ML Commons

MLCommons Supplemental Discussion

MLPerf Inference v6.1 Results Discussion

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

ML Commons

AMD

For the MLPerf Inference v6.1 round, AMD is pleased to share its broadest inference submission to date, expanding from three models in v6.0 to seven models across three AMD Instinct™ GPU platforms: MI355X, MI350X, and MI350P. The submission spans DLRM-v3, Llama 2-70B-99.9, Wan-2.2-T2V-A14B, and GPT-OSS-120B, covering recommendation, large language, reasoning, text-to-video, and multimodal workloads under MLCommons rules. Results span Server, Offline, Interactive, and Single Stream scenarios, providing a broader view of request-driven serving, batch throughput, responsiveness, and media-generation latency.

At single-node scale, the AMD Instinct™ MI355X GPU demonstrates broad capability across an expanded workload set and multiple MLPerf scenarios. The AMD Instinct™ MI350P GPU further broadens deployment options, extending the AMD Instinct portfolio across a wider range of system designs and inference requirements.

This round continues the AMD scale story. A 72-GPU GPT-OSS-120B submission exhibits excellent scalability of 95% in both Offline and Server scenarios and reaches 1 million tokens/second of performance. In addition, in collaboration with Crusoe, we have submitted GPT-OSS-120B and DeepSeek-R1 results using 512 AMD Instinct™ MI355X GPUs, representing the largest scale MLPerf Inference result to date.

Together, these results demonstrate progress in model coverage, competitive performance, and multi-node scale. The same AMD ROCm™ software foundation supports all three GPU platforms, enabling faster model bring-up, optimized kernels and data type paths, efficient communication, and reproducible performance across diverse inference workloads. The submission reinforces the AMD commitment to transparent benchmarking, open software, ecosystem collaboration, and production-ready inference from single-node to clusters.

ML Commons

ASUSTeK

ASUS is proud to present its MLPerf Inference v6.1 results, submitted in the Closed Division, Datacenter category, by the ASUS server engineering team. The submission features the ASUS XA NB3I-E12, an air-cooled AI server built on the NVIDIA HGX B300 platform, pairing eight NVIDIA Blackwell Ultra GPUs with dual Intel® Xeon® 6 processors, and running NVIDIA TensorRT 10.14 with CUDA 13.1.

The system delivered excellent throughput across today's most demanding generative-AI workloads. Highlights include 113,455 tokens/s in the Llama 2-70B Server scenario and 113,007 tokens/s in Offline; more than 109,000 tokens/s on GPT-OSS-120B; and 69,847 tokens/s Offline on the DeepSeek-R1 reasoning model — demonstrating consistently strong performance from high-concurrency interactive serving to peak batch processing.

As an active MLCommons member, ASUS believes transparent, reproducible benchmarking empowers enterprises and researchers to right-size AI infrastructure with confidence and accelerates responsible AI deployment for the broader community. These results reflect close collaboration across ASUS system design, firmware, and thermal engineering teams, and extend the ASUS AI Factory vision of delivering ready-to-deploy, full-stack AI infrastructure — helping more organizations bring large-scale generative AI into production.

ML Commons

Atlas Inference

For the past few years, serious agentic work meant a datacenter round trip. That assumption is what this submission is meant to retire. Local model capabilities and accessible compute have reached a level that everyday users can experience for themselves without the cloud.

The future involves an assistant that's always on, holds context private, and runs sessions for hours with the tool calls touching a real filesystem. This is inherently a different relationship than one you rent by the token. Sovereignty over your own action is what makes this relationship worth securing. Atlas Inference formed and submitted the edge agentic workload on NVIDIA DGX Spark (GB10) and AMD Strix Halo (gfx1151), the bread and butter machines for local AI.

Quantization keeps dense models resident, responsive, and rapidly generating on devices you can put on your desk. The results are excellent across performance and accuracy benchmarks. On DGX Spark, 1007 agentic coding turns completed in under 64 minutes at 20.1 tokens per second with time taken to serve just 40 seconds after a cold start. On Strix Halo, the same engine and recipe delivered 19.63 tokens per second. One engine, two completely different architectures and powerhouse vendors, empowering the community.

A developer choosing an agentic stack should not be choosing an accelerator vendor for the next three years, and an edge result on one device should tell you something useful and reproducible on the next one.

Benchmarking the top of the line local models for real-life use cases on the edge gives the vibrant local AI community a shared definition of a category that did not have one. We are proud to help establish it.

To learn more about Atlas Inference, follow us on X at @atlasinference

For more information, contact debaterishaqui@gmail.com

ML Commons

Cisco

Cisco UCS and Cisco Silicon One™: The Open Architecture for Enterprise AI

Enterprise AI demands peak performance without architectural lock-in. In MLPerf® Inference v6.1, full-stack infrastructure uniting **Cisco UCS® compute** and **Cisco Silicon One™ networking** delivered outstanding choice, linear scale, and maximum efficiency across generative AI.

Breaking Silicon Silos: MLPerf's First Cross-Vendor Heterogeneous Submission

In an MLPerf milestone, Cisco submitted the benchmark's first **cross-vendor heterogeneous accelerator deployment**, unifying eight NVIDIA H200 and eight AMD Instinct™ MI350X GPUs into one inference pool over a Cisco Silicon One G200 fabric with Intelligent Packet Flow.

The deployment achieved **99.9% aggregation efficiency** on Llama 2 70B (99,684 tokens/s) and **99.8%** on Llama 3.1 8B (163,292 tokens/s) over standalone baselines. By eliminating cross-architecture penalties, enterprises can seamlessly combine mixed accelerator investments into one high-performance inference engine.

