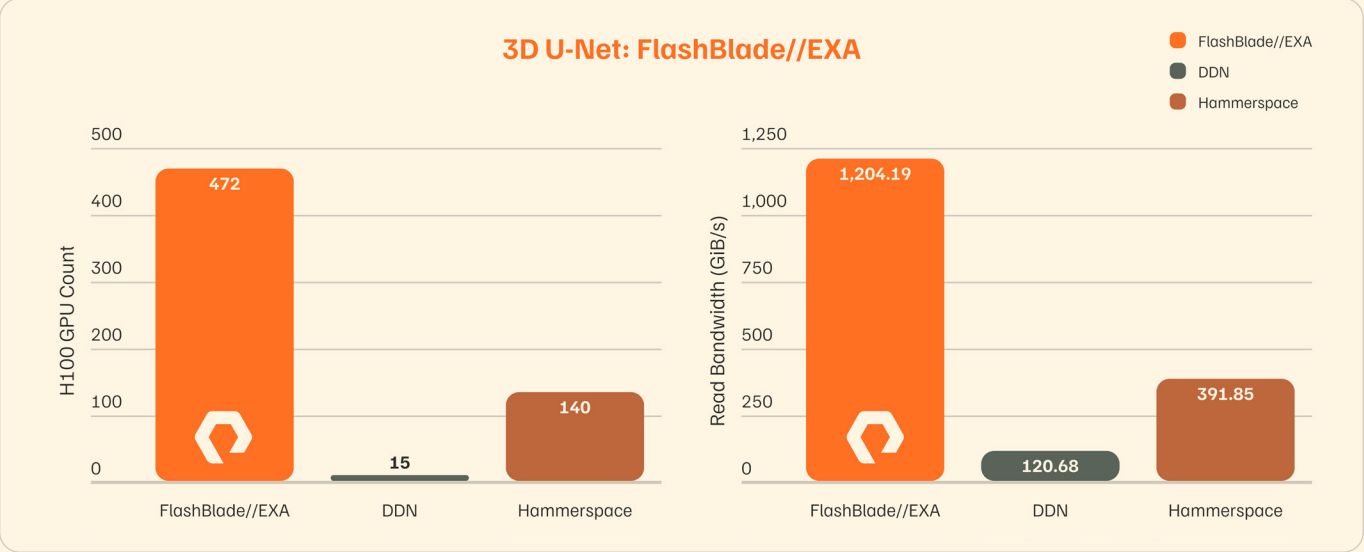


FIGURE 1 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for 3D U-Net | Result not verified by MLCommons Association | Highest recorded reference

CosmoFlow

Description: CosmoFlow trains a 3D convolutional network on cosmology simulation volumes to predict fundamental physical parameters, stressing random-read I/O and large-scale scientific ML throughput.

Use case: It is used to assess systems supporting HPC-scale scientific modeling, including astrophysics simulations, climate modeling, and any workload requiring massive 3D tensor ingestion.

Test results: Across MLPerf Storage v2.0 CosmoFlow results, FlashBlade//EXA demonstrated the highest observed performance and scale, reaching ~590 simulated H100s while maintaining consistent results across a four-times change in storage size (15 to 59 data nodes). At this scale, no directly comparable vendor configurations were identified, and MLCommons guidance discourages normalizing, extrapolating, or projecting results across materially different system sizes. Even so, the data shows FlashBlade//EXA removed storage and network limitations, with scaling ultimately constrained by client-side memory rather than the storage platform. Relative to published submissions, FlashBlade//EXA exceeded the next-best results by ~10–20% and delivered competitive efficiency per storage rack unit without relying on host-side SSD caching.

