

PERFORMANCE ANALYSIS

Accelerating AI/ML Workloads with a Hammerspace Tier 0 Architecture



HAMMERSPACE

Overview

A leading quantitative research firm sought to maximize storage performance across its GPU computing infrastructure and evaluated the performance of a Hammerspace Tier 0 architecture against traditional networked storage solutions. The Tier 0 architecture activates local NVMe storage across a compute cluster, transforming it into a high-performance shared storage tier that delivers significant improvements for GPU operations when compared against I/O that must traverse the network to/from external storage arrays.

The results demonstrate that the Hammerspace Tier 0 architecture delivered superior performance for sequential operations and large-scale parallel workloads, with **2x-4.4x bandwidth improvements and 56% lower latency** compared to the networked storage.

Simulated AI/ML workload performance also showed good comparative results, with the Tier 0 architecture delivering **75% higher throughput for random sampling operations and 66% faster checkpoint saves**.

Hammerspace Tier 0 Architecture Performance Results

4.4x

More Bandwidth for
sequential ops

75%

Higher throughput for random
sampling operations

56%

Lower Latency

66%

Faster checkpoint saves

The Hammerspace Tier 0 solution also demonstrated good scalability characteristics, **maintaining linear performance** increases as additional clients were added to the test environment.

This white paper summarizes the test environment and test methodology, compares performance results and resource utilization, and draws conclusions applicable to designing GPU computing and AI/ML storage architectures.



Test Environment

This testing was conducted by the staff of the quantitative research firm, using equipment from their data center. Hammerspace personnel assisted in the initial deployment and configuration of the solution.

GPU Server Hardware Specifications

- The benchmarking was performed on four GPU servers, each equipped with:
- AMD EPYC 9655 CPUs (96 physical cores per server)
- 8 local NVMe drives
- 6 TB RAM per node
- Dual bonded 200 GbE NICs
- Ubuntu 24.04.2 LTS operating system with kernel 6.8.0-71-generic

Hammerspace Tier 0 Configuration

The Hammerspace Tier 0 architecture was configured as follows:

- 4 NVME drives local to each CPU socket configured in RAID 0
- Each RAID 0 volume is accessed through a bonded NIC connection
- The design leverages CPU socket locality to minimize inter-socket communication overhead, and minimal data path traversal between storage and compute resources
- Hammerspace Tier 0 objectives were configured with durability of three-nines and availability of one-nine

Hammerspace Networked Storage Configuration

- Hammerspace Data Services (DSX) nodes connected using the NFS protocol served as the comparison baseline, representing a more traditional networked storage architecture.
- Hammerspace DSX objectives were configured with durability of three-nines and availability of one-nine, matching the Hammerspace Tier 0 configuration to ensure a fair comparison.



Test Methodology

Benchmark Procedures

Performance testing was conducted using FIO (Flexible I/O Tester) with the following parameters:

- Sequential read and write operations
- Block sizes: 4KB and 1MB
- Jobs: 16 concurrent I/O threads
- Testing scales: 1-4 clients

Additional testing using the elbencho benchmark suite was conducted to evaluate storage performance under simulated AI/ML workloads, which are particularly relevant for GPU computing environments.

Monitoring

System metrics were collected using Prometheus, recording:

- Storage performance metrics
- CPU utilization
- Memory utilization
- Network throughput



Bandwidth, Latency, and IOPS Results (FIO)

Flexible I/O (FIO) benchmark was used to compare bandwidth, IOPs, and latency across different block sizes. The results are summarized at a high level in this table and discussed in more detail below.

In addition to this benchmark testing, Hammerspace demonstrated superior results for a Tier 0 architecture in the latest MLCommons MLPerf Storage benchmark suite, a public benchmark test measuring storage system performance for simulated GPU computing workloads.

