Supplemental Discussion

MLPerf™ Inference 3.0. 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 the MLCommons® Association.**

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Alibaba

Alibaba Cloud Server SinianML 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. SinianML 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 v3.0, Alibaba team demonstrated the efficiency of their hardware and software co-optimization on accelerating benchmarks on RockChip RK3588.

For Edge open division submissions, at the model level, the team leveraged the Sinian architecture-aware model optimization (SinianML) to compress the neural network automatically while satisfying MLPerf's model accuracy requirements. The hardware-aware compression led to more architecture-friendly models at the RKNN runtime. Taking full advantage of RK3588, Alibaba achieved quite impressive results, including 12.53 ms in SingleStream, 29.70 ms in MultiStream, and 971.27 samples/s in Offline, respectively.

For Edge open-power division submissions, the Alibaba team leveraged the capability of RK3588's accelerator to balance the performance and power consumption during ResNet50 inference. Overall, Alibaba team revealed the efficient results, with 77.09 millijoules in SingleStream, 227.31 millijoules in SingleStream, and 14.96 Watts in Offline, respectively.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

ASUSTeK

ASUSTek is delighted to announce its participation in the MLPerf Inference v3.0 benchmark for 2023. In line with its commitment to delivering cutting-edge technology, ASUS has adopted the latest Intel Sapphire Rapid platform (ESC8000-E11) and AMD Genoa Platform (ESC4000A-E12) to run the inference benchmark. Additionally, ASUS has incorporated NVIDIA's newest generation GPU cards, the L4 and H100 PCIe, into the benchmark evaluation to examine various user scenarios and bring MLPerf Inference v3.0 closer to real-world use cases.

ASUS recognizes that the latest benchmark will provide exciting results, particularly as more and more users leverage Large Language Models (LLMs) for creative work. Generative AI has become a burgeoning trend, and it has opened up significant opportunities for server vendors, ASICs, software developers, and other industries. As such, MLCommons benchmarks are relevant to all types of AI applications. By participating in the MLPerf Inference v3.0, ASUS is confident that it will make valuable contributions to advancing AI development.

ASUS is committed to developing comprehensive solutions for a wide range of workloads, including HPC, AI, enterprise, and key verticals. In addition to rack servers, ASUS also offers GPGPU servers and edge solutions to provide a total solution for our customers. Our goal is to offer a complete range of products that can meet the diverse needs of our customers across various industries.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Azure

Azure is pleased to share results from our MLPerf Inference v3.0 submission. For this submission, we benchmarked our NC A100 v4-series, with a full GPU and a single MIG7 instance (1/7th of an NVIDIA A100 GPU). Customers can deploy a full NC A100 v4 virtual machine and create up to seven instances to run jobs concurrently with Multi-Instance GPU (MIG) technology. This offering is powered by the NVIDIA A100 PCIe GPUs and offers great flexibility for a wide range of workloads: one of the outstanding benefits of the NC A100 v4-series is the capacity to run jobs on the full GPUs or to run jobs in parallel on 2, 3, or 7 partitions of the GPU.

These inference benchmark results demonstrate how Azure is committed to empower customers to do more with less. Azure GPU offerings are **in line with on-premises performance**, scale to **adapt to all sizes of AI workloads and needs**, and are **available on-demand in the cloud**.

The NC A100 v4-series are what our Azure customers turn to when adaptability from small-scale to large-scale in AI and ML inference is required. We are excited to see what new breakthroughs our customers will make using these VMs.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

cTuning

The MLCommons taskforce on education and reproducibility is excited to announce the release of the Collective Knowledge Playground (CK) - a free, open-source and technology-agnostic platform to help all MLCommons members, their partners and the community run, optimize, compare and discuss MLPerf benchmarks with any model, data, software and hardware in a fully automated and reproducible way via public and private optimization challenges while slashing their development, experimentation and submission costs.

We are very pleased to announce the successful outcome of the 1st public MLCommons Collective Knowledge challenge to run, reproduce and optimize MLPerf inference v3.0 benchmarks led by cTuning foundation and cKnowledge Ltd: our open-source CK technology has helped to automate, unify and reproduce more than 80% of all submission results including 98% of power results with very diverse technology and benchmark implementations from Neural Magic, Qualcomm, Krai, cKnowledge, cTuning, DELL, HPE, Lenovo, Hugging Face, Nvidia and Apple across diverse CPUs, GPUs and DSPs with PyTorch, ONNX, QAIC, TF/TFLite, TVM and TensorRT using popular cloud providers (GCP, AWS, Azure) and individual servers and edge devices provided by the CK users and contributors.

We invite all MLCommons members to join our open MLCommons taskforce via the public Discord server to improve and adapt our open-source platform to their needs, use-cases and technology, help them automate, optimize and reproduce their end-to-end AI solutions and future MLPerf submissions, and collaborate with ACM, IEEE and the community on the new exciting optimization and reproducibility challenges and competitions that we are organizing in 2023!

