Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Supplemental Discussion

MLPerf™ Training 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 https://mlcommons.org/en/training-normal-20/. The statements **do not reflect the opinions or views of MLCommons®**.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

ASUSTeK

ASUS is proud to have enrolled in MLCommons' influential MLPerf Training 2.0 submission for 2022, allowing our products to be judged for their performance in the competitive field of AI training.

Specifically, we submitted the ASUS ESC N4A-E11 server because we consider it to be the perfect choice for diverse AI workloads in simulation and data analytics, including virtualization, medical diagnoses and voice recognition.

With ESC N4A-E11, ASUS has expertly leveraged the CPU and GPU architecture to maximize performance from powerful combination of the AMD® EPYC 7003 and NVIDIA® HGX platforms. Our optimized hardware design paired with ASUS-exclusive, firmware-based tuning technology speeds up overall efficiency and precision in AI training models. In fact, ESC N4A-E11 is designed especially for high-performance computing (HPC) fields, and has been deployed in numerous HPC projects globally this year alone.

The ESC N4A-E11 architecture, which is engineered for up to four NVIDIA HGX GPUs and a single, 64-core CPU, delivers leading results in multiple training models — including BERT, SSD, RNNT and UNET3D. These models provide a good path for deployment of training models into daily operations in enterprise or data centers.

As demonstrated by the latest MLPerf Training 2.0 benchmarks, ESC N4A-E11's cutting-edge performance and software combine to create an end-to-end solution that can be deployed data centers, in the cloud or at the edge — empowering remarkable results.

ASUS as an integrated-solutions partner delivering leading hardware for the fields of supercomputing and data centers, supported by an extensive AI portal and AI software stack. The MLPerf benchmarks help our engineers to develop an ever-better understanding of AI applications and performance needs — and we're excited to continuing delivering innovations to both shape and grow MLCommons.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Azure

Azure is pleased to share the results of our MLPerf training v2.0 submission. For this submission we used the NDm A100 v4 series virtual machines (VMs) powered by 8 NVIDIA A100 GPUs (80 GB), 8 NVIDIA 200 Gb/s HDR InfiniBand cards, 96 AMD Rome cores, 1.9 TB of RAM, and 8 * 1TB NVMe disks. The NDm A100 v4 allows for high-end AI training needs from 1 – 256+ VMs (8 – 2048+ GPUs).

These benchmark results demonstrate how Azure:

1. is committed to continuously bringing our customers the latest advancements in GPU technology
2. is in line with on-premises performance
3. is committed to democratizing AI at scale in the cloud

To generate the results, we used Azure CycleCloud to orchestrate the cluster environment of 16 VMs. In our previous submission for MLPerf training v1.1 we used it to orchestrate over 264 VMs. We used the Slurm scheduler configured with NVIDIA Pyxis and Enroot to schedule the NVIDIA NGC ML Commons containers***. This setup enabled us to deploy our environment in a timely manner and perform the benchmarks with strong performance and scalability. For more information on how to deploy this setup please see cc-slurm-ngc.

The NDm A100 v4 series VMs are what we and our Azure customers turn to when AI and ML training is required. We are excited to see what new breakthroughs our customers will make using these VMs

*** Special thanks to the NVIDIA team for all their support

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Baidu

Baidu has been working on accelerating large-scale models to benefit more industries. These real-world applications are more complex than training the MLPerf reference models, due to the huge number of parameters, heterogeneous infrastructure and other factors. With MLPerf Training 2.0, we are pleased to present the performance from PaddlePaddle on large-scale models.

PaddlePaddle has continued optimizing on X-MAN, the open computing system powered by NVIDIA A100 80GB GPUs. We submitted the BERT benchmark results using both PaddlePaddle and PyTorch, showing that BERT on PaddlePaddle ranks among the fastest frameworks tested on NVIDIA A100 80GB GPUs.

