# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 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® Training v6.0 submissions and results. The statements **do not reflect the opinions or views of MLCommons.**

**This information is under embargo until the public release at 6/16/26 8:00AM PT**

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# AMD

**MLPerf 6.0 Training Submission Statement**

For the MLPerf 6.0 Training round, AMD is proud to announce its most comprehensive training submission to date, and its first MLPerf Training submission to include multi-node training. This round marks an important milestone in demonstrating AMD Instinct™ GPUs across both single-node and multi-node workloads.

A major highlight is the at-scale Oracle Cloud Infrastructure result on 512 AMD Instinct MI300X GPUs, representing a significant step forward for large-scale training on AMD Instinct infrastructure. Moving from no prior multi-node MLPerf Training submission to a 512-GPU result is a major achievement and an important proof point for production-scale AI training. The submission also includes Flux.1 FP8 multi-node data-parallel training on 64 AMD Instinct MI325X GPUs, extending the scale story across both MI300X and MI325X platforms.

This round also marks the first MLPerf Training submission using the AMD Primus software stack, with results across two language model workloads: Llama2-70B LoRA and Llama3.1-8B pre-training. AMD is also introducing MXFP4 training for the first time on two MLPerf Training models: Llama2-70B LoRA and Llama3.1-8B. Compared to the MLPerf 5.1 MI355X publications, AMD achieved a 19% improvement on Llama2-70B LoRA and a 13% improvement on Llama3.1-8B, highlighting continued progress in software optimization, precision enablement, and training efficiency.

Ecosystem participation remains strong, with confirmed partners including OCI, Dell, HPE, Asus, Cisco, Supermicro, MiTAC, MiTAC with Akash, KRAI and Vultr. Taken as a whole, the AMD MLPerf 6.0 Training submission highlights major progress in multi-node training, software maturity, and ecosystem collaboration.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# ASUSTeK

ASUSTeK Computer Inc. today announced that its high-performance AI server, the XA_NB3I, has delivered outstanding results in the latest round of MLPerf Training benchmarks — the industry-standard AI performance evaluation developed and maintained by MLCommons, a global open engineering consortium dedicated to accelerating machine learning innovation for the benefit of humanity.

As a long term member and contributor of MLCommons, ASUS actively participates in the MLPerf benchmarking program to drive transparency, reproducibility, and continuous advancement across the AI industry. This collaboration reflects ASUS's belief that open, standardized benchmarking is essential to pushing the boundaries of AI technology and ensuring that progress benefits organizations of all sizes worldwide.

The XA_NB3I is powered by an Intel Xeon processor and equipped with dual NVIDIA Blackwell architecture GPUs totaling 8 accelerator cards, purpose-built for demanding generative AI and large language model (LLM) inference workloads. In the latest submission, the system achieved optimized performance in **6.584** on the Llama 2 7B inference task, a competitive **84.85** on the DeepSeek 671B ultra-large model test, and strong results of **74.75** on Llama 3.1 8B and **2.342** on the DLRM-DCN recommendation benchmark.

Through its continued participation in MLCommons and the MLPerf benchmarking program, ASUS remains committed to advancing open, measurable, and reproducible AI performance. By contributing benchmark results and continuously optimizing its AI server platforms, ASUS aims to help enterprises, researchers, and solution partners better evaluate AI infrastructure and accelerate the deployment of responsible AI across a wide range of industries.

Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Azure

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Cisco

## Cisco UCS and Silicon One: Powering Enterprise AI

Cisco's MLPerf® Training v6.0 submission delivers an end-to-end story for high-performance AI infrastructure, pairing high-density Cisco UCS® compute with Cisco-engineered networking powered by Cisco Silicon One™. The submission spans the core workloads driving enterprise AI: LLM pretraining (llama3.1-8B, gpt-oss-20B), parameter-efficient fine-tuning (llama2-70B LoRA), and recommender training (DLRM-DCNv2).

### Cisco UCS Compute Portfolio

Cisco UCS delivers high-density GPU servers purpose-built for demanding AI workloads:

- **Cisco UCS C880A M8** with NVIDIA B300-SXM6 288GB GPUs and NVIDIA BlueField-3 400Gb/s DPUs.
- **Cisco UCS C885A M8** with NVIDIA H200-SXM5 141GB GPUs and NVIDIA BlueField-3 DPUs.
- **Cisco UCS C885A M8** with AMD Instinct™ MI350X OAM accelerators (288GB HBM3e) and AMD Pensando™ Pollara 400 AI NICs, providing cross-architecture flexibility.