The ultimate goal is to use the Collective Knowledge to let anyone automatically generate the most efficient, reproducible and deployable full-stack AI/ML solution using the most suitable SW/HW stack at that time (model, framework, inference engine and any other related dependency) based on their requirements and constraints including costs, throughput, latency, power consumption, accuracy, target devices (cloud/edge/mobile/tiny), environment and data.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Deci

Deci, the deep learning development platform is used by leading AI teams to build, optimize and deploy highly accurate and efficient models to production. Deci demonstrated a record-breaking NLP inference performance on NVIDIA GPUs. Deci submitted its models under the offline scenario in MLPerf's open division in the BERT 99.9% category.

For the submission, Deci leveraged its platform powered by proprietary Automated Neural Architecture Construction technology (AutoNAC). The AutoNAC engine empowers teams to develop hardware aware model architectures tailored for reaching specific performance targets on any inference hardware.

AutoNAC was used by Deci to generate models tailored for various NVIDIA processors and presented unparalleled performance on the NVIDIA A30 GPU, NVIDIA A100 GPU (1 & 8 cores configurations), and NVIDIA H100 GPU.

One of the models generated by Deci, DeciBERT-Large, delivered a record-breaking throughput performance of over 100,000 queries per second on 8 NVIDIA A100 GPUs while also delivering an increase in accuracy compared to the original baseline.

Furthermore, with the throughput performance delivered by Deci on the NVIDIA H100, ML engineers can now achieve a higher throughput on one NVIDIA H100 card than on 8 NVIDIA A100 cards combined. A transition from multi-GPU to a single GPU translates into lower inference cost and reduced engineering efforts. On the NVIDIA A30 GPU, which is a more affordable GPU, Deci delivered accelerated throughput and a 0.4% increase in F1 accuracy compared to an FP32 baseline. By using Deci, teams that previously needed to run on an A100 GPU can now migrate their workloads to the NVIDIA A30 GPU and achieve 3x better performance then they previously had for roughly a third of the compute price.

Leading enterprises and industry leaders are using the Deci platform to boost their deep learning models' performance, shorten development cycles and reduce computing costs.

To learn more about Deci please visit: https://deci.ai/

Hardware	BERT + INT8 throughput	Deci + INT 8 throughput	BERT F1 Accuracy	Optimized F1 Accuracy	Accuracy Increase
1x NVIDIA A30 GPU	**	5885	90.874	91.28163425	0.4076342467
1xNVIDIA A100 GPU	1,756.84	13,377	90.874	91.4300145	0.5560145021
8xNVIDIA A100 GPU	13,967.50	103,053	90.874	91.4300145	0.5560145021
NVIDIA H100 GPU	7,921.10	17,584	90.874	91.34622414	0.4722241354

Figure 1: Deci's results in the Bert-99.9 offline category. Throughput in queries per second.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Dell

The potential of AI-driven applications can transform and revolutionize business, and a powerful foundation of solutions enables businesses to fully maximize performance to accelerate insights.

Our Engineering teams submitted MLPerf Inference v.3.0 results for a wide variety of server CPU-accelerator combinations based on the latest technologies, providing the data you need to make the best choices for your AI workloads and environments.

Dell Technologies works with customers and partners including NVIDIA, Intel, VMware, Qualcomm and AMD, to optimize software-hardware stacks for performance and efficiency, boosting inferencing workloads including generative AI.

Here are some of the latest highlights:

- Our newest server, the Dell PowerEdge XE9680 which delivers cutting-edge performance with 8x NVIDIA H100 SXM GPUs, achieving tremendous gains across many benchmarks, including up to 800% faster versus current 4-way A100 SXM solutions
- New servers with latest 4th Generation Intel Xeon® Scalable processor family (Sapphire Rapids) with Intel AMX ® extensions to boost overall inferencing performance.
- New power efficient results with Qualcomm AIC100 Pro, STD, and Lite variants and Dell PowerEdge R7515, R650, and XR4000 servers. The computer vision ResNet50 offline case on XR4520c with 2x QAIC100 Lite is 124 inferences /sec /watt!
- Fantastic performance in Edge use-cases such as:
  - Preview performance results with NVIDIA L4 on PowerEdge XR7620 and XR5610, including excellent performance per watt
  - New verified results with Dell PowerEdge XR4000 and NVIDIA A30, A2 and Qualcomm AIC100, and higher performance per watt from the ruggedized XR4000 usable at the edge in telco, utilities, and defense.
- Performance increases across many configurations with PCIe GPUs
- Wide variety of results covering virtualized systems, different host operating systems, and all closed scenarios

With the ability to tap into the power of AI and an ever-growing foundation of technology, businesses can deploy a foundation for successful AI initiatives with Dell Technologies solutions.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Gigabyte

GIGABYTE is an industry leader in HPC & AI servers, and uses its hardware expertise, patented innovations, and industry connections to create, inspire, and advance. With over 30 years of motherboard manufacturing excellence and 20 years of design and production of server and enterprise products, GIGABYTE offers an extensive portfolio of data center products for x86 and Arm platforms.

