Supplemental Discussion

MLPerf™ Inference v2.0 Results Discussion

The following descriptions were provided by the submitting organizations 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 Sinian Platform is a heterogeneous hardware acceleration and optimization platform, targeting high execution efficiency for machine learning and data-intensive applications. It is a unified platform to support both machine learning training and inferencing, but fully tailorable for cloud computing, edge computing, and IoT devices. Sinian makes it seamless to build, train, and deploy machine learning models without suffering the loss of performance portability.

In this round of submission to MLPerf Inference v2.0, Alibaba team demonstrates the efficiency of our hardware and software co-optimization on accelerating benchmarks on Alibaba Yitian 710 and Alibaba Haishen.

For open division submissions, at the model level, we leveraged the Sinian architecture-aware model optimizer (SinianML) to automatically compress the neural network while satisfying the MLPerf Datacenter/Edge model accuracy requirement. The hardware-aware compression results in more architecture-friendly models at runtime. Overall, Sinian achieved quite impressive results at both datacenter and edge. On Alibaba Yitian CPU-only hardware, Sinian achieved ~35X performance improvement against the same model without Sinian optimizations while still preserving the same accuracy requirements.

For closed division submissions, we built the system with the capability of each hardware component balanced to maximize the performance for AI workloads. On Alibaba Haishen edge device, the inference time for ResNet50 and ssd-small is as low as 0.75ms and 1.72ms respectively with 15W accelerator TDP constraints.

ASUSTeK

ASUS delivers its first MLPerf submission result today since joining MLCommon community to contribute more innovations in artificial intelligence and applications. With this first result revealed, ASUS feature ESC8000A-E11 and ESC4000A-E11 powered by AMD 3rd Gen EPYC Processors and NVIDIA GPU solutions to construct a solid architecture for high efficiency computing performance. In Inference session, ASUS completed testing items including 3d-unet-99/3d-unet-99.9/bert-99/bert-99.9/dlrm-99/dlrm-99.9/resnet50/rnn-t/ssd-resnet34. With ASUS robust hardware and software design, we delivered ASUS servers that are good at performing complicated and highly intensive AI tasks.

In MLPerf Inference 2.0, we have made 2 model submissions showcasing ASUS capabilities in AI workloads.

Server	CPU	GPU
ASUS ESC8000A-E11	2 x AMD EPYC 7763 processor	8 x 80GB NVIDIA A100 PCIe
ASUS ESC4000A-E11	1 x AMD EPYC 7763 processor	4 x NVIDIA A30

ASUS has years of experience of building supercomputing, cloud solution, AI infrastructures and was involved in the design of Taiwania 2 supercomputing in 2018 which has achieved No.20 in TOP500 ranking. As a total solution provider, ASUS is working closely with partners worldwide to deliver datacenter solutions to customers. Joining with MLPerf community, it is expected to contribute more knowledge to the community and deliver profound product design in AI fields. Based on the first submission, ASUS is ready to deliver more tuning mechanisms to improve MLPerf results in the future.

Azure

Azure is pleased to share results from our first MLPerf inferencing submission. For this submission we benchmarked our NC_A100_v4, NDm_A100_v4, and our ND_v4 offerings. These offerings are our flagship virtual machine (VM) types for AI inferencing and training. These VMs are powered by the latest NVIDIA A100 SXM and PCIe GPUs. These offerings enable our customers to address their inferencing needs from 1 to 8 GPUs.

Some of the highlights from our MLPerf inferencing v2.0 benchmark results are:
1. bert-99: NDm_A100_v4 achieved 27.5K+ samples/s and ~22.5K queries/s for offline and server scenarios
2. resnet: NDm_A100_v4 achieved 300K+ samples/s and ~200K+ queries/s for offline and server scenarios
3. 3d-unet: NDm_A100_v4 achieved 24.87 samples/s for the offline scenario

These inferencing benchmark results demonstrate how Azure:
1. committed to providing our customers with the latest GPU offerings
2. is in line with on-premises performance
3. is committed to democratizing AI at scale in the cloud

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 VM offerings and Azure's Ubuntu 18.04-HPC marketplace image. By following the README.md provided by NVIDIA, we were able to achieve the published results.

