Technical Brief: Hammerspace MLPerf Storage v1.0 Benchmark Results

Delivering the best price/performance storage for GPU computing

Summary

The MLCommons MLPerf Storage benchmark is intended to demonstrate the performance of various storage systems for simulated machine learning workloads, so that technical buyers and decision makers have some criteria when evaluating storage system performance for machine learning, deep learning, and other forms of GPU computing.

This year, Hammerspace submitted results for the MLPerf v1.0 Storage Benchmark for the first time, and this technical brief summarizes the results of that benchmark test, including:

- Background on MLCommons and the MLPerf Storage Benchmark
- A summary of Hammerspace's results relative to other vendors, including the test setup used for the benchmark
- A discussion on the advantages of Hammerspace standards-based parallel file system architecture compared to scale-out NAS and HPC parallel file systems

Results prove the price/performance advantage of Hammerspace for high-throughput, low-latency file and object storage, both on-prem and on-cloud.

Hammerspace Price/Performance Advantage

Hammerspace is the only vendor that demonstrated HPC levels of performance with the standard networking and interfaces of Enterprise storage.

A bit of background… Parallel file systems are needed for efficient high-performance computing. Historically they have required specialized client software for every server accessing the storage, applications needed to be designed to operate with the proprietary interfaces of the system, exotic networking such as Infiniband was needed for performance, the systems were often fragile and suffered from more downtime than enterprises or hyperscalers could tolerate, and the deployment/optimization of the file system could take weeks or even months to fine tune. These complexities have led to many organizations assessing scale-out NAS as an easier to use and more relatable alternative. Unfortunately, scale-out NAS cannot deliver the performance needed for computing at scale.

- Scale-out NAS systems require 2x the number of servers and 2x the number of network ports, relative to parallel file systems that create a direct data path between clients and storage. These types of storage solutions also struggle to deliver performance at scale and have not submitted any results to the MLPerf benchmark.
- Traditional HPC parallel file systems like Lustre require proprietary client software that adds complexity and ongoing administrative burden and costs. They require custom hardware, and exotic and expensive networking technology like Infiniband and Slingshot.

Hammerspace brings the best of these technologies together in its Hyperscale NAS, while overcoming the negative challenges of each, to deliver the best price/performance storage for GPU computing in AI, ML, HPC and deep learning. Hammerspace delivers the best price/performance by using:

- 50% fewer servers and 50% fewer network ports than scale-out NAS architectures such as Dell PowerScale, Qumulo, and VAST
- Standard ethernet connectivity, eliminating the need for a specialized second network, such as Infiniband, which is used in the MLPerf benchmarks from other parallel file system-based results from vendors including DDN, HPE and WEKA
- Existing Linux client servers to connect to the file system without specialized client software
- Existing applications natively designed to interface with NFS or S3 without the need to redesign for a parallel file system interface

By reducing the number of servers and switches, there is a very important corresponding decrease in power and cooling that frees up wattage for the compute environment.

Because Hammerspace software supports any server hardware from any vendor, buyers are free to purchase hardware from any source, including OCP servers such as those being used by Meta in their AI Research Supercluster. Meta chose Hammerspace as their high performance solution for provisioning their Llama 2 and Llama 3 LLM training pipelines, because only Hammerspace demonstrated the linear scalability to achieve over 12TB/sec over standard networking, feeding data between Meta's existing 1,000-node NVMe storage cluster and a 3,000-node GPU cluster with 24,000 GPUs in total. No other vendor came close.

About MLCommons® and the MLPerf Storage Benchmark

"MLCommons is an Artificial Intelligence engineering consortium, built on a philosophy of open collaboration to improve AI systems. Through our collective engineering efforts with industry and academia we continually measure and improve the accuracy, safety, speed, and efficiency of AI technologies–helping companies and universities around the world build better AI systems that will benefit society." (from https://mlcommons.org/about-us/)

Hammerspace has team members on the MLCommons Committee and Board for the MLPerf Storage benchmark.

The MLPerf Storage Benchmark Suite consists of several simulated AI workloads. Hammerspace chose to submit results for two of them:

- **U-Net3D:** A visual ML workload segmenting 3D medical imagery. This is a bandwidth-intensive test that opens large files in small batches and reads them sequentially.
- **ResNet-50:** A deep learning convolutional neural network that excels at image classification – detecting objects within images and classifying them accordingly. This test involves concurrently reading many smaller (~100KB) samples from within a large number (>1000) of larger (>100MB) files. Compared to U-Net3D this workload consists of smaller, more random I/O.