In 2020, GIGABYTE joined MLCommons and submitted its first system- a 2U GPU server with AMD EPYC processors and sixteen NVIDIA T4 GPUs. Since then, we have continued to support the community by submitting our own systems for training and inferencing, as well as providing servers to others. In the past, we submitted systems for AMD EPYC and Intel Xeon processors, but last year we saw our first GIGABYTE systems using the Arm-based processor, Ampere Altra.

For MLPerf Inference v3.0 we have shown the performance that is possible in our G493-SB0 server with two Intel Xeon 8480+ processors and eight NVIDIA H100 PCIe 80GB accelerators. The results for this new generation of CPU and GPU technologies have been very impressive to us. Last, it’s worth mentioning that we were chosen by NVIDIA, Qualcomm, and NEUCHIPS as their first choice of servers for their own testing.

Results compared to v2.1 submission:
Image classification: 33-35% improvement
Object detection: 71-85% improvement
Medical imaging: 47% improvement
Speech-to-text: 11-19% improvement
Natural language processing: 133-182% improvement

To learn more about our solutions, visit: https://www.gigabyte.com/Enterprise

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

HPE

Hewlett Packard Enterprise (HPE) MLPerf 3.0 Inference results represent record setting performance, scalability, and efficiency. These results align to HPE's strategy of delivering AI-at-Scale from edge to cloud and is possible through ongoing collaboration with **Qualcomm Technologies** and **Intel.**

Qualcomm Technologies

HPE submitted best-in-class leadership MLPerf 3.0 Inference results for both edge and datacenter applications working closely with Qualcomm and Krai in producing these results.

- HPE is the first vendor to deliver a datacenter submission on MLPerf 3.0 (ProLiant DL385 Gen10 Plus) using the cost-effective Qualcomm Cloud AI100 **Standard** offering.
- HPE ProLiant DL385 Gen10 Plus with eight Qualcomm Cloud AI100 Standard cards, obtained best-in-class performance and near-linear scalability compared to the HPE Edgeline EL8000 with four AI100 Standard cards.
- HPE Edgeline EL8000 with Qualcomm Cloud AI100 remains the best performing single-socket system designed for true edge environments.
- HPE systems with Qualcomm Cloud AI 100 Standard also demonstrated best in class low lowest latency.

Intel

HPE realizes the importance of CPU AI performance as being an essential element of our customers' edge-to-cloud AI strategy, and to that end we have closely collaborated with Intel for our MLPerf 3.0 inference submissions across five workloads on HPE ProLiant DL380a Gen11 server and the 4th Gen Intel® Xeon® Scalable processor family. Our exceptional CPU results highlight a step-jump in AI performance of 6.6x (a geomean ratio, which is an unofficial measure that is not verified by MLCommons) compared to previous industry submissions on 3rd Gen Intel® Xeon® Scalable processors. This enables our customers to streamline and scale out their AI deployments using the HPE servers that are already running their business.

Information on these HPE Platforms and results can be found at www.hpe.com/ai

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

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 V3.0, Inspur made submissions on two systems: NF5468M6 and NE5260M5.

NF5468M6 is a highly versatile 4U AI server supporting between 4 and 16 NVIDIA single and double-width GPUs, making it ideal for a wide range of AI applications including AI cloud, IVA, video processing and much more. NF5468M6 offers ultra-high storage capacity and the unique function of switching topologies between Balance, Common and Cascade in one click, which helps to flexibly adapt to various needs for AI application performance optimization.

NE5260M5, a flexible, intelligent 2U 2-socket server designed for edge scenarios. Optimized for safe and reliable operations in edge environments, including MEC, 5G / IOT, and AR/VR. A powerful compute node with front I/O access, it can also act as a head node to seamlessly drive an accelerator expansion box with suitable retimer or redriver adapters.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Intel

Intel submitted MLPerf Inference v3.0 results on the 4th Gen Intel® Xeon® Scalable processor product line (codenamed Sapphire Rapids), a platform designed with the flexibility to run any AI code with industry-standard AI frameworks to enable AI practitioners customize models for large scale data center deployments, along with application engineers who want to quickly integrate models into their applications for deployment at the edge.

Our goal with these submissions remains the same: Demonstrate how 4th Gen Xeon processors can run any AI workload – including vision, language processing, recommendation, and speech recognition – and deliver remarkable gains, of up to 5x across these tasks, compared to 3rd Gen Intel® Xeon® Scalable processors – based on v1.1 to v2.x results from last year – due in large part to specialized AI hardware accelerators, including the Intel® Advanced Matrix Extension (Intel® AMX).