The NC_A100_v4, NDm_A100_v4 and ND_v4 offerings are what we and our Azure customers turn to when large-scale AI and ML inferencing is required. We are excited to see what new breakthroughs our customers will make using these VMs.

Deci.ai

Deci's end-to-end deep learning development platform enables AI developers to build, optimize, and deploy faster and more accurate models for any environment. Deci submitted results for both NLP and CV inference on various hardware types for MLPerf 2.0 Datacenter opentrack.

The computer vision submissions were conducted on a 12-core Intel Cascade Lake CPU and two different Intel Ice Lake CPUs with 4 and 32-cores. Models were optimized on a batch size of 32 and quantized to INT8 using OpenVINO. Compared to the 8 INT ResNet50 model, Deci achieved 102.2% accuracy compared to the baseline model (78.14%) and delivered 2.8x to 4x improvement in throughput depending on the hardware type compared to the base ResNet50 model (INT 8).

Computer Vision Submission: Throughput [samples/sec] Results on Different Hardware Types

Hardware	ResNet-50 OpenVINO 32-bit	ResNet-50 OpenVINO 8-bit	DeciNet OpenVINO 8-bit
Intel Ice Lake 4 cores	62.0	259.65	1041.39
Intel Ice Lake 32 cores	415.6	1532.4	4307.71
Intel Cascade Lake 12 cores	212.9	713.2	2243.62

Deci's NLP models, called DeciBert models, were submitted to both BERT99 and BERT99.9 MLPerf tracks and conducted on a 16 core Intel Cascade Lake CPU and 32 core Ice Lake CPU machines. The DeciBertLarge models were optimized on a batch size of 16 and quantized to INT8 using OpenVINO.

Deci's NLP Submission Results:

	Bert-Large OpenVINO 32-bit	DeciBert 1 OpenVINO 8 Bit	DeciBert 2 OpenVINO 8 Bit
SQuAD F1 Score		89.9	91.09
Intel ICE Lake - 32 cores	11.75	121	90.5
Intel Cascade Lake - 16 cores	6.8	66.6	50

These submissions demonstrate the power of Deci's Automated Neural Architecture Construction (AutoNAC) technology, which automatically generated the DeciNets and DeciBert models, thus delivering breakthrough accuracy and throughput performance on CPUs. Deci's AutoNAC is powerful and transferable across deep learning domains. Deci looks forward to producing further results that will make deep learning accessible for the entire AI community.

Dell Technologies

Pushing technology so you can go further, Dell Technologies submitted 53 test results for MLPerf Inference v2.0 focused on AI at the edge.

“Over a given season, we end up with billions of data points. We capture that data at the edge and use all of it to inform our performance strategy,” says Edward Green Head of Commercial Technology, McLaren Racing.

Dell Technologies remains committed to providing customers with the performance data they need for their winning strategies. Dell edge servers crossed the finished line with flying colors in many categories! Look at the PowerEdge XR12 for performance per watt. See the PowerEdge XE2420 performance per GPU, with and without MIG.

We didn’t stop there. Dell Technologies submitted 86 more test results to help customers get the best price/performance for their AI workloads. The team tested edge, core data center and cloud servers with combinations of five different GPUs ranging from the NVIDIA T4 to the A100 SXM4-80GB with TensorRT and Triton. Hear the crowd roar for the PowerEdge R750xa in image recognition, speech-to-text and recommendation engines.

To get the most out of systems, optimization is a must, so we 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 is a leading company of information and communications technology systems. We support our customer’s business by providing robust and reliable ICT systems.

In addition to these features, we also believe that high power efficiency is important, so we entered the closed-power division from this round. In the future, we will continue to study the efficient use of GPU resources in order to realize highly efficient servers.

