Supplemental Results Discussion for MLPerf Training v6.1

MLCommons Supplemental Discussion

MLPerf Training v6.1 Results Discussion

The submitting organizations provided the following 300-word descriptions as a supplement to help the public understand their MLCommons® MLPerf® Training v6.1 submissions and results. The statements **do not reflect the opinions or views of MLCommons.**

Supplemental Results Discussion for MLPerf Training v6.1

AMD

In MLPerf Training 6.1, AMD delivers an average 19% increase in Llama training performance over round 6.0 on the same eight-GPU AMD Instinct™ MI350X and MI355X platforms. Measured across Llama 2 70B LoRA and Llama 3.1 8B, these gains demonstrate continued performance progress on existing hardware.

This round combines that progress with expanded workload coverage and partner results across cloud and server systems. AMD Instinct GPUs, AMD ROCm™ software and AMD Primus support submissions spanning language model training, fine-tuning, image generation and recommendation systems. These results demonstrate continued progress as AMD advances ROCm software optimization, supporting shorter training cycles and faster experimentation.

On FLUX.1, AMD delivers a 3.1X generational performance gain from round 6.0 on 64 AMD Instinct MI325X GPUs to round 6.1 on 64 AMD Instinct MI355X GPUs. This reduces benchmark training time by approximately 68%, demonstrating substantial progress in multi-node image-generation training at the same GPU count.

New GPT-OSS-20B and DLRMv4 results extend coverage submitted by AMD beyond Llama and FLUX.1. These additions broaden the evidence for AMD training capabilities, including recommendation systems relevant to ranking and personalization.

Six partners submitted results using AMD Instinct GPUs: Crusoe, Dell Technologies, HPE, MiTAC, Oracle and Supermicro. Across comparable eight-GPU MI355X submissions on Llama 2 70B LoRA and Llama 3.1 8B, partners achieved an average 94% of AMD reference performance, reinforcing reproducibility across partner systems. Crusoe demonstrated multi-node training with configurations reaching 128 MI355X GPUs.

Together, these results show AMD advancing training performance, workload breadth and ecosystem confidence. AMD Primus and ROCm software provide configurations and recipes to help teams evaluate and reproduce benchmark results. From single-node fine-tuning to multi-node training, AMD continues to strengthen the hardware, software and partner foundation for building and adapting AI models across cloud and on-premises environments.

Supplemental Results Discussion for MLPerf Training v6.1

ASUSTeK

ASUS Submits MLPerf® Training v6.1 Results in the Closed Division Across two Workloads

[Taipei, 2026/11/16] – ASUS today announced its results in the MLPerf Training v6.1 benchmark round from MLCommons®. Submitted in the Closed Division by the ASUS server engineering team, the results cover two large language model workloads: Llama 3.1 8B, and Llama 2 70B LoRA. All results were achieved on **ASUS XA NB3I-E12 with 8× B300 GPU** and dual Intel® Xeon® 6 Scalable processors.

Workload	**Result (time-to-train, minutes)**
Llama 3.1 8B	71.320
Llama 2 70B LoRA	6.570

Because the Closed Division requires submitters to follow the same model and training rules, these results reflect the strength of the underlying system and software stack. Delivering results across three models of different sizes and training characteristics required tuning across the full stack.

MLPerf Training is an industry-standard benchmark suite from MLCommons, an open engineering consortium. With common rules and peer-reviewed submissions, it gives the industry a shared, reproducible reference for evaluating AI training systems. As an MLCommons member, ASUS supports this open and transparent approach, and sees benchmark results as one useful input, alongside workload-specific testing, for organizations planning AI infrastructure.

This work also supports ASUS's AI Factory approach, which brings together servers, software, and services to help organizations build and deploy AI workloads.

Full results are available at MLCommons MLPerf Training Benchmark.

Supplemental Results Discussion for MLPerf Training v6.1

Cisco

Cisco's MLPerf Training v6.1 submission used eight Cisco UCS C880A M8 servers, each with eight NVIDIA B300 GPUs, connected through SONiC-based Cisco 8000 switching with Spectrum technology.

The system achieved:

- **Six benchmark families on one 64-GPU Cisco UCS system**. One configuration completed six of the round's seven benchmark families spanning pretraining, fine-tuning, reinforcement-learning post-training, and recommendation. The results demonstrate the versatility of a common compute and network architecture as enterprise AI programs adopt different training methods and model types.
- **The smallest DeepSeek-V3 training cluster in the round**. The cluster completed the DeepSeek-V3 671B training benchmark on 64 GPUs in 163.19 minutes. This gives enterprises a concrete example of distributing a frontier-scale mixture-of-experts model across eight-GPU server building blocks. Cisco 8000 switching carried activations and gradients between UCS servers, connecting the distributed training stages.
- **A fabric architecture designed for growth**. The quad-plane RDMA over Converged Ethernet (RoCE) design connects each GPU through four 200 Gb/s Ethernet paths, preserving 800 Gb/s of aggregate bandwidth per GPU and 6.4 Tbps of nominal one-direction capacity per server. Lower-speed connections increase the number of GPU interfaces each switch can support, providing a fabric pattern for larger clusters.

