ML Commons

MLPerf Inference v5.0 Supplemental Results Discussion

The submitting organizations provided the following 300 word descriptions as a supplement to help the public understand their MLCommons® MLPerf® Inference v5.0 submissions and results. The statements **do not reflect the opinions or views of MLCommons.**

This information is under embargo until 4/2/25 8:00AM PT

Supplemental Results Discussion for MLPerf Inference v5.0

AMD

AMD is excited to announce strong MLPerf® Inference v5.0 results with the first-ever AMD Instinct™ MI325X submission and the first multi-node MI300X submission by a partner. This round highlights investment from AMD in AI scalability, performance, software advancements and open-source strategy, while demonstrating strong industry adoption through multiple partner submissions.

For the first time, multiple partners—Supermicro (SMC), ASUSTeK, and Giga Computing with MI325X, and MangoBoost with MI300X—submitted MLPerf results using AMD Instinct solutions. MI325X partner submissions for Llama 2 70B achieved comparable performance of AMD's own results, reinforcing the consistency and reliability of our GPUs across different environments.

AMD also expanded its MLPerf benchmarks by submitting Stable Diffusion XL (SDXL) with MI325X, demonstrating competitive performance in generative AI workloads. Innovative GPU partitioning techniques played a key role in optimizing SDXL inference performance.

MangoBoost achieved the first-ever multi-node MLPerf submission using AMD Instinct GPUs, leveraging four nodes of MI300X GPUs. This milestone demonstrates the scalability and efficiency of AMD Instinct for large-scale AI workloads. Additionally, it demonstrated significant scaling from an 8-GPU MI300X submission last round to a 4 node 32-GPU MI300X submission this round, further reinforcing the robustness of AMD solutions in multi-node deployments.

The transition from MI300X to MI325X delivered significant performance improvement in Llama 2 70B inference, enabled by rapid hardware and software innovations.

As AMD continues to drive AI performance forward, we remain committed to transparency, innovation, and delivering industry-leading results through MLPerf benchmarking.

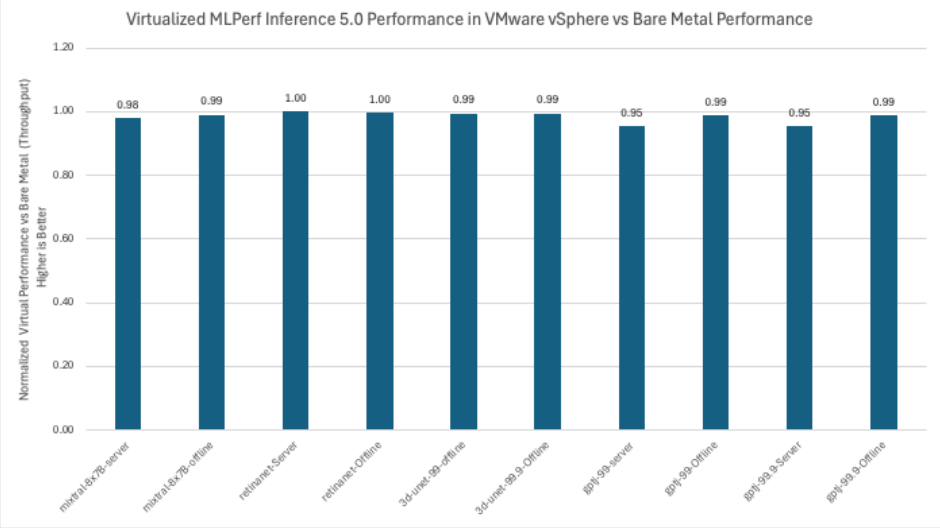
Supplemental Results Discussion for MLPerf Inference v5.0

Broadcom

Broadcom is pioneering VMware Private AI as an architectural approach that balances the business gains from AI and ML with the privacy and compliance needs of organizations. Built on top of the VMware Cloud Foundation (VCF) private cloud platform, this approach ensures privacy and control of data, choice of open-source and commercial AI solutions, optimum cost, performance, compliance, and best-in-class automation and load balancing.

Broadcom brings the power of virtualized NVIDIA GPUs to the VCF private cloud platform to simplify management of AI accelerated data centers and enable efficient application development and execution for demanding AI and ML workloads. VMware software supports various hardware vendors, facilitating scalable deployments.

