China's Top Server-Builders Adopt NVIDIA AI Design for Cloud Computing

Huawei, Inspur, Lenovo Select NVIDIA Volta HGX Architecture to Build AI Systems for Data Centers

GTC China - NVIDIA today announced that China's leading original equipment manufacturers (OEMs) -- including Huawei, Inspur and Lenovo -- are using the NVIDIA® HGX reference architecture to offer Volta architecture-based accelerated systems for hyperscale data centers.

Through the NVIDIA HGX Partner Program, NVIDIA is providing each OEM with early access to the NVIDIA HGX reference architecture for data centers, NVIDIA GPU computing technologies, and design guidelines. HGX is the same data center design used in Microsoft's Project Olympus initiative, Facebook's Big Basin systems and NVIDIA DGX-1™ AI supercomputers.

Using HGX as a starter "recipe," OEM and original design manufacturer (ODM) partners can work with NVIDIA to more quickly design and bring to market a wide range of qualified GPU-accelerated AI systems for hyperscale data centers to meet the industry's growing demand for AI cloud computing.

New HGX server designs coming to market -- with eight NVIDIA Tesla® V100 GPU accelerators in a hybrid cube mesh with NVIDIA NVLink™ interconnect technology -- include Lenovo HG690X and HG695X, Inspur 2U-8GPUAGX-2 and Huawei G-series heterogenous servers.

With GPUs based on the NVIDIA Volta architecture offering three times the performance of its predecessor, manufacturers can meet market demand with new products based on the latest NVIDIA technology.

"As companies increasingly harness the capabilities of artificial intelligence, demand continues to grow for accelerated computing in the data center," said Ian Buck, general manager of Accelerated Computing at NVIDIA. "Our new Volta-based HGX design sets the standard for lightning-fast, energy-efficient data centers that can support the most demanding AI training and inference requirements."

Highly configurable to meet workload needs, HGX can easily combine GPUs and CPUs in a number of ways for high performance computing, deep learning training and inferencing.

The standard HGX design architecture includes eight NVIDIA Tesla GPU accelerators in the SXM2 form factor and connected in a cube mesh using NVIDIA NVLink high-speed interconnects and optimized PCIe topologies. With a modular design, HGX enclosures are suited for deployment in existing data center racks across the globe, using hyperscale CPU nodes as needed.

Industry Support for Volta HGX Platform

"NVIDIA is the world's leading artificial intelligence and accelerated computing technology provider. Inspur has been focusing on the research and development of deep learning and AI computing systems for years; in May this year, Inspur and NVIDIA jointly released an HGX-based system equipped with the latest NVIDIA Tesla V100 GPU and NVLink high-speed interconnect technology, which is designed to provide maximum throughput with higher power consumption efficiency for high performance computing in scientific research and engineering, taking AI computing to the next level."
-- Leijun Hu, vice president of Inspur Group

"NVIDIA HGX architecture provides exceptional capabilities and energy efficiency. Combined with Huawei's strength in computing and connectivity, we can provide an outstanding computing solution with AI capabilities to growing numbers of enterprises."
-- Wu Zhan, vice president of IT Server Product Line at Huawei

"New AI workloads are creating strong demand for high-performance and flexible data center architectures. NVIDIA's modular HGX design offers best-in-class performance for these new AI workloads. Volta HGX designs will help us meet our customers' needs while reducing their energy requirements."
-- Paul Ju, vice president and general manager of Global Hyperscale Segment at Lenovo

NVIDIA also announced today that Alibaba, Baidu and Tencent are incorporating new Volta architecture-based NVIDIA Tesla V100 GPU accelerators into their data centers and cloud-service infrastructures.

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. More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the impact of OEM partners working with NVIDIA; the industry's growing demand for AI cloud computing; the performance and impact of NVIDIA Volta architecture; companies increasingly harnessing artificial intelligence; the growing demand for accelerated computing in the data center; and the benefits of the HGX design 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-Q for the fiscal period ended July 30, 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