Unlocking Linear Scale

Cisco showcased purpose-built infrastructure across every enterprise tier:

- **Enterprise PCIe Density: UCS C845A M8** (8x RTX Pro 6000/4500) and **C240 M8** (5x RTX Pro 4500) delivered balanced execution across LLMs and Whisper speech-to-text.
- **Modular Agility & CPU Nodes:** Modular **UCS X-Series with X580P PCIe nodes** (X210c/X215c M8) achieved exact 2x linear scaling from 4 to 8 RTX Pro GPUs, while 4-socket **UCS X410 M8** servers provided dense CPU inferencing.
- **Multi-Node Fabric Scale:** Multi-node B300 clusters on a dual-plane RoCEv2 fabric scaled DeepSeek-R1 to 137,067 tokens/s at **97.5% linear scaling efficiency** over single-node performance.

From single nodes to fabrics, Cisco delivers an open foundation powering enterprise AI.

ML Commons

CoreWeave

In MLPerf® Inference v6.1, CoreWeave participated in the Data Center Closed division, demonstrating exceptional results across multiple platforms of NVIDIA accelerators. Submissions spanned four frontier architectures including DeepSeek-R1, Qwen3-VL-235B, GPT-OSS-120B, and Llama2-70B, delivering strong, consistent throughput, reliability, and efficiency with frontier models in Server and Offline modes.

- **DeepSeek-R1:** On this 671B parameter Mixture-of-Experts model, CoreWeave achieved competitive per-GPU and aggregate throughput in Server and Offline scenarios on NVIDIA Grace Blackwell and Blackwell platforms.
- **Qwen3-VL-235B:** On this 235B-parameter multimodal model, CoreWeave demonstrated strong per-GPU and total system throughput in Server and Offline modes.
- **GPT-OSS-120B:** CoreWeave delivered exceptional results across Grace Blackwell and Blackwell systems. With GB300 NVL72, CoreWeave achieved an aggregate total throughput of 1.16 million tokens per second, per-GPU throughput improved 17% in Offline and 8% in Server since MLPerf v6.0, while NVIDIA HGX B200 excelled in per-GPU and total throughput.
- **Llama2-70B:** Across Grace Blackwell and Blackwell platforms, CoreWeave delivered strong per-GPU and total throughput with HGX B200 producing exceptional Server and Offline results.

These results reflect the engineering and operational expertise behind CoreWeave's production AI platform. CoreWeave Mission Control provides fleet-wide health monitoring and lifecycle management alongside continuous benchmarking. Topology-aware scheduling pins inference workloads within high-bandwidth NVIDIA NVLink™ domains, while integrated object storage and node-local caching accelerate model startup and stabilize execution.

CoreWeave turns raw accelerator performance into sustained, predictable AI performance at production scale. Across the AI loop from training and inference to observability, evaluation, and continuous improvement, CoreWeave optimizes infrastructure and performance at every stage, helping organizations iterate faster while maximizing throughput, efficiency, and reliability across every generation of accelerators.

ML Commons

Crusoe

Crusoe makes its debut in MLPerf® Inference v6.1, in the Datacenter Closed division with results on Crusoe Cloud AMD Instinct MI355X and NVIDIA GB200 NVL72 systems, covering three of today's most widely deployed open models: DeepSeek-R1, gpt-oss-120b, and Qwen3-VL-235B-A22B.

The centerpiece: the first 512-GPU submissions for DeepSeek-R1 and gpt-oss-120b, run across 64 Crusoe Cloud mi355x-288gb-roce.8x nodes on a standard RoCE Ethernet fabric. The cluster sustained over 5.7 million tokens per second on gpt-oss-120b and over 2.9 million tokens per second on DeepSeek-R1 across Server and Offline scenarios, throughput that reflects what a production fleet delivers when every GPU stays healthy and busy.

For DeepSeek-R1, an SGLang backend ran one 8-GPU replica per node (TP8 / EP8 / DP-attention 8) with MXFP4 expert weights, an FP8 KV cache, and BF16 MLA-absorb. For gpt-oss-120b, vLLM served the model in its native MXFP4 precision as 512 independent single-GPU replicas, with a dedicated head node dispatching queries and relaying token streams. In both cases, throughput scaled near-linearly from a single node to all sixty-four.

On NVIDIA GB200 NVL72, Crusoe submitted Qwen3-VL-235B-A22B with NVIDIA Dynamo disaggregated serving across Server, Offline, and Interactive scenarios, plus gpt-oss-120b with TensorRT-LLM.

Results like these start long before benchmark day. Every Crusoe cluster passes comprehensive burn-in before delivery, and continuous fleet health monitoring keeps it performing throughout its production life. Crusoe Managed Kubernetes, Managed Slurm, and fault-tolerant AutoClusters remove the operational overhead of running AI at scale.

Crusoe is the energy-first AI factory company, designing, building, and operating the data centers, energy, and cloud platform behind next-generation AI. Crusoe Cloud offers everything from on-demand VMs to large-scale clusters, and Crusoe Intelligence Foundry provides Managed AI for open models. Every configuration submitted here is available today, so teams can reproduce these results on the same infrastructure.

ML Commons

Dell

Dell Technologies announced a broad OEM submission portfolio in MLPerf® Inference v6.1, demonstrating its continued commitment to transparent, standards-based benchmarking across an expanding range of AI inference workloads. Dell's submission spans large language models, reasoning, multimodal AI, text-to-video generation, recommendation, computer vision, and emerging generative AI applications.

Dell benchmarked a diverse portfolio of PowerEdge platforms, including the PowerEdge XE9780/XE9780L/XE9780LAP, PowerEdge XE9785L, PowerEdge XE7740/XE7745, and the large-scale NVIDIA GB300 NVL72. The portfolio includes both NVIDIA and AMD accelerators, providing customers with validated performance data across different accelerator architectures, system designs, and deployment scales.

With AMD Instinct accelerators, the PowerEdge XE9785L with 8x AMD Instinct MI355X GPUs delivered exceptional performance across Llama2 70B and GPT-OSS-120B, demonstrating the platform's ability to efficiently serve both established and next-generation generative AI workloads. Dell also expanded coverage with the PowerEdge XE7745 using 8x AMD Instinct MI350P GPUs including MXFP4 (W4A4) weight and FP8 KV-cache precision recipe.

