NVIDIA Announces DGX H100 Systems – World's Most Advanced Enterprise AI Infrastructure

Combined With New NVLink Switch System, Each DGX SuperPOD to Deliver 1 Exaflops of AI Performance; New NVIDIA Eos Supercomputer Expected to Be World's Fastest AI System; Immediate On-Ramp for Customers via Expanded DGX Foundry Service

GTC—NVIDIA today announced the fourth-generation NVIDIA® DGX™ system, the world's first AI platform to be built with new NVIDIA H100 Tensor Core GPUs.

DGX H100 systems deliver the scale demanded to meet the massive compute requirements of large language models, recommender systems, healthcare research and climate science. Packing eight NVIDIA H100 GPUs per system, connected as one by NVIDIA NVLink®, each DGX H100 provides 32 petaflops of AI performance at new FP8 precision — 6x more than the prior generation.

DGX H100 systems are the building blocks of the next-generation NVIDIA DGX POD™ and NVIDIA DGX SuperPOD™ AI infrastructure platforms. The latest DGX SuperPOD architecture features a new NVIDIA NVLink Switch System that can connect up to 32 nodes with a total of 256 H100 GPUs.

Providing 1 exaflops of FP8 AI performance, 6x more than its predecessor, the next-generation DGX SuperPOD expands the frontiers of AI with the ability to run massive LLM workloads with trillions of parameters.

"AI has fundamentally changed what software can do and how it is produced. Companies revolutionizing their industries with AI realize the importance of their AI infrastructure," said Jensen Huang, founder and CEO of NVIDIA. "Our new DGX H100 systems will power enterprise AI factories to refine data into our most valuable resource — intelligence."

Announcing NVIDIA Eos — World's Fastest AI Supercomputer
NVIDIA will be first to build a DGX SuperPOD with the groundbreaking new AI architecture to power the work of NVIDIA researchers advancing climate science, digital biology and the future of AI.

Its "Eos" supercomputer is expected to be the world's fastest AI system after it begins operations later this year, featuring a total of 576 DGX H100 systems with 4,608 DGX H100 GPUs.

NVIDIA Eos is anticipated to provide 18.4 exaflops of AI computing performance, 4x faster AI processing than the Fugaku supercomputer in Japan, which is currently the world's fastest system. For traditional scientific computing, Eos is expected to provide 275 petaflops of performance.

Eos will serve as a blueprint for advanced AI infrastructure from NVIDIA, as well as its OEM and cloud partners.

Enterprise AI Scales Easily With DGX H100 Systems, DGX POD and DGX SuperPOD
DGX H100 systems easily scale to meet the demands of AI as enterprises grow from initial projects to broad deployments.

In addition to eight H100 GPUs with an aggregated 640 billion transistors, each DGX H100 system includes two NVIDIA BlueField®-3 DPUs to offload, accelerate and isolate advanced networking, storage and security services.

Eight NVIDIA ConnectX®-7 Quantum-2 InfiniBand networking adapters provide 400 gigabits per second throughput to connect with computing and storage — double the speed of the prior generation system. And a fourth-generation NVLink, combined with NVSwitch™, provides 900 gigabytes per second connectivity between every GPU in each DGX H100 system, 1.5x more than the prior generation.

DGX H100 systems use dual x86 CPUs and can be combined with NVIDIA networking and storage from NVIDIA partners to make flexible DGX PODs for AI computing at any size.

DGX SuperPOD provides a scalable enterprise AI center of excellence with DGX H100 systems. The DGX H100 nodes and H100 GPUs in a DGX SuperPOD are connected by an NVLink Switch System and NVIDIA Quantum-2 InfiniBand providing a total of 70 terabytes/sec of bandwidth – 11x higher than the previous generation. Storage from NVIDIA partners will be tested and certified to meet the demands of DGX SuperPOD AI computing.

Multiple DGX SuperPOD units can be combined to provide the AI performance needed to develop massive models in industries such as automotive, healthcare, manufacturing, communications, retail and more.

