This information is under embargo until 9/1/26 8:00AM PT

MLCommons Supplemental Discussion

The submitting organizations provided the following 300-word descriptions as a supplement to help the public understand their MLCommons® MLPerf® Storage v3.0 submissions and results. The statements do not reflect the opinions or views of MLCommons.

This information is under embargo until the public release at 9/1/26 8:00AM PT

This information is under embargo until 9/1/26 8:00AM PT

Alluxio

Alluxio Enterprise AI is a data acceleration layer that enables AI workloads to access data in cloud object storage at high throughput and low latency. Deployed alongside GPU infrastructure, Alluxio aggregates local NVMe into a distributed caching layer while keeping object storage as the underlying system of record. Applications can access data through POSIX via FUSE, S3, HDFS, and other interfaces without changing where datasets are persistently stored.

For MLPerf® Storage v3.0, Alluxio submitted closed-division results for training and checkpointing using Amazon S3 as the under-file-system. The submission demonstrates a software-defined architecture that combines cloud object storage with commodity NVMe, scaling from one worker with 25.5 TiB of usable cache to 32 workers with 816 TiB.

In training, Alluxio supported 32 simulated NVIDIA B200 accelerators for RetinaNet and eight simulated B200 accelerators for 3D U-Net, delivering 4.38 GiB/s and 43.68 GiB/s of read bandwidth, respectively.

For Llama 3 checkpointing, Alluxio evaluated 8B, 70B, and 405B models. The largest configuration ran the 405B workload with 512 processes across 32 clients, achieving **147.0 GiB/s write and 124.3 GiB/s read bandwidth**.

These results demonstrate that demanding AI training and large-model checkpointing can be supported while keeping persistent datasets in scalable cloud object storage. By combining Amazon S3 with distributed commodity NVMe, Alluxio provides a high-performance data path to GPU infrastructure **without requiring organizations to move datasets into a separate specialized storage system**, helping simplify AI data infrastructure while preserving the economics and flexibility of object storage.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

Everpure

Everpure's MLPerf Storage v3.0 closed-division submissions evaluate Everpure FlashBlade//EXA for large-model checkpointing and KVCache workloads. The checkpointing results highlight that throughput increases as the number of data nodes increases.

For closed 1T checkpointing, the testbed used 32 client hosts and 1,024 simulated accelerators across four data-node configurations. The measurements are summarized below, sorted by write bandwidth:

# of Hosts	# of data nodes	# of simulated accelerators	Checkpointing - Write B/W (GiB/s)	Checkpointing - Read B/W (GiB/s)
32	30	1,024	877.52	588.28
32	20	1,024	655.60	533.76
32	15	1,024	484.10	439.87
32	10	1,024	327.65	365.12

All checkpointing configurations used Full checkpoint mode and two data-parallel instances.

Closed KVCache submissions used 51 client hosts and 20 data nodes. For Llama 3.1 8B storage-only mode, throughput was 85,736.04 tokens/s. In storage-plus-memory mode, reported throughput was 67,641.97 tokens/s. For Llama 3.1 70B storage-only mode, throughput was 33,402.57 tokens/s.

For the configurations above, 17 reported rack units were attributable to the FlashBlade//EXA three-chassis MDN (three 5 RU chassis plus two 1 RU XFMs). Another 16 reported rack units were attributable to the NVIDIA SN600 data network switches (two NVIDIA SN5600 spine switches and six NVIDIA SN5600 leaf switches). Because testing time was limited, MDN-chassis and data-switch minimization tests were not performed. Therefore, the reported results do not establish the minimum MDN or switch count; fewer MDN chassis or data switches may have been sufficient. The data nodes used in these tests were 1U Supermicro ASG-1115S-NE316R servers.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

FarmGPU

FarmGPU is pleased to present its MLPerf Storage v3.0 submissions across checkpointing, training, and KV cache workloads, with two complementary system designs: **Hickory**, a WEKA-based parallel filesystem cluster, and **Potato**, a fleet of client-local Solidigm QLC NVMe arrays.

Hickory consists of eight 2U WEKA backend nodes, with 18 active Solidigm D7-PS1010 7.68 TB SSDs. The 16-RU system provides 646.58 TiB usable capacity with 5+2 erasure coding and presents one shared POSIX namespace to 15 benchmark clients. All benchmark I/O traverses the network to WEKA; client-local NVMe is excluded from the storage system under test.