Across these platforms, workloads ran with established mixed-precision recipes: FP8 hybrid on NVIDIA stacks and BF16/MXFP4 on AMD, maximizing throughput while preserving convergence quality.

### Cisco Silicon One G200 & Scale-Out Efficiency

Every cluster used Cisco Silicon One™ G200 (a purpose-built 5nm, 51.2 Tbps, 64×800GE switching ASIC) at 400Gb/s per GPU over RoCEv2. The NVIDIA H100 GPU cluster used a rail-optimized topology whereas the AMD MI350X cluster used a 3-stage CLOS.

When the GPU count was doubled, the infrastructure sustained ~76–80% scaling efficiency on LLM pretraining workloads, keeping network overhead tightly controlled at scale. This is enabled by G200 hardware capabilities:

- **Fully shared on-die packet buffer:** Minimizes packet loss and Priority Flow Control (PFC) events, ensuring low-latency RoCEv2 transport.
- **Intelligent Packet Flow:** Uses advanced packet spraying and dynamic load balancing to optimize Job Completion Time (JCT).
- **Hardware-based link failure isolation and rerouting:** Resilient, production-grade fabrics that maintain throughput through link events.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

These MLPerf™ results show how Cisco UCS and Silicon One deliver a full-stack, Cisco-engineered foundation for enterprise AI, built on open, standards-based Ethernet from silicon to system to fabric.

# CoreWeave

CoreWeave submitted MLPerf® Training v6.0 results across multiple system configurations using NVIDIA HGX B200 and NVIDIA GB300 NVL72. CoreWeave delivered exceptional scaling efficiency from 2,048 to 8,192 GPUs, demonstrating near-linear scaling efficiency and training times at different cluster sizes. Customers can scale their clusters with confidence, knowing performance will grow predictably with their workloads.

## DeepSeek-V3-671B on NVIDIA GB300 NVL72

CoreWeave's flagship submission trained DeepSeek-V3-671B, one of the most computationally demanding models in the benchmark, achieving target quality in just 2.02 minutes on 8,192 GB300 GPUs.

As the cluster size doubled from 2,048 to 4,096 to 8,192 GPUs, training time improved predictably from 5.54 to 3.09 to 2.02 minutes respectively, demonstrating near-linear scaling efficiency as a result of CoreWeave's optimizations across the stack.

## Llama-3.1-405B Pretraining on NVIDIA GB300 NVL72

CoreWeave's 1,024-node, 4,096-GPU GB300 deployment reached the Llama-3.1-405B reference quality target in just 9.77 minutes. The run was built on a stack using NVIDIA NeMo Framework Release 26.04, full CUDA graphs, Tensor/pipeline/context-parallel sharding tailored to the GB300 NVL72 topology, and NVIDIA Spectrum-X Ethernet running RoCE for scale-out fabric. Customers can run on this stack without manual configuration and run experiments, iterate models and get to production efficiently on CoreWeave.

## GPT-OSS-20B and Llama-3.1-8B with NVIDIA HGX B200

On a single 8-node, 64-GPU B200 cluster connected via InfiniBand, we trained GPT-OSS-20B in just 26.98 minutes and Llama-3.1-8B in 16.54 minutes. We achieved this result for GPT-OSS-20B via optimizations in our stack delivering performance equivalent to higher-end Blackwell platforms. This validates that our full stack delivers high performance consistently across multiple platforms and deployment sizes.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Dell

Dell's MLPerf Training v6.0 submission demonstrates the breadth and depth of the PowerEdge AI portfolio, covering ten distinct server configurations across NVIDIA GB300, B300 SXM, B200 SXM, H200 NVL, and AMD MI355X accelerators—spanning single-node through 16-node scale-out deployments.

The submission spans four industry-standard training benchmarks: Flux image generation, GPT OSS 20B large language model pre-training, LLaMA 2 70B LoRA fine-tuning, and LLaMA 3.1 8B pretraining. This coverage across image generation, large-scale pre-training, and parameter-efficient fine-tuning workloads reflects the diverse AI training demands customers face today.

