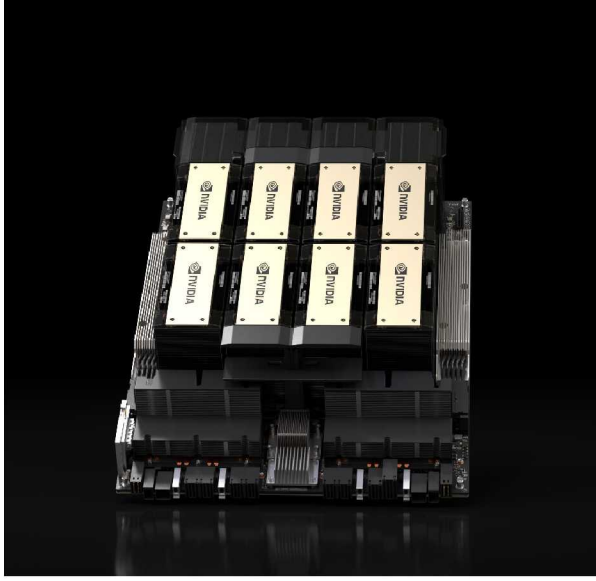
Datasheet

NVIDIA H200 Tensor Core GPU

Supercharging AI and HPC workloads.



Higher Performance With Larger, Faster Memory

The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities.

Based on the **NVIDIA Hopper™ architecture**, the NVIDIA H200 is the first GPU to offer 141 gigabytes (GB) of HBM3e memory at 4.8 terabytes per second (TB/s)—that's nearly double the capacity of the **NVIDIA H100 Tensor Core GPU** with 1.4X more memory bandwidth. The H200's larger and faster memory accelerates generative AI and large language models, while advancing scientific computing for HPC workloads with better energy efficiency and lower total cost of ownership.

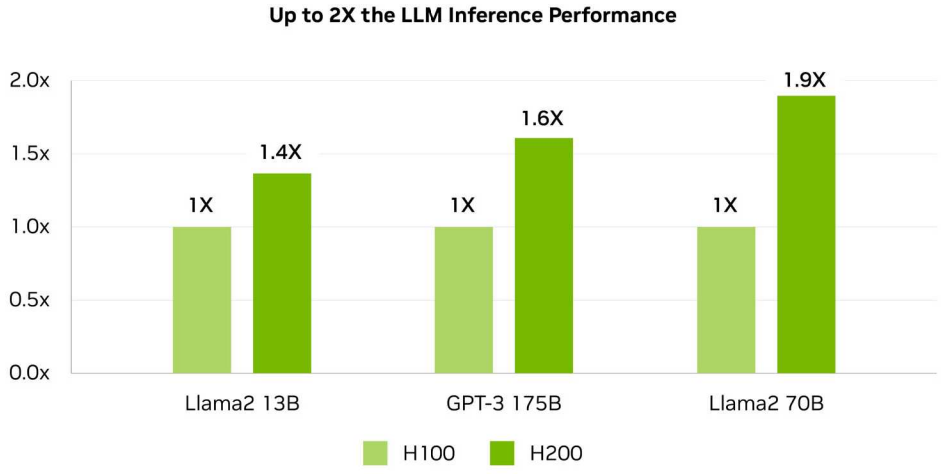
Key Features

- 141GB of HBM3e GPU memory
- 4.8TB/s of memory bandwidth
- 4 petaFLOPS of FP8 performance
- 2X LLM inference performance
- 110X HPC performance

Unlock Insights With High-Performance LLM Inference

In the ever-evolving landscape of AI, businesses rely on large language models to address a diverse range of inference needs. An **AI inference** accelerator must deliver the highest throughput at the lowest TCO when deployed at scale for a massive user base.

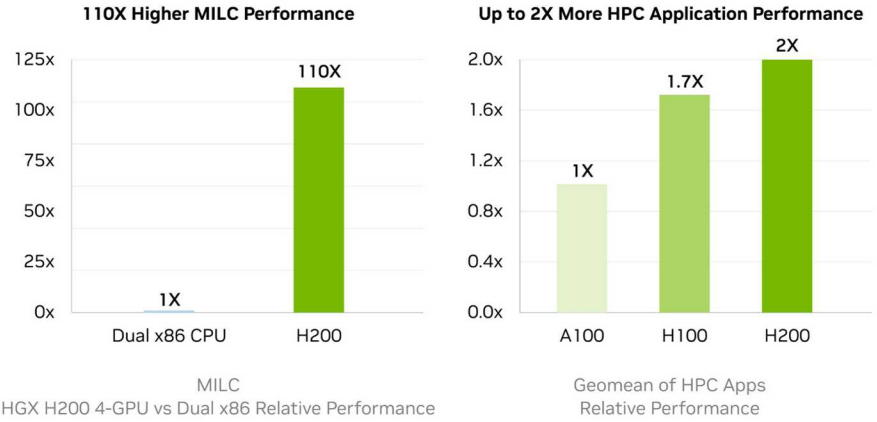
The H200 doubles inference performance compared to H100 GPUs when handling large language models such as Llama2 70B.



Preliminary specifications. May be subject to change.
Llama2 13B: ISL 128, OSL 2K | Throughput | H100 SXM 1x GPU BS 64 | H200 SXM 1x GPU BS 128
GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPUs BS 64 | x8 H200 SXM GPUs BS 128
Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32.

