NVIDIA Smashes Performance Records on AI Inference

NVIDIA Extends Lead on MLPerf Benchmark with A100 Delivering up to 237x Faster AI Inference Than CPUs, Enabling Businesses to Move AI from Research to Production

NVIDIA today announced its AI computing platform has again smashed performance records in the latest round of MLPerf, extending its lead on the industry's only independent benchmark measuring AI performance of hardware, software and services.

NVIDIA won every test across all six application areas for data center and edge computing systems in the second version of MLPerf Inference. The tests expand beyond the original two for computer vision to include four covering the fastest-growing areas in AI: recommendation systems, natural language understanding, speech recognition and medical imaging.

Organizations across a wide range of industries are already tapping into the NVIDIA® A100 Tensor Core GPU's exceptional inference performance to take AI from their research groups into daily operations. Financial institutions are using conversational AI to answer customer questions faster; retailers are using AI to keep shelves stocked; and healthcare providers are using AI to analyze millions of medical images to more accurately identify disease and help save lives.

"We're at a tipping point as every industry seeks better ways to apply AI to offer new services and grow their business," said Ian Buck, general manager and vice president of Accelerated Computing at NVIDIA. "The work we've done to achieve these results on MLPerf gives companies a new level of AI performance to improve our everyday lives."

The latest MLPerf results come as NVIDIA's footprint for AI inference has grown dramatically. Five years ago, only a handful of leading high-tech companies used GPUs for inference. Now, with NVIDIA's AI platform available through every major cloud and data center infrastructure provider, companies representing a wide array of industries are using its AI inference platform to improve their business operations and offer additional services.

Additionally, for the first time, NVIDIA GPUs now offer more AI inference capacity in the public cloud than CPUs. Total cloud AI inference compute capacity on NVIDIA GPUs has been growing roughly 10x every two years.

NVIDIA Takes AI Inference to New Heights

NVIDIA and its partners submitted their MLPerf 0.7 results using NVIDIA's acceleration platform, which includes NVIDIA data center GPUs, edge AI accelerators and NVIDIA optimized software.

NVIDIA A100, introduced earlier this year and featuring third-generation Tensor Cores and Multi-Instance GPU technology, increased its lead on the ResNet-50 test, beating CPUs by 30x versus 6x in the last round. Additionally, A100 outperformed the latest CPUs by up to 237x in the newly added recommender test for data center inference, according to the MLPerf Inference 0.7 benchmarks.

This means a single NVIDIA DGX A100™ system can provide the same performance as about 1,000 dual-socket CPU servers, offering customers extreme cost efficiency when taking their AI recommender models from research to production.

The benchmarks also show that NVIDIA T4 Tensor Core GPU continues to be a solid inference platform for mainstream enterprise, edge servers and cost-effective cloud instances. NVIDIA T4 GPUs beat CPUs by up to 28x in the same tests. In addition, the NVIDIA Jetson AGX Xavier™ is the performance leader among SoC-based edge devices.

Achieving these results required a highly optimized software stack including NVIDIA TensorRT™ inference optimizer and NVIDIA Triton™ inference serving software, both available on NGC™, NVIDIA's software catalog.

In addition to NVIDIA's own submissions, 11 NVIDIA partners submitted a total of 1,029 results using NVIDIA GPUs, representing over 85 percent of the total submissions in the data center and edge categories.

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 https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: NVIDIA extending its lead on MLPerf benchmark and enabling businesses; NVIDIA breaking records; how organizations are using NVIDIA A100 Tensor Core GPUs; the benefits, performance and impact of NVIDIA GPUs, including NVIDIA A100, DGX A100, NVIDIA T4 Tensor Core GPU, NVIDIA Jetson AGX Xavier; industries seeking better ways to apply AI and grow their business; NVIDIA giving companies a new level of AI performance to improve our everyday lives; companies using NVIDIA's AI inference platform to

improve business operations and offer services; and NVIDIA taking AI inference to new heights 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com