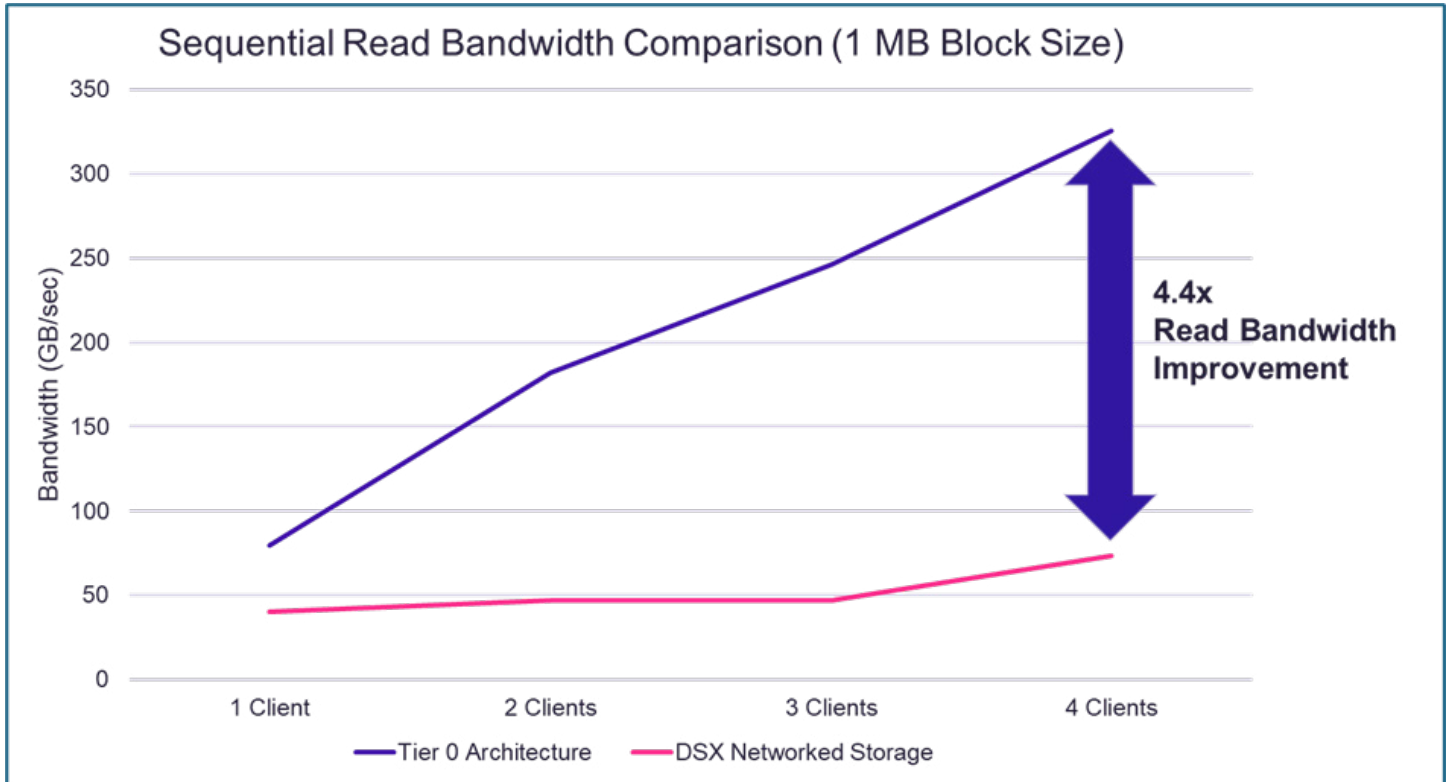
	Hammerspace Tier 0 Architecture	Hammerspace Networked Storage	Tier 0 Improvement
Read Bandwidth (1 MB block size, 4 clients)	325.4 GB/s	73.3 GB/s	4.4x
Write Bandwidth (1 MB block size, 1 client)	28.49 GB/s	12.71 GB/s	2.2x
IOPS (Random read, 4KB block size, 4 clients)	325,349 IOPS	73,333 IOPS	4.4x
Latency (Sequential write, 1 MB block size)	4.26 ms	9.72 ms	56% latency reduction
Latency (Sequential read, 4 KB block size)	0.45 ms	0.63 ms	28.6% latency reduction



Bandwidth Comparison

Sequential Read Performance (1 MB block size)

As shown in the chart below, Tier 0 delivered up to **325.4 GB/s**, vs **73.3 GB/s** on DSX (344% improvement at 4 clients), and demonstrated linear performance scaling.



Sequential Write Performance (1MB Block Size)

For sequential write operations with 1MB block size:

- Hammerspace Tier0 delivers 28.49 GB/s bandwidth
- Hammerspace DSX provides 12.71 GB/s bandwidth
- Result: **124.2% bandwidth improvement** with Hammerspace Tier0



IOPS Comparison

Random Read Performance (4KB Block Size)

For smaller block operations, IOPS were compared and the results are summarized in the table below.
IOPS scaled from **79K (1 client)** to **325K (4 clients)** with Tier 0, compared to **40K to 73K** with DSX.