With NVIDIA accelerators, Dell submitted multiple 8-GPU B300 configurations using TensorRT 10.14, CUDA 13.1, and TensorRT-LLM, alongside RTX PRO 6000 and RTX PRO 4500 platforms. At larger scale, the 18-node, 72-GPU NVIDIA GB300 NVL72 extends Dell's validated inference portfolio to large-scale accelerated infrastructure and emerging workloads such as WAN 2.2 text-to-video generation.

Dell also extended its portfolio to edge AI with the Dell Pro Max powered by NVIDIA GB10, delivering MLPerf Inference Edge YOLO results in a compact, power-efficient platform. This submission demonstrates Dell's focus on bringing validated AI inference performance beyond the datacenter and closer to where AI applications and data are generated.

Together, these MLPerf Inference v6.1 submissions demonstrate Dell Technologies' commitment to rigorous benchmarking and help the broader AI community evaluate infrastructure across the rapidly expanding range of AI inference workloads.

ML Commons

Fujitsu Limited

Fujitsu offers a superior blend of systems, solutions, and expertise, ensuring maximum productivity, efficiency, and flexibility to deliver confidence and reliability. We have been actively participating in MLCommons since 2020. In 2024, we spun off our server, storage, and networking business—including all development, manufacturing, sales, and maintenance operations—into a new company, Fsas Technologies Inc. Since then, we have continued to actively participate in and submit results for both inference and training rounds across data center and edge divisions.

We offer the PRIMERGY CDI series, which leverages Composable Disaggregated Infrastructure (CDI). PRIMERGY CDI stands apart from traditional server products by utilizing a modular architecture comprising computing servers, PCIe fabric switches, and PCIe boxes. Device resources such as GPUs, SSDs, and NICs are housed externally in PCIe boxes rather than within the computing server’s chassis.

In this round, we measured Llama2-70b-99.9 on a PRIMERGY CDI system. By utilizing a single new-generation PCIe Box (10x PCIe slots, 600W, PCIe Gen5) equipped with eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, we successfully outperformed our results from the previous round (5.0-0023; NVIDIA H100-NVL-94GB x8) by approximately 50%. This significant improvement is a direct result of PRIMERGY CDI’s support for the latest generation of GPUs and its enhanced capacity for scaling GPU resources.

The mission of Fujitsu and Fsas Technologies Inc. is to make the world more sustainable by building trust in society through innovation. Leveraging our rich heritage of technological advancement and expertise, we are dedicated to contributing to the prosperity of society and our valued customers. We remain committed to meeting our customers' evolving needs and striving to deliver compelling server systems through our continued contributions to MLCommons.

ML Commons

GigaComputing

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

For MLPerf Inference 6.1, GIGABYTE systems were selected to reflect the broad choices of x86 platforms in the market today. The two 8U systems were optimized for airflow and performance using NVIDIA HGX B300. This round of testing paired the B300 GPU with the 120-core Intel Xeon 6979P (LGA 7529), whereas GIGABYTE's MLPerf Inference 6.0 testing used the 64-core 6767P (LGA 4710). Results became more interesting. Take a look for yourself.

Platforms submitted:

- GIGABYTE G894-ZD3-AAX7, using:

    2x AMD EPYC 9575F (64 core) processors

- GIGABYTE G894-AD3-AAX7, using:

    2x Intel Xeon 6979P (120 core) processors

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

ML Commons

Google

Google intentionally focused the MLPerf 6.1 inference submission on Deepseek-R1 as the industry transitions from standard LLMs to massive Mixture-of-Experts (MoE) architectures. This transition has shifted from raw compute density to communication latency and memory bandwidth.

To meet the changing demands on underlying infrastructure Google provides AI Hypercomputer. AI Hypercomputer is an architecture that combines purpose-built hardware, open software, and flexible consumption models. Each component is carefully co-engineered to work well together, to improve performance, efficiency, and developer productivity.

To further help customers overcome the infrastructure bottlenecks when provisioning and managing at scale we enable users to leverage Slurm + Cluster director NVIDIA Dynamo provides a runtime capable of managing distributed state without introducing synchronization overhead. To scale MoE architectures with A4X and NVIDIA Dynamo check out the published recipes.

ML Commons

HPE

HPE's results in MLPerf Inference v6.1 demonstrate strong performance for AI inference workloads and match the wide variety of inference customer use-cases. HPE supports any size and type of AI workload through a variety of rack-scale systems, high-density AI servers, and enterprise compute servers.

The NVIDIA GB200 NVL72 by HPE is a rack-scale system optimized for large model training, fine-tuning, and real-time inferencing for AI models over 1-trillion parameters. For this round of MLPerf Inference, this system demonstrated scale-out performance and suitability for sparse mixture-of-experts (MoE) reasoning models, like DeepSeek-R1 and GPT-OSS. It delivered high-performance at scale with 126,000 tokens-per-second and a significant 7,800 tokens-per-second per GPU using sixteen GPUs on DeepSeek-R1 (open division). It also delivered 101,000 tokens-per-second and 12,600 tokens-per-second per GPU using eight GPUs on GPT-OSS 20B.

The HPE Compute XD690 and HPE ProLiant Compute XD685 are designed for large AI model training, fine-tuning, and inferencing. Demonstrating high bandwidth per GPU, two HPE Compute XD690 servers with NVIDIA Blackwell Ultra GPUs (NVIDIA HGX™ B300) achieved 136,000 tokens-per-second, and significantly, achieved an 8,500 tokens-per-second per GPU on DeepSeek-R1. A single HPE Compute XD690 achieved 114,000 tokens-per-second and 14,300 tokens-per-second per GPU on GPT-OSS. The HPE ProLiant Compute XD685 with eight AMD Instinct™ MI355X GPUs demonstrated high performance across Llama2 and Llama3.1 LLMs and the sparse MoE reasoning model GPT-OSS.

