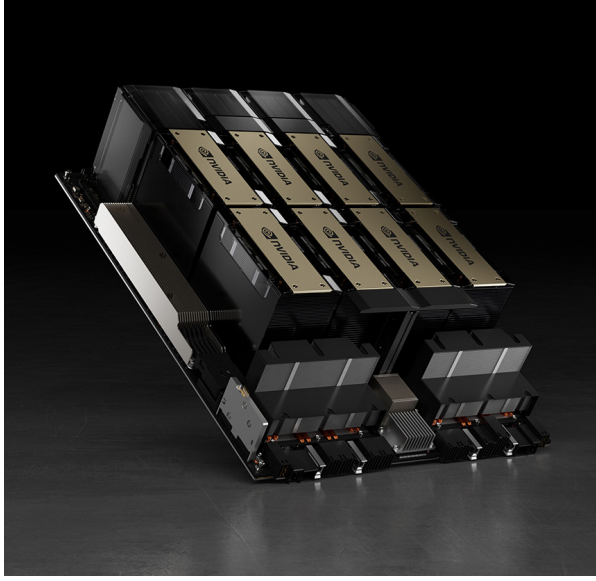
Datasheet

NVIDIA HGX H100 GPU

The most powerful end-to-end AI supercomputing platform.



Purpose-Built for AI, Simulation, and Data Analytics

Exploding model sizes and use cases for AI, complex simulations, and massive datasets require multiple GPUs with extremely fast interconnects and a fully accelerated software stack. The NVIDIA HGX™ AI supercomputing platform brings together the full power of NVIDIA GPUs, NVLink®, networking, and the fully optimized AI and high-performance computing (HPC) software stack from the NGC™ catalog to provide the highest application performance. With its end-to-end performance and flexibility, NVIDIA HGX enables service providers, researchers, and scientists to deliver AI, simulation, and data analytics to drive the fastest time to insights.

Unmatched End-to-End Accelerated Computing Platform

NVIDIA HGX H100 combines **H100 Tensor Core GPUs** with high-speed interconnects to form the world's most powerful servers. With up to eight H100 GPUs, HGX H100 has up to 640 gigabytes (GB) of GPU memory and 24 terabytes per second (TB/s) of aggregate memory bandwidth for unprecedented acceleration. NVIDIA HGX H100 delivers a staggering 32 petaFLOPS, forming the world's most powerful accelerated scale-up server platform for AI and HPC.

HGX H100 includes 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 H100 also includes 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 Training: Performance and Scalability

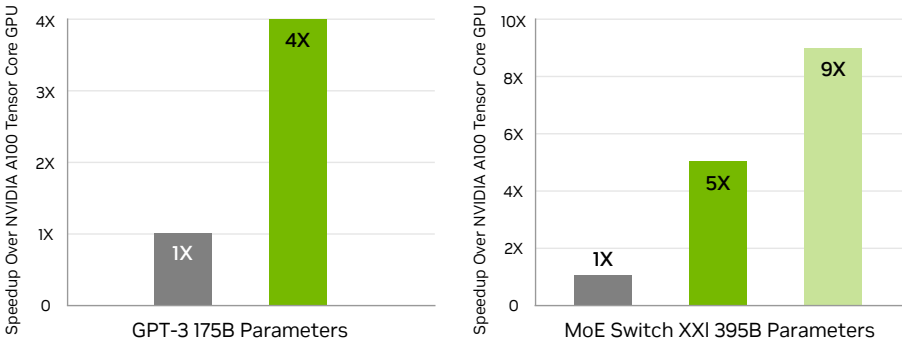
NVIDIA H100 GPUs used in HGX H100 systems feature the Transformer Engine, with FP8 precision, that provides up to 4X faster training over the prior generation for large language models, such as GPT-3 175B. The combination of fourth-generation NVLink, which offers 900GB/s of GPU-to-GPU interconnect, NVLink Switch System, which accelerates collective communication by every GPU across nodes, 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 NVIDIA HGX H100 the most powerful end-to-end AI and HPC data center platform.

Key Features

NVIDIA H100

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

Over 4X Higher AI Training on Largest Models

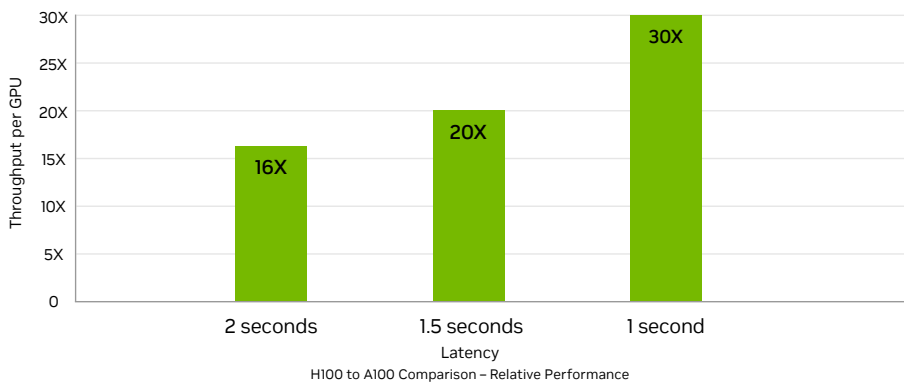


GPT-3 175B training HGX A100 8-GPU cluster: NVIDIA Quantum InfiniBand network, HGX H100 8-GPU cluster: NVIDIA Quantum-2 InfiniBand network | Mixture of Experts (MoE) training Transformer Switch-XXL variant with 395 billion parameters on 1-trillion-token dataset, HGX A100 8-GPU cluster: NVIDIA Quantum infiniBand network, HGX H100 8-GPU cluster: NVIDIA Quantum-2 InfiniBand network or NVLink Switch System where indicated. (Note: H100 systems offering NVLink Switch System aren't currently available. Performance claims are preliminary and subject to change.)

