NVIDIA Introduces HGX-2, Fusing HPC and AI Computing into Unified Architecture

HGX-2 Cloud-Server Platform Accelerates Multi-Precision Workloads; Its Two Petaflops of Processing Power Sets Record for AI Performance

GTC Taiwan--NVIDIA today introduced NVIDIA HGX-2™, the first unified computing platform for both artificial intelligence and high performance computing.

The HGX-2 cloud server platform, with multi-precision computing capabilities, provides unique flexibility to support the future of computing. It allows high-precision calculations using FP64 and FP32 for scientific computing and simulations, while also enabling FP16 and Int8 for AI training and inference. This unprecedented versatility meets the requirements of the growing number of applications that combine HPC with AI.

A number of leading computer makers today shared plans to bring to market systems based on the NVIDIA HGX-2 platform.

“The world of computing has changed," said Jensen Huang, founder and chief executive officer of NVIDIA, speaking at the GPU Technology Conference Taiwan, which kicked off today. “CPU scaling has slowed at a time when computing demand is skyrocketing. NVIDIA's HGX-2 with Tensor Core GPUs gives the industry a powerful, versatile computing platform that fuses HPC and AI to solve the world's grand challenges."

HGX-2-serves as a “building block" for manufacturers to create some of the most advanced systems for HPC and AI. It has achieved record AI training speeds of 15,500 images per second on the ResNet-50 training benchmark, and can replace up to 300 CPU-only servers.

It incorporates such breakthrough features as NVIDIA NVSwitch™ interconnect fabric, which seamlessly links 16 NVIDIA Tesla® V100 Tensor Core GPUs to work as a single, giant GPU delivering two petaflops of AI performance. The first system built using HGX-2 was the recently announced NVIDIA DGX-2™.

HGX-2 comes a year after the launch of the original NVIDIA HGX-1, at Computex 2017. The HGX-1 reference architecture won broad adoption among the world's leading server makers and companies operating massive datacenters, including Amazon Web Services, Facebook and Microsoft.

OEM, ODM Systems Expected Later This Year

Four leading server makers -- Lenovo, QCT, Supermicro and Wiwynn -- announced plans to bring their own HGX-2-based systems to market later this year.

Additionally, four of the world's top original design manufacturers (ODMs) -- Foxconn, Inventec, Quanta and Wistron -- are designing HGX-2-based systems, also expected later this year, for use in some of the world's largest cloud datacenters.

Family of NVIDIA GPU-Accelerated Server Platforms

HGX-2 is a part of the larger family of NVIDIA GPU-Accelerated Server Platforms, an ecosystem of qualified server classes addressing a broad array of AI, HPC and accelerated computing workloads with optimal performance.

Supported by major server manufacturers, the platforms align with the datacenter server ecosystem by offering the optimal mix of GPUs, CPUs and interconnects for diverse training (HGX-T2), inference (HGX-I2) and supercomputing (SCX) applications. Customers can choose a specific server platform to match their accelerated computing workload mix and achieve best-in-class performance.

Broad Industry Support

Top OEMs and ODMs have voiced strong support for HGX-2:

“Foxconn has long been dedicated to hyperscale computing solutions and successfully won customer recognition. We're glad to work with NVIDIA for the HGX-2 project, which is the most promising solution to fulfill explosive demand from AI/DL."

-- Ed Wu, corporate executive vice president at Foxconn and chairman at Ingrasys

“Inventec has a proven history of delivering high-performing and scalable servers with robust innovative designs for our customers who run some of the world's largest datacenters. By rapidly incorporating HGX-2 into our future designs, we'll infuse our portfolio with the most powerful AI solution available to companies worldwide."

-- Evan Chien, head of IEC White Box Product Center, China Business Line Director, Inventec

“NVIDIA's HGX-2 ups the ante with a design capable of delivering two petaflops of performance for AI and HPC-intensive workloads. With the HGX-2 server building block, we'll be able to quickly develop new systems that can meet the growing needs of our customers who demand the highest performance at scale."

-- Paul Ju, vice president and general manager of Lenovo DCG

“As a leading cloud enabler, Quanta is committed to developing solutions for the next generation of clouds for a variety of innovative use cases. As we have seen a multitude of AI applications on the rise, Quanta works closely with NVIDIA to ensure our clients benefit from the latest and greatest GPU technologies. We are thrilled to broaden our GPU compute portfolio with this critical enabler for AI clouds as an HGX-2 launch partner."

-- Mike Yang, senior vice president, Quanta Computer, and president, QCT

“To help address the rapidly expanding size of AI models that sometimes require weeks to train, Supermicro is developing cloud servers based on the HGX-2 platform. The HGX-2 system will enable efficient training of complex models."

-- Charles Liang, president and CEO of Supermicro

“We are very honored to work with NVIDIA as a partner. The demand for AI cloud computing is emerging in today's modern technology environment. I strongly believe the high performance and modularized flexibility of the HGX-2 system will make great contributions to various computing areas, ranging from academics and science to government applications.''

-- Jeff Lin, president of Enterprise Business Group, Wistron

“Wiwynn specializes in delivering hyperscale datacenter and cloud infrastructure solutions. Our collaboration with NVIDIA and the HGX-2 server building block will enable us to provide our customers with two petaflops of computing for computationally intensive AI and HPC workloads.''

-- Steven Lu, Vice President, Wiwynn

About NVIDIA

NVIDIA's 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, performance and abilities of the NVIDIA HGX-2 cloud server platform; HGX-2 providing flexibility and versatility to support the future of computing and ability to meet the requirements of applications combining HPC with AI; computer makers' plans to bring to market systems based on the HGX-2 platform; CPU scaling slowing while computing demand is skyrocketing; HGX-2 giving the industry a powerful and versatile platform to solve the world's grand challenges; HGX-2 serving as a building block for manufacturers to create one of the most advanced systems for HPC and AI and its ability to replace up to 300 CPU-only servers; HGX-1 winning broad adoption among the world's leading server makers and companies operating datacenters; leading server makers' and top original design manufacturers' plans to bring HGX-2-based systems to market later this year; the benefits, performance and abilities of the NVIDIA GPU-Accelerated Server Platforms; HGX-2 being the most promising solution to fulfill demand from AI/DL; HGX-2 being incorporated into Inventec's future designs and it infusing Inventec's portfolio with the most powerful AI solution available; HGX-2's ability to help Lenovo's systems to meet the growing needs of their customers; HGX-2 systems enabling the efficient training of complex models; the demand for AI cloud computing emerging in today's modern technology environment; HGX-2's ability to make great contributions to various computing areas; and HGX-2 enabling Wiwynn to provide customers with two petaflops of computing for AI and HPC workloads 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 April 29, 2018. 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