Leading AI performance and power efficiency with Micron® NVMe SSDs

AI has a power problem. As AI grows, so does its power use. This presents a challenge that needs innovative solutions.

Data centers used less than 485 terawatt-hours (TWh) in 2025 could climb to 946 TWh by the end of 2030, nearly doubling the power demands of data centers in just five years.[1]

To help tame this, we must expand our thinking about designing and building AI platforms. Of course, performance is vital. And so is power efficiency. We can steadily advance our sustainability goals by using extra care to select the essentials of our AI platforms.

Performance: Any well-constructed test can measure SSD performance, but tests that measure workloads simulating real-world AI use cases are far more valuable. MLPerf Storage™,[2] a standardized benchmarking tool, provides just that—SSD performance results across various, often complex, AI use cases.

Power: SSD power consumption should be measured and compared using the average power consumed during the test (expressed in watts) such that SSD power efficiency (GB/s per watt) can be calculated. This provides a more complete picture of the results and their impact on sustainability.

Micron 7600 SSD (E3.S, E1.S, U.2)

Micron 7600 SSD: The right SSDs for machine learning

Three elements of the MLPerf Storage benchmark were utilized to analyze the performance and power use of the Micron 7600 SSD, along with two other mainstream data center SSDs from competing manufacturers.[3]

ResNet50 | Autonomous driving, personalized shopping[5]

37% Better performance

45% Lower power

146% Better efficiency

CosmoFlow | Understand the Universe

29% Better performance

49% Lower power

155% Better efficiency

3D U-Net | Better diagnostics and treatment planning

Similar performance across all tested SSDs[4]

48% Lower power

84% Better efficiency

micron.com/7600

The Micron 7600 SSD offers best-in-class performance and promotes sustainability. It redefines sustainable storage for mainstream workloads, maximizing power efficiency amidst data growth, performance expectations, and environmental concerns.

1. See this page on technologyreview.com for additional information on data center power consumption.
2. We used the MLPerf Storage benchmark suite to compare the SSD performance, power, and efficiency of the Micron 7600 versus two competitors' products. The results shown were run in Micron's data center workload engineering lab and are not official MLPerf Storage results. See this page to learn more about MLCommons, their charter, and the benchmarks they develop.
3. Performance differences calculated as (Micron 7600 value / competitor value) – 1, expressed as a percentage. Power-related differences are calculated as (competitor value – Micron value) / competitor value, expressed as a percentage.
4. The 3D U-Net test is multi-stream large block reads; all three drives perform well from a throughput standpoint in this workload.

The following figures represent test results using the 3D U-Net, CosmoFlow, and ResNet50 benchmarks with the Micron 7600 SSD and two competitive products. Test configurations used the number of supported (virtual) NVIDIA® H100 GPU accelerators, which is noted in each section. All three benchmarks show SSD throughput (GB/s) and power consumption (in watts for the SSD), enabling calculated power efficiency (expressed in GB/s per watt) for the SSD.

These results may give a better picture than other benchmarks because they measure the maximum AI training throughput storage can provide *while maintaining high accelerator utilization.* Micron 7600 SSD results are purple, while competitors' results are black and grey. The MLPerf Storage benchmark results presented in this paper are unofficial results and were not peer reviewed by MLCommons. SSD power measurements are not part of the official MLPerf Storage benchmark results.

ResNet50 analysis | Autonomous driving, personalized shopping

ResNet50 is a deep learning model (tested with 65 H100 virtual accelerators) designed for image classification tasks, making it helpful in comparing user-uploaded images for personalized shopping recommendations and autonomous driving.[5] ResNet50 results are essential because the benchmark tests storage performance for large-scale image recognition tasks, such as those in the ImageNet competition, while maintaining computational efficiency. This makes it valuable for evaluating hardware performance, such as SSDs, in handling intensive AI workloads like autonomous driving or recommendation systems.[6]

37% faster ResNet50 performance helps enable faster AI model execution across various vision-based applications (as seen in Figure 1). Whether powering recommendation engines or real-time perception in autonomous systems, this performance boost is a critical advantage for scaling AI infrastructure in consumer and industrial domains.