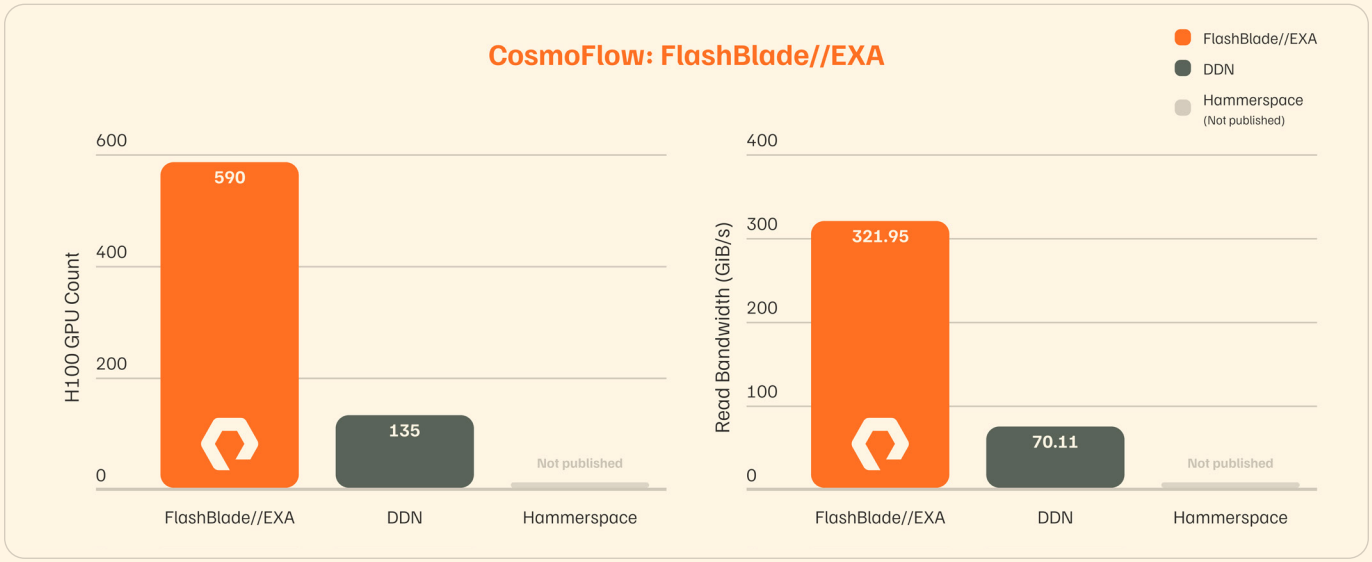


FIGURE 2 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for CosmoFlow | Result not verified by MLCommons Association | Highest recorded reference

ResNet50

Description: ResNet-50 measures image classification performance using a deep residual network that represents a widely adopted baseline for computer vision inference.

Use case: It is commonly used to evaluate general-purpose vision pipelines, edge-to-cloud inference performance, and throughput-sensitive image recognition systems.

Test results: Across MLPerf v2.0 ResNet-50 results, FlashBlade//EXA configurations delivered the highest observed scale and throughput among globally available storage platforms, reaching up to ~2,360 simulated H100s with ~398GiB/s of sustained read bandwidth using three FlashBlade//S500 chassis for metadata services and 15 FlashBlade//EXA data nodes. Performance remained consistent across multiple FlashBlade//EXA data node configurations, indicating the benchmark was host-bound rather than storage-bound at this scale. While some region-specific submissions demonstrated competitive results at smaller footprints, FlashBlade//EXA uniquely combines top-end performance, scalability, and global availability without reliance on specialized or localized solutions.

