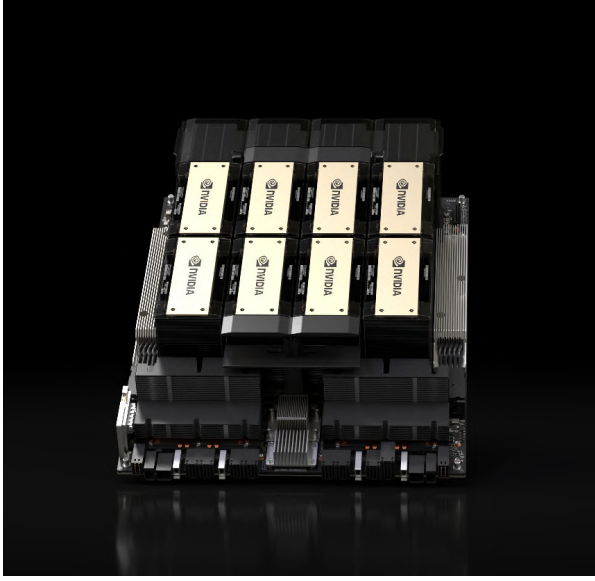
Datasheet

NVIDIA HGX H100 and HGX H200

The world's leading AI computing platform.



Purpose-Built for AI and High-Performance Computing

AI, complex simulations, and massive datasets require multiple GPUs with extremely fast interconnections and a fully accelerated software stack. The NVIDIA HGX™ AI supercomputing platform brings together the full power of NVIDIA GPUs, NVIDIA NVLink™, NVIDIA networking, and fully optimized AI and high-performance computing (HPC) software stacks to provide the highest application performance and drive the fastest time to insights.

Unmatched End-to-End Accelerated Computing Platform

The NVIDIA HGX H200 combines **H200 Tensor Core GPUs** with high-speed interconnects to form the world's most powerful servers. Configurations of up to eight GPUs deliver unprecedented acceleration, with up to 1.1 terabytes (TB) of GPU memory and 38 terabytes per second (TB/s) of aggregate memory bandwidth. This combined with a staggering 32 petaFLOPS of performance creates the world's most powerful accelerated scale-up server platform for AI and HPC.

Both the HGX H200 and HGX H100 include advanced networking options—at speeds up to 400 gigabits per second (Gb/s)—utilizing NVIDIA Quantum-2 InfiniBand and **Spectrum™-X Ethernet** for the highest AI performance. HGX H200 and HGX H100 also include **NVIDIA® BlueField®-3** data processing units (DPUs) to enable cloud networking, composable storage, zero-trust security, and GPU compute elasticity in hyperscale AI clouds.

Deep Learning Inference: Performance and Versatility

AI solves a wide array of business challenges using an equally wide array of neural networks. A great AI inference accelerator has to, not only deliver the highest performance, but also the versatility needed to accelerate these networks in any location that customers choose to deploy them, from data center to edge.

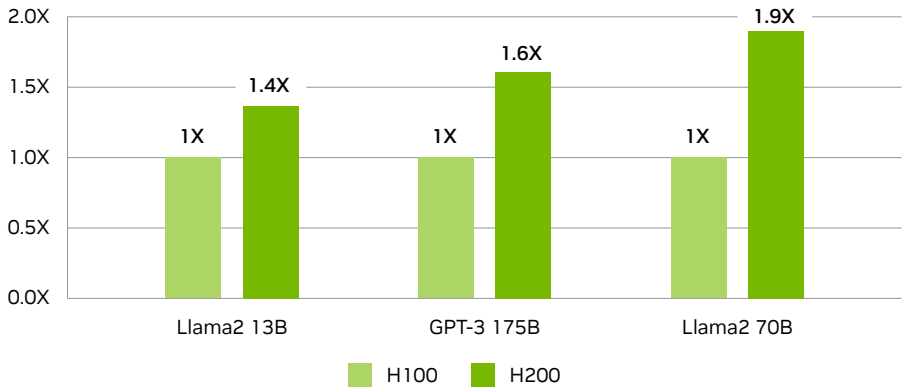
HGX H200 and HGX H100 extend NVIDIA's market-leading inference leadership.

Key Features

NVIDIA HGX H100 and HGX H200

- Transformer Engine
- Fourth-generation NVIDIA NVLink
- NVIDIA Confidential Computing
- NVIDIA Multi-Instance GPU (MIG)
- DPX Instructions

Up to 2X the LLM Inference Performance

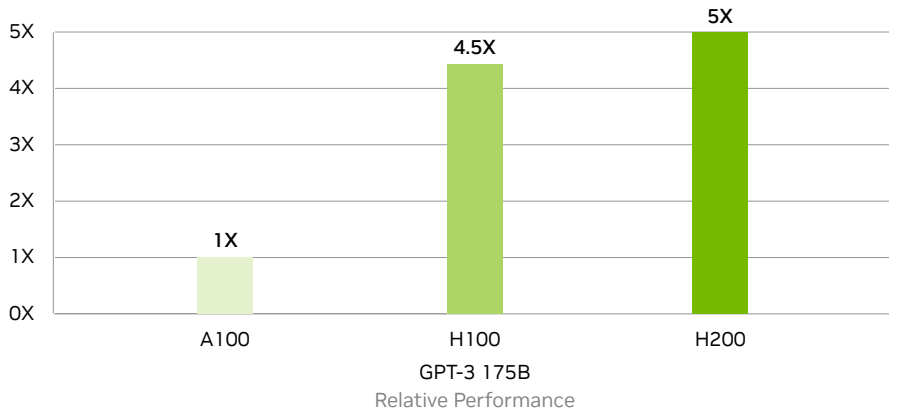


Preliminary specifications. May be subject to change.
Llama 2 13B: ISL 128, OSL 2K | Throughput | H100 SXM 1x GPU BS 64 | H200 SXM 1x GPU BS 128.
GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPUs BS 64 | x8 H200 SXM GPUs BS 128.
Llama 2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM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

