MLCommons Supplemental Discussion

MLPerf Training v5.1 Results Discussion

The submitting organizations provided the following 300 word descriptions as a supplement to help the public understand their MLCommons® MLPerf® Training v5.1 submissions and results. The statements **do not reflect the opinions or views of MLCommons.**

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

AMD

For the MLPerf 5.1 Training round, AMD is proud to announce its first submission using the new AMD Instinct™ MI350 Series GPUs, including both the MI355X and MI350X platforms. This submission represents a significant milestone in delivering high-performance solutions for AI model training across a range of workloads.

Building on the success of their previous MLPerf 5.0 training submission, AMD has doubled the number of submitted benchmarks in v5.1 when compared to the previous round. Alongside continued participation in Llama 2-70B, the latest results include a new submission for Llama 3.1-8B—a benchmark that AMD led and defined within the MLCommons community. The 8-billion-parameter model has emerged as the new pre-training standard, designed to run efficiently on a single node while also scaling seamlessly across multiple nodes—reflecting the evolving balance between accessibility and performance in modern generative AI workloads.

AMD Instinct MI350 Series GPUs demonstrates strong generational performance improvements, delivering up to 2.2X higher performance on Llama 2-70B when comparing the MI355X platform to the MI325X platform, and up to 2.9X higher performance comparing the MI355X platform to the MI300X platform. These results highlight continued progress in architecture, software optimization, and overall system efficiency—enabling faster, more energy-efficient AI model training at scale.

This round also marks a record level of ecosystem participation, with nine partners—Asustek, Cisco, Dell, GigaComputing, Krai, MangoBoost, MiTAC, QCT, and Supermicro—submitting results on AMD Instinct hardware. The breadth of participation underscores the strength and maturity of the AMD AI ecosystem, as well as broad industry confidence in its performance and scalability.

Together, these submissions reaffirm AMD leadership in open benchmarking, ecosystem collaboration, and innovation for the next generation of AI training.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

ASUSTeK

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Cisco

Cisco Silicon One G200: AI Networking at Scale

This multi-node submission evaluated distributed training performance using an 8-node and 16-node cluster running Llama 3.1 8 billion, DLRM DCNv2, and Llama 2 70 billion fine-tuning workloads on NVIDIA H100 GPUs. The nodes were interconnected using Cisco's Silicon One G200 platform, purpose-built for AI-scale networking.

The Silicon One G200 is a high-performance, deterministic, and power-efficient 64x800GE switching platform built on 5nm technology, optimized for AI/ML workloads.

Key capabilities include:

- **High Bandwidth and Low Latency**: With 51.2 Tbps of switching capacity and low-latency, it enables high throughput and reduces training time for large-scale models.
- **Lossless Ethernet with RoCEv2**: Enables high-throughput, low-latency transport critical for distributed AI workloads, supported by advanced congestion control and in-band telemetry.
- **Intelligent Packet Flow:** Combines flowlet-based and packet-spray techniques to optimize path utilization, reduce congestion, and maintain packet ordering.
- **Scalability and Flexibility**: Supports a wide range of port configurations (10G to 800G), enabling efficient scaling across diverse GPU cluster topologies.

Cisco's Compute Portfolio for AI

Cisco offers the UCS C885A M8 server, a high-density GPU rack server aimed at demanding AI workloads, offering powerful performance for model training, deep learning, and inference. Cisco's submission of three MLPerf training results on Cisco UCS C885A M8 platform with NVIDIA HGX H200 Tensor core GPUs, and submission of two MLPerf training results on Cisco UCS C885A M8 platform with AMD Instinct™ MI350X OAM GPU cores.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Cisco has achieved competitive training results with its multi-node configurations. The Cisco UCS C885A M8 platform server has demonstrated exceptional performance running the Llama2-LoRA & Retinanet models with NVIDIA H200-SXM-80GB GPUs and powered by NVIDIA BF3 Ethernet NICs.

