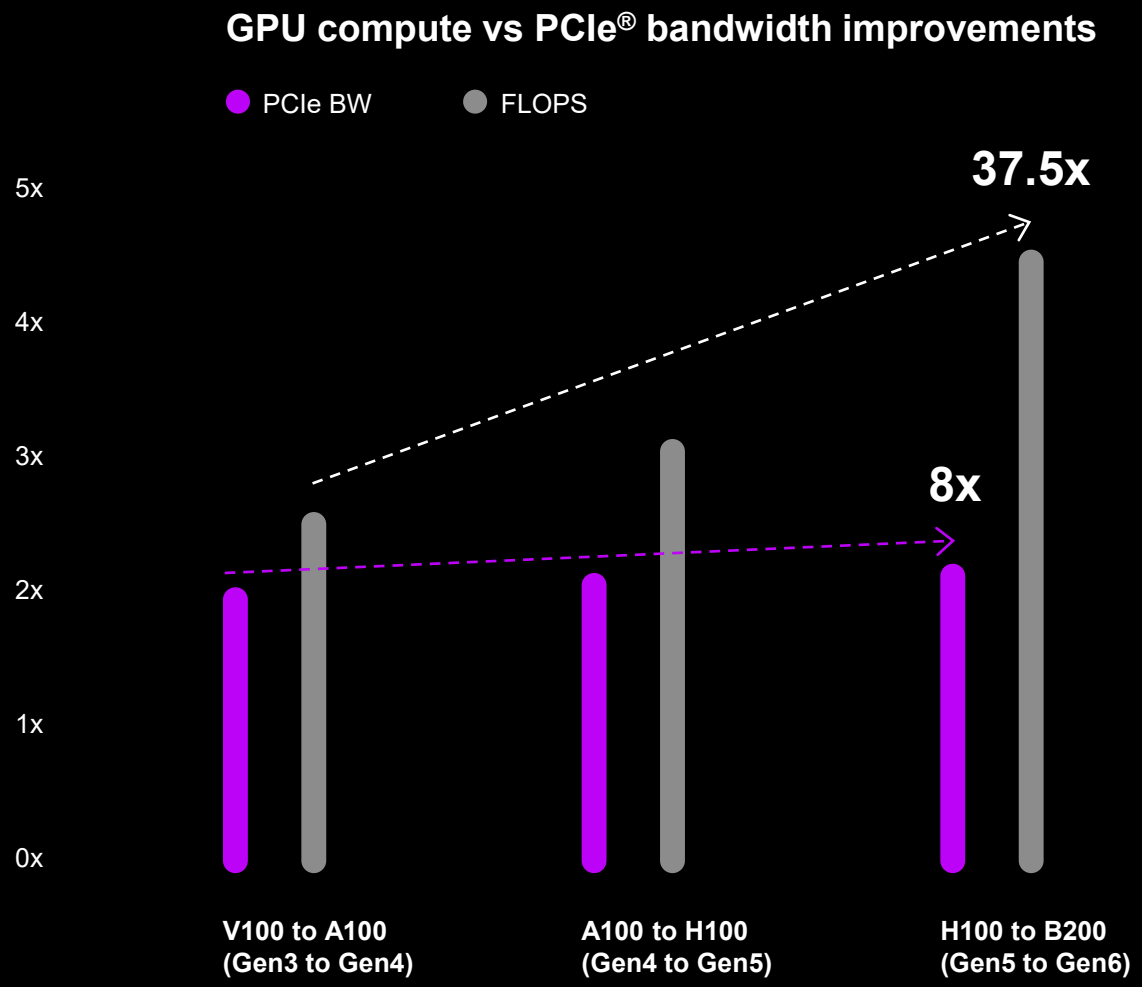
FMS 2025: Real-world AI workloads need fast, efficient storage

Ryan Meredith – Director, Data Center Workload Engineering – Micron

micron FMS

Fast storage is becoming critical for AI training and inference



What is driving increased storage and memory demands?

- Compute is increasing faster than PCIe bandwidth
  - PCIe Gen3 to Gen6 bandwidth increased by 8x
  - In the same timeframe GPU compute increased by 37.5x
- This is putting pressure on memory and storage in AI systems.