Cisco Products Behind the Results

Cisco UCS C880A M8: An eight-GPU enterprise AI building block designed for large-scale model training and other demanding accelerated workloads.

Cisco 8000 switching: The Ethernet foundation connecting every UCS server across four independent planes, combining open network software, high bandwidth, and path diversity for distributed training.

Together, the results provide enterprises with a documented reference configuration for planning capacity in repeatable Cisco UCS server units and designing the Cisco 8000 fabric for both cluster growth and changing workload requirements.

Supplemental Results Discussion for MLPerf Training v6.1

CoreWeave

CoreWeave demonstrates multi-rack NVIDIA Vera Rubin benchmark for MLPerf® Training v6.1

CoreWeave submitted MLPerf® Training v6.1 results on NVIDIA Vera Rubin NVL72 and NVIDIA GB200 NVL72 systems, spanning single rack to multi-rack configurations from 72 to 256 GPUs. Each speedup below compares a CoreWeave's NVIDIA Vera Rubin NVL72 result with CoreWeave's own NVIDIA GB200 NVL72 result at the same GPU count. CoreWeave Mission Control, together with full-stack optimizations across networking, workload orchestration, system software, and cluster operations, delivered the speed, scale, and reliability needed throughout the benchmarks.

DeepSeek-V3-671B benchmark trained in 17.55 minutes on 256 NVIDIA Rubin GPUs

CoreWeave trained DeepSeek-V3-671B benchmark to target quality in just 17.65 minutes on 256 NVIDIA Rubin GPUs, demonstrating a 2.7x generational speedup of Coreweave's NVIDIA Vera Rubin NVL72 over CoreWeave's 256-GPU NVIDIA GB200 NVL72 submission. CoreWeave was the only cloud provider to submit results on this benchmark with NVIDIA Rubin. The Vera Rubin racks were in a production environment, connected by a RoCE-based NVIDIA Spectrum-X™ Ethernet AI fabric, with NVIDIA ConnectX™-9 SuperNICs.

Llama 3.1 8B benchmark training in 7.16 minutes on 72 NVIDIA Rubin GPUs

Llama 3.1 8B benchmark reached target quality in 7.16 minutes on Coreweave's NVIDIA Vera Rubin NVL72, 1.9x faster than CoreWeave's 72-GPU NVIDIA GB200 NVL72 submission, using NVFP4 GEMMs, FP8 attention, and full NVIDIA CUDA™ graphs.

Customers train their frontier-scale models on CoreWeave in the same environment so these results reflect our production setup. These submissions show the value of CoreWeave's purpose-built AI cloud: tightly integrated compute, high-performance networking, intelligent orchestration, deep observability, and rigorous cluster validation engineered to keep infrastructure reliable and productive at scale.

Supplemental Results Discussion for MLPerf Training v6.1

Crusoe

Crusoe Cloud in MLPerf Training v6.1: MI355X, B300, GB300

Crusoe's MLPerf® Training v6.1 submissions span three platforms: AMD Instinct MI355X, NVIDIA B300, and NVIDIA GB300 NVL72. The results cover Llama 3.1 8B pretraining, GPT-OSS 20B, Llama 2 70B LoRA, FLUX.1, and DLRMv4, a broad set of workloads that reflect what AI teams are building today.

Excellent DLRMv4 performance on MI355X

A single MI355X node trained DLRMv4 in 18.8 minutes. Scaling to 64 GPUs across eight nodes on a standard RoCE Ethernet fabric brought training time down to 4.6 minutes. These results show that teams building recommendation systems can get great scaling on open, widely available networking.

Scaling MI355X on standard Ethernet

Crusoe also submitted the largest MI355X configuration in the round, and is the only submitter with multi-node MI355X results on Llama 3.1 8B and GPT-OSS 20B, scaling Llama 3.1 8B to 128 GPUs across 16 nodes over standard RoCE Ethernet.

Full-rack GB300 NVL72 and 128-GPU B300

On a full GB300 NVL72 rack (72 GPUs), Crusoe completed FLUX.1 in 23.6 minutes and Llama 2 70B LoRA in 1.2 minutes, demonstrating great results for image generation and fine-tuning workloads. On B300, GPT-OSS 20B scaled from one node to 16 nodes (128 GPUs), showing consistent scaling as customers grow their clusters.

Benchmark results start with fleet health

Every result was produced on virtualized Crusoe Cloud instances, the same ones customers provision today. Results like these start long before benchmark day: every Crusoe cluster passes comprehensive burn-in before delivery, and continuous fleet health monitoring keeps it performing throughout its production life. Crusoe Managed Kubernetes, Managed Slurm, and fault-tolerant AutoClusters remove the operational overhead of running AI at scale.

