Supplemental Discussion

MLPerf™ Inference v2.1 Results Discussion

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

Alibaba

Alibaba Cloud is one of the top-tier cloud service providers in the world, striving to provide the best customer experience around the globe. In MLPerf Inference v2.1, we demonstrated the efficiency of hardware/software co-optimization tools on Nvidia GPUs and Alibaba Yitian 710 CPU.

In the Network Division, Alibaba made the inaugural submission based on a heterogeneous resource pool of accelerators. By leveraging Alibaba's Sinian vODLA technology which abstracts away the heterogeneity of underlying computing architectures and virtualizes physical compute devices to form a scalable resource pool, we can provide users a single abstract device with arbitrary-sized computation power. Sinian vODLA technology is flexible and optimized to work across many hybrid types of accelerators, going beyond the physical limits of system PCIe slots and pod network links. On the Alibaba EFlops system powered by Sinian vODLA technology, we achieved 107060 QPS for BERT-99 on 4 compute nodes with a total 32 A100 GPUs. It also enabled us to do inference with hybrid types of GPUs.

In the Open Division (datacenter/edge), we used the Sinian full stack optimizer (SinianML) to deeply compress ML models and automatically generate more computation-friendly sparse model architectures at runtime. SinianML also adopts quantization-aware training so that the generated model can benefit greatly from Alibaba Yitian CPU's efficient 'int8' instructions during inference. On the Alibaba Yitian CPU-only system, SinianML gained ~55X speedup against the same model without Sinian optimizations while still satisfying the same accuracy requirements.

In the Closed Division (edge), we leverage AIACC to optimize the ResNet50 within the Single-Stream pattern on Alibaba Cloud. Under equal conditions, the inference latencies were lowered to 0.398ms and 0.377ms on A100 GPU and A10 GPU, respectively.

Alibaba Sinian Platform is a heterogeneous hardware acceleration framework to seamlessly build, train, and deploy ML models without loss of performance portability. AIACC is Alibaba's ECS ApasraCompute Tool for high-performance optimizations on cloud AI workloads.

ASUSTeK

ASUSTeK is committed to joining ML Commons on AI training and inference benchmarks performance. ASUS ESC8000A-E11 and ESC4000-E10S are specifically designed to deliver intensive AI applications across different inference benchmarks and are tested and validated on the latest MLPerf Inference v2.1.

ESC8000A-E11 is powered by AMD EPYC 7003 family processors and 8 NVIDIA A100 PCIe GPUs, and delivers optimized GPU performance to offload critical workloads from CPU. Design thinking is in the corporate DNA at ASUS, and with the innovative hardware design and proprietary tuning technology for server performance and power efficiency, ESC8000A-E11 can boost more performance on AI related workloads.

ESC4000-E10S is equipped with 3rd Gen Intel Xeon Scalable family processors and 8 NVIDIA A2 GPUs, and also achieves great results on MLPerf Inference v2.1. MLPerf Inference V2.1 is a big leap in inference applications, ASUS is firmly impressed by these improvements which solves more AI issues when powering AI frameworks to another level.

Hardware configuration
ESC8000A-E11 with NVIDIA A100 PCIe GPUs and ESC4000-E10S with NVIDIA A2 PCIe GPUs are tested in all Inference V2.1 benchmarks in server and offline scenarios, benchmarks, including ResNet, RetinaNet, 3D-UNet-99, 3D-UNet-99.9, RNN-T, BERT-99, BERT-99.9, DLRM-99, DLRM -99.9. Users could take these benchmarks as reference when evaluating AI workloads based on ASUS GPU server solutions.

Azure

Azure is pleased to share results from our MLPerf Inference v2.1 submission. For this submission, we benchmarked our NC A100 v4-series, NDm A100 v4-series, and NVads A10_v5-series. They are powered by the latest NVIDIA A100 PCIe GPUs, NVIDIA A100 SXM GPUs and NVIDIA A10 GPUs respectively. These offerings are our flagship virtual machine (VM) types for AI inference and training and enable our customers to address their inferencing needs from 1/6 of a GPU to 8 GPUs.

