MLCommons Supplemental Discussion

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

Supplemental Results Discussion for MLPerf Inference v5.1

AMD

AMD Delivers Strong MLPerf Inference Results Across AMD Instinct™ MI355X and MI325X GPUs

The latest MLPerf Inference submission by AMD underscores a rapidly expanding footprint in AI, reflecting growing industry adoption, software-driven performance gains, and competitive strength across diverse workloads. With multiple partner submissions on AMD Instinct™ GPUs, these benchmarks reinforce a reputation for delivering reliable, scalable, high-performing AI solutions.

Headlining this round is the first-ever MLPerf submission using AMD Instinct™ MI355X GPUs—launched just nine weeks ago—delivering impressive results on Llama 2-70B in FP4 numerical precision. Multi-node AMD Instinct MI355X submissions demonstrate strong scalability across configurations, enabling organizations to expand throughput while maintaining cost efficiency.

Model efficiency advances with two Open category AMD Instinct MI355X submissions on Llama 3.1-405B (FP4) using structured pruning techniques. A 21% depth-pruned model and a 31% pruned and fine-tuned version are both designed to reduce compute per token while maintaining accuracy. These optimizations highlight a continued focus on maximizing performance efficiency for emerging AI workloads.

A highlight of this round is AMD submitting several AI models to MLPerf for the first time, expanding workloads represented in AMD results. For the first time, results are being reported on Llama 2-70B Interactive, Mixtral-8×7B, and SD-XL—representing workloads in conversational AI, mixture-of-experts architectures, and advanced generative image models. These additions demonstrate breadth of AMD Instinct GPU capabilities and commitment to enabling emerging AI use cases.

This round also features an MLPerf first: a multi-node heterogeneous inference result powered by AMD Instinct MI300X and MI325X GPUs, underscoring scalability, ecosystem maturity, and cost advantages for enterprise AI deployments.

In addition to our own submission, AMD also collaborated with the following partners in this round: ASUSTek, Dell, GigaComputing, MangoBoost, MiTAC, Quanta Cloud Technology, Supermicro, and Vultr..

Supplemental Results Discussion for MLPerf Inference v5.1

ASUSTek

We are thrilled to announce our participation and significant contributions to the latest MLPerf Inference v5.1 benchmark release! This milestone underscores our unwavering commitment to advancing the state-of-the-art in AI inference performance and efficiency.

For this release, ASUS focused intensively on **optimizing critical models** and **enhancing inference stack** on hardware platforms. We pushed the boundaries of what's possible, demonstrating impressive gains in **latency reduction** and **throughput improvements** for key deep learning workloads. Our engineering efforts targeted meticulous model quantization, innovative kernel optimizations, and a finely-tuned software stack, all designed to extract maximum performance from the underlying hardware.

We successfully submitted results that showcase our dedication to **real-world applicability** and **sustainable AI deployments**. Our contributions highlight not just raw speed, but also the **power efficiency** and **scalability** of our solutions, which are crucial for enterprise and edge applications alike. By consistently participating in MLPerf, we affirm our role as a leader in delivering highly performant and efficient AI inference, empowering developers and users with faster, more responsive AI capabilities.

This achievement reflects the incredible talent and hard work of our engineering teams and our collaborative spirit within the MLPerf community. We look forward to continuing our journey of innovation, pushing the frontiers of AI inference, and making further impactful contributions to future MLPerf benchmarks.

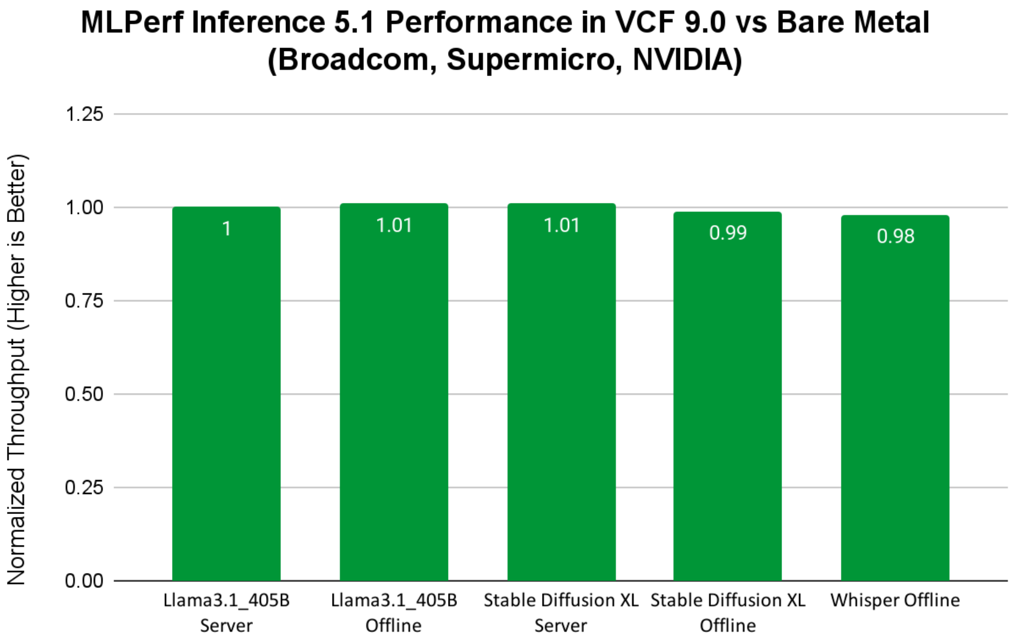
Supplemental Results Discussion for MLPerf Inference v5.1

