NVIDIA Hopper Sweeps AI Inference Benchmarks in MLPerf Debut

In industry-standard tests of AI inference, NVIDIA H100 GPUs set world records, A100 GPUs showed leadership in mainstream performance and Jetson AGX Orin led in edge computing.

Author: Dave Salvator

In their debut on the MLPerf industry-standard AI benchmarks, NVIDIA H100 Tensor Core GPUs set world records in inference on all workloads, delivering up to 4.5x more performance than previous-generation GPUs.

The results demonstrate that Hopper is the premium choice for users who demand utmost performance on advanced AI models.

Additionally, NVIDIA A100 Tensor Core GPUs and the NVIDIA Jetson AGX Orin module for AI-powered robotics continued to deliver overall leadership inference performance across all MLPerf tests: image and speech recognition, natural language processing and recommender systems.

The H100, aka Hopper, raised the bar in per-accelerator performance across all six neural networks in the round. It demonstrated leadership in both throughput and speed in separate server and offline scenarios.

The NVIDIA Hopper architecture delivered up to 4.5x more performance than NVIDIA Ampere architecture GPUs, which continue to provide overall leadership in MLPerf results.

Thanks in part to its Transformer Engine , Hopper excelled on the popular BERT model for natural language processing. It’s among the largest and most performance-hungry of the MLPerf AI models.

These inference benchmarks mark the first public demonstration of H100 GPUs, which will be available later this year. The H100 GPUs will participate in future MLPerf rounds for training.

NVIDIA A100 GPUs, available today from major cloud service providers and systems manufacturers, continued to show overall leadership in mainstream performance on AI inference in the latest tests.

A100 GPUs won more tests than any submission in data center and edge computing categories and scenarios. In June , the A100 also delivered overall leadership in MLPerf training benchmarks, demonstrating its abilities across the AI workflow.

Since their July 2020 debut on MLPerf, A100 GPUs have advanced their performance by 6x , thanks to continuous improvements in NVIDIA AI software.

NVIDIA AI is the only platform to run all MLPerf inference workloads and scenarios in data center and edge computing.

The ability of NVIDIA GPUs to deliver leadership performance on all major AI models makes users the real winners. Their real-world applications typically employ many neural networks of different kinds.

For example, an AI application may need to understand a user’s spoken request, classify an image, make a recommendation and then deliver a response as a spoken message in a human-sounding voice. Each step requires a different type of AI model.

The MLPerf benchmarks cover these and other popular AI workloads and scenarios — computer vision, natural language processing, recommendation systems, speech recognition and more. The tests ensure users will get performance that’s dependable and flexible to deploy.

Users 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 Amazon, Arm, Baidu, Google, Harvard, Intel, Meta, Microsoft, Stanford and the University of Toronto.

In edge computing, NVIDIA Orin ran every MLPerf benchmark, winning more tests than any other low-power system-on-a-chip. And it showed up to a 50% gain in energy efficiency compared to its debut on MLPerf in April.

In the previous round, Orin ran up to 5x faster than the prior-generation Jetson AGX Xavier module, while delivering an average of 2x better energy efficiency.

Orin integrates into a single chip an NVIDIA Ampere architecture GPU and a cluster of powerful Arm CPU cores. It's available today in the NVIDIA Jetson AGX Orin developer kit and production modules for robotics and autonomous systems, and supports the full NVIDIA AI software stack, including platforms for autonomous vehicles (NVIDIA Hyperion), medical devices (Clara Holoscan) and robotics (Isaac).

The MLPerf results show NVIDIA AI is backed by the industry's broadest ecosystem in machine learning.

More than 70 submissions in this round ran on the NVIDIA platform. For example, Microsoft Azure submitted results running NVIDIA AI on its cloud services.

In addition, 19 NVIDIA-Certified Systems appeared in this round from 10 systems makers, including ASUS, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro.

Their work shows users can get great performance with NVIDIA AI both in the cloud and in servers running in their own data centers.

NVIDIA partners participate in MLPerf because they know it's a valuable tool for customers evaluating AI platforms and vendors. Results in the latest round demonstrate that the performance they deliver to users today will grow with the NVIDIA platform.

All the software used for these tests is available from the MLPerf repository, so anyone can get these world-class results. Optimizations are continuously folded into containers available on NGC , NVIDIA's catalog for GPU-accelerated software. That's where you'll also find NVIDIA TensorRT , used by every submission in this round to optimize AI inference.

Read our Technical Blog for a deeper dive into the technology fueling NVIDIA's MLPerf performance .

Original URL: https://blogs.nvidia.com/blog/2022/09/08/hopper-mlperf-inference/