Broadcom partnered with NVIDIA, Supermicro, and Dell Technologies to showcase virtualization's benefits, achieving impressive MLPerf® Inference v5.0 results. We demonstrated near bare-metal performance across diverse AI domains— Computer Vision, Medical Imaging, and Natural Language Processing using a six billion parameter GPT-J language model. We also achieved outstanding results with the Mixtral-8x7B 56 billion parameter large language model (LLM) . The graph below compares normalized virtual performance with bare-metal performance, showing minimal overhead from vSphere 8.0.3 with NVIDIA virtualized GPUs. Refer to the official MLCommons Inference 5.0 results for the raw comparison of queries per second or the tokens per second.



We ran MLPerf Inference v5.0 on eight virtualized NVIDIA SXM H100 80GB GPUs on SuperMicro GPU SuperServer SYS-821GE-TNRT and Dell PowerEdge XE9680. These VMs used only a fraction of the available resources—for example, just 25% of CPU cores, leaving 75% for other workloads. 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 robust data center management. Enterprises now gain both high-performance GPUs and vSphere's operational efficiencies.

Supplemental Results Discussion for MLPerf Inference v5.0

CTuning

The cTuning Foundation is a non-profit organization dedicated to advancing open-source, collaborative, and reproducible research in computer systems and machine learning. It develops tools and methodologies to automate performance optimization, facilitate experiment sharing, and improve reproducibility in research workflows.

cTuning leads the development of the open Collective Knowledge platform, powered by the MLCommons® CM workflow automation framework with MLPerf® automations. This educational initiative helps users to learn how to run AI, ML, and other emerging workloads in the most efficient and cost-effective way across diverse models, datasets, software, and hardware, based on the MLPerf methodology and tools.

In this submission round, we are testing the Collective Knowledge platform with a new prototype of Collective Mind framework (CMX) and MLPerf automations, developed in collaboration with ACM, MLCommons, and our volunteers and participants in open optimization challenges.

To learn more, please visit:

- https://github.com/mlcommons/ck
- https://access.cKnowledge.org
- https://cTuning.org/ae

Supplemental Results Discussion for MLPerf Inference v5.0

Cisco Systems Inc.

With generative AI poised to significantly boost global economic output, Cisco is helping to simplify the challenges of preparing organizations' infrastructure for AI implementation. The exponential growth of AI is transforming data center requirements, driving demand for scalable, accelerated computing infrastructure.

To this end, Cisco recently introduced the Cisco UCS C885A M8, a high-density GPU server designed for demanding AI workloads, offering powerful performance for model training, deep learning, and inference. Built on the NVIDIA HGX platform, it can scale out to deliver clusters of computing power that will bring your most ambitious AI projects to life. Each server includes NVIDIA NICs or SuperNICs to accelerate AI networking performance, as well as NVIDIA BlueField-3 DPUs to accelerate GPU access to data and enable robust, zero-trust security. The new Cisco UCS C885A M8 is Cisco's first entry into its dedicated AI server portfolio and its first eight-way accelerated computing system built on the NVIDIA HGX platform.

Cisco successfully submitted MLPerf® v5.0 Inference results in partnership with Intel and NVIDIA to enhance performance and efficiency, optimizing various inference workloads such as Large language model (Language) , Natural language processing (Language), Image Generation (Image), Generative image (Text to Image), Image classification (Vision), Object detection(Vision), Medical image segmentation (Vision) and Recommendation (Commerce).

Exceptional AI performance across Cisco UCS platforms.

- Cisco UCS C885A M8 platform with 8x NVIDIA H200 SXM GPUs
- Cisco UCS C885A M8 platform with 8x NVIDIA H100 SXM GPUs
- Cisco UCS C245 M8, X215 M8 + X440p PCIe node with 2x NVIDIA PCIe H100-NVL GPUs & 2x NVIDIA PCIe L40S GPUs
- Cisco UCS C240 M8 with Intel Granite Rapid 6787P processors

Supplemental Results Discussion for MLPerf Inference v5.0

CoreWeave

CoreWeave, the AI Hyperscaler™, today announced its MLPerf® Inference v5.0 results, setting outstanding performance benchmarks in AI inference. CoreWeave is the first cloud provider to submit results using NVIDIA GB200 GPUs.