Deep Learning Training: Performance and Scalability

NVIDIA H200 and **H100 GPUs** feature the Transformer Engine, with FP8 precision, that provides up to 5X faster training over the previous GPU generation for large language models. The combination of fourth-generation NVLink—which offers 900GB/s of GPU-to-GPU interconnect—PCIe Gen5, and NVIDIA Magnum IO™ software delivers efficient scalability, from small enterprises to massive unified GPU clusters. These infrastructure advances, working in tandem with the **NVIDIA AI Enterprise software suite**, make HGX H200 and HGX H100 the world's leading AI computing platform.

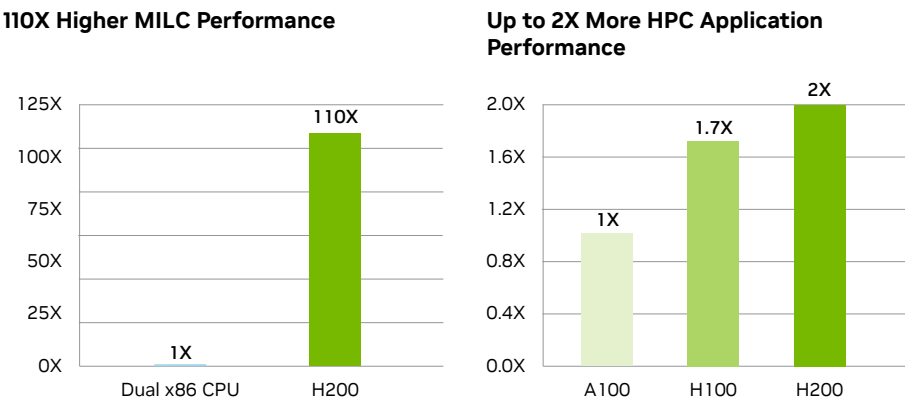
Up to 5X Faster Training at Scale



Projected performance, subject to change.
GPT-3 175B training NVIDIA A100 Tensor Core GPU cluster: NVIDIA Quantum InfiniBand network, H100 cluster: NVIDIA Quantum-2 InfiniBand network

Supercharge HPC Performance

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



Projected performance, 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 | 1x H200.

Accelerating HGX With NVIDIA Networking

The data center is the new unit of computing, and networking plays an integral role in scaling application performance across it. Paired with NVIDIA Quantum InfiniBand, HGX delivers world-class performance and efficiency, which ensures the full utilization of computing resources.

For AI cloud data centers that deploy Ethernet, HGX is best used with the NVIDIA Spectrum-X networking platform, which powers the highest AI performance over Ethernet. It features Spectrum-X switches and BlueField-3 DPUs for optimal resource utilization and performance isolation, delivering consistent, predictable outcomes for thousands of simultaneous AI jobs at every scale. Spectrum-X enables advanced cloud multi-tenancy and zero-trust security. As a reference design, NVIDIA has designed Israel-1, a hyperscale generative AI supercomputer built with Dell PowerEdge XE9680 servers based on the NVIDIA HGX 8-GPU platform, BlueField-3 DPUs, and Spectrum-4 switches.

Technical Specifications				
	HGX H200 4-GPU	**HGX H200 8-GPU**	**HGX H100 4-GPU**	**HGX H100 8-GPU**
Form Factor	4x NVIDIA H200 SXM	8x NVIDIA H200 SXM	4x NVIDIA H100 SXM	8x NVIDIA H100 SXM
FP8 Tensor Core*	16 PFLOPS	32 PFLOPS	16 PFLOPS	32 PFLOPS
INT8 Tensor Core*	16 POPS	32 POPS	16 POPS	32 POPS
FP16/BFLOAT16 Tensor Core*	8 PFLOPS	16 PFLOPS	8 PFLOPS	16 PFLOPS
TF32 Tensor Core*	4 PFLOPS	8 PFLOPS	4 PFLOPS	8 PFLOPS
FP32	270 TFLOPS	540 TFLOPS	270 TFLOPS	540 TFLOPS
FP64	140 TFLOPS	270 TFLOPS	140 TFLOPS	270 TFLOPS
FP64 Tensor Core	270 TFLOPS	540 TFLOPS	270 TFLOPS	540 TFLOPS
Memory	564GB HBM3e	1.1TB HBM3e	320GB HBM3	640GB HBM3
GPU Aggregate Bandwidth	19GB/s	38GB/s	13GB/s	27GB/s
NVLink	Fourth generation	Fourth generation	Fourth generation	Fourth generation
NVSwitch™	N/A	Third generation	N/A	Third generation
NVSwitch GPU-to-GPU Bandwidth	N/A	900GB/s	N/A	900GB/s
Total Aggregate Bandwidth	3.6TB/s	7.2TB/s	3.6TB/s	7.2TB/s

* With sparsity.

Ready to Get Started?

To learn more about NVIDIA HGX, visit
www.nvidia.com/hgx

NVIDIA