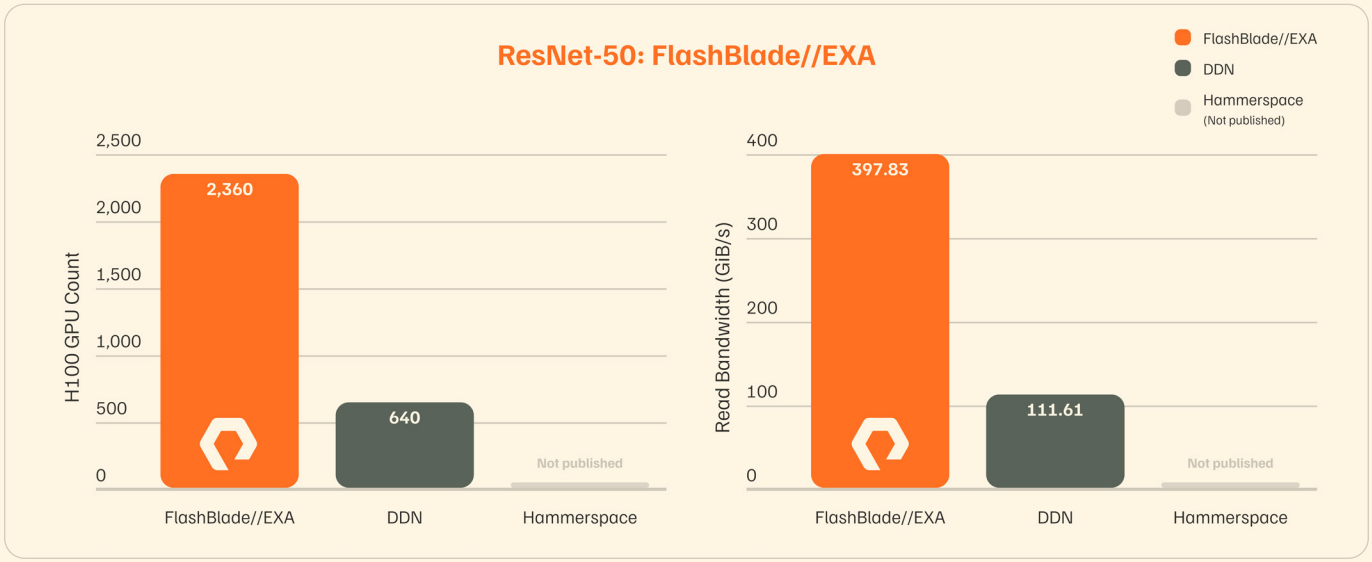


FIGURE 3 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for ResNet-50 | Result not verified by MLCommons Association | Highest recorded reference

Llama 3 8B

Description: Llama 3 8B is a compact, transformer-based LLM benchmark that measures latency and throughput for mid-sized generative AI workloads.

Use case: It is ideal for testing cost-efficient chatbots, on-device assistants, and lightweight enterprise inference where memory footprint and response speed matter.

Test results: Everpure FlashBlade//EXA dramatically outperformed all other platforms both in total and per data node and rack unit of storage consumed.

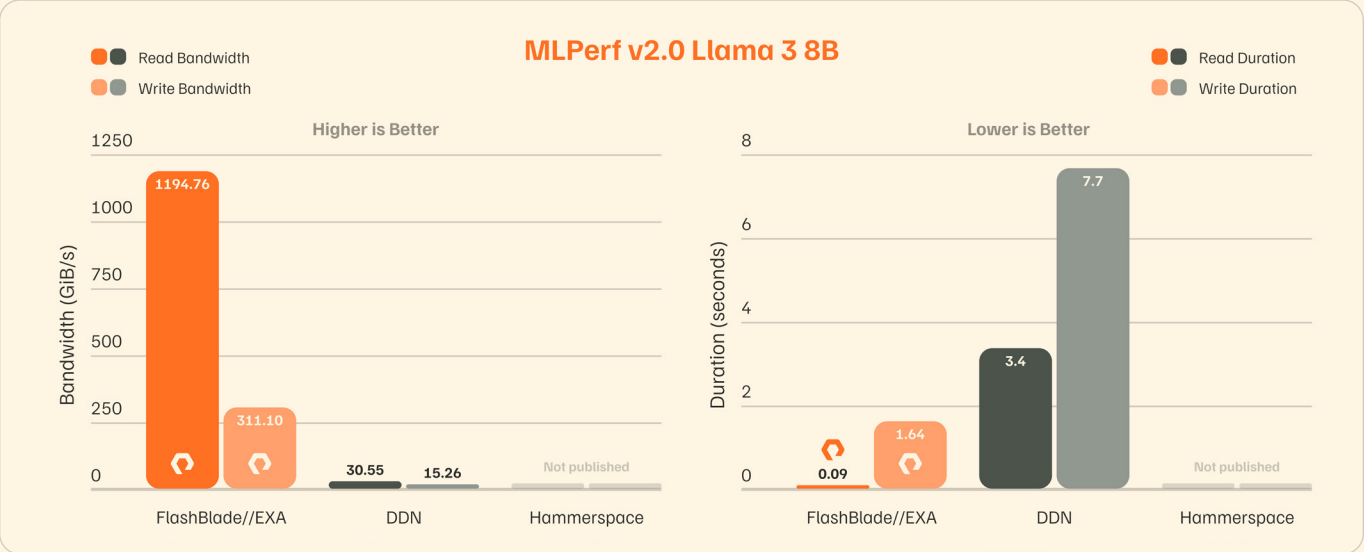


FIGURE 4 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for Llama 3 8B | Result not verified by MLCommons Association | Highest recorded reference

Llama 3 70B

Description: Llama 3 70B evaluates performance on a large, high-capacity LLM that demands substantial memory bandwidth and parallelism for high-quality reasoning and generation.

Use case: It is suited for benchmarking enterprise-grade AI services, multilingual assistants, and high-accuracy text generation systems deployed on multi-GPU or distributed platforms.

Test results: Everpure FlashBlade//EXA outperformed all competitive storage vendors' official MLPerf Storage v2.0 closed submissions.

