Enable AI performance and power efficiency at scale in critical workloads

AI has a power problem. As AI workloads scale, power delivery and power consumption are becoming limiting factors in how systems are deployed and expanded.

Data-center energy use, which was under 415 terawatt-hours (TWh) in 2024, is projected to reach 945 TWh by 2030, more than doubling by the end of the decade.[1] Addressing this challenge requires reconsidering how core architectural choices, including storage architecture, shape both performance and power efficiency at scale.

This analysis evaluates how storage architecture affects AI training performance, scalability, and power efficiency by comparing a high-capacity SSD-based system built with the Micron 6600 ION SSD with capacity-matched HDD architectures[2] using MLCommons® MLPerf® Storage v2.0 benchmark suite workloads.[3]

This analysis shows that as accelerator count increased, the 1x SSD architectures sustained higher samples/sec and training throughput with better power efficiency, elevating storage architecture to a key factor in high-performance AI platforms.[4]

Micron 6600 ION SSD (U.2, E3.L)

Scaling AI training; reduce power and complexity.

This paper shows how storage architecture reshapes AI training at scale, explaining why high-capacity NVMe™ SSDs deliver higher samples/second and better peak throughput while consuming less power.

6.3x
samples/sec

A production-scale SSD configuration with 10x Micron 6600 ION SSDs supported 6.3x the samples per second and 6.3x the peak training throughput of 77x 32TB data center HDDs.[5]

6.3x
training throughput

That same production-scale SSD configuration consumed 84% less storage power at these sample/sec and training throughput rates.

84%
less power

The Micron 6600 ION delivers industry-leading capacity—122TB (E3.S) and 245TB (E3.L)—with PCIe® Gen5 and Micron QLC NAND performance, purpose-built for AI, cloud, and data centers to scale sustainably.[6]

micron.com/6600-ION

1. See "AI is set to drive surging electricity demand from data centers while offering the potential to transform how the energy sector works" iea.org for additional information on data center power consumption.
2. We used 16x 16TB Seagate EXOS data center HDDs.
3. We used the MLCommons MLPerf Storage v2.0 benchmark (3D U-Net medical image segmentation workload) to compare the performance of SSD-based and HDD-based configurations. The results shown were run in Micron's data center workload engineering lab and are not official MLPerf Storage results. See the MLCommons home page to learn more about MLCommons, its charter, and the benchmarks it develops.
4. Power efficiency is calculated as (work done) / (power to do that work).
5. 77x HDD and 8x HDD results are extrapolated from 16x HDD measured data via multiplication of 16TB HDD values (linear scaling). Samples/sec calculated as: (10x SSD) / (77x HDD) = 617.1 / 98.4 ≈ 6.3. Peak training throughput calculated as: (10x SSD) / (77x HDD) = 86,283 / 13,761 ≈ 6.3. Power difference calculated as ((77x HDD power) – (10x SSD power)) / (77x HDD power) = 819/980 ≈ 84% less.
6. Industry-leading statement based on public information on shipping data center SSDs at the time of this document's initial publication.

The following analysis examines 3D U-Net medical image segmentation workload results comparing capacity-focused SSD and HDD architectures across varying numbers of simulated GPU accelerators (of the type commonly deployed in large-scale model training). Results report training and data throughput power consumption, to enable performance and power-efficiency comparisons. The MLPerf Storage results presented are unofficial and were not peer reviewed by MLCommons.

Configurations tested

Modern AI infrastructure decisions are rarely made at a single scale. Data center architects typically evaluate both entry-level deployments, often as proof-of-concept systems, and production-scale platforms, ensuring that architectural choices remain sound as data volumes, system capacity, throughput, and power requirements grow.

For that reason, this study evaluated two capacity-matched configuration pairs to examine how SSD and HDD architectures behave at both proof-of-concept and production scale.

