# NVIDIA Hopper GPUs Expand Reach as Demand for AI Grows

## NVIDIA H100 GPUs Now Being Offered by Cloud Giants to Meet Surging Demand for Generative AI Training and Inference; Meta, OpenAI, Stability AI to Leverage H100 for Next Wave of AI

**GTC—**NVIDIA and key partners today announced the availability of new products and services featuring the NVIDIA H100 Tensor Core GPU — the world's most powerful GPU for AI — to address rapidly growing demand for generative AI training and inference.

Oracle Cloud Infrastructure (OCI) announced the limited availability of new OCI Compute bare-metal GPU instances featuring H100 GPUs. Additionally, Amazon Web Services announced its forthcoming EC2 UltraClusters of Amazon EC2 P5 instances, which can scale in size up to 20,000 interconnected H100 GPUs. This follows Microsoft Azure's private preview announcement last week for its H100 virtual machine, ND H100 v5.

Additionally, Meta has now deployed its H100-powered Grand Teton AI supercomputer internally for its AI production and research teams.

NVIDIA founder and CEO Jensen Huang announced during his GTC keynote today that NVIDIA DGX™ H100 AI supercomputers are in full production and will be coming soon to enterprises worldwide.

"Generative AI's incredible potential is inspiring virtually every industry to reimagine its business strategies and the technology required to achieve them," said Huang. "NVIDIA and our partners are moving fast to provide the world's most powerful AI computing platform to those building applications that will fundamentally transform how we live, work and play."

**Hopper Architecture Accelerates AI**
The H100, based on the NVIDIA Hopper™ GPU computing architecture with its built-in Transformer Engine, is optimized for developing, training and deploying generative AI, large language models (LLMs) and recommender systems. This technology makes use of the H100's FP8 precision and offers 9x faster AI training and up to 30x faster AI inference on LLMs versus the prior-generation A100. The H100 began shipping in the fall in individual and select board units from global manufacturers.

The NVIDIA DGX H100 features eight H100 GPUs connected with NVIDIA NVLink® high-speed interconnects and integrated NVIDIA Quantum InfiniBand and Spectrum™ Ethernet networking. This platform provides 32 petaflops of compute performance at FP8 precision, with 2x faster networking than the prior generation, helping maximize energy efficiency in processing large AI workloads.

DGX H100 also features the complete NVIDIA AI software stack, enabling enterprises to seamlessly run and manage their AI workloads at scale. This offering includes the latest version of NVIDIA AI Enterprise, announced separately today, as well as NVIDIA Base Command™, the operating system of the DGX data center, which coordinates AI training and operations across the NVIDIA DGX platform to simplify and streamline AI development.

**AI Pioneers Adopt H100**
Several pioneers in generative AI are adopting H100 to accelerate their work:

- **OpenAI** used H100's predecessor — NVIDIA A100 GPUs — to train and run ChatGPT, an AI system optimized for dialogue, which has been used by hundreds of millions of people worldwide in record time. OpenAI will be using H100 on its Azure supercomputer to power its continuing AI research.
- **Meta,** a key technology partner of NVIDIA, developed its Hopper-based AI supercomputer Grand Teton system with multiple performance enhancements over its predecessor, Zion, including 4x the host-to-GPU bandwidth, 2x the compute and data network bandwidth, and 2x the power envelope. With this greater compute capacity, Grand Teton can support both the training and production inference of deep learning recommender models and content understanding.
- **Stability AI,** a pioneer in text-to-image generative AI, is an H100 early access customer on AWS. Stability AI plans to use H100 to accelerate its upcoming video, 3D and multimodal models.
- **Twelve Labs**, a platform that gives businesses and developers access to multimodal video understanding, plans to use H100 instances on an OCI Supercluster to make video instantly, intelligently and easily searchable.
- **Anlatan**, the creator of the NovelAI app for AI-assisted story writing and text-to-image synthesis, is using H100 instances on CoreWeave's cloud platform for model creation and inference.

**DGX H100 Around the World**
Innovators worldwide are receiving the first wave of DGX H100 systems, including:

- **CyberAgent**, a leading digital advertising and internet services company based in Japan, is creating AI-produced digital ads and celebrity digital twin avatars, fully using generative AI and LLM technologies.
- **Johns Hopkins University Applied Physics Laboratory**, the U.S.'s largest university-affiliated research center, will use DGX H100 for training LLMs.
- **KTH Royal Institute of Technology**, a leading European technical and engineering university based in Stockholm, will use DGX H100 to provide state-of-the-art computer science programs for higher education.
- **Mitsui**, one of Japan's leading business groups, which has a wide variety of businesses in fields such as energy, wellness, IT and communication, is building Japan's first generative AI supercomputer for drug discovery, powered by DGX H100.
- **Telconet**, a leading telecommunications provider in Ecuador, is building intelligent video analytics for safe cities and language services to support customers across Spanish dialects.