Using a CoreWeave GB200 instance featuring two Grace CPUs and four Blackwell GPUs, CoreWeave delivered 800 tokens per second (TPS) on the Llama 3.1 405B model—one of the largest open-source models. CoreWeave also submitted results for NVIDIA H200 GPU instances, achieving over 33,000 TPS on the Llama 2 70B model benchmark.

“CoreWeave is committed to delivering cutting-edge infrastructure optimized for large-model inference through our purpose-built cloud platform,” said Peter Salanki, Chief Technology Officer at CoreWeave. “These benchmark MLPerf results reinforce CoreWeave’s position as a preferred cloud provider for some of the leading AI labs and enterprises.”

CoreWeave delivers performance gains through its fully integrated, purpose-built cloud platform. Our bare metal instances feature NVIDIA GB200 and H200 GPUs, high-performance CPUs, NVIDIA InfiniBand with Quantum switches, and Mission Control, which helps to keep every node running at peak efficiency.

This year, CoreWeave became the first to offer general availability of NVIDIA GB200 NVL72 instances. Last year, the company was among the first to offer NVIDIA H100 and H200 GPUs and one of the first to demo GB200s.

Supplemental Results Discussion for MLPerf Inference v5.0

Dell Technologies

MLPerf® Inference v5.0: Delivering High-performance AI Inference with Dell.

Dell, in collaboration with NVIDIA, Intel, and Broadcom, has delivered groundbreaking MLPerf Inference v5.0 results, showcasing industry-leading AI inference performance across a diverse range of accelerators and architectures.

Unmatched AI Performance Across Flagship Platforms

- PowerEdge XE9680 and XE9680L (liquid-cooled): These flagship AI inference servers support high-performance accelerators, including NVIDIA H100 and H200 SXM GPUs, ensuring unparalleled flexibility and compute power for demanding ML workloads.
- PowerEdge XE7745: This PCIe-based powerhouse, equipped with 8x NVIDIA H200-NVL or 8x L40S GPUs, demonstrated exceptional performance and industry-leading performance-per-watt efficiency, making it a strong contender for AI scaling at lower energy costs.

Breakthrough Results in Real-World LLM Inference

- Superior Performance Across Model Sizes: Dell's MLPerf submissions include results spanning small (GPT-J 6B), midsize (Llama 2 70B), and large-scale (Llama 3 405B) LLMs, highlighting consistent efficiency and acceleration across different AI workloads.
- Optimized for Latency-Sensitive AI: MLCommon added the Llama 2 70B-Interactive model to show how it works in real-time. We shared data to show how well it performs.
- First-Ever Llama 3 Benchmarks: With Llama 3 introduced in this round, Dell's results provide critical insights into next-generation inference workloads, further substantiating its leadership in AI infrastructure.

Data-Driven Performance Insights for Smarter AI Infrastructure Decisions

- PCIe vs. SXM Performance Analysis: Dell's MLPerf results deliver valuable comparative data between PCIe-based GPUs and SXM-based accelerators, equipping customers with the knowledge to optimize AI hardware selection for their specific workloads.

By delivering energy-efficient, high-performance, and latency-optimized AI inference solutions, Dell sets the benchmark for next-gen ML deployment, helping organizations confidently accelerate AI adoption. Generate higher quality, faster predictions and outputs while accelerating decision-making with powerful Dell Technologies solutions. Test drive them at our worldwide Customer Solution Centers or collaborate in our innovation labs to tap into a Centers of Excellence.

Supplemental Results Discussion for MLPerf Inference v5.0

FlexAI

FlexAI is a Paris-based company founded by industry veterans from Apple, Intel, NVIDIA, and Tesla. It specializes in optimizing and simplifying AI workloads through its Workload as a Service (WaaS) platform. This platform dynamically scales, adapts, and self-recovers, enabling developers to train, fine-tune, and deploy AI models faster, at lower costs, and with reduced complexity.

For this submission round, FlexAI shared with the community a prototype of a simple, open-source tool based on MLPerf® LoadGen to benchmark the out-of-the-box performance and accuracy of non-MLPerf models and datasets from the Hugging Face Hub, using vLLM and other inference engines across commodity software and hardware stacks.

