NVIDIA Launches Blackwell-Powered DGX SuperPOD for Generative AI Supercomputing at Trillion-Parameter Scale

Scales to Tens of Thousands of Grace Blackwell Superchips Using Most Advanced NVIDIA Networking, NVIDIA Full-Stack AI Software, and Storage Features up to 576 Blackwell GPUs Connected as One With NVIDIA NVLink NVIDIA System Experts Speed Deployment for Immediate AI Infrastructure

GTC—NVIDIA today announced its next-generation AI supercomputer — the NVIDIA DGX SuperPOD™ powered by NVIDIA GB200 Grace Blackwell Superchips — for processing trillion-parameter models with constant uptime for superscale generative AI training and inference workloads.

Featuring a new, highly efficient, liquid-cooled rack-scale architecture, the new DGX SuperPOD is built with NVIDIA DGX™ GB200 systems and provides 11.5 exaflops of AI supercomputing at FP4 precision and 240 terabytes of fast memory — scaling to more with additional racks.

Each DGX GB200 system features 36 NVIDIA GB200 Superchips — which include 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs — connected as one supercomputer via fifth-generation NVIDIA NVLink®. GB200 Superchips deliver up to a 30x performance increase compared to the NVIDIA H100 Tensor Core GPU for large language model inference workloads.

“NVIDIA DGX AI supercomputers are the factories of the AI industrial revolution,” said Jensen Huang, founder and CEO of NVIDIA. “The new DGX SuperPOD combines the latest advancements in NVIDIA accelerated computing, networking and software to enable every company, industry and country to refine and generate their own AI.”

The Grace Blackwell-powered DGX SuperPOD features eight or more DGX GB200 systems and can scale to tens of thousands of GB200 Superchips connected via NVIDIA Quantum InfiniBand. For a massive shared memory space to power next-generation AI models, customers can deploy a configuration that connects the 576 Blackwell GPUs in eight DGX GB200 systems connected via NVLink.

New Rack-Scale DGX SuperPOD Architecture for Era of Generative AI
The new DGX SuperPOD with DGX GB200 systems features a unified compute fabric. In addition to fifth-generation NVIDIA NVLink, the fabric includes NVIDIA BlueField®-3 DPUs and will support NVIDIA Quantum-X800 InfiniBand networking, announced separately today. This architecture provides up to 1,800 gigabytes per second of bandwidth to each GPU in the platform.

Additionally, fourth-generation NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology provides 14.4 teraflops of In-Network Computing, a 4x increase in the next-generation DGX SuperPOD architecture compared to the prior generation.

Turnkey Architecture Pairs With Advanced Software for Unprecedented Uptime
The new DGX SuperPOD is a complete, data-center-scale AI supercomputer that integrates with high-performance storage from NVIDIA-certified partners to meet the demands of generative AI workloads. Each is built, cabled and tested in the factory to dramatically speed deployment at customer data centers.

The Grace Blackwell-powered DGX SuperPOD features intelligent predictive-management capabilities to continuously monitor thousands of data points across hardware and software to predict and intercept sources of downtime and inefficiency — saving time, energy and computing costs.

The software can identify areas of concern and plan for maintenance, flexibly adjust compute resources, and automatically save and resume jobs to prevent downtime, even without system administrators present.

If the software detects that a replacement component is needed, the cluster will activate standby capacity to ensure work finishes in time. Any required hardware replacements can be scheduled to avoid unplanned downtime.

NVIDIA DGX B200 Systems Advance AI Supercomputing for Industries
NVIDIA also unveiled the NVIDIA DGX B200 system, a unified AI supercomputing platform for AI model training, fine-tuning and inference.

DGX B200 is the sixth generation of air-cooled, traditional rack-mounted DGX designs used by industries worldwide. The new Blackwell architecture DGX B200 system includes eight NVIDIA Blackwell GPUs and two 5th Gen Intel® Xeon®

processors. Customers can also build DGX SuperPOD using DGX B200 systems to create AI Centers of Excellence that can power the work of large teams of developers running many different jobs.

DGX B200 systems include the FP4 precision feature in the new Blackwell architecture, providing up to 144 petaflops of AI performance, a massive 1.4TB of GPU memory and 64TB/s of memory bandwidth. This delivers 15x faster real-time inference for trillion-parameter models over the previous generation.

DGX B200 systems include advanced networking with eight NVIDIA ConnectX™-7 NICs and two BlueField-3 DPUs. These provide up to 400 gigabits per second bandwidth per connection — delivering fast AI performance with NVIDIA Quantum-2 InfiniBand and NVIDIA Spectrum™-X Ethernet networking platforms.

Software and Expert Support to Scale Production AI
All NVIDIA DGX platforms include NVIDIA AI Enterprise software for enterprise-grade development and deployment. DGX customers can accelerate their work with the pretrained NVIDIA foundation models, frameworks, toolkits and new NVIDIA NIM microservices included in the software platform.

NVIDIA DGX experts and select NVIDIA partners certified to support DGX platforms assist customers throughout every step of deployment, so they can quickly move AI into production. Once systems are operational, DGX experts continue to support customers in optimizing their AI pipelines and infrastructure.

Availability
NVIDIA DGX SuperPOD with DGX GB200 and DGX B200 systems are expected to be available later this year from NVIDIA's global partners.

For more information, watch a replay of the GTC keynote or visit the NVIDIA booth at GTC, held at the San Jose Convention Center through March 21.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing infrastructure company with data-center-scale offerings that are reshaping industry. 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 NVIDIA's products and technologies, including NVIDIA DGX SuperPOD, NVIDIA GB200 Grace Blackwell Superchips, NVIDIA DGX GB200 systems, NVIDIA GB200 Superchips, NVIDIA Grace CPUs, NVIDIA Blackwell GPUs, NVIDIA NVLink, NVIDIA H100 Tensor Core GPU, NVIDIA BlueField-3 DPUs, NVIDIA Quantum-X800 InfiniBand networking, NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) technology, NVIDIA DGX B200 system, NVIDIA B200 Tensor Core GPUs, NVIDIA ConnectX-7 NICs, NVIDIA Quantum-2 InfiniBand, NVIDIA Spectrum-X Ethernet, NVIDIA AI Enterprise software, and NVIDIA NIM; the new DGX SuperPOD enabling every company, industry and country to refine and generate their own AI; and third parties' use or adoption of our products, platforms and technologies and the benefits and impacts thereof 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.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

Allie Courtney

NVIDIA Corporation
+1-408-706-8995
acourtney@nvidia.com