Supplemental Discussion

MLPerf™ Training v2.1 Results Discussion

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

ASUSTeK

ASUSTeK is pleased to announce to keep working with MLCommons on AI training and inference benchmark to improve AI workloads deployed into daily operation. Along with MLPerf™ Training v2.1 benchmark, ASUS disclosed AI training results with ESC8000A-E11 and ESC N4A-E11. In MLPerf™ Training v2.1, ESC8000A-E11 submitted 7 results including BERT, Mask R-CNN, Mini Go, ResNet, SSD, RNN-T, Unet-3D benchmarks and ESC N4A-E11, submitted BERT, DLRM, Mask R-CNN, ResNet, SSD, RNN-T, Unet-3D. Compared to MLPerf™ Training v2.0, ASUS GPU servers got better results which showed algorithm improvement and ASUS server tuning technology in performance.

In MLPerf™ Training v2.1, ASUS kept the same configuration as per MLPerf™ Training v2.0 with NVIDIA A100 PCIE GPU and SXM technologies separately. Besides algorithm updates, ASUS also keeps improving GPU server performance including BIOS settings and overall hardware performance adjustment.

ESC8000A-E11 is ASUS flagship model in PCIE GPU configuration, which can support up to 8 PCIE GPU like NVIDIA A100 GPU. ESC8000A-E11 is fit to apply heavy AI training workloads. Combining with AMD EPYC 3rd Gen processors, ESC8000A-E11 can deliver abundant computing resources no matter in data center and enterprise segments. ESC N4A-E11, powered by NVIDIA SXM technology, is capable of delivering better GPU computing connections through NVLink technologies.

ASUSTeK expects to work with MLCommons in AI fields and contribute more insights into algorithm developments in the coming future.

Azure

Azure is pleased to share results from our MLPerf™ training v2.1 submission. For this submission, we benchmarked our NC A100 v4-series and NDm A100 v4-series offerings, which are our flagship virtual machine (VM) types for running mid-end and high-end AI training workloads, respectively. NC A100 v4-series features up to 4 NVIDIA A100 PCIe GPUs with 80GB memory each, 96 3rd generation AMD EPYC Milan processor cores and 880 GiB of system memory. NDm A100 v4 VMs are powered by 8 NVIDIA A100 SXM 80 GB GPUs (NVLink 3.0) and 96 physical 2nd generation AMD Epyc™ 7V12 CPU cores. These offerings enable our customers to address their AI training needs at scale.

Some of the highlights from our MLPerf™ training v2.1 benchmark results are:

1. Azure is the only cloud provider that demonstrated distributed training at scale with up to 128 Nvidia A100 GPUs.

2. Azure is able to achieve 5.3% improvement in training time from the previous MLPerf training round scores on an 8GPU system for the Mask R-CNN benchmark.

These training results demonstrate how Azure:

- is committed to providing our customers with the latest GPU offerings

- is in line with on-premises performance

- is committed to enabling our customers to run AI at scale in the cloud

Special thanks to our hardware partner NVIDIA for providing the containers from the NVIDIA NGC catalog that enabled us to run these benchmarks. We deployed our environment using the Azure Cyclecloud 8.2, cyclecloud-slurm 2.6.5 scheduler configured with NVIDIA Pyxis and Enroot, and Azure HPC Ubuntu 18.04 and 20.04 marketplace images.

The NC A100 v4-series and NDm A100 v4-series systems are what we and our Azure customers turn to when large-scale AI and ML training is required. We are excited to see what new breakthroughs our customers will make using our offerings.

Azure HazyResearch

The latest MLPerf™ Training v2.1 submission demonstrates the performance gains that can be achieved on Azure by leveraging HazyResearch*** cutting-edge software optimization for BERT and NVIDIA accelerated computing. This collaborative submission demonstrates the potential for optimized heavy workloads combined with Azure's innovative infrastructure.

