NVIDIA Ships World's Most Advanced AI System — NVIDIA DGX A100 — to Fight COVID-19; Third-Generation DGX Packs Record 5 Petaflops of AI Performance

Training, Inference, Data Analytics Unified on One Platform; Each System Configurable from One to 56 Independent GPUs to Deliver Elastic, Software-Defined Data Center Infrastructure

GTC 2020 -- NVIDIA today unveiled NVIDIA DGX™ A100, the third generation of the world's most advanced AI system, delivering 5 petaflops of AI performance and consolidating the power and capabilities of an entire data center into a single flexible platform for the first time.

Immediately available, DGX A100 systems have begun shipping worldwide, with the first order going to the U.S. Department of Energy's (DOE) Argonne National Laboratory, which will use the cluster's AI and computing power to better understand and fight COVID-19.

“NVIDIA DGX A100 is the ultimate instrument for advancing AI,'' said Jensen Huang, founder and CEO of NVIDIA. “NVIDIA DGX is the first AI system built for the end-to-end machine learning workflow -- from data analytics to training to inference. And with the giant performance leap of the new DGX, machine learning engineers can stay ahead of the exponentially growing size of AI models and data.''

DGX A100 systems integrate eight of the new NVIDIA A100 Tensor Core GPUs, providing 320GB of memory for training the largest AI datasets, and the latest high-speed NVIDIA Mellanox® HDR 200Gbps interconnects.

Multiple smaller workloads can be accelerated by partitioning the DGX A100 into as many as 56 instances per system, using the A100 multi-instance GPU feature. Combining these capabilities enables enterprises to optimize computing power and resources on demand to accelerate diverse workloads, including data analytics, training and inference, on a single, fully integrated, software-defined platform.

Immediate DGX A100 Adoption, Support
A number of the world's largest companies, service providers and government agencies have placed initial orders for the DGX A100, with the first systems delivered to Argonne earlier this month.

“We're using America's most powerful supercomputers in the fight against COVID-19, running AI models and simulations on the latest technology available, like the NVIDIA DGX A100,'' said Rick Stevens, associate laboratory director for Computing, Environment and Life Sciences at Argonne. “The compute power of the new DGX A100 systems coming to Argonne will help researchers explore treatments and vaccines and study the spread of the virus, enabling scientists to do years' worth of AI-accelerated work in months or days.''

The University of Florida will be the first institution of higher learning in the U.S. to receive DGX A100 systems, which it will deploy to infuse AI across its entire curriculum to foster an AI-enabled workforce.

“The University of Florida has a vision to be a national leader in artificial intelligence, and NVIDIA is an incredibly valuable partner in our quest to do so,'' said University of Florida President Kent Fuchs. “Across disciplines, our new NVIDIA DGX A100 systems will position our researchers to solve some of our world's most pressing challenges and equip an entire generation of students with the skills that will revolutionize the future workforce.''

Among other early adopters are:

- The Center for Biomedical AI -- at the University Medical Center Hamburg-Eppendorf, Germany -- will leverage DGX A100 to advance clinical decision support and process optimization.
- Chulalongkorn University -- Thailand's top research-intensive university -- will use DGX A100 to accelerate its pioneering research such as Thai natural language processing, automatic speech recognition, computer vision and medical imaging.
- Element AI -- a Montreal-based developer of AI-powered solutions and services -- is deploying DGX A100 to accelerate performance and feature optimization for its Orkestrator GPU scheduler to meet growing AI training and application demands.
- German Research Center for Artificial Intelligence (DFKI) will use the DGX A100 systems to further accelerate its research on new deep learning methods and their explainability while significantly reducing space and energy consumption.
- Harrison.ai -- a Sydney-based healthcare AI company -- will deploy Australia's first DGX A100 systems to accelerate the development of its AI-as-medical-device.
- The UAE Artificial Intelligence Office -- first in the Middle East to deploy the new DGX A100 -- is building a national infrastructure to accelerate AI research, development and adoption across the public and private sector.
- VinAI Research -- Vietnam's leading AI research lab, based in Hanoi and Ho Chi Minh City -- will use DGX A100 to conduct high-impact research and accelerate the application of AI.

Thousands of previous-generation DGX systems are in use around the globe by a wide range of public and private organizations. Among them are some of the world's leading businesses, including automakers, healthcare providers, retailers, financial institutions and logistics companies that are pushing AI forward across their industries.