We have continued to participate in and submit to every inference and training round since 2020. In this round, we submitted benchmark results of two types of servers: PRIMERGY GX2570M6 and PRIMERGY GX2460M1. For the PRIMERGY GX2570M6 system, we also measured power consumption with the same configuration other than GPU power limit. The details of these systems are shown as follows:

1. PRIMERGY GX2570 M6 is a 4U server with Intel (R) Xeon (R) Platinum 8352 V CPUx2 and NVIDIA A100 SXM 40GB x8. It also supports the NVIDIA A100 SXM 80GB x8.
2. PRIMERGY GX2460 M1 is a 2U server with AMD EPYC 7302 CPUx2 and NVIDIA A30 x4/x2.

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.

FuriosaAI

Founded in 2017, FuriosaAI develops high performance AI inference accelerators targeted at data centers and enterprise customers. For the past five years, we have pursued an in-house full stack development, working on not just the AI hardware, but also optimized compiler and cloud-native software. FuriosaAI has so far attracted more than 70 global top-talent professionals—striking an unprecedented balance of HW and SW engineers. We seek to aggressively expand our team by more than double, tapping on the pool of USD 100 million injected by investors.

Following the first FPGA submission towards MLPerf Inference v0.5 in November 2019 and the second Inference v1.1 submission, this was FuriosaAI's third submission. FuriosaAI was the only startup to have submitted chip results in the closed_edge area, and recorded top-notch performance of 1) 0.71ms/2758FPS on ResNet, an area of image classification 2) 13.43ms/80FPS on SSD-Large, an area of object detection, and 3) 0.36ms/8762FPS on SSD-Small, also object detection. Of particular note is the radical improvement of SSD-Small performance—despite using the same silicon; compared to the last v1.1 submissions, latency improved by 15% and offline throughput was enhanced by a striking 113%. Given FuriosaAI's first chip, Warboy's support for depthwise separable convolution and group convolution, the chip can flexibly support tensors of various shapes. Based on such strengths in Warboy's architecture, we were able to achieve substantially enhanced results in SSD-Small only with enhancements in the compiler.

The in-house developed Metamorphic engine enables us to support generic tensor shapes, while flexible reconfiguration of internal interconnection topology depending on tensor shape, size, and operation type maximizes MAC utilization. The optimized compiler takes charge of optimizing data distribution, as well as configuration of optimal topology, thereby enabling Warboy to support more than 300 algorithms on ONNX in the vision area. Currently, FuriosaAI is sampling with customers in a variety of application areas including data centers, OCR, robotics, live streaming, intelligent transportation system, and smart retail. We are also closely coordinating with the two largest IT companies in Korea—Naver and Kakao.

FuriosaAI is currently developing a more comprehensive next-generation AI inference system, with performances that would exceed the company's own Warboy chip by tenfold. This system would not only be able to support vision areas, but also the majority of AI algorithms, and in particular large-scale language and recommendation models.

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 CPUs (AMD, Ampere, and Intel) in configurations to showcase systems that target different markets (x86 and Arm).

For MLPerf Inference v2.0, GIGABYTE has submitted a 4U server for the Intel Xeon Scalable platform supporting an NVIDIA HGX A100 8-GPU solution, and several other configurations using a 2U server for the AMD EPYC platform using Qualcomm Cloud AI 100 accelerators.

While GIGABYTE may not have had the resources to submit their own systems, it is still true that GIGABYTE was the most popular brand of servers used in testing MLPerf Inference v2.0.

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 committed to becoming the most trusted partner of its customers in their quest for business innovation and digital transformation. We offer a full portfolio of Digital Infrastructure products and a comprehensive one-stop digital platform as well as end-to-end technical services.

H3C participates in six model tests of MLPerf Inference 2.0 this time: ResNet50 v1.5, SSD-resnet34, 3D-Unet, RNN-T, BERT, DLRM.