Configuration	Tier 0 IOPS	DSX IOPS	Improvement
1 Client	79,343	40,055	98%
2 Clients	181,903	47,209	285%
3 Clients	245,251	47,173	422%
4 Clients	325,349	73,333	344%

Note: IOPS values directly correspond to the bandwidth measurements, as expected for fixed block size operations.
These measurements demonstrate that a Hammerspace Tier 0 architecture provides higher transaction rates with improvement at the 4-client configuration, making it well-suited for workloads requiring high IOPS across multiple systems

Latency Comparison

Sequential Write Performance (1MB Block Size)

The write latency measurements further emphasize Hammerspace Tier 0’s advantages:

- Hammerspace Tier0: 4.26ms average write latency
- Hammerspace DSX: 9.72ms average write latency
- Result: **56.2% latency reduction** with Hammerspace Tier 0

Sequential Read Operations (4KB Block Size)

For sequential read operations with 4KB block size:

- Hammerspace Tier0: 0.45ms average latency
- Hammerspace DSX: 0.63ms average latency
- Result: **28.6% latency reduction** with Hammerspace Tier 0



Random Read Operations (4KB Block Size)

For random read operations with small block sizes, the comparison reveals a different pattern:

- Hammerspace Tier0: 0.21ms average latency
- Hammerspace DSX: 0.15ms average latency
- Result: **28.6% lower latency** with Hammerspace DSX

These findings highlight the importance of matching storage architecture to specific application workload characteristics. While Hammerspace Tier 0 demonstrates superior performance for sequential and large-block operations, applications dominated by small-block random access patterns may benefit from the Hammerspace DSX architecture in specific scenarios.

Simulated AI/ML Workload Results (elbencho)

In order to assess the performance of a Hammerspace Tier 0 architecture in AI/ML workloads, the **elbencho** benchmark suite was used. Elbencho is a high-throughput, parallelized benchmark suite designed to stress and evaluate the performance of storage systems across real-world AI/ML and HPC workloads, measuring key metrics like IOPS, throughput, and metadata operations to validate architecture scalability and efficiency.

The performance results are summarized and discussed below.

Workload Type	Tier 0 Performance	DSX Performance	Improvement
Bandwidth Comparison			
Random Sampling (128KB)	75.9 GB/s	43.5 GB/s	74.8%
Sharded Training (1MB)	69.5 GB/s	45.1 GB/s	54.1%
Model Loading (40GB)	12.8 GB/s	11.7 GB/s	9.0%
Checkpoint Save (40GB)	5.6 GB/s	1.8 GB/s	209.7%
Latency Comparison			
Random Sampling	10.7 ms	9.7ms	10.3% higher
Sharded Training	1.3 ms	5.8 ms	77.6% lower
Model Loading	4.0 ms	9.6 ms	58% lower
Checkpoint Save	43.9 ms	211.9 ms	79.3% lower



These benchmark results underscore how Hammerspace Tier 0 delivers transformative performance gains for AI workloads at every stage of the pipeline:

- **Training (including sharded and random sampling):**
 - The **74.8% boost in random sampling throughput** significantly accelerates data loading during training epochs, reducing idle GPU cycles and improving overall throughput, especially for deep learning models that rely on randomized dataset shuffling (e.g., image classification, NLP).
 - The **54.1% faster access to sharded training data** is critical in distributed training scenarios, enabling higher GPU utilization and shorter training durations.
- **Checkpointing and Model Management:**
 - A **209.7% improvement in checkpoint saving** directly reduces training interruptions during model state persistence—vital for large model training where frequent, fast checkpointing ensures fault tolerance without derailing performance.
 - With **79.3% lower checkpoint latency**, long-running jobs can checkpoint more frequently without impacting training cadence, supporting more robust recovery and experimentation.
- **RAG and Inference:**
 - Faster **random access and lower model load latencies** are especially valuable in **RAG pipelines**, where embeddings and documents must be retrieved and scored rapidly. Tier 0's performance reduces time-to-first-token (TTFT) in systems where real-time or near-real-time inference is required.
 - For **inferencing**, where models are frequently loaded/unloaded or served from cold starts, Tier 0's **reduced model load times** ensure faster startup and lower tail latency—key for user-facing applications.
- **Latency Impact Across the Pipeline:**
 - Lower latencies for **model loading (58.3% reduction)** and **sharded training (77.6% reduction)** directly benefit dynamic inference environments and elastic training setups, such as Kubernetes-based AI platforms that rapidly spin up and tear down models or services.
 - Even in **random sampling**, where Tier 0 latency was slightly higher (10.3%), the overall throughput advantage outweighs the delta in most practical training scenarios.

