PCIe Gen5 NVMe for AI: MLPerf Storage v2 Findings on Kingston's DC3000ME

Executive Summary

MLCommons MLPerf Storage v2.0 takes a clever approach to testing storage systems for machine learning workloads. Rather than requiring expensive GPU hardware, it uses sleep emulation to simulate the compute portions of ML training while focusing purely on storage performance.

The basic idea works like this: during real ML training, there's a constant cycle of loading data from storage into memory, then processing that data on the GPU. MLPerf recreates this pattern by generating synthetic datasets that mirror real workloads; image files for ResNet-50, cosmological data for CosmoFlow, and medical volumes for 3D U-Net. These files are generated by Pytorch and TensorFlow, industry standard machine learning libraries that match the exact sizes, structures, and access patterns for these models in real production ML training.

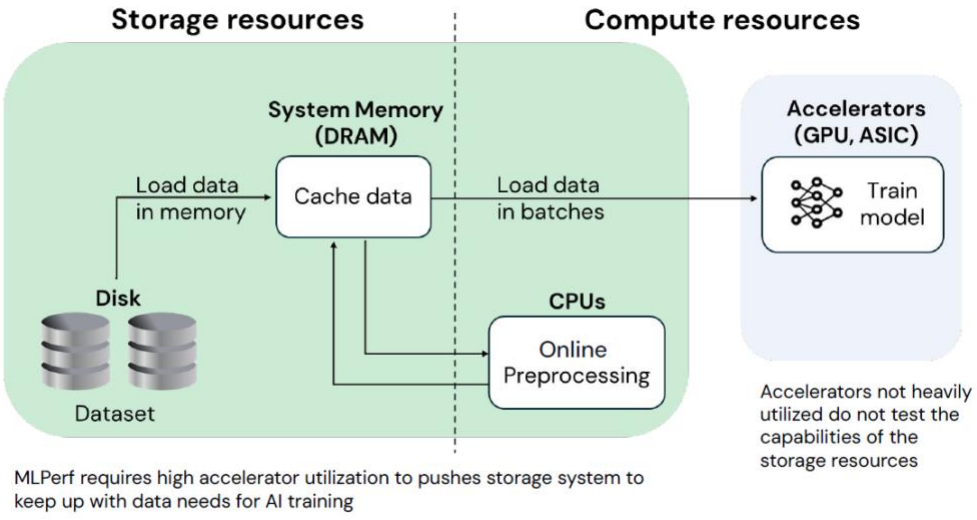


Figure 1 ML training flow visualization. Training Samples-System Memory-CPU-System Memory-GPU Courtesy of MLCommons

MLPerfv2 storage leverages hydra, to pass the yaml configuration parameters for each model to the dlio benchmark. During the training phase, the DLIO benchmark does the sleep emulation and prefetches the data from the filesystem. When running, the benchmark alternates between data loading phases (reading batches from storage) and computation simulation (sleep periods that match real GPU processing times). The sleep durations come from actual performance measurements of A100 and H100 GPUs. Since H100s process data faster, they have shorter sleep times, which means the storage needs to deliver data more quickly to keep up.

Kingston DC3000ME SSDs **(3.84 TB, 7.68 TB, 15.36 TB)** were put through the MLCommons Storage v2.0 benchmark (CLOSED division) to determine per-drive GPU support under representative AI training workloads. MLPerf Storage v2 training measures **how many** A100 or H100 accelerators each drive can **sustain** while maintaining **Accelerator Utilization (AU**) above defined thresholds for **Unet-3D, Resnet-50, and Cosmoflow**. For training workloads, this is the maximum number of GPUs that can satisfy this condition:

- ***ResNet-50 (≥ 90% AU), UNet3D (≥ 90%), CosmoFlow (≥ 70%)***
- ***Where Accelerator Utilization% = (Total Compute Time / Total Benchmark Runtime) × 100%***

Accelerator Utilization, or AU, measures the percentage of time the GPU stays busy versus waiting for data. Slow storage means the GPUs must wait for the batches to be loaded into system memory and results in lower AU scores. Fast storage keeps the data flowing smoothly and hits AU targets of 70-90% depending on the workload. This approach lets storage engineers and infrastructure architects benchmark against realistic ML scenarios without needing actual accelerator hardware, while still getting results that correlate well with real-world performance.