Broadcom

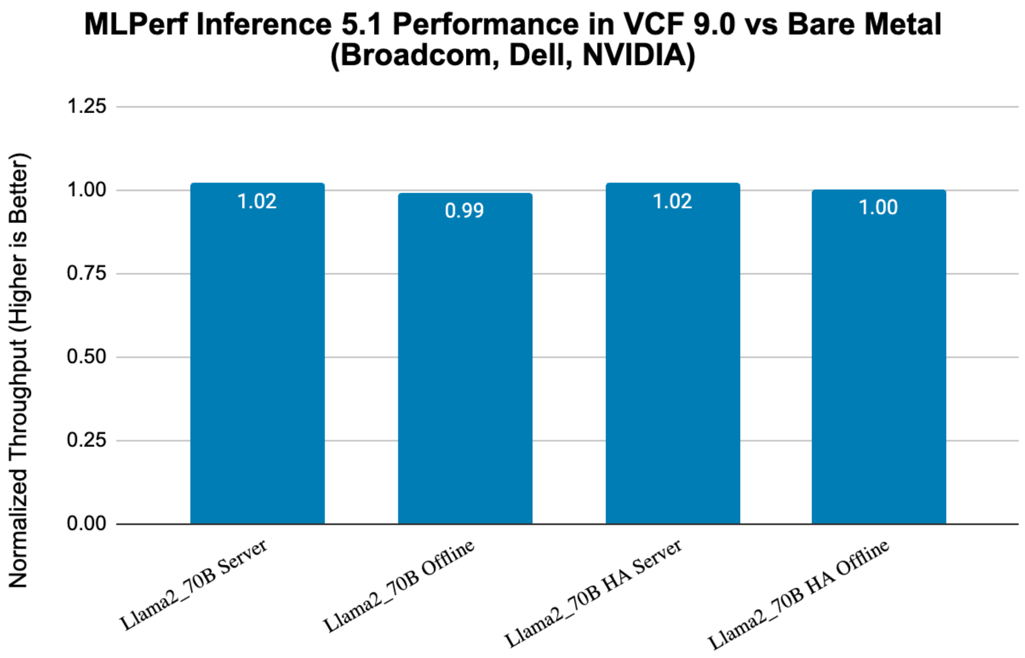
Broadcom is pioneering VMware Private AI as an architectural approach that balances the business gains from AI/ML with the privacy and compliance needs of organizations. Broadcom built private AI directly into the VMware Cloud Foundation (VCF). This provides customers a secure, scalable infrastructure for private AI designed with AI and IT teams' requirements in mind. With VCF, organizations run, move, and govern AI models like any other enterprise application, but do so with a significantly lower TCO due to the benefits of virtualization inclusive of pooling and sharing GPUs, networks, and memory. From fine-tuning to inference, VCF enables private AI to become a scalable, controlled service delivered securely via the private cloud.

Broadcom brings the power of its partners Supermicro, Dell, NVIDIA and Intel to VCF to simplify management of AI accelerated data centers and enable efficient application development and execution for demanding AI/ML workloads. Broadcom supports hardware vendors, facilitating scalable deployments.

Broadcom partnered with NVIDIA, Supermicro, Dell and Intel to showcase virtualization's benefits, achieving impressive MLPerf® Inference v5.1 results. We demonstrated near bare-metal performance across diverse AI domains— Speech-to-Text (Whisper), Text -to-Video (Stable-Diffusion-XL), LLMs (Llama-3.1-405B, Llama2-70B), Graph Neural Networks (RGAT), Computer Vision (RetinaNet). The graphs below compare normalized virtual performance with similar bare-metal performance, showing minimal overhead from VCF 9.0 with NVIDIA virtualized 8xH200 and passthrough 8xB200 and Intel's virtualized dual socket Xeon 6787P. Refer to the official MLCommons Inference 5.1 results for the raw comparison of relevant metrics.

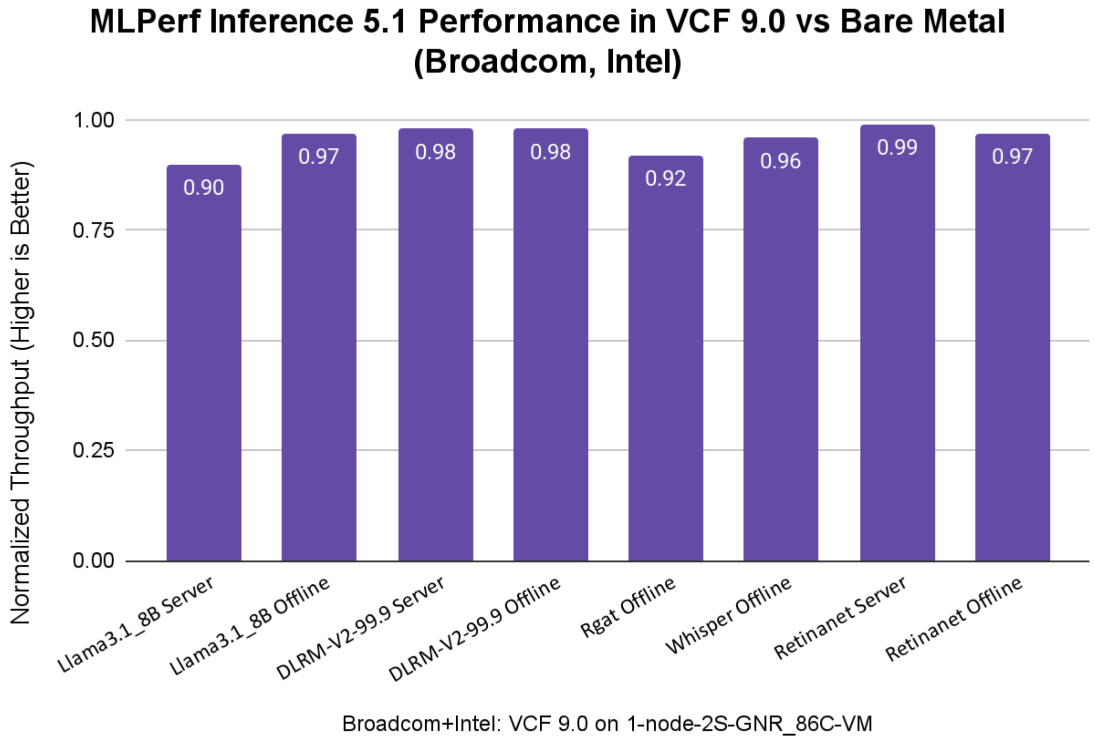

MLPerf Inference 5.1 Performance in VCF 9.0 vs Bare Metal (Broadcom, Supermicro, NVIDIA)