Micron data center NVMe™ SSD portfolio

	High Performance Micron 9000 Series	**Mainstream** Micron 7000 Series	**High Capacity** Micron 6000 Series
	• Designed for AI and performance-critical mixed random workloads	• Designed for the broadest range of applications and workloads	• Hyper dense capacities • TLC and QLC technology options
Capacity	**up to 30.72TB**	**up to 15.36TB**	**up to 245TB**
Interface	**PCIe® Gen6**	**PCIe Gen5**	**PCIe Gen5**

High Performance

GPU as a storage initiator

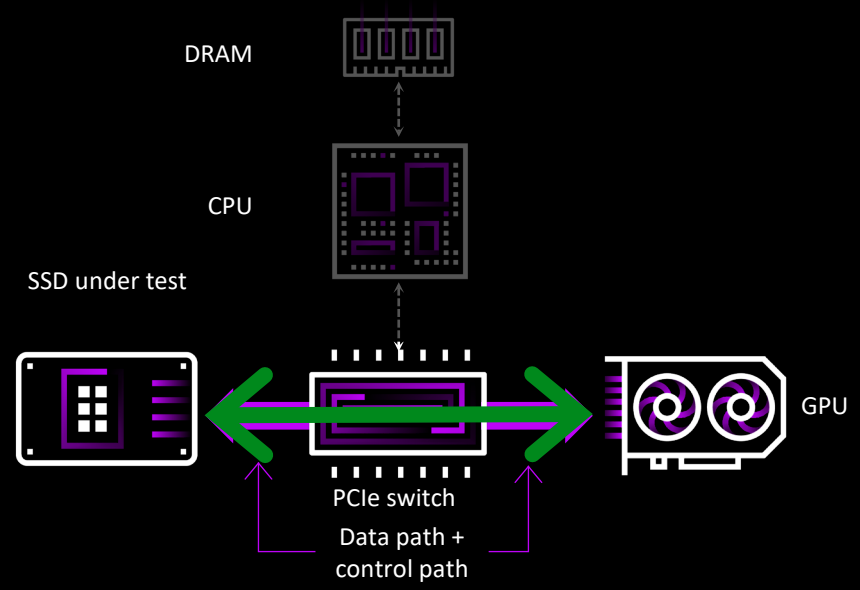
A GPU can drive high levels of IO traffic

- Max IOPs on 1 CPU core is ~1M
- 100M IOPs = 100 cores, just for IO
- AI accelerators have tens of thousands of cores and can use them for massively parallel IO
- Faster SSDs are needed to keep up with AI workload demands

