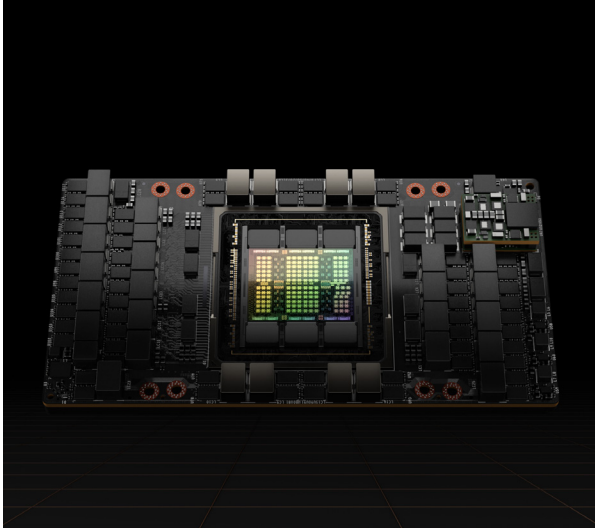
Datasheet

NVIDIA H100 Tensor Core GPU

Extraordinary performance, scalability, and security for every data center.



An Order-of-Magnitude Leap for Accelerated Computing

The NVIDIA H100 Tensor Core GPU delivers exceptional performance, scalability, and security for every workload. H100 uses breakthrough innovations based on the **NVIDIA Hopper™ architecture** to deliver industry-leading conversational AI, speeding up large language models by 30X.

Securely Accelerate Workloads From Enterprise to Exascale

H100 features fourth-generation Tensor Cores and a Transformer Engine with FP8 precision that provides up to 4X faster training over the prior generation for GPT-3 (175B) models. For high-performance computing (HPC) applications, H100 triples the floating-point operations per second (FLOPS) of double-precision Tensor Cores, delivering 60 teraflops of FP64 computing for HPC while also featuring dynamic programming (DPX) instructions to deliver up to 7X higher performance. With second-generation Multi-Instance GPU (MIG), built-in NVIDIA Confidential Computing, and NVIDIA NVLink Switch System, H100 securely accelerates all workloads for every data center, from enterprise to exascale.

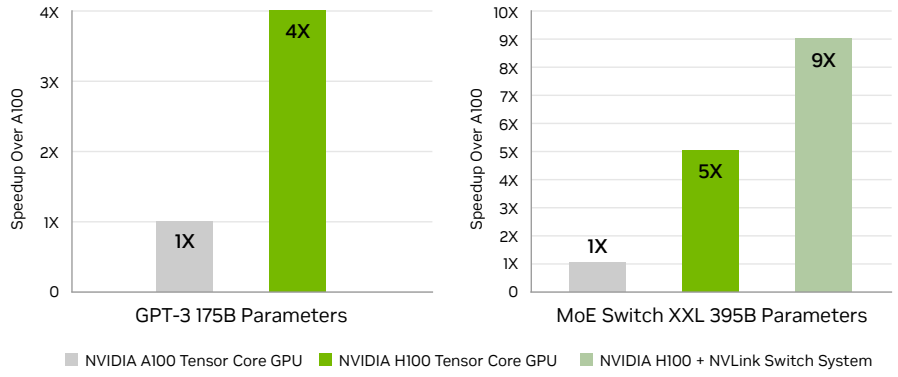
Supercharge Large Language Model Inference With H100 NVL

For LLMs up to 70 billion parameters (Llama 2 70B), the PCIe-based NVIDIA H100 NVL with NVLink bridge utilizes Transformer Engine, NVLink, and 188GB HBM3 memory to provide optimum performance and easy scaling across any data center, bringing LLMs to the mainstream. Servers equipped with H100 NVL GPUs increase Llama 2 70B model performance up to 5X over NVIDIA A100 systems while maintaining low latency in power-constrained data center environments.

Enterprise-Ready: AI Software Streamlines Development and Deployment

NVIDIA H100 NVL is bundled with a five-year **NVIDIA AI Enterprise** subscription and simplifies the way you build an enterprise AI-ready platform. H100 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.