MLPerf v2 storage comes with the addition of checkpointing workload, that simulates a different scenario entirely. Large language model training and fine-tuning run for days or weeks, so ML engineers need to periodically save checkpoints to avoid losing progress if something crashes and load them later to resume training. The benchmark simulates this critical workflow by first reading the model and optimizer weight files from storage into memory buffers, then writing updated weights from memory back to the SSD. In this case, since checkpoints typically occur several times a day, during the fine-tuning process, the time to save the checkpoint to disk and the time to load it back is the correct metric to look at in terms of storage efficiency. Faster, lower latency storage will have lower load and save times along with higher bandwidth while loading and saving the model to disk.

For the checkpoint workloads, Kingston DC3000ME SSDs were put through the llama3-8b (default-mode, 8 processes), 70B, 405B, 1T parameters (subset mode-8 processes). In simple terms, each process represents the number of GPUs in a server. So, in subset mode, a subset of the model's default and optimizer weights are used with 8 processes, to emulate a single server in a cluster with 8 GPUs.

In this paper we walk you through our test method, and test results, and how the results correlate with real world performance.

How we tested

We chose a scientific approach to control as many variables as we could between test runs. We tested all capacities of the DC3000ME in 2 PCIe Gen 5 platforms, SYS-621H-TN12R (Intel Xeon Silver P4510X2), Think system SR-665-V3 (AMD EPYC 9254x2) and 1 Gen4 platform, Dell PowerEdge R7525 (AMD EPYC 7302x2). The goal of the testing is to provide a clear understanding of the performance of DC3000ME with both modern and legacy AI infrastructure deployments.

Operating System	Kernel Version	Open MPI Version	Benchmark
22.04.5 LTS	6.5.0-15-generic	Open MPI 4.1.2	MLperf v2

Platform	CPU	Memory	PCIe Generation
Supermicro SYS-621H-TN12R (Gen5)	Dual Intel Xeon Silver 4510 (24 cores/socket)	256GB (16x16G 4800MHz DDR5 Kingston memory)	5.0
Lenovo SR-665V3 (Gen5)	AMD EPYC 9254 24-Core (24c/48t)	384GB (24x16G 4800MHz DDR5 Kingston memory)	5.0
Dell R7525 (Gen4)	Dual AMD EPYC 7302(16C/32T)	256GB (16x16G 3200MHz DDR4 Kingston memory)	4.0

Table 1: Hardware and Software Used by Kingston for the MLPerf v2 Closed division submission

We developed a script to allow us to batch run multiple accelerators sequentially for each model, with each run following the rules outlined by MLcommons closed rule submission guidelines. An accelerator-count sweep started at one GPU, and we incremented the number of accelerators until the AU dipped below the workload's target. The maximum count at or above threshold yields the per-drive accelerator support, directly guiding storage sizing for GPU fleets.

To determine the AU limit, we start with one emulated GPU and step up the count (`-n`) until the Accelerator Utilization (AU), just dips below the pass mark. Along the way, we log throughput (GB/s), and latency using IOSTAT to help us correlate and tune kernel and drive parameters to achieve the best training performance for each workload. This process is exactly what you should do if you are sizing a real system for your next AI cluster, measuring the storage front lines so your future GPUs never idle data. This makes the MLPerf v2 benchmark the best planning tool for infrastructure architects to understand storage requirement with the most common ML workloads while allowing modifying kernel tunable to maximize IO performance.

