NVIDIA Unveils Next-Generation GH200 Grace Hopper Superchip Platform for Era of Accelerated Computing and Generative AI

World's First HBM3e Processor Offers Groundbreaking Memory, Bandwidth; Ability to Connect Multiple GPUs for Exceptional Performance; Easily Scalable Server Design

SIGGRAPH—NVIDIA today announced the next-generation NVIDIA GH200 Grace Hopper™ platform — based on a new Grace Hopper Superchip with the world's first HBM3e processor — built for the era of accelerated computing and generative AI.

Created to handle the world's most complex generative AI workloads, spanning large language models, recommender systems and vector databases, the new platform will be available in a wide range of configurations.

The dual configuration — which delivers up to 3.5x more memory capacity and 3x more bandwidth than the current generation offering — comprises a single server with 144 Arm Neoverse cores, eight petaflops of AI performance and 282GB of the latest HBM3e memory technology.

"To meet surging demand for generative AI, data centers require accelerated computing platforms with specialized needs," said Jensen Huang, founder and CEO of NVIDIA. "The new GH200 Grace Hopper Superchip platform delivers this with exceptional memory technology and bandwidth to improve throughput, the ability to connect GPUs to aggregate performance without compromise, and a server design that can be easily deployed across the entire data center."

The new platform uses the Grace Hopper Superchip, which can be connected with additional Superchips by NVIDIA NVLink™, allowing them to work together to deploy the giant models used for generative AI. This high-speed, coherent technology gives the GPU full access to the CPU memory, providing a combined 1.2TB of fast memory when in dual configuration.

HBM3e memory, which is 50% faster than current HBM3, delivers a total of 10TB/sec of combined bandwidth, allowing the new platform to run models 3.5x larger than the previous version, while improving performance with 3x faster memory bandwidth.

Growing Demand for Grace Hopper
Leading manufacturers are already offering systems based on the previously announced Grace Hopper Superchip. To drive broad adoption of the technology, the next-generation Grace Hopper Superchip platform with HBM3e is fully compatible with the NVIDIA MGX™ server specification unveiled at COMPUTEX earlier this year. With MGX, any system manufacturer can quickly and cost-effectively add Grace Hopper into over 100 server variations.

Availability
Leading system manufacturers are expected to deliver systems based on the platform in Q2 of calendar year 2024.

Watch Huang's SIGGRAPH keynote address on demand to learn more about Grace Hopper.

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 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 the NVIDIA GH200 Grace Hopper platform, Grace Hopper Superchip, NVIDIA NVLink and NVIDIA MGX; surging demand for generative AI; data centers requiring accelerated computing platforms with specialized needs; and leading system manufacturers delivering systems based on GH200 Grace Hopper Superchip platform in Q2 of calendar year 2024 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com