Crusoe is the energy-first AI factory company, designing, building, and operating the AI factories, energy, and cloud platform behind next-generation AI. Every configuration submitted here is available today at crusoe.ai.

Supplemental Results Discussion for MLPerf Training v6.1

Dell Technologies

Dell Technologies' MLPerf Training v6.1 submission is our broadest yet: 16 systems and 35 benchmark results across five accelerators (NVIDIA GB300, B300, B200 and H200 NVL, plus AMD Instinct MI355X). They run on Dell PowerEdge XE9712, XE9780, XE9785 and XE7740 servers, scaling from a single node to 72 GPUs.

Results span four industry-standard workloads: Llama 3.1 8B pretraining, GPT-OSS 20B, Llama 2 70B LoRA fine-tuning, and Flux.1 image generation. Together they reflect the range of language, generative and recommender workloads customers run today.

Scale-out is where PowerEdge shines. On GB300 systems, Llama 2 70B LoRA fine-tuning completes in just 1.29 minutes across 72 GPUs. On 64 GPUs, Llama 3.1 8B pretraining finishes in 12.7 minutes and GPT-OSS 20B in 18.59 minutes. Scaling from one node to two delivers near-ideal gains: 1.94x on Llama 3.1 8B and 1.96x on Llama 2 70B LoRA.

Single-node performance is equally compelling. Our eight-GPU B300 servers train Llama 2 70B LoRA in 6.6 to 6.9 minutes across multiple XE9780 and XE9785 variants, and complete DLRMv4 in under 34 minutes. Customers get consistent, high-end results whether they choose Intel Xeon 6 or AMD EPYC host platforms.

On the AMD side, PowerEdge XE9785 and XE9785L systems with eight MI355X GPUs, running ROCm 7.0.1, train Llama 3.1 8B in roughly 75 to 78 minutes and Llama 2 70B LoRA in under 8 minutes, giving customers real choice in accelerator architecture.

These results reinforce Dell's commitment to open, vendor-flexible AI infrastructure that is performant, scalable and production-ready, from compact single-node servers to rack-scale clusters.

Supplemental Results Discussion for MLPerf Training v6.1

Firmus

Firmus submitted results to MLPerf® Training v6.1, demonstrating the software engineering capabilities being developed as part of Firmus AI Factory OS, the software and intelligence layer powering its AI Factories.

The submissions covered GPT-OSS-20B and Llama 3.1 8B, using eight NVIDIA GB300 GPUs connected through fifth-generation NVLink on its Supermicro GB300 NVL72 platform.

This round focused on software optimization. Building on NVIDIA's reference recipes, Firmus engineers developed optimised GPU kernels spanning attention, normalisation, multi-layer perceptron (MLP), mixture-of-experts (MoE), low-precision quantization, weight and gradient communication, and optimiser operations.

Firmus engineers combined AI-driven kernel generation with automated performance testing and mathematical validation, enabling rapid evaluation of alternative implementations. This approach identifies computational bottlenecks and removes unnecessary operations while maintaining numerical correctness and model convergence.

For GPT-OSS-20B, Firmus achieved a training time of 52.90 minutes using NVFP4-optimised kernels, a 24% reduction compared with our own baseline of 69.60 minutes, run with the unmodified reference recipe on the same eight GPUs. For Llama 3.1 8B, NVFP4 kernel optimisations achieved 52.40 minutes, a 14% reduction compared with our own baseline of 60.88 minutes.

These results reflect Firmus's approach to engineering a vertically integrated AI Factory, where continuous improvements across software, compute and physical infrastructure work together to optimise performance from model-to-grid.

Supplemental Results Discussion for MLPerf Training v6.1

Fujitsu

Fujitsu offers a fantastic blend of systems, solutions, and expertise to guarantee maximum productivity, efficiency, and flexibility, delivering confidence and reliability. Since 2020, we have been actively participating in and submitting inference and training rounds for both data center and edge divisions.

We offer PRIMERGY CDI series, which utilizes Composable Disaggregated Infrastructure (CDI). PRIMERGY CDI stands apart from traditional server products, comprising computing servers, PCIe fabric switches, and PCIe boxes. Device resources such as GPUs, SSDs, and NICs are stored externally in PCIe boxes rather than within the computing server's chassis.

Our purpose is to make the world more sustainable by building trust in society through innovation. With a rich heritage of driving innovation and expertise, we are dedicated to contributing to the growth of society and our valued customers. Therefore, we will continue to meet the demands of our customers and strive to provide attractive server systems through the activities of MLCommons.

Supplemental Results Discussion for MLPerf Training v6.1

GigaComputing

Giga Computing is the enterprise division (and a subsidiary) of GIGABYTE, responsible for designing, manufacturing, testing and selling GIGABYTE server products. GIGABYTE is a founding member of MLCommons and has contributed high-performance systems to each round of MLPerf Training and MLPerf Inference. By participating, Giga Computing aims to give academia, global technology providers and other organizations a clearer picture of what today's systems can achieve, helping them make better-informed decisions.