H3C UniServer R5300 G5, R5500 G5 and R4900 G5 are adopted this time, which can be flexibly applied to various AI application scenarios including deep learning model training, deep learning reasoning, high performance computing and data analysis:

- ➢ H3C UniServer R5500 G5 supports HGX A100 8-GPU module, and eight A100 GPUs can be fully interconnected with six NVSWITCHes at 600GB/s. It adopts modular design and perfectly matches 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 HGX A100 4-GPU module, with four A100s interconnected with NVLINK at 600GB/s. It can also support 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.

H3C continues to integrate the technical concept of "native intelligence" into the product design, so that the server can effectively perceive the application demand for compute power, provide strong support for various workloads such as deep learning training, inference and data analysis, improve the utilization efficiency of the underlying hardware, and fully release the potential of compute power, thus providing high-quality server products and solutions for the enterprise AI data center.

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.

In MLPerf Inference V2.0, Inspur made submissions on four systems, NF5488A5, NF5688M6 NF5468M6J 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 12 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 DLRM is improved by 1.46% on NF5488A5, and the performance of Bert-99, Bert-99.9,Resnet50, RNN-T and SSD-Large are improved by 28.3%, 28.7%, 36.6%, 41% and 34.4% on NF5468M6J compared with the best performance Inspur achieved in Inference v1.1. In server scenario of datacenter closed division, the performance of DLRM and Resnet50 are improved by 2.88% and 0.34% on NF5688M6 , and the performance of Bert-99, Bert-99.9, Resnet50, RNN-T and SSD-Large are improved by 26.1%, 20.6%, 39.07%, 25.21% and 34.96% on NF5468M6J compared with the best performance that Inspur achieved in Inference v1.1.

In offline scenario of edge closed division, the performance of Resnet50, Bert, RNN-T, SSD-Large and SSD-Large are improved by 8.9%, 7.49%, 9.23%, 10.0% and 2.9%, respectively on NE5260M5 compared with the best performance achieved in Inference v1.1.

Intel

Intel's broad portfolio of hardware offerings enables innovative solutions for optimizing deep learning (DL) models that are specialized for a particular target device. Intel submitted MLPerf Inference v2.0 results on 3rd Gen Intel® Xeon® Scalable processor product line (codenamed Ice Lake). Intel demonstrates, as the only data center CPU vendor to submit MLPerf inference results on a broad set of models using only host CPUs, that it is practical to run DL inference anywhere and alongside other applications on the massive install base of Intel® Xeon® servers.

Our submission is in the datacenter Open Division of MLPerf v2.0 for the optimized ResNet50 benchmark. The recently launched 3rd Gen Intel Xeon Scalable processors (codenamed Ice Lake) delivers more compute and memory capacity/bandwidth than the previous generation (codenamed Cascade Lake). The Ice Lake system used has two Intel® Xeon Platinum 8380 CPUs @ 2.30GHz, 40 cores per socket (80 cores total), 64GB DIMMs, and has hyper-threading and turbo mode enabled.

Intel Labs has developed a BootstrapNAS approach to automatically discover alternative equivalent models optimized for Xeon platforms from a base DL model. In this submission we used Torchvision's existing pre-trained ResNet-50 as base model to produce high-performing alternative models using BootstrapNAS and INT8 acceleration. To showcase this capability, we submitted three models to demonstrate the significant performance gains with essentially no impact in accuracy by automatically generating models using BootstrapNAS. Compared to the ResNet50 base model (INT8) which was already accelerated using Intel DL Boost (VNNI), Model BNAS-A delivered an additional 2.55x higher throughput at 98.53% of FP32 accuracy, Model BNAS-B delivered an additional 1.43x higher throughput at 100.2% of FP32 accuracy, and model BNAS-C delivered an additional 1.89x higher throughput at 99.66% of FP32 accuracy.

All software used for this submission is available through the MLPerf inference repo. BootstrapNAS capability is integrated in the Intel® OpenVINO's Neural Network Compression Framework (NNCF) and Model Optimizer which will be released soon.

