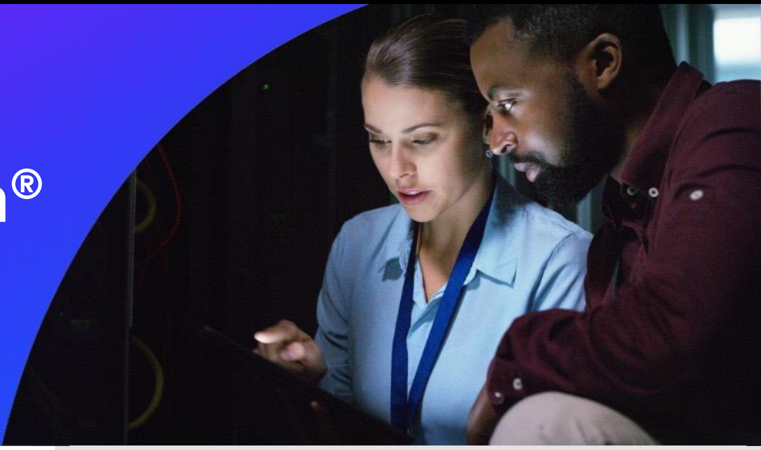
Benchmarking AI performance with Micron® NVMe SSDs



There is a litany of data center SSDs one can choose for storage in machine learning (ML) platforms. Choosing the right SSDs for these platforms is far more complex than simple data sheet specifications and interface rates.

Standard benchmarks like MLPerf™ help provide easy-to-use guidance on SSD selection,[1] evaluate storage performance in AI use cases, and provide direction on the number of AI accelerators one can expect to support with a given SSD.

This document analyzes Micron's MLPerf Storage v1.0 submission and verified results for the performance-focused Micron® 9550 SSD – the first PCIe® Gen5 SSD submission[2] – and the capacity-focused Micron 6500 ION SSD.

About these Micron SSDs

Micron 9550 NVMe™ SSD	
Type	Performance-focused, data center
Capacity	3.2TB to 30.72TB (7.68TB used for testing)
Form factors	U.2 (15mm), E3.S (7.5mm), E1.S (15mm)
The Micron 9550 NVMe™ SSD is a high-performance data center SSD.[3] It is built with industry-leading innovation to deliver superior PCIe Gen5 performance, flexibility, and security for AI and beyond.	

Table 1: Micron 9550 SSD overview

Micron 6500 ION NVMe™ SSD	
Type	Capacity-focused, data center
Capacity	30.72TB
Form factors	U.2 (15mm), E1.L (9.5mm)
Designed to keep pace with the accelerating growth of data, the Micron 6500 ION SSD provides a major advancement for data centers by lowering operating costs and improving storage efficiency.[3]	

Table 2: Micron 6500 ION SSD overview

Micron makes the right SSDs for machine learning

Different ML use cases and deployments require different storage features. The Micron 9550 SSD and 6500 ION SSD are designed to enable choice and flexibility for machine learning.

micron.com/9550

The Micron 9550 NVMe SSD is a high-performance, data center SSD.[3] It is built to manage critical workloads requiring extreme speed, scalability, and power efficiency.

micron.com/6500ION

The Micron 6500 ION SSD redefines scalable storage, maximizing IT budgets amidst data growth, performance expectations, and environmental concerns. It offers best-in-class performance and promotes sustainability.

1. The MLPerf Storage benchmark is from MLCommons, an engineering consortium.
2. The Micron 9550 SSD and the Micron 6500 ION SSD are the first SSDs submitted for v1.0 results.
3. See the Micron 9550 SSD page on micron.com for additional details on this SSD and the Micron 6500 ION page on micron.com for additional details on this SSD.

The following figures represent test results using the Unet3D, CosmoFlow, and ResNet50 benchmarks (organized by benchmark) using NVIDIA® A100 and H100 Tensor Core GPU accelerators. MLPerf Storage differs from other benchmarks in that it measures the maximum AI training throughput that storge can provide *while maintaining high accelerator utilization.* Accelerators are shown on the vertical axis, while the number of accelerators supported is shown on the horizontal axis. Micron 9550 SSD results are shown in purple, while Micron 6500 ION SSD results are shown in blue.