On the hardware side, we benchmarked 1, 8, and 16 virtual machines of the NDm A100 v4-series featuring NVIDIA A100 Tensor Core GPUs on Azure. The cluster used for this effort was powered by NVIDIA A100 80GB Tensor Core SXM GPUs and used the latest versions of the software stack (Ubuntu 20.04-HPC marketplace image and PyTorch NVIDIA release 22.09) and resources (Cycle Cloud 8.2 and slurm 2.6.5). The NDm A100 v4-series instance is what we and our Azure customers turn to for our large-scale AI and ML training workloads.

On the software side, our submissions benefit from algorithmic improvement to the self-attention module at the heart of the transformers. Existing implementations of self-attention tend to be slow and memory-hungry on long sequences. Instead, our FlashAttention leverages tiling and recomputation techniques to reduce the GPU memory reads/writes of attention, without any approximation. Our MLPerf™ benchmark results demonstrate that FlashAttention can yield speedup even in multi-node settings (such as 16 nodes with 128 GPUs). FlashAttention has since been adopted by many research labs and organizations to speed up the training and inference of large language models and image-generative models.

The main highlight from our submission is that Azure-HazyResearch is **achieving a training time below the 2-minute mark** with BERT on 16 virtual machines.

These training benchmark results demonstrate Azure's commitment to **providing our customers with the most efficient and scalable** offerings — that are **available on demand in the cloud** — to allow them to **exceed the on-premises performances** for their AI workloads.

Special thanks to NVIDIA for providing the guidance and containers to run these benchmarks.

**** HazyResearch is represented by Tri Dao for the work on MLPerf™. HazyResearch is a research group from Stanford University led by Professor Chris Ré.*

Baidu

Baidu has been continuously working on enabling large-scale models to benefit more communities. For MLPerf™ Training v2.0, we are glad to share the performance from PaddlePaddle on Transformer NLP model (i.e., BERT).

For this round, we made the submission with PaddlePaddle on NVIDIA GPUs. We used 64 GPUs in v2.1 submission, which is different from that of 8 GPUs in the previous round v2.0. The exceptional performance, once again, showed that PaddlePaddle ranks among the fastest frameworks tested on NVIDIA GPUs.

The continuous leading results come from the significant investments in the PaddlePaddle framework. The distributed training architecture of PaddlePaddle is dedicated to optimizations in hybrid parallel strategies, data loading techniques and work balancing methodologies.

Particularly for the v2.1 submission, we made several optimizations. First, we optimized the group sharded parallel strategy and integrated the SHARP protocol to reach high scaling efficiency in distributed training. Second, in order to resolve the data unbalance issues between different workers, we proposed a hybrid data exchange solution to combine NCCL and MPI communication together. In addition, we integrated the CUDA software stack like cuBLASLt library to improve the operator speed. All these optimizations lead to 15.8% faster than the best 8-node 64-GPU result from MLPerf™ Training v2.0.

It is also worth noting that Baidu has been making submissions with Transformer NLP models since v2.0 submissions. Transformers are playing an important role in large-scale models, with which Baidu has made many efforts. Baidu has released a series of Wenxin large-scale models, which have been widely deployed in manufacturing, energy, finance, communication, media, education and other industries.

Baidu will continue optimizing the performance of large-scale models. We look forward to deploying large-scale models in more industries.

Dell

Put innovation to work with the right technology partner.

For the MLPerf™ training v2.1 benchmark testing, Dell Technologies submitted 46 results across 14 system configurations. Results are available for single-node and multi-node PowerEdge XE8545 and R750xa servers with NVIDIA A100-SXM-40GB, NVIDIA A100-SXM4 80GB, A100-PCIe-80GB, and A30 GPUs on data-driven decision-making training models. This testing demonstrated improved performance for Bert and MaskRCNN results on multiple R750xa/XE8545 based configurations vs previous run v2.0 on same system.

- **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 Rattler supercomputer at the Dell Technologies Edge Innovation Center 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 transport results to your AI or HPC environment with scripts. Dell Technologies engineers created a script for Singularity available for download.

Dig into the 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.

