NVIDIA Partners with World's Top Server Manufacturers to Advance AI Cloud Computing

Foxconn, Inventec, Quanta, Wistron Using NVIDIA HGX Reference Architecture to Build AI Systems for Hyperscale Data Centers

Computex -- NVIDIA today launched a partner program with the world's leading original design manufacturers (ODM) -- Foxconn, Inventec, Quanta and Wistron -- to more rapidly meet the demands for AI cloud computing.

Through the NVIDIA HGX Partner Program, NVIDIA is providing each ODM with early access to the NVIDIA HGX reference architecture, 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," ODM partners can work with NVIDIA to more quickly design and bring to market a wide range of qualified GPU-accelerated systems for hyperscale data centers. Through the program, NVIDIA engineers will work closely with ODMs to help minimize the amount of time from design win to production deployments.

As the overall demand for AI computing resources has risen sharply over the past year, so has the market adoption and performance of NVIDIA's GPU computing platform. Today, 10 of the world's top 10 hyperscale businesses are using NVIDIA GPU accelerators in their data centers.

With new NVIDIA® Volta architecture-based GPUs offering three times the performance of its predecessor, ODMs can feed the market demand with new products based on the latest NVIDIA technology available.

"Accelerated computing is evolving rapidly -- in just one year we tripled the deep learning performance in our Tesla GPUs -- and this is having a significant impact on the way systems are designed," said Ian Buck, general manager of Accelerated Computing at NVIDIA. "Through our HGX partner program, device makers can ensure they're offering the latest AI technologies to the growing community of cloud computing providers."

Flexible, Upgradable Design
NVIDIA built the HGX reference design to meet the high-performance, efficiency and massive scaling requirements unique to hyperscale cloud environments. Highly configurable based on workload needs, HGX can easily combine GPUs and CPUs in a number of ways for high performance computing, deep learning training and deep learning 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.

Both NVIDIA Tesla P100 and V100 GPU accelerators are compatible with HGX. This allows for immediate upgrades of all HGX-based products once V100 GPUs become available later this year.

HGX is an ideal reference architecture for cloud providers seeking to host the new NVIDIA GPU Cloud platform. The NVIDIA GPU Cloud platform manages a catalog of fully integrated and optimized deep learning framework containers, including Caffe2, Cognitive Toolkit, MXNet and TensorFlow.

"Through this new partner program with NVIDIA, we will be able to more quickly serve the growing demands of our customers, many of whom manage some of the largest data centers in the world," said Taiyu Chou, general manager of Foxconn/Hon Hai Precision Ind Co., Ltd., and president of Ingrasys Technology Inc. "Early access to NVIDIA GPU technologies and design guidelines will help us more rapidly introduce innovative products for our customers' growing AI computing needs."

"Working more closely with NVIDIA will help us infuse a new level of innovation into data center infrastructure worldwide," said Evan Chien, head of IEC China operations at Inventec Corporation. "Through our close collaboration, we will be able to more effectively address the compute-intensive AI needs of companies managing hyperscale cloud environments."

"Tapping into NVIDIA's AI computing expertise will allow us to immediately bring to market game-changing solutions to meet the new computing requirements of the AI era," said Mike Yang, senior vice president at Quanta Computer Inc. and president at QCT.

"As a long-time collaborator with NVIDIA, we look forward to deepening our relationship so that we can meet the increasing computing needs of our hyperscale data center customers," said Donald Hwang, chief technology officer and president of the Enterprise Business Group at Wistron. "Our customers are hungry for more GPU computing power to handle a variety of AI workloads, and through this new partnership we will be able to deliver new solutions faster."

"We've collaborated with Ingrasys and NVIDIA to pioneer a new industry standard design to meet the growing demands of the new AI era," said Kushagra Vaid, general manager and distinguished engineer, Azure Hardware Infrastructure, Microsoft Corp. "The HGX-1 AI accelerator has been developed as a component of Microsoft's Project Olympus to achieve extreme performance scalability through the option for high-bandwidth interconnectivity for up to 32 GPUs."

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 benefits and impact of NVIDIA's partnership programs, NVIDIA Volta architecture-based GPUs and the HGX reference design; and the availability of V100 GPUs 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 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