Supporting more accelerators with a specific SSD reflects that SSD's higher throughput (relative to an SSD that supports fewer accelerators) and its ability to reduce bottlenecks and return results faster. This can be critical in several real-world use cases:

- **Deep learning and AI training:** High-performance SSDs can feed data to multiple accelerators simultaneously, improving the training of complex neural networks used in applications like natural language processing, image recognition, and autonomous driving.[4]
- **Scientific research:** In fields such as genomics, climate modeling, and astrophysics, large datasets need to be processed quickly. Efficient data transfer between SSDs and accelerators can significantly speed up simulations and data analysis.[5]
- **Financial modeling:** High-frequency trading and risk assessment models require rapid data processing. Enhanced SSD-to-accelerator ratios help ensure that these models can run in real time, providing timely insights and decisions.[6]
- **Media and entertainment:** Rendering high-resolution graphics and video editing demands substantial data throughput. Multiple accelerator support can help handle these tasks more efficiently, reducing rendering times and improving productivity.[7]

Unet3D analysis

The Micron 9550 SSD supports up to eight A100 accelerators while Micron 6500 ION SSD results show it supports up to three. With H100 accelerators, the Micron 9550 SSD results show support for up to four while the Micron 6500 ION SSD supports one, as seen in Figure 1.

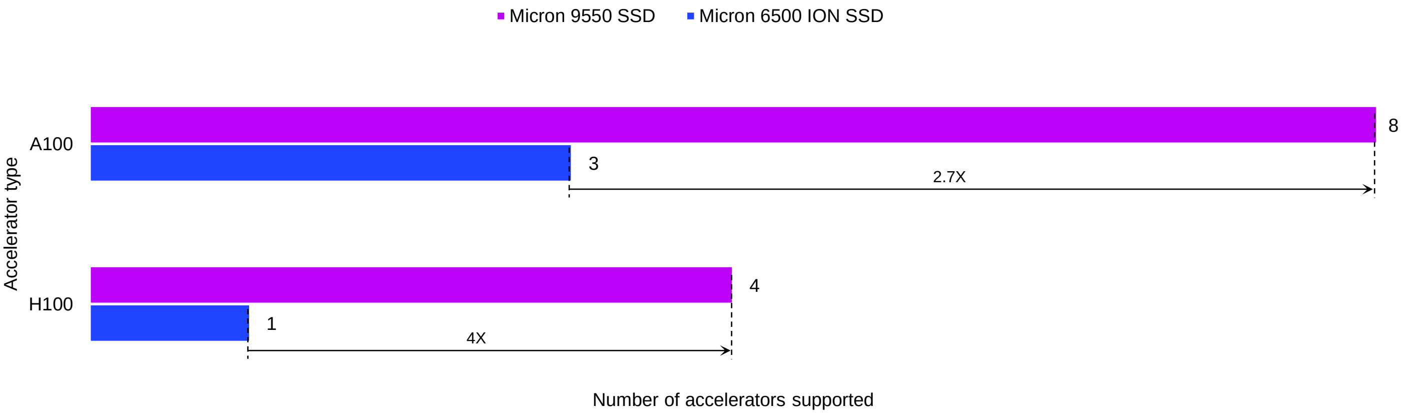


Figure 1: Unet3D results

[4] See this page on the an NVIDA developer website to learn more about how storage performance can affect deep learning.
[5] See this page on dataversity.net for additional information on the effect of fast SSDs impacts deep learning.
[6] This page on kinetics.com discusses how GPU acceleration can affect financial analytics.
[7] This page on howtogeek.com discusses accelerator support requirements in image processing and offers an example of real-world training based on high-resolution media.