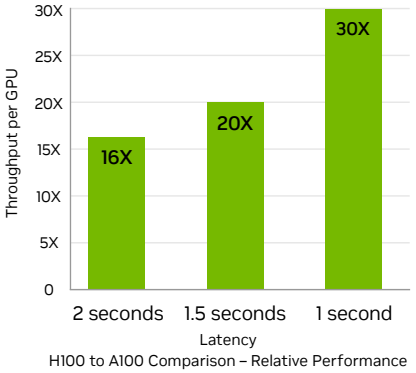
Up to 4X Higher AI Training on GPT-3



Projected performance subject to change. GPT-3 175B Training A100 cluster: HDR IB network, H100 cluster: NDR IB network | Mixture of Experts (MoE) Training Transformer Switch-XXL variant with 395B parameters on 1T token dataset, A100 cluster: HDR IB network, H100 cluster: NDR IB network with NVLink Switch System where indicated.

Up to 30X Higher AI Inference Performance on the Largest Model

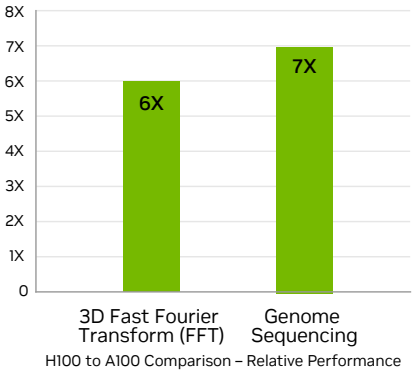
Megatron chatbot inference (530 billion parameters)



H100 to A100 Comparison – Relative Performance

Projected performance subject to change. Inference on Megatron 530B parameter model based chatbot for input sequence length=128, output sequence length =20 | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB

Up to 7X Higher Performance for HPC Applications



H100 to A100 Comparison – Relative Performance

Projected performance subject to change. 3D FFT (4K^3) throughput | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB | Genome Sequencing (Smith-Waterman) | 1 A100 | 1 H100

Technical Specifications

	H100 SXM	H100 NVL
FP64	34 teraFLOPS	30 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	60 teraFLOPS
FP32	67 teraFLOPS	60 teraFLOPS
TF32 Tensor Core*	989 teraFLOPS	835 teraFLOPS
BFLOAT16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS	3,341 teraFLOPS
INT8 Tensor Core*	3,958 TOPS	3,341 TOPS
GPU Memory	80GB	94GB
GPU Memory Bandwidth	3.35TB/s	3.9TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	Up to 700W (configurable)	350-400W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGS @ 12GB each
Form Factor	SXM	PCIe dual-slot air-cooled
Interconnect	**NVIDIA NVLink™:** 900GB/s **PCIe Gen5:** 128GB/s	**NVIDIA NVLink:** 600GB/s **PCIe Gen5:** 128GB/s
Server Options	NVIDIA HGX H100 Partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1–8 GPUs
NVIDIA Enterprise	Add-on	Included

*With sparsity

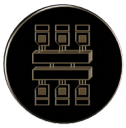
Explore the Technology Breakthroughs of NVIDIA Hopper

NVIDIA H100 Tensor Core GPU

Built with 80 billion transistors using a cutting-edge TSMC 4N process custom tailored for NVIDIA's accelerated compute needs, H100 features major advances to accelerate AI, HPC, memory bandwidth, interconnect, and communication at data center scale.

Transformer Engine

The Transformer Engine uses software and Hopper Tensor Core technology designed to accelerate training for models built from the world's most important AI model building block, the transformer. Hopper Tensor Cores can apply mixed FP8 and FP16 precisions to dramatically accelerate AI calculations for transformers.



NVLink Switch System

The NVLink Switch System enables the scaling of multi-GPU input/output (IO) across multiple servers at 900 gigabytes per second (GB/s) bidirectional per GPU, over 7X the bandwidth of PCIe Gen5. The system delivers 9X higher bandwidth than InfiniBand HDR on the NVIDIA Ampere architecture.

NVIDIA Confidential Computing

NVIDIA H100 brings high-performance security to workloads with confidentiality and integrity. Confidential Computing delivers hardware-based protection for data and applications in use.

Second-Generation Multi-Instance GPU (MIG)

The Hopper architecture's second-generation MIG supports multi-tenant, multi-user configurations in virtualized environments, securely partitioning the GPU into isolated, right-size instances to maximize quality of service (QoS) for 7X more secured tenants.

DPX Instructions

Hopper's DPX instructions accelerate dynamic programming algorithms by 40X compared to CPUs and 7X compared to NVIDIA Ampere architecture GPUs. This leads to dramatically faster times in disease diagnosis, real-time routing optimizations, and graph analytics.

Ready to Get Started?

To learn more about the NVIDIA H100 Tensor Core GPU, visit:
www.nvidia.com/h100

NVIDIA