Broadcom+Supermicro: VCF 9.0 on SYS-422GA-NBRT-LCC (Passthrough 8x B200-SXM-180GB)


MLPerf Inference 5.1 Performance in VCF 9.0 vs Bare Metal (Broadcom, Dell, NVIDIA)

Broadcom+Dell: VCF 9.0 on PowerEdge XE9680 (vGPU 8x H200-SXM-141GB)

Supplemental Results Discussion for MLPerf Inference v5.1



Broadcom's GPU-enabled VMs typically used a fraction of the available resources—50% of CPU cores, leaving 50% for other tenant workloads with full isolation. This efficient utilization maximizes hardware investment, allowing concurrent execution of other applications. This translates to significant cost savings on AI/MLinfrastructure while leveraging vSphere's datacenter management. Enterprises gain high-performance GPUs and VCF's operational efficiencies.

Supplemental Results Discussion for MLPerf Inference v5.1

Cisco Systems Inc.

As generative AI transforms the global economy, Cisco is simplifying the complex infrastructure required to power it. We deliver the scalable, accelerated computing solutions essential for modern data centers.

For Demanding, Large-Scale AI

The **Cisco UCS C885A M8** is a high-density, eight-way GPU server engineered for the most demanding AI workloads like model training and distributed inference. Built on the NVIDIA HGX platform, it provides immense, scalable computing power for your most ambitious AI projects. Each server integrates NVIDIA NICs or SuperNICs and BlueField-3 DPUs to ensure top-tier networking performance, accelerated data access for GPUs, and robust zero-trust security.

For Flexible, Enterprise-Grade AI

For broader enterprise applications, the **Cisco UCS C845A M8 Rack Server** brings powerful AI capabilities to a flexible and customizable platform. Based on the NVIDIA MGX reference design, this server supports two to eight NVIDIA PCIe GPUs, making it adaptable for various data-center workloads. It features NVIDIA ConnectX-7 Smart NICs or BlueField-3 DPUs for high-speed networking and efficient scaling.

Proven Performance

Cisco, in partnership with NVIDIA and Intel, demonstrated exceptional performance and efficiency in the latest **MLPerf Inference v5.1** benchmark submissions. Notably, our submissions showcased near-linear scaling for multi-server, multi-GPU inference by pairing **Cisco UCS C885A M8** servers with a high-performance network fabric built on **Cisco Silicon One G200** Ethernet switches. Our platforms excelled across a range of critical AI workloads, including:

- Large Language Models (LLM)
- Text-to-Image Generation
- Object Detection
- Recommendation

This high performance is consistent across our UCS portfolio, highlighted by standout results from:

- Cisco UCS C885A M8 platform with 8x NVIDIA H200 SXM GPUs
- Cisco UCS C845A M8 platform with 8x NVIDIA H200 NVL
- Cisco UCS C845A M8 platform with 8x NVIDIA L40S GPUs
- Cisco UCS X210 M8 with Intel Granite Rapid 6787P processors
- Cisco UCS C240 M8 with Intel Granite Rapid 6787P processors

CoreWeave

CoreWeave, the AI Hyperscaler™, today announced its MLPerf® Inference v5.1 results, reinforcing its leadership in setting new industry standards for AI inference performance. CoreWeave is the first cloud provider to submit MLPerf results for NVIDIA's flagship GB300 GPUs, once again demonstrating high speed in bringing NVIDIA's most advanced hardware to market—including both GB200 NVL72 and GB300 NVL72 GPUs. By consistently delivering cutting-edge technology quickly, CoreWeave empowers customers and organizations to capitalize on performance gains immediately.

CoreWeave achieved impressive per-chip offline scenario results for DeepSeek R1, delivering 6,005 tokens per second per GPU. This benchmark underscores the excellent performance of the GB300 architecture when paired with CoreWeave's AI cloud platform. Combined with the NVL8 spanning 2 nodes and Slurm on Kubernetes (SUNK) orchestration, these results represent excellence for inference throughput.

CoreWeave's integrated platform delivers this performance through a combination of cutting-edge cloud infrastructure, CKS (CoreWeave Kubernetes Service), SUNK, and Mission Control, each purpose-built to deliver maximum performance and efficiency for AI workloads. CKS provides a fully managed Kubernetes environment, while SUNK enhances it with native Slurm integration for seamless workload orchestration. The topology-aware scheduling in SUNK takes full advantage of GB300's high-bandwidth NVLINK domain, enabling exceptionally fast multi-node inference. This ensures large-scale inference tasks run with optimal locality, maximizing communication speed and minimizing latency across nodes. Mission Control further elevates performance by automating node and cluster provisioning as well as monitoring, ensuring the entire fleet operates at peak efficiency.

"Today is another example of our relentless pursuit to deliver a highly performant cloud infrastructure for AI workloads," said Peter Salanki, Chief Technology Officer at CoreWeave. "By combining NVIDIA's most advanced hardware with CoreWeave's tightly integrated cloud platform, we are delivering incredible performance and scalability for our customers' most demanding workloads."

Supplemental Results Discussion for MLPerf Inference v5.1

Dell Technologies

Dell Technologies continues to push the boundaries of AI performance and innovation, as demonstrated in our latest results from MLCommons® MLPerf® Inference v5.1. These achievements reflect the power of strong industry collaboration, cutting-edge engineering, and a commitment to advancing AI infrastructure for organizations everywhere.