How to use these Unet3D results

Unet3D benchmark results are useful because they help characterize the complex task of 3D image segmentation, which is crucial in medical imaging for accurately identifying and delineating structures within volumetric data. By providing standardized performance metrics, Unet3D helps researchers and developers optimize their algorithms and hardware, helping lead to more effective and efficient healthcare solutions.[4]

When comparing Unet3D results, it is evident that the Micron 9550 SSD can support 2.7 to 4 times as many accelerators, making the Micron 9550 SSD particularly well-suited for high-demand AI applications. The Unet3D benchmark results benefit significantly from the increased parallel processing capabilities of the Micron 9550 SSD. This allows for faster training times and more efficient handling of large datasets, making it ideal for large-scale ML tasks, real-time data processing, and complex simulations where high throughput is critical. The Micron 6500 ION SSD, while supporting fewer accelerators, excels in scenarios where SSD capacity is paramount. The Unet3D benchmark still demonstrates the Micron 6500 SSD's effectiveness in smaller-scale AI projects.

Unet3D SSD details: Figures 2 and 3 below represent throughput and sample rate results for each SSD training Unet3D.

The accelerators used (A100 and H100) are shown on the vertical axis of each figure. Throughput is shown on the horizontal axis in Figure 2, and samples per second are shown on the horizontal axis in Figure 3.

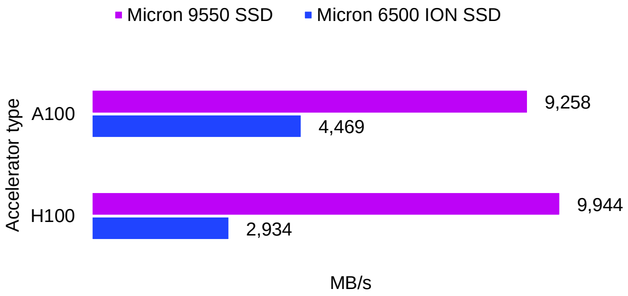


Figure 2: Unet3D SSD throughput

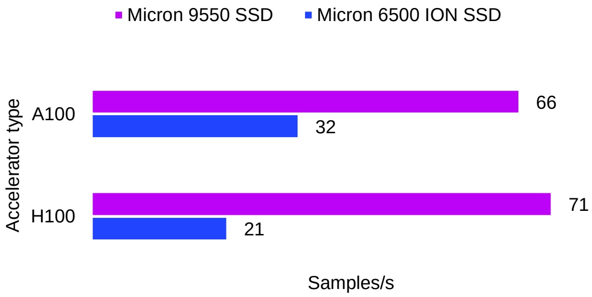


Figure 3: Unet3D sample rates

CosmoFlow analysis

CosmoFlow results are also affected by the performance characteristics of SSDs. Differences between the Micron 9550 SSD and the Micron 6500 ION SSD in the CosmoFlow workload can be seen in Figure 4.