Hickory achieved (closed division):

- Checkpointing:
  - ~5.3 TB 405B checkpoint: 165.21 GiB/s write & 283.65 GiB/s read
  - ~15.4 TB 1.25T checkpoint: 169.67 GiB/s write & 252.11 GiB/s read
- UNet3D training:
  - ~406.72 GiB/s w/ 75 simulated NVIDIA B200s @ ~92% utilization

Each **Potato** node utilizes 24 Solidigm D5-P5336 122.88 TB QLC NVMe drives in a single array, putting dense, cost-efficient Solidigm capacity directly behind the compute node for latency-sensitive KV cache workloads; the eight nodes total 192 Solidigm drives and roughly 23.6 PB of raw QLC capacity.

The eight-node Potato configuration achieved the following:

- In **closed** division:
  - Llama 3.1 8B storage-only: 15,522 tokens per second & 248.47 GiB/s read
- In **open** division:
  - Llama 3.1 8B: 451.02 GiB/s of read bandwidth & 75,975 tokens per second
  - Llama 3.1 70B: 888.80 GiB/s read bandwidth & 30,896 tokens per second

With this submission, FarmGPU demonstrates the value of both a shared parallel filesystem, and Solidigm SSDs for different points in the AI data lifecycle, supporting the broader community's work scaling large model training and serving on efficient storage infrastructure.

This information is under embargo until 9/1/26 8:00AM PT

Holmes AI

HolmesAI Limited is pleased to present its MLPerf Storage v3.0 closed-division training results for StorMeshAI, an available shared, remote parallel filesystem designed to provide scalable, high-throughput data access for AI workloads. StorMeshAI presents a POSIX file interface and uses a distributed chunk pool with three-way replication across three storage nodes. The submitted system provides 42.84 TiB of usable capacity from 128.52 TiB of raw NVMe capacity.

File payload traffic is carried over two independent 400 Gb/s NDR InfiniBand/RDMA links on the client and each storage node. StorMeshAI distributes payload traffic across the two links through per-chunk hash-based path selection, while a separate 50 Gb/s Ethernet bond supports management, control-plane, and filesystem metadata operations. This separation allows file payload traffic and filesystem coordination traffic to be managed independently.

In closed RetinaNet training, StorMeshAI supported 20 simulated NVIDIA B200 accelerators from one client and delivered 2.59 GiB/s of training read bandwidth. In closed UNet3D training, the system supported four simulated NVIDIA B200 accelerators from one client and delivered 21.42 GiB/s of training read bandwidth.

For the broader AI community, this submission provides a transparent example of how a shared remote filesystem, protected NVMe capacity, and high-speed RDMA networking can be combined to support AI training data paths. HolmesAI Limited hopes these results help infrastructure teams evaluate practical storage designs for scalable AI training infrastructure.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

HPE

HPE is committed to producing high-performance products and capabilities for demanding AI workloads. As a continuation of HPE's results published in MLPerf Storage v1.0 and 2.0, HPE Cray Supercomputing Storage Systems demonstrated high-performance results throughout all tests for MLPerf Storage v3.0. This round includes new performance results for:

- HPE Cray Supercomputing Storage Systems E2000 with embedded open-source Lustre file system
- HPE Cray Supercomputing Storage Systems K3000 with embedded open-source distributed asynchronous object storage (DAOS) software

HPE Cray Supercomputing Storage Systems are high-performance solutions designed for optimal input / output to address large-scale HPC and AI workloads where low-latency, and support for multi-tenant access, is critical. The new KV cache benchmark measures how storage systems handle key-value (KV) cache data during large language model (LLM) inference and multi-turn conversational AI. The checkpointing benchmark measures how fast and efficiently a storage system can load and save large AI model states during training. The training benchmark measures how fast data storage systems can supply training data to AI accelerators during machine learning training.