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 both making actual and "dry-run" MLPerf submissions. Truth be told, even great engineering teams working on great products can have blind spots. 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 third time round. In addition, we assisted Qualcomm's partners Alibaba and Gigabyte. 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 2,350 benchmarking results on our own, with over 2,000 results accompanied by power measurements. Our submissions demonstrate accuracy/performance/power trade-offs across a range of Edge platforms, workloads and engines.

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 understand the complexity of deploying AI (Artificial Intelligence) solutions that solve significant business challenges, but we believe choosing the best infrastructure for your workloads should not be a barrier to starting your AI initiatives. Our goal through MLPerf Inference v2.0 is to bring clarity to infrastructure decisions so our customers can focus on the success of their AI deployment overall.

In this round, Lenovo has demonstrated AI performance across a broad range of configurations and is the only vendor to submit results on NVIDIA A16 and NVIDIA Triton Inference Server running on liquid-cooled CPUs, demonstrating the performance benefits of our latest Lenovo Neptune™ systems.

NVIDIA A16 is a double-wide GPU that contains 4 separate chips, providing improved density and price/performance compared to other GPUs. With NVIDIA Triton Inference Server running on CPUs, we're showcasing the versatility of Triton Inference Server to run across both accelerated and heterogeneous platforms, deliver Exascale grade performance in a standard, dense data center footprint. Lenovo also showcased the efficiency and performance of air-cooled systems, providing both PCIe and HGX deployment options in a standard data center platform that can easily be deployed by enterprises of all sizes.

Discover all Lenovo has to offer including software and services solutions to accelerate your AI initiatives through the Lenovo AI Innovation Centers.

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.

With a focus on servers for 15 years, Nettrix has developed server products for industries including Internet, telecommunications, finance, medical care and education.

Nettrix X660 G45 is a high-performance computing platform specially developed for deep learning training. It is equipped with 8 NVIDIA Tesla SXM4 A100 GPUs.

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.

Neuchips

NEUCHIPS is an AI ASIC solution provider founded by a team of veteran IC and software design experts in 2019. With decades of experience from leading semiconductor companies, Neuchips holds patents in signal processing, neural networks, and circuit design. NEUCHIPS's mission is to develop purpose-built AI solutions for the Cloud which provides the most energy-efficient and cost-efficient deep learning inference accelerators that deliver the lowest TCO for data centers.

In 2020, Neuchips joined the MLPerf v0.7 for the 1st Recommendation inference-datacenter and kept submitting its improving result by optimizing its software, algorithm, and logic circuit. In MLPerf 2.0, Neuchips team pushed the hardware utilization to the limit by adapting recommendation queries with their patent design. The performance result shows RecAccel™-FPGA enhanced 42.6% in server-mode and 46.9% in offline-mode compared to previous submissions with the same hardware platform.

NEUCHIPS is developing the purpose-built recommender chip, RecAccel™ N3000, on the advanced process technology and ONNX support. Designed to be scalable and has an optimized hardware and software stack to deliver superior performance while being energy efficient and offering low TCO. The samples plan to be released in 2022H2 for boosting recommendation performance in data centers.

NVIDIA

Overall, NVIDIA continued to show great performance across all workloads and scenarios in this fifth round of the industry-standard benchmark for AI inference. In MLPerf Inference version 2.0, our new Jetson Orin SoC made its debut as a preview submission, showing excellent performance on all MLPerf Inference Edge workloads, and delivering up to 5x higher performance than our previous generation, Jetson AGX Xavier, while delivering an average of 2x better energy efficiency. Our NVIDIA A100 Tensor Core GPUs again delivered exceptional performance and efficiency results across all tests.

