# NVIDIA Announces DGX GH200 AI Supercomputer

## New Class of AI Supercomputer Connects 256 Grace Hopper Superchips Into Massive, 1-Exaflop, 144TB GPU for Giant Models Powering Generative AI, Recommender Systems, Data Processing

**COMPUTEX—**NVIDIA today announced a new class of large-memory AI supercomputer — an NVIDIA DGX™ supercomputer powered by NVIDIA® GH200 Grace Hopper Superchips and the NVIDIA NVLink® Switch System — created to enable the development of giant, next-generation models for generative AI language applications, recommender systems and data analytics workloads.

The NVIDIA DGX GH200's massive shared memory space uses NVLink interconnect technology with the NVLink Switch System to combine 256 GH200 superchips, allowing them to perform as a single GPU. This provides 1 exaflop of performance and 144 terabytes of shared memory — nearly 500x more memory than the previous generation NVIDIA DGX A100, which was introduced in 2020.

"Generative AI, large language models and recommender systems are the digital engines of the modern economy," said Jensen Huang, founder and CEO of NVIDIA. "DGX GH200 AI supercomputers integrate NVIDIA's most advanced accelerated computing and networking technologies to expand the frontier of AI."

**NVIDIA NVLink Technology Expands AI at Scale**
GH200 superchips eliminate the need for a traditional CPU-to-GPU PCIe connection by combining an Arm-based NVIDIA Grace™ CPU with an NVIDIA H100 Tensor Core GPU in the same package, using NVIDIA NVLink-C2C chip interconnects. This increases the bandwidth between GPU and CPU by 7x compared with the latest PCIe technology, slashes interconnect power consumption by more than 5x, and provides a 600GB Hopper architecture GPU building block for DGX GH200 supercomputers.

DGX GH200 is the first supercomputer to pair Grace Hopper Superchips with the NVIDIA NVLink Switch System, a new interconnect that enables all GPUs in a DGX GH200 system to work together as one. The previous-generation system only provided for eight GPUs to be combined with NVLink as one GPU without compromising performance.

The DGX GH200 architecture provides 48x more NVLink bandwidth than the previous generation, delivering the power of a massive AI supercomputer with the simplicity of programming a single GPU.

**A New Research Tool for AI Pioneers**
Google Cloud, Meta and Microsoft are among the first expected to gain access to the DGX GH200 to explore its capabilities for generative AI workloads. NVIDIA also intends to provide the DGX GH200 design as a blueprint to cloud service providers and other hyperscalers so they can further customize it for their infrastructure.

"Building advanced generative models requires innovative approaches to AI infrastructure," said Mark Lohmeyer, vice president of Compute at Google Cloud. "The new NVLink scale and shared memory of Grace Hopper Superchips address key bottlenecks in large-scale AI and we look forward to exploring its capabilities for Google Cloud and our generative AI initiatives."

"As AI models grow larger, they need powerful infrastructure that can scale to meet increasing demands," said Alexis Björlin, vice president of Infrastructure, AI Systems and Accelerated Platforms at Meta. "NVIDIA's Grace Hopper design looks to provide researchers with the ability to explore new approaches to solve their greatest challenges."

"Training large AI models is traditionally a resource- and time-intensive task," said Girish Bablani, corporate vice president of Azure Infrastructure at Microsoft. "The potential for DGX GH200 to work with terabyte-sized datasets would allow developers to conduct advanced research at a larger scale and accelerated speeds."

**New NVIDIA Helios Supercomputer to Advance Research and Development**
NVIDIA is building its own DGX GH200-based AI supercomputer to power the work of its researchers and development teams.

Named NVIDIA Helios, the supercomputer will feature four DGX GH200 systems. Each will be interconnected with NVIDIA Quantum-2 InfiniBand networking to supercharge data throughput for training large AI models. Helios will include 1,024 Grace Hopper Superchips and is expected to come online by the end of the year.

**Fully Integrated and Purpose-Built for Giant Models**
DGX GH200 supercomputers include NVIDIA software to provide a turnkey, full-stack solution for the largest AI and data analytics workloads. NVIDIA Base Command™ software provides AI workflow management, enterprise-grade cluster

management, libraries that accelerate compute, storage and network infrastructure, and system software optimized for running AI workloads.

Also included is NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform. It provides over 100 frameworks, pretrained models and development tools to streamline development and deployment of production AI including generative AI, computer vision, speech AI and more.

**Availability**
NVIDIA DGX GH200 supercomputers are expected to be available by the end of the year.

Watch Huang discuss NVIDIA DGX GH200 supercomputers during his keynote address at COMPUTEX.

**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 the creation of the industrial metaverse. NVIDIA is now a full-stack computing 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 our products, services and technologies, including NVIDIA Grace Hopper Superchips and supercomputer, NVIDIA DGX and DGX GH200, NVLink including the NVLink Switch System and NVLink interconnect technology, DGX H100, NVIDIA Grace CPU, H100 Tensor Core GPU, Helios supercomputer, Quantum-2 InfiniBand, NVIDIA Base Command and NVIDIA AI Enterprise; our collaborations with Google Cloud, Meta and Microsoft and the benefits, impact, performance, features and availability thereof; generative AI, recommender systems and data analytics being engines of the modern economy, requiring unprecedented scale, speed and efficiency; and NVIDIA's intention to provide the DGX GH200 design as a blueprint to cloud service providers and other hyperscalers so they can further customize it for their infrastructure 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
NVIDIA Corporation
+1-310-920-9642
smcphee@nvidia.com