DATASHEET

NVIDIA DGX H100

The Gold Standard for AI Infrastructure

Artificial intelligence has become the go-to approach for solving difficult business challenges. Whether improving customer service, optimizing supply chains, extracting business intelligence, or designing cutting-edge products and services across nearly every industry, AI gives organizations the mechanism to realize innovation. And as a pioneer in AI infrastructure, NVIDIA DGX™ systems provide the most powerful and complete AI platform for bringing these essential ideas to fruition.

NVIDIA DGX H100 powers business innovation and optimization. The latest iteration of NVIDIA's legendary DGX systems and the foundation of NVIDIA DGX SuperPOD™, DGX H100 is an AI powerhouse that features the groundbreaking NVIDIA H100 Tensor Core GPU. The system is designed to maximize AI throughput, providing enterprises with a highly refined, systemized, and scalable platform to help them achieve breakthroughs in natural language processing, recommender systems, data analytics, and much more. Available on-premises and through a wide variety of access and deployment options, DGX H100 delivers the performance needed for enterprises to solve the biggest challenges with AI.

The Cornerstone of Your AI Center of Excellence

AI has bridged the gap between science and business. No longer the domain of experimentation, AI is used day in and day out by companies large and small to fuel their innovation and optimize their business. As the fourth generation of the world's first purpose-built AI infrastructure, DGX H100 is designed to be the centerpiece of an enterprise AI center of excellence. It's a fully optimized hardware and software platform that includes full support for the new range of NVIDIA AI software solutions, a rich ecosystem of third-party support, and access to expert advice from NVIDIA professional services. DGX H100 offers proven reliability, with DGX systems being used by thousands of customers around the world spanning nearly every industry.

Break Through the Barriers to AI at Scale

As the world's first system with the NVIDIA H100 Tensor Core GPU, NVIDIA DGX H100 breaks the limits of AI scale and performance. It features 6X more performance, 2X faster networking with NVIDIA ConnectX®-7 smart network interface cards (SmartNICs) and NVIDIA BlueField®-3 data processing units (DPUs), and high-speed scalability for NVIDIA DGX SuperPOD. The next-generation architecture is supercharged for the largest, most complex AI jobs, such as natural language processing and deep learning recommendation models.

SPECIFICATIONS

GPU	**8x NVIDIA H100 Tensor Core GPUs**
GPU memory	**640GB total**
Performance	**32 petaFLOPS FP8**
NVIDIA® NVSwitch™	**4x**
System power usage	**~10.2kW max**
CPU	**Dual x86**
System memory	**2TB**
Networking	**4x OSFP ports serving 8x single-port NVIDIA ConnectX-7** **400Gb/s InfiniBand/Ethernet** **2x dual-port NVIDIA BlueField-3 DPUs VPI** **1x 400Gb/s InfiniBand/Ethernet** **1x 200Gb/s InfiniBand/Ethernet**
Management network	**10Gb/s onboard NIC with RJ45** **50Gb/s Ethernet optional NIC** **Host baseboard management controller (BMC) with RJ45** **2x NVIDIA BlueField-3 DPU BMC (with RJ45 each)**
Storage	**OS: 2x 1.9TB NVMe M.2** **Internal storage: 8x 3.84TB NVMe U.2**
System software	**DGX H100 systems come preinstalled with DGX OS, which is based on Ubuntu Linux and includes the DGX software stack (all necessary packages and drivers optimized for DGX).** **Optionally, customers can install Ubuntu Linux or Red Hat Enterprise Linux and the required DGX software stack separately.**
Operating temperature range	**5–30°C (41–86°F)**

Leadership-Class Infrastructure on Your Terms

AI for business is about more than performance and capabilities. It's also about fitting neatly into an organization's IT envelope and practices. DGX H100 can be installed on-premises for direct management, colocated in NVIDIA DGX-Ready data centers, rented in NVIDIA DGX Foundry, and accessed through NVIDIA-certified managed service providers. And with the DGX-Ready Lifecycle Management program, organizations get a predictable financial model to keep their deployment at the leading edge. This makes DGX H100 as easy to use and acquire as traditional IT infrastructure, with no additional burden on busy IT staff—which lets organizations leverage AI for their businesses today instead of waiting for tomorrow.

Ordering Information

NVIDIA DGX H100 is coming late 2022. Contact your NVIDIA sales representative for more details.