acecloud by rtds

NVIDIA

DATASHEET

NVIDIA H200 Tensor Core GPU on AceCloud

Supercharging Generative AI and HPC Workloads

The **NVIDIA H200 Tensor Core GPU**, powered by the NVIDIA Hopper™ architecture, delivers breakthrough performance for **generative AI, large language models (LLMs), and high-performance computing (HPC)**. Available on AceCloud's enterprise-grade infrastructure, the H200 provides **larger memory, faster bandwidth, and unparalleled inference efficiency**, enabling enterprises to scale AI and scientific workloads with confidence.

Higher Performance with Larger, Faster Memory

The NVIDIA H200 is the first GPU to feature **141 GB of HBM3e memory** with **4.8 TB/s of memory bandwidth** nearly double the capacity of the H100 and 1.4X more bandwidth

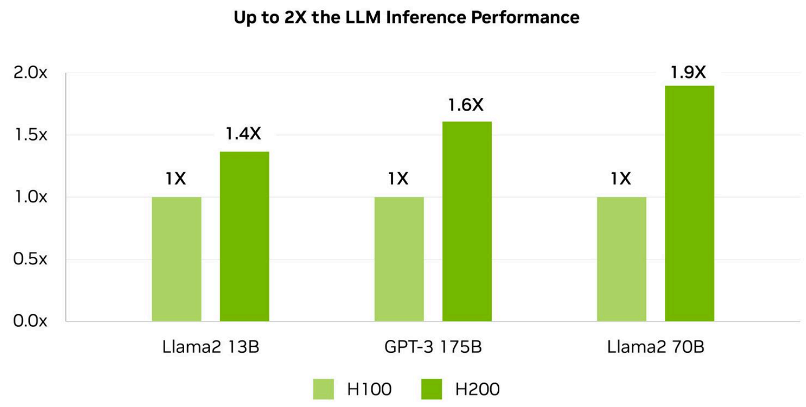
This enhanced memory architecture accelerates **LLMs, generative AI applications, and HPC simulations**, while reducing energy usage and total cost of ownership. Organizations gain faster results, more efficient scaling, and the ability to handle even the most demanding AI pipelines.

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

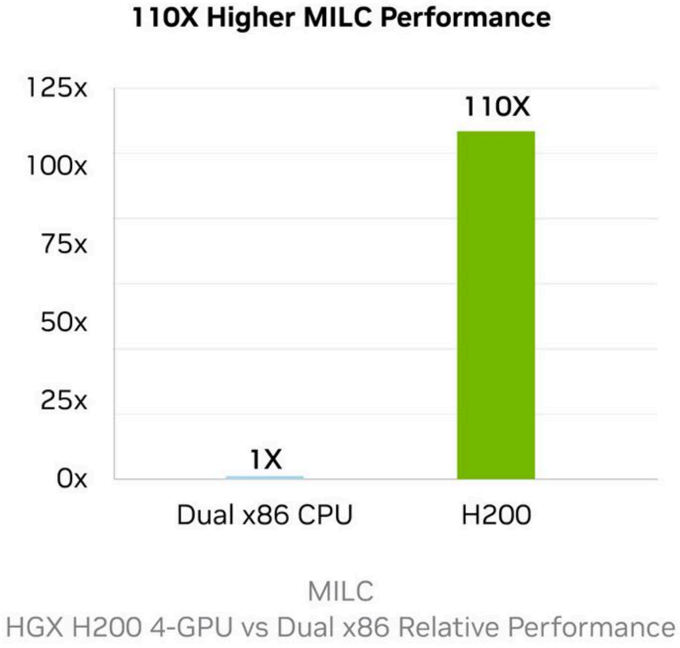
For large-scale AI adoption, inference throughput is critical. The H200 delivers **2X the inference performance of the H100**, making it ideal for running advanced LLMs like **Llama 2 70B** at scale This makes the H200 a powerful choice for enterprises deploying AI-driven services to millions of users, ensuring performance, scalability, and cost efficiency in production environments.

