Micron® 6550 ION NVMe™ SSD performance with NVIDIA® Magnum IO™ GPUDirect® Storage

Imagine a data superhighway built for speed and efficiency — a route for direct memory access that connects GPUs directly to NVMe™ SSD storage. That's what NVIDIA® Magnum IO™ GPUDirect® Storage (GDS) is in the world of artificial intelligence (AI).

Part of the powerful framework of NVIDIA's Magnum IO™ acceleration technologies, GDS is a game-changer, especially when combined with the Micron® 6550 ION NVMe SSD.[1]

The result? A dramatic surge in bandwidth combined with a significant reduction in power used across different training workloads.[2]

This document analyzes the performance (in GB/s) and power efficiency (in GB/s per watt) of the Micron 6550 ION SSD compared to a competitor's. During all testing, the Micron 6550 ION SSD's maximum power consumption was limited to just 20 watts; the SSD was used in power state one (PS1), while the competitor could use its maximum power consumption of 25 watts with power state 0 (PS0).[3]

Test results show the Micron 6550 ION SSD enables greater storage performance and power efficiency for AI workloads across four transfer sizes: 4KB, 16KB, 128KB, and 1MB. Both SSDs offer an advertised capacity of 61.44TB.[4]

Note that AI training is more complex than simply large-block, random IO. Micron engineering traces of AI workloads show significant 4KB transfers depending on the type of files used for storing samples. Additional detail is available in the Future of Memory and Storage 2024 presentation, MLPerf Storage - Enabling easy Storage for AI benchmarking, by Wes Vaske.

Key findings

As AI models have grown rapidly in size and complexity, managing their power consumption has become increasingly difficult. Performance and power efficiency have become paramount in AI use cases.[5]

The Micron 6550 ION SSD enabled higher performance and better power efficiency at every tested transfer size — all while running in a lower power state.

20% **Lower power**

Because the Micron 6550 ION SSD was configured to use power state 1 (PS1, 20-watt maximum), its maximum power consumption limit for all tests was 20% lower than the maximum power draw for the competitor (PS0, 25-watt maximum).

By reducing power consumption in the SSDs, additional power budget is freed for the rest of the system configuration.

104% **Better power efficiency**

One way to help reduce power consumption is to build more power-efficient hardware, like the Micron 6550 ION SSD. During testing, the Micron 6550 ION SSD demonstrated up to 104% better power efficiency (measured in GB/s per watt).

147% **Higher performance**

Head-to-head testing demonstrated the Micron 6550 ION SSD delivered up to 147% better SSD performance (GB/s).

micron.com/6550ION

1. See the NVIDIA GPUDirect Storage Overview Guide for additional information on GDS.
2. Micron internal engineering analysis of AI training workloads shows that different IO sizes are seen depending on model and data formats. Therefore, this document demonstrates small (4KB), medium (16KB and 128KB), and large (1MB) transfer size results with each SSD.
3. NVMe power states limit the maximum power an SSD can consume (actual power consumption depends on many factors). To learn more about NVMe power states, see this page on nvmexpress.org.
4. Unformatted. 1 GB = 1 billion bytes. Formatted capacity is less.
5. To learn more about the growth of power requirements in AI use cases and data centers, see this document by the US Department of Energy.

GDS – a direct path between GPUs and NVMe SSDs[6]

GPUDirect Storage creates a direct data path between local or remote storage, such as NVMe (or NVMe over Fabrics [NVMe-oF]; this discussion focuses on local NVMe) and GPU memory. By enabling a direct-memory access (DMA) engine near the network adapter or storage, data is moved into or out of GPU memory — without burdening the CPU (the control path still relies on the CPU). This direct IO path is represented in Figure 2.

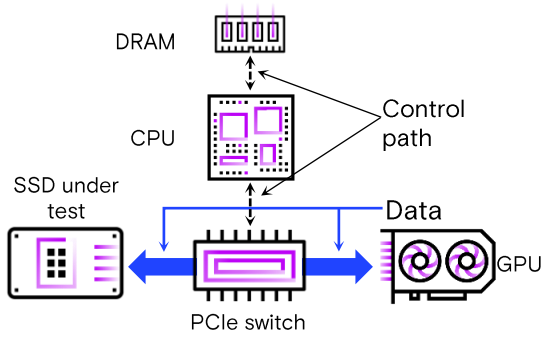

Figure 2: GDS IO path

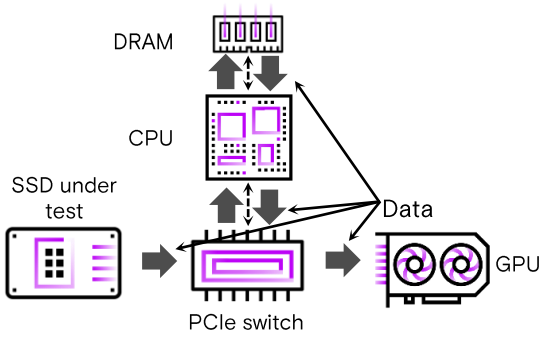

Figure 3: Legacy IO path

This more direct data path between storage and GPU memory is designed to avoid extra copies through a bounce buffer in the CPU's memory required by a legacy IO path represented in Figure 3. All tests use the GDSIO tool to measure results using the GDS IO path.[7]