For each test, the goal is to demonstrate the maximum number of simulated GPUs ("Accelerators" in MLPerf terminology) that the storage system can simultaneously supply with data, keeping the utilization of every simulated GPU at 90% or higher. Total throughput is also reported in average MB/s.

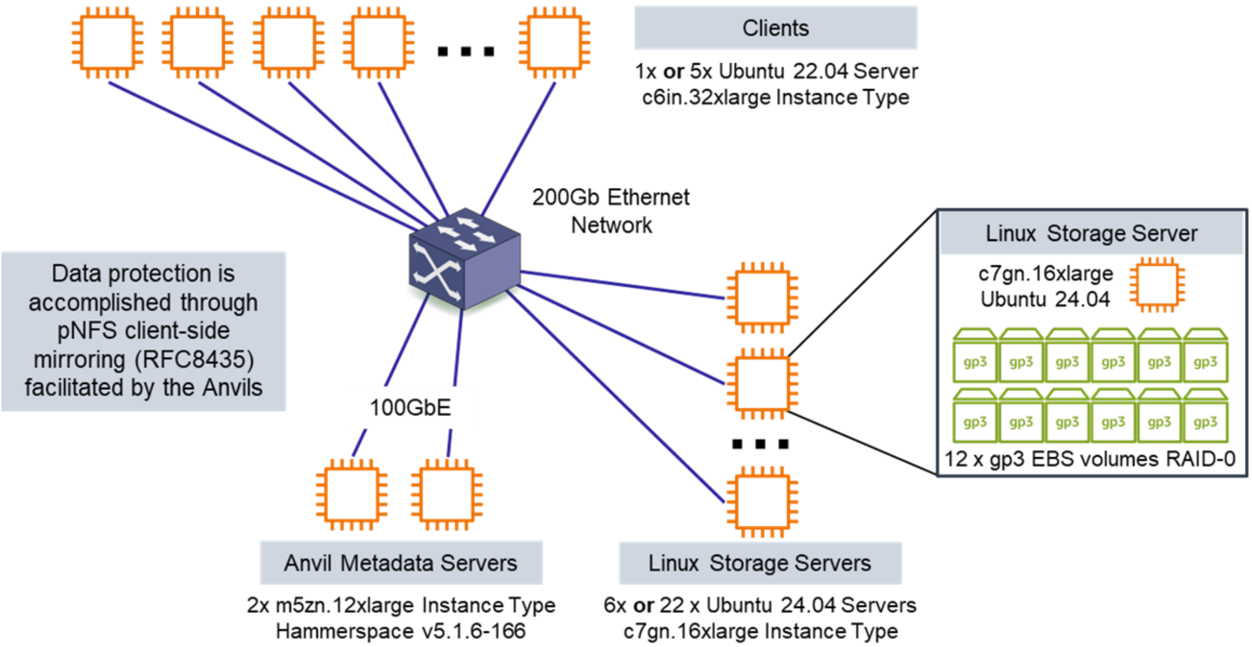
Results Summary

- **Hammerspace Demonstrated Excellent Performance Results:** Hammerspace delivered excellent performance results as measured by the number of simulated GPUs, or Accelerators, and throughput that could be driven from a Hammerspace storage system
- **Hammerspace Does Not Require a Proprietary File System Client:** Unlike other HPC parallel file systems that submitted results to MLPerf, including Lustre and Weka, Hammerspace does not require a proprietary file system client. Instead, Hammerspace uses capabilities built into the NFSv4.2 client to deliver low-latency, high-throughput performance without the complexities of proprietary client software or by connecting to any storage via standard NFSv3

- **Hammerspace Does Not Require Exotic Networking Technologies:** Unlike the other HPC parallel systems which submitted results based on Infiniband or HPE Slingshot networking, Hammerspace results are based on standard ethernet networking.
- **Hammerspace Delivers Performance Both On-Cloud and On-Premises:** Notably, Hammerspace was the only parallel file system vendor that submitted results based on a cloud-native test environment, demonstrating that flexibility of Hammerspace to act as a high performance file system whether deployed on physical hardware or cloud environments.

Test System Configuration

Testing was performed using Amazon Web Services (AWS) public cloud infrastructure. Cloud was selected for this MLPerf submission to show the performance that can be achieved in standard cloud instances without specialized hardware designs. Hammerspace will publish benchmarks run on physical hardware later in 2024.