HPE's enterprise server, HPE ProLiant Compute DL380a Gen12, achieved high performance in Llama2 and Llama3.1 LLMs, GPT-OSS, and Whisper using NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. With NVIDIA RTX PRO 4500 Blackwell Server Edition GPUs, the HPE ProLiant Compute DL345 Gen12, HPE ProLiant Compute DL380 Gen12, and HPE ProLiant DL145 Gen11 achieved high performance across You Only Look Once (YOLO) computer-vision, Whisper speech-to-text, and Qwen3.6 multi-turn Agentic edge workloads that demand low latency and low power.

ML Commons

Intel

Intel is pleased to share MLPerf Inference v6.1 results across our portfolio of AI solutions, led by Intel® Xeon® 6 processors and Intel® Arc™ Pro B-Series GPUs, demonstrating accessible AI inference across datacenter and workstation deployments.

Intel® Xeon® 6 delivered its largest generational gains this round from software alone. On the same silicon and same socket count as v6.0, Llama3.1-8B Server throughput on Intel® Xeon® 6980P rose 2.4× (+142%) and Offline throughput rose 56%. Partners measured the same improvements independently on their own Xeon 6 platforms. Xeon coverage also broadened, growing from two benchmarked Xeon 6 SKUs in v6.0 to five in v6.1, and from 24 to 35 CPU inference results.

A single node configured with 4X Intel® Arc™ Pro B70 GPUs provides 128GB of VRAM, enabling Intel to submit Llama3.1-8B, Llama2-70B, gpt-oss-120B, Whisper and E2E-RAG.On this configuration gpt-oss-120B improved 36% in Server and 27% in Offline over v6.0 on the same system, reflecting continued software stack maturity. Intel also introduced results on the new v6.1 E2E-RAG benchmarks, on a Xeon 6787P plus 4× Arc Pro B70 configuration. In this submission, CPU and GPU are both on the critical path in a single measured run.

Customer and partner participation on Intel platforms grew to 41 results. Oracle entered with its first Intel-based submission, Red Hat with its first Xeon CPU inference submission, and Intel® Arc™ Pro B70 drew its first partner submissions from Quanta Cloud Technology and Supermicro. Intel® Xeon® is the head node for 47 of 111 systems submitted in v6.1.

ML Commons

Inventec Corporation

Inventec Demonstrates Engineering Strength with NVIDIA HGX B300 Servers in MLPerf Inference v6.1

Inventec Corporation demonstrated its hardware engineering strength and strong NVIDIA partnership in the **MLPerf Inference v6.1** benchmarks. The company showcased two flagship NVIDIA HGX™ B300 8-GPU servers powered by dual Intel® Xeon® 6 processors:

- **P5800G7 (5U, Liquid-Cooled):** Engineered with direct liquid cooling for high-density, energy-efficient AI data centers (P5800G7 Specs).
- **P9000G7 (10U, Air-Cooled):** Designed with precision air cooling and modular CPU/GPU thermal zoning for flexible deployment (P9000G7 Specs).

These benchmark-proven platforms deliver top-tier AI performance, giving enterprise customers versatile liquid and air-cooling choices.

About Inventec

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

ML Commons

KRAI

Founded in 2020 in Cambridge, UK ("The Silicon Fen"), KRAI is a purveyor of premium benchmarking and optimization solutions for AI Systems.

KRAI is proud to be a Founding Member of MLCommons. This MLPerf round is KRAI's 23rd round since 2021 (12x Inference, 8x Training, 2x Endpoints, 1x Tiny), a record matched only by NVIDIA.

In this round, we submitted the first result from our KRAI/Gentic, our automated agentic full-stack optimization framework. We applied KRAI/Gentic to the Qwen3-VL benchmark to automatically search for faster kernels and integrate them with vLLM. We conducted the search and validated the integrated solutions on HPE Cray XD670 servers with 8x NVIDIA H200 GPUs.

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

ML Commons

Lambda

NVIDIA Blackwell Architecture and MoE kernel optimizations enable lightning-fast inference

Lambda partnered with NVIDIA for this round of MLPerf® Inference, running benchmarks on Lambda's cloud infrastructure with two NVIDIA Blackwell systems: NVIDIA B200 GPUs, each with 180GB SXM memory, 112 CPU cores, 2.9 TB RAM, and 28 TB SSD, as well as NVIDIA GB300s with 279GB HBM3e memory and 36 Grace CPUs.

Our benchmarks with Qwen3-VL-235B-A22B-Instruct and GPT-OSS-120B on the NVIDIA B200 and GB300 chips deliver extremely fast throughput, with GPT-OSS-120b particularly providing up to an 8.85% increase in throughput over the same workload during MLPerf Inference v6.0.

Additionally, this round of inference marks the second Open Division submission for Lambda. We deployed the Agentic Inference workload on a single NVIDIA B200 node, marking the first Agentic workload for MLCommons recorded on data center hardware. We replaced the Qwen3-27B model with Kimi 2.6 from Moonshot AI, with a staggering one trillion total model parameters.

The throughput achieved on NVIDIA GB300 and NVIDIA B200 systems validates that our infrastructure delivers excellent performance for demanding inference workloads. Furthermore, our open submission of Kimi 2.6 shows that Lambda’s team of engineers is ready to deploy large-scale inference workloads. With Lambda, AI teams get proven performance, the fastest route to compute, and the professional support required to deploy inference at scale.

About Lambda

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

Founded in 2012 by machine learning engineers who published at NeurIPS and ICCV, Lambda builds supercomputers for AI training and inference.

ML Commons

MangoBoost

In the MLPerf Inference 6.1 submission round, MangoBoost is proud to announce groundbreaking results that push the frontier of LLM Inference Serving. Powered by our LLMBoost software, we achieved AMD GPUs' first-ever Prefill-Decode disaggregated inference submission, alongside the largest multi-region cluster submission to date. Focusing on the massive GPT-OSS 120B model, LLMBoost demonstrated unmatched versatility in both latency-sensitive interactive scenarios and high-throughput environments.

