NVIDIA Announces Hopper Architecture, the Next Generation of Accelerated Computing

The New Engine for World's AI Infrastructure, NVIDIA H100 GPU Makes Order of Magnitude Performance Leap

GTC—To power the next wave of AI data centers, NVIDIA today announced its next-generation accelerated computing platform with NVIDIA Hopper™ architecture, delivering an order of magnitude performance leap over its predecessor.

Named for Grace Hopper, a pioneering U.S. computer scientist, the new architecture succeeds the NVIDIA Ampere architecture, launched two years ago.

The company also announced its first Hopper-based GPU, the NVIDIA H100, packed with 80 billion transistors. The world's largest and most powerful accelerator, the H100 has groundbreaking features such as a revolutionary Transformer Engine and a highly scalable NVIDIA NVLink® interconnect for advancing gigantic AI language models, deep recommender systems, genomics and complex digital twins.

"Data centers are becoming AI factories -- processing and refining mountains of data to produce intelligence," said Jensen Huang, founder and CEO of NVIDIA. "NVIDIA H100 is the engine of the world's AI infrastructure that enterprises use to accelerate their AI-driven businesses."

H100 Technology Breakthroughs
The NVIDIA H100 GPU sets a new standard in accelerating large-scale AI and HPC, delivering six breakthrough innovations:

- World's Most Advanced Chip -- Built with 80 billion transistors using a cutting-edge TSMC 4N process designed for NVIDIA's accelerated compute needs, H100 features major advances to accelerate AI, HPC, memory bandwidth, interconnect and communication, including nearly 5 terabytes per second of external connectivity. H100 is the first GPU to support PCIe Gen5 and the first to utilize HBM3, enabling 3TB/s of memory bandwidth. Twenty H100 GPUs can sustain the equivalent of the entire world's internet traffic, making it possible for customers to deliver advanced recommender systems and large language models running inference on data in real time.
- New Transformer Engine -- Now the standard model choice for natural language processing, the Transformer is one of the most important deep learning models ever invented. The H100 accelerator's Transformer Engine is built to speed up these networks as much as 6x versus the previous generation without losing accuracy.
- 2nd-Generation Secure Multi-Instance GPU -- MIG technology allows a single GPU to be partitioned into seven smaller, fully isolated instances to handle different types of jobs. The Hopper architecture extends MIG capabilities by up to 7x over the previous generation by offering secure multitenant configurations in cloud environments across each GPU instance.
- Confidential Computing -- H100 is the world's first accelerator with confidential computing capabilities to protect AI models and customer data while they are being processed. Customers can also apply confidential computing to federated learning for privacy-sensitive industries like healthcare and financial services, as well as on shared cloud infrastructures.
- 4th-Generation NVIDIA NVLink -- To accelerate the largest AI models, NVLink combines with a new external NVLink Switch to extend NVLink as a scale-up network beyond the server, connecting up to 256 H100 GPUs at 9x higher bandwidth versus the previous generation using NVIDIA HDR Quantum InfiniBand.
- DPX Instructions -- New DPX instructions accelerate dynamic programming -- used in a broad range of algorithms, including route optimization and genomics -- by up to 40x compared with CPUs and up to 7x compared with previous-generation GPUs. This includes the Floyd-Warshall algorithm to find optimal routes for autonomous robot fleets in dynamic warehouse environments, and the Smith-Waterman algorithm used in sequence alignment for DNA and protein classification and folding.

The combined technology innovations of H100 extend NVIDIA's AI inference and training leadership to enable real-time and immersive applications using giant-scale AI models. The H100 will enable chatbots using the world's most powerful monolithic transformer language model, Megatron 530B, with up to 30x higher throughput than the previous generation, while meeting the subsecond latency required for real-time conversational AI. H100 also allows researchers and developers to train massive models such as Mixture of Experts, with 395 billion parameters, up to 9x faster, reducing the training time from weeks to days.