We validated this prototype in our open submission with the DeepSeek R1 and Llama 3.3 models.

An advanced version of this tool—with full automation and optimization workflows—is available in the FlexAI cloud. To learn more, please visit: https://flex.ai.

Supplemental Results Discussion for MLPerf Inference v5.0

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 to inference and training rounds for both data center and edge divisions.

In this round, we focused on PRIMERGY CDI, equipped with external boxes compatible with PCIe Gen.5, which has 8x H100 NVL GPUs, submitting two divisions: the data center closed division and its power division.

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 chassis. The most remarkable feature of PRIMERGY CDI is the ability to freely allocate devices within the PCIe boxes to multiple computing servers. For instance, you can reduce the number of GPUs for inference tasks during the day and increase them for training tasks at night. This flexibility in GPU allocation allows for reduced server standby power without occupying GPUs for specific workloads.

In this round, the PRIMERGY CDI system equipped with 8x H100 NVL GPUs achieved outstanding results across seven benchmark programs, including mixtral-8x7b, which could not be submitted in the previous round, as well as newly added llama 2 70b-interactive and RGAT.

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 Inference v5.0

GATEOverflow

GATEOverflow, an education initiative based in India, is pleased to announce its first MLPerf® Inference v5.0 submission, reflecting our ongoing efforts in machine learning benchmarking. This submission was driven by the active involvement of our students, supported by GO Classes, fostering hands-on experience in real-world ML performance evaluation.

Our results—over 15,000+ performance benchmarks—were generated using MLCFlow, the automation framework from MLCommons®, and deployed across a diverse range of hardware, including laptops, workstations, and cloud platforms such as AWS, GCP, and Azure.

Notably, GATEOverflow contributed to over 80% of all closed Edge submissions. We are also the only power submitter in the Edge category and the sole submitter for the newly introduced PointPainting model.

This submission underscores GATEOverflow's dedication to open, transparent, and reproducible benchmarking. We extend our gratitude to all participants and contributors who played a role in developing MLCFlow automation and ensuring the success of this submission.

With this achievement, we look forward to further innovations in AI benchmarking and expanding our collaborations within the MLPerf community.

Supplemental Results Discussion for MLPerf Inference v5.0

Giga Computing

Giga Computing, a GIGABYTE subsidiary, specializes in server hardware and advanced cooling solutions. Operating independently, it delivers high-performance computing for data centers, edge environments, HPC, AI, data analytics, 5G, and cloud. With strong industry partnerships, it drives innovation in performance, security, scalability, and sustainability. Giga Computing leverages the widely recognized GIGABYTE brand at expos, participating under the GIGABYTE banner.

As a founding member of MLCommons®, Giga Computing continues to support the community's efforts in benchmarking server solutions for AI training and inference workloads. In the latest MLPerf® Inference v5.0 benchmarks, Giga Computing submitted test results based on the GIGABYTE G893 air-cooled series equipped with the most advanced accelerators, including the AMD Instinct™ MI325X and the NVIDIA HGX™ H200. These systems showcase industry-leading performance and provide comprehensive test results across all mainstream AI platforms.

These systems excel in high data bandwidth, large memory capacity, optimized GPU resource allocation, and unique data transfer solutions such as InfiniBand, Infinity Fabric™, and all-Ethernet designs. With a thoroughly optimized thermal design and verified systems, our results speak for themselves—delivering outstanding efficiency while maintaining top-tier performance across all benchmarked tasks.

At Giga Computing, we are committed to continual improvement, offering remote testing and public benchmarks for system evaluations. We also lead in advanced cooling technologies, including immersion and direct liquid cooling (DLC), to address the growing power demands of modern computing. Stay tuned as we push the boundaries of computing excellence with Giga Computing.

Supplemental Results Discussion for MLPerf Inference v5.0

Google Cloud

For MLPerf® Inference v5.0, Google Cloud submitted 15 results, including its first submission with A3 Ultra (NVIDIA H200) and A4 (NVIDIA HGX B200) VMs, and a second-time submission for Trillium, the sixth-generation TPU. The strong results demonstrate the performance of its AI Hypercomputer, bringing together AI optimized hardware, software, and consumption models to improve productivity and efficiency.