Parameter	SSD configuration	HDD configuration[5]
Drive type	Micron 6600 ION SSD 245TB	Capacity-focused, data center HDD
Capacity per drive[7]	245TB	16TB
Interface	NVMe (PCIe Gen5)	SATA 6Gb/s
Baseline configuration	1x 245.76TB Micron 6600 ION SSD (245TB)	16x 16TB HDD (256TB) (8x 32TB HDD)
Scale configuration	10x 245.76TB Micron 6600 ION SSD (2,457TB)	77x 32TB HDD (2,464TB)

Table 1: Configurations overview

Baseline configuration: This configuration represents the smallest scale at which high-capacity flash can be evaluated as an alternative to HDD-based storage. At roughly 250TB, it reflects a common entry point for AI training and imaging workloads, enabling a direct comparison of SSD and HDD architectures without the noise or complexity of large-scale systems. Throughout this document, baseline configurations are referred to as "1x SSD," "16x 16TB HDD," and "8x 32TB HDD."

Production-scale configuration: This configuration describes an AI infrastructure at scale. At multi-petabyte capacity, storage becomes a systems problem, where device count directly drives power, cooling, density, and operational complexity. By scaling flash and HDD-based architectures to equivalent capacity, this configuration highlights how system behavior changes when flash replaces HDDs at scale. Throughout this document, production-scale configurations are referred to as "10x SSD" and "77x 32TB HDD."

3D U-Net analysis | Better diagnosis and treatment planning

Simulating real-world workloads (and virtual accelerators) with a standard benchmark tool can help healthcare providers make storage architecture decisions that deliver data faster for more accurate diagnoses.

Figure 1 shows how training throughput can be scaled with the number of simulated accelerators for both baseline configurations.

At one accelerator, the 1x SSD configuration delivered 2,232 MB/s, compared to 2,859 MB/s for the 16x HDD array, indicating comparable performance at low I/O demand (one accelerator).

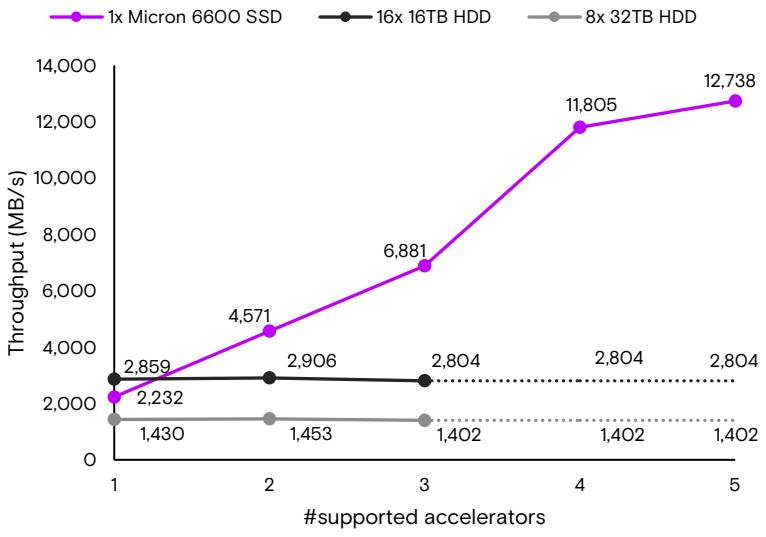

Figure 1: Baseline configuration performance

7. All capacity statements refer to rated capacity. Formatted capacity will be less. 1GB = 1 billion bytes. 1x 32TB HDD and 77x 32TB HDD throughput was extrapolated from the measured 16TB HDD throughput multiplication (linearscaling) of 16TB HDD values

However, as accelerator count increased, the 1x SSD configuration continued to scale, while the 16x 16TB HDD quickly plateaued at three accelerators. With five accelerators, 1x SSD reached 12,738 MB/s, delivering ≈ 4.4x advantage over the 16x HDD peak of 2,906 MB/s (12,738 / 2,906 ≈ 4.4) at only two accelerators.

These results are summarized in Table 2.

