# MLCommons Supplemental Discussion

## MLPerf Inference v4.1 Results Discussion

The submitting organizations provided the following 300 word descriptions as a supplement to help the public understand their MLPerf® Inference v4.1 submissions and results. The statements **do not reflect the opinions or views of MLCommons®.**

# Supplemental Results Discussion for MLPerf Inference v4.1

## AMD

AMD is thrilled to announce the great performance of the AMD Instinct™ MI300X accelerators in our debut submission. The AMD Instinct MI300X accelerators delivered exceptional results in the LLaMA2-70B benchmark, achieving high throughput in both real-time inference and batch processing, all while maintaining the target accuracy of responses. A key factor contributing to this exceptional performance is the support for the FP8 datatype within our LLM software stack powered by AMD ROCm.

One of the standout results was powered by a high-frequency server processor from the forthcoming 5th Gen AMD EPYC™ CPU family, designed expressly for optimal GPU hosting, combined with proven AMD Instinct MI300X GPU accelerators. This platform combination effectively demonstrated the robustness and efficiency of AMD AI-optimized hardware alongside the strength of AMD AI software solutions.

The benchmark utilized one of the latest versions of AMD ROCm™ (6.1.2), which played a pivotal role in boosting AMD Instinct MI300X accelerator performance through optimized math libraries, advanced compiler and runtime efficiencies, effective memory management, and seamless AI framework integration. Additionally, our partner, Dell Technologies, has validated the AMD Instinct MI300X accelerator capabilities by submitting their platform results.

Leveraging advanced hardware design, extensive memory capacity, FP8 datatype support, and the optimized ROCm software stack, the AMD Instinct MI300X accelerators deliver outstanding performance, making it an ideal choice for a wide range of AI inference applications. AMD is committed to transparency and trust, ensuring that enterprise customers can make informed decisions when integrating AI technologies. The MLPerf® benchmarks are crucial for standardizing AI performance evaluation, and participation from AMD underscores our dedication to providing the AI community with objective, reliable data.

As we continue to innovate and push the boundaries of AI hardware and software, AMD remains committed to publishing transparent performance results and shaping the future of AI and machine learning. We look forward to setting new industry benchmarks in the years ahead.

# Supplemental Results Discussion for MLPerf Inference v4.1

# ASUSTek

ASUS, a global technology leader renowned for its cutting-edge innovations, today announced its enthusiastic involvement in MLCommons® latest round of MLPerf® v4.1 inference benchmarks. This move underscores ASUS's steadfast commitment to pushing the boundaries of machine learning (ML) performance and fostering collaboration within the ML community.

ASUS's participation in the MLPerf v4.1 inference benchmarks, a cornerstone of MLCommons' initiatives, allows the company to showcase its hardware and software solutions under rigorous real-world conditions. By measuring performance across a diverse range of ML tasks, ASUS gains invaluable insights that drive continuous improvement and innovation.

ASUS is excited to be an active contributor to MLCommons' efforts, We believe that collaboration is key to accelerating ML advancements, and MLCommons provides a vital platform for knowledge sharing and collective progress. Our participation in these benchmarks reflects our dedication to delivering top-tier ML solutions to users worldwide. ASUS's contributions to MLCommons extend beyond benchmarking. We actively participate in discussions and working groups, offering valuable perspectives on industry trends and challenges. This involvement allows ASUS to shape the future of ML performance and ensure that its solutions remain at the forefront of technological innovation.

ASUS's commitment to ML innovation is evident in its extensive portfolio of AI-powered products and solutions. From high-performance laptops and desktops designed for demanding ML workloads to cutting-edge AI software and tools, ASUS empowers users across various industries to harness the potential of ML. In MLperf v4.1 Inference, ASUS leveraged the latest AI server to do MLPerf Inference 4.1 including ESC4000A-E12 (4xH100-NVL-94GB), ESC8000A-E12 (8x H100-PCIe-80GB, TensorRT), and ESC-N8-E11 (8x H100-SXM-80GB, TensorRT), we tried to cover as much as we can in inference use scenario with different kinds hardware configuration.

By joining MLCommons' latest efforts, ASUS reaffirms its position as a leader in the ML space. The company's dedication to collaboration, innovation, and performance will continue to drive advancements in ML technology and deliver exceptional experiences to users worldwide.

Supplemental Results Discussion for MLPerf Inference v4.1

# Cisco Systems

Enterprises across all industries are recognizing the true potential of AI/ML. Artificial intelligence (AI) and machine learning (ML) are opening up new ways for enterprises to solve complex problems.