Some of the highlights from our MLPerf inferencing v2.1 benchmark results are:
- NC A100 v4-series achieved 54.2K+ samples/s for RNN-T offline scenario
- NDm A100 v4-series achieved 26+ samples/s for 3D U-Net offline scenario
- NVads A10 v5-series achieved 24.7K+ queries/s for ResNet50 server scenario

These inference benchmark results demonstrate how Azure is committed to providing our customers with the latest GPU offerings, that are in line with on-premises performance and available on-demand in the cloud, and scales to adapt to all sizes of AI workloads and needs. Special thanks to our hardware partner NVIDIA for providing the instructions and containers that enabled us to run these benchmarks. We deployed our environment using the aforementioned offerings and Azure's Ubuntu 18.04-HPC marketplace image.
The NC A100 v4-series, NDm A100 v4-series, and NVads A10 v5-series are what we and our Azure customers turn to when large-scale AI and ML inference is required. We are excited to see what new breakthroughs our customers will make using these VMs.

Biren

Founded in 2019, Biren Technology is a provider of high performance accelerated computing products for AI and graphics.

We are pleased to share the results from our first MLPerf Inference submission. Our submission is in the Datacenter Close Division, and the system is powered by our in-house Bili 104 PCIe accelerators (based on BR104 chips) and the BIRENSUPA software development platform. Some of the result highlights include

1. BERT 99.9: achieved 22.1K+ samples/s with a 8-accelerator system in the offline scenario
2. Resnet50: achieved 424.6K+ samples/s with a 8-accelerator system in the offline scenario

The BIRENSUPA software development platform provides a complete development environment for working with Biren devices, including a novel programming model & language, compiler, acceleration libraries, deep learning frameworks, development tools, and SDKs for different areas. Bili 104, together with the BIRENSUPA, provides customers with flexible and powerful solutions to address computation challenges in the AI domain.

The above MLPerf results clearly demonstrate the performance and capabilities that our products can offer. Biren continues to integrate new technologies and optimizations into our products. We are excited to work with customers, and help them to achieve more.

Dell Technologies

Bring your brightest ideas to life and drive continuous innovation with intelligent Dell Technologies. For MLPerf Inference v.2.1, Engineering submitted results for a wide variety of server CPU-accelerator combinations to provide the data you need to make the best choices for your workloads and environments.

Working with Dell Technologies: “It really is a partnership,” says Ralph Zottola, Ph.D. and Assistant VP at UAB. “Dell has the ability to work holistically, to take a big-picture engineering approach. It’s not just about the hardware. They work to identify the right type of resources, connections and services that we will need. But most importantly, they are a partner who helps us think through problems and find ideal solutions.”

Dell Technologies works with customers and partners including AMD, Deci, Intel, NVIDIA and Qualcomm to optimize software-hardware stacks for performance and efficiency. Take a closer look at the PowerEdge R750xa performance per GPU numbers. Zoom in on the Dell PowerEdge XE8545 performance per watt in nine categories! Don’t miss the ruggedized PowerEdge XR12 for performance per watt at the edge in telco, utilities, marine and defense.

You've got the power . . . numbers. Across multiple servers and accelerators in different configurations, Dell Technologies submitted power consumption metrics across Datacenter and Edge suites in Open and Closed divisions. With this data, you can get insight into operating costs and total cost of ownership, while creating an opportunity to compute more and use less.

To get the most out of systems, optimization is a must, so we openly share tips, scripts and best practices. Come take a test drive in one of our worldwide Customer Solution Centers. Collaborate with our Innovation Lab and/or tap into one of our Centers of Excellence.

Fujitsu

Fujitsu offers a fantastic blend of systems, solutions, and expertise to guarantee maximum productivity, efficiency, and flexibility delivering confidence and reliability. We have continued to participate in and submit to every inference and training round for the data center division since 2020.
In this round, Fujitsu measured benchmark programs for the edge division with PRIMERGY RX2540 M6 for the first time, not just for the data center division with PRIMERGY GX2570 M6. We also measured power consumption with the same configuration other than GPU power limit with PRIMERGY GX2570 M6. The details of these systems are shown as follows:

1. PRIMERGY GX2570 M6 is a 4U rack-mount server with Intel (R) Xeon (R) Platinum 8352V CPUx2 and NVIDIA A100 SXM 80GB x8. This server performs far better in our PRIMERGY server portfolio across heavy-duty Deep Learning (AI), Data Science and HPC workloads.