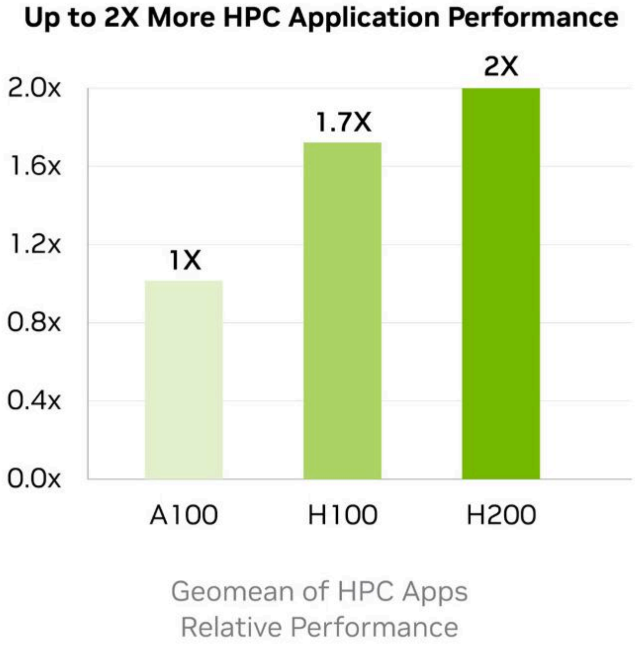


Preliminary specifications. May be subject to change.
Llama2 13B: ISL 128, OSL 2K | Throughput | H100 SXM 1x GPU BS 64 | H200 SXM 1x GPU BS 128
GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPUs BS 64 | x8 H200 SXM GPUs BS 128
Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x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

Supercharge High-Performance Computing

With its **4.8 TB/s bandwidth**, the H200 removes bottlenecks in data-intensive HPC applications such as molecular dynamics, quantum chemistry, and large-scale simulations. Tests show up to **110X HPC performance improvements compared to CPUs**





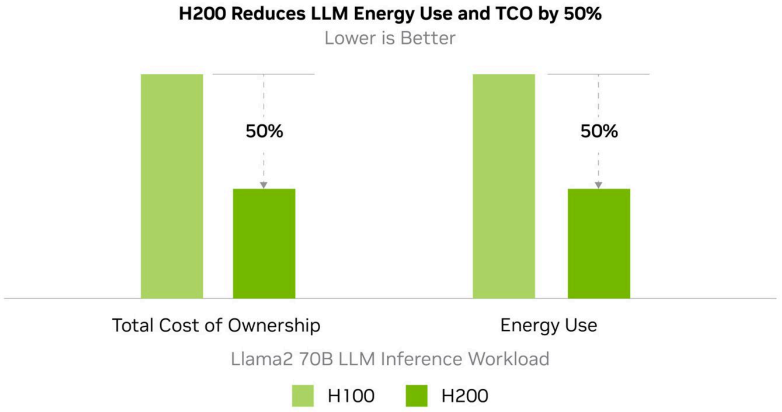
Enterprise-Ready with NVIDIA AI Software

Deployed through AceCloud, the H200 includes access to **NVIDIA AI Enterprise**, featuring frameworks, pre-trained models, and NIM microservices to simplify generative AI deployment. This provides enterprises with **production-ready AI environments**, backed by AceCloud's **99.99% uptime SLA, MIG-enabled isolation, and secure multi-cloud architecture.**

Reduce Energy and TCO

The NVIDIA H200 Tensor Core GPU brings a new level of efficiency to enterprise AI and HPC. Designed with the same power profile as the H100, the H200 not only delivers higher performance but also helps organizations **cut energy consumption and total cost of ownership (TCO) by up to 50%.**

For enterprises running large-scale AI factories or supercomputing workloads, this translates to systems that are both **faster and more eco-friendly**, giving businesses a clear economic and sustainability advantage.



Use Cases

- **Generative AI & LLM inference**
- **HPC workloads & simulations**
- **Enterprise AI applications**
- **AI factories & data centers**
- **Hybrid & multi-cloud deployments**

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

NVIDIA H200 Tensor Core GPU – Technical Specifications

SYSTEM SPECIFICATIONS

Specification	H200 SXM	H200 NVL
FP64	34 TFLOPS	30 TFLOPS
FP64 Tensor Core	67 TFLOPS	60 TFLOPS
FP32	67 TFLOPS	60 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS	835 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS	1,671 TFLOPS
FP16 Tensor Core[2]	1,979 TFLOPS	1,671 TFLOPS
FP8 Tensor Core[2]	3,958 TFLOPS	3,341 TFLOPS
INT8 Tensor Core[2]	3,958 TFLOPS	3,341 TFLOPS
GPU Memory	141 GB	141 GB
GPU Memory Bandwidth	4.8 TB/s	4.8 TB/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 (MIGs)	Up to 7 MIGs @ 18 GB each	Up to 7 MIGs @ 16.5 GB each
Form Factor	SXM	PCIe (Dual-slot air-cooled)
Interconnect	NVIDIA NVLink: 900 GB/s PCIe Gen5: 128 GB/s	2- or 4-way NVLink bridge: 900 GB/s per GPU PCIe Gen5: 128 GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ (4 or 8 GPUs)	NVIDIA MGX™ H200 NVL partner and NVIDIA-Certified Systems™ (up to 8 GPUs)
NVIDIA AI Enterprise	Add-on	Included

[1] Preliminary specifications. May be subject to change.

[2] With sparsity.

Launch H200 on AceCloud Today

- ✓ Instant provisioning of NVIDIA H200 GPUs
- ✓ CUDA & container-ready environments
- ✓ MIG-enabled multi-tenant isolation
- ✓ 99.99% uptime SLA

Scan to launch your H200 VM →

Start with H200 now