Notably, HPE Cray Supercomputing Storage Systems K3000 demonstrated high performance throughout all three training, KV cache, and checkpointing tasks by achieving:

- Over 345 gibibytes per second (GiB/s) read bandwidth across 15 clients in training
- Over 8,700 tokens per second and over 151 GiB/s across 16 clients in KV caching
- Over 83 GiB/s read bandwidth across 8 clients in checkpointing

Similarly, the HPE Cray Supercomputing Storage Systems E2000 delivered high performance in training and checkpointing by achieving:

- Over 167 GiB/s read bandwidth across 15 clients in training
- Over 110 GiB/s read bandwidth across 16 clients in checkpointing

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

Inspur Data

Inspur Data is pleased to participate in the MLCommons MLPerf Storage V3.0 benchmark. The AS13000G7, a distributed symmetric architecture product, is built for AI and HPC scenarios, serving as a data foundation for model training, inference, and massive data storage.

In the Checkpoint 70B model training scenario, the product achieved single-node read bandwidth of 269 GB/s and write bandwidth of 195 GB/s. High-throughput checkpoint persistence is essential for large-scale model training workflows, where frequent state snapshots are required to protect training progress and enable fault recovery. These metrics indicate that the product can sustain demanding write-intensive operations without becoming a bottleneck, allowing training tasks to proceed with minimal interruption.

In the KVCache 8B inference scenario, the product achieved single-node throughput of 14,579.22 tok/s. During large language model inference, KV cache access can place significant pressure on both GPU memory and the storage data path. By sustaining high-bandwidth KV cache reads, the product helps reduce data movement overhead and alleviate memory constraints, contributing to more efficient token generation and inference pipeline execution.

Across the training and inference workloads evaluated in this benchmark, the measured results reflect the product's capability to handle data-intensive AI tasks. These outcomes provide a data-driven reference for understanding how AS13000G7 performs as part of an end-to-end large-model infrastructure.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

Microsoft

Microsoft's MLPerf Storage v3.0 submission reflects our commitment to advancing AI infrastructure through transparent, reproducible performance results. The submission demonstrates the performance of Azure Managed Lustre (AMLFS), a fully managed parallel file system available in Azure, across the training and checkpointing benchmark categories. Azure Managed Lustre can be provisioned on demand close to GPU clusters in Azure fleet, enabling low-latency access and flexible hourly consumption pricing. The results help the broader AI community understand how Azure Managed Lustre supports modern AI workloads with sustained throughput, predictable performance, and tier-based deployment choices.

In the training category, the evaluated configurations maintained consistent throughput and accelerator utilization for five consecutive hours across all 13 submitted results, with filesystem capacities ranging from tens of TiBs to tens of PiBs. This predictable data delivery supports high accelerator utilization throughout long-running AI workloads. Azure Managed Lustre also demonstrated checkpointing performance of up to 642 GiB/s write throughput and 490 GiB/s recovery throughput, enabling checkpoint creation and restoration of some of the industry's largest AI models consistently in just tens of seconds or less. Together, these results demonstrate highly efficient storage performance for sustained training data access and rapid protection and recovery of model state, helping maximize accelerator utilization and minimize wasted GPU hours.

Across both categories, Microsoft's submission spans five Azure Managed Lustre service tiers, the breadth of configurations available to support the most demanding AI workloads. AI infrastructure requirements vary by model size, accelerator count, data volume, and I/O pattern. Testing across five tiers provides insight into how customers can balance throughput, capacity, and cost as models, datasets, and compute clusters evolve.

Accurate sizing of an Azure Managed Lustre filesystem unlocks faster time to results and an optimal performance balance, with the built-in durability and resiliency of a cloud-managed filesystem.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

Nebius

For MLPerf® Storage v3.0, Nebius submitted Closed division results for its fully S3-compatible Object Storage service. All tests used the Enhanced Throughput storage class. The benchmarking covered two workload categories: training and checkpointing.

Training submissions ran the RetinaNet workload on 32 client nodes with 768 simulated NVIDIA B200 accelerators, and the Unet3D workload on 21 client nodes with 21 simulated B200 accelerators. Checkpointing submissions covered two Llama 3 model scales: 8B on a single node with eight simulated accelerators, and 70B on eight nodes with 64. All runs read from and wrote to Object Storage directly through the S3 API.

Nebius tests, tunes and optimizes its storage infrastructure to ensure AI workloads receive consistent, scalable I/O performance in a cloud environment. This work spans multiple layers of the storage stack, from hardware qualification to storage software optimization and multi-node cluster validation. These efforts help ensure stable performance and predictable system behavior across workload types and model scales.