Pioneering Next-Generation Architecture LLMBoost successfully deployed PD disaggregation in the Interactive scenario, optimizing for latency where it matters most. By isolating prefill and decode phases, we achieved exceptional, comparable performance across both inter-node and intra-node topologies. Furthermore, LLMBoost orchestrated an unprecedented 32-GPU, multi-region cluster across four distinct global environments, including Azure, TensorWave, MangoBoost, and Dell servers. Combining MI300X and MI355X GPUs into a single serving pool, we achieved a staggering 285,454 tokens/second in Offline mode with 97% scaling efficiency, and 253,501 tokens/second in Server mode, flawlessly bridging intercontinental network gaps.

Maximizing TCO with Effortless Integration Partnering with Dell, our joint submission highlights the robust capabilities of both air-cooled and liquid-cooled MI355X servers in the Closed division. Additionally, MangoBoost unlocked GPT-OSS 120B on single-node MI300X architectures by independently preparing the quantized model and recipe. This provides a transformative Total Cost of Ownership (TCO) advantage, proving customers can seamlessly integrate and scale AMD-powered servers into existing pipelines using LLMBoost without sacrificing efficiency.

Turn-Key AI Infrastructure LLMBoost provides a turn-key solution bridging the gap from idea to production, allowing developers to serve open models at any scale in minutes. Beyond LLMBoost, MangoBoost offers advanced DPU- and intelligence-driven AI Solutions:

- **Mango Kesar:** High-density, disaggregated storage platform for high-bandwidth, low-latency NVMe access
- **Mango Alphonso:** Full-stack system purpose-built to maximize AI performance and tokens/$
- **Mango BoostX™:** DPU for offloading networking and storage for AI workloads

ML Commons

MiTAC

It is with great honor that **MiTAC**, an established server platform designer, manufacturer, and a subsidiary of **MiTAC Holdings Corporation (TWSE:3706)**, announces its performance results from its internal validation utilizing the MLPerf® Inference v6.1 benchmark suite. Our **G8825Z5**, **G4826Z5,** and **G4520G6 AI/HPC server** series have demonstrated their capability, successfully running a comprehensive suite of demanding generative AI (GenAI) and Large Language Model (LLM) architectures. The readiness of our AI portfolio is built upon purpose-built hardware and system designs, showing validation across open-source benchmarks:

• **MiTAC G8825Z5 (8U 8 AMD Instinct™ MI350X GPUs HPC/AI Server):** Engineered for massive scale-out AI Inference workloads, this platform integrates high-performance **AMD EPYC™ 9575F and AMD EPYC™ 9755** processors with 8x AMD Instinct™ MI350X accelerators (featuring 4th Gen CDNA™ architecture and 288GB HBM3e per GPU). In our MLPerf® Inference v6.1 internal runs, the platform demonstrated versatility and consistent latency across reasoning and text generation tasks by successfully completing: **Llama3.1-8b Llama2-70b-99 Llama2-70b-99.9 Gpt-oss-120b**

• MiTAC G4826Z5 (4U 8x AMD Instinct™ MI355X Liquid-Cooled AI Training& Inference Server): Optimized for space-constrained data centers requiring dense compute without power throttling, this 4U chassis integrates advanced liquid cooling technology to ensure optimal thermal efficiency. It couples **AMD EPYC™ 9575F and AMD EPYC™ 9755** processors with 8x AMD Instinct™ MI355X next-generation accelerators, establishing standard validation by handling an expanded matrix of four GenAI Inference workloads: **Llama3.1-8b Llama2-70b-99 Llama2-70b-99.9 Gpt-oss-120b**

• MiTAC G4520G6 (4U Powerhouse for Generative AI& LLM): Built for enterprise-grade AI deployment and reasoning tasks, this system is equipped with **NVIDIA RTX PRO 6000 Blackwell Server Edition**. It demonstrated structural efficiency and compute continuity, driving Inference workloads under monitoring by successfully completing: **Llama2-70b-99 Llama2-70b-99.9**

ML Commons

Nebius

For MLPerf® Inference v6.1, Nebius submitted a Datacenter Closed division preview result for the Nebius VR200 NVL72, which is built on NVIDIA Vera Rubin NVL72. Nebius was one of two submitters with next-generation silicon. Available category results also covered the Nebius GB300 NVL72 rack-scale platform, the Nebius B300 and Nebius B200 systems, and a server with 8 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. Together, these entries include a broad mix of system types and scales.

The Nebius VR200 NVL72 preview ran on 36 GPUs across nine nodes. Nebius GB300 NVL72 submissions spanned an 18-node full 72 GPU rack, a two-node 8 GPU configuration, and a single-node 4 GPU configuration. The Nebius B300, Nebius B200, and Nebius RTX PRO 6000 systems each ran as single 8 GPU nodes on the Nebius virtualized cloud. Nebius benchmarked widely used open-source models, DeepSeek R1, gpt-oss 120B, and Qwen3-VL 235B, in the Server and Offline scenarios.

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

The submitted results reflect this ongoing work. Nebius recorded results in both scenarios at every scale, from single node to full rack. These outcomes confirm that Nebius' cloud infrastructure supports demanding inference workloads on current and next-generation NVIDIA platforms.

The MLPerf® Inference benchmarks from MLCommons provide a standardized framework for evaluating AI inference performance across systems and platforms. Nebius' participation in MLPerf contributes to this effort and reflects the capability of its cloud infrastructure to support modern AI workloads with scalable and dependable performance.

ML Commons

NVIDIA

In MLPerf Inference v6.1, NVIDIA debuted the Vera Rubin NVL72 platform. Vera Rubin NVL72 delivered up to 2.5x higher token throughput compared to GB300 NVL72 on DeepSeek-R1 across offline, server, and interactive scenarios using TensorRT LLM. On the Qwen3 vision language model, Vera Rubin NVL72 achieved up to 3.7x higher throughput compared to GB300 NVL72, using vLLM with NVIDIA Dynamo.