2. PRIMERGY RX2540 M6 is a 2U rack-mount server with Intel(R) Xeon(R) Platinum 8352Y CPUx2 and NVIDIA A30 x1.

Compared with our past submissions, performance has improved over time, partly because GPU performance has evolved. Our purpose is to make the world more sustainable by building trust in society through innovation. We have a long heritage of bringing innovation and expertise, continuously working to contribute to the growth of society and our customers. Therefore, we will continue to meet the demands of our customers and strive to provide attractive server systems through the activities of MLCommons.

GIGABYTE

GIGABYTE is an industry leader in high-performance servers, and uses hardware expertise, patented innovations, and industry leadership to create, inspire, and advance. With over 30 years of motherboard manufacturing excellence and 20 years of server and enterprise products, GIGABYTE offers an extensive portfolio of enterprise products.
Over the years, GIGABYTE has submitted benchmark results for both Training and Inference. As well, the submitted servers were equipped with various brands of accelerators (NVIDIA and

Qulacomm) and processors (AMD, Ampere, and Intel) in configurations to showcase systems that target different markets (x86 and Arm).
For MLPerf Inference v2.1, in closed Data Center, all tasks were run in a GIGABYTE 4U server using the Intel platform supporting NVIDIA SXM GPUs.

Server	CPU	GPU	Testing
G492-ID0	Intel Xeon 8380	NVIDIA SXM-A100 80GB 8-GPU	closed, datacenter

GIGABYTE will continue optimization of product performance to provide products with high expansion capability, strong computational ability, and applicable to various applications at data centers. GIGABYTE solutions are ready to help customers upgrade their infrastructure.

H3C

H3C is pleased to share results from our MLPerf Inference v2.1 submission.
We benchmarked H3C UniServer R5300 G5, R5500 G5 and R4900 G5 which can be applied to AI inference and training:

- H3C UniServer R5500 G5 supports NVIDIA HGX A100 8-GPU module, and eight NVIDIA A100 GPUs can be fully interconnected with six NVSWITCH at 600GB/s. It adopts modular design and perfectly matches NVIDIA A100 GPU in heat dissipation, power supply and I/O expansibility, giving full play to the strong performance of A100.
- H3C UniServer R5300 G5 supports NVIDIA HGX A100 4-GPU module, with four A100s interconnected with NVLINK at 600 GB/s. It also supports various PCIe AI accelerators, with up to 8 dual-width or 20 single-width accelerators supported. The CPU and GPU mount ratio can be in various topology configurations, including 1:4 and 1:8.

Some of the highlights from our MLPerf Inference v2.1 benchmark results are:
1. H3C UniServer R5300 G5 achieved 13.04 samples/s for 3D-UNET offline scenario.
2. H3C UniServer R5500 G5 achieved 314368 queries/s for ResNet50 server scenario, improved by 11.5% compared with the performance H3C achieved in Inference v2.0.
3. H3C UniServer R5500 G5 with one A100-SXM-80GB achieved extremely low latency of 1.53ms for BERT SingleStream scenario.

These inference benchmark results demonstrate H3C's strong technical power. We're committed to providing our customers with the latest GPU offerings to adapt to all sizes of AI workloads and needs.

HPE

Hewlett Packard Enterprise (HPE) has returned to MLPerf Inference submissions with a strong presence. Aligning to HPE's edge-to-cloud and heterogeneity goals, our submission has been developed partnering with Nvidia and Qualcomm Technologies.

This year, HPE continued to expand its submissions which highlights the versatility of the HPE Apollo 6500 in supporting a variety of NVIDIA A100 accelerator hardware configurations while maintaining excellent performance. HPE has doubled the number of submissions in this round as compared to 1.1, adding several new closed, datacenter submissions with Apollo-based configurations as well as new submissions in the closed, edge category with Edgeline-based configurations. These Apollo-based configurations include third generation AMD EPYC™ processors with eight A100-SXM-80GB accelerators (with and without MIG) and eight A100-SXM-40GB accelerators, as well as second generation AMD EPYC™ processors with four A100-SXM-40GB accelerators.