Our systems utilize BF16 mixed-precision training with FP8 acceleration on supported Hopper and Blackwell GPUs, taking full advantage of NVIDIA Transformer Engine to maximize throughput without sacrificing convergence quality. Multi-node configurations leverage high-bandwidth NVLINK and InfiniBand networking to deliver strong scaling efficiency across our 2- to 16-node deployments. On the AMD side, our PowerEdge XE9785 systems running MI355X accelerators are optimized using ROCm-based software stacks, enabling Dell customers to benefit from AMD's latest GPU performance for large language model training. Dell's participation across both NVIDIA and AMD GPU architectures reflects our commitment to open, vendor-flexible AI infrastructure—ensuring customers can build on the platform that best fits their needs.

We are excited to share our results with the industry and our customers. These results affirm Dell's ability to deliver high-performance, accessible, scalable and production-ready AI training infrastructure across the full spectrum of enterprise needs - from compact single-node deployments to large multi-node clusters.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# 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.

In this round, we measured Llama2-70b-lora on a PRIMERGY CDI system. The training time required was approximately 30 minutes. This time is half of the results we have measured in the past round (4.1-0033; PRIMERGY GX2560M7). This is attributed to the PRIMERGY CDI's support for newer generation GPUs and its ability to accommodate a larger number of GPUs.

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 v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# GigaComputing

Giga Computing is a GIGABYTE subsidiary that serves as the company's enterprise division responsible for designing, manufacturing, testing, and selling GIGABYTE server products. Giga Computing is one of the founding members that has continually contributed unique systems in testing for each round of MLPerf Training and MLPerf Inference. We hope these results can allow academia, global technology providers, and others to have better expectations of what systems can achieve to drive results.

For MLPerf Training 6.0, GIGABYTE systems were selected to reflect the broad choices of accelerated computing systems in the market today. The two 8U systems were optimized for airflow and performance using 8-GPU OAM and HGX baseboards. An NVIDIA GB300 NVL72 was also introduced for AI training benchmarking.

- GIGABYTE G893-ZX1-AAX4, using:
  - 2x AMD EPYC 9575F (64 core) processors
  - AMD Instinct MI355X
- GIGABYTE G894-SD3-AAX7, using:
  - 2x Intel Xeon 6767P (64 core) processors
  - NVIDIA HGX B300
- GIGABYTE Compute Nodes in NVIDIA GB300 NVL72
  - 16 nodes (64x Blackwell GPUs, 32x Vera CPUs)
  - 18 nodes (72x Blackwell GPUs, 36x Vera CPUs)

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

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Google

Google's MLPerf® Training v6.0 submission highlights exceptional scaling performance and architectural alignment with the most advanced generative AI workloads in production today.  Google intentionally focused on the newly introduced DeepSeek-V3 671B as the first large-scale Mixture-of-Experts (MoE) workload in the suite. DeepSeek-V3 directly mirrors the sparse, high-capacity model architectures that our customers rely on to deploy highly efficient foundation models. Our performance results demonstrate superb computational efficiency across multiple NVLink domains across ROCE.

Trust in our infrastructure begins with reliability, Google engineers systems that demonstrate inherent resilience. The AI Hypercomputer is a supercomputing system optimized to support your AI/ML workloads. It's an integrated system of performance-optimized hardware, open software, ML frameworks, and flexible consumption models. The AI Hypercomputer system incorporates best practices and systems-level design to boost efficiency and productivity across AI pre-training, tuning, and serving. To enable easy creation and management of training, Cluster Director configures compute, networking, and storage resources for your training clusters to maximize performance and minimize downtime

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# HPE

HPE continues to demonstrate differentiated AI performance across its diverse portfolio through consistent benchmarking. For MLPerf Training v6.0, HPE submitted its highest number of AI training results to-date for a single round and achieved a significant result – an order of magnitude improvement in time-to-train for a rack-scale AI system.

Results submitted include recommendation, LLM and MoE on HPE ProLiant Compute XD servers and NVIDIA GB200 NVL72 by HPE and NVIDIA GB300 NVL72 by HPE rack-scale systems. HPE benchmarked air-cooled and direct-liquid cooled configurations powered by AMD EPYC™ CPUs, Intel Xeon 6 processors and NVIDIA Grace CPUs, as well as NVIDIA Blackwell Ultra and AMD Instinct™ GPUs.

NVIDIA GB200 NVL72 by HPE and GB300 NVL72 by HPE rack-scale systems featured LLM Llama3.1-8B pretraining and Llama2-70B LoRA finetuning. NVIDIA GB300 NVL72 by HPE featured multiple LLMs, including the new GPT-OSS-20B MoE and the only Llama3.1-405B results on a single rack.