For MLPerf Training 6.1, GIGABYTE systems were selected to reflect the choices of cooling technologies for accelerated computing systems in the market today. For this round of testing, systems were submitted using NVIDIA B300 GPU, paired with 120 core Intel Xeon 6 CPUs (liquid cooled system), and 64 core AMD EPYC 9005 CPUs (air cooled system).

Learn More: https://www.gigabyte.com/Enterprise and https://www.gigacomputing.com/en/

Supplemental Results Discussion for MLPerf Training v6.1

Google

Google is excited to present DeepSeek-V3 (671b) pretraining and new Qwen3.5-397b Reinforcement Learning (RL) results on Ironwood (TPU7x), our seventh-generation TPU, and A4X and A4X Max instances accelerated by NVIDIA GB200 NVL72 and GB300 NVL72 systems.

On DeepSeek-V3 pretraining, Ironwood reached target quality in 1.70 minutes on 8,192 chips in one inter-chip interconnect (ICI) domain, our largest TPU submission to date. At 2,048 and 4,096 chips, training took 5.03 and 2.62 minutes (~96% scaling efficiency to 4,096; 77% from 4,096 to 8,192). Runs used MaxText's mixed-precision implementation (FP8 expert and attention-projection matrix multiplications, BF16 core attention, FP32 master weights) with a batch-split schedule overlapping one half's expert communication with the other's compute. Timed-run effective model FLOPs utilization (EMFU) against Ironwood's peak BF16 was 56%, 54%, and 42% at 2,048, 4,096, and 8,192 chips.

For RL, Ironwood reached target quality in 18.02 minutes on 640 chips: a 128-chip (4x4x8) trainer slice and 128 4-chip sampler engines. Built from scratch for long-horizon agentic workloads, our new asynchronous, decoupled RL stack combines MaxText training, vLLM with tpu-inference for high-throughput sampling, Google Kubernetes Engine (GKE) agentic sandboxing, and TPU-sync weight synchronization.

On A4X and A4X Max, running NVIDIA's published FP8 recipes with Cluster Director (Slurm) and Managed Lustre across multiple domains over RoCE, 256 GB200 GPUs trained RL in 21.03 minutes and 256 GB300 GPUs trained DeepSeek-V3 in 31.96 minutes (63% EMFU against GB300 peak BF16 on 256 GPUs), with single NVL72 domain RL runs at 56 and 64 GPUs.

These submissions bring the best of Google Cloud TPUs and GPUs to an open-source stack: JAX, MaxText, Tunix, vLLM, and NVIDIA software. We thank MLCommons for introducing the agentic RL benchmark and providing rigorous, peer-reviewed performance data.

Supplemental Results Discussion for MLPerf Training v6.1

HPE

HPE reinforced its diverse portfolio for large-scale AI pretraining in MLPerf v6.1, spanning NVIDIA and AMD accelerators across direct liquid-cooled (DLC) and air-cooled configurations. The results offer customers strong proof points on flexible scale, accelerator, and cooling options optimized for model training, natural language processing, and multimodal training.

HPE rack-scale AI systems continued momentum from its MLPerf v6.0 results on the NVIDIA GB300 NVL72 by HPE with improved pretraining performance for large language models (LLMs) and reasoning models, including Llama3.1-8B pretraining in under 12 minutes when using 64x NVIDIA GB300 superchips. Each rack consumes approximately 132 kW to 140 kW rack of power, with 90% liquid cooled and 10% air cooled components1. This design allows for sustained peak performance of trillion parameter models, while lowering overall cooling and power costs.

HPE ProLiant Compute XD685 demonstrated efficient DLC performance and flexible 8-way accelerator choice with two configurations, including one with NVIDIA Blackwell Ultra (B300) 288GB GPUs and the other with AMD InstinctTM MI355X 288GB GPUs. Both configurations use 5th Generation AMD EPYCTM processors. This is HPE's first submission with the HPE ProLiant Compute XD685 and NVIDIA Blackwell Ultra (HGXTM B300) 288GB GPUs with DLC. Both configurations are designed to maintain peak AI pretraining performance in DLC environments, while lowering energy consumption and power costs. This is demonstrated by pretraining Llama3.1-8B with 8x AMD Instinct MI355X GPUs in 76.41 minutes and with 8x NVIDIA B300 in 77.39 minutes compared to 78.25 minutes using 8x air-cooled NVIDIA B300 GPUs.

HPE Compute XD690 continued momentum with improved pretraining performance for LLMs and reasoning models using Intel® Xeon® 6 processors and NVIDIA Blackwell Ultra (HGXTM B300) 288GB air-cooled accelerators. Both single-node and multi-node configurations were included in this round to demonstrate scale-up performance. This configuration is a strong choice for scaling up AI pretraining performance in air-cooled environments.