Figure 1: ResNet50 SSD throughput

With 45% lower power consumption in ResNet50 testing seen in Figure 2, the Micron 7600 SSD reduces heat output and cooling demands, which can help cut operational costs and support more sustainable AI deployments. It's a critical advantage for organizations prioritizing both performance and environmental responsibility.

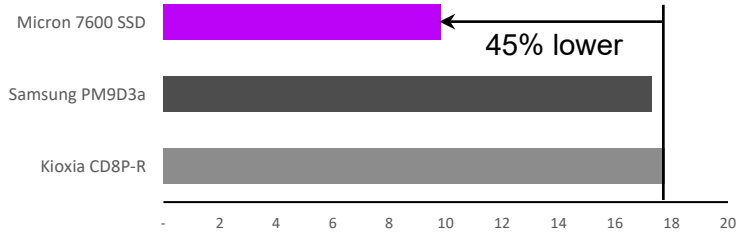

Figure 2: ResNet50 SSD power

The Micron 7600 SSD shows up to 146% better power efficiency, enabling data centers to run more AI inference tasks per watt consumed, as seen in Figure 3.

This reduces energy costs and cooling requirements while maintaining high throughput for vision-based workloads. It's a key enabler for sustainable AI infrastructure across industries.

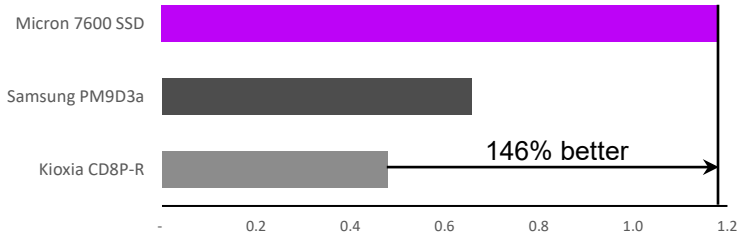

Figure 3: ResNet50 SSD power efficiency

5. See this page on the thinkingstack.ai website and this page on the linkspringer.com webpage for additional information about using ResNet50 in these examples (recommendation systems may use image or pattern matching).
6. See this paper posted to the rjpn.org website and this paper posted to the norma.ncirl.ie website for additional information on using ResNet50 in recommendation systems. See this paper posted to the ijcrt.org website and this blog on the ultralytics.com website for a discussion about using ResNet50 to enable image classification in real-world applications like autonomous systems (such as driving).

CosmoFlow analysis | Understand the Universe

The CosmoFlow benchmark simulates training a deep learning model on 3D cosmological data to predict universe parameters (22 H100 virtual accelerators), such as dark energy and matter density. It is crucial for evaluating the performance of AI systems on large-scale scientific workloads, particularly those involving volumetric data and complex numerical simulations.[7]

By pushing the boundaries of computational tools used in astrophysics, CosmoFlow contributes to broader scientific discovery and enhances our understanding of the universe. The insights gained from CosmoFlow can lead to more accurate models of cosmic evolution, helping scientists unravel the mysteries of the cosmos.[8]

With up to 29% faster performance in CosmoFlow workloads, our SSD enables faster AI model convergence for complex simulations like cosmological parameter estimation. This performance gain enables faster training of cosmological models, helping researchers explore dark matter and energy more efficiently.

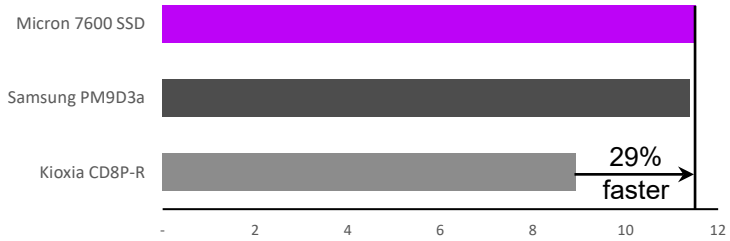

Figure 4: CosmoFlow SSD throughput

CosmoFlow testing also showed that the Micron 7600 SSD demonstrated up to 49% lower power usage than competitors, lowering power demands in AI-driven data centers. This reduction can translate to lower operational costs and improved thermal management, helping support a sustainable scaling of high-performance computing infrastructure for scientific and medical research.