Cisco successfully submitted 56 results for MLPerf® 4.1 Inference for datacenter category in partnership with Intel and NVIDIA to enhance performance and efficiency, optimizing inference workloads such as LLM summarization (Language), Language processing (Language), Image Generation (Image), Image classification (Vision), Object detection(Vision), Medical image segmentation (Vision)  and Recommendation (Commerce) across UCS rack & blade platforms Cisco UCS C240 M7 system with Intel Xeon 5th gen processors & NVIDIA L40S GPUs, UCS X210c M7 system with Intel Xeon 5th gen processors & NVIDIA L40S GPUs and UCS C245 M8 system with NVIDIA L40S & H100 GPUs.

Cisco servers with AI capabilities can dynamically adjust resource allocation based on workload demands, improving overall performance and efficiency.

Supplemental Results Discussion for MLPerf Inference v4.1

# Connect Tech Inc

Connect Tech is a global leader in embedded computing and a proud Elite partner in the NVIDIA® Partner Network. Our edge AI solutions meet the evolving needs of modern Edge AI applications across industries such as industrial automation, public safety, robotics, transportation, and more.

In our second consecutive year, we submitted to the MLPerf® Inference v4.1 category, running the GPT-J 6 billion parameter large language model benchmark. Utilizing our Anvil Embedded System powered by the NVIDIA Jetson AGX Orin™ 64 GB module, we have delivered a powerful AI inference performance that pushes edge computing boundaries.

For edge devices like the NVIDIA Jetson AGX Orin, the MLPerf Inference: Edge benchmarks are crucial as they simulate the types of workloads these devices would encounter in real-life applications. The Anvil system has demonstrated outstanding capabilities in handling these demanding AI workloads.

Connect Tech customers have utilized the Jetson AGX Orin's performance in many applications, including:

- Autonomous forklifts in warehouse and logistics, enabling rapid heavy load lifting with object detection and avoidance software.
- Heavy vehicles for industry and mining, connecting GPS, vision, and remote management systems.
- Last-mile delivery robots, empowering remote management, sensors, and object detection and avoidance.
- Self-driving vehicles, both in fleet management and deployment, integrating complex cameras and sensors.

Connect Tech is honored to be included in the MLPerf Inference v4.1 contingent, alongside esteemed organizations such as NVIDIA, Google, Dell, Intel, and many others. The limits of innovation are endless, driven by the momentum provided by MLCommons® in benchmarking performance across diverse AI platforms.

Supplemental Results Discussion for MLPerf Inference v4.1

# cTuning Foundation

In this submission round, the cTuning foundation and cKnowledge.org have completed transferring ownership and maintenance of their Collective Mind workflow automation framework (CM) with virtualized MLOps (CM4MLOps) to MLCommons®: github.com/mlcommons/ck.

The goal of this partnership is to help everyone run and optimize MLPerf® benchmarks in a unified and automated way across diverse models, data sets, software and hardware from different vendors: docs.mlcommons.org/inference.

Our next objective is to work with students, researchers and engineers to reproduce, compare and improve MLPerf results using the common CM interface as a part of our educational and reproducibility initiatives at AI, ML and systems conferences including ACM/IEEE SuperComputing and MICRO (cTuning.org/ae) and open optimization challenges (access.cKnowledge.org/challenges).

We are also setting up an open science lab in collaboration with FlexAI to learn how to build and run AI in the most efficient and cost-effective way using the most suitable models, software and hardware based on user requirements and constraints, MLPerf and CM virtualization for AI systems.

Supplemental Results Discussion for MLPerf Inference v4.1

# Dell Technologies

Dell Technologies has once again proven its commitment to advancing AI workloads in the latest MLPerf® Inference v4.1 results. With 130 submissions across 11 different system configurations, Dell continues to make a significant contribution in both data center and edge closed submissions.

Key highlights of this round include:

- The PowerEdge XE9680 server, equipped with 8x NVIDIA H200 Tensor Core GPUs, delivered exceptional performance across various AI workloads.
- The PowerEdge XE8640 and its liquid-cooled counterpart, the XE9640, each featuring 4x NVIDIA H100 Tensor Core GPUs, excelled in a wide range of tasks, further showcasing Dell's leadership in AI acceleration.
- The PowerEdge R760xa server, configured with 4x NVIDIA L40S and 4x H100 GPUs, also stood out, demonstrating its capability to handle demanding AI workloads efficiently. This 2RU server is optimized for PCIe GPU density, with enhanced power and airflow management.
- This round also marks Dell's debut focus on 8-way OAM AI acceleration, with new options like the XE9680 featuring 8x MI300X GPUs and 1.5 TB of HBM3 memory, underscoring Dell's ongoing innovation in delivering powerful AI solutions.

