NVIDIA Supercharges Hopper, the World's Leading AI Computing Platform

HGX H200 Systems and Cloud Instances Coming Soon From World's Top Server Manufacturers and Cloud Service Providers

SC23—NVIDIA today announced it has supercharged the world's leading AI computing platform with the introduction of the NVIDIA HGX™ H200. Based on NVIDIA Hopper™ architecture, the platform features the NVIDIA H200 Tensor Core GPU with advanced memory to handle massive amounts of data for generative AI and high performance computing workloads.

The NVIDIA H200 is the first GPU to offer HBM3e — faster, larger memory to fuel the acceleration of generative AI and large language models, while advancing scientific computing for HPC workloads. With HBM3e, the NVIDIA H200 delivers 141GB of memory at 4.8 terabytes per second, nearly double the capacity and 2.4x more bandwidth compared with its predecessor, the NVIDIA A100.

H200-powered systems from the world's leading server manufacturers and cloud service providers are expected to begin shipping in the second quarter of 2024.

"To create intelligence with generative AI and HPC applications, vast amounts of data must be efficiently processed at high speed using large, fast GPU memory," said Ian Buck, vice president of hyperscale and HPC at NVIDIA. "With NVIDIA H200, the industry's leading end-to-end AI supercomputing platform just got faster to solve some of the world's most important challenges."

Perpetual Innovation, Perpetual Performance Leaps
The NVIDIA Hopper architecture delivers an unprecedented performance leap over its predecessor and continues to raise the bar through ongoing software enhancements with H100, including the recent release of powerful open-source libraries like NVIDIA TensorRT™-LLM.

The introduction of H200 will lead to further performance leaps, including nearly doubling inference speed on Llama 2, a 70 billion-parameter LLM, compared to the H100. Additional performance leadership and improvements with H200 are expected with future software updates.

NVIDIA H200 Form Factors
NVIDIA H200 will be available in NVIDIA HGX H200 server boards with four- and eight-way configurations, which are compatible with both the hardware and software of HGX H100 systems. It is also available in the NVIDIA GH200 Grace Hopper™ Superchip with HBM3e, announced in August.

With these options, H200 can be deployed in every type of data center, including on premises, cloud, hybrid-cloud and edge. NVIDIA's global ecosystem of partner server makers — including ASRock Rack, ASUS, Dell Technologies, Eviden, GIGABYTE, Hewlett Packard Enterprise, Ingrasys, Lenovo, QCT, Supermicro, Wistron and Wiwynn — can update their existing systems with an H200.

Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure will be among the first cloud service providers to deploy H200-based instances starting next year, in addition to CoreWeave, Lambda and Vultr.

Powered by NVIDIA NVLink™ and NVSwitch™ high-speed interconnects, HGX H200 provides the highest performance on various application workloads, including LLM training and inference for the largest models beyond 175 billion parameters.

An eight-way HGX H200 provides over 32 petaflops of FP8 deep learning compute and 1.1TB of aggregate high-bandwidth memory for the highest performance in generative AI and HPC applications.

When paired with NVIDIA Grace™ CPUs with an ultra-fast NVLink-C2C interconnect, the H200 creates the GH200 Grace Hopper Superchip with HBM3e — an integrated module designed to serve giant-scale HPC and AI applications.

Accelerate AI With NVIDIA Full-Stack Software
NVIDIA's accelerated computing platform is supported by powerful software tools that enable developers and enterprises to build and accelerate production-ready applications from AI to HPC. This includes the NVIDIA AI Enterprise suite of software for workloads such as speech, recommender systems and hyperscale inference.

Availability
The NVIDIA H200 will be available from global system manufacturers and cloud service providers starting in the second quarter of 2024.

Watch Buck's SC23 special address on Nov. 13 at 6 a.m. PT to learn more about the NVIDIA H200 Tensor Core GPU.

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, performance, specifications, impact and availability of the NVIDIA HGX H200 and NVIDIA Hopper architecture; the processing requirements to create intelligence with generative AI and HPC applications; the ease for partner server makers to update H100-based systems with H200; and the first cloud service providers expected to deploy H200-based instances 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