4KB transfers: Up to 147% higher performance and up to 104% better power efficiency

With over 16,000 cores, the NVIDIA H100 Tensor Core GPU drives high parallelism in AI workloads. In testing GDS on an AI system, the number of GDSIO workers was increased until reaching maximum SSD performance. Figure 4 represents performance results (in GB/s) with a 4KB workload. The Micron 6550 ION SSD is shown in purple while the competitor is shown in gray.

Each SSD shows maximum performance at 512 GDSIO workers, with the maximum performance difference at 256 GDSIO workers.[8] The Micron 6550 ION SSD shows 147% higher performance than the competitor at 256 GDSIO workers. Figure 5 represents power efficiency in GB/s per watt, with efficiency data at 256 GDSIO workers highlighted. Here, the Micron 6550 ION shows 104% higher power efficiency than the competitor.[9]

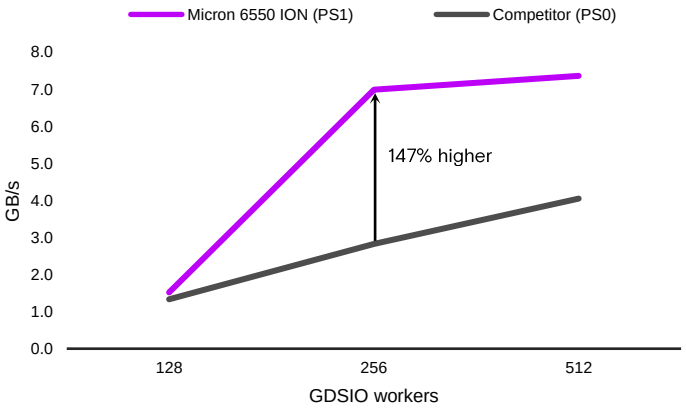

Figure 4: 4KB transfer performance (GB/s, higher is better)

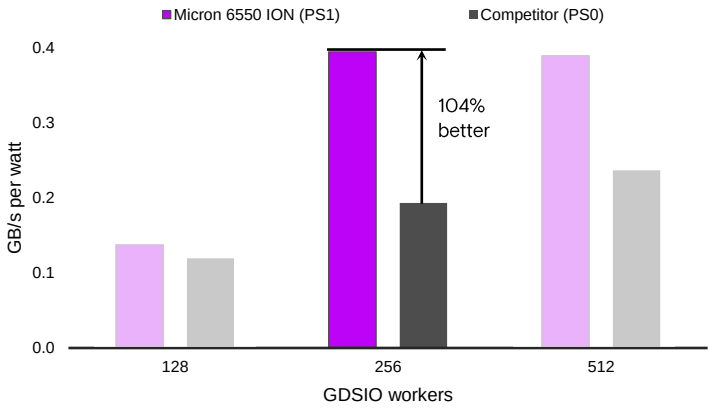

Figure 5: 4KB transfer power efficiency (GB/s per watt, higher is better)

6. See https://developer.nvidia.com/gpudirect-storage for additional details on the IO path differences.
7. GDSIO is a tool that simulates increasing IO parallelism and bandwidth requirements of a GPU. See NVIDIA GPUDirect Storage Benchmarking and Configuration Guide for additional information on this tool.
8. Performance and power efficiency differences are calculated as (higher value / lower value) – 1, expressed as a percentage.
9. Power data for 256 GDSIO workers is highlighted in Figure 5. In each of the subsequent figures, the highlighted value is the number of GDSIO workers at best results seen.

16KB transfers: Up to 30% higher performance and up to 31% better power efficiency

Figure 6 represents performance results with a 16KB workload. The Micron 6550 ION SSD is shown in purple while the competitor is shown in gray. The Micron 6550 ION SSD shows maximum performance at 512 GDSIO workers, where it shows 30% higher performance than the competitor.

Figure 7 represents power efficiency (in GB/s per watt) with power efficiency data for 512 GDSIO workers highlighted. The Micron 6550 ION shows 31% better power efficiency than the competitor.