To simplify the use of these new accelerator engines and extract the best performance, we remain laser-focused on software. Our customers tend to use the mainstream distributions of the most popular AI frameworks and tools, and Intel's AI experts have been working for years with the AI community to co-develop and optimize a broad range of open and free-to-use tools, optimized libraries, and industry frameworks to deliver the best out-of-the-box performance regardless of the Xeon generation.

Compared to our 4th Gen Xeon preview results last August, we were able to improve performance by 1.2x and 1.4x for Offline and Server scenarios respectively, which underscores how we are continuously optimizing the AI software stack to realize full utilization of the unique capabilities of our Xeon CPUs. We've also closely collaborated with our OEM partners – namely HPE, Dell, and Quanta – by helping them tune their submissions using our recipes, further showcasing AI performance scalability and portability on servers powered by Intel® Xeon processors.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex .

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Krai

KRAI, headquartered in Cambridge, UK, is an industry leader in benchmarking and optimization. Our team has been actively contributing to MLPerf since its inception in 2018, having prepared over 60% of all Inference and over 85% of all Power results to date. This feat was possible thanks to our unique workflow automation approach that unlocks the full potential of AI, software, and hardware co-design. We are proud to partner with companies such as Qualcomm, HPE, Dell, Lenovo, and more.

To support our partners' submissions in this and future rounds, we have released a special "MLPerf edition" of our KRAI Inference Library Template (KILT) codebase under a permissive open-source license. The released code supports Qualcomm Cloud AI 100 accelerators out-of-the-box, and can be customized to support new Inference accelerators. KILT has been battle-hardened over the last two years to produce some of the fastest and most power-efficient results in the history of MLPerf Inference. For example, in this round we used KILT to achieve up to 242 QPS/Watt on Datacenter servers and up to 316 QPS/Watt on Edge appliances on the ResNet50 benchmark. KILT features a highly efficient implementation of the variable-workload Server scenario, achieving on average 95% of the peak performance under the Offline scenario. In addition, KILT has powered 2 out of 3 submissions under the up-and-coming Network category.

We embrace the growing need for benchmarking and optimization of full-stack AI solutions. We are committed to continuing our efforts of pushing the boundaries of MLPerf benchmarking. In addition to KILT, we are excited to pre-announce releasing the KRAI X automation technology in the near future. KRAI technologies deliver fully optimized end-to-end AI solutions while slashing the required engineering resources, development time, and operating costs. We are thrilled to help more companies develop, benchmark and optimize their AI solutions.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Lenovo

Lenovo delivers smarter technology for all, for hardware, software and more importantly solutions that customers have come to know of. Being smarter requires the research, testing and benchmarking that MLPerf Inference v.3.0 provides. Allowing you to see first-hand how delivering smarter technology means producing best-in-class results.

Since partnering with MLCommons, Lenovo has been able to showcase such results quarterly in the MLPerf benchmarking. Achieving these leading results is one side of the benefit, whereas the other side allows us to constantly be improving our own technology for our customers based on our leading benchmarks by closely working with our partners in optimizing the overall performance.

We are excited to announce with our partners NVIDIA and Intel, that we competed on multiple important AI tasks such as Image Classification, Medical Image Segmentation, Speech-to-text, and Natural Language Processing running on our ThinkSystem SR670 V2 Server with 8x 80GB PCIe A100. Along with our ThinkSystem SE350 and SE450 Edge Servers through our Qualcomm partnership. These partnerships have allowed us to consistently achieve improving results.

This partnership with MLCommons is key to the growth of our products by providing insights into where we stand against the competition, baselines for customer expectations, and the ability to improve. MLCommons allows Lenovo to collaborate and engage with industry experts to create growth and in the end produce better products for our customers. Which in today’s world is the number one focus here at Lenovo, customer satisfaction.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Moffett AI

Moffett AI is the leader in sparse AI computing, dedicated to providing AI computing platforms and services, with a mission to continue evolving the frontiers of AI performance using sparse computing.

Following the outstanding performance in the MLPerf Inference v2.1 benchmarks last year, Moffett AI has once again submitted impressive results for the Sparse Processor Unit (SPU) Family, including the S10 Accelerator, the S30 Accelerator, and the new S40 Accelerator. These results were achieved on ResNet50 and Bert-99.9 models in server and offline modes in the Data Center Open Division of MLPerf v3.0.

The SPU Family is powered by Moffett's Antoum processor - the world's First AI computing Accelerators with 32x sparsity. Moffett's patented deep dual sparsity algorithm and the co-designed Antoum chip and software platform architecture drastically increase the computing performance of the accelerators, increasing throughput, reducing latency and power, maintaining accuracy, while also significantly reducing the Total Cost of Ownership (TCO).