GPU Initiated Storage

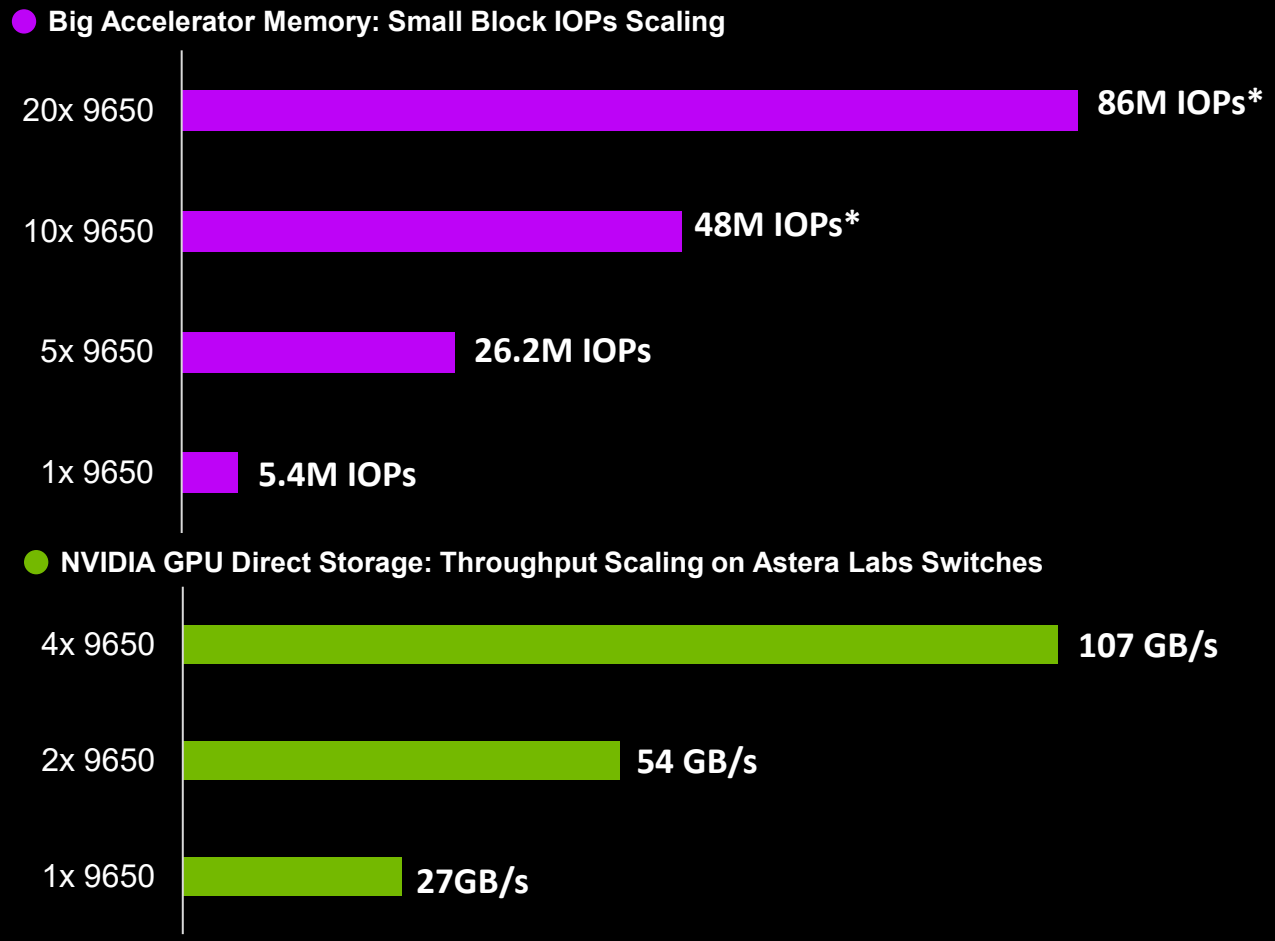
New storage software is required to avoid system bottlenecks

- **NVIDIA GDS:** Data flows direct to GPU, control plane travels over CPU+DRAM
- **Big Accelerator Memory (BAM):** Data and control plane traffic flow directly to GPU over PCIe switch complex
- **NVIDIA SCADA:** Like BAM, plus client server architecture and advanced features



Micron 9650 | GPU initiated storage needs PCIe Gen6

PCIe Gen6 scaling on H100 and Micron 9650 SSDs (Preliminary Results)



Big Accelerator Memory on a Gen6 capable server

- H3 Platform System:
  - Intel 8568Y+, 512GB DDR5
  - 3x Broadcom 144 lane PCIe Gen6 Switches
  - 20x Micron 9650 Gen6 NVMe SSD, E1.S 7.68TB
  - H100 NVL 96GB HBM3

* Preliminary Results: Scaling up to 86M IOPs.