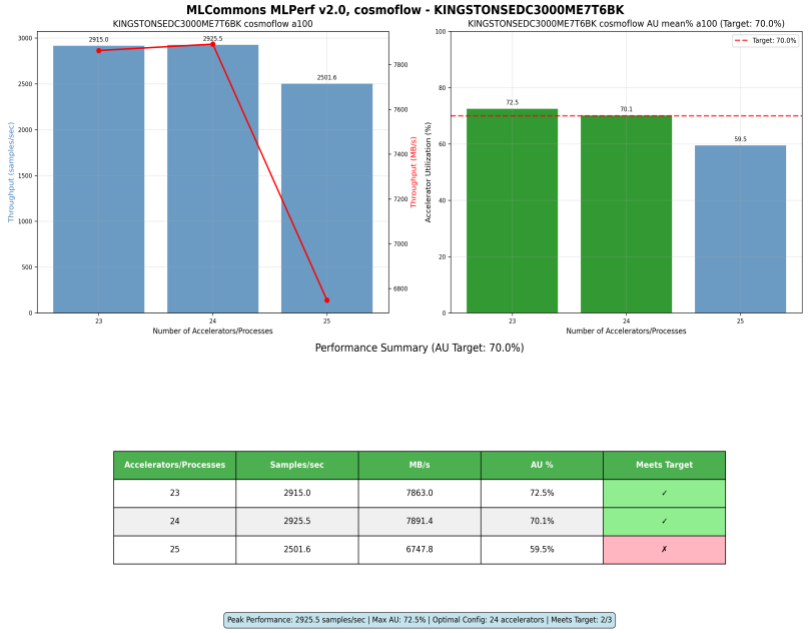


Figure 2: Powerful visualization showing our test method

Key results summary - Training

SEDC3000ME/15.36TB

Workload	Best Configuration	Max Accelerators	Peak Throughput	Peak Bandwidth	Platform
ResNet-50	A100	127	106,090 samples/sec	11,601 MB/s	SR-665 V3 AMD Lenovo Gen5
ResNet-50	H100	67	111,522 samples/sec	12252 MB/s	SR-665 V3 AMD Lenovo Gen5
UNet3D	A100	9	90 samples/sec	12,591 MB/s	SYS-621H-TN12R Intel SMC Gen5
UNet3D	H100	4	85 samples/sec	11,933 MB/s	SYS-621H-TN12R Intel SMC Gen5
CosmoFlow	A100	23	2,921 samples/sec	7,878 MB/s	SR-665 V3 AMD Lenovo Gen5
CosmoFlow	H100	19	3,849 samples/sec	10,383 MB/s	SR-665 V3 AMD Lenovo Gen5

Key results summary – Checkpointing SEDC300ME 7.68 TiB

Model	Max Parallel Processes	Min Load BW(GiB/s)	Min Save BW(GiB/s)
Llama3-8b	8	11.5	9.2
Lllama3-70b	8	11.6	8.2
Lllama3-405b	8	10.8	7.6
Lllama3-1t	8	10.6	5.6