Our success is built on powerful partnerships. Through close collaboration with NVIDIA, joint submissions with MangoBoost leveraging AMD Instinct™ MI300X GPUs, and work with Broadcom on a virtualized submission using NVIDIA H200 GPUs, Dell showcased the versatility of the PowerEdge portfolio to meet diverse AI needs.

In this round, Dell submitted **12 systems** featuring an impressive mix of accelerators, including AMD Instinct MI300X, NVIDIA L4, L40S, H100, H200, and the new NVIDIA B200 GPUs. This breadth demonstrates our ability to support everything from generative AI to computer vision and speech recognition workloads.

Standout performances include:

- **LLaMA 2 70B Interactive** ran on thePowerEdge XE9680L system equipped with NVIDIA B200 GPUs and TensorRT, demonstrating exceptional performance in interactive workloads.
- **LLaMA 3.1 8B Server** showcased outstanding capabilities on the PowerEdge XE9685L with the NVIDIA B200, delivering top-tier results.
- **Stable Diffusion XL** executed on the PowerEdge XE9685L, achieving high performance across both offline and server queries.
- **Whisper Speech-to-Text** on the PowerEdge XE9680L delivered excellent performance.

These results highlight Dell's ability to deliver high performance across a variety of AI models, scaling from interactive inference to image generation and speech recognition—all on proven enterprise platforms. As AI workloads evolve, Dell Technologies remains dedicated to delivering the infrastructure that helps customers innovate faster, operate efficiently, and unlock the full potential of their data.

Supplemental Results Discussion for MLPerf Inference v5.1

GATEOverflow

As part of our mission to make **MLPerf Inference** benchmarking more practical and reproducible on everyday hardware, we've successfully run submissions on **NVIDIA RTX 4090 GPUs** as well as a wide range of **AMD and Intel CPUs —** leveraging implementations from **NVIDIA, cTuning, and MLCommons**.

To simplify reproducibility at scale, we've also enhanced the **MLCFlow automation framework** from MLCommons, which powered all of our submissions and ensures **scalable, reliable performance benchmarking**.

We're excited to collaborate with partners who share our vision of making **MLPerf benchmarking accessible, repeatable, and impactful** across diverse hardware platforms.

Supplemental Results Discussion for MLPerf Inference v5.1

Giga Computing

The MLPerf® benchmark submitter - Giga Computing - is a GIGABYTE subsidiary that made up GIGABYTE's enterprise division that designs, manufactures, and sells GIGABYTE server products.

The GIGABYTE brand has been recognized as an industry leader in HPC & AI servers and has a wealth of experience in developing hardware for all data center needs, while working alongside technology partners: NVIDIA, AMD, Ampere Computing, and Intel.

In 2020, GIGABYTE joined MLCommons® and submitted its first system. In the latest **MLPerf Inference v5.1** (closed division) benchmarks, the submitted 8U GIGABYTE systems demonstrated competitive performance across multiple training applications using two unique platforms.

Testing was done using 8U air-cooled servers for AMD EPYC + AMD Instinct MI325X and Intel Xeon + NVIDIA HGX B200 - they've been thermally optimized for greater compute density, ensuring maximum performance and efficiency in high-demand environments.

- **GIGABYTE G894-SD1**:
- 2x Intel Xeon 6745P (32 cores) CPUs
- 8x NVIDIA HGX B200 GPUs

- **GIGABYTE G893-ZX1:**
- 2x AMD EPYC 9575F (64 cores) CPUs
- 8x AMD Instinct MI325X GPUs

Supplemental Results Discussion for MLPerf Inference v5.1

HPE

HPE is committed to benchmarking excellence through our partnership with MLCommons. This round, HPE highlighted some of the newest AI servers featuring updated workload performance, configurations, and the latest NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. The HPE performance results for MLPerf Inference v5.1 include servers suitable for both datacenter and edge inferencing.

HPE datacenter servers demonstrated strong inference performance in large language models (LLMs) for chat Q&A, text summarization, text generation, math and code generation, and the new DeepSeek R1 workload. HPE servers also demonstrated powerful performance in other datacenter workloads including computer vision object detection and recommendation. These datacenter server configurations include:

- HPE Cray XD670 is a versatile and scalable server engineered for high performance AI distributed training, fine-tuning, and inference.
- HPE ProLiant Compute DL380a Gen12 is engineered for multi-GPU AI fine-tuning and inference. This was the first time HPE submitted benchmark tests on a configuration of this server featuring 8x NVIDIA RTX Pro 6000 GPUs.
- HPE ProLiant Compute DL384 Gen12 is engineered for high-throughput AI inference per GPU.
- HPE ProLiant DL385 Gen11 is designed with maximum core counts to accelerate demanding AI workloads.

The HPE ProLiant ML30 Gen11, an HPE server engineered for power-efficient edge inferencing, also delivered inference results in computer vision object detection scenarios.

Supplemental Results Discussion for MLPerf Inference v5.1

Intel

Intel is excited to share MLPerf 5.1 AI inference results for Intel's inference workstations, code-named Project Battlematrix, configured with newly-released Arc Pro B-Series GPUs. Intel GPU Systems demonstrate an accessible solution addressing emerging AI workloads across high-end workstations and edge applications.

Configuring a system with 8 Arc Pro B60 GPUs provides 192GB of VRAM to serve 70b parameter models, like Llama2-70b, to multiple concurrent users. The new companion software stack containerizes driver, framework, and tool support layers together to simplify installation and optimize for multi-GPU inference performance scaling for enterprises. The Intel Arc Pro B60 graphics card variants are available from many partners either standalone or as part of an OEM pre-built and validated inference workstation.