	MB/s			Samples/sec.		
Accelerators	1x SSD	16x 16TB HDD	8x 32TB HDD	1x SSD	16x 16TB HDD	8x 32TB HDD
1	2,232	2,859	1,430	15.97	20.45	10.23
2	4,571	2,906	1,453	32.7	20.78	10.39
3	6,881	2,804	1,402	49.22	20.06	10.03
4	11,805	2,804	1,402	84.44	20.06	10.03
5	12,738	2,804	1,402	91.11	20.06	10.03

Table 2: Baseline configuration results details

In workloads such as medical imaging (CT, MRI, PET scans), scientific and industrial 3D imaging, and volumetric analysis pipelines, the baseline SSD architecture enabled accelerator-driven scaling. In contrast, baseline HDD architectures quickly capped performance and utilization. Thus, the 1x SSD configuration was both faster and better suited for scalable AI training pipelines (at scale, HDD-based architectures could not sustain accelerator-driven workloads).

HDD capacity growth | Larger HDDs, same array capacity, lower array throughput

Having established how SSD and HDD architectures behave at a common entry capacity, we extended the analysis to similar capacity configurations where the 16x HDD capacity was doubled.

This evaluation quantifies the effects of per-HDD capacity growth on workload throughput while keeping the HDD array capacity constant.

The results in Figure 2 showed an additional structural disadvantage of HDD-based scaling: as individual HDD capacity increases, the number of drives per the same-capacity array decreases.

This HDD count reduction (from 16 to 8) reduced both training throughput (samples/s) and peak throughput (MB/s), working against HDD array performance.[8]

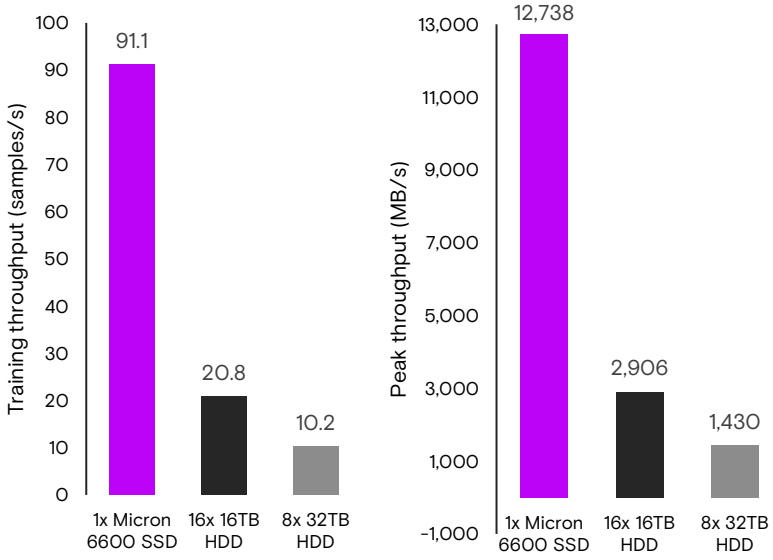

Figure 2: Throughput for SSD, 16TB HDD, and 32TB HDD capacity (same array capacity)

Production-scale | Where architecture determines success

The baseline results so far establish how SSD and HDD architectures behave as capacity increases at smaller scales. We now move to production-level configurations to see how those architectural differences compound when AI infrastructure is deployed in multi-petabyte environments.

This section evaluates storage behavior at true production capacity, comparing the peak throughput, training performance, and power efficiency of a 10x Micron 6600 ION SSD configuration (2,458TB) against a capacity-matched 77x 32TB HDD array (2,464TB). Here, the impact of device count, power efficiency, and throughput becomes decisive, as they directly shape the performance and efficiency of large AI training platforms.