The submitted results reflect this ongoing work. Training runs sustained high accelerator utilization, and checkpointing bandwidth scaled with model and cluster size. All client nodes ran as publicly available Nebius Cloud virtual machines over Ethernet, with capped memory. The outcomes confirm that Nebius Object Storage supports demanding AI data pipelines, from streaming training data to writing and restoring checkpoints, without a separate filesystem layer.

The MLPerf® Storage benchmarks from MLCommons provide a standardized framework for evaluating storage system performance for AI workloads. Nebius's results contribute to this effort and highlight the readiness of its Object Storage service to support training and checkpointing workloads at scale.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

NewFW

NewFW, an independent storage engineering firm and Solidigm's benchmarking partner, has announced its MLPerf Storage v3.0 submissions featuring Solidigm's latest enterprise NVMe SSDs: the D7-PS1010 (PCIe Gen5, TLC, 7.68 TB) and the D5-P5336 (PCIe Gen4, QLC, 61.44 TB). The pairing spans AI storage's two extremes: a Gen5 TLC performance flagship and a Gen4 ultra-high-capacity QLC drive that keeps large AI datasets on a single device. NewFW ran both drives in the CLOSED division, in single-host, single-drive configurations with directly attached NVMe storage — the only ones of their kind in MLPerf Storage v3.0.

The D7-PS1010 was benchmarked across all four v3.0 workloads, delivering 11.6 GiB/s read bandwidth on Unet3D training I/O with 2 emulated B200 accelerators, 13.1 GiB/s read on Llama 3.1 checkpointing, 871 tokens/s at 14.5 GiB/s read and 3.1 GiB/s write bandwidth on the Llama 3.1 8B KV-cache workload, and 140 queries/s at 9.0 ms latency on a 10M-vector DiskANN vector-database workload. The D7-PS1010 combines high performance with low latency — a Solidigm Gen5 TLC drive purpose-built for AI workloads.

The D5-P5336 delivered 5.8 GiB/s training reads with an emulated B200 accelerator and 480 tokens/s at 8.0 GiB/s read and 1.8 GiB/s write bandwidth on the Llama 3.1 8B KV-cache workload — a compelling data point for capacity-optimized AI storage.

NewFW applied deep, NVMe-specific optimization — XFS formatting, mount options, kernel VFS tuning — plus workload-specific parameter refinement on a dual-AMD EPYC 9655 server (Ubuntu 24.04). The results give customers credible, reproducible reference numbers for sizing AI storage on Solidigm drives, demonstrating Solidigm's commitment to AI-optimized storage — a validated path from Gen5 TLC performance to 61.44 TB QLC capacity.

This information is under embargo until 9/1/26 8:00AM PT

Nutanix

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

NVIDIA

NVIDIA's MLPerf Storage v3.0 submission evaluates AIStore, an open-source, deploy-anywhere distributed cache for object-store data specifically designed for AI workloads. AIStore scales linearly, supports fast tiering across remote cloud backends, and exposes a unified namespace. NVIDIA tested its S3-compatible interface in training and checkpointing benchmarks.

Our AIStore deployment on Oracle Cloud Infrastructure (OCI) demonstrated near-linear scalability. UNet3D training throughput nearly doubled whenever the storage-node count doubled: 29.15 GiB/s with three nodes, 58.22 GiB/s with six, and 115.58 GiB/s with twelve. The twelve-node configuration supported 20 simulated NVIDIA B200 accelerators at 98.02% average utilization.

Checkpoint recovery showed similarly strong scaling. Llama 3 1T checkpoint recovery reads scaled from 34.20 GiB/s with three nodes to 70.01 GiB/s with six and 136.54 GiB/s with twelve. Write throughput rose from 33.62 to 64.50 and 133.27 GiB/s. At twelve nodes, read and write throughput reached 97.7% and 95.4% of the storage nodes' aggregate nominal network line rate, showing that provisioned network bandwidth was the practical performance limit.

The submission also demonstrates AIStore's cloud portability. NVIDIA also tested three-node UNet3D configurations on AWS, Google Cloud, and OCI through the same S3-compatible interface. Because instance types, networks, CPUs, client counts, and tuning varied, these results demonstrate consistent cross-cloud performance rather than a direct provider comparison.