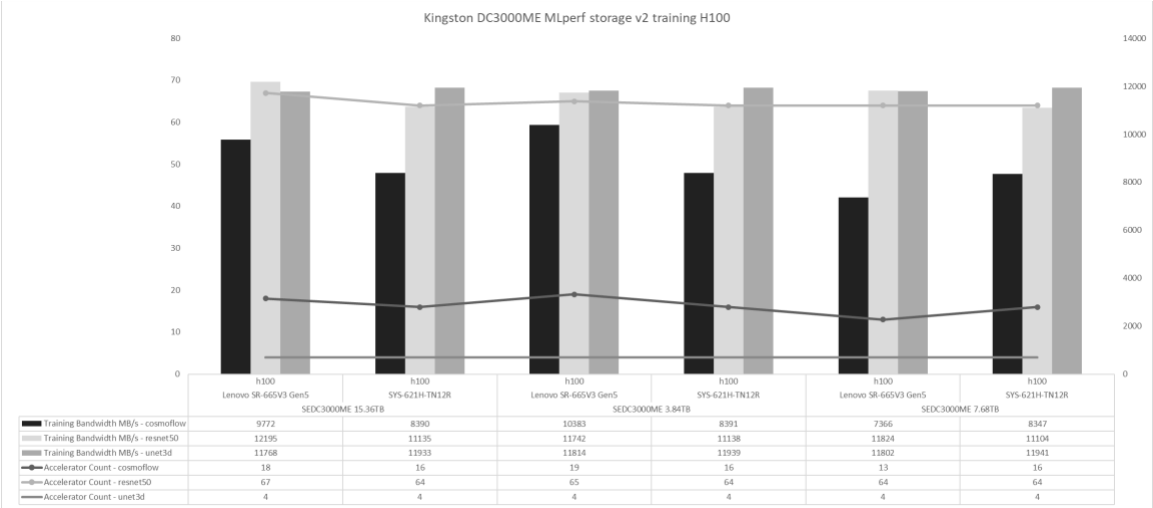


Figure 3 SEDC3000ME H100 results, Gen5

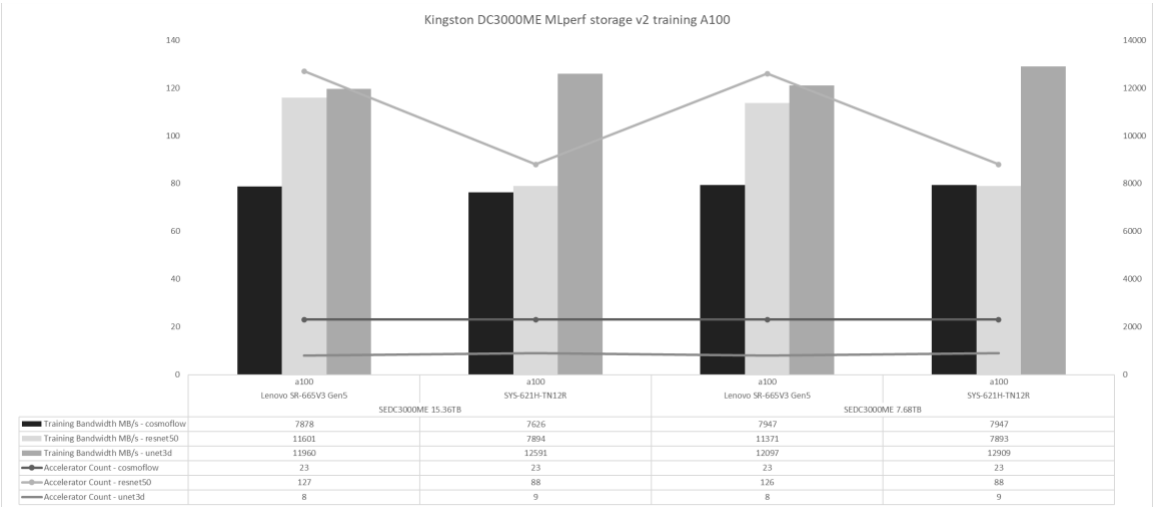


Figure 4 SEDC3000ME A100 results, Gen5

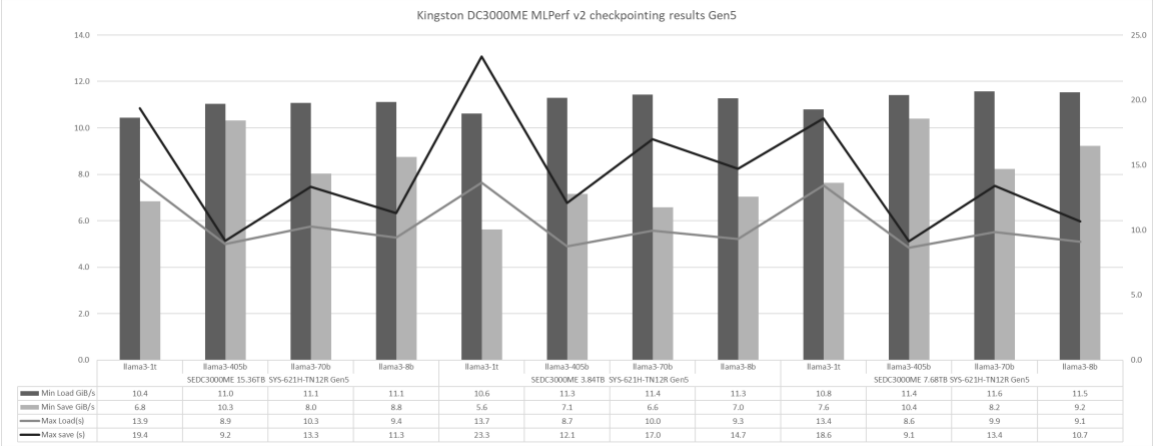


Figure 5 SEDC3000ME checkpointing results Gen5

Training Synopsis

So *why* is Gen5 storage important for ML training workflows? It's ***performance per bay efficiency***. MLPerfv2 storage has proven that fast storage plays a role in ML training workflows, especially for certain workloads.