Dell Technologies continues to push the boundaries of AI acceleration, offering solutions that deliver higher quality predictions, faster time-to-value, and enhanced decision-making. Discover these cutting-edge technologies at our global Customer Solution Centers or collaborate with us at our Centers of Excellence. Experience the power of our innovations firsthand and accelerate your decision-making with Dell Technologies.

# Supplemental Results Discussion for MLPerf Inference v4.1

# 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 submitted results to the Datacenter Closed division with two systems. The first system is PRIMERGY CDI, equipped with 16x NVIDIA L40S GPUs installed in two external PCIe boxes. The second is PRIMERGY GX2560 M7, equipped with 4x NVIDIA H100 Tensor Core GPUs inside the server.

PRIMERGY CDI is a departure from traditional server products. It consists of compute 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 compute server chassis. The most remarkable feature of PRIMERGY CDI is the ability to freely allocate devices within the PCIe boxes to multiple compute servers. For instance, during a day, you can allocate a small number of GPUs for inference tasks, and at night, you can increase the number of GPUs for training tasks.

In this round, the PRIMERGY CDI system, equipped with 16x NVIDIA L40S, achieved significantly better results in the retinanet and 3d-unet benchmarks compared to the GX2560 M7 equipped with 4x NVIDIA H100s. Moreover, we are pleased to announce that we successfully executed and submitted all benchmarks in the datacenter division, including the newly added mixtral-8x7b, using the GX2560 M7 equipped with 4x NVIDIA H100-SXM-80GB.

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 v4.1

# GIGABYTE (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, Intel, and Qualcomm.

In 2020, GIGABYTE joined MLCommons® and submitted its first system. In the latest **MLPerf Inference v4.1 (closed division)** benchmarks, the submitted GIGABYTE G593 Series platform demonstrated competitive performance across multiple inference applications using the latest NVIDIA Tensor Core H200 GPUs. It showed up to 1.3x and 1.4x performance improvements in stable-diffusion-xl and llama 2 70b 99.9, respectively, and overall enhancements of 4% to 8% in vision and language tasks compared to the previous benchmark tested with NVIDIA H100 GPUs.

With the inclusion of the NVIDIA H200 in the G593 product line, GIGABYTE now offers flexible GPU platforms, covering OAM designs from AMD and the HGX platform from NVIDIA, accommodating both air-cooled and liquid-cooled solutions. Additionally, the design in a 5U form factor has been thermally optimized for greater compute density and rack density, ensuring maximum performance and efficiency in high-demand environments.

- GIGABYTE G593-SD1: Dense accelerated computing in a 5U server
- 2x Intel Xeon 8480+ CPUs
- 8x NVIDIA H200 Tensor Core GPUs with NVIDIA® NVLink® and NVSwitch™
- Optimized for baseboard GPUs

To learn more about our solutions, visit: https://www.gigabyte.com/Enterprise
Giga Computing's website is still being rolled out: https://www.gigacomputing.com/

Supplemental Results Discussion for MLPerf Inference v4.1

# Google Cloud

For MLPerf® Inference v4.1, Google Cloud submitted nine results demonstrating the performance of its AI Hypercomputer, a supercomputing architecture bringing together an integrated system of performance-optimized hardware, open software and frameworks, and flexible consumption models. AI Hypercomputer employs systems-level codesign to boost efficiency and productivity for training and serving modern AI workloads as they continue to evolve.

Seven of Google Cloud's submitted results were on A3 Mega VMs, which are powered by NVIDIA 80GB H100 Tensor Core GPUs in a single VM. Now generally available, A3 Mega VMs offer double the GPU-to-GPU networking bandwidth of A3 VMs. In MLPerf Inference v4.1, **A3 Mega VMs achieved extremely competitive results** across the board for generative AI models (Stable Diffusion XL, Llama 2 70B, and Mixtral 8x7B), and demonstrated stand out performance on the DLRM v2 benchmark for ranking and recommendation models. These strong results are a testament to Google Cloud's close partnership with NVIDIA to build workload-optimized infrastructure for modern AI workloads.