8. This was observed by comparing the peak throughput of the 16x 16TB and 5x 32TB HDD arrays in Figure 2.

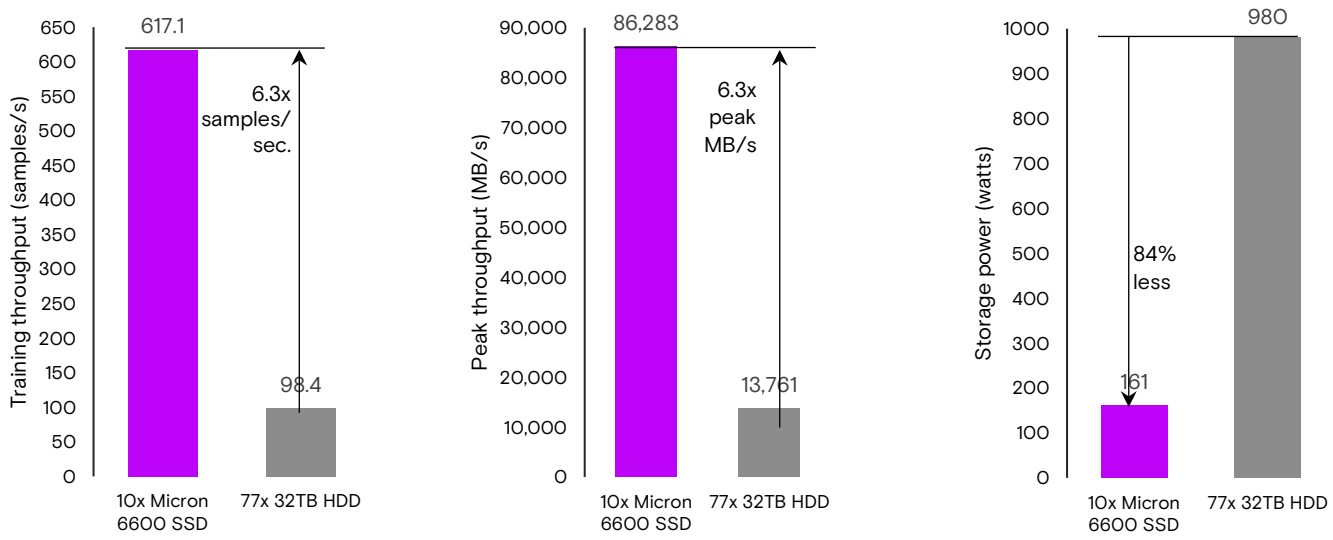

Figure 3: Production scale configuration training throughput, peak throughput, and power use (same

Figure 3 shows that at a production scale, a 10x Micron 6600 ION SSD configuration delivered substantially higher training and peak throughput while consuming far less power than a capacity-matched HDD array, demonstrating that architectural efficiency should be used to determine the right balance of AI platform performance and power use.

The Micron 6600 ION delivers higher power efficiency

If performance and scale define what an AI platform can do, power efficiency determines whether it can be deployed responsibly.

Figure 4 reframes the prior performance and storage power analysis through the lens of storage power efficiency.

It shows that the same architectural efficiencies that enabled the 1x SSD configuration to deliver higher training throughput and peak throughput at dramatically better power efficiency also enabled superior storage power efficiency.

Comparing the 1x SSD to the 16x 16TB HDD showed a difference of (627.5 MB/s per watt / 13.1 MB/s per watt ≈ 47.9x). At production-scale, we saw a difference of (536.6 MB/s pwer watt / 14.0 MB/s per watt ≈ 38.3x).

These differences are not incremental optimizations, but structural efficiency gains. For power-constrained AI environments, storage power efficiency can become a gating factor for sustainable deployment.

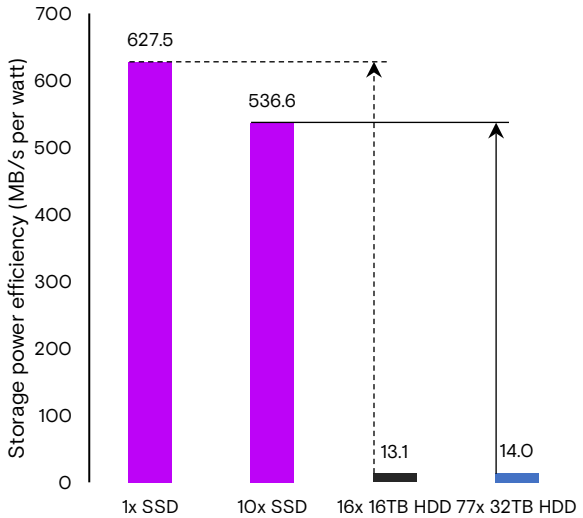

Figure 4: Power efficiency, all configurations

Conclusion

This analysis shows that storage architecture is no longer a secondary design choice; it is a defining factor in AI platform performance, scalability, and sustainability.