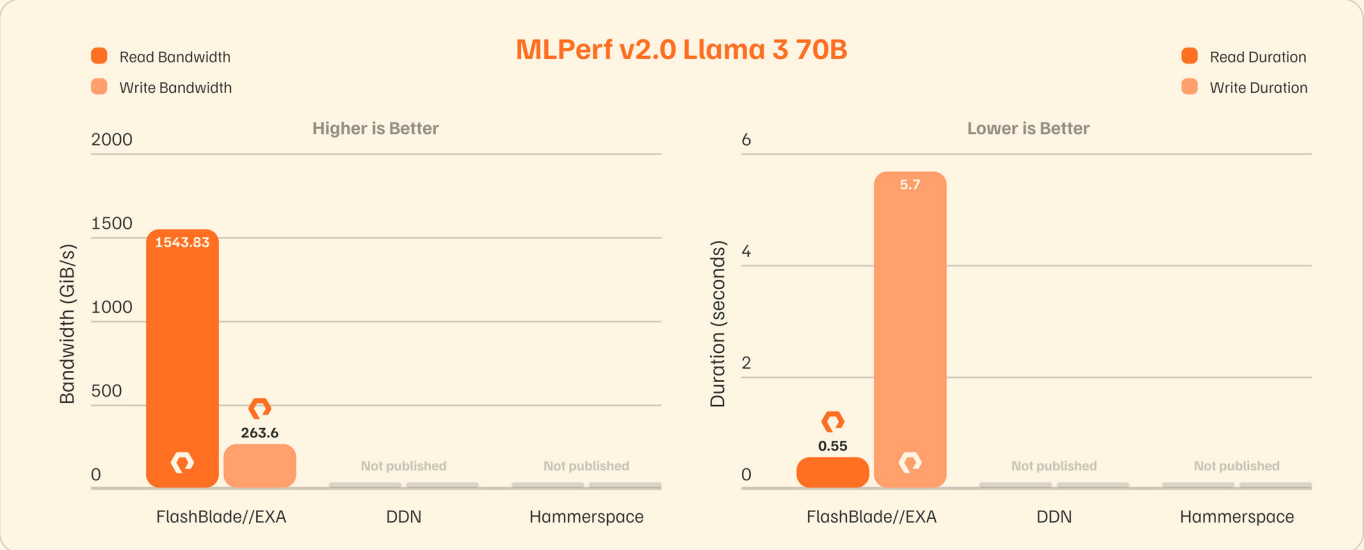


FIGURE 5 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for Llama 3 70B | Result not verified by MLCommons Association | Highest recorded reference

Llama 3 405B

Description: Llama 3 405B represents an ultra-large-scale LLM benchmark that stresses extreme model parallelism, checkpointing bandwidth, and long-context generative reasoning.

Use case: It is used to validate infrastructure for frontier-scale AI training, large-context retrieval-augmented generation, and research-class supercomputing environments.

Test results: Everpure FlashBlade//EXA dramatically outperformed all other platforms in total.