This is also Google Cloud's first MLPerf submission using Trillium (launching later this year), the sixth-generation Tensor Processor Unit (TPU), and the most performant yet. Google Cloud achieved standout performance on the Stable Diffusion XL model, with **Trillium delivering a 3.1x and 2.9x throughput improvement for samples/second and queries/second, respectively**, compared to its predecessor, TPU v5e. (Four TPU v5e chips delivered 1.75 offline samples/second and 1.55 server queries/second, while four Trillium chips delivered 5.44 offline samples/second and 4.49 server queries/second). This is primarily driven by a combination of Trillium's purpose-built architecture and advancements in the software stack to harness the increased compute power (demonstrated by MaxDiffusion's implementations of core components of diffusion models such as cross attention, convolutions, and high-throughput image data loading).

# Supplemental Results Discussion for MLPerf Inference v4.1

# Hewlett Packard Enterprise

Hewlett Packard Enterprise (HPE) ran benchmarks on servers designed for AI inference.

**Highlights include:**
HPE's latest server, the HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2, delivered outstanding performance on all models submitted owing to the 144GB HBM3e memory available per Superchip and NVIDIA's collaboration on this submission. HPE and NVIDIA proved that this platform delivers high performance for generative AI (GenAI) inference. The HPE ProLiant Compute DL384 Gen12 server is built around two NVLink® connected NVIDIA GH200 Superchip processors for low-latency datacenter inference. HPE is first to submit performance results with the NVIDIA GH200 NVL2[1].

HPE Cray XD670 with eight NVIDIA H100-SXM GPUs and Intel® Xeon® processors delivered HPE's highest MLPerf inference performance to-date in computer vision, GenAI, LLMs, and language models for a single node system, hosting all datasets on HPE's high-performance storage for supercomputing. This is well-suited for GenAI inference requiring high batch sizes and parallelism. With up to 22% performance gains compared to prior v4.0 results, HPE Cray XD670 continues to demonstrate high throughput across all AI inferencing scenarios.

The HPE ProLiant DL380a Gen11 server with four NVIDIA H100-NVL 94GB GPUs and Intel Xeon processors delivered HPE's highest GenAI inference throughput to-date from a 2U server platform and was a top performer among systems using four of these GPUs for computer vision and LLMs. It was the only PCIe-based server in this configuration to submit Mixtral-8x7B performance[2]. HPE submitted a second HPE ProLiant DL380a Gen11 server configuration with four NVIDIA L40S GPUs and Intel Xeon processors. This configuration performed well in all submitted categories: computer vision, 3D medical imaging, and natural language processing.

HPE and Intel also partnered to submit our first results for HPE ProLiant DL380 Gen11 with Intel Xeon processors which included computer vision, 3D medical imaging, natural language processing, and the GPT-J 6B large language model.

[1] Based on results for NVIDIA GH200 NVL2 with 144GB HBM3e memory.

[2] Based on servers with four GPUs: ResNet-Server, RetinaNet-Server, Llama2-70B-Server (99.0 and 99.9), Llama2-70B-Offline (99.0 and 99.9), and Mixtral-8x7B (Open division)

# Supplemental Results Discussion for MLPerf Inference v4.1

# Intel

With the latest MLPerf® results, Intel continues to show Xeon's strength for AI inference and general-purpose AI workloads. For the MLPerf Inference v4.1 round, Intel submitted results for 5th Gen Xeon Scalable Processors and, for the first time, Xeon 6 processors with Performance-cores. The results demonstrate Intel Xeon's strength for AI workloads, including classical machine learning, small- and mid-size models, and vector search embedding.

In total, Intel submitted six MLPerf benchmarks for Xeon 6 processors and 5th Gen Xeon processors. Across all six benchmarks, Xeon 6 processors have improved by 1.9x in AI performance compared to 5th Gen Xeon processors.

Intel has significantly invested in AI and improved performance with its CPUs over the past four years. Since Intel began submitting Xeon to MLPerf in 2021 with 3rd Gen Xeon Scalable Processors, Intel has improved its AI performance up to 11x on ResNet50 and up to 17x on BERT. Further, Intel remains the only vendor to submit server CPU MLPerf results.

Intel will have more Xeon 6 performance results to share at launch. For more details, please see MLCommons.org.

Supplemental Results Discussion for MLPerf Inference v4.1

# Juniper Networks

Juniper is thrilled to continue to collaborate with MLCommons® to accelerate AI innovation and make data center infrastructure simpler, faster and more economical to deploy.