HPE also showcased a new edge offering which delivered more than 77,500 inferences/second on the ResNet50 computer vision workload and 2,800 inferences/second on the BERT NLP workload with Qualcomm Cloud AI 100. These results were delivered on HPE Edgeline, optimized for 0-55C operating environments while consuming only about 420-450 Watts. HPE also delivered leading results for high-performance power-efficiency in this category. HPE Edgeline EL8000 systems with Qualcomm Cloud AI 100 accelerators are available now where high-performance, power efficiency, and server ruggedization is needed. HPE worked with partners Qualcomm and Krai for the MLPerf v2.1 benchmark submissions powered by Collective Knowledge v2.6.1.

HPE demonstrated in this round a strong compute portfolio for inference by going beyond our big investments in machine learning tools such as Machine Learning Development Environment (MLDE) to train accurate models faster, Swarm learning bringing training to the data sources, and a full platform in Machine learning Development System (MLDS) for scaling AI models. Information on these products can be found at https://www.hpe.com/us/en/solutions/artificial-intelligence.html.

Inspur

Inspur Electronic Information Industry Co., LTD is a leading provider of data center infrastructure, cloud computing, and AI solutions, ranking among the world's top 3 server manufacturers. Through engineering and innovation, Inspur delivers cutting-edge computing hardware design and extensive product offerings to address important technology arenas like open computing, cloud data center, AI, and deep learning.

In MLCommons Inference V2.1, Inspur made submissions on five systems, NF5488A5, NF5688M6, NF5468M6J, NF5468M6 and NE5260M5.

NF5488A5 is Inspur's flagship server with extreme design for large-scale HPC and AI computing. It contains 8 A100-500W GPUs with liquid cooling. NF5688M6 based on 3rd Gen Intel® Xeon® scalable processors increases performance by 46% from Previous Generation, and can support 8 A100 500W GPUs with air cooling. NF5468M6J supports up to 24 A100 GPUs and can be widely used in Internet AI public cloud, enterprise-level AI cloud platform, smart security, video codec, etc.
NE5260M5 is an edge server with building blocks optimized for edge AI applications, and 5G applications with capability of operating at temperatures between -5℃~50℃.

In offline scenario of datacenter closed division, the performance of Bert-99, Bert-99.9, 3D-UNET-99 and 3D-UNET-99.9, are improved by 93.8%, 96.7%, 106.3% and 106.3% on NF5468M6J compared with the best performance that Inspur achieved in Inference v2.0. In server scenario of datacenter closed division, the performance of Bert-99 and Bert-99.9 are improved by 100% and 102.4% on NF5468M6J compared with the best performance that Inspur achieved in Inference v2.0. Inspur also submitted Bert and Resnet on NF5468M6 with four BR104 accelerators, and achieved very good performance.

In offline scenario of edge closed division, the performance of 3D-Unet-99, 3D-Unet-99.9, bert-99 and RNNT are improved by 3.97%, 3.97%, 1.9% and 0.35%, respectively on NE5260M5 compared with the best performance achieved in Inference v2.0.

Intel

Intel submitted MLPerf Inference v2.1 results on the 4th Gen Intel® Xeon® Scalable processor product line (codenamed Sapphire Rapids) as a submission in preview of Intel's largest, most comprehensive AI launch ever, a platform designed with the flexibility to address the needs from data scientists who want to customize models for large scale data center deployments to the needs of applications engineers who want to quickly integrate models into applications and deploy at the edge. As the only data center CPU vendor to submit MLPerf inference results on a broad set of models, we demonstrate the practicality of running DL inference anywhere on the massive existing install base of Intel Xeon servers alongside other applications.

This preview submission covers the full suite of MLPerf data center benchmarks including image processing, natural language processing (NLP), and recommendation systems. This new platform will deliver significantly more compute and memory capacity/bandwidth than the previous generation, offering a performance boost of 3.9X - 4.7X in Offline and 3.7X - 7.8X in Server, compared to Intel's previous submissions.

New hardware innovations include Intel® Advanced Matrix Extensions (Intel® AMX) for INT8-based submissions on all workloads, spanning multiple frameworks. This all-new AI accelerator engine sits on every core and delivers 8x operations per clock compared to the previous generation.