* Hardware & software stack tuning for Gen6 ongoing, we expect 100M+ IOPs

NVIDIA GDS on Astera Labs PCIe Gen6 Switches

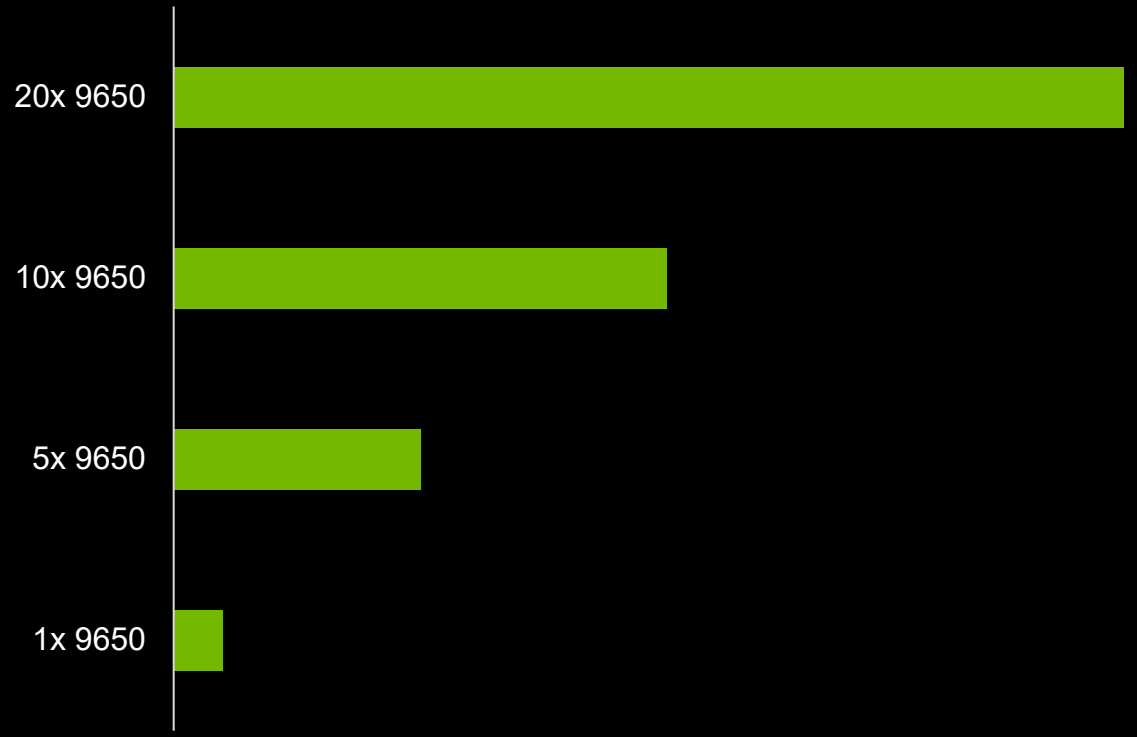
- Linear bandwidth scaling to 4x 9650 using 2x H200 AI Accelerators
- Check out the demo in our booth.

Performance results using Big Accelerator Memory (BAM) and GPU Initiated Direct Storage (GIDS) for NVME block bench workloads..
System under test: H3 Platform PoC system, 1x Intel 8568Y+, 512GB DDR5, 3x Broadcom A0 PCIe Gen6 switches, 20x Micron 9650 E1.S 7.68TB, NVIDIA H100NVL-96GB PCIe Gen5x16
BaM performance testing completed by Micron's Data Center Workload Engineering team in the Longmont, CO lab.
GDS testing completed by the Astera Labs performance engineering team in San Jose, CA.

Micron 9650 | NVIDIA SCADA

Early SCADA code shows strong performance and linear scaling on H3 Platform (Preliminary Results)

- **NVIDIA SCADA (Preview Software)**



Early NVIDIA SCADA code drives impressive small block random read IOPs through 20 Micron 9650 Gen6 NVMe SSDs

- **Linear performance scaling from 1 to 20 drives**
- H3 Platform System:
  - Intel 8568Y+, 512GB DDR5
  - 3x Broadcom 144 lane PCIe Gen6 Switches
  - 20x Micron 9650 Gen6 NVMe SSD, E1.S 7.68TB
  - H100 NVL 96GB HBM3

* Preliminary Results:

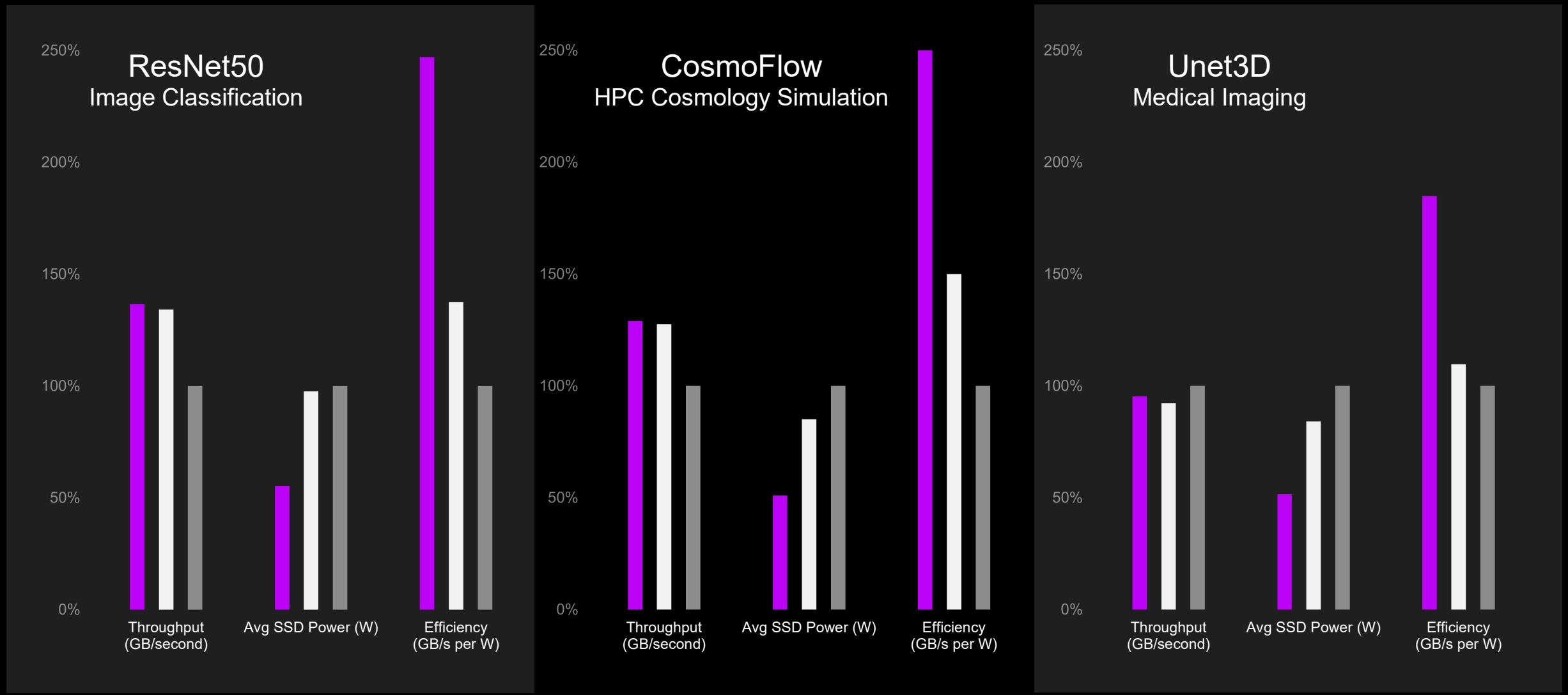
  * Hardware & software stack tuning for ongoing

SCADA test results collected on pre-production code.
System under test: H3 Platform PoC system, 1x Intel 8568Y+, 512GB DDR5, 3x Broadcom A0 PCIe Gen6 switches, 20x Micron 9650 E1.S 7.68TB, NVIDIA H100NVL-96GB PCIe Gen5x16,
Workload is 512B random read initiated from H100 GPU.
Performance testing completed by Micron's Data Center Workload Engineering team.

Mainstream

Micron 7600 | MLPerf Storage

Mainstream performance and incredible energy efficiency for AI workloads



Test Notes

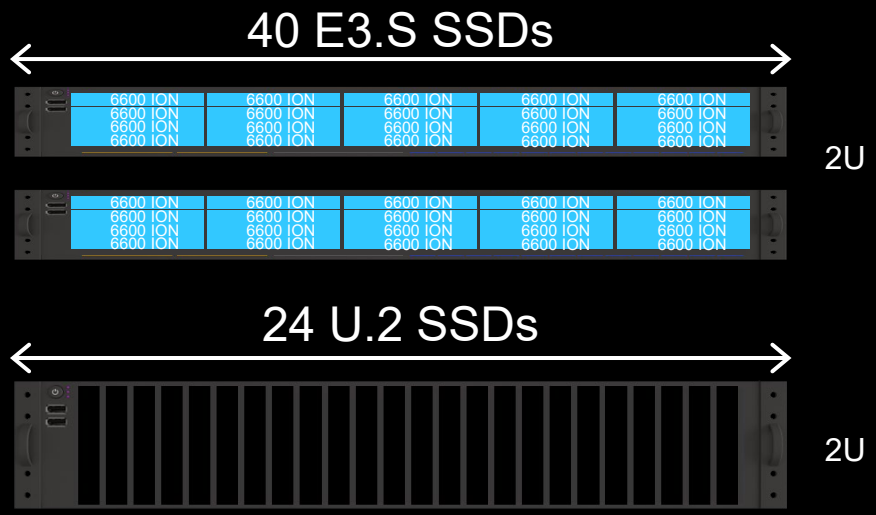
1. The results shown were run in Micron's data center workload engineering lab and are not official MLPerf Storage results.