A3 Ultra VM is powered by eight NVIDIA H200 Tensor Core GPUs and offers 3.2 Tbps of GPU-to-GPU non-blocking network bandwidth and twice the high bandwidth memory (HBM) compared to A3 Mega with NVIDIA H100 GPUs. Google Cloud's **A3 Ultra demonstrated highly competitive performance across LLMs, MoE, image, and recommendation models.** In addition, **A4 VMs in preview, powered by NVIDIA HGX B200 GPUs, achieved standout results among comparable GPU submissions.** A3 Ultra and A4 VMs deliver powerful inference performance, a testament to Google Cloud's continued close partnership with NVIDIA to provide infrastructure for the most demanding AI workloads.

Google Cloud's Trillium, the sixth-generation TPU, delivers our highest inference performance yet. Trillium continues to achieve standout performance on compute-heavy workloads like image generation, further improving Stable Diffusion XL (SDXL) throughput by 12% since the MLPerf v4.1 submission. **Trillium now delivers 3.5x throughput improvement for queries/second on SDXL** compared to the performance demonstrated in the last MLPerf round by its predecessor, TPU v5e. This is driven by Trillium's purpose-built architecture and advancements in the open software stack, specifically on inference frameworks, to leverage the increased compute power.

Supplemental Results Discussion for MLPerf Inference v5.0

Hewlett Packard Enterprise

This is the eighth round Hewlett Packard Enterprise (HPE) has joined since v1.0 and our growing portfolio of high-performance servers, storage, and networking products have consistently demonstrated strong AI inference results. In this round, HPE submitted multiple new configurations from our Compute and High Performance Computing (HPC) server families with our partner, NVIDIA.

Our HPE ProLiant Compute Gen12 portfolio offers servers optimized for performance and efficiency to support a wide variety of AI models and inference budgets. Highlights include:

- HPE ProLiant Compute DL380a Gen12 – supporting eight NVIDIA H200 NVL, H100 NVL, or L40S GPUs per server – and NVIDIA TensorRT LLM, more than doubled inference throughput since the last round thanks to upgrades and performance optimizations.
- HPE ProLiant Compute DL384 Gen12 with dual-socket (2P) NVIDIA GH200 144GB demonstrated twice the performance of our single-socket (1P) NVIDIA GH200 144GB submission – a significant first in HPE's MLPerf results.
- HPE's first-ever submission of a HPE Compute Scale-up Server 3200 featured a 1-to-1 mapping between four Intel CPUs and four NVIDIA H100 GPUs, offering high-performance reliability and scalable inference.

The HPE Cray XD portfolio delivers high-performance for a variety of training and inference use cases. Highlights on inference results include:

- HPE Cray XD670 air-cooled servers with 8-GPU NVIDIA HGX H200 and H100 baseboards delivered our highest MLPerf inference performance to date.
- All HPE Cray results used HPE GreenLake for File Storage or HPE Cray ClusterStor E1000 Storage Systems to host datasets and model checkpoints, proving that high-throughput inference can be obtained without moving datasets and checkpoints to local disks.

HPE would like to thank the MLCommons® community and our partners for their continued innovation and for making MLPerf® the industry standard to measure AI performance.

Supplemental Results Discussion for MLPerf Inference v5.0

Intel

The latest MLPerf® Inference v5.0 results reaffirm Intel® Xeon® 6 with P-cores strength in AI inference and general-purpose AI workloads.

Across six MLPerf benchmarks, 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, including classical machine learning, small- to mid-size models, and relational graph node classification.

Since first submitting Xeon to MLPerf in 2021 (with 3rd Gen Intel® Xeon® Scalable Processors), Intel has achieved up to a 15x performance increase, driven by hardware and software advancements, with recent optimizations improving results by 22% over v4.1.

Notably, Intel is the only vendor submitting server CPU results to MLPerf. And Xeon also continues to be the host CPU of choice for accelerated systems.

Intel also supported multiple customers with their submissions and collaborated with OEM partners – Cisco, Dell Technologies, Quanta, and Supermicro – to deliver MLPerf submissions powered by Intel Xeon 6 with P-cores.