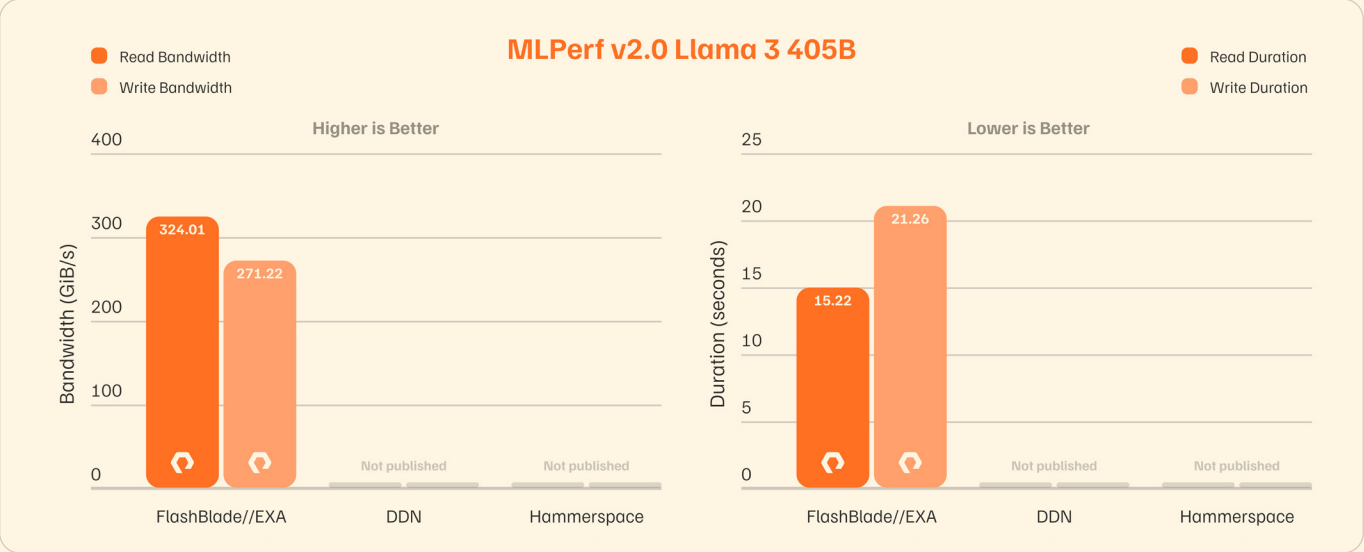


FIGURE 6 Comparison of FlashBlade//EXA to best published DDN and Hammerspace results for Llama 3 405B | Result not verified by MLCommons Association | Highest recorded reference

Training test summary

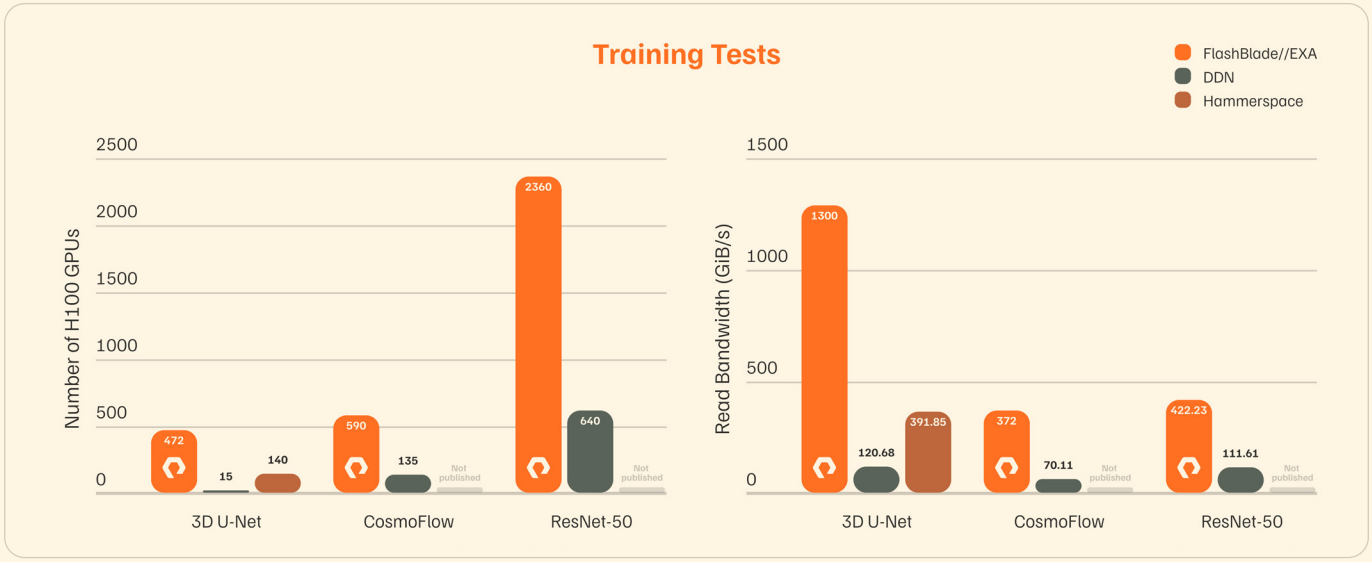


FIGURE 7 Summary of training tests comparing FlashBlade//EXA to best published Hammerspace and DDN results | Result not verified by MLCommons Association | Highest recorded reference