Fujitsu

Fujitsu offers a fantastic blend of systems, solutions, and expertise to guarantee maximum productivity, efficiency, and flexibility delivering confidence and reliability. We have continued to participate in and submit to every inference and training round since 2020.

In this training round, Fujitsu measured benchmark programs for the closed division with PRIMERGY RX2540 M6 in addition to PRIMERGY GX2570 M6. The details of these systems are shown as follows:

1. PRIMERGY GX2570 M6 is a 4U rack-mount server with Intel (R) Xeon (R) Platinum 8352V CPUx2 and NVIDIA A100 SXM 80GB x8. This server is a performance-oriented server in PRIMERGY for high-grade AI, Data Science and HPC workloads.

2. PRIMERGY RX2540 M6 is a 2U rack-mount server with Intel(R) Xeon(R) Platinum 8352Y CPUx2. In this benchmark, two types of accelerators are measured: NVIDIA A100 PCIe 80GB x2 and NVIDIA A30 x2. This server forms the standard in every modern data center and enables the running of nearly every workload from the most basic to business-critical applications.

Regarding GX2570 M6, we were able to reduce the training time to the same or less than our previous submission by improving system settings.

Our purpose is to make the world more sustainable by building trust in society through innovation. We have a long heritage of bringing innovation and expertise, continuously working to contribute to the growth of society and our customers.

Therefore, we will continue to meet the demands of our customers and strive to provide attractive server systems through the activities of MLCommons.

GIGABYTE

GIGABYTE is an industry leader in high-performance servers, and uses hardware expertise, patented innovations, and industry leadership to create, inspire, and advance. With over 30 years of motherboard manufacturing excellence and 20 years of server and enterprise products, GIGABYTE offers an extensive portfolio of enterprise products.

Over the years, GIGABYTE has submitted benchmark results for both Training and Inference. As well, the submitted servers have been equipped with various brands of accelerators (NVIDIA and Qulacomm) and processors (AMD, Ampere, and Intel) in configurations to showcase compute performance on different platforms (x86 and Arm).

For MLPerf Training v2.1, in closed division, all tasks were run in a GIGABYTE 4U server, G492–ZD2, that is NVIDIA-Certifed and prepared with hardware: dual AMD EPYC 7713 CPUs and NVIDIA A100 SXM GPU (Delta).

Frameworks: PyTorch, Tensorflow, hugectr, and mxnet

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.

Habana Labs

We’re pleased to deliver the second set of results for the Habana® Gaudi®2 second-generation deep learning processor, a purpose-built deep learning processor in Intel’s AI XPU portfolio. Gaudi2 was launched in May and is selling through Supermicro this quarter and also Inspur in the first quarter. In this MLPerf™ Gaudi2 results were also submitted in the PyTorch framework, as well! Habana continues to expand its software stack, submitting language (BERT) and vision (ResNet-50) model results in both TensorFlow and PyTorch. Both frameworks were submitted under the closed division and in the available category.

Compared with the previous Gaudi2 submission, both TensorFlow models achieved a 10% time-to-train reduction and our first PyTorch results are 5% better than the June MLPerf™ 2.0 TensorFlow results. The fast progress of Gaudi®2 is enabled due to the increasing maturity of our SynapseAI® software stack and its unique architecture. The Gaudi®2 processor, launched in May 2022, is implemented in a 7nm process, with a Memory subsystem that includes 96 GB of HBM2E memories, delivering 2.45 TB/sec bandwidth per Gaudi2. Gaudi®2 based servers provide customers with excellent performance as the 8-processor MLPerf™ results reflect. We look forward to continuing to advance performance, benchmark model coverage, and end-user ease of use with the Gaudi platform. We are pleased to continue to support the MLPerf™ organization and are committed to progressing means by which customers can reliably compare Habana solutions against our peers to further advance AI applications.

We embrace the MLPerf competition and look forward to the next submission!