1 Accelerate AI at scale with NVIDIA GB300 NVL72 by HPE

Supplemental Results Discussion for MLPerf Training v6.1

Inventec

Inventec Powers Large-Scale AI Training with NVIDIA HGX B300 in MLPerf Training v6.1

Inventec Corporation highlighted its engineering excellence in the MLPerf Training v6.1 benchmarks, delivering powerful results with its NVIDIA HGX B300 8-GPU server portfolio powered by dual Intel Xeon 6 processors:

● **P5800G7 (5U, Liquid-Cooled):** High-density direct liquid cooling with integrated ConnectX-8 SuperNICs for maximum cluster training throughput (P5800G7 Specs).

● **P9000G7 (10U, Air-Cooled):** Precision air cooling with modular architecture and high-bandwidth NVLink interconnects for reliable deep learning performance (P9000G7 Specs).

These benchmark-proven platforms deliver top-tier AI training performance, giving enterprise customers versatile liquid- and air-cooling choices for modern LLM and deep learning infrastructure.

About Inventec

Inventec specializes in server design and AI hardware innovation. Learn more at ebg.inventec.com.

Supplemental Results Discussion for MLPerf Training v6.1

KRAI

Founded in 2020 in Cambridge, UK (“The Silicon Fen”), KRAI is a purveyor of premium benchmarking and optimization solutions for AI systems. KRAI is proud to be a Founding Member of MLCommons. This MLPerf round is KRAI’s 25th round since 2021 (9x Training + 12x Inference + 3x Endpoints + 1x Tiny), a record only matched by NVIDIA.

In this v6.1 round, we applied systematic agentic engineering to optimize Llama2-70B-LoRA finetuning on a Dell PowerEdge XE9680 server with 8x AMD Instinct MI300X GPUs. In v6.0, we achieved a time-to-train of 28.65 minutes on the same server using AMD’s optimized implementation from v5.1. Starting from the same implementation, we reduced the time-to-train to 22.83 minutes: a 20% improvement.

We achieved this by optimizing the software stack, assisted by Claude Opus 5.5 agents. Most notably, AMD’s implementation saves memory by discarding the activations of some layers during the forward pass and recomputing them during the backward pass. By storing activations in FP8 and packing dropout masks, we freed enough GPU memory to keep all activations, eliminating recomputation entirely. This alone cut the training step time by 12%. We also fused the LoRA adapter computation into the epilogue of the base GEMM kernel, and performed other improvements.

This shows that substantial headroom remains even in vendor-optimized implementations, and that rigorous optimization can unlock significant extra performance from long-deployed hardware. We are now generalizing these techniques in The Collective, our product for automatically optimizing AI infrastructure using specialized cooperating agents.

Finally, we submitted results of benchmarking Isambard AI, the UK’s National AI Research Resource, still the only results from any Sovereign AI infrastructure.

We cordially thank our partners Dell Technologies and Hewlett Packard Enterprise for their long-term support, which allows us to continue pushing the boundaries of performance benchmarking and optimization.

Lambda

About the Benchmarks: NVIDIA GB300 NVL72 on Lambda AI Cloud

Supplemental Results Discussion for MLPerf Training v6.1

Lambda partnered with NVIDIA for the MLPerf® Training v6.1 round. We benchmark two different workloads on a bare-metal Lambda Cloud cluster with 64 NVIDIA Blackwell Ultra GPUs (GB300-SXM-288GB).

Our NVIDIA GB300 GPT-OSS-20b runs converge 7.15% faster than our previous best MLPerf Training result on a larger scale (72 GB300s) with the same accelerator (17.046 minutes vs. 18.359 minutes). Those gains confirm NVIDIA's latest hardware and software improvements on Lambda's bare-metal infrastructure.

These results were produced on the same bare-metal infrastructure available to Lambda customers: teams training frontier-scale models can expect similar production performance. Lambda supports orchestration through Kubernetes or Slurm, including managed options on 1-Click Clusters.

About Lambda
Lambda, The Superintelligence Cloud, is a leader in AI cloud infrastructure serving tens of thousands of customers. Our customers range from AI researchers to enterprises and hyperscalers. Lambda's mission is to make compute as ubiquitous as electricity and give everyone the power of superintelligence. One person, one GPU.