Additionally, Cisco achieved exceptional results on llama2-70b LoRA & llama3.1-8b models with AMD Instinct™ MI350X OAM GPUs and also powered by NVIDIA CX7 Ethernet NICs.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

DataCrunch

DataCrunch is a European provider of cloud infrastructure for efficient and scalable AI training and inference. Our services adhere to European compliance standards and utilize 100% renewable energy sources.

Taking part in MLPerf Training v5.1 by MLCommons, DataCrunch conducted the Llama-3.1-8B benchmark with an 8-node system equipped with 8×**B200** GPUs each. This workload was exercised in our environment, as provided and tested by NVIDIA.

Our central objective was to demonstrate a clear and reproducible path from cluster provisioning to a compliant run. We executed the benchmark using the containerized framework of NeMo in conjunction with the training artifacts from Megatron-LM, and orchestration with Slurm and Pyxis. Each node offered high-bandwidth intra-GPU connectivity via NVLink and PCIe Gen5 storage ingress for dataset and checkpoint traffic.

This submission was launched through **Instant Clusters**, DataCrunch's service that enables the allocation of homogeneous node configurations with the latest GPUs. Instant Clusters enable rapid and repeatable provisioning of the 8×B200 setup used for this submission, reducing environment drift and simplifying re-runs for validation. Instant Clusters are equipped with the latest hardware fabric, flexible storage options, and a full suite of development environments for ML workloads.

As for future work, we propose to extend this experimentation to a broader range of larger configurations, between 64× and 128× B200 GPUs, and thus using proportional scaling of models.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Dell

Dell Technologies continues to advance the frontiers of AI performance and choice in infrastructure with its latest MLPerf Training v5.1 submissions. These results underscore Dell's ongoing commitment to supporting open benchmarking, collaboration, and innovation through the MLCommons community.

- In this round, the PowerEdge XE9780 system delivered proven performance across Llama 3 8B runs on the B300 GPU, itGPU, it represents Dell's continued drive to validate emerging architectures that enhance efficiency and throughput for AI training.
- The inaugural submission of the PowerEdge XE9785L with the MI355X system on Llama3-8B and Llama2-LoRA-70B benchmarks showcases Dell's deep engineering expertise across AMD ecosystems.
- Performance gains were particularly notable in the FP4 runs of Llama 2-LoRA 70B model, where Dell's PowerEdge XE9680L system achieved a 27% performance uplift over previous MLPerf v5.0 submissions. This progress reflects Dell's ability to optimize both air- and liquid-cooled platforms for demanding AI workloads, supported by the latest software advancements in deep learning frameworks.
- Dell's PowerEdge XE7745 server achieved another industry milestone as the only RTX PRO 6000 submission, reinforcing Dell's focus on delivering flexibility by supporting a wide array of configurations and innovation across PCIe-based accelerators.
- Finally, Dell submitted updated results for the NVIDIA H200 GPU, continuing to demonstrate strong performance and efficiency across established platforms.

Together, these submissions showcase Dell's broad and balanced AI infrastructure portfolio. Spanning from cutting-edge GPU innovation to sustainable, production-ready server design. Dell Technologies remains committed to advancing open, trusted AI benchmarking. By supporting MLCommons initiatives, Dell helps the global AI community build, test, and deploy solutions that accelerate innovation and deliver real-world impact.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

GigaComputing

Giga Computing is a GIGABYTE subsidiary that serves as the company's enterprise division responsible for designing, manufacturing, and selling GIGABYTE server products.

For MLPerf Training 5.1, two GIGABYTE systems were prepared using an 8U chassis that is optimized for airflow and performance using 8-GPU OAM and HGX baseboards (supporting up to 1400W per GPU). Frameworks: DGL, HugeCTR, and PyTorch.