NVIDIA DGX Foundry Speeds Customer Success With DGX SuperPOD
The NVIDIA DGX Foundry hosted development solution is expanding worldwide to give DGX SuperPOD customers

immediate access to advanced computing infrastructure while their systems are being installed. New locations added in North America, Europe and Asia offer remote access to DGX SuperPODs, or a portion of one.

DGX Foundry includes NVIDIA Base Command™ software, which lets customers easily manage the end-to-end AI development lifecycle on DGX SuperPOD infrastructure.

Qualified enterprises can experience NVIDIA Base Command and DGX systems at no charge through curated labs available through NVIDIA LaunchPad hosted at Equinix International Business Exchange™ (IBX®) data centers around the world.

MLOps, Enterprise AI Software Support Customers' Growing AI Adoption
To support DGX customers who are operationalizing AI development, MLOps solutions from NVIDIA DGX-Ready Software partners including Domino Data Lab, Run:ai and Weights & Biases are joining the "NVIDIA AI Accelerated" program.

MLOps applications from participating partners will be validated to provide DGX customers with enterprise-grade workflow and cluster management, scheduling and orchestration solutions.

Additionally, NVIDIA DGX systems now include the NVIDIA AI Enterprise software suite, which newly supports bare-metal infrastructure. DGX customers can accelerate their work with the pretrained NVIDIA AI platform models, toolkits and frameworks included in the software suite, such as NVIDIA RAPIDS™, NVIDIA TAO Toolkit, NVIDIA Triton Inference Server™ and more.

DGX-Ready Managed Services Program Simplify AI Deployments
As enterprise AI adoption grows, customers are seeking more options for adding the infrastructure needed to transform their businesses. NVIDIA is introducing a new DGX-Ready Managed Services program to support customers who would like to work with service providers to oversee their infrastructure.

Deloitte is the first global provider to be teaming with NVIDIA in the program and will be certified to support customers in Europe, North America and Asia, along with regional providers CGit, ePlus inc., Insight Enterprises and PTC System.

"The business breakthroughs made possible with AI can only be realized if enterprises have the ability to integrate the technology into their operations," said Jim Rowan, principal and AI and Data Operations offering leader at Deloitte Consulting LLP. "With the new DGX-Ready Managed Services program, clients can easily adopt world-leading AI with NVIDIA DGX systems and software managed by Deloitte experts around the world."

DGX-Ready Lifecycle Management Program Enables Easy Upgrades
Customers now have the ability to upgrade their existing DGX systems with the newest NVIDIA DGX platform through the new DGX-Ready Lifecycle Management program.

NVIDIA channel partners participating in the DGX-Ready Lifecycle Management program will be able to refresh the previous-generation DGX systems for purchase by new customers, expanding access to the world's universal systems for AI infrastructure.

Availability
NVIDIA DGX H100 systems, DGX PODs and DGX SuperPODs will be available from NVIDIA's global partners starting in the third quarter.

Customers can also choose to deploy DGX systems at colocation facilities operated by NVIDIA DGX-Ready Data Center partners including Cyxtera, Digital Realty and Equinix IBX data centers.

To learn more about NVIDIA DGX systems, watch Huang's GTC 2022 keynote, and register for GTC for free to attend sessions with NVIDIA and industry leaders.

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, specifications, performance and availability of Hopper-based DGX H100 systems, NVIDIA H100 GPUs, NVIDIA DGX POD, NVIDIA DGX SuperPOD, NVIDIA Eos, NVIDIA BlueField-3 DPUs, NVIDIA ConnectX-7 Quantum-2 InfiniBand networking adapters, NVLink, NVSwitch, the NVIDIA DGX Foundry, NVIDIA Base Command software and the NVIDIA AI Enterprise software suite; AI being essential to building systems; storage from NVIDIA partners being tested and certified to meet the demands of DGX SuperPOD AI computing; MLOps applications being validated to provide DGX customers with workflow and cluster management, scheduling and orchestration solutions; customers seeking more options for adding infrastructure as enterprise AI adoption grows; Deloitte being certified to support customers; and customers' and NVIDIA channel partners' ability to upgrade and refresh existing DGX systems 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