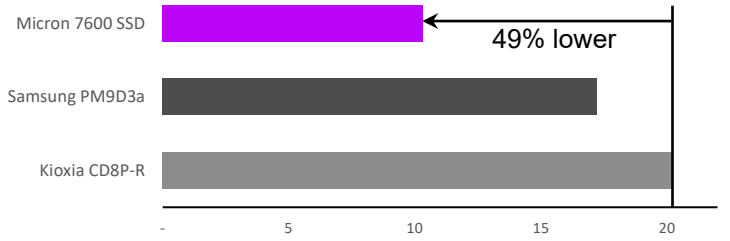

Figure 5: CosmoFlow SSD power

The Micron 7600 SSD delivered 155% better performance efficiency (GB/s per watt), enabling high-throughput AI workload execution with significantly less power. This means faster data processing without the power penalty.

This empowers researchers to run more simulations, faster, within the same power envelope.

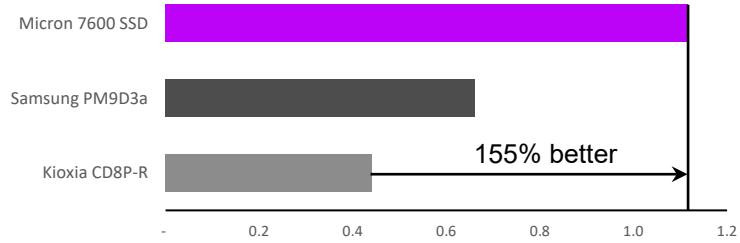

Figure 6: CosmoFlow SSD power efficiency

7. Learn more about the CosmoFlow benchmark from this page on github.com.
8. See this page on arxiv.org to learn how CosmoFlow uses deep learning to further our understanding of the universe at scale.

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

The MLPerf Storager 3D U-Net benchmark evaluates how efficiently AI systems can train 3D medical image segmentation models. These are critical for identifying tumors and other anomalies in volumetric scans like CT or MRI.

Simulating real-world workloads (four H100 virtual accelerators) can help healthcare providers select storage and compute systems that deliver faster, more accurate diagnoses. This enables clinicians to develop effective treatment plans based on precise, AI-assisted imaging insights.[9]

Figure 7 shows that all three SSDs performed similarly in 3D U-Net throughput.

The 3D U-Net workload utilizes Direct I/O, which helps enable the system to get full performance out of NVMe SSDs.[10]

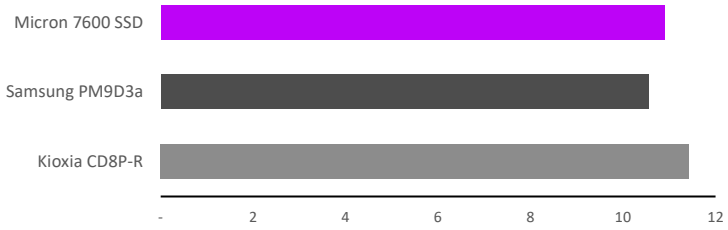

Figure 7: 3D U-Net SSD throughput

Figure 8 shows that the Micron 7600 SSD achieved up to 48% lower SSD power consumption than the competitor's products. This translates to significant power savings in data centers running AI-driven medical diagnostics. The Micron 7600 SSD helps support healthcare innovation and sustainability goals by reducing environmental impact while maintaining high throughput.

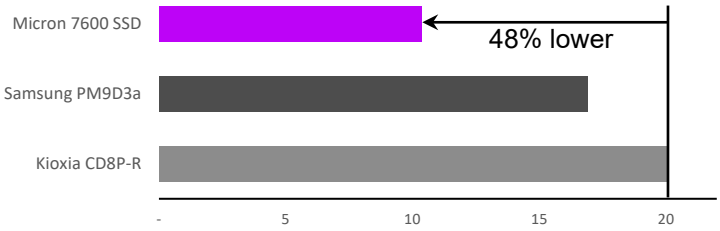

Figure 8: 3D U-Net SSD power

Figure 9 shows that the Micron 7600 SSD enables up to an 84% power efficiency advantage in the 3D U-Net workload. This can help healthcare IT teams scale AI infrastructure without exceeding thermal or energy budgets, for broader deployment of AI-assisted diagnostics across facilities.