In the latest MLPerf® Inference submissions, Juniper Networks showcased a Llama2-70b model running on two and four nodes of NVIDIA H100 Tensor Core GPUs. Building on our success in the Inference 4.0 round, this submission marks our first closed submission to MLPerf Inference, where we highlighted the scalability of our network interconnect across both Offline and Server scenarios. Leveraging Juniper's AI-Optimized Ethernet fabric with RoCEv2 for inter-GPU communication, we achieved linear throughput scaling for both use cases without any loss of accuracy. The network topology used in this submission is now available as a Juniper Validated Design (JVD). All experiments and testing were conducted in Juniper's in-house Ops4AI Lab.

Juniper is committed to an operations-first approach to help customers manage the entire data center lifecycle with market-leading capabilities in intent-based networking, AIOps and 800Gb Ethernet. Open technologies such as Ethernet and our Apstra data center fabric automation software eliminate vendor lock-in, take advantage of the industry ecosystem to push down costs and drive innovation, and enable common network operations across AI training, inference, storage and management networks. In addition, rigorously pre-tested, validated designs are critical to ensure that customers can deploy secure data center infrastructure on their own.

# Supplemental Results Discussion for MLPerf Inference v4.1

# KRAI

Established in 2020 in "Silicon Fen" (Cambridge, UK), KRAI are the purveyors of premium benchmarking and optimization solutions tailored for engineering ultra-efficient and cost-effective Computer Systems for Artificial Intelligence. The KRAI team led by Dr Anton Lokhmotov has participated in all 10 out of 10 MLPerf® Inference rounds, a feat achieved by only three other submitters out of 70+ since 2019. Crucially, KRAI's open-source technologies - KILT for efficient and scalable inference and X for reproducible workflow automation - have been honed and battle-tested over the years through thousands of submissions from KRAI's customers such as Qualcomm, Google, Hewlett Packard Enterprise, Dell Technologies and Lenovo.

In this tenth anniversary submission round of MLPerf Inference, KRAI partnered with Untether AI to showcase speedAI 240, their innovative second-generation at-memory inference accelerator. By expertly integrating support for speedAI 240 into KILT, KRAI ensured that Untether AI benefitted from many lessons learnt in creating highly competitive and fully compliant submissions, and submitted jaw-dropping performance and energy efficiency results.

In the Datacenter Power category, a 2U rack server equipped with six speedAI 240 Slim accelerators running ResNet50 Offline delivered 334,462 samples per second, while consuming only 1,021 Watts, thus achieving 327 samples per second per Watt.

In the Edge Power category, a workstation equipped with three speedAI 240 Slim accelerators running ResNet50 Offline delivered 168,720 samples per second, while consuming only 398 Watts, thus achieving 424 samples per second per Watt.

Moreover, complemented by the low-overhead KILT software, the speedAI 240 hardware achieved ultra-low ResNet50 latencies:

- For a single sample (SingleStream): 0.12 milliseconds on one accelerator;
- For a batch of eight samples (MultiStream): 0.29 milliseconds on one accelerator, 0.21 milliseconds on two accelerators, and 0.17 milliseconds on four accelerators.

# Supplemental Results Discussion for MLPerf Inference v4.1

# Lenovo

**Leveraging MLPerf Inference v4.1, Lenovo Drives AI Innovation**

At Lenovo, we're dedicated to empowering our customers with cutting-edge AI solutions that transform industries and improve lives. To achieve this vision, we invest in rigorous research and testing using the latest MLPerf® Inference v4.1 benchmarking tools.

**Benchmarking Excellence: Collaborative Efforts Yield Industry-Leading Results**

Through our strategic partnership with MLCommons®, we're able to demonstrate our AI solutions' performance and capabilities quarterly, showcasing our commitment to innovation and customer satisfaction. Our collaborations with industry leaders like NVIDIA and AMD on critical AI applications such as image classification, medical image segmentation, speech-to-text, and natural language processing have enabled us to achieve outstanding results.

**ThinkSystem SR685A v3 with 8x NVIDIA H200 Tensor Core GPUs (141GB) and the SR675 v3 with 8x NVIDIA H100 NVL GPUs: Delivering AI-Powered Solutions**

We're proud to have participated in these challenges using our ThinkSystem SR685A v3 with 8x NVIDIA H200 (141GB) GPUs and the SR675 v3 with 8x NVIDIA H100 NVL GPUs. These powerful systems enable us to develop and deploy AI-powered solutions that drive business outcomes and improve customer experiences.