Supercharge High-Performance Computing

Memory bandwidth is crucial for HPC applications, as it enables faster data transfer and reduces complex processing bottlenecks. For memory-intensive HPC applications like simulations, scientific research, and artificial intelligence, the H200's higher memory bandwidth ensures that data can be accessed and manipulated efficiently, leading to 110X faster time to results.



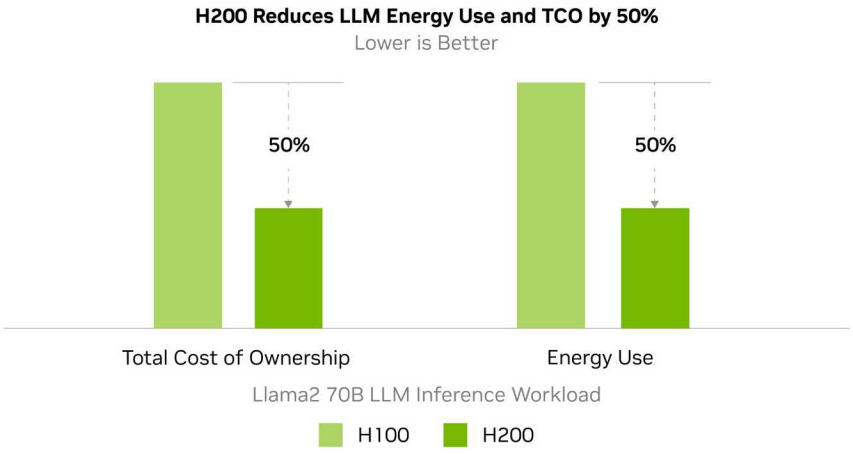
Preliminary specifications. May be subject to change.

HPC MILC- dataset NERSC Apex Medium | HGX H200 4-GPU | dual Sapphire Rapids 8480

HPC Apps- CP2K: dataset H2O-32-RI-dRPA-96points | GROMACS: dataset STMV | ICON: dataset r2b5 | MILC: dataset NERSC Apex Medium | Chroma: dataset HMC Medium | Quantum Espresso: dataset AUSURF112 | 1x H100 SXM | 1x H200 SXM.

Reduce Energy and TCO

With the introduction of H200, energy efficiency and TCO reach new levels. This cutting-edge technology offers unparalleled performance, all within the same power profile as the **H100 Tensor Core GPU**. AI factories and supercomputing systems that are not only faster but also more eco-friendly deliver an economic edge that propels the AI and scientific communities forward.



Preliminary specifications. May be subject to change.

Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32

Enterprise-Ready: AI Software Streamlines Development and Deployment

NVIDIA H200 is bundled with a five-year **NVIDIA AI Enterprise** subscription and simplifies the way you build an enterprise AI-ready platform. H200 accelerates AI development and deployment for production-ready generative AI solutions, including computer vision, speech AI, retrieval augmented generation (RAG), and more. NVIDIA AI Enterprise includes **NVIDIA NIM™**, a set of easy-to-use microservices designed to speed up enterprise generative AI deployment. Together, deployments have enterprise-grade security, manageability, stability, and support. This results in performance-optimized AI solutions that deliver faster business value and actionable insights.

Unleashing AI Acceleration for Mainstream Enterprise Servers With H200 NVL

The NVIDIA H200 NVL is the ideal choice for customers with space constraints within the data center, delivering acceleration for every AI and HPC workload regardless of size. With a 1.5X memory increase and a 1.2X bandwidth increase over the previous generation, customers can fine-tune LLMs within a few hours and experience LLM inference 1.8X faster.

Technical Specifications		
Form Factor	**H200 SXM[1]**	**H200 NVL[1]**
FP64	34 TFLOPS	34 TFLOPS
FP64 Tensor Core	67 TFLOPS	67 TFLOPS
FP32	67 TFLOPS	67 TFLOPS
TF32 Tensor Core	989 TFLOPS[2]	989 TFLOPS[2]
BFLOAT16 Tensor Core	1,979 TFLOPS[2]	1,979 TFLOPS[2]
FP16 Tensor Core	1,979 TFLOPS[2]	1,979 TFLOPS[2]
FP8 Tensor Core	3,958 TFLOPS[2]	3,958 TFLOPS[2]
INT8 Tensor Core	3,958 TFLOPS[2]	3,958 TFLOPS[2]
GPU Memory	141GB	141GB
GPU Memory Bandwidth	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Confidential Computing	Supported	Supported
Max Thermal Design Power (TDP)	Up to 700W (configurable)	Up to 600W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @16.5GB each	Up to 7 MIGs @16.5GB each
Form Factor	SXM	PCIe
Interconnect	NVIDIA NVLink®: 900GB/s; PCIe Gen5: 128GB/s	2- or 4-way NVIDIA NVLink bridge: 900GB/s; PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs	NVIDIA MGX™ H200 NVL partner and NVIDIA-Certified Systems with up to 8 GPUs
NVIDIA AI Enterprise	Add-on	Included

1. Preliminary specifications. May be subject to change.
2. With sparsity.

Ready to get started?

To learn more about the NVIDIA H200 Tensor Core GPU, visit **nvidia.com/h200**

NVIDIA

NORTHERN DATA GROUP