This submission covers only the AIStore project; AIStore is not an NVIDIA-managed storage service. NVIDIA is publishing the results to help the AI infrastructure community design scalable object-storage caches for training and checkpointing.

This information is under embargo until 9/1/26 8:00AM PT

OpenLake

OpenLake Technologies, a AI data infrastructure provider for accelerated computing, today announced its MLPerf Storage v3.0 Closed Division results, powered by its Infinity Core I/O Engine.

OpenLake submitted results for Llama 3.1 8B checkpointing, and achieved a 6.72 GiB/s write bandwidth and 11.55 GiB/s read bandwidth from a single benchmark client. This was achieved by OpenLake's asynchronous io_uring engine and pinned thread per core architecture, allowing it to use CPU resources efficiently while sustaining high levels of parallel I/O. The team applied XFS filesystem tuning, fine grained I/O coalescing, and workload specific optimizations to saturate the NVME drives.

The Llama 3.1 8B measures fast checkpointing and measures the ability of training systems to maintain progress without getting stalled on I/O. OpenLake ran the benchmark on an HPC cluster with a gateway on a RAID0 NVMe disk array connected through a 400 Gb/s InfiniBand network. The setup is complemented by GPU native lossless compression to compress data on the GPU before it is written to storage, reducing the storage footprint and network traffic for data intensive workloads such as KV cache offloading.

Together, these results demonstrate a new level of efficiency for object storage for demanding AI workloads. The architecture minimizes data movement through RDMA, on-GPU lossless compression, and small file packing, which reduce network traffic and CPU overhead.

Beyond checkpointing, the Infinity Core Engine also powers AI training, KV cache offloading, vector search, context storage, and other data intensive applications. By publishing these results through MLCommons, OpenLake aims to guide the AI infrastructure community with data points for reducing infrastructure spends and improving hardware efficiency.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

Samsung

Samsung is proud to announce our submission to MLPerf Storage v3.0, showcasing the AI storage performance of the Samsung PM1763 NVMe SSD across the full AI workload pipeline.

MLPerf Storage is the industry-standard benchmark for evaluating storage systems under real AI workloads. In v3.0, Samsung submitted results across all benchmark categories on a single-node configuration using the PM1763, Samsung's PCIe Gen6 TLC NVMe SSD, with 16 TiB of usable capacity.

Training: The PM1763 sustained training data delivery across two workloads on NVIDIA B200 accelerators. On RetinaNet, Samsung achieved 85.9% Accelerator Utilization (AU) across 40 simulated B200s — demonstrating that a single PM1763 can keep a large GPU cluster consistently busy. On Unet3D, the drive delivered 26.59 GiB/s read bandwidth, ensuring storage never became a bottleneck for the compute pipeline.

Checkpointing: On the LLaMA3-8B checkpoint benchmark, the PM1763 achieved 15.91 GiB/s write and 17.77 GiB/s read bandwidth, completing a full checkpoint write in under 6.7 seconds — minimizing recovery overhead in large-scale LLM training.

KV Cache (new in v3.0): KV Cache workloads are not read-only — storage must handle simultaneous reads and writes as tokens are continuously prefetched and evicted. Under this mixed I/O pressure, Samsung achieved 813 tok/s throughput with 13.54 GiB/s read and 2.87 GiB/s write bandwidth concurrently for LLaMA3.1-8B, demonstrating the PM1763's ability to sustain high inference throughput under real-world bidirectional storage demand.

These results confirm that the Samsung PM1763 PCIe Gen6 NVMe SSD delivers the throughput, responsiveness, and reliability demanded across the full AI pipeline. Samsung remains committed to advancing storage performance to power the next generation of AI infrastructure.

This information is under embargo until 9/1/26 8:00AM PT

Suzhou Zishan Longlin

GalaxSphere is an all-flash data intelligence engine powered by the GLFS Data Operating System. It is designed for enterprise GPU clusters, private large language models, RAG, agents, multimodal AI, HPC, and AI-HPC convergence environments where data access must be visible, manageable, and highly responsive.

In the MLPerf Storage v3.0 submission, GalaxSphere showed strong results across AI training, checkpointing, KVCache, and vector database workloads. These results reflect the system's ability to serve data-intensive AI pipelines through a unified data plane that combines an all-flash hot data layer with compute-node local NVMe Tier0 acceleration.