- GIGABYTE G893-ZX1-AAX4 (to be released):
    - 2x AMD EPYC 9575F (64 core) processors
    - AMD Instinct MI355X
- GIGABYTE G894-AD1
    - 2x Intel Xeon 6960 (72 core) processors
    - NVIDIA HGX B200

To learn more about our solutions, visit: https://www.gigabyte.com/Enterprise and https://www.gigacomputing.com/en/

HPE

As a founding member of MLCommons, HPE remains dedicated to HPC and AI performance and innovation. We appreciate MLCommons for their impact and continued support to the community.

Once again, HPE demonstrated strong MLPerf Training v5.1 performance using HPE Cray XD servers for large language models (LLMs). In this round, HPE added to our previous MLPerf Training v5.0 LLM and GenAI results of HPE Cray XD670 by producing impressive performance for the new MLPerf Llama3.1-8B pretraining benchmark. We obtained a Llama3.1-8B pretraining time-to-train result of 37.75 minutes when scaling across 64 NVIDIA H200 141GB GPUs on an eight-node HPE Cray XD670 cluster with HPE Cray ClusterStor E1000.

HPE continues to showcase the performance, scalability, and versatility through our contributions to MLPerf.

HPE also thanks Krai and NVIDIA for their contributions to scalable AI performance. Specifically, HPE is proud to highlight our collaborative work with Krai which produced strong scalable performance using an eight-node HPE Cray XD670 cluster with 64 NVIDIA H200 141GB GPUs and HPE Cray ClusterStor E1000.

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, having participated in all MLPerf Inference rounds and more than half of MLPerf Training rounds to date.

Krai collaborated with HPE on optimizing the Llama2-70B-LoRA fine-tuning benchmark on an eight-node cluster of HPE Cray XD670 servers with 8x NVIDIA H200 GPUs. This resulted in the first ever Llama2-70B result submitted to MLPerf Training using 64x NVIDIA H200 GPUs, with the score of 4.41 minutes. Krai also submitted exceptional results using 16x and 32x NVIDIA H200 GPUs, with the scores of 12.45 and 6.83 minutes, respectively, which demonstrated marked improvements from results from the previous rounds.

Krai also collaborated with AMD and Dell on optimizing the Llama2-70B benchmark on a Dell PowerEdge XE9680 server with 8x AMD Instinct MI300X GPUs. Krai obtained the scores of 29.22 minutes and 28.65 minutes, based on code that AMD prepared for the v5.0 and for the v5.1 rounds, respectively.

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

This information is under embargo until 11/12/25 8:00AM PT

Lambda

About the Benchmarks: NVIDIA GB300 NVL72 on Lambda AI Cloud

Lambda partnered with NVIDIA for the latest MLPerf® Training v5.1 round, benchmarking on a bare-metal Lambda Cloud cluster featuring 72 NVIDIA Blackwell GPUs (GB300-SXM-280GB), 144 CPU cores, 2.0 TB RAM, and 22 TB local NVMe storage.

Lambda is one of only three participants — alongside NVIDIA and Supermicro — to successfully submit GB300 results in the latest MLPerf® Training v5.1 round. Our Llama benchmarks achieved state-of-the-art time-to-solution.

Compared to the best-in-class performance from the previous MLPerf® Training round, our GB300 NVL72 Llama 2-70B run converges 1.66 times faster than the best 64× H200 run and 1.3 times faster than the best GB200 NVL72 run — showcasing major performance gains enabled by NVIDIA's latest hardware and software innovations.

These results position Lambda AI Cloud with NVIDIA GB300 NVL72 as a premier platform for training frontier-scale AI models—delivering performance and flexible orchestration through Kubernetes or Slurm (managed or unmanaged).

About Lambda

Lambda, The Superintelligence Cloud, builds gigawatt-scale AI factories for training and inference. From prototyping to serving billions of users in production, we build the underlying infrastructure that powers AI. Lambda was founded in 2012 by published AI engineers. Lambda’s mission is to make compute as ubiquitous as electricity and give everyone in America the power of superintelligence. One person, One GPU.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Lenovo

Lenovo delivers smarter technology for all, for hardware, software and more importantly solutions that customers have come to know of. Being smarter requires the research, testing and benchmarking that MLPerf Training v5.1 provides. Allowing you to see first-hand how delivering smarter technology translates into tangible performance metrics.