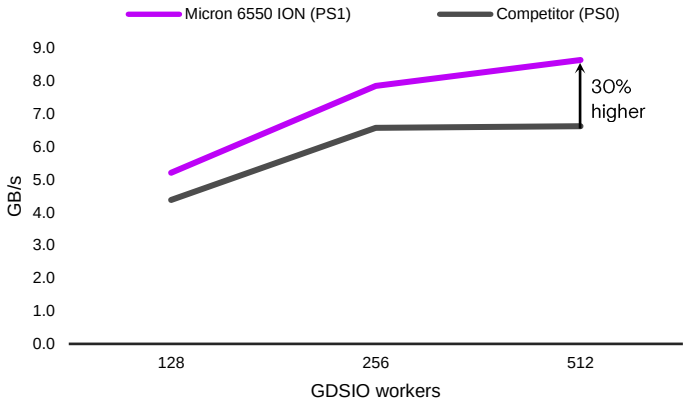

Figure 6: 16KB transfer performance (GB/s, higher is better)

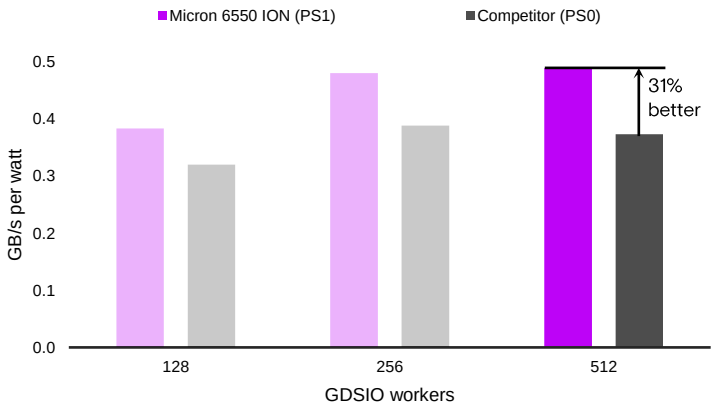

Figure 7: 16KB transfer power efficiency (GB/s per watt, higher is better)

128KB transfers: Up to 42% higher performance and up to 36% better power efficiency

Figure 8 represents performance results with a 128KB workload. The Micron 6550 ION SSD and the competitor are shown in the same colors as earlier. These SSDs show maximum performance at 128 GDSIO workers. The Micron 6550 ION SSD shows 42% higher performance than the competitor.

Figure 9 represents power efficiency (in GB/s per watt) with maximum power efficiency (at 64 GDSIO workers) highlighted. The Micron 6550 ION shows 36% better power efficiency than the competitor.

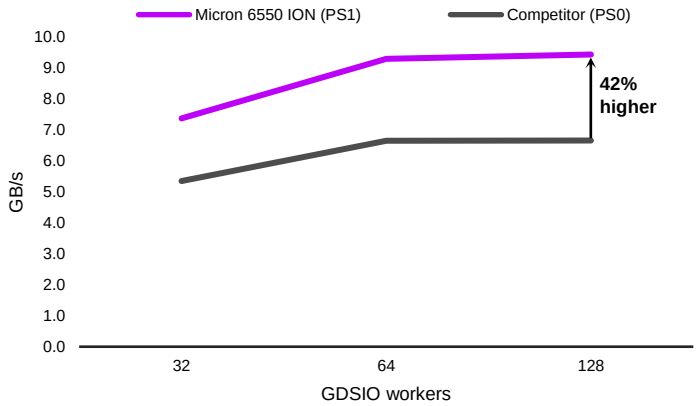

Figure 8: 128KB transfer performance (GB/s, higher is better)

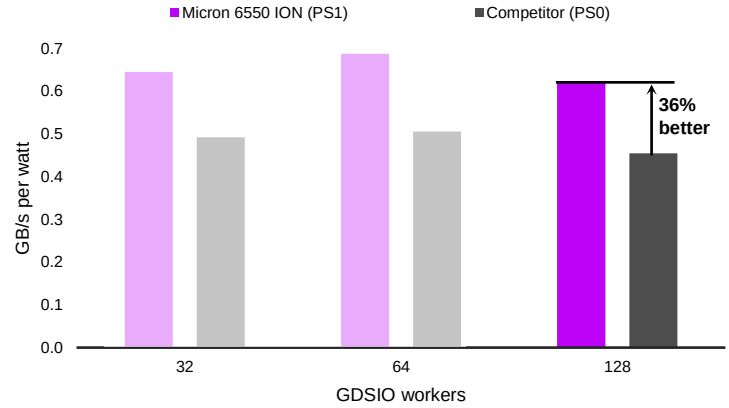

Figure 9: 128KB transfer power efficiency (GB/s per watt, higher is better)

1MB transfers: Up to 26% higher performance and up to 16% better power efficiency

Figure 10 represents performance results in GB/s with a 1MB workload. The Micron 6550 ION SSD and the competitor are shown in the same colors as earlier. Each SSD shows maximum performance at 32 GDSIO workers. The Micron 6550 ION SSD shows 26% higher performance than the competitor.