Moffett AI's submissions showcase the remarkable benefits of the SPU Family:

Enabling extraordinarily low TCO: Moffett AI accelerators perform computing workloads with less infrastructure, simplified deployment, and reduced Total Cost of Ownership (TCO) to achieve overall cost reduction and efficiency.

Excellent performance in demanding scenarios for high throughput and low latency: As a result of the hardware and software co-design, the SPU Family delivers outstanding results in offline and server scenarios alike.

Continued advancement and dominance in product performance: on ResNet-50, the S30 accelerator shows further improvement building on its best-in-class result from last year. The S40 accelerator makes its debut and achieves a new record high performance (127,375 FPS).

Scalable highly performant across data center ecosystem: Moffett’s partners Supermicro and others also submitted results for the S30 accelerator running on their respective servers. The S30 accelerator performed exceptionally well in single-card and 4-card mode across all these servers.

Moffett's deep sparse industry leading performance is optimal for generative AI workloads such as chatGPT. These AI models are massively increasing in size and the demand for computing performance is skyrocketing, as well as the need to reduce power, latency and TCO.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

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.
In MLCommons Inference v3.0, Nettrix has completed submissions on four systems: X620 G40, X620 G50, X640 G40 and X640 G50.

Nettrix X620 G40 is a high-end 2U rack-mounted server based on a third-generation Intel Xeon scalable processor. The product supports a variety of different GPU topologies, and optimizes GPU interconnection for different applications and models. At the same time, the X620 G40 is fully competent for desktop virtualization, cloud gaming, video flow code multi-purpose requirements, and it offers a combination of specialization and many abilities.

As the next generation product, Nettrix X620 G50 uses the latest generation of Intel Xeon expandable processors to improve performance and support PCIe 5.0 high-speed bus technology. The X620 G50 perfectly supports the latest GPU acceleration services of NVIDIA and it is an ideal AI inference platform. Resources can be flexibly configured based on service requirements to meet application requirements of various industries. Perfect for machine learning, AI inference, cloud computing, high-performance computing, virtualization, office system and other application scenarios.

Nettrix X640 G40 is an all-around GPU server with both training and inference 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 to meet the flexible needs of customers.

X640 G50 is a new generation of AI server products to be released soon, which can provide customers with more efficient AI services and more powerful AI computing power.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

NEUCHIPS

NEUCHIPS, an AI ASIC solution provider, was established in 2019 by a team of veteran IC and software design experts with decades of experience from leading semiconductor companies. With patents in signal processing, neural networks, and circuit design, NEUCHIPS is committed to designing purpose-built AI solutions for the Cloud that deliver the most energy-efficient deep learning inference accelerators and yield the lowest TCO for data centers.

NEUCHIPS recently showcased its energy-efficient inference processing capabilities using its first purpose-built RecAccel™ N3000 ASIC platform during the MLPerf™ V3.0 benchmark testing. The results were remarkable, with maximum performance of 856,398 Queries/Sec for deep learning recommendation models (DLRM) and excellent energy efficiency of 1,060 Queries/Sec/Watt. NEUCHIPS' optimized design also demonstrated nearly 100% scalability across eight RecAccel™-N3000-32G-PCIe cards. This exceptional performance is attributed to NEUCHIPS' calibrator, which delivers the highest 99.9% accuracy inferences for both Closed and Closed power divisions.

NEUCHIPS' RecAccel™ N3000 inference platform is a game-changing solution with the potential to revolutionize datacenter recommendation systems. We are excited to witness the impact this technology will have on the industry, and we are confident that it will set a new standard for energy-efficient and high-performance deep learning inference accelerators.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Neural Magic

While pursuing research at MIT, Nir Shavit and Alexander Matveev felt constrained by GPUs and hardware options to achieve the results they needed with deep learning. This frustration led them to develop technology that unfetters innovation with machine learning and to the start of Neural Magic. Now, with the use of software and algorithms, customers can attain the outcomes they need for deep learning deployments on commodity CPUs.

Our open-source compression framework, SparseML, unifies state-of-the-art sparsification algorithms for easy use and application across ML use cases like computer vision (CV) and natural language processing (NLP). DeepSparse is our sparsity-aware inference runtime that delivers GPU-class performance on commodity CPUs, purely in software, anywhere.

For this year's MLPerf Inference v3.0 benchmark testing, we are pleased to be the first organization to leverage the new 4th Gen AMD EPYC™ Processors for our ML workloads. DeepSparse demonstrates an impressive increase in performance, decrease in file size, and improved accuracy, all while executing on commodity CPUs. We confirmed an overall 6X performance uplift over our results from our previous submission, once again validating the power of sparse execution on CPUs.

For ResNet-50 image classification, we sparsified the weights to 11 MB and improved performance 13X to an impressive 19,632 samples/sec. For BERT-99 question answering, we sparsified the weights by two orders of magnitude to 10 MB and improved performance 1000X to 5,578 samples/sec.