Checkpoint benchmark results summary

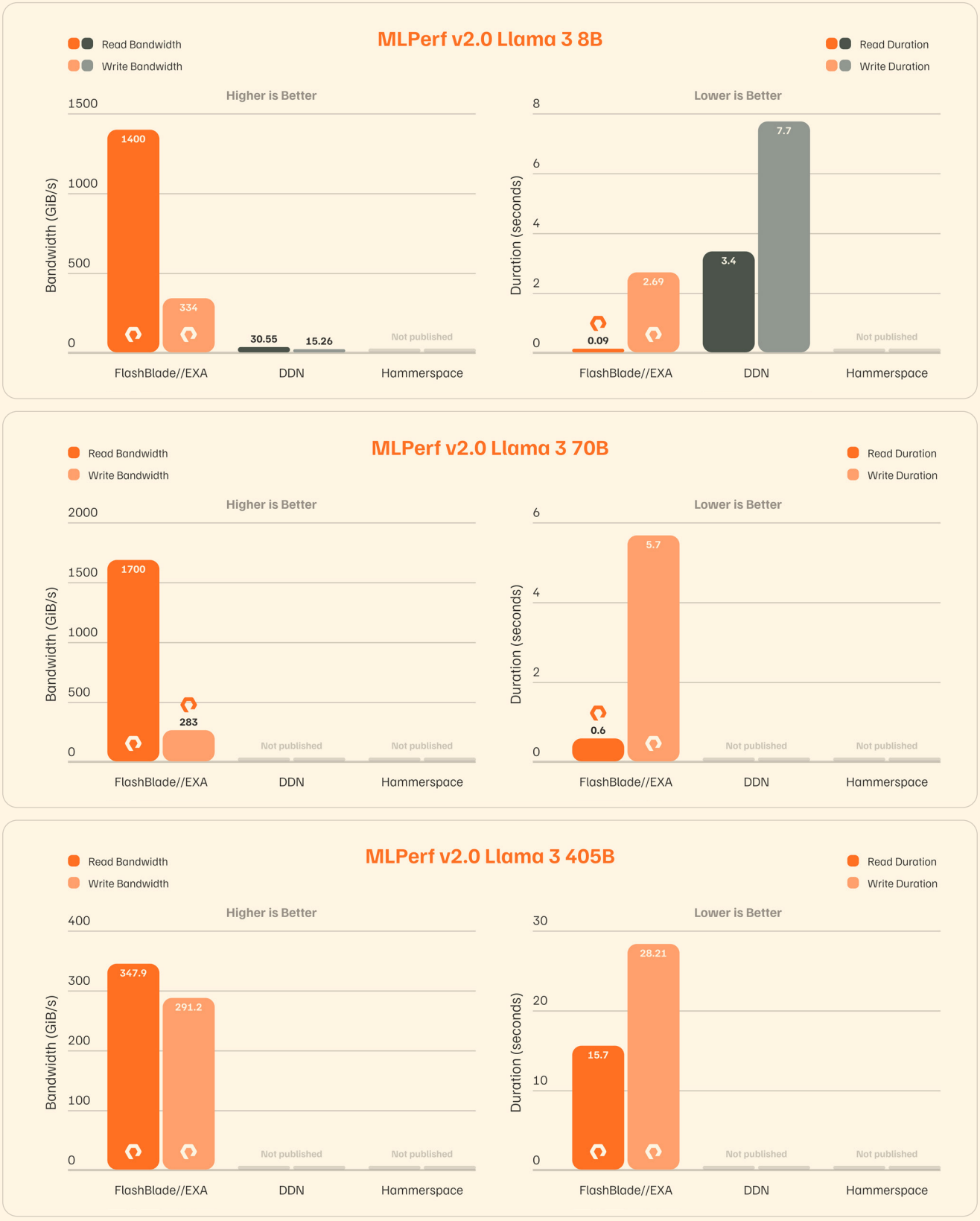


FIGURE 8 Summary of checkpoint benchmark tests comparing FlashBlade//EXA to best published Hammerspace and DDN results | Result not verified by MLCommons Association | Highest recorded reference

Conclusion

Everpure FlashBlade//EXA delivers groundbreaking advantages by leveraging the innovative data architecture of Everpure to address the escalating demands of modern AI and HPC environments. Built on a proven, high-efficiency metadata core, FlashBlade//EXA provides industry-leading performance levels specifically designed to meet the challenges of modern large-scale HPC and AI workloads.

FlashBlade//EXA helps ensure data access, with industry-leading read performance exceeding 10TB/s in a single namespace and write performance as high as 50% of reads. This unmatched performance is underpinned by eight years of metadata core innovation. With support for trillions of metadata operations and more than 20 times the number of file systems in a single namespace compared to alternative solutions, FlashBlade//EXA delivers multidimensional performance that powers the most advanced AI models, which include billions of parameters and multimodal data sets. Enterprises can benefit from reduced data processing times and enhanced engagement with AI models, driving more immediate and insightful business outcomes.

FlashBlade/EXA improves manageability at scale by eliminating traditional parallel storage system complexity and redefines simplicity and efficiency in large-scale data environments through its sophisticated yet straightforward design, minimizing the complexity traditionally associated with parallel file systems. Installation times can be reduced by up to 50% compared to competitive systems. FlashBlade//EXA can optimize the ever-growing power and cooling costs associated with energy-hungry GPU environments.