HPE can help customers gain efficiency in pretraining AI models and showcased that smart hyper-parameter tuning can make a significant difference. Two of HPE's Open-division Llama3.1-405B results demonstrated time-to-train efficiency gains of 10X on a single NVIDIA GB300 NVL72 by HPE. Starting from a time-to-train of 475 minutes, HPE showed a 3X improvement of 158 minutes by reducing warm-up steps to 150. By reducing warm-up steps and global batch size, the time-to-train improved by 10X to 50 minutes, which translates to 10X energy savings. These results prove it is possible to conserve energy, make better use of system resources, and improve time-to-train through careful selection of hyperparameters to fit cluster sizes.

Through deep expertise and confident deployments with performance validation, HPE delivers differentiated AI performance to customers.

Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Inventec

Inventec, a global leader in server manufacturing, proudly announces its successful participation and highly competitive results in the MLPerf Training Benchmark v6.0. The Inventec team achieved great performance benchmarks across demanding large language models (LLMs), including "llama2_70b_lora", "llama31_8b", and "gpt_oss_20b".

These achievements were powered by Inventec's latest flagship AI hardware, the P9000G7(AC) GPU server. Engineered to tackle the massive computational demands of modern generative AI, the P9000G7(AC) integrates state-of-the-art NVIDIA Blackwell Ultra B300 GPUs. This integration delivers a monumental LLM training performance boost of multiple times over previous-generation servers, allowing enterprises to drastically shorten training cycles and optimize data center efficiency.

As AI models expand in complexity, Inventec remains committed to delivering robust, scalable infrastructure. This latest MLPerf validation underscores Inventec's ability to engineer world-class hardware that meets strict industry standards for speed and reliability.

To explore the P9000G7(AC) and Inventec's comprehensive lineup of high-performance AI solutions, please visit the official website at https://ebg.inventec.com/en/product/Server.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# 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 20th round since 2021 (8x Training + 11x Inference + 1x Tiny), a record that very few MLCommons members have achieved.

In this round, we submitted results of benchmarking Isambard AI, the UK's National AI Research Resource. To the best of our knowledge, this is the first time MLPerf will feature results from any Sovereign AI infrastructure.

Isambard AI is composed of 5,448 NVIDIA GH200 Grace Hopper Superchips housed in 1,368 compute nodes across 12 HPE Cray EX4000 cabinets. In our experiments, we used up to 16 nodes with 4x GH200 Superchips each.

We based our Llama2-70B-LoRA submissions on NVIDIA's code from the v5.0 round, which originally used the enroot and pyxis plugins for Slurm. As these were not supported on Isambard AI, we refactored the code to use Podman-HPC and then focused on optimizing the performance.

We also collaborated with HPE on optimizing the same benchmark on an eight-node cluster of HPE Cray XD670 servers with 8x NVIDIA H200 GPUs each.

Considering hardware configurations with the same number of GPUs (16, 32 and 64), this benchmark takes about 30–35% longer to run on GH200 than on H200. During training, we observed GH200 consuming about 510 Watts and H200 consuming about 700 Watts, so the difference in power consumption largely explains the difference in performance.

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

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Lambda

**About the Benchmarks: NVIDIA GB300 NVL72 on Lambda AI Cloud**

Lambda partnered with NVIDIA for the MLPerf® Training v6.0 round, benchmarking on a bare-metal Lambda Cloud cluster with 72 NVIDIA Blackwell GPUs (GB300-SXM-288GB), as well as a single node of eight NVIDIA Blackwell GPUs (B200-SXM-180GB).

Lambda's GB300 NVL72 Llama 3.1 8B run converges 18.7% faster than our previous best MLPerf Training result (11.59 minutes vs. 14.25 minutes). Those gains confirm NVIDIA's latest hardware and software improvements on Lambda's bare-metal infrastructure.

On the single B200 node, Lambda posted competitive results across two workloads. GPT-OSS-20B, introduced for the first time this round, converged to target accuracy loss in 96.46 minutes. That result shows Lambda Cloud keeping pace with the workloads AI teams are actively running.

Lambda AI Cloud is built for training frontier-scale AI models. Orchestrate through Kubernetes or Slurm, managed or unmanaged.