In closed KVCache, GalaxSphere delivered high read and write bandwidth across multiple Llama 3.1 operating modes. The 4-client Llama 3.1 8B storage-only submission reached 286.7 GiB/s read bandwidth and 54.6 GiB/s write bandwidth. In Llama 3.1 8B storage-plus-memory mode, it reached 22.5 GiB/s read bandwidth and 26.3 GiB/s write bandwidth. For Llama 3.1 70B storage-only, the submission reached 308.5 GiB/s read bandwidth and 55.8 GiB/s write bandwidth.

GalaxSphere presents distributed and heterogeneous data from different storage systems, clusters, and business environments through a unified namespace. Applications can continue to use stable file paths while GalaxSphere automatically orchestrates data placement across storage tiers and compute environments according to workload demand, data temperature, lifecycle policies, and task requirements.

The platform uses event-driven and policy-driven automation to support data discovery, classification, prefetching, acceleration, archiving, and governance. This helps AI and HPC teams reduce manual data movement, improve data accessibility, and keep critical datasets close to compute resources for performance-sensitive workloads.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

TTA

The Telecommunications Technology Association (TTA) is a non-profit organization established in 1988, offering testing, certification, and standardization services for Artificial Intelligence (AI).

TTA helps ensure the safety, reliability, and interoperability of AI technologies by validating performance, evaluating compliance, and developing technical standards for trustworthy AI deployment. As a testing and certification body, TTA regularly evaluates server and storage systems under standardized conditions, and recognizes the value of publishing results through platforms like MLCommons.

For this round, TTA submitted results for Seahorse-storage, a software-defined storage system evaluated in our test environment.

Seahorse-storage is built on standard x86 servers and commodity NVMe SSDs, presenting a single shared POSIX namespace to all client nodes. Its data path uses end-to-end RDMA over InfiniBand: an in-kernel client moves data directly between client memory and storage-node memory with zero-copy, one-sided transfers, accelerating both direct and buffered I/O. From the same shared namespace it serves large sequential transfers, such as checkpoint save and restore, alongside highly concurrent small reads, such as vector-index search. Capacity and aggregate throughput scale with additional storage or client nodes, and the metadata service scales out independently to accelerate metadata-intensive workloads with many small files, letting one platform cover training-data feeding, LLM checkpointing, KV-cache offloading, and vector-database search.

The evaluated configuration comprises three storage nodes serving four client nodes over 400 Gb/s NDR InfiniBand. The tests spanned four v3.0 workload families in a single submission: 3D U-Net training, checkpointing for Llama 3.1 70B, KV cache for Llama 3.1 8B, and vector database search on a DiskANN index.

Covering this range in one round showed how a single storage design responds to very different access patterns, and TTA hopes these data points are useful to the MLPerf community.

We believe that open benchmarking and participation in collaborative initiatives such as MLPerf are key to fostering innovation and broad community benefit.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

TuringData

TuringData is a Singapore-based AI data infrastructure company focused on distributed storage and caching for AI workloads. Our systems are designed to provide high-throughput data movement between persistent flash storage and GPU memory, supporting both training data delivery and inference use cases.

For our first MLPerf® Storage v3.0 submission, we submitted benchmark results with TuringData Flash, our all-flash distributed storage appliance built on the TuringData Platform, an all-flash distributed storage appliance built on the TuringData Platform. TuringData Flash is powered by PCIe 5.0 NVMe SSDs and a high-speed 400Gbps network, with support for RoCE and NVIDIA GPUDirect Storage to accelerate AI data access.

Using only a three-node deployment, TuringData Flash delivered strong performance across representative AI training and inference benchmarks:

- 3D U-Net: Up to 531 GiB/s aggregate read bandwidth while supporting 96 simulated accelerators.
- Checkpoint-70B: Up to 307.12 GiB/s write bandwidth and 539.94 GiB/s read bandwidth.
- Checkpoint-8B: Up to 108 GiB/s write bandwidth and 166 GiB/s read bandwidth.
- KVCache: Up to 54 GiB/s write bandwidth and 270 GiB/s read bandwidth.