Scaling efficiency — how effectively additional GPUs translate to throughput gains — is a key measure of AI infrastructure productivity. NVIDIA's DeepSeek-R1 submission scaled to 288 GPUs across four GB300 NVL72 racks with near-linear scaling efficiency of 99% relative to the 72-GPU single-rack baseline for the offline scenario.

Compared to last round, NVIDIA demonstrated performance gains across models through software optimization and rack-scale deployments. For example, Qwen3-VL performance improved by up to 1.6x on GB300 systems through use of lower KV cache precision, more fusion, better kernels, and disaggregated serving with vLLM and Dynamo. For WAN 2.2 text-to-video, NVIDIA scaled GB300 NVL72 to a full rack domain, reaching a peak throughput of 0.65 720p videos/second and a latency of 5.7 seconds/video — 9x higher throughput and 7.5x lower latency than single-node.

NVIDIA delivered excellent results on the Edge-Agentic benchmark, introduced for the first time in v6.1, running the Qwen3.6-27B model on Jetson AGX Thor using TRT Edge-LLM.

The NVIDIA partner ecosystem participated broadly, with 19 partners — 8 of them on multi-node Blackwell NVL72 systems — including ASUSTeK, Azure, Cisco, CoreWeave, Crusoe, Dell, Fujitsu, GigaComputing, HPE, Inventec, Lambda, MiTAC, Nebius, Oracle Cloud, Quanta Cloud Technology, Red Hat, ScitiX, Supermicro, and Wiwynn, demonstrating excellent performance.

ML Commons

Oracle

Oracle delivered a comprehensive submission to the MLPerf® Inference 6.1 suite, reporting results across a broad set of benchmark workloads and model families. The submission demonstrates a practical, scalable OCI AI inference environment designed for the workloads organizations are deploying today: language processing, advanced reasoning, speech-to-text, text-to-video generation, and multimodal product understanding.

To demonstrate flexibility across deployment scales and accelerator architectures, Oracle's GB300 NVL72 submissions spanned a full 18-node, 72-GPU rack; a two-node, 8-GPU configuration; and a single-node, 4-GPU configuration. Oracle also submitted results on an 8-GPU HGX B300 node, as well as single-node systems equipped with eight AMD Instinct MI355X GPUs and Intel X12. Together, these results reflect OCI's ability to offer customers a range of infrastructure choices for diverse inference performance, scale, and deployment requirements.

OCI provides a broad GPU portfolio optimized for different workload profiles. Customers can run demanding AI training and inference workloads on NVIDIA A10, A100 80GB, H100, H200, B300, GB200, and GB300 GPUs, scaling from individual nodes to clusters with tens of thousands of GPUs. OCI also offers AMD Instinct MI300X and MI355X accelerators, expanding customer choice for advanced training and inference using both open-weight and proprietary models. Intel X12 further broadens the range of available AI infrastructure options.

Generative AI workloads have performance characteristics and infrastructure requirements that differ materially from traditional cloud applications. Oracle has therefore engineered a purpose-built GenAI infrastructure stack on OCI, including high-bandwidth, low-latency cluster networking with RDMA support for efficient distributed workloads, together with high-performance Managed Lustre storage to help keep large models and datasets moving efficiently.

Oracle's participation in MLPerf® Inference 6.1 reflects its commitment to advancing practical AI infrastructure and supporting a broad range of AI innovation. Through scalable cloud infrastructure, accelerator choice, and purpose-built capabilities for distributed AI inference, OCI helps organizations accelerate model development, reduce operational complexity, and bring AI-powered applications into production more efficiently.

ML Commons

Orrick Industries

Orrick Industries is a research and development company advancing computational performance and efficiency across artificial intelligence and other demanding workloads.

In our first MLPerf Inference submission, Orrick Industries submitted two Open division results, both measured in the Offline scenario:

- Llama-3.1-8B — 274,927 output tokens per second
- Llama-2-70B (99.9) — 130,928 output tokens per second

Both results were produced on a single node of six NVIDIA B200 accelerators, with our patent-pending inference technology applied to stock published model weights in NVFP4 precision with an FP8 KV cache.

Inference throughput of this kind translates directly into practical benefits for the organizations deploying these models. Higher throughput means lower response latency for end users, and more concurrent users served from the same infrastructure. It means the same production workload can be met with fewer servers — reducing capital expenditure, datacenter footprint, power draw, and cooling requirements. As inference rather than training comes to dominate the lifetime cost and energy consumption of deployed AI systems, efficiency gains at this layer compound across every query served. We believe this is where much of the remaining headroom in AI deployment economics lies, and where progress most directly widens access to capable models.

Language model inference is one application of our research. We are developing technology across a range of computationally hard problems, with further work to be announced soon.

To learn more about Orrick Industries, visit https://orrickindustries.com and follow us on LinkedIn at https://www.linkedin.com/company/orrick-industries.

For more information, contact info@orrickindustries.com

ML Commons

Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global provider of data center and AI infrastructure solutions, announced its participation in the latest MLPerf® Inference v6.1 benchmark submissions in the Data Center Closed division. The submissions demonstrate QCT's continued commitment to delivering validated, scalable, and flexible infrastructure for the rapidly evolving demands of modern AI inference.

QCT submitted results based on two AI platforms addressing distinct deployment, performance, and scaling requirements:

- **QuantaGrid D75H-10U** – A high-density accelerated computing platform powered by dual Intel® Xeon® 6 processors and NVIDIA HGX™ B300 GPUs with an SXM-based architecture. Equipped with eight 800Gb/s OSFP networking ports, the D75H-10U provides high-bandwidth, low-latency connectivity designed to support scale-out AI infrastructure for demanding training and inference workloads.
- **QuantaGrid D75E-4U** – A flexible 4U platform powered by dual Intel® Xeon® 6 processors and supporting up to eight PCIe GPU accelerators. Its versatile architecture enables customers to tailor AI infrastructure to diverse workload and deployment requirements, from CPU-based inference accelerated by Intel® AMX to cost-efficient GPU inference with PCIe accelerators such as Intel® Arc™ Pro B70 GPUs.