Notably, Intel remains the only vendor submitting server CPU results to MLPerf, reinforcing its leadership in host CPUs as demonstrated by the submission systems. MLPerf Inference v5.1 results demonstrate the strength of Intel® Xeon® 6 with P-cores in AI inference and general-purpose AI workloads. Across five MLPerf benchmarks, Intel® Xeon® 6 CPUs delivered a 1.9x improvement in AI performance boost over its previous generation, 5th Gen Intel® Xeon® processors. The results highlight Xeon's capabilities in AI workloads, from traditional machine learning, small- to mid-size models, to LLM serving.

The combination of Intel® Xeon® 6 and Intel's Arc Pro B-Series GPUs represent our investment to expand customer choice and value, offering real-world solutions that address both LLM models as well as traditional machine learning workloads. Intel also supported partners - Cisco, Dell Technologies, Lenovo, Quanta, Supermicro, and VMWare - with benchmark submissions on their Xeon 6 based products demonstrating great performance.

Thank you, MLCommons, for creating a trusted standard for measurement and accountability.

For more details, please see MLCommons.org.

Notices & Disclaimers Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex . Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software or service activation.

Supplemental Results Discussion for MLPerf Inference v5.1

KRAI

Founded in 2020 in Cambridge, UK ("The Silicon Fen"), KRAI is a purveyor of premium benchmarking and optimization solutions for AI Systems. Our experienced team has participated in all 12 MLPerf ® Inference rounds (as only 3 other submitters), having contributed to some of the fastest and most energy efficient results in MLPerf history in collaboration with leading vendors.

In this round, we benchmarked LLMs via the OpenAI API with realistic overheads (online tokenization/detokenization, networking latency, etc.), as proposed by the new MLPerf Endpoints taskforce.

For example, we benchmarked the performance of a pre-quantized Llama3.1-70B model, as well as dynamically quantized models, on a single server with 8x H200 GPUs.

Using the pre-quantized model, we achieved Offline scores of 31,391 TPS with SGLang v0.4.9, and 26,319 TPS with vLLM v0.9.2. For comparison, using the same pre-quantized model and methodology in the previous v5.0 round, we achieved 27,950 TPS with SGLang v0.4.3, and 21,372 TPS with vLLM v0.6.4.

Using dynamic quantization, we achieved 27,697 TPS with SGLang v0.4.9, and 30,893 TPS with vLLM v0.9.2.

Over the past 6 months, the performance of open-source inference engines on the H200 has improved from ~60-80% to ~90% of the performance demonstrated by submissions using NVIDIA's implementation (c. 35,000 TPS).

We further scaled the performance of Llama3.1-70B with vLLM v0.9.2 to multiple servers, achieving 58,617 TPS on 2 servers and 87,334 TPS on 3 servers (respectively, 1.9x and 2.8x speedup over the single server performance).

We cordially thank Hewlett Packard Enterprise for providing access to Cray XD670 servers with 8x NVIDIA H200 GPUs, and Dell Technologies for providing access to PowerEdge XE9680 servers with 8x AMD MI300X GPUs and 8x NVIDIA H100 GPUs.

Lambda

About the Benchmarks: NVIDIA HGX B200 throughput up double digits

Supplemental Results Discussion for MLPerf Inference v5.1

We partnered with NVIDIA for this round of MLPerf® inference submissions, running benchmarks on a Lambda Cloud 1-Click Cluster™ equipped with eight NVIDIA B200 180GB SXM GPUs, 112 CPU cores, 2.9 TB RAM, and 28 TB SSD.

Our inference benchmarks on models such as Llama 2-70B, Llama 3.1-405B, and Stable Diffusion XL deliver state-of-the-art time-to-solution.

Thanks to NVIDIA software stack improvements, we observed significant throughput gains over the previous MLPerf® round—up to 7% higher for Stable Diffusion XL and up to 15% higher for Llama 3.1-405B.

These results position Lambda's 1-Click Clusters™ with NVIDIA HGX B200 GPUs as a premier choice for AI teams tackling the most demanding inference workloads: proven best-in-class performance coupled with on-demand availability and choice of orchestration (Kubernetes and Slurm, offered as managed or unmanaged).

These clusters range from 16 to 1,536 NVIDIA GPUs and can be self-served on Lambda's Public Cloud–from a single week for prototyping, to multiple years for training and inference production.

About Lambda

Lambda, The Superintelligence Cloud, builds Gigawatt-scale AI Factories for Training and Inference.

Lambda is where AI teams find infinite scale to produce intelligence: from prototyping on on-demand compute to serving billions of users in production, we guide and equip the world's most AI-advanced organizations to securely build and deploy AI products.

Lambda is the trusted AI Infrastructure advisor to the world's top AI Labs, Enterprises and Hyperscalers. With over a decade of experience co-engineering and deploying high-stake cloud solutions at speed and at scale, our teams bring the deepest levels of AI expertise, engineering, management, operations and support – bar none.

Lambda was founded in 2012 by published AI engineers with the vision to enable a world where Superintelligence enhances human progress, by making access to computation as effortless and ubiquitous as electricity.

AI has been Lambda's sole focus since its founding.

Supplemental Results Discussion for MLPerf Inference v5.1

Supplemental Results Discussion for MLPerf Inference v5.1

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 Inference v5.1 provides. Allowing you to see first-hand how delivering smarter technology means producing best-in-class results.

Since partnering with MLCommons, Lenovo has been able to showcase such results quarterly in the MLPerf benchmarking. Achieving these results is one side of the benefit, whereas the other side allows us to constantly be improving our own technology for our customers based on our 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 Large Language Model, Generative Image, Image Classification, Medical Image Segmentation, Speech-to-text, and Natural Language Processing running on our ThinkSystem SR680a V3, ThinkSystem SR780a V3, ThinkSystem SR650 V4 and ThinkEdge SE100. 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 where we stand against the competition, 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.

MangoBoost

In the MLPerf Inference 5.1 submission round, MangoBoost is proud to announce groundbreaking results that push the frontier of LLM Inference Serving on multiple fronts.