**Partnership for Growth: MLCommons Collaboration Enhances Product Development**

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

Supplemental Results Discussion for MLPerf Inference v4.1

# Neural Magic

Neural Magic is a leader in optimized enterprise inference solutions that maximize performance and increase hardware efficiency, across both CPU and GPU infrastructure. When it comes to the deployment of large language models (LLMs), Neural Magic is a key player as the top commercial contributor to vLLM, the leading open-source LLM inference server.

One of Neural Magic's most notable contributions to vLLM is its implementation of FP8 W8A8 support. This advancement allows for up to 2x reduction in latency under load with minimal accuracy degradation, enabling more efficient utilization of modern GPUs like NVIDIA H100 and AMD MI300x. The FP8 implementation has been adopted by other MLPerf submitters, showcasing its impact on the broader AI community.

Neural Magic also develops LLM Compressor, a unified library for creating compressed models for faster inference with vLLM. This tool enables various optimizations, including activation and weight quantization algorithms like SmoothQuant and GPTQ, which can lead to up to 3x faster server/throughput deployments.

For enterprises looking to deploy at scale, Neural Magic offers nm-vllm, an enterprise distribution of vLLM. This package includes stable builds with bug fixes, enterprise support with SLAs, tools for applying model optimizations, a pre-optimized model registry, and Kubernetes reference architectures for production deployments.

For this MLPerf® round, Neural Magic leveraged the open-source Collective Knowledge (CK) framework to integrate vLLM as a reference backend for Llama 2 70B. This integration allows users to produce their own performant and reproducible results across various hardware accelerators, democratizing access to high-performance LLM inference.

Neural Magic's contributions to vLLM and the broader open-source LLM ecosystem demonstrate their commitment to pushing the boundaries of AI performance and efficiency. By offering both cutting-edge open-source tools and enterprise-grade support, Neural Magic enables organizations to deploy state-of-the-art LLMs on their infrastructure efficiently and effectively.

# Supplemental Results Discussion for MLPerf Inference v4.1

# NVIDIA

NVIDIA is excited to share outstanding results in the latest round of MLPerf® Inference benchmarks. In its MLPerf debut, the NVIDIA Blackwell architecture delivered up to 4X higher Llama 2 70B throughput, per-GPU, compared to the prior-generation H100 GPU. These great results were powered by Blackwell's second-generation Transformer Engine, which combines Blackwell's FP4 Tensor Core and TensorRT-LLM software innovations, to deliver significant large language model (LLM) inference performance leaps while maintaining model accuracy.

Our submissions using eight NVIDIA H200 Tensor Core GPUs, connected via high-bandwidth NVLink and NVSwitch, delivered great performance on every workload, including on the newly added mixture-of-experts (MoE) benchmark based on Mixtral 8x7B. NVIDIA H200-based systems also delivered up to 1.5X more inference performance compared to NVIDIA H100 Tensor Core GPU-based systems. And, through ongoing software improvements, H200 performance increased by up to 27% compared to its prior round.

NVIDIA also submitted results using the feature-rich, open-source NVIDIA Triton Inference Server running on H200 GPUs, delivering virtually identical performance compared to bare-metal results run without Triton Inference server.

And, for edge AI and robotics, new TensorRT-LLM software innovations increased performance on GPT-J by up to 6.2x compared to the prior round, and Stable Diffusion XL performance by 1.3X. Jetson AGX Orin and the NVIDIA software stack make running demanding generative AI workloads at the edge possible.

NVIDIA is also pleased to see great submissions from 13 partners this round, including ASUSTek, Cisco, Connect Tech Inc, Dell, Fujitsu, GigaComputing, Google Cloud, Hewlett Packard Enterprise, Lenovo, Oracle Cloud Infrastructure, Quanta Cloud Technology, Supermicro, and Sustainable Metal Cloud.

NVIDIA would also like to thank MLCommons® for its ongoing commitment to bringing benchmarking best practices to the rapidly-evolving field of artificial intelligence and generative AI.

## Supplemental Results Discussion for MLPerf Inference v4.1

# Oracle

Oracle Cloud Infrastructure (OCI) offers AI Infrastructure, Generative AI, AI Services, ML Services, and AI in our Fusion Applications. Our AI infrastructure portfolio includes bare metal instances powered by NVIDIA H100 Tensor Core GPUs, NVIDIA L40S GPUs, NVIDIA A100 Tensor Core GPUs, and NVIDIA A10 GPUs. OCI also provides virtual machines powered by NVIDIA A10 GPUs.