NVIDIA Builds Next-Gen 700 Petaflops DGX SuperPOD
NVIDIA also revealed its next-generation DGX SuperPOD, a cluster of 140 DGX A100 systems capable of achieving 700 petaflops of AI computing power. Combining 140 DGX A100 systems with Mellanox HDR 200Gbps InfiniBand interconnects, NVIDIA built the DGX SuperPOD AI supercomputer for internal research in areas such as conversational AI, genomics and autonomous driving.

The cluster is one of the world's fastest AI supercomputers -- achieving a level of performance that previously required thousands of servers. The enterprise-ready architecture and performance of the DGX A100 enabled NVIDIA to build the system in less than a month, instead of taking months or years of

planning and procurement of specialized components previously required to deliver these supercomputing capabilities.

To help customers build their own A100-powered data centers, NVIDIA has released a new DGX SuperPOD reference architecture. It gives customers a blueprint that follows the same design principles and best practices NVIDIA used to build its DGX A100-based AI supercomputing cluster.

DGXpert Program, DGX-Ready Software

NVIDIA also launched the NVIDIA DGXpert program, which brings together DGX customers with the company's AI experts; and the NVIDIA DGX-Ready Software program, which helps customers take advantage of certified, enterprise-grade software for AI workflows.

DGXperts are AI-fluent specialists who can help guide clients on AI deployments, from planning to implementation to ongoing optimization. These individuals can help DGX A100 customers build and maintain state-of-the-art AI infrastructure.

The NVIDIA DGX-Ready Software program helps customers quickly identify and take advantage of NVIDIA-tested third-party MLOps software that can help them increase data science productivity, accelerate AI workflows and improve accessibility and utilization of AI infrastructure. The first program partners certified by NVIDIA are Allegro AI, cnvrg.io, Core Scientific, Domino Data Lab, Iguazio and Paperspace.

DGX A100 Technical Specifications

- Eight NVIDIA A100 Tensor Core GPUs, delivering 5 petaflops of AI power, with 320GB in total GPU memory with 12.4TB per second in bandwidth.
- Six NVIDIA NVSwitch™ interconnect fabrics with third-generation NVIDIA NVLink® technology for 4.8TB per second of bi-directional bandwidth.
- Nine Mellanox ConnectX-6 HDR 200Gb per second network interfaces, offering a total of 3.6Tb per second of bi-directional bandwidth.
- Mellanox In-Network Computing and network acceleration engines such as RDMA, GPUDirect® and Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ to enable the highest performance and scalability.
- 15TB Gen4 NVMe internal storage, which is 2x faster than Gen3 NVMe SSDs.
- NVIDIA DGX software stack, which includes optimized software for AI and data science workloads, delivering maximized performance and enabling enterprises to achieve a faster return on their investment in AI infrastructure.

A single rack of five DGX A100 systems replaces a data center of AI training and inference infrastructure, with 1/20th the power consumed, 1/25th the space and 1/10th the cost.

Availability

NVIDIA DGX A100 systems start at $199,000 and are shipping now through NVIDIA Partner Network resellers worldwide. Storage technology providers DDN Storage, Dell Technologies, IBM, NetApp, Pure Storage and Vast plan to integrate DGX A100 into their offerings, including those based on the NVIDIA DGX POD and DGX SuperPOD reference architectures.

NVIDIA DGX-Ready Data Center partners offer colocation services in more than 122 locations across 26 countries to help customers seeking cost-effective facilities to host their DGX infrastructure. Customers can take advantage of these services to house and access DGX A100 infrastructure inside validated, world-class data center facilities.

Further information, including detailed technical specifications and ordering details, is available at www.nvidia.com/DGX-A100.

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 https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance, features and availability of our products, technologies and services, including NVIDIA DGX A100, NVIDIA A100 Tensor Core GPUs, DGX SuperPOD, NVIDIA Mellanox InfiniBand interconnects, DGX SuperPOD reference architecture, NVIDIA DGXpert program, NVIDIA DGX-Ready Software program, NVSwitch interconnect fabrics, NVLink technology, Mellanox's In-Network Computing and network acceleration engines, NVIDIA DGX software stack, and NVIDIA DGX-Ready Data Center partners 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.

Shannon McPhee
+1-310-920-9642
smcphee@nvidia.com