With LLMBoost and the optimized ROCm software, MangoBoost becomes the first-ever submitter to submit heterogeneous GPU systems with up to 4-node MI300X + 2-node MI325X GPU and 1xMI300X GPU + 2xMI325X GPU configurations that reach 169k tok/s for llama2-70b. MangoBoost is the first-ever third-party vendor to submit results on the latest AMD MI355X GPU, which highlights the power of AMD MI355X Instinct GPUs achieving 648k tok/s.

MangoBoost collaborates with multiple key partners from AMD, Supermicro and Dell, and showcases performance scalability across many multi-node scenarios. LLMBoost's multi-node optimization ensures MangoBoost's serving engine is aware of the GPUs' architecture, resources, the model's characteristics, and automatically deploys architecture-, model-specific parallelization and autoscaling strategies regardless of GPU architectures and LLM models. LLMBoost also optimizes the communication library to allow an optimal strategy that can overlap compute and data transfer. With these optimizations working together, LLMBoost is able to achieve a near-perfect 92%-97% performance scaling on llama2-70B.

LLMBoost's turn-key solution helps bridge the gap between idea into production, allowing developers, researchers and data scientists to serve multiple powerful open-model at any scale in minutes on any infrastructure whether you are experimenting with a small Qwen2.5-0.5B model on an edge device, or a llama4-Maverick model on a large cluster. Beyond LLMBoost, MangoBoost offers hardware acceleration solutions based on Data Processing Unit (DPUs) for AI and cloud infrastructure. Mango GPUBoost™ – RDMA acceleration for multi-node inference/training via RoCEv2. Mango NetworkBoost™ – Offloads TCP/IP stack to free up CPU resources. Mango StorageBoost™ – High-performance NVMe initiator/target stack for scalable AI storage.

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

Supplemental Results Discussion for MLPerf Inference v5.1

the latest MLPerf® Inference 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 leadership in several categories. Standout performances include:

- **LLaMA 2 70B Interactive:** The G8825Z5 with AMD MI325X GPUs delivered impressive performance in the highly demanding interactive category, demonstrating its capability for real-time conversational AI with a score of 18,846.1 tokens/s.
- **LLaMA 2 70B Server:** The G8825Z5 ranked third, showcasing its robust performance for high-throughput, latency-optimized server workloads.
- **Mixtral 8x7B Server:** The G8825Z5 delivered excellent performance in this complex benchmark, proving its versatility and efficiency across a wide range of LLM architectures.

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.

Supplemental Results Discussion for MLPerf Inference v5.1

Nebius

For MLPerf® Inference v5.1, Nebius submitted results for three AI systems powered by the latest NVIDIA Blackwell and Hopper platforms currently available on the market: NVIDIA GB200 NVL72, HGX B200, and HGX H200. These submissions demonstrate exceptional performance when running Llama 2 70B and Llama 3.1 405 models on single-host installations — with 8 GPUs for the HGX platforms and 4 GPUs for the GB200.

This benchmarking round gave Nebius the opportunity to showcase their engineering expertise in delivering consistent performance for modern AI workloads in general, and inference scenarios in particular. The platform equipped with four GB200 GPUs delivered outstanding results for the Llama 3.1 405B model in both server and offline modes, with 596.11 and 855.82 tokens/s respectively. Similarly, B200- and H200-powered servers achieved solid inference throughput across both models: Llama 2 70B and Llama 3.1 405B.

These results confirm Nebius' ability to run AI workloads in highly-efficient virtualized environments while delivering performance on par with bare metal installations. They reflect the company's commitment to exceptional AI infrastructure, built on custom hardware, proprietary software, and energy-efficient data centers. This combination gives AI labs and enterprises supercomputer-level performance and reliability, coupled with the flexibility and simplicity of a hyperscaler.

The MLPerf® Inference benchmarks from MLCommons help AI companies and ML teams make informed decisions when selecting an AI infrastructure provider. Nebius's latest results reinforce its position as a compelling choice for large-scale inference and highlight its dedication to advancing AI infrastructure through continuous optimization and innovation.

Supplemental Results Discussion for MLPerf Inference v5.1

NVIDIA

In MLPerf Inference v5.1, the NVIDIA platform delivered excellent results across all existing and newly added benchmarks. This round, NVIDIA also made the world's first submissions of the GB300 NVL72 rack-scale system in the available category, which unifies 72 NVIDIA Blackwell Ultra GPUs and 36 Arm-based NVIDIA Grace CPUs in a single platform optimized for reasoning inference. This available submission comes just one round after the first available category submission of the Blackwell-based GB200 NVL72 system.

The Blackwell Ultra-powered GB300 NVL72 system delivered up to 40% more DeepSeek-R1 reasoning inference throughput compared to the prior generation. These gains were driven by several architectural enhancements targeted specifically at reasoning inference, including 15 petaFLOPS of dense NVFP4 throughput, greatly increased attention-layer acceleration, and up to 288GB of HBM3e memory.

NVIDIA also made its first submissions using NVIDIA Dynamo in the open division using the Blackwell-based GB200 NVL72 rack-scale system. By dedicating sets of Blackwell GPUs to each of the prefill and decode phases of inference and using high bandwidth NVLink and NVLink Switch for fast KV Cache transfer, disaggregated serving using NVIDIA Dynamo enabled up to 1.5x increase in throughput compared to traditional serving on the Llama 3.1 405B interactive benchmark. This means higher revenue potential for AI factories and lower cost per million tokens using the same GPU architecture.

This round, many NVIDIA partners submitted great results across the full range of NVIDIA GPU architectures, including ASUSTeK, Azure, Broadcom, Cisco, CoreWeave, Dell, GigaComputing, Google, HPE, Lambda, Lenovo, Nebius, Oracle, Quanta Cloud Technology, Supermicro, and the University of Florida.

Finally, we would like to commend MLCommons for their ongoing stewardship of the MLPerf benchmarks, ensuring that the industry has fair and transparent measures of AI performance that they can rely on to make important infrastructure decisions.