In many cases, PCIe Gen5 storage is necessary. For example, for a server with 8 H100 GPUs, you need at least 2 Gen5 NVMe SSDs to maximize accelerator utilization for unet3d medical imaging ML training workflows. To get the same level of performance from Gen4 drives, you need to double the number of drives. If we take this theory and apply it to a cluster of servers; 8 DC3000ME drives can service a cluster of 4 servers, fully populated with 8 H100s.

Our testing confirmed that Kingston's DC3000ME drives have robust accelerator support for all the training workloads:

- **ResNet-50**: DC3000ME maintained 127 A100 accelerators at 91.6% accelerator utilization on the SR-665V3 system with 11,601 MB/s sustained throughput. H100 setups provided 109,527 samples/sec with 11,977 MB/s bandwidth with 65 accelerators at 93.4% AU, significantly higher than the 90% threshold requirement.
- **UNet3D**: Peak bandwidth achieved 12,591 MB/s for 9 A100 accelerators at 91.2% AU on the SYS-621H-TN12R system. H100 performance delivered 11,933 MB/s with 4 accelerators sustaining 98.8% accelerator utilization, showing great efficiency for medical imaging workloads.

- **CosmoFlow**: The drives sustained 19 H100 accelerators at 73.4% AU with 10,383 MB/s throughput for cosmological simulation workloads. A100 configurations maintained 23 accelerators at 7,878 MB/s while surpassing the 70% AU threshold.

Checkpointing Synopsis

Checkpointing is a critical workflow in the LLM fine-tuning process, and the underlying storage must be both performant and reliable in the fine-tuning workflow. In a typical fine-tuning workflow, checkpoints are saved and loaded multiple times a day. PCIe Gen5 Enterprise SSDs, like SEDC3000ME, provide both reliability and performance, and allow fine-tuning workflows to a) take more frequent checkpoints so progress is never lost; and b) eliminate time overhead caused by saving and loading checkpoints because the underlying IO subsystem is no longer a bottleneck.

In the checkpointing workloads, DC3000ME delivered excellent load performance, consistently achieving > 10GiB/s worst case load bandwidth across all SKUs and model weights, along with fast worst case model load times (8.6-13.9s), making it a great option for restoring checkpoints seamlessly. In the checkpoint saves, throughput varies more with capacity and model size; llama3-405b sustains > 10 GiB/s on 7.68 TB and 15.36 TB, while the 3.84 TB SKU drops to 5.6–7.1 GiB/s, with a (9.1-12.1s) max save time.