Benchmark	File Size	ImageNet Top1 Accuracy	Offline Throughput
ResNet-50 ONNX Runtime	97.7 MB	76.456% (R=100.00%)	1,488.4
ResNet-50 DeepSparse	**11.0 MB**	**75.712% (R=99.02%)**	**19,632.1**

Benchmark	File Size	SQuAD v1.1 F1 Score	Offline Throughput
BERT-Large ONNX Runtime	1.3 GB	90.874 (R=100.00%)	4.6
oBERT-Large DeepSparse	38.2 MB	90.03 (R=99.07%)	1,367.14
oBERT-MobileBERT 99.9% DeepSparse	19.4 MB	90.80 (R=99.92%)	3,275.62
oBERT-MobileBERT 99% DeepSparse	**9.56 MB**	**90.41 (R=99.49%)**	**5,578.73**

Fueled by these results, Neural Magic is excited to help customers unlock the full potential of their ML environment, to accommodate their deep learning requirements without the added complexity of specialized hardware or cost.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

NVIDIA

In MLPerf Inference 3.0, we are excited to make our first DGX H100 submission, which delivered up to 54% more performance per accelerator compared to our first H100 submission last round. Additionally, we made power submissions using H100 PCIe, demonstrating up to a 2.4x uplift in performance/watt compared to A100 PCIe.

We were also pleased to make our first submission using the L4 Tensor Core GPU, powered by the NVIDIA Ada Lovelace architecture with new 4th Gen Tensor Cores with FP8. It delivered up to 3x more performance compared to our prior-generation T4 product while retaining the same single-slot, low-profile PCIe card form factor.

In the new Network Division, we made submissions on both ResNet-50 and BERT using our DGX A100 system running over InfiniBand, delivering nearly equal performance versus running the workloads entirely within the node.

The NVIDIA AI platform delivers innovation across the full stack, accelerates the entire AI workflow end-to-end – from data preparation to model training to deployed inference – and achieves great performance across a broad range of AI models. It is also available from every major cloud and server maker, and offers the quickest path to production AI and enterprise-grade support with NVIDIA AI Enterprise.

We are thrilled to see 11 NVIDIA partners submit great inference results, including first-time submissions from two of our partners, with both on-prem and cloud solutions spanning the breadth of our data center GPU portfolio.

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

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Qualcomm Technologies, Inc.

Qualcomm Technologies'® MLPerf™ v3.0 inference results demonstrate its commitment to the MLCommons® MLPerf™ Benchmarks by extending their submission scope to all benchmark divisions including closed, open and network, and all power submissions. All benchmarks show an increase in performance and power efficiency for NLP and Computer Vision networks. For the first time, Qualcomm® introduced the Qualcomm® Cloud AI 100 PCIe Lite accelerator, which is configurable from 35-55W TDP addressing AI Edge market requirements. The newly introduced platforms include the HPE ProLiant DL385 server, Dell PowerEdge XR4520c, Dell PowerEdge R650 and Lenovo ThinkEdge SE450. Qualcomm collaborates closely with KRAI for its MLPerf™ Benchmark submissions, who has also submitted several Cloud AI 100 based ML benchmarks.

Qualcomm Technologies' benchmark results have exceptional peak offline performance, power efficiency, and lower latencies in all categories. The 2U datacenter server platform with 18x Qualcomm Cloud AI 100 Pro (75W TDP) accelerators achieved the ResNet50 Offline peak performance of 430K+ inference per second and achieved power efficiency of 241 inference/second/watt on 8x Cloud AI100 platform. The Gloria Edge Appliance reached its peak performance and power efficiency for all three submitted neural network categories. It recorded power efficiency (inference/second/Watt) 315 for ResNet50, 5.64 for RetinaNet and 12.13 for BERT. The RetinaNet Offline and Server performance across all platforms has been optimized up to 110%.

For the first time, Qualcomm Technologies made network division submissions applicable for the datacenter and achieved nearly the same results as the Closed division.

In the Open division, Qualcomm Technologies submitted 3 BERT Networks (Block Pruned 99, Block Pruned 99.9 and DistilBERT) that demonstrated 3 to 4 times higher performance (inference/second) as well as power efficiency reaching up to 39.34 inference/second/watt.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Qualcomm and Qualcomm Robotics are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Rebellions

Rebellions Inc. is a Korean AI startup established in 2020, leading the design of AI hardware accelerators and software solutions. Rebellions' latest inference SoC, ATOM™, is built to deliver the best performance and energy efficiency and natively supports various models such as CNNs, RNNs, and Transformers. Equipped with Rebellions' proprietary full-stack software suite, ATOM™ plans to be deployed in KT (Korea Telecom) Cloud datacenters in 2023.