We are happy Graphcore made a submission with PaddlePaddle on IPUs with outstanding performance. As for BERT training performance on IPUs, PaddlePaddle is in-line with Graphcore's PopART framework. It shows PaddlePaddle's hardware ecosystem is expanding, and PaddlePaddle performs excellently on more and more AI accelerators.

These remarkable results come from both optimizations within PaddlePaddle and ongoing collaboration with hardware vendors. Driven by industrial practices, the distributed training architecture of PaddlePaddle fully demonstrates its key attributes of low-storage and high-performance training with heterogeneous hardware. It leads to several featured innovations: the world's first general-purpose heterogeneous parameter server architecture, the world's first 4D hybrid parallel training technology, and an end-to-end adaptive distributed training architecture. These innovations enabled the release of PCL-BAIDU Wenxin: the world's first Knowledge-Enhanced 100-billion-scale pretrained language model, and a series of big industrial models, such as State Grid-Baidu·Wenxin, SPD-Baidu·Wenxin and HELIX-Fold. Moreover, PaddlePaddle is collaborating with hardware partners to provide excellent performance and user experiences. PaddlePaddle v2.3 was released recently with support for both NVIDIA GPUs and Graphcore IPUs. As such, PaddlePaddle brings more options for training large-scale models to the community.

Baidu-PaddlePaddle looks forward to accelerating large-scale models to benefit the world with innovative technologies in distributed training and cooperation with hardware providers.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

CASIA

CASIA (Institute of Automation, Chinese Academy of Sciences) constantly focuses on researches of intelligent science and technology, which has sufficient innovation capabilities of chip R&D, system integration and algorithm design. We have rich experiences in the field of finance, transportation, medical treatment, etc. Meanwhile, our institute maintains long-term cooperative relationships with Dell EMC, H3C, NVIDIA, etc.

As a new member and first-time submitter of MLCommons, CASIA participated in 5 tasks of MLPerf Training V2.0, and successfully submitted 7 results across 3 configurations of our high-density AI server with excellent performances. We are one of the first to release the single node 16 NVIDIA A100 server named Xiangxue-3B in MLPerf developed by CASIA and Chip-hop Ltd.(www.chiphop.cn). It supports up to 16x NVIDIA A100 GPUs, dual 64 Core AMD EPYC 7773X processors, etc., and provides 1280GB HBM2E, 4TB DDR4 system memory and 1.5GB CPU cache. Xiangxue-3B has the flexibility of various PCIE topologies, which can satisfy general and extensive application needs. The results demonstrate its extraordinary performances in all of the 5 different tasks of Minigo, SSD, ResNet, Mask R-CNN and 3D U-Net, supplying referential performance data to AI users. With its high GPU expanding capabilities, users can quickly accelerate the speed of training tasks without building up complicated distributed systems. Xiangxue-3B supports various types of PCIE accelerators, and all the 16 cards' IO interfaces are available on the chassis, making it suitable for both high-density computation and IO upgrading.

CASIA is a comprehensive research institute and has top algorithm experts, software and hardware engineers. As the developer and user at the same time, we have deep understandings on the actual needs in the academic and applicationable scenarios. It is our capability to provide various types of customized software and hardware products, which can be developed efficiently and further upgraded in fast speed. We embrace more opportunities of innovative cooperation and will constantly support the development of MLCommons.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Dell Technologies

Put innovation to work with the right technology partner.

For data-driven decision making, Dell Technologies submitted 42 results across 12 system configurations. Results are available for single-node and multi-node PowerEdge XE8545, R750xa and DSS8440 servers with NVIDIA A100 SXM 40GB and A100 SXM 80GB, A100 PCIe 80GB and A30 GPUs on MLPerf training models.