The inference benchmark results for the high-end NVIDIA GH200 Grace Superchip instance demonstrate that OCI provides high performance that at least matches that of other deployments for both on-premises and cloud infrastructure. NVIDIA GH200 as evidenced by the benchmark helps customers maximize their investments on GPU Infrastructure.

Supplemental Results Discussion for MLPerf Inference v4.1

# Quanta Cloud Technology

Quanta Cloud Technology (QCT), a global data center solutions provider enabling diverse HPC and AI workloads, has been recognized in the latest MLPerf® inference results released by MLCommons®.

QCT participated in the latest round of MLPerf Inference v4.1, submitting results in the data center closed division across various system configurations.

One of the highlighted submissions was the QuantaGrid S74G-2U, equipped with the NVIDIA® Grace™ Hopper™ Superchip. This platform, featuring coherent memory between the CPU and GPU via the NVLink® C2C interconnect, excels in handling memory-intensive AI inference tasks. QCT achieved great performance across various AI tasks in the data center category.

Another configuration was the QuantaGrid D54U-3U, an acceleration server tailored for AI and HPC workloads. Powered by two 4th Gen Intel Xeon Scalable processors, this 3U system supports up to four dual-width or eight single-width accelerator cards, offering a versatile architecture optimized for a wide range of AI and HPC applications. In this round, QCT validated the system using four NVIDIA H100 Tensor Core GPUs (PCIe cards) and four NVIDIA L40s GPUs (PCIe cards), respectively.

Additionally, QCT showcased the QuantaGrid D54X-1U, featuring Intel 5th Gen Xeon Scalable Processors, in CPU-only inference scenarios. This configuration was validated for its exceptional performance in general-purpose AI workloads, leveraging Intel AMX instruction sets.

Looking ahead, QCT remains committed to providing comprehensive hardware systems, solutions, and services to both academic and industrial users. The company will continue to share its MLPerf results with the MLCommons community, contributing to the ongoing evolution of MLPerf inference and training benchmarks.

# Supplemental Results Discussion for MLPerf Inference v4.1

# Red Hat

Red Hat®, the world's leading provider of enterprise open source solutions, is proud to demonstrate impressive performance with Llama 2 70b in MLPerf® Inference v4.1, utilizing Red Hat® OpenShift® AI with vLLM runtime on a Dell R760xa server with 4x NVIDIA L40S GPUs. Our llama-2-70b results showcase a cost-effective way to run AI inference, powered by open source innovation. Red Hat OpenShift AI with vLLM runs on Red Hat OpenShift offering great performance, ease of use and a hardware agnostic solution for model serving and inference. Get fast inferencing with Red Hat OpenShift AI and the vLLM runtime option:

- Red Hat OpenShift makes creating, scaling and monitoring your AI/ML workloads easier and more secure. OpenShift Operators discover, configure and monitor your GPUs, storage devices, and network devices, providing ease of use, flexibility and security.
- Red Hat OpenShift AI is a flexible, scalable MLOps platform with tools to build, tune, deploy, and manage AI models at scale. Built using open source technologies, it provides trusted, operationally consistent capabilities for teams to experiment, serve models, and deliver innovative apps. Red Hat OpenShift AI supports the full lifecycle of AI/ML 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 library for LLM inference and serving.  vLLM is provided as a supported model serving runtime in Red Hat OpenShift AI as of version 2.10.
- We demonstrate 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 FP8 quantization support, and chunked prefill.
- Thanks to the GPU memory savings of FP8 quantization, we were also able to demonstrate performance of Llama-2-70b 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 OpenShift AI here.

# Supplemental Results Discussion for MLPerf Inference v4.1

# Supermicro

Supermicro, a global leader in providing AI infrastructure servers and storage systems to AI factories, is participating in and submitting results for the MLPerf® Inference v4.1 benchmark suite. The range of Supermicro AI servers spans from very high-density liquid-cooled servers to edge devices, incorporating GPUs from various vendors. This building block architecture allows Supermicro to deliver leading-edge, performance-optimized servers with varying form factors.

For the highest density AI training systems, the 4U liquid-cooled systems, the SYS-421GE-TNHR2-LCC and AS -4125GS-TNHR2-LCC, deliver outstanding results for several of these benchmarks. With its extremely compact design at only 4U, up to eight of these servers can be installed in a single rack, including coolant distribution manifolds and a coolant distribution unit. These systems contain eight GPUs and dual processors with either Intel or AMD CPUs.