Across baseline and production scale configurations tested, the Micron 6600 ION SSD consistently delivered higher training throughput, improved accelerator utilization, dramatically lower device counts, and far superior power efficiency than capacity-matched HDD architectures.

As AI systems scale into the multi-petabyte range, these advantages can compound, making flash-based storage not just faster, but fundamentally better suited to the demands of modern AI.

The conclusion is clear: SSD-based architecture can help provide the performance headroom and power efficiency required to scale AI more sustainably.

How we tested

The MLPerf Storage benchmark results presented in this paper are unofficial and were not peer-reviewed by MLCommons. The testing was completed by Micron in our Longmont, Colorado labs. MLPerf Storage benchmarks are designed to simulate real-world machine learning workloads, consistently stressing the storage system to allow for an accurate assessment of its performance. This design helps ensure reproducible results and focuses on storage performance.

The configuration details for the system under test are in Table 1 for reference.

Storage configuration details

Parameter	SSD configuration	HDD configuration
Drive	Micron 6600 ION 245TB	Seagate EXOS X18 16TB
Manufacturer	Micron	Seagate
Model	6600	ST16000NM000J
Capacity per drive	245TB	16TB
Interface	NVMe PCIe Gen5	SATA
Form factor	U.2	3.5"
Baseline configuration	1x Micron 6600 ION SSD (245TB)	16x 16TB HDD JBOD RAID-0 (256TB)
Production scale configuration	10x Micron 6600 ION SSD (2,450TB)	77x 32TB HDD (2,464TB, calc.)

Table 3: Complete test platform configurations

Test Platform

Parameter	Value
Simulated accelerators	GPU accelerator (as commonly deployed in large-scale model training)
Client memory	256 GB Micron DDR5
Benchmark tool	MLPerf Storage (MLPerf Storage v2.0)
ML model	3D U-Net (3D Medical Image Segmentation)
Server Model	Supermicro AS-1115CS-TNR

Table 4: Test platform details

Test methodology

MLPerf Storage Benchmark

MLPerf Storage v2.0 is an industry-standard benchmark developed by MLCommons to measure storage system performance for machine learning training workloads. Unlike traditional storage benchmarks that measure raw I/O metrics, MLPerf Storage evaluates how well a storage subsystem can feed GPU accelerators with training data in realistic ML training scenarios.

The benchmark simulates the data loading pipeline of popular ML training frameworks, generating I/O patterns that match real-world training workloads. It reports both I/O throughput (MB/s) and training throughput (samples/second), providing a direct measure of how storage performance translates to training efficiency.

3D U-Net Workload

The 3D U-Net workload simulates training a 3D medical image segmentation model, which is characterized by:

- **Large sample sizes:** ≈ 140 MB per training sample
- **Sequential read patterns:** Each sample is read sequentially, but samples are accessed randomly
- **High sustained throughput:** Requires continuous high-bandwidth data delivery
- **I/O sensitivity:** Training throughput is directly limited by storage I/O bandwidth

MLPerf Storage configuration

Parameter	Value
Benchmark	MLPerf Storage v2.0 (mlpstorage)
Model	3D U-Net
Accelerator type	GPU accelerator, similar to those commonly deployed in large-scale model training
Accelerator counts tested	1–40 (SSD), 1–3 (HDD)
Direct I/O	Enabled (reader.odirect=true)
Training files (1-drive SSD)	9,375–17,500
Training files (10-drive SSD)	70,000–140,000
Training files (HDD)	9,375
Epochs	5
Steps per Epoch	500
Run selection	Best-performing run per configuration

Table 5: MLPerf Storage 3D U-Net test parameters

micron.com/6600-ION