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

For more details, please see MLCommons.org.

Supplemental Results Discussion for MLPerf Inference v5.0

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 11 out of 11 MLPerf® Inference rounds, having contributed to some of the fastest and most energy efficient results in MLPerf history in collaboration with leading vendors. As we know firsthand how much effort (months to years) goes into preparing highly competitive and fully compliant MLPerf submissions, we set out to demonstrate what can be achieved with more realistic effort (days to weeks).

As the only submitter of the state-of-the-art **DeepSeek-v3-671B** Open Division workload, we compared it against **Llama 3.1 405B** on **8x MI300X-192GB GPUs**. Surprisingly, (sparse) DeepSeek-v3 was slower than (dense) Llama 3.1 405B in terms of tokens per second (TPS: 243.45 vs 278.65), as well as less accurate (e.g. rougeL: 18.95 vs 21.63).

Furthermore, we compared the performance of **Llama 3.1 70B** using several publicly released Docker images on **8x H200-141GB GPUs**, achieving Offline scores of up to 30,530 TPS with **NIM** v1.5.0, 27,950 TPS with **SGLang** v0.4.3, and 21,372 TPS with **vLLM** v0.6.4. Curiously, NIM was slower than SGLang in terms of queries per second (QPS: 94.26 vs 95.65). This is explained by NIM also being less accurate than SGLang (e.g. rouge1: 46.52 vs 47.57), while generating more tokens per sample on average. In the new **interactive Server** category with much stricter latency constraints, NIM achieved 15,960 TPS vs 28,421 TPS for non-interactive Server and 30,530 TPS for Offline.

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.

Supplemental Results Discussion for MLPerf Inference v5.0

Lambda

About Lambda

Lambda was founded in 2012 by AI engineers with published research at the top machine learning conferences in the world. Our goal is to become the #1 AI compute platform serving developers across the entire AI development lifecycle.

Lambda enables AI engineers to easily, securely and affordably build, test and deploy AI products at scale. Our product portfolio spans from on-prem GPU hardware to hosted GPUs in the cloud - in both Public and Private settings.

Lambda's mission is to create a world where access to computation is as effortless and ubiquitous as electricity.

About the Benchmarks

This is our team's first participation in the MLPerf® inference round of submissions, in partnership with NVIDIA. Our benchmarks are run on two Lambda Cloud 1-Click Clusters™:

- 1-Click Cluster with eight NVIDIA H200 141GB SXM GPUs, 112 CPU Cores, 2 TB RAM, and 28 TB SSD.
- 1-Click Cluster with eight NVIDIA B200 180GB SXM GPUs, 112 CPU Cores, 2.9 TB RAM, and 28 TB SSD.

Our inference benchmark runs for models like DLRM, GPTJ, Llama 2 and Mixtral on 8 GPUs demonstrate state-of-the-art time-to-solution on Lambda Cloud 1-Click Clusters. We noted significant throughput improvements on our clusters, with performance levels making them prime systems for our customers' most demanding inference workloads.

Lenovo

Empowering Innovation: Lenovo's AI Journey

At Lenovo, we're passionate about harnessing the power of AI to transform industries and lives. To achieve this vision, we invest in cutting-edge research, rigorous testing, and collaboration with industry leaders through MLPerf® Inference v5.0.

Driving Excellence Through Partnership and Collaboration

Our partnership with MLCommons® enables us to demonstrate our AI solutions' performance and capabilities quarterly, showcasing our commitment to innovation and customer satisfaction. By working together with industry pioneers like NVIDIA on critical applications such as image classification and natural language processing, we've achieved outstanding results.

Unlocking Innovation: The Power of Lenovo ThinkSystem

Our powerful ThinkSystems SR675 V3, SR680a V3, SR6780a V3, and SR650a V4, equipped with NVIDIA GPUs, enable us to develop and deploy AI-powered solutions that drive business outcomes and improve customer experiences. We're proud to have participated in these challenges using our ThinkSystem alongside NVIDIA GPUs.

Partnership for Growth: Enhancing Product Development

Our partnership with MLCommons provides valuable insights into how our AI solutions compare against the competition, sets customer expectations, and empowers us to continuously enhance our products. Through this collaboration, we can work closely with industry experts to drive growth and deliver better products for our customers.