Clients
1x **or** 5x Ubuntu 22.04 Server
c6in.32xlarge Instance Type
200Gb Ethernet
Network
Data protection is accomplished through pNFS client-side mirroring (RFC8435) facilitated by the Anvils
100GbE
Linux Storage Server
c7gn.16xlarge
Ubuntu 24.04
gp3
12 x gp3 EBS volumes RAID-0
Anvil Metadata Servers
2x m5zn.12xlarge Instance Type
Hammerspace v5.1.6-166
Linux Storage Servers
6x **or** 22 x Ubuntu 24.04 Servers
c7gn.16xlarge Instance Type

The Hammerspace system consisted of two redundant Anvil metadata servers in an active/passive configuration and up to 22 Linux storage server (LSS) nodes. The Anvil servers are responsible for metadata operations and cluster coordination tasks, while the LSS nodes serve test data using associated solid-state EBS volumes storage devices. Single-client tests used six LSS nodes, and multiple-client tests used all 22.

It's important to emphasize that the LSS nodes are just standard Linux servers exporting NFSv3, with no added software.

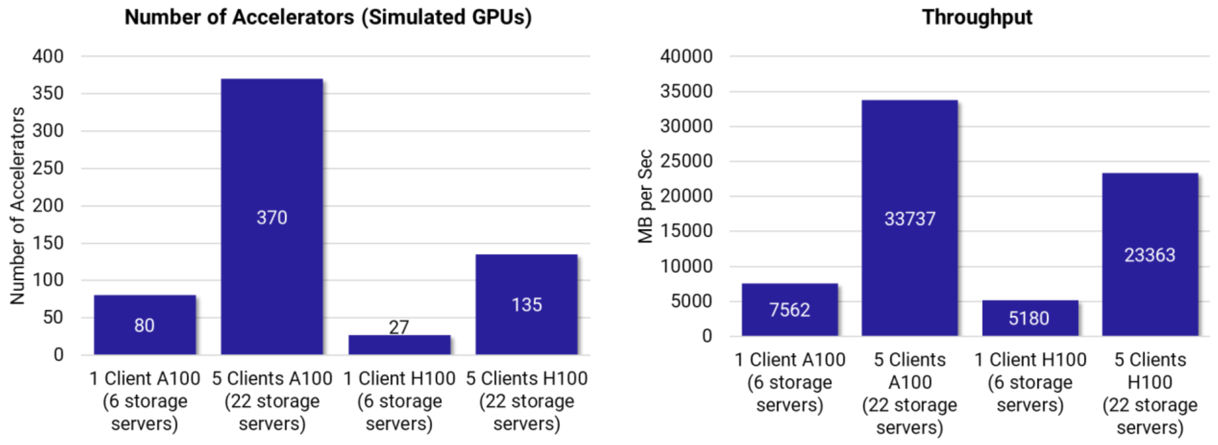
All client systems mounted a Hammerspace share using standard pNFSv4.2. This is the Hyperscale NAS architecture.

Clients and storage servers were connected to the network using 200GbE interfaces. Anvil nodes were connected via 100GbE. Since Anvils are only involved in metadata communication (no data flows through them), 200GbE was not necessary.

Hammerspace Results

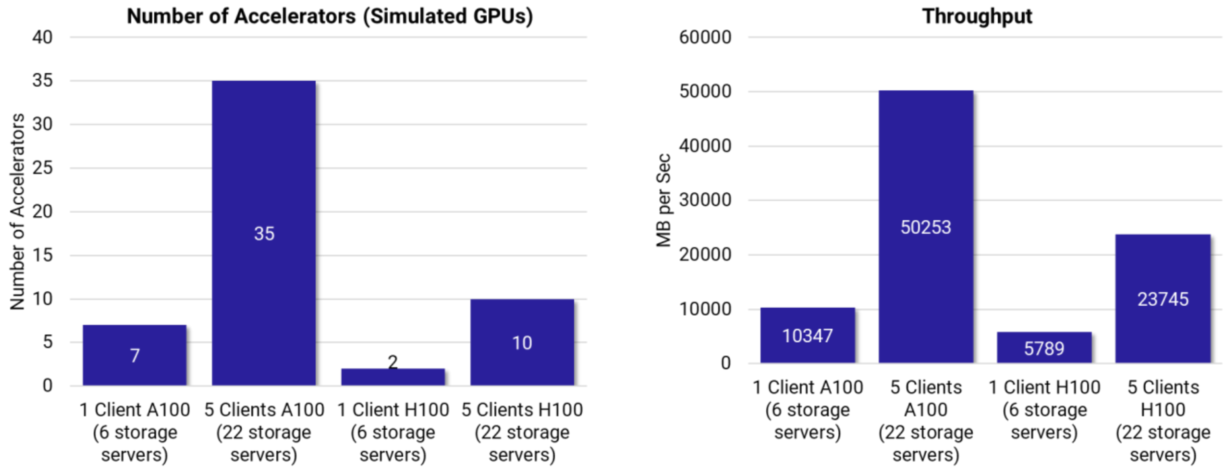
Hammerspace test results are shown in the figures below.