In summary, Hammerspace Tier 0's benchmarked performance gains directly translate into **faster training cycles, higher GPU efficiency, resilient model management, and lower latency inference**—all of which are essential for scaling real-world AI workloads efficiently across hybrid cloud environments.



System Resource Utilization

As part of this performance analysis, CPU and memory utilization on the GPU servers were compared.

CPU Utilization

The FIO sequential read test with 1MB blocks showed:

- Hammerspace Tier0: 48.5% average CPU utilization
- Hammerspace DSX: 13.3% average CPU utilization

The higher CPU utilization for Hammerspace Tier 0 is expected and represents the tradeoff for direct local processing of I/O operations versus the offloaded network processing model of networked storage.

When evaluating AI/ML workloads through elbencho testing, we observed similar patterns:

- Random Sampling: 72% CPU (Hammerspace Tier0) vs 19% CPU (Hammerspace DSX)
- Sharded Training: 13% CPU (Hammerspace Tier0) vs 9% CPU (Hammerspace DSX)
- Model Loading: 4% CPU (Hammerspace Tier0) vs 4% CPU (Hammerspace DSX)

The significantly higher CPU utilization for random sampling operations with Hammerspace Tier 0 correlates with its substantially higher throughput, representing efficient use of local compute resources to achieve maximum I/O performance.

Memory Utilization

Memory usage was minimal and comparable between configurations:

- Hammerspace Tier 0: 1.1% memory utilization
- Hammerspace DSX: 1.0% memory utilization

These figures indicate that neither solution places significant pressure on system memory resources.

Conclusion

The Hammerspace Tier 0 storage architecture turns GPU server-local storage into a tier of high-performance shared storage for AI/ML and HPC workloads and demonstrates significant performance advantages over traditional networked storage.

Highlights from this performance analysis include:

- **Exceptional Performance for Sequential Operations**, with bandwidth improvements of 4.4x for reads and 2.2x for writes compared to Hammerspace DSX. The latency advantages are similarly substantial, with reductions of 29% for reads and 56% for writes.



- **Faster Training Cycles, Maximized GPU Utilization, Faster Time-to-First-Token (TTFT) for AI/ML Workloads:** By delivering higher throughput and significantly lower latencies for critical operations like random sampling, sharded data access, model loading, and checkpointing, Tier 0 enables faster training cycles, maximized GPU utilization, and reduced time-to-first-token (TTFT) in production inference pipelines.
- **Linear Scalability:** One of Hammerspace Tier 0's most compelling advantages is its linear scaling capability as clients are added to the configuration. This characteristic is particularly valuable for environments requiring the aggregation of storage performance across multiple systems.
- **Resource Utilization Considerations:** The performance advantages come with significantly higher CPU utilization as summarized above. During sequential write operations, Hammerspace Tier 0 showed 48.5% CPU utilization compared to just 13.3% for Hammerspace DSX. This represents a fundamental architectural tradeoff: the Hammerspace Tier 0 approach utilizes local CPU resources to process I/O operations directly, whereas the Hammerspace DSX solution offloads much of this processing to the storage server. While this higher CPU consumption might be a consideration for CPU-bound workloads, for most I/O-intensive applications the tradeoff is advantageous, as the substantially improved storage performance typically outweighs the increased CPU overhead.
- **Workload Considerations:** The performance profile is not uniformly superior across all workload types. For small-block random read operations, Hammerspace DSX demonstrated lower latency and higher IOPS in specific test cases, suggesting potential advantages for certain database or transaction-processing workloads.

The results highlight the importance of matching storage architecture to specific application requirements. The Hammerspace Tier 0 architecture is particularly well-suited for data-intensive applications with sequential access patterns or requiring high aggregate throughput, such as scientific computing, AI/ML model training and large-scale analytics.

For environments where storage performance constitutes a primary bottleneck and data locality can be managed appropriately, the Hammerspace Tier 0 architecture provides a compelling solution that significantly outperforms traditional networked storage in most scenarios.

To learn more, visit [Hammerspace.com/tier-0](https://hammerspace.com/tier-0).

