NVIDIA Wins New AI Inference Benchmarks

NVIDIA Turing GPUs and NVIDIA Xavier Achieve Fastest Results on MLPerf Benchmarks Measuring Data Center and Edge AI Inference Performance

NVIDIA today posted the fastest results on new benchmarks measuring the performance of AI inference workloads in data centers and at the edge -- building on the company's equally strong position in recent benchmarks measuring AI training.

The results of the industry's first independent suite of AI benchmarks for inference, called MLPerf Inference 0.5, demonstrate the performance of NVIDIA Turing™ GPUs for data centers and NVIDIA Xavier™ system-on-a-chip for edge computing.

MLPerf's five inference benchmarks -- applied across a range of form factors and four inferencing scenarios -- cover such established AI applications as image classification, object detection and translation.

NVIDIA topped all five benchmarks for both data center-focused scenarios (server and offline), with Turing GPUs providing the highest performance per processor among commercially available entries[1]. Xavier provided the highest performance among commercially available edge and mobile SoCs under both edge-focused scenarios (single-stream and multi-stream)[2].

“AI is at a tipping point as it moves swiftly from research to large-scale deployment for real applications," said Ian Buck, general manager and vice president of Accelerated Computing at NVIDIA. “AI inference is a tremendous computational challenge. Combining the industry's most advanced programmable accelerator, the CUDA-X suite of AI algorithms and our deep expertise in AI computing, NVIDIA can help data centers deploy their large and growing body of complex AI models."

Watch a video of Buck discussing the MLPerf inference benchmarks: https://youtu.be/G3nsTSPY4LI

Highlighting the programmability and performance of its computing platform across diverse AI workloads, NVIDIA was the only AI platform company to submit results across all five MLPerf benchmarks. In July, NVIDIA won multiple MLPerf 0.6 benchmark results for AI training, setting eight records in training performance.

NVIDIA GPUs accelerate large-scale inference workloads in the world's largest cloud infrastructures, including Alibaba Cloud, AWS, Google Cloud Platform, Microsoft Azure and Tencent. AI is now moving to the edge at the point of action and data creation. World-leading businesses and organizations, including Walmart and Procter & Gamble, are using NVIDIA's EGX edge computing platform and AI inference capabilities to run sophisticated AI workloads at the edge.

All of NVIDIA's MLPerf results were achieved using NVIDIA TensorRT™ 6 high-performance deep learning inference software that optimizes and deploys AI applications easily in production from the data center to the edge. New TensorRT optimizations are also available as open source in the GitHub repository.

New Jetson Xavier NX
Expanding its inference platform, NVIDIA today introduced Jetson Xavier NX, the world's smallest, most powerful AI supercomputer for robotic and embedded computing devices at the edge. Jetson Xavier NX is built around a low-power version of the Xavier SoC used in the MLPerf Inference 0.5 benchmarks.

1. MLPerf v0.5 Inference results retrieved from www.mlperf.org on Nov. 6, 2019, from entries Inf-0.5-15, Inf-0. 5-16, Inf-0.5-19, Inf-0.5-21. Inf-0.5-22, Inf-0.5-23, Inf-0.5-27. Per-processor performance is calculated by dividing the primary metric of total performance by number of accelerators reported.
2. MLPerf v0.5 Inference results retrieved from www.mlperf.org on Nov. 6, 2019, from entries Inf-0.5-24, Inf-0.5-28, Inf-0.5-29.

About NVIDIA
NVIDIA’s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to the benefits, impact, and performance of NVIDIA Turing GPUs for data centers, the NVIDIA Xavier system-on-a-chip for edge, NVIDIA’s TensorRT 6, and Jetson Xavier NX; and NVIDIA’s ability to help data centers deploy complex AI models are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Media Contacts

Kristin Bryson
+1-203-241-9190
kbryson@nvidia.com