To simplify the use of these new accelerator engines and extract the best performance we focused on software. Our customers tend to use the mainstream distributions of the most popular AI frameworks and tools, so Intel's AI experts have been working for years with the AI community to co-develop this AI Platform integrated with a broad range of open and free-to-use tools, optimized libraries and industry frameworks to deliver a seamless experience deploying AI solutions across generations, getting the best performance out of the box. All software used is available through MLPerf Inference repo, DockerHub, and optimizations are integrated into Intel-optimized PyTorch.

Krai

We often get asked about the value of submitting benchmarking results to MLPerf. Potential submitters, especially ML hardware startups, are understandably wary of committing precious engineering resources to optimizing industry benchmarks instead of actual customer workloads.

MLPerf is the Olympics of ML optimization and benchmarking. While consulting several leading ML hardware companies as their "Olympic coach", we have witnessed first-hand the value that our customers extracted from making both actual and "dry-run" MLPerf submissions. As we have discovered, the MLPerf Inference suite is sufficiently diverse that nearly every benchmark presents its own unique challenges, especially when scaling to multiple accelerators and/or hundreds of thousands queries per second. So if nothing else, an intense focus on the MLPerf benchmarks is a health check that serves the broader performance cause, as it often helps resolve similar challenges with customer workloads.

We are proud to have supported Qualcomm's MLPerf Inference submissions for the fourth time round. In this v2.1 round we have also collaborated with Qualcomm's partners Dell, HPE and Lenovo, in addition to Alibaba and Gigabyte in the previous v2.0 round. To this end, we have implemented and optimized the Computer Vision and Natural Language Processing benchmarks across a range of Datacenter and Edge platforms powered by the Qualcomm Cloud AI 100 accelerators. As every Olympic coach is proud of their athletes winning Olympic

medals, we are proud of the achieved results demonstrating industry-leading performance and energy efficiency.

Finally, we submitted over 4,200 benchmarking results on our own, with over 2,800 results accompanied by power measurements. Our submissions demonstrate accuracy/performance/power trade-offs across a range of Edge platforms, workloads and engines.

Overall, we prepared 81% of all performance results and 97% of all power results in this round, thanks to our powerful workflow automation. We also provided a preview of our next generation workflow technology called KRAI-X.

Lenovo

Lenovo delivers Smarter Technology for All, to enrich the lives of people, advance research and usher in a new era of digital transformation for organizations of all sizes. We aim to deliver a diverse portfolio of compute platforms specifically designed for AI workloads ranging from the data center to the edge and deliver AI solutions that open new possibilities for enterprises and tech companies of all sizes. Choosing the best infrastructure for your AI applications should not create a barrier to launching or scaling your AI Initiatives. Our goal through MLPerf inference v2.1 is to bring clarity to infrastructure decisions so our customers can focus on the success of their AI deployment overall.

Our submissions covered 13 different benchmarks across 6 submissions with 3 variations of platforms and acceleration choices. Our Lenovo ThinkSystem SR670 V2 produced more than three-hundred thousand ResNet50 Inferences a second in a 3U data center footprint leveraging eight NVIDIA A100 PCIe GPUs while the ThinkSystem SR670 V2 HGX solution demonstrated high performance inferencing in a hybrid cooled server design with our latest Lenovo Neptune™ liquid-cooling technology.

MLPerf Inference v2.1 was also Lenovo's debut of the Qualcomm AI Cloud 100 accelerator featured in the Lenovo ThinkEdge SE350 platform, delivering more than 23,000 ResNet50 inferences a second in ruggedized edge server with a footprint less than 3 liters.

We are very proud of the results and advancements that MLPerf has contributed to AI performance improvements, standards and resources for the data science community.

Moffett

The inventor of the dual sparsity algorithm, Moffett AI has the world's leading sparse computing techniques with more than 30 patents worldwide. The company creates a new generation of AI computing platform with hardware and software co-design to achieve order-of-magnitude acceleration of computing performance, reducing latency and low TCO. The results of S4, S10,

and S30 in MLPerf v2.1 have demonstrated the potential of sparse computing in inference performance and energy efficiency, which leads to a lower total cost of ownership (TCO).