- **See how AI training scales.** As AI training continues to scale with the need for speed, the Dell Technologies Innovation Lab team submitted training results with up to 32x PowerEdge XE8545 servers with 128 NVIDIA A100 SXM GPUs in the TOP500 Rattler supercomputer to show scalable performance.
- **Evaluate price/performance**. As results vary across different domains, the Innovation Lab team tested different CPUs, operating systems, NVIDIA GPU interconnects and more, so you have the data to select the right technologies for your workloads and your budget.
- **Leverage Dell engineering expertise**. Easily run multi-node AI training in your traditional HPC environment with a new Singularity script available in the community for download.

Dig into the Dell engineering test results. Test for yourself in one of our worldwide Customer Solution Centers. Collaborate with our HPC & AI Innovation Lab and/or tap into one of our HPC & AI Centers of Excellence.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Fujitsu

Fujitsu is a leading company of information and communication technology systems. We support our customer's business by providing robust and reliable ICT systems.

In this round, Fujitsu measured seven benchmark programs except for minigo with our latest server, PRIMERGY GX2570M6. The system is 4U rack mount and has an NVIDIA HGX A100 board that includes eight GPUs. As for storage, we configured RAID0 with PCIe connected NVMe SSDs to feed enough data during training. Compared with our past submissions, the performance is proportional to the number of GPUs.

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.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

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 GIGABYTE 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).

GIGABYTE submissions for MLPerf Training v2.0:

- GIGABYTE 4U server: G492-ID0
- Dual 3rd Gen Intel Xeon Scalable processors - Platinum 8380
- NVIDIA HGX A100 80GB 8-GPU system
- Frameworks: mxnet, Tensorflow, hugectr, and PyTorch

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.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Google

In MLPerf v2.0, Google's submissions demonstrated exceptional performance across all five of the benchmarks we submitted.

Four of Google's submissions were in the cloud category based on our publicly available Cloud TPU v4. This means that all of Google's ML infrastructure is available to ML practitioners and innovators around the world.

Two of our submissions this time around were at the "full TPU v4 pod" 4096 chip scale with each pod delivering 1.1 exaflop/s of peak performance. As we've noted before, each Cloud TPU v4 pod consists of 4096 chips connected together via an ultra-fast interconnect network with the equivalent of an industry-leading 6 terabits per second (Tbps) of bandwidth per host, enabling rapid training for the largest models.

Google's TPU 2.0 submissions were made on Tensorflow and are a significant improvement over our 1.0 submissions last year, reflecting the significant investments in our software stack that we have made, including optimizations made to the TPU compiler.

Finally, it is worth noting that the Cloud TPU v4 pods powering our ML perf results are located in our Oklahoma data center which operates at 90% carbon-free energy. In fact, with a PUE of 1.10, the Oklahoma data center is one of the most energy efficient data centers in the world. Additionally, the TPU v4 chip itself is highly energy efficient with 3x the peak FLOPs per watt of the prior v3 generation.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Graphcore

Since the last MLPerf Training round, Graphcore has focused on building out breadth and depth in our model garden of real-world applications driven by customers. These include emerging models, like GNNs and graph networks which are well-suited to the IPU's highly differentiated, designed-for-AI architecture, large Transformer NLP models like GPT and more accurate vision models like EfficientNet.

With MLPerf Training 2.0, we are pleased to show continued significant performance improvements for MLPerf reference models BERT and ResNet and a new submission for speech transcription, Transformer-Transducer. Other highlights include the first PaddlePaddle submission from Baidu on IPU and the introduction of our new Bow Pod systems, straight into the available on-prem and cloud category.

In March 2022, we launched a new Bow IPU processor and Bow Pod platforms, delivering up to 40% better performance and up to 16% better power efficiency. We are delighted to show results for our new Bow systems, just a few months later, from Bow Pod16 through to our large Bow Pod256 system, designed for large scale distributed training. These systems all feature in the MLPerf 'available' category as they are already shipping to customers and are available in the public cloud.

Our software continues to go from strength to strength, with additional software improvements on top of those delivered by our new Bow systems, giving an overall decrease in time-to-train of 37% for BERT and 31% for Resnet50, as validated by MLPerf.