Supplemental Results Discussion for MLPerf Training v6.1

MiTAC

It is with great honor that MiTAC, an established server platform designer, manufacturer, and a subsidiary of MiTAC Holdings Corporation (TWSE:3706), announces its performance results from its internal validation utilizing the MLPerf® Training v6.1 benchmark suite. These results validate our commitment to delivering cutting-edge AI infrastructure that enhances compute density and scalability. Our G8825Z5 and G4826Z5 AI/HPC server series have demonstrated their high capability, successfully running a comprehensive suite of demanding generative AI (GenAI) and Large Language Model (LLM) architectures. The readiness of our AI portfolio is built upon purpose-built hardware and system designs, showing validation across open-source benchmarks:

• **MiTAC G8825Z5 (8U 8x AMD Instinct™ MI350X GPUs HPC/AI Server):** Built specifically for massive scale-out AI training and high-throughput LLM operations, this system pairs high-performance AMD EPYC™ 9575F or 9755 processors with 8x AMD Instinct™ MI350X accelerators, featuring the 4th Gen CDNA™ architecture and 288GB of HBM3e memory per GPU. During our internal MLPerf® Training v6.1 evaluation runs, the platform exhibited robust adaptability and stable execution across reasoning and text generation tasks by successfully completing:

- llama2_70b_lora

• **MiTAC G4826Z5 (4U 8x AMD Instinct™ MI355X Liquid-Cooled AI Training & Inference Server)**: Tailored for high-density, space-conscious data centers that demand sustained computing power free from thermal throttling, this 4U chassis integrates sophisticated liquid cooling mechanics to guarantee peak thermal efficiency. It combines dual AMD EPYC™ 9575F or 9755 processors alongside 8x AMD Instinct™ MI355X accelerators, achieving comprehensive validation through the successful execution of an expanded three-workload GenAI matrix:

- llama2_70b_lora

Supplemental Results Discussion for MLPerf Training v6.1

Nebius

For MLPerf® Training v6.1, Nebius submitted Closed division results for two NVIDIA systems, the Nebius GB300 n16 (16x NVIDIA GB300 NVL4) and the Nebius VR200 n16 (16x NVIDIA VR200 NVL4). The GB300 entry is in the Available cloud category, the VR200 entry in Available on-premise. The benchmarking covered two workloads, GPT-OSS-20B and Qwen3.5-397B GRPO.

Each system ran as a 16-node cluster with 64 GPUs. The GB300 system paired 64 NVIDIA Blackwell Ultra GPUs with 32 Neoverse-V2 host processors. The VR200 system paired 64 NVIDIA Rubin GPUs with 32 NVIDIA Vera CPUs at a 2,300 W TGP per GPU.

Nebius tests, tunes, and optimizes its infrastructure to ensure the latest NVIDIA hardware platforms deliver high performance in a cloud environment. This work spans multiple layers of the AI infrastructure, from thorough server acceptance testing to cloud software optimization and multi-node cluster validation. These efforts help ensure stable performance and consistent system behavior across different hardware platforms, deployment sizes, and workload types.

The submitted results reflect this ongoing work. The two systems ran identical workloads at matching node and GPU counts, so the results show two NVIDIA platforms completing the same training work on Nebius infrastructure. Across all four runs, the systems achieved high GPU utilization and stable throughput. The outcomes confirm that Nebius' virtualized environment sustains efficiency close to bare-metal configurations, which requires deliberate infrastructure work to achieve and maintain at scale.

The MLPerf® Training benchmarks from MLCommons provide a standardized framework for evaluating AI system performance. Nebius' results contribute to this effort and highlight the readiness of its infrastructure to support demanding AI training workloads on NVIDIA GB300 and VR200 systems at 16-node scale.

Supplemental Results Discussion for MLPerf Training v6.1

NVIDIA

In MLPerf Training v6.1, NVIDIA submitted its first Vera Rubin NVL72 results, delivering excellent performance across the complete benchmark suite. Covering language model pretraining and fine-tuning, agentic reinforcement learning, recommender systems, and image generation, the Vera Rubin NVL72 and GB300 NVL72 results highlight NVIDIA's ability to accelerate diverse AI training workloads across thousands of GPUs with efficient scaling.

Vera Rubin NVL72 delivered up to 2x faster training than NVIDIA's GB300 NVL72 at matching GPU counts. The platform also demonstrated nearly linear scaling on DeepSeek-V3 671B, achieving approximately 97% scaling efficiency as GPU count increased from 256 to 1,024. With four times as many GPUs, training time fell from 16.12 to 4.14 minutes — a 3.9x speedup — illustrating how Vera Rubin turns additional GPU resources into shorter training times and faster experimentation for mixture-of-experts workloads. These early results highlight NVIDIA's accelerated pace of innovation, with further performance improvements expected through continuous software optimization.

NVIDIA and Azure made the largest Blackwell-based submissions this round, scaling DeepSeek-V3 671B training across 10,240 GPUs using GB300 NVL72 and GB200 NVL72 systems, respectively.

These results reflect how NVIDIA's extreme co-design of accelerated computing, NVLink fabric, InfiniBand and Spectrum-X scale-out networking, and optimized training software helps developers shorten training times.

NVIDIA ecosystem partners participated extensively, with submissions from 20 organizations: ASUSTeK, Azure, Cisco, CoreWeave, Crusoe, Dell Technologies, Firmus, Fujitsu, GigaComputing, Google, Hewlett Packard Enterprise, Inventec, KRAI, Lambda, Nebius, Oracle, Quanta Cloud Technology, Supermicro, TTA, and the University of Florida. These submissions illustrate how NVIDIA's programmable platform with open-source libraries and frameworks enables partners to optimize AI training performance and bring their own innovations to the ecosystem.