For MLPerf Inference v2.1, Moffett has submitted results of three high-sparsity accelerators (S4, S10, and S30) running ResNet-50 and Bert-large 99.9%, respectively.

Submission for: Datacenter, Offline scenario, Open division

	S30	S4	S10
resnet50	95784 FPS	31679 FPS	65593 FPS
Bert-99.9	3837 SPS	1219 SPS	2536 SPS

Besides the performance results, energy efficiency is another significant highlight of these three sparse computing accelerators. For example, the peak power consumption of S30 is merely 250W, and 75W for S4.

The Antoum architecture through hardware and software co-design and Moffett's original sparsity algorithm are the reasons to achieve great performance with high energy efficiency.

The accelerators for AI inference applications in data centers are equipped with Moffett's 1st generation Antoum processor - the first commercial AI processor with 32x sparsity in the world. Besides the sparse processing units (SPU) for native sparse convolution and matrix computing in Antoum, the processor also integrates a Vector Processing Unit (VPU), which enables flexible programmability to keep up with the fast evolution of AI models.

Also, the on-chip Video Codec, which supports 192-way 1080p video decoding at 30 FPS, and the JPEG decoder, which supports 1080p image decoding up to 6960 FPS, provide an end-to-end capability for video and image inference workloads.

Nettrix

Nettrix Information Industry (Beijing) Co., Ltd. (Hereinafter referred to as Nettrix) is a server manufacturer integrating R&D, production, deployment, and O&M, as well as an IT system solution provider. It aims to provide customers industry-wide with various types of servers and IT infrastructure products such as common rack based on artificial intelligence, multiple nodes, edge computing and JDM life cycle customization. Nettrix has developed server products for industries including Internet, telecommunications, finance, medical care and education.

Nettrix X640 G40 is an all-round GPU server with both training and reasoning functions. It supports up to 8 training GPUs, and provides comprehensive performance support for high-density GPU computing. The product supports a variety of different GPU topologies, and optimizes GPU interconnection for different applications and models. It is an efficient and all-round computing platform. At the same time, it has been adapted to the mainstream GPUs

on the market, and is perfectly compatible with a variety of GPU types. Meets the flexible needs of customers.

Nettrix X660 G45 LP is a liquid cooled GPU server. It is a high-performance computing platform specially developed for deep learning training and is designed for huge computing clusters. As a liquid cooled server as well, X660 G45 LP is designed with cold plate liquid cooling for the main heat sources, including CPU, GPU, and memory, which can achieve component-level precise cooling so as to release the full performance of CPU and GPU and reduce energy consumption. The liquid cooling technology has played a significant role in this test, effectively reducing the GPU temperature, ensuring the optimal working frequency, and giving full play to the GPU computing power.

Nettrix has been focusing on server business for 15 years. Each liquid cooled server has experienced different sorts of strict tests from R&D, manufacturing to delivery inspection, including impact resistance test, vibration resistance test, signal integrity test, and electromagnetic compatibility test, which can ensure its stable and reliable operations. In the AI era, Nettrix is committed to creating rich AI server products for customers, continuously improving computing power and energy efficiency returns, and helping customers' intelligent transformation.

Neural Magic

Neural Magic introduces performant model sparsity to MLPerf. It brings an open-source compression framework that unifies state-of-the-art algorithms such as pruning and quantization (SparseML), and a sparsity-aware inference engine for CPUs (DeepSparse). The power of sparse execution enables neural networks to be run accurately, at orders of magnitude faster speeds, using commodity CPU hardware. This marks an important milestone in efficient machine learning powered by better algorithms and software.

When applied to BERT-Large SQuAD v1.1 question answering task, Neural Magic's benchmarks:

- Maintain >99% of its original F1 score
- Decrease model size by orders of magnitude from 1.3 GB to ~10 MB
- Improve throughput performance from ~10 samples/second to up to 1,000 samples/second when executed in the sparsity-aware DeepSparse Engine.

The benchmark was evaluated using a server with two 40-core Intel Xeon Platinum 8380 CPUs. More details on each of the models and methods can be found on GitHub under Neural Magic's BERT-Large DeepSparse MLPerf Submission.