FlashBlade//EXA empowers organizations to keep pace with AI innovation with a highly configurable and disaggregated architecture that enables seamless adaptation to evolving AI workloads, making it a flexible and future-proof investment for enterprises. The disaggregated architecture allows independent scaling of metadata and data nodes, eliminating traditional bottlenecks and ensuring optimal performance irrespective of workload changes. This configurability supports the rapid pace of AI model training and deployment, empowering organizations to stay ahead in the competitive landscape by continuously innovating and upgrading their AI capabilities.

Learn More About FlashBlade//EXA.

Appendix: FlashBlade//EXA technical specifications

The following are hardware descriptions for the test equipment and rack space usage.

Metadata core

- **Scalability**: 1–10 chassis, 10 blades per chassis
- **Capacity:** 1–4 DirectFlash Modules per blade, 37.5TB DirectFlash Module
- **Connectivity with 2 XFMs**: 16x 400GbE uplinks

Physical

Metadata chassis

- Dimensions (per chassis): 5U
- Power: 2600W (nominal at full configuration)

Pair of XFMs

- Dimensions (per XFM): 1U
- Power: 310W (nominal at full configuration)

Data nodes

Scalability

- Unlimited

Minimum CPU and memory requirements

- CPU: 32 cores
- DRAM: 192GB

Capacity per node

- NVMe drives: 12–16 (PCIe Gen4+)
- Drive capacity: 3.8TB–61.44TB
- PCIe Gen5 drives for best performance

Connectivity

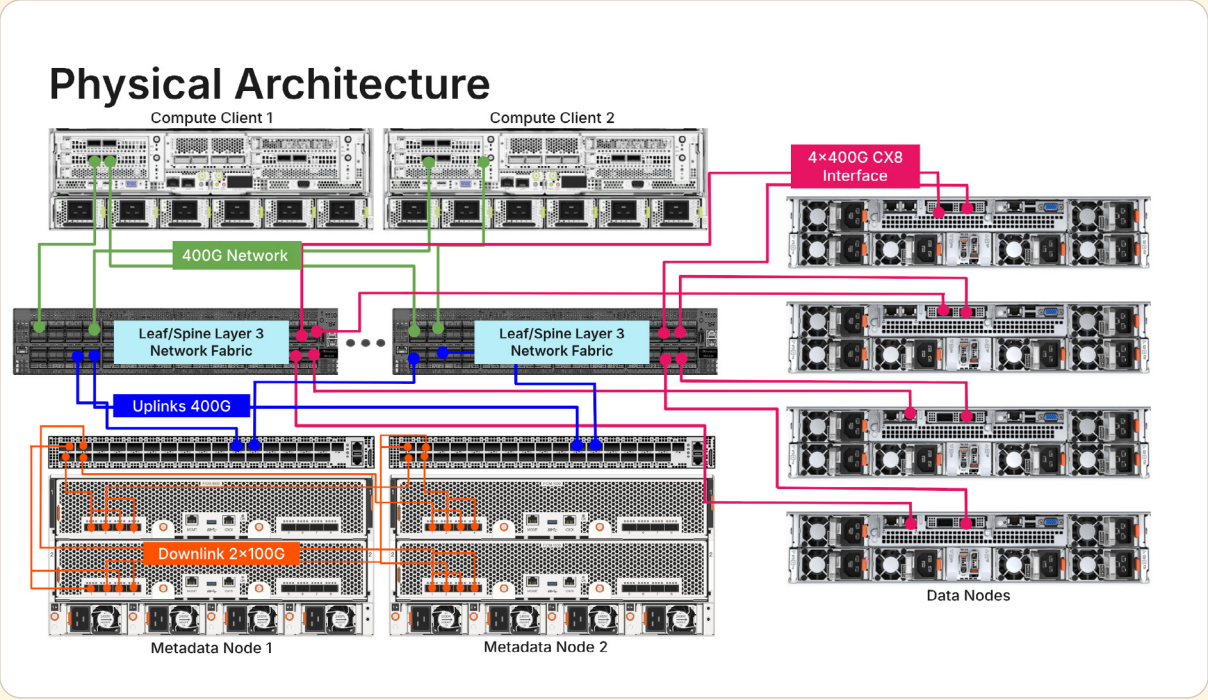
- 2x 400GbE NICs for best performance

Physical

- Minimum dimensions: 1U
- Drive form factor: determined by data node
- Power: determined by data node