As our software matures, our software ecosystem expands as new partners are able to integrate with our open-source APIs and libraries in the Poplar SDK. We are delighted that Baidu has made a submission with its PaddlePaddle framework using Graphcore IPUs for the first time and has received outstanding results. The submissions for BERT on Bow Pod16 and Bow Pod64, achieve near identical performance to the results submitted for Graphcore's PopART framework. This is a testament to the consistent performance that can be achieved across multiple frameworks on the Bow IPU platforms, and the efficiency in model development using PaddlePaddle.

We made a new submission for a speech transcription transformer-transducer model in this round. Working with our customer, Gridspace, we applied this 100M parameter model to accelerate training on real world data - over 700GB/10k hours of speech - to a state-of-the-art word error rate. By scaling the model to run on a Bow Pod64 training time on this dataset was reduced from weeks to days.

All software used for our submissions is available from the MLPerf repository, to allow anyone to reproduce our results. The Graphcore Github repository also covers many other new and

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

emerging models where the IPU's unique architecture can enable innovators to create the next breakthroughs in machine intelligence.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

H3C

As a digital solution leader, H3C is committed to providing intelligent software management platform, diversified smart computing system and unified artificial intelligence platform for computing power scheduling, so as to achieve easier product deployment, more reliable operation, more powerful monitoring and more convenient troubleshooting, and escort the production of AI customers.

Complex deep learning models often consume lots of computing and storage resources. Different AI workloads have different requirements for hardware bandwidth, topology, single-point computing power and the like due to their diverse capabilities in data processing, data storage and data interaction. H3C introduces H3C UniServer 5500 G5 8HGX A100 server model (medium and large scale), R5300 G5 PCIe GPU server model (medium and large scale) and 2U R4900 G5 universal server model (small scale) for different AI workload scenarios, in order to meet customers' demand for computing power resources in different scenarios.

In this benchmark test of MLPerf Training v2.0, H3C participated in the training evaluation for eight task with six types of models, including R5500 /R5300/R4900 G5, and has made excellent achievements in natural language processing, target detection, medical image processing and other fields.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Habana Labs

We're pleased to deliver Habana's third training submission results, including the recently launched second-generation Gaudi®2 deep learning training processor, our new purpose-built AI processor that further enhances Intel's AI XPU portfolio. Gaudi®2 was launched in May and shows dramatic advancements in time-to-train, which translate to 3x speed-up for the vision model (ResNet-50) and 4.7x for the language model (BERT) relative to first-gen Gaudi performance. These advances can be attributed to the transition to 7nm process from 16nm, the Gaudi2 memory subsystem which contains 96 GB of HBM2E memories delivering 2.45 TB/sec bandwidth and additional architecture advances. The quick and seamless launch of Gaudi®2 and our ability to submit Gaudi2 in this current MLPerf evaluation were enabled due to the continued maturity of our SynapseAI® software stack.

In addition, we submitted performance of 128 and 256 accelerator configurations of our first-generation Gaudi solution that demonstrate great performance and impressive linear scaling. The performance of both generations of Gaudi processors is achieved without product enhancements or special software manipulations that differ from our commercial solutions available to Habana customers out of the box. As a result, customers can expect to achieve MLPerf-comparable results in their own Gaudi systems. Both generations of Gaudi have been designed at inception to deliver exceptional AI deep learning efficiency—so we can provide customers with the excellent performance reflected in these MLPerf results, while maintaining very competitive pricing.

In addition, we are pleased that Supermicro submitted Gaudi's first OEM server results, based on the first-generation Gaudi training solution. We look forward to continuing to advance performance, efficiency, and end-user ease of use with the Gaudi platform and continue to support the MLPerf organization, our partners and our peers in leading the future of AI.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

HazyResearch

Transformer models (e.g., BERT, GPT) have emerged as the most widely used architecture in applications such as natural language processing. Transformers have grown larger and deeper, but equipping them with longer context remains difficult, since the self-attention module at their heart has time and memory complexity quadratic in sequence length. Many approximate attention methods aimed to alleviate these issues do not display wall-clock speedup against standard attention, as they focus on FLOP reduction and tend to ignore overheads from memory access (IO).

