Now Shipping: DGX H100 Systems Bring Advanced AI Capabilities to Industries Worldwide

Customers from Tokyo to Stockholm will plug into NVIDIA's latest AI supercomputers to advance workloads that include generative AI across manufacturing, healthcare, robotics and more.

Author: Tony Paikeday

Customers from Japan to Ecuador and Sweden are using NVIDIA DGX H100 systems like AI factories to manufacture intelligence.

They're creating services that offer AI-driven insights in finance, healthcare, law, IT and telecom — and working to transform their industries in the process.

Among the dozens of use cases, one aims to predict how factory equipment will age, so tomorrow's plants can be more efficient.

Called Green Physics AI , it adds information like an object's CO2 footprint, age and energy consumption to SORDI.ai , which claims to be the largest synthetic dataset in manufacturing.

The dataset lets manufacturers develop powerful AI models and create digital twins that optimize the efficiency of factories and warehouses. With Green Physics AI, they also can optimize energy and CO2 savings for the factory's products and the components that go into them.

Imagine a robot that could watch you wash dishes or change the oil in your car, then do it for you.

Boston Dynamics AI Institute (The AI Institute), a research organization which traces its roots to Boston Dynamics, the well-known pioneer in robotics, will use a DGX H100 to pursue that vision. Researchers imagine dexterous mobile robots helping people in factories, warehouses, disaster sites and eventually homes.

"One thing I've dreamed about since I was in grad school is a robot valet who can follow me and do useful tasks — everyone should have one," said Al Rizzi, CTO of The AI Institute.

That will require breakthroughs in AI and robotics, something Rizzi has seen firsthand. As chief scientist at Boston Dynamics, he helped create robots like Spot, a quadruped that can navigate stairs and even open doors for itself.

Initially, the DGX H100 will tackle tasks in reinforcement learning , a key technique in robotics. Later, it will run AI inference jobs while connected directly to prototype bots in the lab.

"It's an extremely high-performance computer in a relatively compact footprint, so it provides an easy way for us to develop and deploy AI models," said Rizzi.

You don't have to be a world-class research outfit or Fortune 500 company to use a DGX H100. Startups are unboxing some of the first systems to ride the wave of generative AI .

For example, Scissero , with offices in London and New York, employs a GPT-powered chatbot to make legal processes more efficient. Its Scissero GPT can draft legal documents, generate reports and conduct legal research.

In Germany, DeepL will use several DGX H100 systems to expand services like translation between dozens of languages it provides for customers, including Nikkei, Japan's largest publishing company. DeepL recently released an AI writing assistant called DeepL Write.

Many of the DGX H100 systems will advance healthcare and improve patient outcomes.

In Tokyo, DGX H100s will run simulations and AI to speed the drug discovery process as part of the Tokyo-1 supercomputer . Xeureka — a startup launched in November 2021 by Mitsui & Co. Ltd., one of Japan’s largest conglomerates — will manage the system.

Separately, hospitals and academic healthcare organizations in Germany, Israel and the U.S. will be among the first users of DGX H100 systems.

Universities from Singapore to Sweden are plugging in DGX H100 systems for research across a range of fields.

A DGX H100 will train large language models for Johns Hopkins University Applied Physics Laboratory . The KTH Royal Institute of Technology in Sweden will use one to provide state-of-the-art computer science programs for higher education.

Among other use cases, Japan’s CyberAgent, an internet services company, is creating smart digital ads and celebrity avatars. Telconet, a leading telecommunications provider in Ecuador, is building intelligent video analytics for safe cities and language services to support customers across Spanish dialects.

Each NVIDIA H100 Tensor Core GPU in a DGX H100 system provides on average about 6x more performance than prior GPUs. A DGX H100 packs eight of them, each with a Transformer Engine designed to accelerate generative AI models.

The eight H100 GPUs connect over NVIDIA NVLink to create one giant GPU. Scaling doesn’t stop there: organizations can connect hundreds of DGX H100 nodes into an AI supercomputer using the 400 Gbps ultra-low latency NVIDIA Quantum InfiniBand , twice the speed of prior networks.

DGX H100 systems run on NVIDIA Base Command , a suite for accelerating compute, storage, and network infrastructure and optimizing AI workloads.

They also include NVIDIA AI Enterprise , software to accelerate data science pipelines and streamline development and deployment of generative AI, computer vision and more.

The DGX platform offers both high performance and efficiency. DGX H100 delivers a 2x improvement in kilowatts per petaflop over the DGX A100 generation.

NVIDIA DGX H100 systems, DGX PODs and DGX SuperPODs are available from NVIDIA’s global partners .

Manuvir Das, NVIDIA’s vice president of enterprise computing, announced DGX H100 systems are shipping in a talk at MIT Technology Review’s Future Compute event today. Watch the video of his talk below.

Original URL: https://blogs.nvidia.com/blog/2023/05/01/dgx-h100-systems-shipping/