These results are meaningful because the workloads reflect different stages of the AI pipeline. Training requires sustained data reads, checkpointing introduces bursty writes and recovery reads, and KV cache offload represents inference-related I/O. Evaluating these patterns under a common published methodology helps infrastructure teams better understand storage behavior across end-to-end AI systems, not only in isolated scenarios.

We appreciate MLCommons and the Storage working group for expanding the benchmark suite to reflect how AI infrastructure is evolving. We hope these results provide useful reference points for teams planning storage capacity and data paths across training and inference environments.

This information is under embargo until 9/1/26 8:00AM PT

UBIX

UBIX TECHNOLOGY CO., LTD.

UBIX Delivered Outstanding Performance in the MLPerf® Storage v3.0 Benchmarks

UBIX Technology Co., Ltd. is a high-tech enterprise focused on innovative storage products and solutions. The company independently developed UBIXFS, a high-performance distributed file system with full intellectual property rights, and launched distributed all-flash storage, distributed massive storage, and AI Computing Appliance.

UbiPower 18000 distributed All-flash storage, relying on an innovative two-layer storage pool architecture, RDMA high-speed transmission and storage virtualization, builds a hierarchical, highly efficient resource pool. UBIX's innovative AI storage products have been commercially deployed in multiple supercomputing and AI data centers, demonstrating excellent performance in data preprocessing, massive data access, and large-scale checkpoint read/write operations.

In this MLPerf test, UBIX deployed 3 storage nodes and participated in 3 training tasks, 4 checkpoint tasks, and 3 kvcache tasks, totaling 10 scenarios, achieving outstanding performance in the following 7:

Training:
- RetinaNet: served 688 B200 accelerators
- RetinaNet: served 736 Mi355 accelerators

Checkpoint:
- Llama 3 1T: delivered 461.98 GiB/s of read bandwidth and 293.61 GiB/s of write bandwidth (cluster)
- Llama 3 405B: delivered 459.73 GiB/s of read bandwidth and 293.76 GiB/s of write bandwidth (cluster)

KV Cache:
- Llama 3.1-8B-200USER: delivered 443.67 GiB/s of read bandwidth and 96.95 GiB/s of write bandwidth (cluster)
- Llama 3.1-8B-100USER-CPU-4GB: delivered 57.77 GiB/s of read bandwidth and 64.58 GiB/s of write bandwidth (cluster)
- Llama 3.1-70B: delivered 401.82 GiB/s of read bandwidth and 92.77 GiB/s of write bandwidth (cluster)

In training scenarios, UBIX demonstrated outstanding large-scale GPU concurrency; in checkpoint scenarios, ultra-fast read/write for hundred-billion-parameter models; and in KV

cache scenarios, industry-leading high bandwidth with low latency. This synergistic breakthrough gives UBIX a comprehensive technological moat across the full AI-model lifecycle — training, checkpoint resumption, and inference acceleration.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

XSKY

XSKY AIMesh and MeshFS Deliver High-Performance Data Access for Diverse AI Workloads

XSKY presents its MLPerf Storage v3.0 closed-division results for AIMesh, XSKY's AI data infrastructure for AI workloads.

At its core is MeshFS, an AI parallel file system purpose-built for AI workloads. MeshFS provides a shared POSIX namespace and uses a decentralized distributed metadata architecture to spread metadata processing across the cluster. This parallelizes file-open, directory-lookup, and namespace operations so clients can access million-file training datasets concurrently without a centralized metadata bottleneck. Applications retain familiar file access while MeshFS scales data and metadata services across storage nodes.

The submitted system used three storage nodes with 48 7.68 TB NVMe drives and a RoCE data path. In RetinaNet training, AIMesh supported 675 simulated B200 accelerators, delivered 319,705 samples per second and 98.47 GB/s of training I/O, and maintained 94.0% accelerator utilization. RetinaNet contains 1,170,301 JPEG files, making it a demanding test of file-open rate, metadata processing, random reads, and concurrent access. The results show that a compact MeshFS cluster can supply flagship GPU-class systems with sustained training data.

AIMesh also demonstrated large-model checkpoint performance. In the Llama 3 405B workload, the three-node system delivered 379.9 GiB/s of load bandwidth for 5.29 TB of model state. At the measured transfer rate, reading the checkpoint payload takes about 14 seconds. This reduces data waits during restart, rollback, recovery, and model loading.