Supplemental Results Discussion for MLPerf Inference v5.1

Oracle

Oracle has delivered exceptional results in the MLPerf Inference v5.1 benchmark, highlighting the strengths of Oracle Cloud Infrastructure (OCI). Oracle submitted competitive benchmarks across a wide range of workloads. These benchmarks were executed on NVIDIA B200, and GB200 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, 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.

Supplemental Results Discussion for MLPerf Inference v5.1

Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global leader in data center solutions, continues to advance high-performance computing (HPC) and artificial intelligence (AI) with innovative system designs. In the latest **MLPerf™ Inference v5.1 benchmark**, QCT submitted results in the data center closed division, demonstrating diverse system configurations optimized for modern workloads.

- **QuantaGrid D75E-4U**: A flexible 4U system supporting up to eight GPU accelerators with dual Intel® Xeon® 6 processors. Tested in both CPU-only and GPU-accelerated modes, the CPU-only system with Intel® AMX delivers cost-efficient AI inference. The GPU configuration with four NVIDIA H200 NVL GPUs provides scalable performance for demanding HPC and AI applications.
- **QuantaGrid D74H-7U**: A powerful 7U platform supporting up to eight NVIDIA H200 SXM5 GPUs. Featuring NVLink® interconnect for high-speed GPU-to-GPU communication and GPUDirect Storage for direct, low-latency access between GPUs and storage, it eliminates data bottlenecks and delivers ultra-fast pipelines for large-scale AI inference and training.
- **QuantaGrid D75T-7U**: Designed for heterogeneous AI computing, the system is powered by dual AMD EPYC™ 9005 processors and up to eight AMD Instinct™ MI325X accelerators, each with 256GB HBMe memory, to handle the most compute-intensive AI workloads.

By participating in MLPerf, the industry's trusted AI benchmark, QCT underscores its commitment to validated performance, transparency, and customer confidence. These results empower enterprises, research institutions, and service providers to make data-driven infrastructure decisions with confidence.

Supplemental Results Discussion for MLPerf Inference v5.1

Red Hat, Inc

Red Hat, a leading provider of enterprise open source solutions, is proud to demonstrate impressive performance for the Llama-3.1-8B-FP8 model in MLPerf Inference v5.1 results, utilizing vLLM for inference on a Dell PowerEdge XE8640 server with H100 GPUs, and with an IBM Cloud gx3-48x240x2l40s instance with L40s GPUs. Our Llama-3.1-8b results showcase a cost-effective way to run AI inference powered by open source innovation. Red Hat AI Inference Server runs on Red Hat Enterprise Linux and Red Hat OpenShift offering enhanced performance, ease of use and a hardware agnostic solution for model serving and inference powered by Red Hat's fully supported, and production-ready vLLM builds.

Get fast inferencing with Red Hat AI Inference Server, powered by vLLM:

- Red Hat AI Inference Server allows you to run any open source generative AI model on any hardware accelerator on-premise and in the public cloud.
- Red Hat AI Inference Server optimizes model inference across the hybrid cloud. Built using open source technologies, it provides trusted, operationally consistent capabilities for teams to experiment, serve models, and deliver innovative apps. Red Hat AI Inference Server supports a wide array of hardware types with experiments and models, on-premise and in the public cloud.
- This submission demonstrates the inference performance of vLLM. vLLM is a fast and easy-to-use open source library for LLM inference and serving. vLLM is provided as a supported model serving runtime in Red Hat AI Inference Server as of version 3.
- This submission demonstrates the rapid advancement of the open source vLLM project. We were able to achieve impressive performance thanks to recent optimizations in the vLLM project including.
  - Integration of custom CUTLASS kernels for FP8 optimized low-level GEMM operations that maximize throughput on H100 and L40S tensor cores.
  - Adoption of FlashAttention-3 for more efficient attention computation, especially for long-context workloads.
  - Introduction of the vLLM v1 engine with improved scheduling, asynchronous execution, and reduced CPU overhead for better end-to-end throughput.
- Thanks to FP8 quantization, we were also able to demonstrate improved performance of Llama-3.1-8b on a small hardware footprint (NVIDIA L40S GPU) which is competitive with other state-of-the-art model runtimes.
- Get access to a 60 day free trial of Red Hat AI Inference Server here.

Supplemental Results Discussion for MLPerf Inference v5.1

Supermicro

Supermicro, Inc., a Total IT Solution Provider for AI, HPC, and Cloud, offers an entire line of NVIDIA HGX B200 systems, with both air-cooled and liquid-cooled options, with a choice of AMD EPYC or Intel® Xeon® 6 CPUs. We have submitted MLPerf 5.1 inference benchmarks for these HGX B200 systems. In addition, Supermicro is also publishing benchmark results for AMD Instinct™MI325X GPU systems using one and two-node configurations.

Supermicro demonstrated state-of-the-art inference performance, along with other vendors using the HGX-B200 8-GPU system for

- LLAMA 3.1 8B (server, offline, interactive)
- LLAMA 2 70B-99 (server, offline, interactive)
- LLAM2 2 70B-99.9 (server, offline, interactive)
- Whisper offline
- These results demonstrate the excellent inference performance of HGX-B200-8-GPU systems.

Supermicro was able to produce similar results in B200 systems using Intel CPU, AMD CPU, in both liquid and air cooling.

Working together, Supermicro and Mangoboost the highest results using 16 MI300X and 16 MI325X GPUs, in a networked configuration for the following:

- LLAMA 2-70B-99 (server, offline)
- LLAMA 2-70B-99.9 (server, offline)
- These results showed excellent performance using 32 x H100-SXM GPUs
- The AMD MI325X systems show the benefits of alternative GPU options and using the network to scale performance for inference.