Through continued participation in MLPerf Training and Inference benchmark submissions, QCT reinforces its commitment to open industry standards, independently validated performance, and scalable AI infrastructure. These results demonstrate QCT's ability to provide a broad portfolio of AI platforms addressing the performance, scalability, flexibility, and cost-efficiency requirements of enterprises, cloud service providers, and research organizations.

ML Commons

Red Hat

Red Hat, the open hybrid cloud technology leader delivering a trusted, consistent and comprehensive foundation for transformative IT innovation and AI applications, is proud to showcase MLPerf® Inference v6.1 results. These results demonstrate the breadth, flexibility, and performance of the Red Hat AI platform across models, hardware architectures, and deployment environments.

Our submissions span a diverse set of open models, including gpt-oss-120b, Qwen3-VL-235B-A22B, Llama-3.1-8B, and Whisper-Large-v3, and represent generative AI, multimodal vision-language, and speech workloads. They were evaluated on two very different classes of hardware: NVIDIA GB200 NVL4 Grace-Blackwell systems and, in a joint submission with Intel, CPU-only Intel Xeon 6 servers without accelerators. Red Hat AI's inference stack is grounded in vLLM, the leading open source inference engine, which delivers high performance through efficient attention kernels, continuous batching, asynchronous execution, and quantization optimizations. This round, vLLM produced valid results across both GPU and CPU architectures without changing the software layer.

Highlights of this submission include the only Kubernetes-based result in the closed datacenter division: gpt-oss-120b served by vLLM on Red Hat OpenShift on GB200 NVL4, demonstrating that cloud-native orchestration and peak inference performance are no longer a trade-off. Our Qwen3-VL-235B-A22B results, in partnership with NVIDIA and Supermicro, combined vLLM with NVIDIA Dynamo to serve a demanding multimodal vision-language workload on the latest Grace-Blackwell hardware. And our CPU-only Llama-3.1-8B and Whisper-Large-v3 results with Intel show that enterprises can serve small language and speech models, within MLPerf's strict latency constraints, on the general-purpose servers they already own.

Together, these results demonstrate how Red Hat empowers enterprises to deploy any model on any hardware, from Grace-Blackwell superchips to x86 CPUs, on-premise or in the cloud, while scaling seamlessly on Kubernetes - all powered by fully open source technology.

Learn more about Red Hat AI and try out its integrated and optimized inference stack with Red Hat AI's free 60-day trial.

ML Commons

ScitiX

ScitiX is pleased to announce our submission of MLPerf® Inference v6.1 benchmark results, based on the NVIDIA HGX B200 (8xB200-SXM-180GB) system. These results showcase the exceptional performance and scalability of ScitiX's optimized infrastructure.

For the DeepSeek-R1 benchmark, the measured throughput reached 60,411 tokens/s in the Offline scenario and 59,668 tokens/s in the Server scenario, strictly under the accuracy and latency constraints defined by MLPerf.

Serving a large-scale Mixture-of-Experts model within one node, our engineering work covered the following areas:

- **Precision:** Utilized NVFP4 and FP8 for weights, activations, and KV cache, retaining BF16 for quantization-sensitive layers, ensuring strict accuracy compliance.
- **Batching & Scheduling:** Implemented in-flight batching with chunked prefill. The scheduler was fine-tuned for time-to-first-token (TTFT) and time-per-output-token (TPOT) boundaries in Server, and optimized with specific warmup steps for steady-state Offline throughput.
- **Parallelism:** Combined tensor and expert parallelism across 8 GPUs, optimizing expert placement to keep all-to-all traffic on intra-node NVLink and overlapping communication with computation.
- **Kernel & Graph:** Applied CUDA graph capture on the decode path, fused GEMM, normalization, and activation sequences, and autotuned kernels for this workload's tensor shapes.
- **Host & Platform:** Configured NUMA-aware memory placement and CPU core pinning to ensure host-side handling does not bottleneck GPU utilization.

These optimizations are delivered in the standard ScitiX GPU Cloud Platform, allowing users to reproduce this configuration seamlessly. Full system specifications and run parameters accompany this submission.

ML Commons

Supermicro

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

Supermicro provides the latest computing and AI systems, reducing the time-to-online for customers worldwide. These include CPU and GPU systems from Intel, AMD, NVIDIA, and Arm. In this Inference v6.1 benchmark, Supermicro provided results for the NVIDIA B300, AMD Instinct MI355X, Intel Arc Pro B70, and Intel Xeon 6/6+ CPUs. Some of these systems were supported by liquid cooling, while others were supported by air cooling.

ML Commons

Telecommunications Technology Association

The Telecommunications Technology Association (TTA) is a non-profit organization established in 1988, providing testing, certification, and standardization services for Information and Communication Technology. TTA supports trustworthy AI deployment by validating performance and developing standards for the safety, reliability, and interoperability of AI technologies.

To support the Korean AI infrastructure ecosystem and the global visibility of Korean platform vendors, TTA values publishing performance results through MLCommons. On this occasion, TTA characterized large language model serving on the DeepGadget dg5W.

The Deep gadget dg5W-PRO6000-4 is a single-node, direct-liquid-cooled (DLC) server built around four PCIe Gen5-attached Blackwell-generation accelerators (96 GB each), paired with a 16-core server-class CPU and 512 GB of DDR5 memory. Its fully enclosed, self-contained liquid-cooling design keeps GPU temperatures stable without dedicated data-center HVAC, letting it run comfortably in a standard office-level environment while supporting sustained high-density compute in a compact chassis. Built-in dual-pump, dual-reservoir cooling and real-time monitoring of temperature, flow, and coolant level ensure consistent performance and long-term reliability, making it a great fit for single-node, PCIe-based AI inference deployments.