Since partnering with MLCommons, Lenovo has been able to showcase such results quarterly in the MLPerf benchmarking. The insights gained from these benchmarks are one side of the benefit, whereas the other side allows us to constantly be improving our own technology for our customers based on our leading benchmarks by closely working with our partners in optimizing the overall performance.

We are excited to announce with our partners NVIDIA and Intel, that we competed on multiple important AI tasks such as Llama3.1 8B, Llama2 70B, Retinanet, DLRM and RGAT running on our ThinkSystem SR680aV3, SR680aV4 and SR780aV3 with NVIDIA B200 and B300. These partnerships have allowed us to consistently achieve improving results.

This partnership with MLCommons is key to the growth of our products by providing insights into our performance, baselines for customer expectations, and the ability to improve. MLCommons allows Lenovo to collaborate and engage with industry experts to create growth and in the end produce better products for our customers. Which in today’s world is the number one focus here at Lenovo, customer satisfaction.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

MangoBoost

MangoBoost advances MLPerf Training 5.1 with scalable, reproducible results powered by Mango LLMBoost. In the 5.1 round, we delivered performance and scalability improvements over our 5.0 submissions across AMD Instinct™ MI300X and MI325X systems. This round marked several milestones: MangoBoost was the only organization to submit multi-node AMD results, the first to showcase multi-node MI325X performance, and the first to complete an official co-submission on AMD GPUs for Training, highlighting our collaboration with Supermicro. The submission leveraged the combined strength of Mango LLMBoost and GPUBoost RNIC, pairing an optimized AI software stack with high-throughput networking to enable efficient multi-node performance at scale.

These results are driven by Mango LLMBoost, our ready-to-deploy AI solution for training, finetuning, inference, RAG, and agentic AI workloads. LLMBoost is designed to deliver high performance with flexibility to run GenAI workload across diverse AMD and NVIDIA GPUs on a variety of models, as exemplified by our MLPerf results.

LLMBoost provides developers the freedom to scale from a single GPU to massive, multi-node clusters without friction. With one-line deployment, robust OpenAI and REST API integration, and reproducible performance (via our Docker image), LLMBoost effortlessly scales GenAI workloads across on-premise and cloud environments including Azure, AWS, and GCP.

Together with our hardware acceleration suite, LLMBoost forms the foundation of MangoBoost's end-to-end full-stack AI infrastructure solutions. MangoBoost is an AI solution company improving datacenter efficiency by offloading network and storage tasks. Our DPU and RNIC solutions complement our software stack:

- **Mango GPUBoost:** Provides RDMA acceleration for multi-node training/inference.
- **Mango BoostX:** Offloads TCP stacks to reduce CPU utilization and delivers high-performance NVMe solutions for scalable AI storage.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

MiTAC

It is with great honor that MiTAC, a leading server platform designer, manufacturer, and a subsidiary of MiTAC Holdings Corporation (TWSE:3706), announces its outstanding results from the latest MLPerf® Training v5.1 benchmark suite. These results validate our commitment to delivering cutting-edge AI infrastructure that pushes the boundaries of performance and efficiency. Our flagship G8825Z5 AI/HPC server series, including G8825Z5U2BC-325X-755 and G8825Z5U2BC-325X-575 powered by AMD MI325X GPUs, has demonstrated its prowess, with several standout performances in key Large Language Model (LLM) benchmarks.

The G8825Z5's remarkable achievements are a direct result of its purpose-built hardware and optimized design, which showcased excellence in several categories. Standout performances include:

- **LLaMA2_70B_Lora:** The G8825Z5 showcased robust performance for high-throughput, Time-To-Train-optimized server workloads.