**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 v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# 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 v6.0 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 Large Language Model (LLM) architectures.

- **MiTAC G8825Z5 (8U Dual-Socket 8-GPU Enterprise AI Platform):** Engineered for massive scale-out AI training and high-throughput LLM workloads, this platform integrates high-performance **AMD EPYC™ 9575F and AMD EPYC™ 9755** processors with **8x AMD Instinct™ MI350X** accelerators.The platform demonstrated strong versatility and stable compilation across next-generation pretraining, parameter-efficient fine-tuning, and high-fidelity visual AI by successfully completing:
    - **Llama 3.1-8B**
    - **Llama 2-70B LoRA**
- **MiTAC G4826Z5 (4U Multi-GPU Liquid-Cooled Server):** Optimized for space-constrained data centers requiring dense compute without power throttling, this 4U chassis integrates **advanced liquid cooling technology** to ensure optimal thermal efficiency. It couples **AMD EPYC™ 9575F and AMD EPYC™ 9755** processors with **8x AMD Instinct™ MI355X** next-generation accelerators:
    - **Llama 3.1-8B**
    - **Llama 2-70B LoRA**

These comprehensive validation metrics affirm that MiTAC’s AI servers stand as a future-proof, robust infrastructure solution for enterprises and research institutions seeking to deploy high-performance, real-time GenAI applications at scale.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Nebius

For MLPerf® Training v6.0, Nebius submitted results for NVIDIA Blackwell Ultra systems: NVIDIA HGX B300 and NVIDIA GB300 NVL72. The benchmarking focused on evaluating configurations for foundation model training across different cluster sizes. The HGX B300 system was tested on the following configurations: single node (8 GPUs), four nodes (32 GPUs), and eight nodes (64 GPUs). The GB300 platform was evaluated on the following configurations: two nodes (8 GPUs total), eight nodes (32 GPUs total), and a full-rack configuration (72 GPUs total)

Nebius submitted results for two large language models, Llama-3.1-8B and GPT-OSS-20B, and one text-to-image model: FLUX.1. LLM benchmarks were conducted on single-node and eight-node HGX B300 and single-node and full-rack GB300 NVL72 systems. For FLUX.1 model training, larger configurations were used: four-node and eight-node HGX B300, and eight-node and full-rack GB300 NVL72 systems.

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 systems achieved high GPU utilization and stable throughput across all tested configurations, from single-node to full-rack deployments, and across both LLM training and image generation workloads. The outcomes confirm that Nebius's virtualized environment sustains efficiency and system behavior 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’s results contribute to this effort and highlight the readiness of its infrastructure to support demanding AI training and fine-tuning workloads with reliable scalability and resource efficiency.

Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Netweb Technologies India LTD

Netweb Technologies announced its participation in **MLCommons MLPerf® Training v6.0**, demonstrating its capabilities in delivering high-performance infrastructure for large-scale AI model training. This submission highlights Netweb's focus on enabling efficient, scalable training platforms optimized for modern deep learning workloads across enterprise and research environments.

For MLPerf Training v6.0, Netweb utilized its **Tyrone Camarero PDI200A2HG-810 platform**, configured with **8× NVIDIA B200 GPUs** featuring **HBM3e high-bandwidth memory** and interconnected via **NVLink Gen5**, enabling high-speed GPU-to-GPU communication. The system is powered by **dual Intel Xeon Platinum 8562Y+ processors** and supported by **2 TB of system memory**, providing a balanced architecture for compute-intensive training workloads and data pipeline orchestration.

The submission includes representative large-scale AI training workloads spanning both full model pretraining and parameter-efficient fine-tuning. Netweb executed **pretraining workloads for GPT-OSS-20b and Llama 3.1 8B**, as well as **LoRA-based fine-tuning for Llama 2 70B**. These workloads reflect key enterprise AI use cases such as foundation model development, domain adaptation, and efficient customization of large language models.

The software stack is based on **PyTorch NVIDIA Release 26.04**, along with **CUDA 13.2**, cuDNN, NCCL, and other optimized GPU libraries, ensuring efficient scaling and high utilization across multiple accelerators. The system runs on **Ubuntu 24.04 LTS**, providing a stable and production-ready environment for AI training workloads.

