NVIDIA Sets AI Inference Records, Introduces A30 and A10 GPUs for Enterprise Servers

NVIDIA AI Platform Smashes Every MLPerf Category, from Data Center to Edge

NVIDIA today announced that its AI inference platform, newly expanded with NVIDIA® A30 and A10 GPUs for mainstream servers, has achieved record-setting performance across every category on the latest release of MLPerf.

MLPerf is the industry's established benchmark for measuring AI performance across a range of workloads spanning computer vision, medical imaging, recommender systems, speech recognition and natural language processing.

Debuting on MLPerf, NVIDIA A30 and A10 GPUs combine high performance with low power consumption to provide enterprises with mainstream options for a broad range of AI inference, training, graphics and traditional enterprise compute workloads. Cisco, Dell Technologies, Hewlett Packard Enterprise, Inspur and Lenovo are expected to integrate the GPUs into their highest volume servers starting this summer.

NVIDIA achieved these results taking advantage of the full breadth of the NVIDIA AI platform — encompassing a wide range of GPUs and AI software, including TensorRT™ and NVIDIA Triton™ Inference Server — which is deployed by leading enterprises, such as Microsoft, Pinterest, Postmates, T-Mobile, USPS and WeChat.

"As AI continues to transform every industry, MLPerf is becoming an even more important tool for companies to make informed decisions on their IT infrastructure investments," said Ian Buck, general manager and vice president of Accelerated Computing at NVIDIA. "Now, with every major OEM submitting MLPerf results, NVIDIA and our partners are focusing not only on delivering world-leading performance for AI, but on democratizing AI with a coming wave of enterprise servers powered by our new A30 and A10 GPUs."

MLPerf Results
NVIDIA is the only company to submit results for every test in the data center and edge categories, delivering top performance results across all MLPerf workloads.

Several submissions also use Triton Inference Server, which simplifies the complexity of deploying AI in applications by supporting models from all major frameworks, running on GPUs, as well as CPUs, and optimizing for different query types including batch, real-time and streaming. Triton submissions achieved performance close to that of the most optimized GPU implementations, as well as CPU implementations, with comparable configurations.

NVIDIA also broke new ground with its submissions using the NVIDIA Ampere architecture's Multi-Instance GPU capability by simultaneously running all seven MLPerf Offline tests on a single GPU using seven MIG instances. The configuration showed nearly identical performance compared with a single MIG instance running alone.

These submissions demonstrate MIG's performance and versatility, which enable infrastructure managers to provision right-sized amounts of GPU compute for specific applications to get maximum output from every data center GPU.

In addition to NVIDIA's own submissions, NVIDIA partners Alibaba Cloud, Dell Technologies, Fujitsu, GIGABYTE, HPE, Inspur, Lenovo and Supermicro submitted a total of over 360 results using NVIDIA GPUs.

NVIDIA's Expanding AI Platform
The NVIDIA A30 and A10 GPUs are the latest additions to the NVIDIA AI platform, which includes NVIDIA Ampere architecture GPUs, NVIDIA Jetson AGX Xavier™ and Jetson Xavier NX, and a full stack of NVIDIA software optimized for accelerating AI.

The A30 delivers versatile performance for industry-standard servers, supporting a broad range of AI inference and mainstream enterprise compute workloads, such as recommender systems, conversational AI and computer vision.

The NVIDIA A10 GPU accelerates deep learning inference, interactive rendering, computer-aided design and cloud gaming, enabling enterprises to support mixed AI and graphics workloads on a common infrastructure. Using NVIDIA virtual GPU software, management can be streamlined to improve the utilization and provisioning of virtual desktops used by designers, engineers, artists and scientists.

The NVIDIA Jetson platform, based on the NVIDIA Xavier™ system-on-module, provides server-class AI performance at the edge, enabling a wide variety of applications in robotics, healthcare, retail and smart cities. Built on NVIDIA's unified architecture and the CUDA-X™ software stack, Jetson is the only platform capable of running all the edge workloads in compact designs while consuming less than 30W of power.

Availability

NVIDIA A100 GPUs are available in servers from leading manufacturers and in the cloud from all major cloud service providers. Additionally, A100 GPUs are featured across the NVIDIA DGX™ systems portfolio, including the NVIDIA DGX Station A100, NVIDIA DGX A100 and NVIDIA DGX SuperPOD.

The A30 and A10, which consume just 165W and 150W, are expected in a wide range of servers starting this summer, including NVIDIA-Certified Systems™ that go through rigorous testing to ensure high performance across a wide range of workloads.

The Jetson AGX Xavier and Jetson Xavier NX system-on-module are available from distributors globally.

NVIDIA Triton and NVIDIA TensorRT are both available on NGC™, NVIDIA's software catalog.

About NVIDIA

NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company's pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: NVIDIA setting and smashing records; the benefits, performance and impact of our products and technologies, including its AI inference and AI platforms, A30 GPUs, A10 GPUs, Triton Inference Server, Multi-Instance GPUs, NVIDIA virtual GPU software and NVIDIA Jetson; the companies expected to integrate GPUs into their servers this summer; the entities deploying our products; AI transforming every industry; the importance of MLPerf; which OEMs are submitting results; democratizing AI with NVIDIA; what MIG enables; and the availability of NVIDIA A100, A30 and A10 GPUs, Jetson AGX Xavier, Jenson Xavier NX system-on-a-module, NVIDIA Triton and NVIDIA TensorRT 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