Detailed model methods:

- BERT-Large Prune OFA - Prune Once for All: Sparse Pre-Trained Language Models
- oBERT-Large: The Optimal BERT Surgeon applied to the BERT-Large model
- oBERT-MobileBERT: The Optimal BERT Surgeon applied to the MobileBERT model

NVIDIA

In MLPerf 2.1 Inference, we are excited to make our first H100 submission, demonstrating up to 4.5X higher performance than our A100 GPU. NVIDIA H100, based on the groundbreaking NVIDIA Hopper Architecture, supercharges the NVIDIA AI platform for advanced models, providing customers with new levels of performance and capabilities for the most demanding AI and HPC workloads. Hopper features our Transformer Engine, which applies per-layer intelligence to the use of FP8 precision, delivering optimal performance for both AI training and inference workloads while preserving model accuracy.

The NVIDIA AI platform delivers great performance on a broad range of models, accelerates the end-to-end AI workflow from data prep to training to deployed inference, and is available from every major cloud and server maker. We make these resources available to the developer community via NGC, our container repository.

We are excited to see our 12 NVIDIA partners submit great inference results on A100-based systems across all tests, both for on-prem as well as cloud platforms. A100 continues to deliver excellent inference performance across the full suite range of MLPerf tests, spanning image, speech, reinforcement learning, natural language and recommender systems.

We also wish to commend the ongoing work MLCommons is doing to bring benchmarking best practices to AI and HPC, enabling peer-reviewed apples-to-apples comparisons of AI and HPC platforms to better understand and compare product performance.

OctoML

Speaker: Grigori Fursin, Vice President of MLOps and the author of the Collective Knowledge framework

OctoML is a Series-C startup developing a machine learning deployment platform to automatically optimize, benchmark and deploy models in production from any deep learning framework to any target hardware based on user requirements and constraints.

Since 2021, OctoML is supporting and extending the Collective Knowledge framework (CK) to make it easier for companies to use, customize and extend the MLPerf benchmarking infrastructure, lower the barrier of entry for new MLPerf submitters and reduce their associated costs.

After donating the CK framework to the MLCommons, OctoML continued developing this portable workflow automation technology as a community effort within the open education workgroup to modularize MLPerf and make it easier to plug in real-world tasks, models, data sets, software and hardware from the cloud to the edge.

We are very glad that more than 80% of all performance results and more than 95% of all power results were automated by the MLCommons CK v2.6.1 in this round thanks to submissions from Qualcomm, Krai, Dell, HPE and Lenovo!

Furthermore, the feedback from the users has helped our workgroup to develop the next generation of the CK technology called Collective Mind available to everyone under Apache 2.0 license at https://github.com/mlcommons/ck/tree/master/cm - we used CM in this round to test the new framework and demonstrate how to simplify and automate submissions.

We invite you to join our open education workgroup to continue developing this portable workflow automation framework as a community effort to help everyone:

- learn how to use, customize and extend the MLPerf benchmarking infrastructure;
- modularize MLPerf and make it easier to plug in real-world tasks, models, data sets, software and hardware from the cloud to the edge;
- lower the barrier of entry for new MLPerf submitters and reduce their associated costs.

Qualcomm Technologies, Inc.

Speaker: John Kehrli, Senior Director of Product Management at Qualcomm Technologies, Inc.

Qualcomm's® MLPerf v2.1 inference results continue to demonstrate our performance-to-power efficiency leadership, from edge to cloud with incremental upside in performance for NLP and computer-vision networks. We expanded Qualcomm® Cloud AI 100 benchmark submission scope with new partners, Dell, HPE and Lenovo, and new edge and datacenter platforms, including HPE Proliant e920d, Dell PowerEdge 7515 and Lenovo ThinkSystem SE350. For the first time PCIe HHHL-Standard (350 TOPs) accelerator-based submissions are made in addition to PCIe HHHL-Pro (400 TOPs) accelerators. We introduced three new edge device platforms based on Snapdragon® with Qualcomm Cloud AI 100 accelerators, including Foxconn Gloria High-end with 200 TOPs, and Thundercomm TurboX EB6 and Inventec Heimdall each with a 70 TOPs accelerator - covering more than 200 MLPerf v2.1 results based on Qualcomm Cloud AI 100 accelerators.