By empowering innovation and driving excellence in AI development, we're committed to delivering unparalleled experiences and outcomes for our customers. Our partnerships with leading organizations like NVIDIA have been instrumental in achieving these goals. Together, we're shaping the future of AI innovation and making it accessible to everyone.

Supplemental Results Discussion for MLPerf Inference v5.0

MangoBoost

MangoBoost is a DPU and system software company aiming to offload network and storage tasks to improve performance, scalability and efficiency of modern datacenters. We are excited to introduce **Mango LLMBoost, a ready-to-deploy AI inference serving solution that delivers high performance, scalability, and flexibility**. LLMBoost software stack is designed to be seamless and flexible across a diverse range of GPUs. In this submission, MangoBoost partnered up with AMD and showcased the performance of LLMBoost as a highly-scalable multi-node inference serving solution. **Our results for LLaMA2-70B on 32 MI300x GPUs showcase linear performance scaling in both server and offline scenarios, achieving 103k and 93k TPS** respectively. This result highlights how LLMBoost enables effortless LLM scaling—from a single-GPU proof of concept to large-scale, multi-GPU deployments. Best of all, anyone can replicate our results with ease by downloading our Docker image (https://hub.docker.com/r/llmboost/mb-llmboost) and running a single command.

Mango LLMBoost has been rigorously tested and proven to support inference serving on multiple GPU architectures from AMD and NVIDIA, and is compatible with popular open models such as the Llama, Qwen, DeepSeek, and multi-modal such as Llava. LLMBoost's one-line deployment, robust OpenAI API and REST API integration enable developers and data scientists to rapidly incorporate LLMBoost inference serving into their existing software ecosystems. **LLMBoost is available and free to try in leading cloud environments including Azure, AWS, and GCP as well as on-premise.**

MangoBoost's suite of DPU and hardware solutions provides complementary hardware acceleration for inference serving, significantly reducing system overhead in modern AI/ML workloads. Mango GPUBoost RDMA accelerates multi-node inference and training over RoCEv2 networks, while Mango NetworkBoost offloads web-serving and TCP stacks, dramatically reducing CPU utilization. Mango StorageBoost delivers high-performance storage initiator and target solutions, enabling AI storage systems—including traditional and JBoF systems—to scale efficiently.

Supplemental Results Discussion for MLPerf Inference v5.0

NVIDIA

In MLPerf® Inference v5.0, NVIDIA and its partners submitted many outstanding results across both the Hopper and Blackwell platforms.

NVIDIA made its first MLPerf submission using the GB200 NVL72 rack-scale system, featuring 72 Blackwell GPUs and 36 Grace CPUs in a rack-scale, liquid-cooled design. It features a 72-GPU NVLink domain that acts as a single, massive GPU. This system achieved Llama 3.1 405B token throughput of 13,886 tokens per second in the offline scenario and achieved 30X higher throughput in the server scenario than the NVIDIA submission using eight Hopper GPUs.

NVIDIA also submitted results using a system with eight B200 GPUs, connected over NVLink, including on the Llama 2 70B Interactive benchmark, a version of the Llama 2 70B benchmark with tighter time-to-first token and token-to-token latency constraints. On this benchmark, eight Blackwell GPUs delivered triple the token throughput compared to the same number of Hopper GPUs.

Hopper performance also saw increases, bringing even more value to the Hopper installed base. Through full-stack optimizations, the Hopper architecture delivered a cumulative performance improvement of 1.6X in one year on the Llama 2 70B benchmark among results in the “available” category. Hopper also continued to deliver great results on all benchmarks, including on the new Llama 2 70B Interactive, Llama 3.1 405B, and GNN benchmarks.

15 NVIDIA partners submitted great results on the NVIDIA platform, including ASUSTek, Broadcom, CoreWeave, Cisco, Dell Technologies, Fujitsu, GigaComputing, Google, HPE, Lambda, Lenovo, Oracle, Quanta Cloud Technology, Supermicro, and Sustainable Metal Cloud.

NVIDIA would also like to commend MLCommons® for their ongoing commitment to develop and promote objective and useful measurements of AI platform performance.