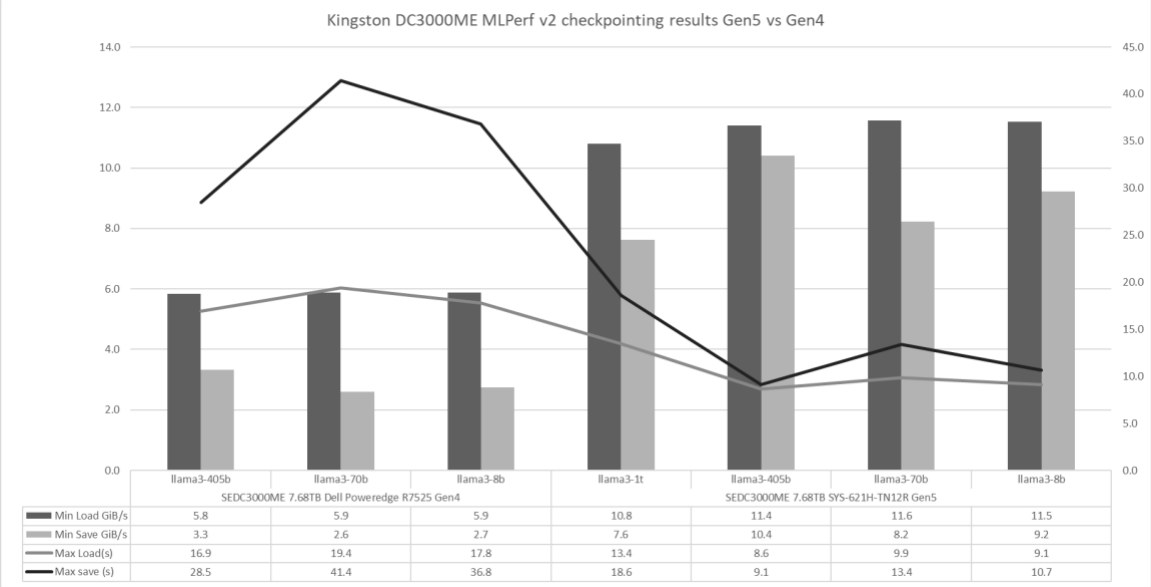
Theoretically, this checkpoint performance can scale with the number of drives and the number of servers. In this scenario, we are emulating 8 GPUs running on a single server, but the performance should scale almost linearly with the number of drives, and the number of servers. So, in theory 4 DC3000ME/7.68T per server in a 400Gb/s 4 node cluster can saturate the throughput of 1 400Gb/s uplink (~50GB/s).

PCIe Gen5 vs Gen4 Training and Checkpointing Synopsis

In our closed division submission, we also wanted to provide some quantifiable numbers to provide some insight into training and checkpointing on Gen4 servers, to analyze the impact of the interconnect as a bottleneck to ML training workflows. The Gen5 interconnect provided ~2x the number of accelerators supported for unet3d training workloads (4 vs 8 a100s),2x faster checkpoint load times (16.9 vs 8.6s for llama3-405b) and 3x faster checkpoint save times (28.5 vs 9.1s). From a business perspective, this translates to more GPU utilization in fine-tuning workflows, and less time wasted saving and loading checkpoints.

cosmoflow - Training Bandwidth MB/s					6019.62936	
resnet50 - Training Bandwidth MB/s	6461.8	6472.3	6472.0	6475.4	6505.2	9094.7
unet3d - Training Bandwidth MB/s	6064		6070		6069	
cosmoflow - Accelerator Count					18	
resnet50 - Accelerator Count	71	36	71	36	71	36
unet3d - Accelerator Count	4		4		4	

Figure 6 SEDC3000ME Gen4 training results



Min Load GiB/s	5.8	5.9	5.9	10.8	11.4	11.6	11.5
Min Save GiB/s	3.3	2.6	2.7	7.6	10.4	8.2	9.2
Max Load(s)	16.9	19.4	17.8	13.4	8.6	9.9	9.1
Max save (s)	28.5	41.4	36.8	18.6	9.1	13.4	10.7