In this round of MLPerf inference v3.0, Rebellions' first submission, the team achieved very strong results demonstrating its versatile yet extremely low latency core architecture. Rebellions showed impressive results across both vision and language models, including 0.24ms in ResNet50 SingleStream and 4.30ms in BERT-Large SingleStream respectively. The Rebellions team accomplished this impressive feat while only having two months to work on ATOM™ silicon. To clarify, our MLPerf v3.0 results do not yet include software optimization. We are excited and looking forward to submitting software-optimized results in the next MLPerf inference round.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

SiMa

SiMa.ai is a startup and new-comer to MLPerf™ that focuses on push-button power efficiency for any embedded edge computer vision AI model. Power efficiency for “Single Stream System Energy” of 15.29 millijoules tells the story of SiMa’s focus on power draw and battery life for any model at “the edge” - such as cars, drones and robots. SiMa’s hardware is purpose built for edge computer vision applications and targets a different customer than Datacenter AI and Cloud AI specific chips.

SiMa’s secret to success is their purpose-built Machine Learning System-on-Chip (MLSoC) features a 50 TOPS proprietary Machine Learning Accelerator (MLA). The MLA is paired with an ARM Cortex 65 and EV74 processor which all share 16 GB RAM and are packaged up into a Dual M.2 form factor. This first generation architecture at a 16nm process promises significant gains from future process shrinks.

It’s very rare to see startups compete in MLPerf and SiMa is the first edge hardware startup to throw their hat in the ring on edge power efficiency. SiMa is excited to stand toe-to-toe with giants and showcase what makes the Dual M.2 special: it’s streaming performance normalized by power of 65 FPS/Watt (1000 / 15.29) in “Single Stream” and 137 FPS/Watt (8000 / 58.05) in “MultiStream”.

Not captured in the MLPerf’s result table is the ease of use of SiMa’s software Palette. These results were made possible by SiMa’s ML Model compiler and SDK which can bring any computer vision model to the power efficient chip at the push of a button.

SiMa.ai had a MLPerf Inference: Edge v3.0 Image Classification score of 2190.27 FPS and a measured power of 16.11W. Therefore, SiMa.ai achieved 136fps/W.

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

Supermicro

Supermicro has a long history of designing a wide array of products for various AI use cases. In MLPerf Inference v3.0, Supermicro has submitted three systems in the closed division and three in the open division in the data center and edge inference category.

Supermicro's mission is to provide application optimized systems for a spectrum of workloads. For example, Supermicro designs and manufactures systems for the NVIDIA HGX H100 8-GPU platform, such as the SYS-821GE-TNHR, customizable for customers' various requirements and workload needs through our building block approach. Now we also offer liquid cooling options for the latest NVIDIA HGX based-systems, as well as PCIe-based systems to help deployments use higher TDP CPUs and GPUs without thermal throttling.

Supermicro's GPU A+ Server, the AS-4125GS-TNRT has flexible GPU support and configuration options: with active & passive GPUs, and dual root or single-root configurations for up to 10 double-width, full-length GPUs. Furthermore, the dual root configuration features directly attached 8 GPUs without PLX switches to achieve the lowest latency possible and improve performance, which is hugely beneficial for demanding scenarios our customers face with machine learning (ML) and HPC workloads.

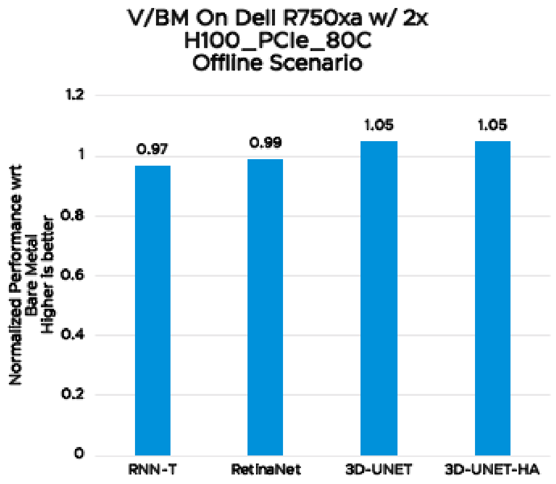
Supermicro's rack-mountable tower/workstation form factor GPU SuperServer, the SYS-741GE-TNRT accommodates up to four double-width GPUs, including the NVIDIA H100 or the A100 PCIe GPUs. This system can be used in an office environment or a data center rack. Moreover, the new SYS-751GE-TNRT, another AI development platform with Supermicro's advanced liquid cooling design, offers an unprecedented level of power and efficiency with a whisper-quiet noise level suited for offices, labs, or even to home offices.