Supplemental Results Discussion for MLPerf Inference v5.0

Oracle

Oracle has delivered stellar MLPerf® Inference v5.0 results, showcasing the strengths of OCI's Cloud Offering in providing industry-leading AI inference performance. Oracle successfully submitted results across various inference workloads such as image classification, object detection, LLM – Q&A, summarization, text and code generation, node classification, recommendation for commerce. The inference benchmark results for the high-end NVIDIA H200 bare metal instance demonstrate that OCI provides high performance and throughput that are essential to delivering fast, cost-effective, and scalable LLM inference across various applications.

Oracle Cloud Infrastructure (OCI) offers AI Infrastructure, Generative AI, AI Services, ML Services, and AI in our Fusion Applications. OCI has various options for GPUs for workloads ranging from small-scale to mid-level and large-scale AI. For the smallest workloads, we offer both bare metal(BM) and VM instances with *NVIDIA A10 GPUs*. For mid-level scale-out workloads, we offer the *NVIDIA A100 VMs, BMs and NVIDIA L40S Bare Metal with GPUs*. For the largest inference and foundational model training workloads, we offer *NVIDIA A100 80GB, H100 GPU, NVIDIA H200, NVIDIA GB200 and NVIDIA GB300* that can scale from one node to tens of thousands of GPUs.

Generative AI workloads drive a different set of engineering tradeoffs than traditional cloud workloads. So, we designed a purpose-built GenAI network tailored to the needs of the best-of-breed Gen AI workloads. Oracle Cloud Infrastructure (OCI) offers many unique services, including cluster network, an ultra-high performance network with support for remote direct memory access (RDMA) to support high-throughput and latency-sensitive nature of Gen AI workloads as well as high performance storage solutions.

Supplemental Results Discussion for MLPerf Inference v5.0

Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global leader in data center solutions, continues to drive innovation in HPC and AI with best-in-class system designs tailored to meet the evolving demands of modern computational workloads. As part of its commitment to delivering cutting-edge performance, QCT participated in the latest MLPerf® Inference v5.0 benchmark, submitting results in the data center closed division across various system configurations.

In this round of MLPerf Inference v5.0 submissions, QCT showcased systems based on different compute architectures, including CPU-centric and GPU-centric systems.

CPU-Centric Systems:

- **QuantaGrid D55X-1U** – A high-density 1U server powered by dual Intel® Xeon® 6 processors as the primary compute engines. With AMX instruction set support, it delivers optimized inference performance across general-purpose AI models such as ResNet-50, RetinaNet, 3D-UNet, and DLRMv2, as well as small language models like GPT-J-6B—offering a cost-efficient alternative to GPU-based solutions.

GPU-Centric Systems:

- **QuantaGrid D54U-3U** – A highly flexible 3U x86_64-based platform supporting up to four dual-width or eight single-width PCIe GPUs, allowing users to customize configurations based on their workload requirements.
- **QuantaGrid D74H-7U** – A powerful 7U x86_64-based system designed for large-scale AI workloads, supporting up to eight NVIDIA H100 SXM5 GPUs. Leveraging NVLink® interconnect, it enables high-speed GPU-to-GPU communication and supports GPUDirect Storage, ensuring ultra-fast, low-latency data transfers.
- **QuantaGrid S74G-2U** – A next-generation 2U system based on the NVIDIA® Grace™ Hopper™ Superchip. The integration of CPU and GPU via NVLink® C2C establishes a unified memory architecture, allowing seamless data access and improved computational efficiency.

QCT's comprehensive AI infrastructure solutions cater to a diverse user base, from academic researchers to enterprise clients, who require robust and scalable AI infrastructure. By actively participating in MLPerf benchmarking and openly sharing results, QCT underscores its commitment to transparency and reliability, empowering customers to make data-driven decisions based on validated performance metrics.

Supermicro

Supermicro offers a comprehensive AI portfolio with over 100 GPU-optimized systems, both air-cooled and liquid cooled options, with choice of Intel, AMD, NVIDIA, and Ampere CPUs. These systems can be integrated into racks, along with storage and networking for deployment in data centers. Supermicro is shipping NVIDIA HGX 8-GPU B200 systems, as well as NVL72 GPU systems, along with an extensive list of other systems.