We commend MLCommons for its work advancing AI benchmarking best practices and providing the industry with reliable, peer-reviewed performance data.

Supplemental Results Discussion for MLPerf Training v6.1

Oracle

Oracle submitted results to the MLPerf® Training v6.1 benchmark suite across language model pretraining and fine-tuning, agentic reinforcement learning, and image generation. The submission included NVIDIA GB300 systems at multiple GPU scales and an eight-GPU AMD Instinct MI355X configuration, demonstrating OCI's support for diverse training workloads and deployment scales.

Oracle's submissions offer insight into training performance across the tested configurations, including how performance changes as NVIDIA GB300 GPU counts increase. These results help customers evaluate infrastructure options against their model development, training performance, and scale requirements.

Beyond the benchmarked configurations, OCI offers a broad portfolio of infrastructure for AI training and inference. NVIDIA options include A10, A100 80GB, H100, H200, B300, GB200, GB300, and Vera Rubin NVL72 platforms. OCI also offers AMD Instinct MI300X and MI355X accelerators for customers developing and deploying open-weight and proprietary models. Deployment options range from individual instances to large GPU clusters, depending on the platform and regional availability. OCI X12 Standard Acceleron Compute instances, powered by Intel Xeon 6 processors, complement this GPU portfolio with CPU-based options for suitable inference and supporting application workloads.

Generative AI workloads place distinct demands on compute, networking, and storage. OCI addresses these requirements through an infrastructure stack designed for distributed AI workloads, combining accelerated compute with high-bandwidth, low-latency cluster networking that supports remote direct memory access (RDMA). OCI File Storage with Lustre, a fully managed parallel file system, provides high-throughput access to training data and model checkpoints. Together, these capabilities are designed to reduce communication and data-access bottlenecks during model development and training.

Oracle's participation in MLPerf® Training v6.1 reflects its commitment to transparent, standardized performance measurement. Combined with OCI's accelerator portfolio and distributed computing capabilities, these submissions help organizations assess infrastructure choices for developing and training AI models across a range of workloads and deployment scales.

Supplemental Results Discussion for MLPerf Training v6.1

Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global provider of data center and AI infrastructure solutions, announced its participation in the MLPerf™ Training v6.1 benchmark submissions in the Data Center Closed division, reaffirming its commitment to delivering high-performance, scalable AI infrastructure for next-generation AI workloads.

QCT's submission features the NVIDIA GB300 NVL72 rack-scale platform, integrating 18 QCT QuantaGrid D75U-1U servers powered by the NVIDIA Grace Blackwell Ultra architecture. Each server incorporates dual NVIDIA GB300 Grace Blackwell Ultra Superchips, while the rack-scale platform combines NVIDIA Blackwell Ultra GPUs, NVIDIA Grace CPUs, NVIDIA ConnectX®-8 SuperNICs, and NVIDIA BlueField®-3 DPUs. Featuring NVIDIA NVLink™ Switch technology and a fully liquid-cooled design, the platform provides a high-density computing architecture engineered for demanding AI workloads.

For MLPerf Training v6.1, QCT submitted results using 8, 16, 64, and 72 GPUs within the NVL72 architecture, tailoring GPU configurations to the computational requirements of individual AI models. This approach demonstrates the platform's flexibility in accommodating diverse AI training workloads while emphasizing the importance of workload-aware resource allocation.

For computationally intensive AI training workloads, including large language model (LLM) pretraining and post-training, the NVIDIA GB300 NVL72 architecture combines substantial computing capacity, high-bandwidth GPU interconnectivity, and large GPU memory capacity to support increasingly complex AI models. Beyond training, the platform is also designed to address the growing computational demands of large-scale AI reasoning and inference.

QCT remains committed to delivering high-performance, scalable, and workload-optimized AI infrastructure through continuous innovation and industry-standard performance validation. By leveraging transparent benchmark results, QCT aims to help enterprises, cloud service providers, and research organizations make informed, data-driven decisions and accelerate their AI infrastructure deployments.

Supplemental Results Discussion for MLPerf Training v6.1

Supermicro

Supermicro is a global technology leader committed to delivering first-to-market innovation for AI, Cloud, Storage, and 5G/Edge IT Infrastructure. We are a Rack-Scale Total IT Solutions provider that designs and builds an environmentally friendly, energy-efficient range of servers, storage systems, and switches. We also offer a comprehensive portfolio of software and global support services.

We rapidly incorporate the latest technological innovations, NVIDIA GPU, NVL72 Vera Rubin, AMD, Intel, ARM technologies, NVMe, PCIe, networking, into our products. We work closely with technology partners to deliver the latest generation of cutting-edge solutions across server, storage, and networking platforms.

