NVIDIA and Microsoft Boost AI Cloud Computing with Launch of Industry-Standard Hyperscale GPU Accelerator

NVIDIA with Microsoft today unveiled blueprints for a new hyperscale GPU accelerator to drive AI cloud computing.

Providing hyperscale data centers with a fast, flexible path for AI, the new HGX-1 hyperscale GPU accelerator is an open-source design released in conjunction with Microsoft's Project Olympus.

HGX-1 does for cloud-based AI workloads what ATX -- Advanced Technology eXtended -- did for PC motherboards when it was introduced more than two decades ago. It establishes an industry standard that can be rapidly and efficiently embraced to help meet surging market demand.

The new architecture is designed to meet the exploding demand for AI computing in the cloud -- in fields such as autonomous driving, personalized healthcare, superhuman voice recognition, data and video analytics, and molecular simulations.

"AI is a new computing model that requires a new architecture," said Jen-Hsun Huang, founder and chief executive officer of NVIDIA. "The HGX-1 hyperscale GPU accelerator will do for AI cloud computing what the ATX standard did to make PCs pervasive today. It will enable cloud-service providers to easily adopt NVIDIA GPUs to meet surging demand for AI computing."

"The HGX-1 AI accelerator provides extreme performance scalability to meet the demanding requirements of fast-growing machine learning workloads, and its unique design allows it to be easily adopted into existing data centers around the world," wrote Kushagra Vaid, general manager and distinguished engineer, Azure Hardware Infrastructure, Microsoft, in a blog post.

For the thousands of enterprises and startups worldwide that are investing in AI and adopting AI-based approaches, the HGX-1 architecture provides unprecedented configurability and performance in the cloud.

Powered by eight NVIDIA® Tesla® P100 GPUs in each chassis, it features an innovative switching design -- based on NVIDIA NVLink™ interconnect technology and the PCIe standard -- enabling a CPU to dynamically connect to any number of GPUs. This allows cloud service providers that standardize on the HGX-1 infrastructure to offer customers a range of CPU and GPU machine instance configurations.

Cloud workloads are more diverse and complex than ever. AI training, inferencing and HPC workloads run optimally on different system configurations, with a CPU attached to a varying number of GPUs. The highly modular design of the HGX-1 allows for optimal performance no matter the workload. It provides up to 100x faster deep learning performance compared with legacy CPU-based servers, and is estimated at one-fifth the cost for conducting AI training and one-tenth the cost for AI inferencing.

With its flexibility to work with data centers across the globe, HGX-1 offers existing hyperscale data centers a quick, simple path to be ready for AI.

Collaboration to Bring Industry Standard to Hyperscale
Microsoft, NVIDIA and Ingrasys (a Foxconn subsidiary) collaborated to architect and design the HGX-1 platform. The companies are sharing it widely as part of Microsoft's Project Olympus contribution to the Open Compute Project, a consortium whose mission is to apply the benefits of open source to hardware and rapidly increase the pace of innovation in, near and around the data center and beyond.

Sharing the reference design with the broader Open Compute Project community means that enterprises can easily purchase and deploy the same design in their own data centers.

NVIDIA Joins Open Compute Project
NVIDIA is joining the Open Compute Project to help drive AI and innovation in the data center. The company plans to continue its work with Microsoft, Ingrasys and other members to advance AI-ready computing platforms for cloud service providers and other data center customers.

Keep Current on NVIDIA
Subscribe to the NVIDIA blog, follow us on Facebook, Google+, Twitter, LinkedIn and Instagram, and view NVIDIA videos on YouTube and images on Flickr.

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. Today, NVIDIA is increasingly known as "the AI computing company." More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the performance, impact and benefits of the HGX-1 hyperscale GPU accelerator; and NVIDIA joining the Open Compute Project 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 reports NVIDIA files with the Securities and Exchange Commission, or SEC, including its Form 10-K for the fiscal period ended January 29, 2017. 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