The award-winning portfolio of Server Building Block Solutions® allows customers to optimize for their exact workload and application by selecting from a broad family of systems built from our flexible and reusable building blocks that support a comprehensive set of form factors, processors, storage, GPUs and other choices. Check out https://www.supermicro.com/en/products/gpu.

Supplemental Results Discussion for MLPerf Inference v5.1

TheStage AI

We are TheStage AI. We build tools that make neural networks run faster without losing what makes them unique. Our focus is simple – automate the hard parts of optimization so models can go from idea to production in hours, not months. From generative AI to vision systems, we work where speed and quality have to live side by side.

To make this practical, we developed ANNA (Automated Neural Networks Accelerator). Think of it like a finely tuned slider: it adjusts automatically to find the optimal balance between inference speed and output quality for each model. This approach allows teams to optimize without manually tuning dozens of parameters, saving both time and effort.

In this MLPerf® Inference v5.1 submission, we used Stable Diffusion XL (SDXL), a state-of-the-art text-to-image latent diffusion model, to address a clear challenge: improve inference performance for data centers while keeping image quality within a strict target range. This reflects real-world constraints, where a model's output must remain true to its baseline without unwanted shifts.

ANNA applies an automated search to quickly identify the point where speed and quality meet the target. Once the target point is reached, it is deployed using our compiler-based optimization, which integrates with existing pipelines in just a few lines of code. This makes the process reproducible, production-ready, and easy to adopt in large-scale environments.

While demonstrated with SDXL, the method is model-agnostic. It can be applied to other neural networks – from generative AI to computer vision – wherever inference efficiency and output consistency are critical. On an 8xH100 SXM GPU node, ANNA lets SDXL generate 18.1 images per second, illustrating its effectiveness for high-throughput, high-fidelity workloads. As a result, it is a practical solution for a wide range of data center applications.

Supplemental Results Discussion for MLPerf Inference v5.1

University of Florida

The University of Florida (UF) is making significant strides in AI research and education with the NVIDIA DGX B200 SuperPOD as part of the university's HiPerGator supercomputer. Our inaugural submission of MLPerf Inference results from this system confirms its AI performance is measurable, comparable, reproducible and accountable with healthy comparisons between peer institutions and industry. The results show our system delivers AI applications under realistic service constraints and provides an evidence base for future improvements to the computing infrastructure. Our benchmark results met MLPerf closed-division requirements and latency targets in the Server scenario with good scalability. This submission highlights the following significance and contributions:

1. UF's submission broadens the MLCommons' benchmarking ecosystem, validates the capabilities of university high performance computing (HPC) environments, and creates a pathway for students, scholars and public research labs to participate in rigorous, comparable AI model performance benchmarking.
2. All benchmarks were executed on the UF's NVIDIA B200 DGX SuperPOD using Apptainer to run a partner-supplied container image. This workflow integrates with the SLURM batch scheduler, preserving the integrity of the closed-division software stack while operating under realistic HPC constraints, shared parallel file systems, network and process isolation, and multi-tenancy.
3. Using Apptainer ensured a rootless container workflow suitable for multi-user clusters that did not provide Docker or Sudo. The recipes avoided host configuration changes and required standard CUDA and driver support, streamlining the method for academic and enterprise operators to incorporate and reproduce routine performance validation.

As the first academic institution to submit MLPerf Inference results, UF expanded the system architectures and operational realities in the MLPerf benchmarking community. We are committed to sharing operational experience, collaborating with peers and enabling more institutions to run compliant AI workloads on shared HPC infrastructure.

Supplemental Results Discussion for MLPerf Inference v5.1

Vultr

MLPerf® benchmark submitter Vultr is on a mission to make high-performance GPU cloud infrastructure easy to use, affordable, and locally accessible for enterprises and AI innovators around the world. Vultr is trusted by hundreds of thousands of active customers across 185 countries for its flexible, scalable, and global Cloud Compute, Cloud GPU, Bare Metal, and Cloud Storage solutions.

With a full line of the latest AMD and NVIDIA data center GPUs available, Vultr offers affordable, secure and compliant cloud GPU computing.

For Vultr's first MLPerf benchmark submission in MLPerf Inference v5.1, the submitted Supermicro AS-8126GS-TNMR with 8x AMD Instinct MI325X accelerators, delivered through Vultr Cloud GPU, showcased competitive performance across benchmarked AI models.

With the confirmation of these benchmarks, Vultr Cloud GPU customers can be confident that they are harnessing the latest in GPU acceleration for maximum performance and efficiency.

Supplemental Results Discussion for MLPerf Inference v5.1

Amitash Nanda

Amitash Nanda is a first-time individual contributor and MLPerf benchmark submitter. As a 3rd-year UC San Diego Ph.D. student, contributing to open source projects will broaden his knowledge in AI research and performance inference. His submission demonstrates that high-quality ML inference can be performed on widely accessible, energy-efficient consumer hardware. The benchmark was performed on an Apple MacBook Pro with the M1 Pro chip (10-core CPU, 16-core GPU, and 16 GB unified memory). The inference results are reliable and accurate for the offline scenario. The system utilized ONNX runtime v1.19.2 with Apple's CoreML execution to run inference on the integrated GPU and Apple Neural Engine (ANE). It showcases the potential of ARM-based platforms to contribute meaningfully to ML performance and accessibility. He implemented the official MLCommons reference implementation, with dataset preprocessing, model handling, and reproducibility managed entirely through mlcr automation tools.

The submission accuracy is higher than the targeted accuracy for an edge-class category. This submission is a valuable reference for the broader community seeking to evaluate or deploy ML workloads on general-purpose edge systems. By using a reproducible, standard pipeline and fully open automation tools, this work can help educators, students, developers, and researchers explore ML performance in real-world settings without needing specialized infrastructure. He is excited about the future of high-efficiency, high-portability ML inference and proud to contribute to the MLPerf Inference v5.1 results that promote transparency, collaboration, and community-driven benchmarking.