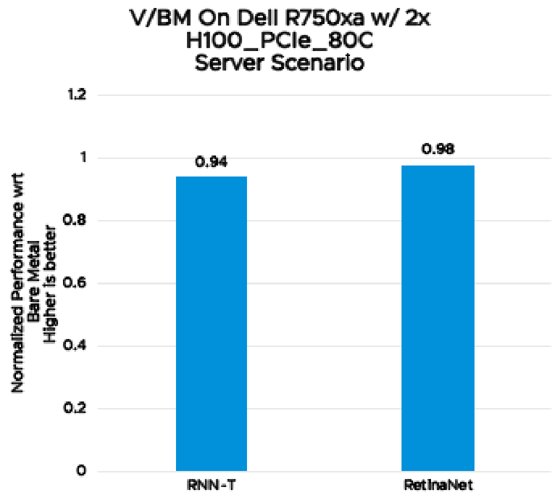
Supermicro delivers a variety of GPU systems for any environment. The results show high performance in multiple MLPerf tests. Supermicro will continue to tune the various systems to bring optimized experience and performance to customers.

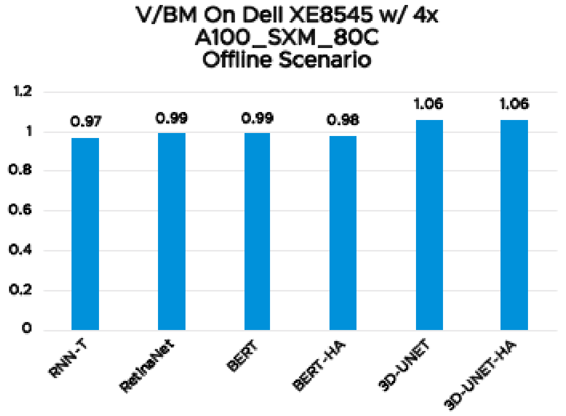
UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

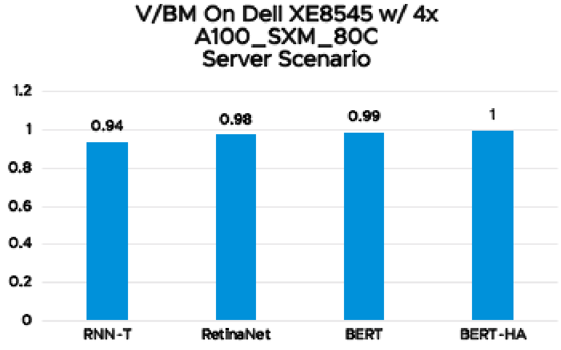
VMWare

Twenty-five years ago, VMware virtualized x86-based CPUs and has been a leader in virtualization technologies since then. VMware is again repeating its magic act in collaboration with NVIDIA and Dell in virtualizing accelerators for Machine Learning. We are announcing near bare metal or better than bare metal performance for MLPerf Inference 3.0 benchmarks. We ran MLPerf Inference v3.0 benchmarks on Dell XE8545 with 4x virtualized NVIDIA SXM A100-80GB and Dell R750xa with 2x virtualized NVIDIA H100-PCIE-80GB both with only 16 vCPUs out of 128. Now you can run ML workloads in VMware vSphere with virtualized NVIDIA GPUs and combine the power of virtualized NVIDIA GPUs with the virtualization benefits of VMware vSphere for managing data centers. The comparison of our results with bare metal is presented below.


V/BM On Dell R750xa w/ 2x H100_PCIe_80C Offline Scenario


V/BM On Dell R750xa w/ 2x H100_PCIe_80C Server Scenario


V/BM On Dell XE8545 w/ 4x A100_SXM_80C Offline Scenario


V/BM On Dell XE8545 w/ 4x A100_SXM_80C Server Scenario

UNDER EMBARGO UNTIL Wednesday April 05 at 10:00am Pacific.

XFusion

xFusion Digital Technology Co., Ltd. is committed to becoming the world's leading provider of computing power infrastructure and services. We adhere to the core values of "customer-centric, striver-oriented, long-term hard work, and win-win cooperation", continue to create value for customers and partners, and accelerate the digital transformation of the industry.

In this performance competition of MLPerf inference v3.0, we used a new generation of GPU server product FusionServer G5500 V7 to conduct performance tests on all benchmarks under various GPU configurations and achieved excellent results.

FusionServer G5500 V7 (G5500 V7) is a new-generation 4U 2-socket GPU server. It supports a maximum of 10 x double-width GPU cards. We use Intel® Xeon® Platinum 8458P CPU x2 and 8~10 A30 or L40 GPU configurations to test all evaluation items. It has made excellent achievements in Language processing and Recommendation etc, especially the server scenario. 19 test results can achieve the best performance under the same GPU hardware configuration.

FusionServer G5500 V7 features high performance, flexible architecture, high reliability, easy deployment, and simplified management. It accelerates applications such as AI training, AI inference, high-performance computing (HPC), image and video analysis, and database, and supports enterprise and public cloud deployment.