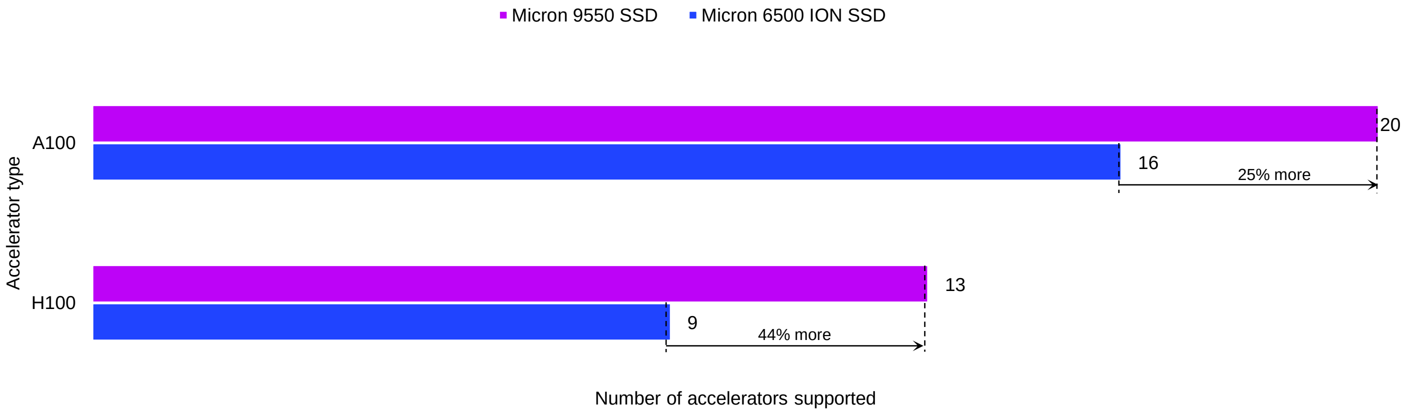


Figure 4: CosmoFlow results

[8] See this page on the arxiv.org website for additional information on an example use case.

How to use these CosmoFlow results

The CosmoFlow test is a highly scalable deep learning application designed to predict cosmological parameters from 3D simulations of the universe. It uses efficient 3D convolution and pooling techniques on the TensorFlow framework to manage the heavy computational load.[9] CosmoFlow offers valuable insight as it provides standardized performance metrics for machine learning tasks in cosmological simulations. Its results can help researchers and developers optimize algorithms and hardware, helping ensure robust and efficient solutions for analyzing large-scale structures of the universe.

When comparing CosmoFlow benchmark results, it is again clear that the Micron 9550 SSD supports more accelerators than does the Micron 6500 ION SSD – 25% more A100 accelerators and 44% more H100 accelerators, both of which position the Micron 9550 SSD as a powerhouse for high-performance AI applications.

The CosmoFlow benchmark, which evaluates the performance of deep learning models in cosmology simulations, benefits from the performance of the Micron 9550 SSD, helping enable faster and more efficient processing of complex simulations making the Micron 9550 SSD ideal for research-intensive environments, large-scale scientific computations, and real-time data analysis. Conversely, the Micron 6500 ION SSD – although supporting fewer accelerators – helps manage deployments where cost-efficiency and per-SSD capacity is paramount.

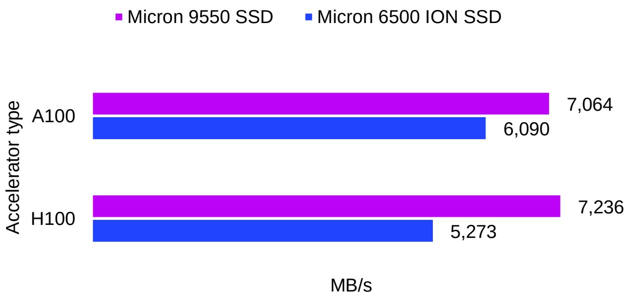


Figure 5: CosmoFlow SSD throughput

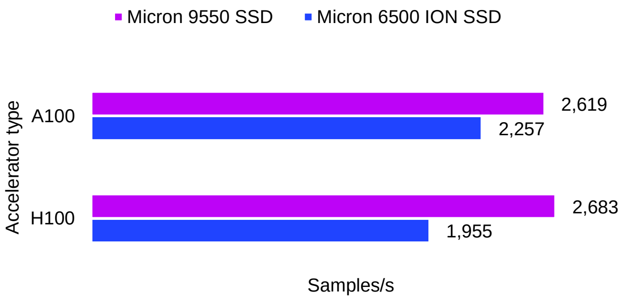


Figure 6: CosmoFlow sample rate

CosmoFlow SSD details

The figures below represent throughput and sample rate results for the CosmoFlow benchmark. A100 and H100 accelerators are again shown on the vertical axis, while the throughput is shown on the horizontal axis in Figure 5 and samples per second on the horizontal axis in Figure 6.

ResNet50 analysis

ResNet50 results are influenced by SSD performance factors like those affecting CosmoFlow results. However, higher-performing SSDs still help ensure that the GPUs are continuously supplied with data, maximizing the ResNet50 results.