ResNet-50 Results



In the ResNet-50 image classification workload simulation, a Hammerspace system with 22 flash-based Linux storage servers (LSSs) drove 370 simulated A100 GPUs and 135 simulated H100 GPUs to > 90% utilization, delivering 33.7GB/s and 23.3GB/s aggregate read performance, respectively.

U-Net3D Results



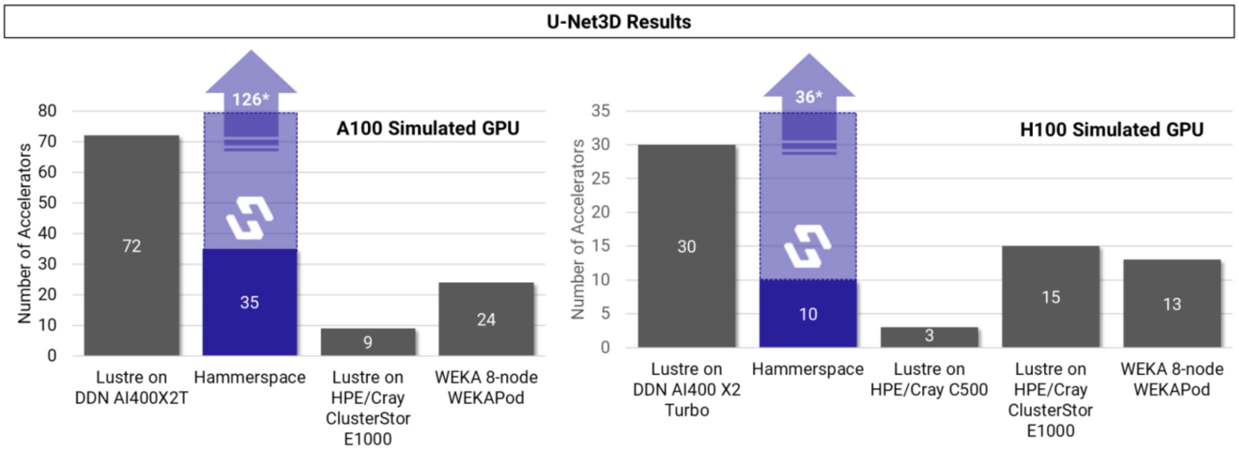
With the U-Net3D simulated image segmentation workload, this system drove 35 simulated A100s and 10 simulated H100s, delivering 50.3GB/s and 23.7GB/s, respectively.

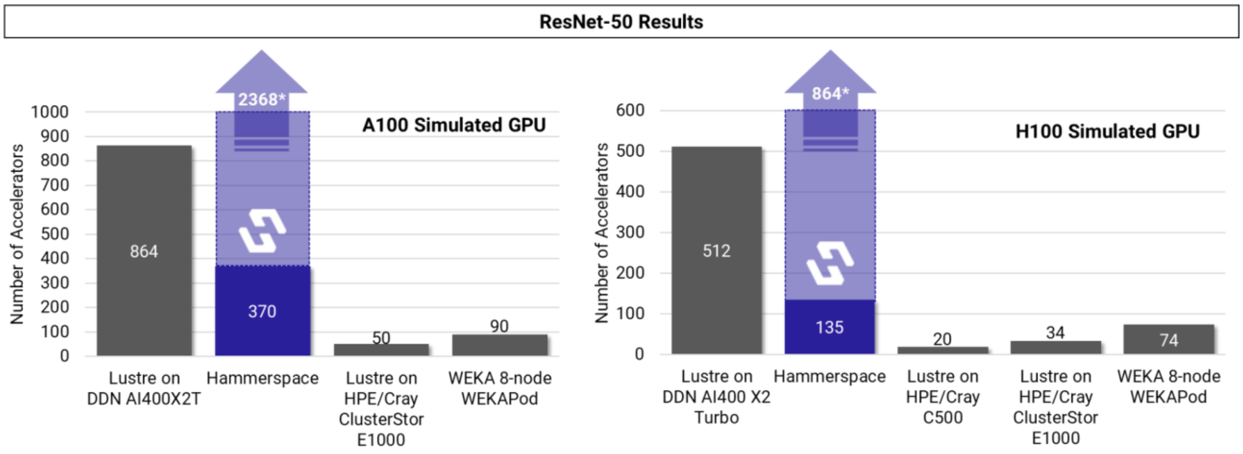
A smaller configuration with six LSSs performed admirably as well, demonstrating the scalability of the system. It supported 80 simulated A100s at 7.6GB/s and 27 simulated H100s at 5.2GB/s in the ResNet-50 test, and 7 simulated A100s and 2 simulated H100s at 10.3GB/s and 5.8GB/s, respectively, in the U-Net3D test.