We argue that a missing principle is making attention algorithms **IO-aware**--carefully accounting for reads and writes to different levels of fast and slow memory (e.g., between fast GPU SRAM and relatively slow GPU HBM). On modern GPUs, compute speed has out-paced memory speed, and most operations in Transformers are bottlenecked by memory accesses.

We propose FlashAttention, a new algorithm that computes exact attention with far fewer HBM accesses. This requires (i) computing softmax without access to the whole input (ii) not storing large intermediate matrices for the backward. We apply two well-established techniques to address these challenges. (i) We split the input into blocks and incrementally perform the softmax (**tiling**). (ii) We store the softmax normalization from the forward to quickly **recompute** attention in the backward.

Our algorithm both **runs faster** (4x) and **uses less memory** (10x) than Pytorch standard attention, thanks to the massively reduced amount of HBM access. This brings substantial speedup and memory saving to many Transformer models (BERT, GPT2, ViT). Our BERT submission to MLPerf 2.0 is 11% faster than the best one-node result from MLPerf 1.1.

We look forward to future work on applying these general techniques to other accelerators. We're really excited to see what new capabilities will emerge from being able to use longer context in Transformers.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

HPE

HPE makes AI that is data-driven, production-oriented, and cloud-enabled, available anytime, anywhere and at any scale. We understand that successfully deploying AI workloads requires much more than hardware. That's why we deliver a full complement of offerings that enable customers to embark on their AI journey with confidence. Award-winning HPE AI Transformation Services make some of the brightest data scientists in the industry available to assist with everything from planning, building, and optimizing to implementation, and we offer continuing support through PointNext. Built upon the widely popular open source Determined.AI Training Platform, the Machine Learning Development Environment (MLDE) software and the Machine Learning Development System (MLDS) integrated hardware and software solution from HPE help developers and scientists focus on innovation by removing the complexity and cost associated with machine learning model development. Our platform accelerates time-to-production by removing the need to write infrastructure code, and makes it easy to set-up, manage, secure, and share AI computing clusters. With MLDE, customers are training models faster, building more accurate models, managing GPU costs, and tracking experiments, while HPE's Greenlake provides seamless integration of cloud and on-premises data hosting and analysis.

Today we are publishing MLPerf Training results based on the HPE Apollo 6500 Gen 10+. For these results, dual AMD EPYC 7763 processors and eight NVIDIA HGX A100 80GB GPUs delivered leading results across multiple categories, including image detection, classification, and segmentation. Mounted within our newest internal cluster to enter the Top500, Champollion, the Apollo 6500 Gen 10+ is a highly flexible system that supports either modular or PCIe GPUs from multiple vendors and dedicated workload profiles that allow users to optimize for power or throughput. As a founding member of MLCommons, HPE is committed to delivering benchmark results that provide our customers with guidance on the platforms best suited to support a variety of workloads.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

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 TrainingV2.0, Inspur made submissions on two systems: NF5488A5 and NF5688M6.

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. NF5488A5 system is capable of high temperature tolerance with operating temperature up to 40℃. It can be deployed in a wide range of data centers with 4U design, greatly helps to lower cost and increase operation efficiency.

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. It accommodates more than 10 PCIe Gen4 devices, and brings about a 1:1:1 balanced ratio of GPUs, NVMe storage and NVIDIA Mellanox InfiniBand network.