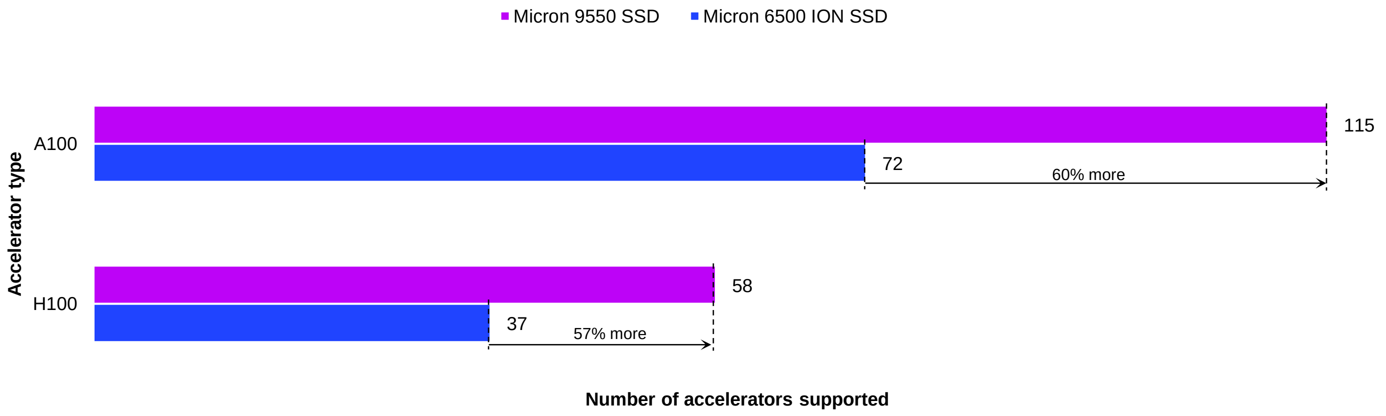


Figure 7: ResNet50 results

[9] Learn more about the CosmoFlow benchmark from this page on Github.

How to use these ResNet50 results

ResNet50 is a deep learning model designed for image classification tasks, known for its ability to manage very deep neural networks through residual learning. ResNet50 results are important because the benchmark achieves high accuracy on large-scale image recognition tasks, such as those in the ImageNet competition, while maintaining computational efficiency.[10] This makes it valuable for evaluating the performance of hardware like SSDs in handling intensive AI workloads and gauging SSD-to-accelerator ratios.[11]

The benchmark achieves high accuracy on large-scale image recognition tasks, such as those in the ImageNet competition, while maintaining computational efficiency.

ResNet50 benchmark results show that the Micron 9550 SSD supports 115 A100 accelerators and 58 H100 accelerators. The Micron 6500 SSD, while supporting fewer accelerators, excels in deployments where cost efficiency and per-SSD capacity is paramount. The ResNet50 benchmark highlights the Micron 6500 ION SSD's effectiveness in smaller-scale AI projects.

ResNet50 SSD details

The figures below represent throughput and sample rates for the ResNet50 benchmark. A100 and H100 accelerators are shown on the vertical axis, while throughput is shown on the horizontal axis in Figure 8 and samples per second on the horizontal axis in Figure 9.

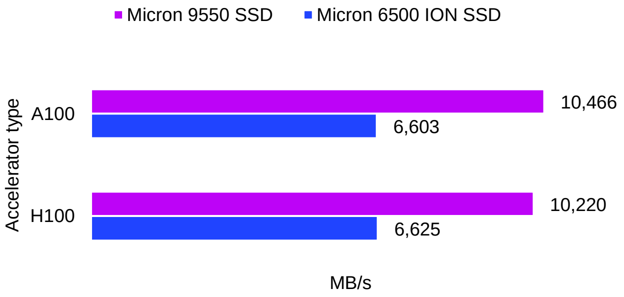


Figure 8: ResNet50 SSD throughput

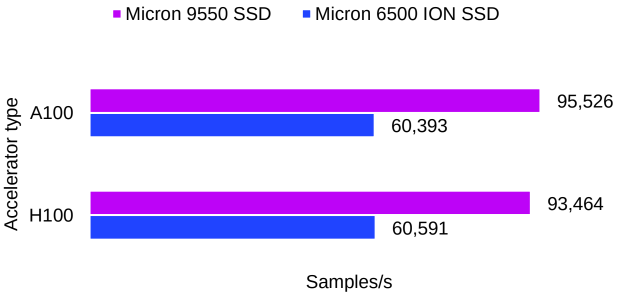


Figure 9: ResNet50 sample rate

Conclusion

Choosing the right SSD for AI workloads involves more than simple performance claims and capacity. One must look beyond these traditional, standard values and instead focus on relevant, specific benchmark results like those in the MLPerf benchmark suite.