Both Hammerspace configurations used standard Ethernet networking and standard pNFSv4.2, requiring no special client-side software or agents. Neither was tested to its limits.

Hammerspace Results Relative to Other Vendors

Hammerspace results relative to other parallel file system vendors are shown in the figures below.

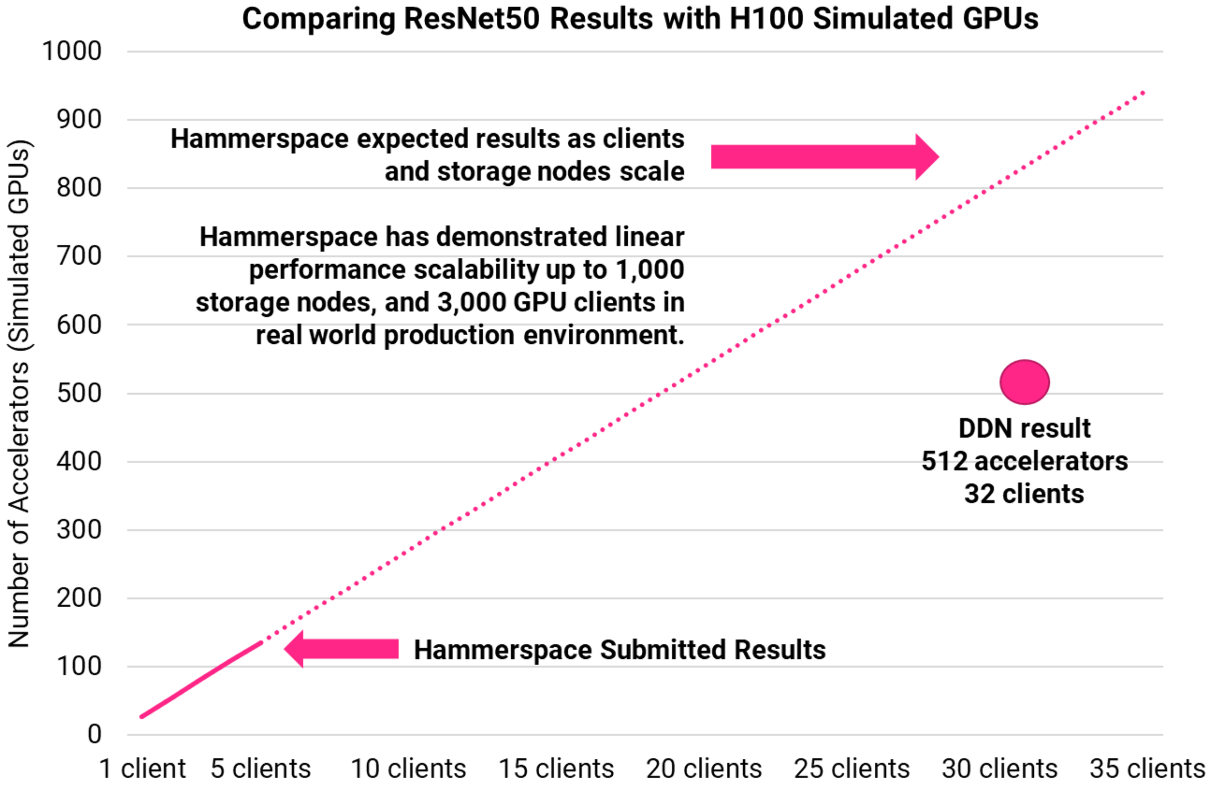




**Hammerspace performance scales linearly by adding more clients and storage nodes. These figures show the expected result with the same number of clients used in DDN’s submission.*

Hammerspace has demonstrated linear performance scalability in real world production environments up to 1000 storage nodes, and 8,000 GPU clients with 24,000 GPUs.