Topology	**# of Gaudi2s**	**V2.0 Submission [min]**	**V2.1 Submission [min]**
ResNet-50 TF	8	18.35	16.60
BERT-Large TF	8	17.2	15.56
ResNet50 PyTorch	8	-	17.23
BERT-Large PyTorch	8	-	16.45

HPE

HPE makes AI/ML solutions that are data-driven, production-oriented, and cloud-enabled, available anytime, anywhere and at any scale. We understand that successfully deploying AI/ML models 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 HPE Pointnext and HPE's Greenlake services.

Built upon the widely popular open source Determined.AI Training Platform, the HPE Machine Learning Development Environment software and the HPE Machine Learning Development System 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. With HPE Machine Learning Development Environment, customers are training models faster, building more accurate models, managing GPU costs, and tracking experiments.

HPE's MLCommons Training v2.1 results are based on the HPE Apollo 6500 Gen 10 Plus with support for up to 10 double-wide (16 single-wide) PCIe GPUs as part of our newest internal cluster to enter the Top500^{TM} List, Champollion. HPE Apollo systems also support dedicated workload profiles, allowing users to optimize for power or throughput. Since Training 2.0, we have upgraded our storage to HPE's Parallel File System Storage (PFSS), a hybrid solid-state and hard drive storage system. With these newest results, Champollion is demonstrating exceptional multi-node scaling (with 8x NVIDIA HGX A100-SXM-80GB accelerators on each node) across multiple AI workloads, from natural language processing to computer vision and recommendation. 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 AI/ML workloads.

Intel

Intel submitted MLPerf™ Training v2.1 results on the 4th Gen Intel® Xeon® Scalable processor product line (codenamed Sapphire Rapids) across a range of workloads, once again demonstrating it is the best general-purpose CPU for AI training, which enables customers to use their shared infrastructure to train anywhere, anytime. The 4th Gen Intel® Xeon® Scalable processors with Intel® Advanced Matrix Extensions (Intel® AMX) deliver significant out-of-box performance improvements that span multiple frameworks, along with support for end-to-end data science tools and a broad ecosystem of smart solutions. This new built-in AI accelerator engine sits on every core and delivers 8x operations per clock compared to the previous generation.

But this cycle's results aren't just about the CPU, as they bring together the entire Intel platform portfolio – and not just AI – to optimize the scaling efficiency with Intel Ethernet Network Adapters, Intel® Tofino™ based switches, Intel® Optane SSDs, and Intel® Silicon Photonics.

Training on Intel Xeon Scalable processors lets enterprises avoid the cost and complexity of introducing special-purpose hardware for AI training, and benefit from being able to maintain a common data and MLOps pipeline by performing data preparation, model training, and inference on the same familiar Xeon Scalable technology.

Intel's results show that 4th Gen Intel® Xeon® Scalable processors are expanding the reach of general-purpose CPUs for AI training, so customers can do more with the Xeons that are already running their business. This is especially true for training medium to small models or transfer learning (aka fine tuning). The DLRM results are great examples of where we were able to train the model in less than 30 mins (26.73 mins) with only 4 nodes! Even for mid- sized and larger models, 4th Gen Xeons could train BERT and ResNet-50 models in less than 50 mins (47.26 mins) and less than 90 mins (89.07 mins), respectively. Speaking in relative terms, that means you can train small models over a coffee break and mid-sized models over lunch!

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.

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 TrainingV2.1, Inspur made submissions on three models with system: NF5468M6J.

NF5468M6J, submitted in the RDI category, supports up to 24*A40 GPUs and can be widely used in Internet AI public cloud, enterprise-level AI cloud platform, smart security, video codec, etc. It 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.

In the closed division, the Inspur's single node performance of Resnet and MaskRCNN are improved by 9.1% and 13.13%, compared with the best performance Inspur achieved in Training v2.0. Inspur submitted the SSD (RetinaNet) model for the first time and got good grades.

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 another foray into MLPerf™ Training. Both for ML Inference and Training, we partner with hardware designers, OEMs, and end-users to solve the real-world pains of ML/software/hardware benchmarking, optimization and deployment.