This MLPerf Training v6.0 submission reinforces Netweb Technologies' capability to support **end-to-end AI training workflows**, combining advanced GPU acceleration, high-speed interconnects, and optimized software frameworks. Through participation in MLCommons benchmarking, Netweb demonstrates its commitment to transparent, standards-based evaluation and readiness for next-generation AI deployments.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# NVIDIA

This round, NVIDIA submitted results across the complete MLPerf Training suite, including two newly introduced pretraining workloads: GPT-OSS-20B and DeepSeek-V3 671B, the first large mixture-of-experts (MoE) model in the suite. NVIDIA scaled DeepSeek-V3 training across 8,192 GPUs using GB200 NVL72 systems, the largest-scale Blackwell-based submission to MLPerf Training.

NVIDIA also continued to submit excellent results across every training benchmark, including dense and MoE LLM workloads, recommender systems, and image generation, with GB300 NVL72 systems delivering up to 1.6x improvement over GB200 NVL72 at the same scale.

This round's results highlight the efficiency, scalability, and reliability of the NVIDIA platform delivered through capabilities such as NVFP4 precision, NVLink fabric, Spectrum-X Ethernet and Quantum InfiniBand networking platforms, and the Reliability, Availability, and Serviceability (RAS) Engine. Paired with continuous software optimization and an expansive developer ecosystem, NVIDIA provides model developers with a versatile platform ideal for training and deploying the next generation of AI models.

The NVIDIA ecosystem participated extensively this round, with compelling submissions from 19 organizations: ASUSTeK, Azure, Cisco, CoreWeave, Dell Technologies, Fujitsu, GigaComputing, Google, Hewlett Packard Enterprise, Inventec, Krai, Lambda, Nebius, Netweb Technologies India Ltd., Oracle, Quanta Cloud Computing, Scitix, Supermicro, and TTA. The largest-scale Blackwell-based partner submissions came from CoreWeave, who scaled DeepSeek-V3 training to 8,192 GPUs on GB300 NVL72 systems, and Azure, who scaled Llama 3.1 405B training to 8,192 GPUs on GB200 NVL72 systems. These submissions reflect NVIDIA's deep co-engineering with cloud partners and the reliability and scalability of its platform across the largest AI training workloads.

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

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Oracle

Oracle delivered a comprehensive submission to the MLPerf Training 6.0 suite, with results across the full range of benchmarks. This submission highlights a practical, scalable AI training environment that supports today's most important model families. It spans language model pretraining, large model fine-tuning, and text-to-image generation, mirroring the diverse training patterns used by researchers, developers, and enterprises building modern AI systems. To demonstrate flexibility at multiple scales, Oracle submitted NVIDIA GB300 Oracle Cloud Infrastructure (OCI) Supercluster results on configurations with 8, 32, 72, and 512 GPUs, as well as single-node AMD Instinct MI355X systems with 8 GPUs. This range shows how OCI can power everything from experimental projects to large-scale production training.

OCI provides a broad portfolio of GPU options tuned to different workload profiles. For high-end training needs, OCI supports some of the most demanding training and inference scenarios with NVIDIA A100 80GB, H100, H200, GB200, and GB300 GPUs, scaling from single nodes to clusters with tens of thousands of GPUs. OCI also supports AMD Instinct MI300X and MI355X accelerators, expanding choice for customers building advanced large-scale training and inference for both open and proprietary models.

Because Generative AI workloads have distinct performance characteristics and infrastructure needs compared with traditional cloud applications, Oracle has engineered a purpose-built GenAI infrastructure stack on OCI. This includes high-bandwidth, low-latency Cluster Networking with RDMA support for efficient distributed training, and high-performance storage powered by Managed Lustre to keep large models and datasets flowing smoothly.

Oracle's participation in MLPerf Training 6.0 reflects its commitment to advancing practical AI infrastructure and supporting a broad range of AI innovation. By providing scalable cloud infrastructure, accelerator choice, and purpose-built capabilities for distributed AI training, OCI enables organizations of all sizes to accelerate model development, reduce operational complexity, and bring AI-powered solutions into production more efficiently.

Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global provider of data center and AI infrastructure solutions, announced its participation in the latest MLPerf® Training v6.0 benchmark submissions in the data center closed division. The submissions highlight QCT's continued investment in scalable AI and HPC infrastructure for large-scale model training and next-generation AI workloads.

QCT submitted results based on two AI platforms designed for different deployment and scaling requirements.