In the closed division, Inspur performance of Bert, RNNT, DLRM, Resnet, MaskRCNN and UNet3D are improved by 18.15%, 13.84%, 5.95%, 1.24%, 10.40% and 7%, compared with the best performance Inspur achieved in Training v1.1.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Krai

Comparing "entry-level" options for training neural networks is of interest to many organizations with limited resources. This is why Krai, a staunch supporter of MLPerf Inference, is making a foray into MLPerf Training in this round.

Our ResNet50 result of 284 minutes on two NVIDIA RTX A5000 GPUs consuming up to 460 Watts can be usefully compared with 236 minutes on two NVIDIA A30 GPUs consuming up to 330 Watts. Our result consumes 39% more power and is 20% slower, which translates into 67% more energy consumed. However, the A5000 is 2-3 times cheaper than the A30, making the less efficient option still attractive for small organizations. Furthermore, taking into account the power of other system components (say, 200 Watts), the total energy consumed may only be 25% higher.

Another interesting comparison is with 396 minutes on a single NVIDIA A100 PCIe 80GB GPU consuming up to 300 Watts. While our result consumes 53% more power, it is 28% faster, which translates into only 10% more energy consumed. However, a pair of A5000 GPUs is 3-4 times cheaper than a single A100 PCIe 80GB.

Seymour Cray once famously quipped: "If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens?" We at Krai are not arguing with Cray: we are not expecting to deliver training on a thousand Raspberry Pi's any time soon. Still, our motto is "horses for courses": using horses may be just fine if oxen are hard to come by (or buy), and elephants live on an entirely different continent.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

MosaicML

MosaicML, a startup on a mission to make ML training efficient for everyone, is pleased to share our MLPerf Training 2.0 results. Our results demonstrate the ability to accelerate training through software and algorithms, and the ease with which enterprise ML teams can realize more efficient training.

Our algorithmic efficiency methods achieve a 4.5x speed-up on training the Image Classification benchmark compared to our own baseline. Image Classification benchmark (open division) is trained in 23.8 minutes on 8x A100 NVidia GPUs, compared to our baseline without efficiency methods taking 110.5 minutes.

Our submission uses our open source training framework Composer, built on PyTorch, instead of heavily customized benchmark-specific code. Enterprise ML engineers can easily apply these same techniques to their own datasets. With just a few lines of code, train faster on existing hardware, or use the MosaicML Cloud, our purpose built platform for efficient ML training, for an optimized experience.

We believe that algorithmic efficiency is key to making training of large scale models more accessible to the broader community, and our submission to MLPerf is a first step in fulfilling that goal.

Details

We submitted to the Image Classification benchmark in the Open division with two configurations:

- **Baseline**: Our baseline used common hyperparameter settings from (Goya et al, 2017), and provides the standard configuration used in research. We include popular optimizations such as mixed precision and channels last.
- **Baseline+Methods**: With no changes to the existing hyperparameters, we applied a recipe of efficiency methods from our Composer library.

MosaicML believes in lowering the barrier to MLPerf benchmarks, and our software includes an MLPerf logger plug-in to automatically generate compliant submission logs. We invite the broader community to use our provided tools to further optimize the algorithms and make their own submissions to future MLPerf benchmarks.

Contact: customers@mosaicml.com

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

NVIDIA

Overall, NVIDIA continued to show great performance across every MLPerf Training 2.0 workload in this most recent round of the industry-standard benchmark for AI training.

Our NVIDIA A100 Tensor Core GPUs again delivered exceptional performance and efficiency results across all tests, as demonstrated in our own submissions as well as those by over a dozen of our partners. These include submissions by cloud provider Microsoft Azure, as well as system makers ASUSTek, Baidu, CASIA, Dell, Fujitsu, Gigabyte, H3C, Hewlett-Packard Enterprise, Inspur, Lenovo, Nettrix, and Supermicro.

Our submissions on all eight tests ran the gamut from single-node all the way up to 4,216 GPUs across 527 nodes – the largest at-scale submission seen in this round. Our platforms delivered excellent per-GPU AI training acceleration, while our at-scale submission showcased not only our multi-node GPU scalability, but also our full datacenter platform including our InfiniBand networking technology.