For this submission, XSKY combined commercially available servers and NVMe SSDs with integrated hardware and software optimization. This integrated approach helps customers meet AI training and checkpoint requirements while minimizing total cost of ownership and maximizing infrastructure value across computer vision, autonomous driving, robotics, and industrial AI workloads.

ML Commons Supplemental Results Discussion for MLPerf Storage v3.0

This information is under embargo until 9/1/26 8:00AM PT

YanRongTech

YanRong is a leading AI-native storage provider in China, specializing in high-performance storage infrastructure across the entire AI pipeline. Built on a native AI storage architecture with an AI-native distributed file system at its core, YanRong is deeply optimized for large-scale model training, checkpoint acceleration, and long-context LLM inference. We are proud to participate in the MLPerf Storage v3.0 benchmark with our F9000X all-flash storage appliance, demonstrating outstanding performance across both AI training and inference workloads.

Highlights from the 2026 submission:

- Up to 544 GiB/s aggregate read bandwidth in the 3D U-Net benchmark with a three-node storage cluster, supporting 99 simulated accelerators.
- Up to 313.7 GiB/s write bandwidth and 538.7 GiB/s read bandwidth in the Checkpoint-70B benchmark.
- Up to 105 GiB/s write bandwidth and 168 GiB/s read bandwidth in the Checkpoint-8B benchmark.
- Up to 58.8 GiB/s write bandwidth and 283.2 GiB/s read bandwidth in the KVCache benchmark.

These benchmark results demonstrate YanRong's ability to deliver high-performance AI storage across the entire AI lifecycle, meeting the demanding requirements of both large-scale model training and LLM inference.

YanRong enables high-bandwidth, low-latency data access to keep GPUs continuously supplied with data during distributed training while efficiently supporting intensive checkpoint operations. For inference, YanRong extends KV cache beyond GPU HBM memory, overcoming GPU memory capacity constraints for long-context and high-concurrency LLM serving, thereby reducing Time to First Token (TTFT), increasing token throughput, and improving overall token economics.

YanRong remains committed to advancing AI storage innovation, empowering organizations to build faster, more scalable AI infrastructure. From model training to enterprise-scale inference, our solutions improve infrastructure efficiency, accelerate AI innovation, and maximize the value of AI investments.

This information is under embargo until 9/1/26 8:00AM PT

ZettaLane

ZettaLane Systems presents its MLPerf(R) Storage v3.0 closed-division results for MayaNAS and MayaScale, two cloud-native storage engines that ran entirely on standard Google Cloud virtual machines.

MayaNAS is a novel storage substrate presenting parallel and network filesystems (Lustre, pNFS, and NFS), accessed through open-source clients with no proprietary software to install.

Each OST is an OpenZFS dataset on regional, Standard-class Google Cloud Storage buckets: bulk data is read and written directly to object storage, while a small NVMe holds only metadata and small blocks. In MLPerf Storage v3.0, MayaNAS is the only submission running Lustre directly on a cloud object-storage tier. It runs active-active and scales by adding storage-node HA pairs. MayaScale is a companion NVMe-over-TCP disaggregated block engine.

A single approach served training, checkpointing, and inference-cache workloads:

- Checkpointing (MayaNAS): Llama 3 8B reached 14.43 GB/s write and 10.40 GB/s read on one client, scaling with two clients to 32.42 GB/s write and 21.20 GB/s read on Llama 3 70B.
- Training: 3D U-Net held ~92% accelerator utilization on MayaNAS (3 simulated NVIDIA B200) and 93.31% on MayaScale (4 B200); on MayaScale a single client fed 72 simulated B200 for RetinaNet at 86.68%, a high per-client density that kept accelerators and the client network highly utilized.
- KVCache (MayaNAS): Llama 3.1 8B storage-only reached 524.65 tokens per second at 9.14 GB/s read.

For the broader community, these results show economical cloud object storage as the durable tier of a full POSIX parallel filesystem, comfortably serving the AI data pipeline without staging copies in and out The filesystem, buckets, and keys stay in the operator's own cloud account, meeting data-residency and sovereignty requirements. Our solutions are available on the Google Cloud Marketplace and the open-lustre-cloud GitHub project for GCP, Azure, and AWS.