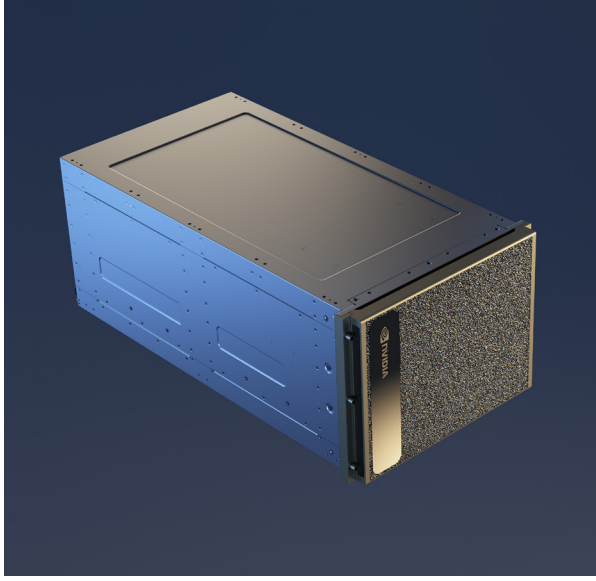
Datasheet



NVIDIA DGX H200

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 leading-edge products and services with generative AI and other transformer models, AI gives organizations across nearly every industry the mechanism to realize innovation. And as a pioneer in AI infrastructure, NVIDIA DGX™ provides the most powerful and complete AI platform for bringing these essential ideas to fruition.

NVIDIA DGX H200 powers business innovation and optimization. As a part of NVIDIA's legendary **DGX platform** and the foundation of **NVIDIA DGX SuperPOD™** and **DGX BasePOD™**, DGX H200 is an AI powerhouse that features the groundbreaking **NVIDIA H200 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 H200 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 a part of the world's first purpose-built AI infrastructure portfolio, DGX H200 is designed to be the centerpiece of an enterprise AI center of excellence. It's a fully optimized hardware and software platform that includes **NVIDIA Enterprise Support** for the new range of NVIDIA AI software solutions, a rich ecosystem of third-party support, and access to expert advice from NVIDIA specialists, allowing organizations to solve the biggest and most complex business problems with AI. DGX H200 offers proven reliability, with the DGX platform being used by thousands of customers around the world spanning nearly every industry.

Break Through the Barriers to AI at Scale

NVIDIA DGX H200 breaks the limits of AI scale and performance. It delivers 32 petaFLOPS of AI performance, 2X faster networking than DGX A100 with NVIDIA ConnectX®-7 smart network interface cards (SmartNICs), and high-speed scalability for NVIDIA DGX SuperPOD and DGX BasePOD. DGX H200 is supercharged with

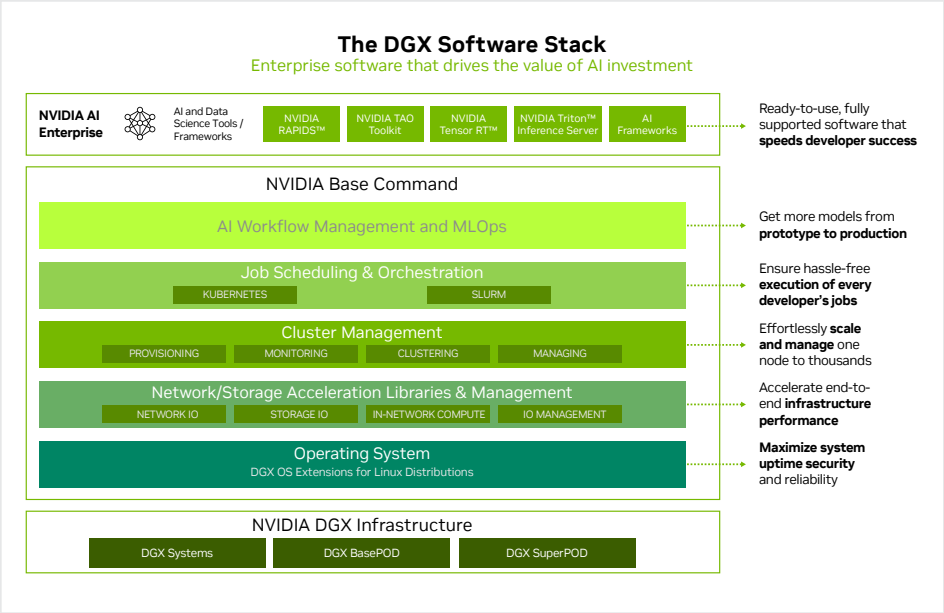
Specifications	
GPU	8x NVIDIA H200 Tensor Core GPUs, with 141GB of GPU memory each
GPU memory	1,128GB total
Performance	32 petaFLOPS FP8
NVIDIA NVSwitch™	4x
System power usage	10.2kW max*
CPU	Dual Intel® Xeon® Platinum 8480C Processors 112 Cores total, 2.00 GHz (Base), 3.80 GHz (Max Boost)
System memory	2TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	OS: 2x 1.92TB NVMe M.2
Internal storage:	8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI software NVIDIA Base Command – Orchestration, scheduling, and cluster management DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky – Operating System
Support	Comes with 3-year business-standard hardware and software support
System weight	287.6lbs (130.45kgs)
Packaged weight	376lbs (170.45kgs)
System dimensions	Height: 14.0in (356mm) Width: 19.0in (482.2mm) Length: 35.3in (897.1mm)
Operating temperature range	5–30°C (41–86°F)

*10.2kW for standard configuration. DGX H200 CTS (Custom Thermal Solution) supports up to 14.3kW.

1,128GB of GPU memory for the largest, most complex AI training and inference jobs, such as generative AI, natural language processing and deep learning recommendation models.

Powered by NVIDIA Base Command

NVIDIA Base Command powers the DGX platform, enabling organizations to leverage the best of NVIDIA software innovation. Enterprises can unleash the full potential of their DGX infrastructure with a proven platform that includes enterprise-grade orchestration and cluster management, libraries that accelerate compute, storage and network infrastructure, and an operating system optimized for AI workloads. Additionally, DGX infrastructure includes NVIDIA AI Enterprise, offering a suite of software optimized to streamline AI development and deployment.



NVIDIA DGX AI Software Stack

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 H200 can be installed on premises for direct management, colocated in **NVIDIA DGX-Ready Data Centers**, 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 H200 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.

Ready to Get Started?

To learn more about NVIDIA DGX H200, visit **nvidia.com/DGX-H200**