In this evaluation, the accelerators communicate over the host PCIe Gen5 fabric only, with no external or multi-node network fabric. The submission targets Llama 3.1 8B in the Closed Division under the Offline and Server scenarios. The precision recipe quantizes weights and activations to NVFP4 with an FP8 KV cache, calibrated solely on the MLCommons-provided calibration set.

Both scenarios produced valid results within the benchmark's accuracy and latency constraints. We are pleased to share these results as a reference point for single-node LLM serving on PCIe-attached accelerators. TTA remains committed to open benchmarking and initiatives such as MLPerf.

ML Commons

VibeHPC

VibeHPC is a high-performance AI systems organization focused on making advanced AI faster, more efficient, and more accessible. We build and optimize AI systems across the full technology stack, from customized models and high-performance GPU kernels to end-to-end inference systems. Our goal is to help organizations deploy demanding AI workloads with higher throughput, lower latency, and greater infrastructure efficiency.

For MLPerf Inference v6.1, VibeHPC contributed results to the Datacenter Closed division on NVIDIA B300 GPU platforms. Our submission spans Llama 3.1-8B, Llama 2-70B (99.9%), GPT-OSS-120B, DeepSeek-R1, Whisper, and WAN 2.2, covering language generation, large-scale reasoning, speech recognition, and text-to-video generation.

Our eight-GPU B300 system delivered 53,267.2 samples/second on Whisper Offline and 0.0783824 samples/second on WAN 2.2 Offline. For Llama 3.1-8B, the same system achieved 156,867 tokens/second in Server and 161,767 tokens/second in Offline. VibeHPC also demonstrated strong single-GPU inference, achieving 20,420.4 tokens/second in Server and 21,266.1 tokens/second in Offline for Llama 3.1-8B. On a single B300, the system further delivered 14,495.6 and 14,636.8 tokens/second for Llama 2-70B (99.9%), and 13,854.0 and 14,130.6 tokens/second for GPT-OSS-120B in Server and Offline, respectively.

These results reflect VibeHPC's end-to-end approach to AI inference optimization. Efficient inference requires co-optimization across numerical precision, GPU kernels, memory utilization, batching and scheduling, inference software, and accelerator architecture. Our work focuses on improving GPU utilization, increasing throughput, reducing latency, and lowering the infrastructure cost and complexity of deploying advanced AI models.

Beyond benchmarking, VibeHPC is building practical technologies spanning efficient AI models, high-performance GPU kernels, and optimized inference systems. Through MLPerf, we aim to contribute reproducible performance data to the broader AI community while translating innovations in high-performance computing and AI systems into efficient, scalable, and deployable AI infrastructure.

ML Commons

Wiwynn

Wiwynn is an innovative 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 workloads, including advanced AI inference.

In MLPerf Inference v6.1, Wiwynn reported GPT-OSS-120B benchmark results on a single Wiwynn OneW GB300 NVL72 node with four NVIDIA GB300-288GB GPUs, leveraging NVIDIA NVLink and NVLink-C2C connectivity. With Wiwynn's system integration expertise and TensorRT-LLM's inflight-batching harness, the system delivers fast, accurate large language model (LLM) inference for real-world deployment. This submission reflects Wiwynn's strengths in system design, hardware/software co-optimization, and firmware fine-tuning, achieving high throughput, low latency, and preserved output quality through FP4 weight precision.

This node reached 63,991 tokens/second in the Offline scenario and 60,466 tokens/second in the Server scenario, with Server latency meeting MLPerf requirements for both time-to-first-token (TTFT) and time-per-output-token (TPOT). In particular, p99 TTFT was 336 ms against a 3,000 ms limit, with accuracy above 83% across both scenarios. These results underscore the benefits of Wiwynn's AI-optimized infrastructure for customers deploying LLM inference applications.

Guided by our vision to unleash the power of digitalization and ignite the innovation of sustainability, Wiwynn will continue to develop IT solutions with optimized TCO, workload performance, and energy efficiency. By supporting the MLCommons community and contributing to advancing open, measurable, and reproducible AI performance, Wiwynn is committed to empowering hyperscalers and cloud service providers to accelerate the deployment of reliable, production-ready AI.

For more information, please visit Wiwynn website, Facebook and Linkedin.

ML Commons

Individual Contributor: Naeem Khoshnevis

Naeem Khoshnevis is a Lead ML Research Engineer at the Kempner Institute at Harvard University and an individual contributor to MLPerf Inference v6.1. His work focuses on machine learning systems, high-performance computing, and efficient large-scale AI workloads. This submission evaluates the inference performance of Llama 3.1-8B on a single NVIDIA H200 SXM 141GB GPU using both the Offline and Server scenarios in the Closed Division, Datacenter category.

The system achieved 7,796.59 tokens/s in the Offline scenario and 7,376.60 tokens/s in the Server scenario. The benchmark evaluated the full 13,368-sample CNN/DailyMail validation set with zero failed samples, demonstrating both strong throughput and reliable inference performance on a single accelerator.

The submission uses FP8 inference, with weights and activations quantized to FP8 (W8A8) using NVIDIA ModelOpt static per-tensor post-training quantization, together with an FP8 KV cache. The inference engine was built with TensorRT-LLM 1.0.0 and served using trtllm-serve, with the MLCommons endpoints (API-centric harness) client communicating through an OpenAI-compatible HTTP interface. The benchmark was performed on a single NVIDIA H200 SXM 141GB GPU installed in a Lenovo ThinkSystem SD665-N V3 server.

This submission provides a reproducible reference for researchers and engineers interested in understanding the performance of modern large language model inference on a single high-end datacenter GPU. By documenting the quantization, calibration, engine-building, and serving configuration, the work contributes to transparent and reproducible benchmarking and helps the broader ML systems community evaluate practical approaches to high-performance LLM inference.

Naeem is pleased to contribute to the MLPerf Inference v6.1 results and to the broader MLCommons effort to promote transparent, reproducible, and community-driven evaluation of machine learning systems.