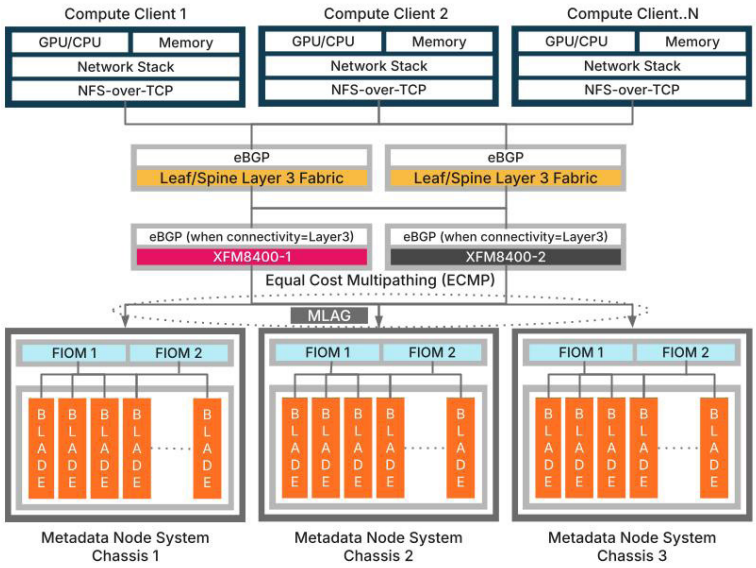
Test setup

- Testing was completed using only three out of a possible 10 metadata nodes.
- Data nodes were tested at 15–59 to verify performance scalability.



Everpure

Metadata Node Connectivity Stack



Metadata I/O Stack

Compute Clients

- Primary supercomputing clustered systems
- Distributed parallel computing applications
- Mix of GPU/CPU-enabled data processing
- pNFS over TCP network stack with eBGP

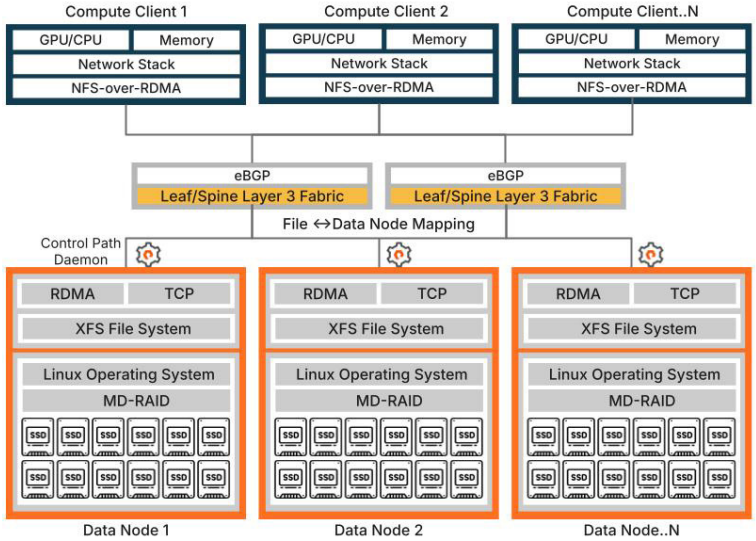
Leaf/Spine BGP Fabric

- Customer supplied Leaf/Spine BGP networks
- Connects compute and storage clusters
- Backbone of hyperscale supercomputing
- Highest network availability and resiliency

Metadata Nodes

- //S500 chassis, Purity OE & FIOMs are MLAG
- Every XFM port has downlink (4×100G) to FIOM ports
- XFM uplinks (4×400G) to Leaf/Spine switches
- XFM complements eBGP network topology

Data Node Connectivity Stack



Data I/O Stack

Compute Clients

- Primary supercomputing clustered systems
- Distributed parallel computing applications
- Mix of GPU/CPU-enabled data processing
- NFS-over-RDMA connection with data nodes, Linux distribution (recommended Linux kernel higher than 6.8 & 6.14); this needs to be patched with Everpure developed patches

Leaf/Spine BGP Fabric

- Customer supplied Leaf/Spine BGP networks
- Connects compute and storage clusters
- Backbone of hyperscale supercomputing
- Highest network availability and resiliency

Data Nodes

- Off-the-shelf servers with commodity SSDs & the FlashBlade//EXA data node operating system (Purity//DN) with XFS filesystem
- NFS-over-RDMA connection with compute
- Everpure supplied thin-client package installed

Visit Our Website | **800.379.PURE**

ID17777-01-en-03/26