The QCT **QuantaGrid D75H-10U** server is powered by dual Intel® Xeon® 6 processors and NVIDIA HGX™ B300 GPUs in an SXM-based architecture, delivering a high-density accelerated computing platform for AI training and inference. Equipped with eight 800Gb/s OSFP networking ports, the system is designed to support large-scale GPU clusters with high-bandwidth, low-latency interconnects for demanding AI and HPC environments. The platform targets workloads including large language model (LLM) training, agentic AI, reasoning models, and multimodal AI applications.

QCT also submitted results based on the NVIDIA GB300 NVL72 rack-scale platform integrating QCT **QuantaGrid D75U-1U** servers. Built on the NVIDIA GB300 Grace Blackwell Ultra architecture, the platform combines NVIDIA Blackwell Ultra GPUs, NVIDIA Grace CPUs, NVIDIA ConnectX®-8 SuperNICs, and NVIDIA BlueField®-3 DPUs to support large-scale AI training and inference workloads.

For the MLPerf Training v6.0 submissions, QCT utilized a 64-GPU configuration within the NVL72 architecture to align with common multi-node cluster deployment practices and enable more consistent comparisons with conventional 8-GPU server cluster designs. Leveraging NVIDIA NVLink™ Switch technology and a fully liquid-cooled rack-scale design, the platform is engineered for high-density AI infrastructure deployments.

By participating in MLPerf Training v6.0, QCT continues to demonstrate its commitment to open industry benchmarks, validated infrastructure performance, and scalable AI system design for enterprises, cloud service providers, and research organizations.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# SCITIX

ScitiX is pleased to announce our submission of MLPerf® Training v6.0 benchmark results, which are based on the **NVIDIA HGX B200** system. These results are intended to showcase the exceptional performance and scalability of ScitiX's optimized infrastructure in a multi-node environment.

The focus of this submission is to evaluate our capability to train foundation models in scaled clusters. We tested a configuration consisting of **8 nodes with a total of 64 B200 GPUs**. Through careful tuning and optimization of our infrastructure and software stack, we successfully completed the training of the **Llama-3.1-8B** large language model on this configuration, achieving **excellent results**. This achievement demonstrates the high efficiency and stable throughput that the ScitiX platform delivers when executing demanding AI training tasks.

To ensure high performance in multi-node deployments, our infrastructure efforts covered several key areas:

- **Precision Scheme**: We utilized **BF16 mixed-precision training** and leveraged **FP8 acceleration technology** to maximize GPU throughput without sacrificing model convergence quality.
- **Network Setup**: The multi-node configuration employed high-bandwidth **InfiniBand interconnect technology**, guaranteeing efficient, low-latency communication within the cluster at the 64-GPU scale, thus achieving robust scaling efficiency.

The ScitiX results prove that **our standardized GPU Cloud Platform** provides performance and consistent system behavior close to bare-metal configurations **without requiring specialized tuning**, ensuring we can offer reliable and scalable AI training infrastructure across various deployment sizes and large language model training workloads. We are committed to advancing the efficiency and reliability of AI systems and providing AI developers with a ready-to-use platform that supports high-demand AI training workloads.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# 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 are the leading GPU server and storage vendor that designs, develops, and manufactures the majority of our development in the United States – at our headquarters in San Jose, Calif.

We have submitted MLPerf training v6.0 benchmark results for the following GPU systems:

**Advanced Liquid Cooling**:

NVIDIA Blackwell Ultra GPU (B300 SXM):

Supermicro SYS-222GS-NB3OT-ALC, with Intel(R) Xeon(R) 6776P CPU

**Liquid Cooling**:

NVL72 with NVIDIA Blackwell Ultra (B300) GPU with NVIDIA Grace GPU:

Supermicro SRS-GB300-NVL72-M1 (18x ARS-121GL-NB3), in liquid cooling rack

NVIDIA Blackwell Ultra GPU (B300 SXM):

Supermicro AS-4126GS-NB3RT-LCC, with AMD EPYC 9575F CPU

Supermicro SYS-422GS-NB3RT-LCC, with Intel(R) Xeon(R) 6776P CPU

NVIDIA Blackwell GPU (B200 SXM):

Supermicro AS-4126GS-NBR-LCC, with AMD EPYC 9965 CPU

AMD MI355X GPU:

Supermicro AS-4126GS-NMR-LCC, with AMD EPYC 9575F CPU

**Air Cooling**:

NVIDIA Blackwell Ultra GPU (B300 SXM):