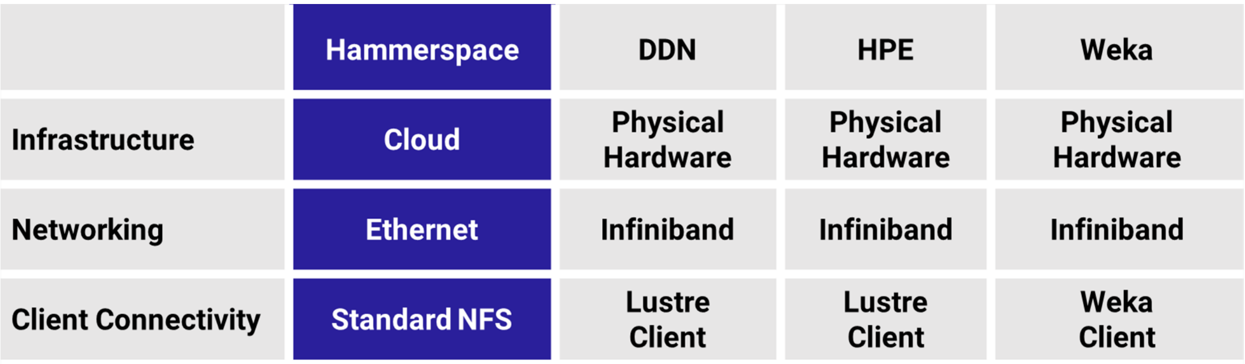
By adding more clients and storage nodes, Hammerspace would be able to deliver results that surpass the leading results, as shown in the diagram below.



Lustre has historically been considered the “gold standard” for high-performance file systems. It is a parallel file system that requires proprietary client software and typically runs on Infiniband or Slingshot networking. Lustre adoption has been limited primarily to HPC organizations in research as it requires proprietary client software, added networking, plus applications need to be designed to work with Lustre’s proprietary interfaces and architecture, which makes it difficult to deploy and optimize.

Hammerspace results are within the same range as the Lustre submissions from DDN and HP. This provides empirical evidence that Hammerspace is as fast as the gold standard for HPC-class parallel file systems, and it also includes Enterprise standard RAS features that Lustre can’t provide.

The figure below summarizes the key architectural differences between Hammerspace and the other parallel file system vendor test set ups.



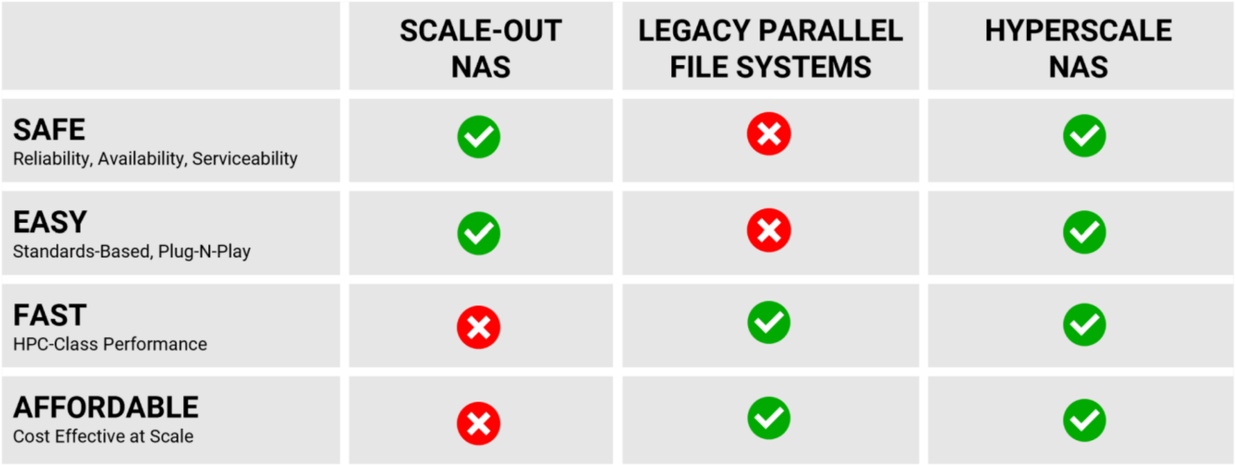
	Hammerspace	DDN	HPE	Weka
Infrastructure	Cloud	Physical Hardware	Physical Hardware	Physical Hardware
Networking	Ethernet	Infiniband	Infiniband	Infiniband
Client Connectivity	Standard NFS	Lustre Client	Lustre Client	Weka Client