The NVIDIA AI platform again attracted a large number of MLPerf submissions from a broad ecosystem of partners, including Microsoft Azure, and system makers ASUStek and H3C, all of whom made their MLPerf debut in this round with submissions using the NVIDIA AI platform. They joined returning system makers Dell, Fujitsu, Gigabyte, Hewlett-Packard Enterprise, Inspur, Lenovo, Nettrix, and Supermicro that submitted results on more than two dozen NVIDIA-Certified Systems.

Software plays a pivotal role in making great AI inference performance happen, and two key components that enabled our inference results — NVIDIA TensorRT for optimizing AI models and NVIDIA Triton Inference Server for deploying them efficiently — are available free on NGC, our catalog of GPU-optimized software. TensorRT's extensive library of optimized GPU kernels has been extended to support Jetson Orin, and the plugins used in MLPerf networks have all been ported to Orin and added to TensorRT 8.4. Triton eases deployment of trained models, is tightly integrated with Kubernetes and can manage AI inference work on GPUs as well as x86 and Arm CPUs.

We look forward to upcoming rounds of MLPerf testing for AI training, AI inference and HPC. MLPerf's broad set of tests spans a wide array of today's most popular AI workloads and scenarios. These diverse results help IT decision marks make data-driven platform investments best suited to their particular needs.

Qualcomm

Qualcomm® Cloud AI 100 delivers solutions from 70 TOPS on edge devices to 100+ Peta-ops in datacenter racks.

Qualcomm MLPerf v2.0 inference results builds on MLPerf ™ v1.0 and v1.1 demonstrated performance-to-power efficiency leadership, from edge to cloud. We have expanded Cloud AI 100 benchmark submission scope with 3 new partners, Gigabyte, Alibaba and Krai. Including partners' results, submissions have more than doubled to 200+ from previous submission. We continue to innovate and optimize the solutions across all our submissions

For example, on ResNet-50, power efficiency has increased by 17% to 230 inference/Sec/Watt on servers configured with 8x accelerator cards. The 18x-platform for ResNet-50 delivers offline: 418,794 queries/s, whereas 16x platform delivers 371,473 queries/S within a compact 2U rack server from Gigabyte. A single datacenter rack configured with 20x servers with 16x accelerators can provide 7.4+ million inferences per second, while also achieving industry-leading power efficiency at that scale.

In MLPerf v2.0 datacenter power submissions, we have optimized Energy Efficiency (EE) further to demonstrate power efficiency at its peak.

Qualcomm Technologies has a history of demonstrating leadership in power efficiency. We have introduced commercial edge appliance solution Gloria from Foxconn powered by Snapdragon with Qualcomm® Cloud AI 100 DM.2e accelerator while delivering maximum performance per watt. The measured efficiency has now attained 300 Inf/sec/Watt for Resnet-50 and 154 inf/sec/watt on SSD Small networks. Qualcomm’s BERT Power submission showcases 9.71 inf/sec/watt. All our datacenter and edge power submissions have set new standards for power efficiency.

On edge server submissions we have achieved extremely low latency of 0.33ms and improved all vision network latencies upto 4x times. The low multistream latencies showcase compute efficiencies of Cloud AI 100.

Qualcomm Technologies worked with partner Krai for the MLPerf v2.0 benchmark submission and all submissions are powered by Collective Knowledge v2.6.1.

Supermicro

Supermicro has its long history of providing a broad portfolio of products for different AI use cases. In MLPerf Inference v2.0, we have submitted three systems with six 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 v2.0, we submitted results for IoT SuperServer SYS-220HE-FTNR, a compact high performer featuring short depth chassis, with the option of telecom-standard NEBS Level 3 compliance AC or DC power versions. 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.

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.

ZhejiangLab

Zhejiang Lab is a research institute established jointly by the Zhejiang province government, Zhejiang University and Alibaba group. We belong to the advanced computer research center in Zhejiang Lab. We aim to develop advanced computer hardware and software systems for various dedicated industry fields.

This is the first time we submit results to the MLPerf. We improve the inference speed by optimizing some operators and reducing data movement costs. These graph level optimizations will reduce the inference time as much as possible.