We continue to innovate and optimize AI solutions across submissions. The 2U server platform with 18x Qualcomm Cloud AI 100 accelerators achieved ResNet-50 offline peak performance of 428K+ inference per second (418K in v2.0). ResNet-50 power efficiency for Foxconn Gloria High-end peaked at 311 inferences/second/Watt. BERT performance across platforms has increased up to 15% and BERT power efficiency on Foxconn Gloria Entry improved by 15%.

We expanded network coverage with the RetinaNet object detector in the Open division achieving 187 inference/second for PCIe-HHHL-Pro accelerator. Among Qualcomm Cloud AI 100 accelerator based RetinaNet submissions, the highest power efficiency for datacenter submissions was measured at 2.65 inferences/second/Watt, whereas Gloria High-end achieved 4.16 inferences/second/Watt.

All submissions are prepared in collaboration with Krai and are powered by Collective Knowledge v2.6.1.

SAPEON

SAPEON Inc. is a corporation founded by SK Telecom, SK Hynix, and SK Square to make inroads in the global AI semiconductor market.
The AI accelerator project, SAPEON, started in 2017 at SK Telecom with the goal of enhancing the internal AI infrastructure for SK Group. It has since been evolved to the second-generation product, the X220. The accelerators have been used for various applications like NUGU, Korea's first AI speaker, and our intelligent video security system SK Shieldus, the second largest physical security company in Korea.
After being spun off from SK Telecom, SAPEON Inc. disclosed the X220's performance results for the first time, which have been validated by MLPerf in order to introduce its product to the market. SAPEON X220, the second-generation product debuted in 2020, accelerates deep learning workloads in data centers.
SAPEON is optimized for real-time interactive AI inferences for a large number of users at one time, and the results are presented in MLPerf "Inference: Datacenter" benchmarks to measure the performance of those types of services.
The highlights from our MLPerf 2. 1 benchmark results are:
- SAPEON X220-Compact achieved 6,145 queries/s and 6,769 samples/s for ResNet 50-v1.5 in server and offline scenario respectively.
- SAPEON X220-Enterprise achieved 12,036 queries/s and 13,101 samples/s for ResNet 50-v1.5 in server and offline scenario respectively.

The cost and energy-efficiency of the X220 will be superior in the next-generation, X330, released in the second half of 2023. The X330 is a superior performing AI Accelerator for large-scale AI inference systems and provides precise operations for more accurate services. It can also be used for training infrastructure.
SAPEON is targeting the global AI markets with vertical solutions which integrates AI semiconductor chips and software such as AI algorithms and services with a proven track record in SK Group businesses such as media, security, smart factories, etc.

Supermicro

Supermicro has its long history of providing a broad portfolio of products for different AI use cases. In MLPerf Inference v2.1, we have submitted three systems with nine configurations in the datacenter and edge inference category. These are to address the performance for multiple use cases, including medical image segmentation, general object detection, recommendation systems, and natural language processing in centralized datacenter and distributed edges. Supermicro's DNA is to provide the most optimal hardware solution for your workloads and services. For example, we provide four different systems for NVIDIA's HGX A100 8GPU platform and HGX A100 4GPU respectively. Customers can configure the CPU and GPU baseboards based on their needs. Furthermore, we provide upgraded power supply versions to give you choices on using our cost-effective power solutions or genuine N+N redundancy to maximize your TCO. Supermicro also offers liquid cooling for HGX based-systems to help you deploy higher TDP GPU baseboards without thermal throttling.

For customers looking for PCIe platforms, Supermicro provides even more choices. In MLPerf v2.1, we submitted results for IoT SuperServer SYS-E403-12P-FN2T, a compact high performer with box PC form factor. With multiple selections of GPUs, the system is a perfect fit for edge AI applications such as predictive analysis, and network operation monitoring and management. The system is currently shipping worldwide.
Supermicro's SYS-120GQ-TNRT is a high density 1U server that accommodates 4 double width GPUs. It is designed for customers looking for compact but high AI performance. It can achieve 50% of the performance while only using 25% of the rack unit of our high end GPU servers.
We are happy to see all the results we ran on MLPerf using our portfolio of systems, and we will keep optimizing the solutions for customer's different requirements to help achieve the best TCO.