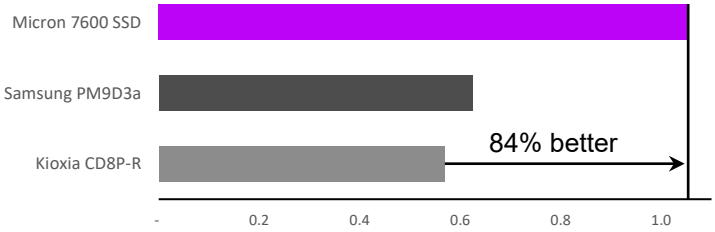

Figure 9: 3D U-Net SSD power efficiency

Conclusion

To help address the growing power consumption challenges of AI, it is crucial to prioritize both performance and power efficiency in AI platform design. We've seen that the Micron 7600 SSD does both helping deliver workload performance and SSD power efficiency.

ResNet50: In this image classification model, the Micron 7600 provided up to 37% better performance using up to 45% less power and delivered as much as 146% better power efficiency, leading to safer autonomous driving and more personalized shopping.

CosmoFlow: Up to 29% better performance, 49% lower power, and 155% better power efficiency for more sustainable cosmological research.

3D U-Net: The Micron 7600 SSD used up to 48% less power and delivered up to 84% better power efficiency in workloads that help identify tumors and other anomalies in volumetric scans for more effective treatment plans.

9. See this page on followinghealthcare.com for additional information on how AI-assisted medical imaging can improve treatment through enhanced diagnostic accuracy, workflow optimization, reducing human error, and enabling broader access.
10. Using Direct I/O bypasses the page cache for READ and WRITE operations, which can reduce latency, enhance consistency, and help optimize large transfers. See this page on springer.com and this page on medium.com for more information. See this page on kernels.org to learn more about Direct I/O. The 3D U-Net test is multi-stream large block reads; all three drives perform well from a throughput standpoint in this workload.

How we tested

The MLPerf Storage benchmark results presented in this paper are unofficial results 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.

Rules for submitting MLPerf Storage results are described in the MLPerf Storage Benchmark Rules page of mlcommons.org. Official peer-reviewed results for the Micron 7600 can be found in this repository on the GitHub/MLCommons page.[11] The configuration details for the system under test are in Table 1 for reference.

	SSD test platform
Server platform	Supermicro® AS-1115CS-TNR
CPU	AMD EPYC™ 9654
Memory	256GB (320GB for ResNet50 with H100)[12]
Server Storage (7.68TB)[13]	Micron 7600 SSD Samsung PM9D3a Kioxia CD8P-R
Boot, Applications Drive	Micron 7450 SSD 960GB
Operating System	Ubuntu 20.04.6 LTS (Focal Fossa)

Table 1: Server configuration

Appendix: Engineer's notes

MLPerf Storage defines three training workloads; but how do we define a training workload? Under the hood, MLPerf Storage uses the DLIO tool developed by Argonne National Labs. DLIO allows us to simulate real AI training workloads by defining a sample size, a container format, a framework, and an emulated accelerator (itself described as a batch size and computation time). It generates a dataset using the actual AI framework (Pytorch, TensorFlow, DALI, and so on). It then executes the data ingest operations in the same way as the actual workload.

MLPerf Storage v2.0 defines the following workloads:

Workload / model	Sample size	Samples per file	Container	Batch size	Framework	H100 seconds per batch
3D U-Net	142 MB	1	Serialized NumPy Arrays (npz)	7	Pytorch	0.3230
ResNet50	128 KB	1,251	TFRecord	400	TensorFlow	0.2240
CosmoFlow	2.8 MB	1	TFRecord	1	TensorFlow	0.0035

While there are only three workloads, and each is an AI training workload, the following make these three workloads significantly different from one another:

- Framework variation
- Number of samples per file
- DirectIO vs. filesystem caching
- Different batch sizes
- Sample size differences
- Time between batches

These workload differences result in different IO patterns, block sizes, and workload intensities (queue depths).

11. See this page on mlcommons.org to learn more about the MLPerf Storage benchmark.
12. Despite the benchmark being designed so that system DRAM does not affect the results, system DRAM is a reporting requirement, as noted in the benchmark-specific sections on this page on github.com. The values used are shown for clarity.
13. The rated capacity and formatted capacity will be less. 1GB = 1 billion bytes.

micron.com/ssd