Deep Learning Inference: Performance and Versatility

HGX H100 further extends NVIDIA's market-leading inference leadership with several advancements that accelerate inference by up to 30X over the prior generation on Megatron 530-billion-parameter chatbots. Fourth-generation Tensor Cores speed up all precisions, including FP64, TF32, FP32, FP16, and INT8, and the Transformer Engine utilizes FP8 and FP16 together to reduce memory usage and increase performance while maintaining accuracy for large language models.

Megatron Chatbot Inference (530 Billion Parameters)

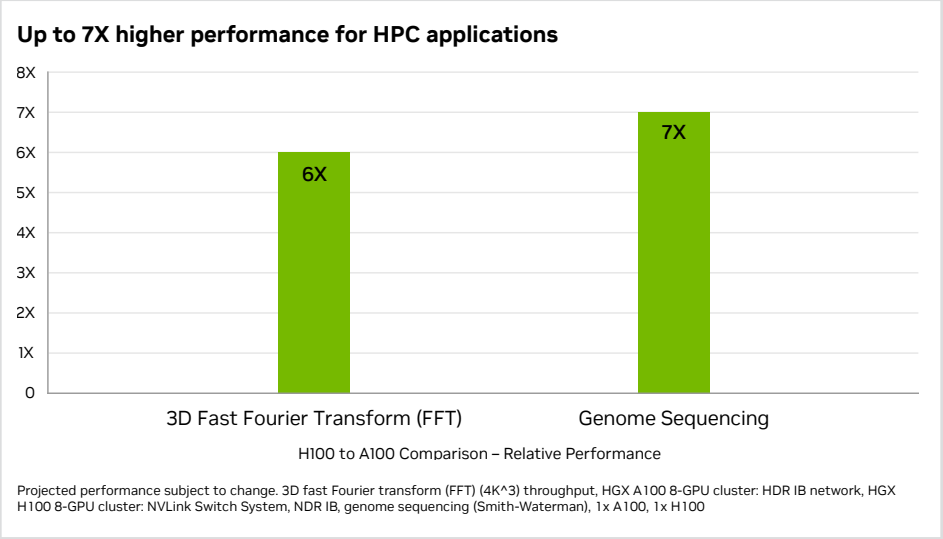


Measured performance. Inference on Megatron 530B model chatbot for input sequence length = 128, output sequence length = 20, HGX A100 8-GPU cluster: NVIDIA Quantum InfiniBand network, HGX H100 8-GPU cluster: NVIDIA Quantum-2 InfiniBand network for 2x HGX H100 configurations, 4x HGX A100 vs. 2x HGX H100 for 1 and 1.5 seconds, 2x HGX A100 vs. 1x HGX H100 for 2 seconds.

HPC: Faster Double-Precision Tensor Cores and Dynamic Programming

HGX H100 triples the floating-point operations per second (FLOPS) of double-precision Tensor Cores, delivering up to 535 teraFLOPS of FP64 computing for HPC in the eight-way configuration or 268 teraFLOPs in the four-way configuration. AI-fused HPC applications can also leverage H100's TF32 precision to achieve nearly 8,000 teraFLOPS of throughput for single-precision matrix-multiply operations, with zero code changes.

H100 features new DPX instructions that speed up dynamic programming algorithms—such as Smith-Waterman used in DNA sequence alignment and protein alignment for protein structure prediction—by 7X over the NVIDIA Ampere architecture. By increasing the throughput of diagnostic functions like gene sequencing, H100 will enable every clinic to offer accurate, real-time disease diagnosis and precision medicine prescription, democratizing healthcare as we know it.



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. NVIDIA Quantum leads the way with in-network computing acceleration, remote direct-memory access (RDMA), and advanced quality-of-service (QoS) capabilities.

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 400Gb/s Ethernet. Featuring NVIDIA Spectrum™-4 switches and BlueField-3 DPUs, Spectrum-X delivers consistent, predictable outcomes for thousands of simultaneous AI jobs at every scale through optimal resource utilization and performance isolation. Spectrum-X enables advanced cloud multi-tenancy and zero-trust security. With Spectrum-X, cloud service providers can accelerate the development, deployment, and time to market of AI solutions, while improving return on investment.

System Specifications

Peak Performance		
	HGX H100 4-GPU	**HGX H100 8-GPU**
FP64	134 TFLOPS	268 TFLOPS
FP64 Tensor Core	268 TFLOPS	535 TFLOPS
FP32	268 TFLOPS	535 TFLOPS
TF32 Tensor Core	3,958 TFLOPS*	7,915 TFLOPS*
FP16 Tensor Core	7,915 TFLOPS*	15,830 TFLOPS*
FP8 Tensor Core	15,830 TFLOPs*	31,662 TFLOPS*
INT8 Tensor Core	15,830 TOPS*	31,662 TOPS*
GPU memory	320GB	640GB
Aggregate GPU memory bandwidth	13TB/s	27TB/s
Maximum number of MIG instances	28	56
NVIDIA NVLink	Fourth-generation NVLink 900GB/s	Fourth-generation NVLink 900GB/s
NVIDIA NVSwitch	N/A	Third-generation NVSwitch
NVSwitch GPU-to-GPU bandwidth	N/A	900GB/s
In-network compute	N/A	3.6 TFLOPS
Total aggregate network bandwidth	3.6TB/s	7.2TB/s

* With sparsity.

Ready to Get Started?

To learn more about NVIDIA HGX H100, visit:
www.nvidia.com/hgx

NVIDIA