The Micron 9550 NVMe SSD offers superior performance, especially with H100 accelerators, while the Micron 6500 ION SSD may provide a more cost-effective solution for certain use cases.

Visit the data center SSD page on micron.com now to get started.

[10] See this page on the pytorch.org website for additional information about ResNet50.
[11] See this page on the debuggercafe.com website for addition information on using ResNet50 and image classification.

How we tested

MLPerf Storage benchmark results should be viewed in a different context than many other benchmark results because MLCommons results are peer-reviewed before publication. According to MLCommons, the organization behind the MLPerf benchmark, the latest benchmark round received submissions from multiple organizations and released over 1,800 peer-reviewed performance results for machine learning systems.[12]

This peer-review process helps ensure the credibility and reliability of the results, making them a trustworthy source. Additionally, using MLPerf Storage benchmark results is governed by the MLPerf Storage V1.0 Benchmark Rules (as found on this page on github.com).

MLPerf Storage benchmarks are designed such that specific system configurations used during testing do not significantly impact the benchmark outcomes.[13] This is because the MLPerf Storage benchmark is designed to be hardware-agnostic, providing standardized tests and metrics that ensure comparability across diverse hardware setups. Because this benchmark simulates real-world machine learning workloads, the benchmark consistently stresses the storage system, allowing for an accurate assessment of its performance. This design helps ensure that the results are reproducible on any compliant system, thereby isolating the storage performance as the primary focus of evaluation.[14,15]

Rules for submitting MLPerf Storage results are described in MLPerf Storage V1.0 Benchmark Rules page of mlcommons.org. The configuration details in Table 3 are supplied for reference.

	Micron 9550 SSD platform	Micron 6500 ION SSD platform
Server platform	Supermicro® AS-1115CS-TNR	Dell Technologies® PowerEdge® R7525
CPU	AMD EPYC™ 9654	2x AMD EPYC™ 7713
Memory	256GB (320GB for ResNet50 with A100)[16]	256GB
Server Storage	Micron 9550 SSD 7.68TB	Micron 6500 ION SSD 30.72TB
Boot, Applications Drive	Micron 7450 SSD 960GB	Dell ExpressFlash SSD
Operating System	Ubuntu 20.04.6 LTS (Focal Fossa)	Alma Linux 9.3

Table 3: Server configurations

[12] See this page on mlcommons.org for additional details.
[13] See this page on github.com to learn more about the benchmark's design.
[14] Additional background on platform-agnostic results can be found on this page on mlcommons.org.
[15] See this readme page on mlcommons.org to learn more about reproducibility.
[16] Despite the benchmark being designed such that system DRAM used does not affect the results, system DRAM is a reporting requirement as noted on this the benchmark-specific sections seen on this page on github.com. The values used are shown for clarity.

Appendix: Engineer’s notes

MLPerf Storage defines 3 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 defined as a batch size and computation time). It does this by using the actual AI framework (Pytorch, TensorFlow, DALI, etc.) to generate a dataset. Then it executes the data ingest operations in the same way as the actual workload.

MLPerf Storage v1.0 has defined the following workloads:

Workload / model	Sample size	Samples per file	Container	Batch size	Framework	H100 seconds per batch
Unet3D	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 other:

- Framework variation
- Containerization
- Sample size differences
- Number of samples per file
- Different batch sizes
- Time between batches

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

micron.com/ssd

Rev. A 09/2024 CCM004-676576390-11767