Software plays a large role in delivering results both in single-node and at-scale. Submissions this round brought up to 13% more performance than our same system submitted in the previous round, which highlights the ongoing benefit customers get from our continuous optimization of our software tools. Beyond MLPerf, NVIDIA accelerates many additional models, which are freely available to developers both from NGC as well as our GitHub repository.

We congratulate our 13 partners on their outstanding submissions, and commend MLCommons for their ongoing work on these industry-standard benchmarks. We look forward to upcoming rounds of MLPerf testing for AI inference, HPC and training. 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 makers towards data-driven platform investments best suited to their particular needs.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Samsung

This is Samsung's second participation in MLPerf Training with better performance than the previous round v1.1. We delivered an extremely strong performance on BERT training, 22.3 seconds on 1024 Nvidia A100 GPUs and 21.4 seconds on 1368 Nvidia A100 GPUs. This is a 12% improvement TTT (Total Time on Test) over our v1.1 performance in 1024 GPUs (25.06 seconds).

The system used for BERT training consists of 171 nodes, which have two AMD EPYC 7543 processors and eight NVIDIA Tesla A100s as accelerators, which are connected with NVLinks and have their own 80GB memory in HBM. This system is the same as the previous round v1.1, but we just expanded the scalability from 128 nodes to 171 nodes. All hardware and software components we used are available in public, so we changed our system status from research to on-premise.

Based on PyTorch NVidia Release 21.09, we changed the optimizer from LAMB to ADAM and focused on the large batch training with computation and communication overlap. We also tested BERT training with other optimizers, including the standard optimizer LAMB, but we could see ADAM optimizer showed the best performance in our approach.

Our key optimizations are:

1. Complete usage of Pytorch DDP and ADAM optimizer for large batch training with communication/computation overlap

2. Bucket-wise gradient clipping before all-reduce that combines the advantages of clipping before all-reduce and clipping after all-reduce

3. Efficient load balancing of input data for increasing GPU utilization.

In addition to AI acceleration in mobile devices, Samsung is actively researching on the scalable and sustainable AI computing. We will work to solve the scaling challenge between computing capability and memory/storage bandwidth through innovation in memory and storage computings such as HBM-PIM, CXL-Memory, and CXL-SSD.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines

Supermicro

Supermicro has its long history of providing a broad portfolio of AI-enabled products for different use cases. In MLPerf Training v2.0, we have submitted results based on four high performance systems to address multiple compute intensive use cases, including medical image segmentation, general object detection, recommendation systems, and natural language processing.

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 8 GPU platform and HGX A100 4 GPU 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. If customers are looking for rack scale design to cluster systems for large machine learning training problems, we can offer rack integration in air cooled solution, RDHx and DLC liquid cooling solution to suit your plug and play need.

Supermicro's SYS-420GP-TNAR, AS-4124GO-NART, AS-2124GQ-NART and upcoming SYS-220GQ-TNAR with NVIDIA's HGX A100 GPUs can pass data directly from GPU to GPU, to avoid the pass-through overhead from processors and system memory. By shortening the data path to the accelerator, it shortens the training time for applications such as computer vision and recommendation system.

In addition, Supermicro keeps tuning the current platforms' performance and aims to extend the benchmark efforts to large portfolios, including Habana, AMD, OpenVino, and Arctic Sounds. MLPerf Training v2.0 is the first time Supermicro submitted Image Classification on SYS-420GH-TNGR with Habana first generation GAUDI accelerators. With Supermicro 400Gbps switches, the MLPerf Training benchmark results show excellent performance and superior scalability to meet the challenges of distributed AI training on multiple nodes.

With multiple configurations of processors, accelerators, system form factors, cooling solutions, and scale out options, Supermicro would like to provide our customers the most comprehensive and convenient solutions to solve the AI problems. 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.