Supermicro AS-8126GS-NB3RT, with AMD EPYC 9575F CPU, 2 x systems, 16 GPUs

NVIDIA Blackwell GPU (B200 SXM)g:

Supermicro SYS-A22GA-NBRT, with Intel(R) Xeon(R) 6979P CPU

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Tinycorp

tiny corp maintains tinygrad, a pure-python, 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 v6.0 benchmark round, we submitted LLaMa 8B pretraining results on single-node AMD MI350 machines. This submission leveraged tinygrad with its python-based userspace driver for the MI350, which delivered a performance uplift over the kernel driver through reduced overhead and more control over the hardware. To demonstrate the flexibility of tinygrad, this submission also employed a mixed FP8 and BF16 precision recipe. This allowed us to utilize the memory and throughput advantages of FP8 combined with the numerical stability of BF16 for LLM pretraining.

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 v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# TTA

The Telecommunications Technology Association (TTA) is a non-profit organization established in 1988, providing testing, certification, and standardization services for Artificial Intelligence (AI). TTA plays a key role in ensuring the safety, reliability, and interoperability of AI technologies by validating performance, evaluating compliance, and developing technical standards that support trustworthy AI deployment across industries.

As part of its broader effort to support the Korean AI infrastructure ecosystem and expand the visibility of Korean server and accelerator platform vendors in global markets, TTA recognizes the value of publishing performance results through platforms such as MLCommons. Following our earlier participation in MLPerf Storage v2.0, this submission marks TTA's first contribution to MLPerf Training, expanding the scope of our benchmarking activities from storage I/O characterization to end-to-end model training performance.

For this round, TTA evaluated a single-node, air-cooled training system equipped with eight PCIe-attached Blackwell-generation accelerators (96 GB each), dual 96-core server-class CPUs, and 1.5 TB of DDR5 host memory. All inter-accelerator communication occurs over a PCIe Gen5 fabric within a single NUMA domain, with no external interconnect or multi-node fabric involved. The submission targets the Llama 3.1 8B pretraining benchmark in the Closed Division, using the NeMo and Megatron-LM reference software stack with the prescribed Closed Division hyperparameters and an FP8 hybrid precision recipe combined with BF16 mixed precision.

This submission provides a useful reference point for single-node, PCIe-based on-premises training configurations, and we are excited to share these results with the community. TTA remains committed to open benchmarking and to active participation in collaborative initiatives such as MLPerf, which we view as essential to fostering innovation and ensuring broad community benefit.

# Supplemental Results Discussion for MLPerf Training v6.0

**This information is under embargo until 6/16/26 8:00AM PT**

# Vultr

Vultr's MLPerf Training v6.0 submission demonstrates high-performance, cloud-accessible AI training infrastructure powered by AMD Instinct MI355X accelerators. This marks Vultr's first MLPerf Training submission, following its first MLPerf Inference submission in August 2025, and extends Vultr's participation in open, standardized AI benchmarking from inference into training.

For this round, Vultr submitted Closed Division, single-node results for three large language model workloads: Llama 3.1 8B, Llama 2 70B LoRA, and GPT-OSS-20B. Results were obtained on Supermicro AS-4126GS-NMR-LCC systems configured with eight AMD Instinct MI355X accelerators. Each accelerator includes 288 GB of HBM3e memory, providing 2.3 TB of total GPU memory in a single system. The platform also includes two AMD EPYC 9575F CPUs, 3 TB of system memory, AMD Infinity Fabric and PCIe Gen5 connectivity, and eight 400 Gbps RoCE interfaces for cluster-ready deployments.

The benchmark software stack used ROCm 7.2.0 with Primus and Megatron Bridge, and the submitted workloads used an FP8 hybrid precision recipe. Because this submission focused on single-node benchmarking, RoCE networking was not used for benchmark communication; intra-node communication used the system's local GPU and platform interconnects. The available 8x400 Gbps RoCE interfaces support customers building larger GPU clusters on the same infrastructure foundation.

Vultr makes high-performance GPU cloud infrastructure easy to use, affordable, and locally accessible for enterprises and AI innovators. Trusted by hundreds of thousands of active customers across 185 countries, Vultr provides flexible Cloud Compute, Cloud GPU, Bare Metal, and Cloud Storage solutions. These MLPerf Training results give customers transparent performance data for evaluating MI355X-powered infrastructure for large-scale training, fine-tuning, inference, and high-performance computing workloads.