Figure 7 SEDC3000ME Gen4 vs Gen5 training results

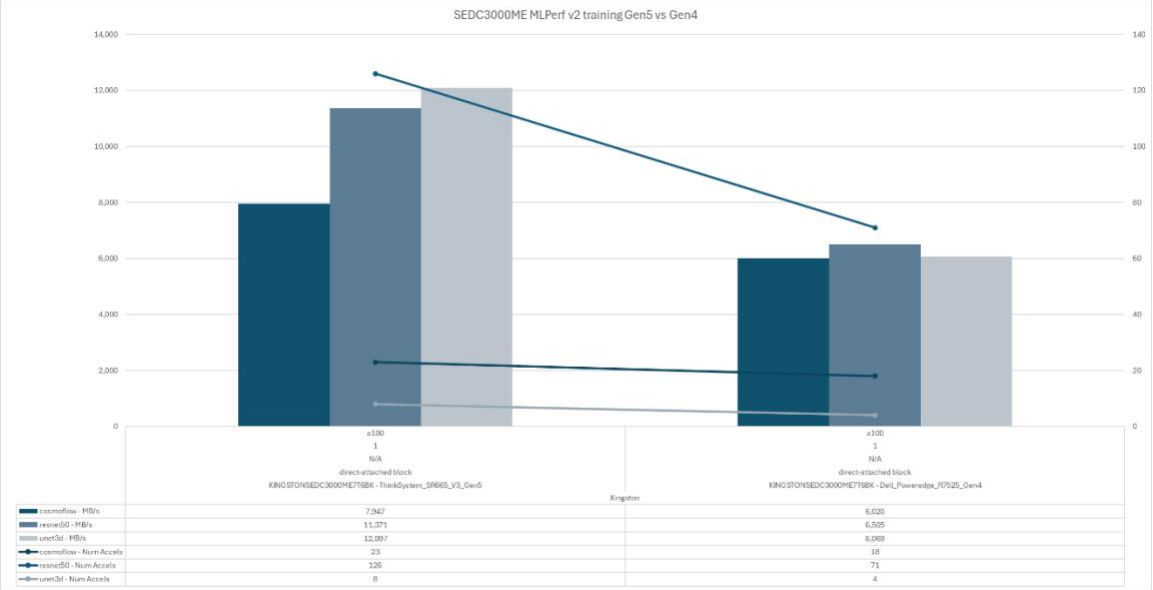


Figure 8 SEDC3000ME Gen4 vs Gen5 training results

Thinksystem SR-665V3 Power State Clarification & Post Deadline results

The SR-665V3 platform supports firmware/BIOS selectable system power operating points (according to Lenovo Press LP2210):

Maximum Efficiency: perf/watt bias: Determinism=Performance, cTDP/PPL=Auto (~SKU TDP), Power Profile=Efficiency; P-states enabled (DVFS), xGMI=20 GT/s (2S) → lower boost, lower fabric BW/power draw.

Maximum Performance: throughput bias: Determinism=Power, cTDP/PPL=Max, Power Profile=High Performance; P-states disabled (lock P0), xGMI=32 GT/s → higher sustained clocks & fabric bandwidth at the cost of watts/heat and variability

In short, Max-Performance essentially trades energy efficiency for **higher sustained clocks**, and **faster inter-socket memory bandwidth**, and **looser determinism**.

During the MLPerf Storage v2 CLOSED submission cycle, the SR-665V3 system was inadvertently left in its default **Maximum Efficiency mode**. As a result, all closed Lenovo SR-665V3 results in the preceding Training table show a **conservative** power profile. After the submission deadline closed, we reran the training and checkpointing workloads in **Maximum Performance mode** to evaluate the impact of proper power tuning on the scalability of per-drive accelerators.

Performance deltas (Maximum Performance vs submitted Maximum Efficiency):

- ResNet-50 (H100): Accelerator ceiling increased from **65 to 75** (+15%), throughput from **11.98 to 13.29 GB/s**.
- ResNet-50 (A100): **127 to 140 accelerators (+10%),** throughput **11.60 to 12.79 GB/s**.
- CosmoFlow (H100): **19 to 20 accelerators (marginal ~5% improvement)**, same bandwidth (~10.4 to 10.27 GB/s; workload variability dominates).
- CosmoFlow (A100): **23 to 30 accelerators (+30%)**, **throughput 7.95 to 9.86 GB/s.**
- UNet3D: Higher counts (5 H100 / 10 A100) failed AU ≥90%; previous maxima (4 H100 / 8 A100) still representative, meaning the workload was storage-saturated instead of power-limited.