Figure 11 represents power efficiency (in GB/s per watt) with maximum power efficiency (at 16 GDSIO workers) highlighted. The Micron 6550 ION shows 16% better power efficiency than the competitor.

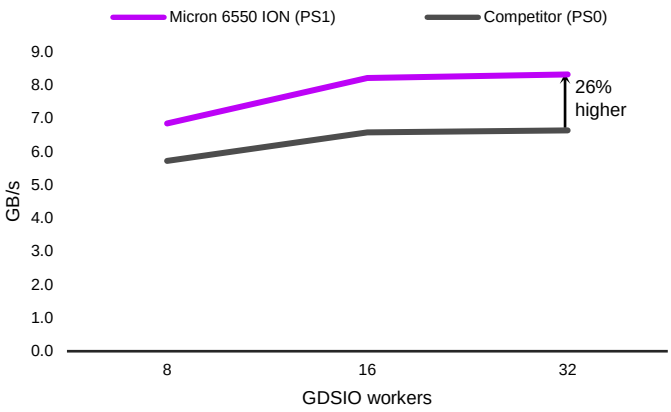

Figure 10: 1MB transfer performance (GB/s, higher is better)

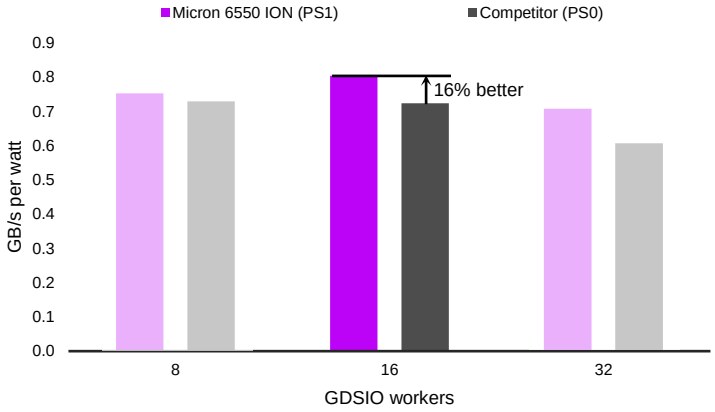

Figure 11: 1MB transfer power efficiency (GB/s per watt, higher is better)

Conclusion

Using the Micron 6550 ION NVMe SSD with GDS shows that the SSD offers significant performance and efficiency advantages compared to the competitor. The Micron 6550 ION SSD demonstrated up to 104% better power efficiency and up to 147% better performance.

SSD performance and power efficiency are crucial in AI systems:

- AI workloads can require substantial computational power and data processing, which can lead to significant energy consumption. Efficient SSDs help reduce this energy consumption, leading to lower operational costs and a smaller carbon footprint, aligning with sustainability goals and regulatory requirements.
- Higher SSD performance can enhance the overall performance of AI systems, while superior power efficiency can help reduce their energy consumption. This, in turn, helps lower both power and cooling costs.
- Power-efficient SSDs can help extend AI infrastructure lifespan, minimizing wear and tear caused by excessive heat.

To learn more about the Micron 6550 ION SSD and the complete line of data center SSDs, visit micron.com/ssd.

How we tested

Tables 1 and 2 outline the test system and software configurations used.

Supermicro server with NVIDIA H100 configuration	
Server	Supermicro® SYS-521GE-TNRT
CPU	2X Intel® Xeon® Platinum 8568Y+ 48-core processors
Memory	16X Micron 96GB DDR4 RDIMMs @ 5600MT/s (1.5TB total memory)
Network	1X NVIDIA® H100-NVL 96GB
Micron SSDs	Micron 61.44TB 6550 ION NVMe SSD
Competitor's SSD	61.44 TB capacity-focused, data center SSD

Table 1: Hardware configuration

Software configuration	
OS	Ubuntu 20.04.6 LTS
Kernel	5.4.0-182-generic
CUDA	12.4
NVIDIA Mellanox OFED version	24.04-0.7.0
NVIDIA GDSIO tool	1.11
Filesystem	XFS
Data Layout	48GB file per GDSIO worker and 1,024 total files equates to 48TB used

Table 2: Software configuration