NVIDIA H100 GPUs Set Standard for Generative AI in Debut MLPerf Benchmark

In a new industry-standard benchmark, a cluster of 3,584 H100 GPUs at cloud service provider CoreWeave completed a massive GPT-3-based benchmark in just 11 minutes.

Author: Dave Salvator

Leading users and industry-standard benchmarks agree: NVIDIA H100 Tensor Core GPUs deliver the best AI performance, especially on the large language models (LLMs) powering generative AI .

H100 GPUs set new records on all eight tests in the latest MLPerf training benchmarks released today, excelling on a new MLPerf test for generative AI. That excellence is delivered both per-accelerator and at-scale in massive servers.

For example, on a commercially available cluster of 3,584 H100 GPUs co-developed by startup Inflection AI and operated by CoreWeave , a cloud service provider specializing in GPU-accelerated workloads, the system completed the massive GPT-3-based training benchmark in less than eleven minutes.

“Our customers are building state-of-the-art generative AI and LLMs at scale today, thanks to our thousands of H100 GPUs on fast, low-latency InfiniBand networks,” said Brian Venturo, co-founder and CTO of CoreWeave. “Our joint MLPerf submission with NVIDIA clearly demonstrates the great performance our customers enjoy.”

Inflection AI harnessed that performance to build the advanced LLM behind its first personal AI, Pi , which stands for personal intelligence . The company will act as an AI studio, creating personal AIs users can interact with in simple, natural ways.

“Anyone can experience the power of a personal AI today based on our state-of-the-art large language model that was trained on CoreWeave’s powerful network of H100 GPUs,” said Mustafa Suleyman, CEO of Inflection AI.

Co-founded in early 2022 by Mustafa and Karén Simonyan of DeepMind and Reid Hoffman, Inflection AI aims to work with CoreWeave to build one of the largest computing clusters in the world using NVIDIA GPUs.

These user experiences reflect the performance demonstrated in the MLPerf benchmarks announced today .

H100 GPUs delivered the highest performance on every benchmark, including large language models, recommenders, computer vision, medical imaging and speech recognition. They were the only chips to run all eight tests, demonstrating the versatility of the NVIDIA AI platform.

Training is typically a job run at scale by many GPUs working in tandem. On every MLPerf test, H100 GPUs set new at-scale performance records for AI training.

Optimizations across the full technology stack enabled near linear performance scaling on the demanding LLM test as submissions scaled from hundreds to thousands of H100 GPUs.

In addition, CoreWeave delivered from the cloud similar performance to what NVIDIA achieved from an AI supercomputer running in a local data center. That’s a testament to the low-latency networking of the NVIDIA Quantum-2 InfiniBand networking CoreWeave uses.

In this round, MLPerf also updated its benchmark for recommendation systems.

The new test uses a larger data set and a more modern AI model to better reflect the challenges cloud service providers face. NVIDIA was the only company to submit results on the enhanced benchmark.

Nearly a dozen companies submitted results on the NVIDIA platform in this round. Their work shows NVIDIA AI is backed by the industry's broadest ecosystem in machine learning.

Submissions came from major system makers that include ASUS, Dell Technologies, GIGABYTE, Lenovo, and QCT. More than 30 submissions ran on H100 GPUs.

This level of participation lets users know they can get great performance with NVIDIA AI both in the cloud and in servers running in their own data centers.

NVIDIA ecosystem partners participate in MLPerf because they know it's a valuable tool for customers evaluating AI platforms and vendors.

The benchmarks cover workloads users care about — computer vision, translation and reinforcement learning, in addition to generative AI and recommendation systems .

Users can rely on MLPerf results to make informed buying decisions, because the tests are transparent and objective. The benchmarks enjoy backing from a broad group that includes Arm, Baidu, Facebook AI, Google, Harvard, Intel, Microsoft, Stanford and the University of Toronto.

MLPerf results are available today on H100, L4 and NVIDIA Jetson platforms across AI training, inference and HPC benchmarks. We'll be making submissions on NVIDIA Grace Hopper systems in future MLPerf rounds as well.

As AI's performance requirements grow, it's essential to expand the efficiency of how that performance is achieved. That's what accelerated computing does.

Data centers accelerated with NVIDIA GPUs use fewer server nodes, so they use less rack space and energy. In addition, accelerated networking boosts efficiency and performance, and ongoing software optimizations bring x-factor gains on the same hardware.

Energy-efficient performance is good for the planet and business, too. Increased performance can speed time to market and let organizations build more advanced applications.

Energy efficiency also reduces costs because data centers accelerated with NVIDIA GPUs use fewer server nodes. Indeed, NVIDIA powers 22 of the top 30 supercomputers on the latest Green500 list .

NVIDIA AI Enterprise , the software layer of the NVIDIA AI platform, enables optimized performance on leading accelerated computing infrastructure. The software comes with the enterprise-grade support, security and reliability required to run AI in the corporate data center.

All the software used for these tests is available from the MLPerf repository, so virtually anyone can get these world-class results.

Optimizations are continuously folded into containers available on NGC , NVIDIA's catalog for GPU-accelerated software.

Read this technical blog for a deeper dive into the optimizations fueling NVIDIA's MLPerf performance and efficiency.

Original URL: https://blogs.nvidia.com/blog/2023/06/27/generative-ai-debut-mlperf/