	SEDC3000ME 3.84 resnet50 h100	SEDC3000ME 3.84 resnet50 a100	SEDC3000ME 3.84 cosmoflow h100	SEDC3000ME 3.84 cosmoflow a100	SEDC3000ME 7.68 resnet h100	SEDC3000ME 7.68 resnt a100	SEDC3000ME 7.68 unet3d a100	SEDC3000ME 15.36 resnet h100	SEDC3000ME 15.36 resnet a100	SEDC3000ME 15.36 cosmoflow a100
Accelerator Ceiling (MaxPerf)	75	141	21	29	75	140	9	75	140	30
Accelerator Ceiling (Eff)	67	127	20	23	67	126	8	67	127	23

Figure 9 Max performance vs Max Efficiency number of accelerators impact

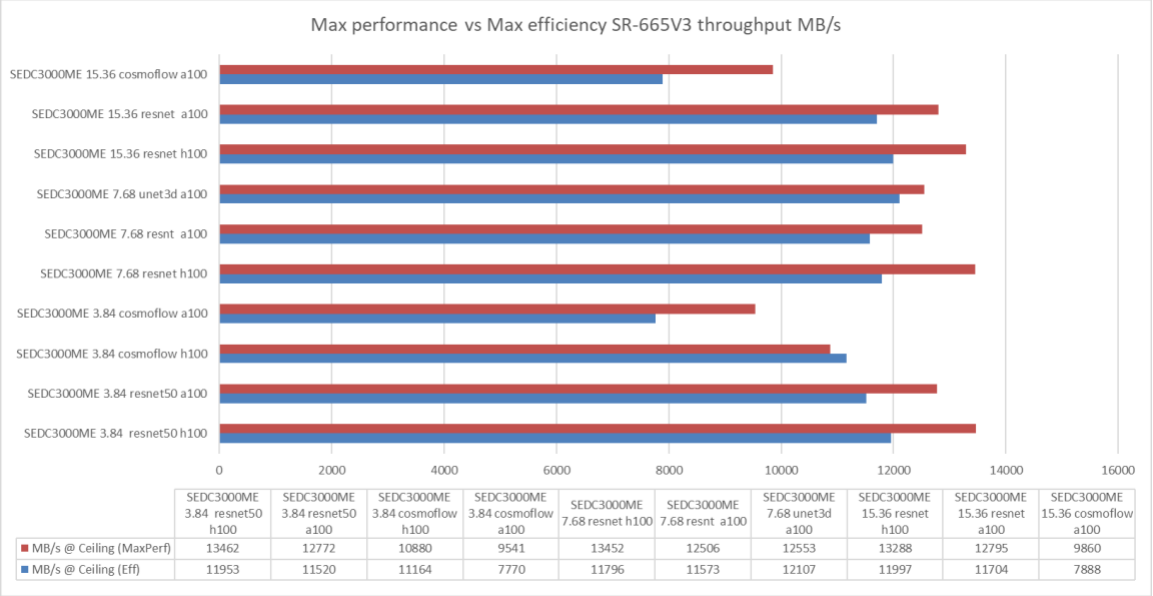


Figure 10 BW ceiling Max performance vs Max efficiency SR-665V3

Conclusion

MLperf storage V2 has proven that fast storage plays a critical efficiency role in machine learning training and checkpointing. PCIe Gen5 Enterprise SSDs like SEDC3000ME serve as building block for AI GPU clusters to eliminate the I/O bottleneck and maximize accelerator utilization. They help organizations maximize their GPU hours by shortening checkpoint load and save times. In the closed division results, SEDC3000ME showed very strong, competitive performance in all training workloads, and some of the fastest load times in checkpoint workloads, making it a high impact storage solution that dramatically accelerates AI training workloads.

Engineer's Note on Tunables

We applied these tunables in our closed division submission:

- *NVMe interrupt coalescing set to feature-id 0x8 = 0x070a (700 µs timeout, 11-I/O threshold) reduces CPU overhead under bursty loads.*
- *ResNet-50 (random small reads): max_sectors_kb=32, read_ahead_kb=32, nomerges=2, 4 reader threads, improves random I/O throughput.*
- *UNet3D, CosmoFlow, and checkpointing (sequential workloads): max_sectors_kb=128, read_ahead_kb=4096, nomerges=0, uses large sequential tunables for steady bandwidth. param reader.read_threads=16 for unet3d & cosmoflow*
- *System-wide settings: rq_affinity=2 (NUMA-aware interrupt routing), wbt_lat_usec=0 (disable write-back throttling), add_rand (scatter I/O to avoid alignment hotspots).*

Author: Hazem Awadallah, Senior Systems Engineer, Kingston Technology