These benchmark results affirm the G8825Z5's capability as a powerful solution for enterprises and research institutions seeking to deploy high-performance, real-time LLM applications at scale

Nebius

For MLPerf® Training v5.1, Nebius submitted results for two AI systems built on the NVIDIA Blackwell platform: NVIDIA HGX B300 and HGX B200. The benchmarking focused on evaluating these configurations for foundation model training across different cluster sizes. The B300 system was tested with 8 GPUs on a single host, while the B200 was assessed in three configurations — one, two, and four nodes with 8 GPUs each.

Nebius submitted results for three models: Llama 2 70B (LoRA fine-tuning), Llama 3.1 8B, and Flux.1. Both Llama models were trained on all tested systems (8x B300, 8x B200, 16x B200, and 32x B200), while Flux.1 was trained only on the largest setup with 32x B200 GPUs.

The results show that the HGX B300 system achieved faster training times than the HGX B200 in comparable configurations, likely due in part to its higher memory capacity — 275 GB per GPU compared to 180 GB on the B200 — and increased FP4 performance. This improvement is supported by the newest implementations of the Llama 2 70B and Llama 3.1 8B benchmarks, which feature FP4 mixed-precision training and leverage the latest capabilities of the Blackwell architecture. Increasing the cluster size from 8 to 32 GPUs also led to predictable improvements in training performance, confirming good scalability across hosts.

Overall, the submission demonstrates solid and consistent performance across all systems and models. The outcomes indicate that Nebius's virtualized environment can sustain high GPU utilization rates, delivering efficiency and stability close to bare-metal configurations.

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 v5.1

This information is under embargo until 11/12/25 8:00AM PT

NVIDIA

This round marks the MLPerf Training debut of the GB300 NVL72 rack-scale system in the available category, featuring 72 Blackwell Ultra GPUs connected as one giant GPU using fifth-generation NVLink. Multiple GB300 NVL72 systems were interconnected using the NVIDIA Quantum-X800 InfiniBand platform – the world's first end-to-end 800 Gb/s networking platform comprising the ConnectX-8 SuperNIC, Quantum-X800 InfiniBand switch, and LinkX cables and transceivers. NVIDIA also made the first-ever training submissions this round using NVFP4 precision, which combines innovations across numerics, hardware architecture, and software to accelerate time to train. Together, these technologies nearly doubled Llama 3.1 405B training performance at the same 512-GPU scale NVIDIA submitted using GB200 NVL72 just five months ago.

NVIDIA also submitted GB200 NVL72 results at 5,120 GPU scale on the Llama 3.1 405B benchmark, more than doubling the maximum scale submitted last round. The combination of significantly larger scale, NVFP4 precision that allows faster math, and accompanying software optimizations yielded a 2.7x improvement in Blackwell training performance at scale.

NVIDIA also continued to submit excellent results on every training benchmark – spanning large and small LLMs, LLM fine-tuning, recommender systems, graph neural networks, and image generation. The performance and programmability of NVIDIA CUDA GPUs, the high bandwidth and scalability enabled by NVIDIA networking technologies, and expansive NVIDIA developer ecosystem provides model developers with a rich platform ideal for inventing the next generation of AI.

The NVIDIA ecosystem participated extensively this round, with compelling submissions from 15 organizations including ASUSTeK, Cisco, Datacrunch, Dell Technologies, Giga Computing, HPE, Krai, Lambda, Lenovo, Nebius, Oracle, Quanta Cloud Technology, Supermicro, University of Florida, and Wiwynn.

We commend MLCommons for its ongoing work to bring benchmarking best practices to AI computing and provide the industry with reliable, peer-reviewed performance data.

Oracle

Oracle has delivered exceptional results in the MLPerf Training v5.1 benchmark, highlighting the strengths of Oracle Cloud Infrastructure (OCI) in delivering excellent AI training performance. Oracle submitted competitive benchmarks across a wide range of workloads—including small large language models (LLMs) pretraining and text to image. These benchmarks were executed on NVIDIA B200 accelerators across various cluster scales —demonstrating the flexibility and scalability of OCI's AI infrastructure.