Our submissions uniquely use NVIDIA RTX A5000 GPUs. While A100 and A30 GPUs are more performant and efficient for ML Training, A5000 GPUs are more affordable for small organizations.

Another interesting point to note is that adopting a more recent software release does not necessarily bring performance improvements. Comparing our ResNet50 submissions, we observe that the NVIDIA NGC MxNet Release 22.04 (based on CUDA 11.6) container enabled 12% faster training than the Release 22.08 (based on CUDA 11.7) on the dual A5000 system we tested.

Seymour Cray once famously quipped: "If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens?" While we at KRAI are not expecting to deliver training on a thousand Raspberry Pi's any time soon, our motto is "horses for courses": using heavy horses may be just fine if oxen are hard to come by.

MosaicML

MosaicML, a startup on a mission to make ML training efficient for everyone, is pleased to share our MLPerf™ Training 2.1 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 system and algorithmic optimizations achieve a 2.7x speed-up on training the Natural Language Processing benchmark compared to a Hugging Face baseline with the same hyperparameters. The Natural Language Processing benchmark (open division) is trained in 7.9 minutes on 8x A100 NVIDIA GPUs, compared to 21.4 minutes for the baseline.

Our submission uses Composer, our open source training framework 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. They can train faster on their own hardware, or have an optimized experience on MosaicML Cloud, our purpose-built platform for efficient ML training.

Details

Our submission to the Natural Language Processing benchmark in the Open division includes two configurations:

- **Baseline**: We provide a strong baseline by training a widely used BERT model available from the popular Hugging Face repository. We use mixed precision training for the baseline.
- **Baseline+Methods**: With no changes to the existing hyperparameters, we apply a recipe of efficiency methods from our Composer library to a slightly modified model, as well as some private efficiency methods.

This NLP submission improves performance by combining algorithmic and system optimizations, the same approach our previous submission used to achieve a 4.5x speed-up on image classification. We invite the community to use Composer, which includes a plug-in to generate MLPerf™-compliant submission logs automatically, to further optimize the algorithms and make their own submissions to future MLPerf™ benchmarks.

Contact: customers@mosaicml.com

NVIDIA

We are excited to make our debut H100 submission for MLPerf™ Training 2.1 in the preview category, demonstrating up to 6.7X higher training performance than our A100 GPU when it was first submitted in the available category. NVIDIA H100 is based on the groundbreaking NVIDIA Hopper Architecture and supercharges the NVIDIA AI platform for advanced models, empowering customers with new levels of performance and capabilities for the most demanding AI and HPC workloads. Hopper features Transformer Engine, which applies per-layer intelligence to the use of FP8 precision, delivering optimal performance for both AI training and inference workloads while preserving model accuracy.

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

We are excited to see our 10 NVIDIA partners submit great training results on A100-based systems across all tests, both for on-prem as well as cloud platforms. A100 has seen up to 2.5x more performance from software improvements over time, and continues to deliver excellent training performance across the full suite range of MLPerf™ tests, spanning image, speech, reinforcement learning, natural language and recommender systems.

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

Samsung

This is Samsung's third participation in MLPerf Training with high performance. We delivered an extremely strong performance on BERT training in open division, 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.

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 benchmark test of MLPerf™ Training v2.1, xFusion participated in the training evaluation for four task with three types of models for the closed division with FusionServer G5500 V6. The details of the system is shown as follows:

FusionServer G5500 V6 is a 4U 2-socket GPU server with Intel (R) Xeon (R) Platinum 8380 CPU x2 and NVIDIA A30 x8. Compared with the models with the same configuration, it has made excellent achievements in natural language processing, Recom-mendation, Image classification and Speech recognition etc.

FusionServer G5500 V6 is suitable for many scenarios such as AI inference, AI training, HPC, video analysis and database acceleration by virtue of excellent computing performance, flexible expansion, and easy operation and maintenance.It is optimized for business types such as HPC and supports both enterprise and public cloud deployments.