Among NVIDIA HGX H100 8-GPU equipped systems, the Supermicro 4U liquid-cooled system has achieved the highest llama2-70b-99 performance of 23,699.70 queries per second.

Supermicro also offers many server product families that contain a range of GPUs. From 1U and 2U servers that can contain up to four GPUs, to dedicated GPU servers with up to 10 PCIe GPUs, and to fully integrated systems with eight interconnected GPUs, Supermicro brings new technologies to market faster than most vendors. A number of the GPU-optimized servers can be liquid-cooled, with some product lines optimized for air cooling. Supermicro works closely with leading GPU suppliers such as NVIDIA, AMD, and Intel to bring the latest technologies to market. Supermicro has been delivering dense AI servers for many years and continues to innovate with complete liquid-cooling solutions designed by Supermicro engineers and built in manufacturing facilities worldwide.

Supermicro continues to deliver a range of solutions to customers worldwide, with leading-edge CPUs and GPUs for maximum performance for various AI applications.

Visit https://www.supermicro.com/en/solutions/ai-deep-learning for more information.

Supplemental Results Discussion for MLPerf Inference v4.1

# Sustainable Metal Cloud (SMC)

Sustainable Metal Cloud (SMC), developed by Firmus Technologies, is at the forefront of building and scaling the most sustainable AI infrastructure.

Our GPU cloud is run on the Firmus proprietary immersion-cooling platform, which uses substantially less energy than legacy clouds.

Building on our world-class training results, we're proud to announce exceptional performance in the MLPerf® Inference v4.1 benchmark. Our Firmus Immersion modified servers with 8x NVIDIA H100 GPUs matched industry leaders across all AI models and scenarios, solidifying our position as a top global performer in AI model processing. We combine industry-leading performance with a dedication to sustainability, all accessible via our cloud solutions.

These benchmark results validate the Firmus immersion cooling platform as a scalable, long-term solution to the industry's energy challenges. Our GPU cloud, built on this technology, uses substantially less energy than traditional air-cooled systems, which means lower costs for our customers whilst offering a sustainable path forward for next-generation AI infrastructure buildout.

Dr Daniel Kearney, CTO of Firmus Technologies: “The submission of our first MLPerf Inference results together with our previous training results, affirms SMC’s ability to deliver a performant, secure and reliable GPU cloud for customers to run the most demanding accelerated compute workloads from pilot to production. Not all Clouds are built equal. Customers running on SMC, not only benefit from state of the art compute infrastructure but also from the platform’s inherent cost competitiveness due to the lower energy requirement and class-leading sustainability footprint.”

As part of MLCommons®, we aim to showcase our progressive technologies, set benchmarks for best practices, and advocate for long-term energy-saving initiatives.

Supplemental Results Discussion for MLPerf Inference v4.1

# Untether AI

Untether AI, the leader in energy-centric AI inference acceleration, is thrilled to announce our participation in MLPerf® inference v4.1, where we demonstrated the exceptional performance and energy efficiency of our AI inference accelerator cards. These results reflect our unwavering commitment to delivering industry-leading AI inference accelerators, from the cloud to the edge.

This marks our debut submission to MLPerf, showcasing the capabilities of our speedAI 240 cards across multiple system configurations in both Datacenter and Edge categories. Importantly, Untether AI submitted into the Power categories for both Datacenter and Edge, as power consumption is a critical factor in deploying AI. It has not only an ecological impact as AI usage increases, but goes directly to total cost of ownership of AI acceleration solutions. Untether AI applauds MLPerf for quantifying this import benchmark.

“MLPerf submission requires operating hardware, shipments to customers, computational accuracy, and a mature software stack that can be peer reviewed. It also requires companies to declare how many accelerators are used in their submission. These factors are what makes these benchmarks so objective,” commented Bob Beachler, VP of Product at Untether AI. “The work of MLPerf is crucial for ensuring a fair and transparent evaluation process, allowing innovative technology like Untether AI’s to be tested and compared on a global scale. Untether AI’s results in MLPerf v4.1 are a testament to our team’s dedication to pushing the boundaries of what’s possible in AI acceleration.”

Untether AI’s submissions were made possible with the invaluable support from KRAI, and we extend our thanks to the KRAI team and MLCommons for their ongoing efforts.