Hammerspace Architecture

The Hammerspace Data Platform has three key technologies all in a single software solution, with a single all-inclusive license:

1. Multiprotocol Global Namespace
2. Parallel Global File System
3. Automated Data Orchestration

The high-performance storage portion of Hammerspace technology, which combines our parallel file system and global namespace, is delivered as a Hyperscale NAS. Hyperscale NAS uses standard Ethernet and requires no proprietary client like a traditional enterprise-class scale-out NAS, , but couples this with the parallel file system performance and linear scalability found in industry-leading parallel file systems in the HPC space.

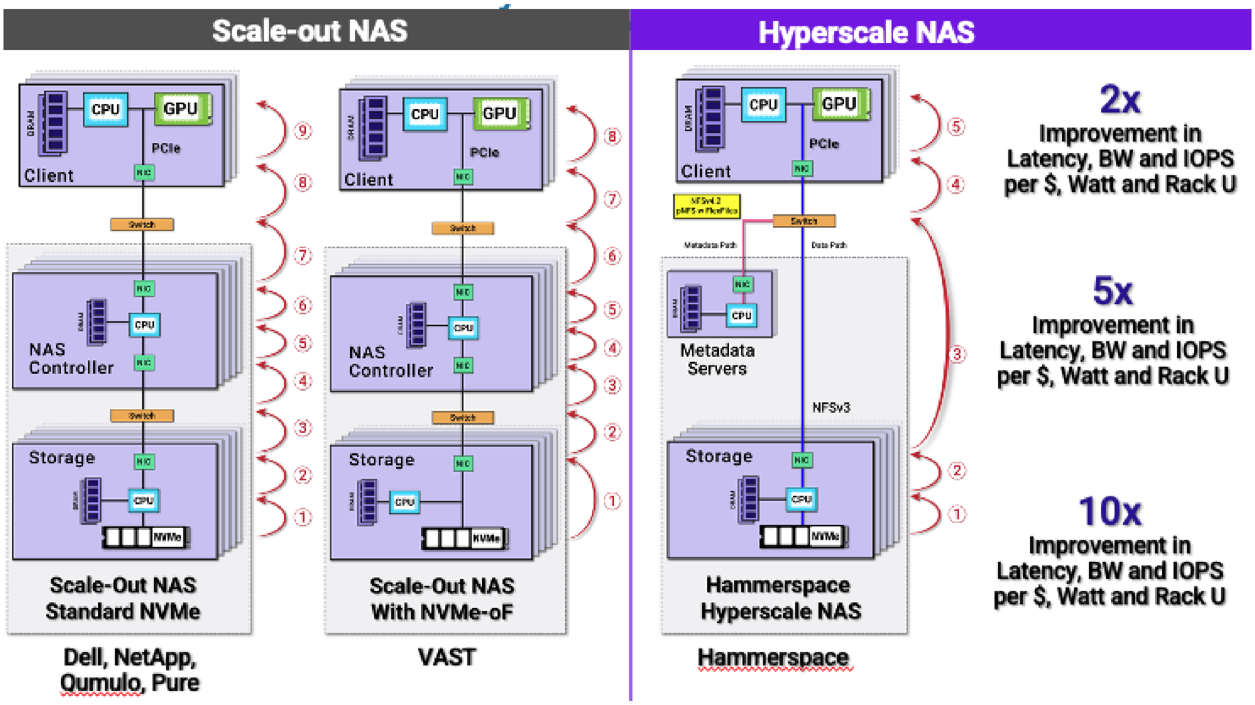


	SCALE-OUT NAS	LEGACY PARALLEL FILE SYSTEMS	HYPERSCALE NAS
SAFE Reliability, Availability, Serviceability	✓	✗	✓
EASY Standards-Based, Plug-N-Play	✓	✗	✓
FAST HPC-Class Performance	✗	✓	✓
AFFORDABLE Cost Effective at Scale	✗	✓	✓

Hammerspace has invested heavily into enhancing the standard NFS protocol to include a fast, feature rich, parallel file system client that is built into the Linux kernel of all commercial distributions.

Advantages of Hammerspace Hyperscale NAS Architecture

It is notable that no scale-out NAS vendor submitted results as part of the MLPerf Storage Benchmark. Well known NAS vendors like Dell, NetApp, Qumulo, and VAST Data are absent. Why wouldn't these companies submit results? Most likely it is because there are too many performance bottlenecks in the I/O paths of scale-out NAS architectures to perform well in these benchmarks.