We deliver AI systems and pods, directly or through partners, to end customers. For MLPerf training v6.1, Supermicro submitted an NVL72 Vera Rubin submission, demonstrating it’s time to market. Supermicro also submitted 16-B300 and 8xMI355X submissions. Furthermore, Supermicro had multiple partners who submitted Supermicro systems for the training benchmarks.

Supplemental Results Discussion for MLPerf Training v6.1

Tinycorp

tiny corp maintains tinygrad, a from-scratch, backend-agnostic neural network library that breaks down the most complex networks into a set of basic operations, which can then be highly optimized for various hardware accelerators.

For the latest MLPerf Training v6.1 benchmark round, tiny corp submitted Llama 3.1 8B pretraining results on single and dual node AMD MI350X machines in addition to GPT-OSS 20B pretraining results on single and dual node AMD MI350X machines. This submission leveraged tinygrad with its python-based userspace driver for the MI350X, which delivered a performance uplift over the kernel driver through reduced overhead and more control over the hardware.

tiny corp will continue to push the envelope of machine learning with a focus on democratizing access to high performance compute.

Supplemental Results Discussion for MLPerf Training v6.1

TTA

Telecommunications Technology Association (TTA) is a government-affiliated testing and certification body in the Republic of Korea. We operate shared AI computing infrastructure and help domestic organizations prepare for internationally recognized benchmark certification. In this round we submitted a Closed-division result for Llama 3.1 8B together with a power measurement.

The system is a single server with four NVIDIA H100 NVL GPUs connected over NVLink. Power was sampled at the power-supply level once per second across all ten runs, and the reported energy follows the MLCommons power methodology, including the AC-to-DC conversion factor defined by the rules.

ExaStor-MT is a parallel file system for AI training workloads, with an architecture designed to support parallel data access, memory pool burst buffering, workload-aware prefetching, and multilevel tiering. ExaStor-MT can be deployed directly on compute nodes in a hyperconverged infrastructure (HCI) configuration, integrating computing and storage within the same cluster. In this submission it ran on the training server itself. Its design supports efficient dataset loading, data placement, and checkpoint handling throughout training. A stackable architecture enables integration with heterogeneous storage backends, while unified management provides visibility into storage resources and I/O activity.

We submitted with power because energy is becoming as meaningful a dimension of AI infrastructure as time. Producing a measured energy figure under a common methodology required instrumentation, logging and verification practice that we can now reuse and pass on to the organizations we support.

We thank the MLCommons Training working group, and the reviewers whose questions sharpened our methodology and documentation during this round.

TTA intends to keep participating in MLPerf Training and to expand measured-energy reporting in future rounds, so that more organizations in Korea can evaluate and present AI system performance on an internationally comparable basis.

Supplemental Results Discussion for MLPerf Training v6.1

University of Florida

The University of Florida (UF) continues its commitment to advancing AI research and education through its second MLPerf Training submission. Our v6.1 results evaluate NVIDIA DGX Blackwell 200 (B200) SuperPOD and Blackwell 300 (B300) systems on HiPerGator, UF's supercomputing cluster, across a broad range of training workloads. These results demonstrate that HiPerGator supports demanding AI model training workloads and provide an academic computing perspective on transparent, reproducible performance measurement.

UF ran all seven MLPerf Training v6.1 benchmarks: Llama-3.1-8B, FLUX.1, Llama-2-70B-LoRA, GPT-OSS-20B, DeepSeek-V3-671B, Qwen3.5-397B-GRPO and DLRMv4, using NVIDIA's MLPerf containers lightly adjusted for our environment. To evaluate training scalability and parallel strategies, we ran each benchmark on the B200 from its smallest feasible node configuration up to 32 nodes (256 GPUs), producing full scaling curves across the benchmark suite. Smaller workloads, such as Llama-3.1-8B, Llama-2-70B-LoRA and DLRMv4, show strong multi-node scaling. Of the large models, Qwen3.5-397B-GRPO sustains distributed training efficiency from 16 to 32 nodes, and DeepSeek-V3-671B trains successfully at 32 nodes. These outcomes confirm HiPerGator can effectively manage frontier-scale AI training and post-training under realistic, shared HPC constraints. UF also submitted single-node (eight GPUs) B300 results for four benchmarks that fit within one node, expanding coverage across accelerator platforms.

UF ran benchmarks on only half of the B200 SuperPOD (32 nodes) while the other half continued serving daily workloads from active HiPerGator users. This demonstrates HiPerGator's stability and effective resource isolation in a multi-tenant production environment. Our workflow uses Apptainer containers, SLURM scheduling and a parallel Lustre file system. Containers run securely without root privileges, which keeps the closed-division software stack intact and gives other institutions a practical, easy-to-adopt way to run compliant AI workloads on shared academic infrastructure. Through continued participation in MLCommons, UF shares results and operational experience that support trustworthy performance evaluation.