OCI provides a comprehensive AI platform that includes AI Infrastructure, Generative AI services, Machine Learning services, and embedded AI across Oracle Fusion Applications. Customers can choose from a wide spectrum of GPU options tailored to different workload sizes. For entry-level use cases, OCI offers both bare metal and virtual machine instances powered by NVIDIA A10 GPUs. For mid-scale distributed training and inference workloads, customers can leverage NVIDIA A100 VM and bare metal instances, as well as NVIDIA L40S bare metal systems. At the high end, OCI supports the most demanding training and inference workloads with access to NVIDIA A100 80GB, H100, H200, GB200, B300 and GB300 GPUs—scaling from single nodes to clusters with tens of thousands of GPUs. OCI also supports AMD Instinct MI300X accelerators, enabling advanced large-scale training and inference capabilities for open and proprietary models.

Recognizing that Generative AI workloads have fundamentally different performance characteristics and infrastructure requirements than traditional cloud applications, Oracle has engineered a purpose-built GenAI infrastructure stack. This includes a high-bandwidth, low-latency Cluster Networking architecture with Remote Direct Memory Access (RDMA) support—crucial for distributed training efficiency—as well as high-performance storage solutions powered by Managed Lustre. Together, these innovations ensure that OCI delivers the performance, scalability, and reliability needed to support the next generation of AI workloads.

Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global data center solution provider, continues to advance AI and HPC innovation with high-performance systems built for today's most demanding workloads.

In the latest MLPerf™ Training v5.1 benchmark organized by MLCommons, QCT participated in the Closed Division with two powerful 8-GPU systems- the **QuantaGrid D75T-7U** and **QuantaGrid D74H-7U**: showcasing outstanding performance for various AI training workloads.

The QuantaGrid D74H-7U, powered by **dual Intel® Xeon® Scalable processors** and **eight NVIDIA® H200 SXM5 GPUs**, integrates non-blocking GPUDirect RDMA and GPUDirect Storage technologies. This design enables exceptional data throughput and system efficiency for large-scale AI model training.

The QuantaGrid D75T-7U, designed for next-generation AI applications, features **dual AMD EPYC™ 9005 Series processors** and a UBB 2.0 baseboard supporting up to **eight AMD Instinct™ MI325X accelerators**, each equipped with **256 GB of HBM3e** memory. Featuring eight ×16 PCIe® Gen 5 host I/O links and an **AMD Infinity Fabric™ mesh interconnect**, the platform ensures high intra-GPU bandwidth and low-latency communication within a single node—eliminating data-path bottlenecks for intensive AI training workloads.

These results reaffirm QCT's engineering excellence and commitment to providing optimized AI infrastructure for enterprise, research, and hyperscale deployment. By participating in MLPerf, QCT strengthens its dedication to performance transparency, open benchmarking, and empowering customers to make data-driven decisions with confidence.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Supermicro

Supermicro's innovative Data Center Building Block Solutions® (DCBBS) simplifies and shortens the time for global-scale buildouts of AI factories. In addition to enabling systems racking, power, and cooling, Supermicro offers a broad set of high-quality GPU systems. For MLPerf training v5.1 benchmarking, Supermicro submitted benchmarks on the various offerings using the NVIDIA B200, B300, GB300, MI325, MI350, and MI355X GPUs.

These systems are built to scale hundreds of thousands of GPUs to support immense AI factories for AI training and inference. Supermicro has submitted benchmark results using these systems that are built to scale.

As the need for AI infrastructure continues to scale, AI training clusters require thousands of GPUs to develop foundation models. As members of Supermicro's AI Factory DCBBS package, these systems tackle the rising AI computational requirements. These Supermicro systems feature a modular building block approach, composed of three hierarchical levels: the system-level, rack-level, and data center level, giving customers unparalleled design options in determining a system-level bill of materials, down to selecting individual components, including CPUs, GPUs, DIMMs, drives, and NICs.