Scale-out NAS
Hyperscale NAS
CPU
GPU
DRAM
PCIe
Client
NIC
Switch
NAS Controller
Storage
NVMe
Scale-Out NAS Standard NVMe
Dell, NetApp, Qumulo, Pure
Scale-Out NAS With NVMe-oF
VAST
NFSv4.2 pNFS w FlexFiles
Metadata Path
Data Path
Metadata Servers
NFSv3
Hammerspace Hyperscale NAS
Hammerspace
①
②
③
④
⑤
⑥
⑦
⑧
⑨
2x
Improvement in Latency, BW and IOPS per $, Watt and Rack U
5x
Improvement in Latency, BW and IOPS per $, Watt and Rack U
10x
Improvement in Latency, BW and IOPS per $, Watt and Rack U

A traditional Scale-out NAS design requires a NAS controller in the storage system. There are two key areas where this causes issues:

1) Performance, even at small scale

a) The eight or nine hops a single bit needs to take for a read or write operation in scale-out NAS architectures introduces latency

b) The data and the metadata in those systems are sharing the same network path, both trying to squeeze in to a fixed amount of network capacity (kind of like a traffic jam in Los Angeles – only so many cars can fit on the highway to go straight (Data path), and, when you add in cars merging to get to exits or the HOV lane, additional friction is created (metadata path)

c) If anything breaks or goes offline in the data path, the environments become fragile and slow, or go down completely, because they don't have the client-side intelligence and the telemetry feedback loop to automatically reroute around blockages which is part of the pNFS v4.2 standard.

2) Performance at large scale

Scale-out NAS architectures face performance thresholds at scale due to several key limitations inherent in their design:

1. **Metadata Bottleneck**: File system metadata is typically managed by a limited set of central controller servers. As the number of files and the volume of data grow, these controllers can become bottlenecks, slowing down file access and system performance that are also bottlenecked through the same controller.
2. **Network Overhead**: Since scale-out NAS architectures distribute data across multiple nodes, they rely heavily on network infrastructure to connect those nodes to each other behind the centralized controller. As the system scales, the internal network traffic and cache contention (for data retrieval, replication, synchronization, etc.) increases, and standard Ethernet or other networking technologies may not provide sufficient bandwidth, creating a performance ceiling.
3. **Concurrency Limits**: Scale-out NAS systems often face challenges with concurrent access, particularly when many users or applications are accessing the same data. Locking mechanisms to ensure data consistency (such as file locks) can cause delays and limit scalability, as more clients try to access or modify the same files simultaneously.

4. **Data Distribution Overheads**: As the system scales, managing where data is stored across the storage nodes becomes increasingly complex. Some scale-out NAS systems distribute data using techniques like hashing or striping across multiple nodes, which adds overhead as the system tries to keep track of where each piece of data resides.
5. **Protocol Inefficiencies**: Traditional client-side NAS protocols like NFSv3 or SMB, while widely used, are not optimized for massive scalability. They have a serious performance problem of excessive chattiness between the client and the server, because there is no intelligence in the client to retain state. This leads to overhead in terms of session management, data transfer, and security features that may limit throughput as the system scales.
6. **Non-Linear Scaling of I/O**: While additional storage nodes increase capacity, they can't linearly scale to improve input/output (I/O) performance. Data movement between nodes, replication overhead, and inter-node communication can cause I/O bottlenecks, limiting scalability of scale-out NAS architectures from all vendors.
7. **Latency Amplification**: As the number of storage nodes increases, latency from factors like network hops, inter-node communication, and coordination overhead grows, impacting performance.

Hammerspace Delivers High Performance On-Premises and In the Cloud

An NVIDIA survey to their customer base in 2024 indicated that 49% of their customers plan to run AI projects ***both on-premises and in-cloud.*** In other words, about half of NVIDIA's customers will require high-performance file and object storage that can run on-premises (when GPU clusters are local), and in the cloud (to access GPU resources in the public cloud). Hammerspace MLPerf results demonstrate excellent performance in a 100% cloud-based environment, showcasing the ability for enterprises to achieve low-latency, high-throughput file storage no matter where it runs.

And because Hammerspace is a global file system that can span sites and multiple clouds, with data orchestration services that automate the flow of data, it means that customers can bring their data to the compute resources as needed, whether those compute resources are local or cloud-based.