NVIDIA and Global Partners Launch New HGX A100 Systems to Accelerate Industrial AI and HPC

Wide Range of HPC Systems and Cloud Services Powered by HGX Now Supercharged with NVIDIA A100 80G PCIe, NVIDIA NDR 400G InfiniBand, NVIDIA Magnum IO

ISC—NVIDIA today announced it is turbocharging the NVIDIA HGX™ AI supercomputing platform with new technologies that fuse AI with high performance computing, making supercomputing more useful to a growing number of industries.

To accelerate the new era of industrial AI and HPC, NVIDIA has added three key technologies to its HGX platform: the NVIDIA® A100 80GB PCIe GPU, NVIDIA NDR 400G InfiniBand networking, and NVIDIA Magnum IO™ GPUDirect™ Storage software. Together, they provide the extreme performance to enable industrial HPC innovation.

Atos, Dell Technologies, Hewlett Packard Enterprise (HPE), Lenovo, Microsoft Azure and NetApp are among dozens of partners using the NVIDIA HGX platform for next-generation systems and solutions.

“The HPC revolution started in academia and is rapidly extending across a broad range of industries,” said Jensen Huang, founder and CEO of NVIDIA. “Key dynamics are driving super-exponential, super-Moore’s law advances that have made HPC a useful tool for industries. NVIDIA’s HGX platform gives researchers unparalleled high performance computing acceleration to tackle the toughest problems industries face.”

Industry Leaders Use HGX Platform to Power Innovation Breakthroughs

The HGX platform is being used by high-tech industrial pioneer General Electric, applying HPC innovation to computational fluid dynamics simulations that guide design innovation in large gas turbines and jet engines. The HGX platform has achieved order-of-magnitude acceleration for breakthrough CFD methods in GE’s GENESIS code, which employs Large Eddy Simulations to study the effects of turbulent flows inside turbines that are composed of hundreds of individual blades and require uniquely complex geometry.

Besides driving industrial HPC transformation, the HGX platform is also accelerating scientific HPC systems around the world, including the next next-generation supercomputer at the University of Edinburgh, also announced today.

NVIDIA A100 80GB PCIe Performance Enhancements for AI and HPC
NVIDIA A100 Tensor Core GPUs deliver unprecedented HPC acceleration to solve complex AI, data analytics, model training and simulation challenges relevant to industrial HPC. A100 80GB PCIe GPUs increase GPU memory bandwidth 25 percent compared with the A100 40GB, to 2TB/s, and provide 80GB of HBM2e high-bandwidth memory.

The A100 80GB PCIe’s enormous memory capacity and high-memory bandwidth allow more data and larger neural networks to be held in memory, minimizing internode communication and energy consumption. Combined with faster memory bandwidth, it enables researchers to achieve higher throughput and faster results, maximizing the value of their IT investments.

A100 80GB PCIe is powered by the NVIDIA Ampere architecture, which features Multi-Instance GPU (MIG) technology to deliver acceleration for smaller workloads such as AI inference. MIG allows HPC systems to scale compute and memory down with guaranteed quality of service. In addition to PCIe, there are four- and eight-way NVIDIA HGX A100 configurations.

NVIDIA partner support for the A100 80GB PCIe includes Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, H3C, HPE, Inspur, Lenovo, Penguin Computing, QCT and Supermicro. The HGX platform featuring A100-based GPUs interconnected via NVLink is also available via cloud services from Amazon Web Services, Microsoft Azure and Oracle Cloud Infrastructure.

Next-Generation NDR 400Gb/s InfiniBand Switch Systems

HPC systems that require unparalleled data throughout are supercharged by NVIDIA InfiniBand, the world’s only fully offloadable in-network computing interconnect. NDR InfiniBand scales performance to tackle the massive challenges in industrial and scientific HPC systems. The NVIDIA Quantum™-2 fixed-configuration switch systems deliver 64 ports of NDR 400Gb/s InfiniBand per port (or 128 ports of NDR200), providing 3x higher port density versus HDR InfiniBand.

The NVIDIA Quantum-2 modular switches provide scalable port configurations up to 2,048 ports of NDR 400Gb/s InfiniBand (or 4,096 ports of NDR200) with a total bidirectional throughput of 1.64 petabits per second — 5x over the previous-generation. The 2,048-port switch provides 6.5x greater scalability over the previous generation, with the ability to connect more than a million nodes with just three hops using a DragonFly+ network topology.

The third generation of NVIDIA SHARP In-Network Computing data reduction technology boosts performance for high-

performance industrial and scientific applications with 32x higher AI acceleration power compared to the previous generation.

Advanced management features include self-healing network capabilities and NVIDIA In-Network Computing acceleration engines. Data center downtime is further minimized with the NVIDIA UFM® Cyber-AI platform.

Based on industry standards, the NVIDIA Quantum-2 switches — which are expected to sample in the third quarter — are backward- and forward-compatible, enabling easy migration and expansion of existing systems and software.

Industry-leading infrastructure manufacturers — including Atos, DDN, Dell Technologies, Excelero, GIGABYTE, HPE, Lenovo, Penguin Computing, QCT, Supermicro, VAST and WekaIO — plan to integrate the Quantum-2 NDR 400Gb/s InfiniBand switches into their enterprise and HPC offerings. Cloud service providers including Azure are also taking advantage of InfiniBand technology.

Introducing Magnum IO GPUDirect Storage

Providing unrivaled performance for complex workloads, Magnum IO GPUDirect Storage enables direct memory access between GPU memory and storage. The direct path enables applications to benefit from lower I/O latency and use the full bandwidth of the network adapters while decreasing the utilization load on the CPU and managing the impact of increased data consumption.

Industry leaders supporting Magnum IO GPUDirect Storage, which is available now, include DDN, Dell Technologies, Excelero, HPE, IBM Storage, Micron, NetApp, Pavilion, ScaleFlux, VAST and WekaIO. A full list of storage partners is available at https://developer.nvidia.com/gpudirect-storage.

Tune in to the NVIDIA ISC 2021 Special Address at 9:30 a.m. PT to get an overview of the latest news from NVIDIA's Marc Hamilton, followed by a live Q&A panel with NVIDIA HPC experts.

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: the benefits, impact, performance, features and availability of our products and services; NVIDIA turbocharging the NVIDIA HGX AI supercomputing platform; partners using the NVIDIA HGX platform for next-generation systems and solutions; NVIDIA partners supporting the A100 80GB PCIe; NDR InfiniBand scaling performance to tackle the massive challenges in industrial and scientific HPC systems; the expected timing of the sampling of NVIDIA Quantum-2 switches; industry-leading infrastructure manufacturers planning to integrate the Quantum-2 NDR 400Gb/s InfiniBand switches into their enterprise and HPC offerings; cloud service providers taking advantage of InfiniBand technology; and industry leaders supporting Magnum IO GPUDirect Storage 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.

Alex Shapiro
Enterprise Networking
1-415-608-5044
ashapiro@nvidia.com