**Ecosystem Support**
"We are fully focused on AI innovation and AI-first products. NVIDIA H100 GPUs are state-of-the-art machine learning accelerators, giving us a significant competitive advantage within the machine learning industry for a wide variety of applications from model training to model inference." — *Eren Doğan, CEO of Anlatan*

"AWS and NVIDIA have collaborated for more than 12 years to deliver large-scale, cost-effective GPU-based solutions on demand. AWS has unmatched experience delivering GPU-based instances that push the scalability envelope with each successive generation. Today, many customers scale machine learning training workloads to more than 10,000 GPUs. With second-generation EFA, customers can scale their P5 instances to more than 20,000 H100 GPUs, bringing on-demand supercomputer capabilities to any organization." — *David Brown, vice president of Amazon EC2 at AWS*

"AI is at the core of everything we do at Google Cloud. NVIDIA H100 GPU and its powerful capabilities, coupled with our industry-leading AI products and services, will enable our customers to break new ground. We are excited to work with NVIDIA to accelerate enterprises in their effort to tap the power of generative AI." — *Amin Vahdat, vice president of Systems & Services Infrastructure at Google Cloud*

"As we build new AI-powered experiences — like those based on generative AI — the underlying AI models become increasingly more sophisticated. Meta's latest H100-powered Grand Teton AI supercomputer brings greater compute, memory capacity and bandwidth, further accelerating training and inference of Meta's AI models, such as the open-sourced DLRM. As we move into the next computing platform, H100 also provides greater compute capabilities for researching Meta's future content recommendation, generative AI and metaverse needs." — *Alexis Bjorlin, vice president of Infrastructure, AI Systems and Accelerated Platforms at Meta*

"As the adoption of AI continues to accelerate, the way businesses operate and succeed is fundamentally changing. By bringing NVIDIA's Hopper architecture to Microsoft Azure, we are able to offer unparalleled computing performance and functionality to enterprises looking to scale their AI capabilities." — *Scott Guthrie, executive vice president of the Cloud + AI group at Microsoft*

"The computational power of the NVIDIA H100 Tensor Core GPU will be vital for enabling our efforts to push the frontier of AI training and inference. NVIDIA's advancements unlock our research and alignment work on systems like GPT-4." — *Greg Brockman, president and co-founder of OpenAI*

"OCI is bringing AI supercomputing capabilities at scale to thousands of organizations of all sizes. Our strong collaboration with NVIDIA is providing great value to customers, and we're excited by the power of H100." — *Greg Pavlik, CTO and senior vice president at Oracle Cloud Infrastructure*

"As the world's leading open-source generative AI model company, Stability AI is committed to providing consumers and enterprises with the world's best tools for multimodal creation. Harnessing the power of the NVIDIA H100 provides unprecedented computing power to fuel the creativity and research capabilities of the surging numbers of those looking to benefit from the transformative powers of generative AI. It will unlock our video, 3D and other models that uniquely benefit from the higher interconnect and advanced architecture for exabytes of data." — *Emad Mostaque, founder and CEO of Stability AI*

"Twelve Labs is excited to leverage Oracle Cloud Infrastructure Compute bare-metal instances powered by NVIDIA H100 GPUs to continue leading the effort in bringing video foundation models to market." — *Jae Lee, CEO of Twelve Labs*

**Availability**
NVIDIA DGX H100 supercomputers are in full production and orderable from NVIDIA partners worldwide. Customers can trial DGX H100 today with NVIDIA DGX Cloud. Pricing is available from NVIDIA DGX partners worldwide.

NVIDIA H100 in the cloud is available now from Azure in private preview, Oracle Cloud Infrastructure in limited availability, and generally available from Cirrascale and CoreWeave. AWS announced H100 will be available in the coming weeks in limited preview. Google Cloud along with NVIDIA's cloud partners Lambda, Paperspace and Vultr plan to offer H100.

Servers and systems featuring NVIDIA H100 GPUs are available from leading server makers including Atos, Cisco, Dell

Technologies, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro.

Pricing and other details are available directly from NVIDIA partners.

Watch Huang discuss the NVIDIA Hopper architecture in his GTC keynote.

**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 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, collaborations, partnerships and technologies, including Hopper GPUs, H100 Tensor Core GPUs, DGX H100, A100, NVLink high-speed interconnects, Quantum InfiniBand, Spectrum Ethernet, NVIDIA AI software stack, NVIDIA AI Enterprise, NVIDIA Base Command, and the DGX platform including DGX Cloud; NVIDIA DGX H100 AI supercomputers being in full production and coming soon to enterprises worldwide; innovators worldwide receiving the first wave of DGX H100; and Mitsui building the world's first generative AI supercomputer for drug discovery 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.



Allie Courtney
NVIDIA Corporation
+1-408-706-8995
acourtney@nvidia.com