Broad NVIDIA H100 Adoption
NVIDIA H100 can be deployed in every type of data center, including on-premises, cloud, hybrid-cloud and edge. It is expected to be available worldwide later this year from the world's leading cloud service providers and computer makers, as

well as directly from NVIDIA.

NVIDIA's fourth-generation DGX™ system, DGX H100, features eight H100 GPUs to deliver 32 petaflops of AI performance at new FP8 precision, providing the scale to meet the massive compute requirements of large language models, recommender systems, healthcare research and climate science.

Every GPU in DGX H100 systems is connected by fourth-generation NVLink, providing 900GB/s connectivity, 1.5x more than the prior generation. NVSwitch™ enables all eight of the H100 GPUs to connect over NVLink. An external NVLink Switch can network up to 32 DGX H100 nodes in the next-generation NVIDIA DGX SuperPOD™ supercomputers.

Hopper has received broad industry support from leading cloud service providers Alibaba Cloud, Amazon Web Services, Baidu AI Cloud, Google Cloud, Microsoft Azure, Oracle Cloud and Tencent Cloud, which plan to offer H100-based instances.

A wide range of servers with H100 accelerators are expected from the world's leading systems manufacturers, including Atos, BOXX Technologies, Cisco, Dell Technologies, Fujitsu, GIGABYTE, H3C, Hewlett Packard Enterprise, Inspur, Lenovo, Nettrix and Supermicro.

NVIDIA H100 at Every Scale
H100 will come in SXM and PCIe form factors to support a wide range of server design requirements. A converged accelerator will also be available, pairing an H100 GPU with an NVIDIA ConnectX®-7 400Gb/s InfiniBand and Ethernet SmartNIC.

NVIDIA's H100 SXM will be available in HGX™ H100 server boards with four- and eight-way configurations for enterprises with applications scaling to multiple GPUs in a server and across multiple servers. HGX H100-based servers deliver the highest application performance for AI training and inference along with data analytics and HPC applications.

The H100 PCIe, with NVLink to connect two GPUs, provides more than 7x the bandwidth of PCIe 5.0, delivering outstanding performance for applications running on mainstream enterprise servers. Its form factor makes it easy to integrate into existing data center infrastructure.

The H100 CNX, a new converged accelerator, couples an H100 with a ConnectX-7 SmartNIC to provide groundbreaking performance for I/O-intensive applications such as multinode AI training in enterprise data centers and 5G signal processing at the edge.

NVIDIA Hopper architecture-based GPUs can also be paired with NVIDIA Grace™ CPUs with an ultra-fast NVLink-C2C interconnect for over 7x faster communication between the CPU and GPU compared to PCIe 5.0. This combination -- the Grace Hopper Superchip -- is an integrated module designed to serve giant-scale HPC and AI applications.

NVIDIA Software Support
The NVIDIA H100 GPU is supported by powerful software tools that enable developers and enterprises to build and accelerate applications from AI to HPC. This includes major updates to the NVIDIA AI suite of software for workloads such as speech, recommender systems and hyperscale inference.

NVIDIA also released more than 60 updates to its CUDA-X™ collection of libraries, tools and technologies to accelerate work in quantum computing and 6G research, cybersecurity, genomics and drug discovery.

Availability
NVIDIA H100 will be available starting in the third quarter.

To learn more about NVIDIA Hopper and H100, watch the GTC 2022 keynote from Jensen Huang, and register for GTC 2022 for free to attend sessions with NVIDIA and industry leaders.

About NVIDIA
NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing, and artificial intelligence. The company's pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. 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 and availability of the NVIDIA Hopper architecture, the NVIDIA H100 GPU and DGX H100; data centers becoming AI factories; NVIDIA H100 providing the next engine for the world's data centers that will enable enterprises to use AI to transform their businesses; the ability for customers to apply confidential computing to federated learning; NVIDIA H100's ability to be deployed in every type of data center; and the software support for NVIDIA H100 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