High Capacity

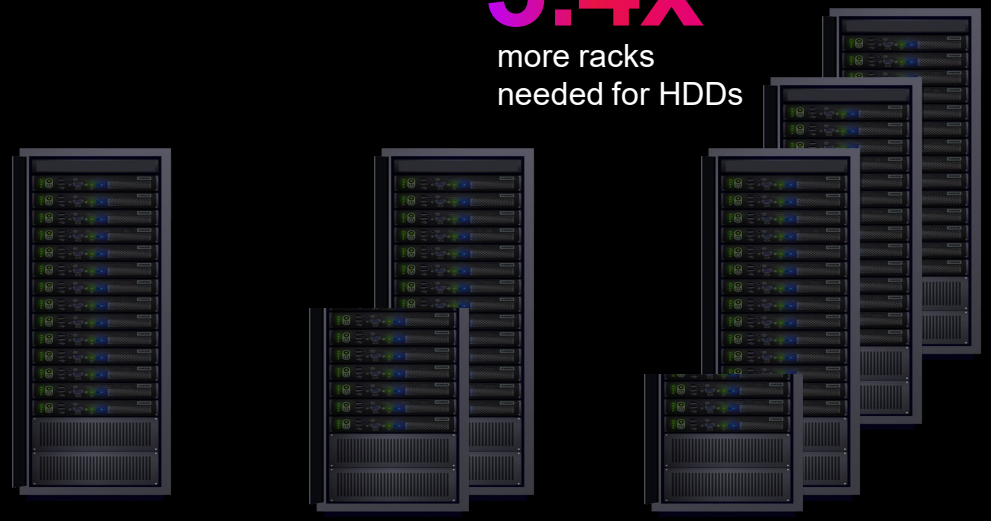
Micron 6600 | Storage Density

Store up to 67% more than U.2 122TB SSDs and 242% more capacity per rack than HDDs

67% increased chassis capacity per 2U enabled by E3.S 122TB density



3.4x more racks needed for HDDs



	Micron 6600 ION	122TB U.2	3.5" HDD
Capacity (TB)	122.88	122.88	36
Form factor[1]	E3.S 1T	U.2	3.5"
Drives per 2U[2]	40	24	40
Drives in 36U	720	432	720
36U rack capacity (PB)[3,4]	**88.5**	**53.1**	**25.9**
Density (PB/U)	2.5	1.5	0.7

1. 6600 ION SSD comparisons are based on currently in-production and available Gen5 high-capacity data center SSDs from the top five competitive suppliers of OEM data center SSDs by revenue as of May 2025, as per Forward Insights analyst report, "SSD Supplier Status Q1/25. Solidigm offers 122TB drives in U.2 and E1.L only.
2. SSD system comparison using 20 slot SSD E3.S based platform in 1U vs. 24 slot U.2 based platform in 2U.
3. SSD rack comparison using Micron E3.S drives with 20 SSDs per U x 36 servers (36U) x 122TB equaling 88.5PB vs the competitor U.2 with 24 SSDs per 2U x 18 servers (36U) x 122TB equaling 53 PB, leaving 6U for other equipment in each rack
4. HDD comparison is based on 42U rack with 36U allocated for server/storage. Each 2U accommodates 40 Micron 6600 ION SSDs (122.88TB each) or 100 HDDs in 5U server/storage bay and a theoretical quantity of 720 HDDs.

Real-world AI workloads need fast, power-efficient storage

Micron is ahead of the curve on components required for efficient AI systems

Storage is important at scale
- Compute is outpacing PCIe
- Many AI workloads are becoming storage intensive

High Performance
- Micron is a leader in Gen6 development
- GPU initiated storage needs Gen6 SSDs
- 107 GB/s on 4x 9650's with GDS on Astera Labs switches
- 86M IOPs from 20x 9650's SSDs to H100

Mainstream
- Power efficiency is a key differentiator
- Enterprise AI will use mainstream SSDs

Capacity
- 122TB E3.S will drive storage density, 67% higher than U.2
- 245TB in E1.L and U.2

Differentiated NVMe swim lanes are required to meet the massive demand of AI workloads.

Would you like to know more?

Visit the Micron team in booth 107