System-level customization ensures the ability to meet specialized hardware requirements for a particular data center workload and applications, and allows for granular fine-tuning of data center resources. For the systems supporting AMD MI325, MI355X/MI350X, and NVIDIA B200, both air and liquid-cooled Systems are available to support different data center requirements.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

University-of-Florida

The University of Florida (UF) is the nation's first and leading AI university. The recently completed installation of NVIDIA DGX B200 SuperPOD as part of **HiPerGator** fourth generation, UF's supercomputer, underscores this commitment. By submitting its MLPerf Training v5.1 results, UF demonstrates that large-scale AI training on academic supercomputers can be measurable, comparable, reproducible and accountable — highlighting both infrastructure capability and readiness to handle demanding AI workloads from research, education, and industry.

UF ran seven MLPerf Training benchmarks: SSD-RetinaNet, DLRMv2, Llama-3.1-405B, Llama-2-70B-LoRA, Llama-3.1-8B, RGAT and Flux.1, using NVIDIA's MLPerf containers lightly adjusted for our environment. To evaluate training scalability and throughput across a range of model sizes and parallel strategies, we executed each benchmark across node configurations ranging from one node to 56 nodes (eight to 448 B200 GPUs), depending on the model. The results exhibit strong multi-node scaling efficiency. For example, SSD-RetinaNet and Llama-2-70B-LoRA demonstrate near-linear scaling from one to eight nodes, while Llama-3.1-405B sustains distributed training efficiency up to 56 nodes. These outcomes validate HiPerGator's capacity for frontier-scale AI model training under realistic shared HPC constraints.

All runs used Apptainer containers with SLURM batch scheduling and a parallel Lustre file system, preserving the integrity of the closed-division MLPerf software stack. The workflow employs secure, rootless container execution without Docker or elevated privileges, offering a practical methodology for operating large AI systems in shared academic computer environments. This submission confirms that reproducible AI benchmarking can be performed on multi-tenant HPC systems using standard infrastructure and widely adoptable workflows.
As the sole academic institution in this submission round, UF contributes operational insights to MLCommons, helping advance transparent and trustworthy AI performance measurement. UF is committed to sharing our experience, collaborating with peers and enabling more institutions to run compliant AI workloads on shared HPC infrastructure.

Supplemental Results Discussion for MLPerf Training v5.1

This information is under embargo until 11/12/25 8:00AM PT

Wiwynn

Wiwynn is a leading provider of high-efficiency IT infrastructure solutions. We specialize in powering hyperscale data centers, designing and producing high-quality servers and integrated systems for a wide range of demanding applications, including advanced Artificial Intelligence.

In MLPerf Training v5.1, Wiwynn reported Llama2 70B LoRA benchmark results on the Wiwynn Kinabalu system. By integrating NVIDIA's GB200 NVL72 platform with Wiwynn's purpose-built AI cloud platform and a pioneering approach to advanced AI workloads, the system delivers fast, scalable foundation-model training. Its carefully selected, cutting-edge components and technologies are designed to minimize energy consumption, maximize performance, and ensure efficient power management, leveraging optimized processors that balance throughput and power usage.

We conducted efficiency testing on the Wiwynn Kinabalu system configured with 576 NVIDIA Blackwell GPUs on the NVIDIA GB200 NVL72 platform. These results build on Wiwynn's MLPerf Training v5.1 performance and underscore the benefits of our AI-optimized infrastructure for customers building, fine-tuning, or deploying AI applications.

Wiwynn's corporate mission is to "Provide the Best TCO, Workload and Energy Optimized IT Solutions from Edge to Cloud". Wiwynn will continue to work towards this goal and participate in community activities. Our commitment to innovation and excellence is reflected in our participation in industry benchmarks such as MLPerf Training v5.1, where we strive to demonstrate the capabilities of our products and contribute to the advancement of the field.