NVIDIA Hopper in Full Production

World’s Leading Computer Makers Dell Technologies, HPE, Lenovo, Supermicro, Plus Cloud Service Providers AWS, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure Building H100-Based Offerings; Availability Begins Next Month

GTC—NVIDIA today announced that the NVIDIA H100 Tensor Core GPU is in full production, with global tech partners planning in October to roll out the first wave of products and services based on the groundbreaking NVIDIA Hopper™ architecture.

Unveiled in April, H100 is built with 80 billion transistors and benefits from a range of technology breakthroughs. Among them are the powerful new Transformer Engine and an NVIDIA NVLink® interconnect to accelerate the largest AI models, like advanced recommender systems and large language models, and to drive innovations in such fields as conversational AI and drug discovery.

“Hopper is the new engine of AI factories, processing and refining mountains of data to train models with trillions of parameters that are used to drive advances in language-based AI, robotics, healthcare and life sciences,” said Jensen Huang, founder and CEO of NVIDIA. “Hopper’s Transformer Engine boosts performance up to an order of magnitude, putting large-scale AI and HPC within reach of companies and researchers.”

In addition to Hopper’s architecture and Transformer Engine, several other key innovations power the H100 GPU to deliver the next massive leap in NVIDIA’s accelerated compute data center platform, including second-generation Multi-Instance GPU, confidential computing, fourth-generation NVIDIA NVLink and DPX Instructions.

A five-year license for the NVIDIA AI Enterprise software suite is now included with H100 for mainstream servers. This optimizes the development and deployment of AI workflows and ensures organizations have access to the AI frameworks and tools needed to build AI chatbots, recommendation engines, vision AI and more.

Global Rollout of Hopper

H100 enables companies to slash costs for deploying AI, delivering the same AI performance with 3.5x more energy efficiency and 3x lower total cost of ownership, while using 5x fewer server nodes over the previous generation.

For customers who want to immediately try the new technology, NVIDIA announced that H100 on Dell PowerEdge servers is now available on NVIDIA LaunchPad, which provides free hands-on labs, giving companies access to the latest hardware and NVIDIA AI software.

Customers can also begin ordering NVIDIA DGX™ H100 systems, which include eight H100 GPUs and deliver 32 petaflops of performance at FP8 precision. NVIDIA Base Command™ and NVIDIA AI Enterprise software power every DGX system, enabling deployments from a single node to an NVIDIA DGX SuperPOD™ supporting advanced AI development of large language models and other massive workloads.

H100-powered systems from the world’s leading computer makers are expected to ship in the coming weeks, with over 50 server models in the market by the end of the year and dozens more in the first half of 2023. Partners building systems include Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro.

Additionally, some of the world’s leading higher education and research institutions will be using H100 to power their next-generation supercomputers. Among them are the Barcelona Supercomputing Center, Los Alamos National Lab, Swiss National Supercomputing Centre (CSCS), Texas Advanced Computing Center and the University of Tsukuba.

H100 Coming to the Cloud

Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure will be among the first to deploy H100-based instances in the cloud starting next year.

“We look forward to enabling the next generation of AI models on the latest H100 GPUs in Microsoft Azure,” said Nidhi Chappell, general manager of Azure AI Infrastructure. “With the advancements in Hopper architecture coupled with our investments in Azure AI supercomputing, we’ll be able to help accelerate the development of AI worldwide”

“By offering our customers the latest H100 GPUs from NVIDIA, we’re helping them accelerate their most complex machine learning and HPC workloads,” said Karan Batta, vice president of product management at Oracle Cloud Infrastructure. “Additionally, using NVIDIA’s next generation of H100 GPUs allows us to support our demanding internal workloads and helps our mutual customers with breakthroughs across healthcare, autonomous vehicles, robotics and IoT.”

NVIDIA Software Support

The advanced Transformer Engine technology of H100 enables enterprises to quickly develop large language models with a higher level of accuracy. As these models continue to grow in scale, so does the complexity, sometimes requiring months to train.

To tackle this, some of the world's leading large language model and deep learning frameworks are being optimized on H100, including NVIDIA NeMo Megatron, Microsoft DeepSpeed, Google JAX, PyTorch, TensorFlow and XLA. These frameworks combined with Hopper architecture will significantly speed up AI performance to help train large language models within days or hours.

To learn more about NVIDIA Hopper and H100, watch Huang's GTC keynote. Register for GTC for free to attend sessions with NVIDIA and industry leaders.

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 and ignited the era of modern AI. 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, specifications, performance, features and availability of our products and technologies, including NVIDIA H100 Tensor Core GPUs, NVIDIA Hopper architecture, NVIDIA AI Enterprise software suite, NVIDIA LaunchPad, NVIDIA DGX H100 systems, NVIDIA Base Command, NVIDIA DGX SuperPOD and NVIDIA-Certified Systems; a range of the world's leading computer makers, cloud service providers, higher education and research institutions and large language model and deep learning frameworks adopting the H100 GPUs; the software support for NVIDIA H100; large language models continuing to grow in scale; and the performance of large language model and deep learning frameworks combined with NVIDIA Hopper architecture 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