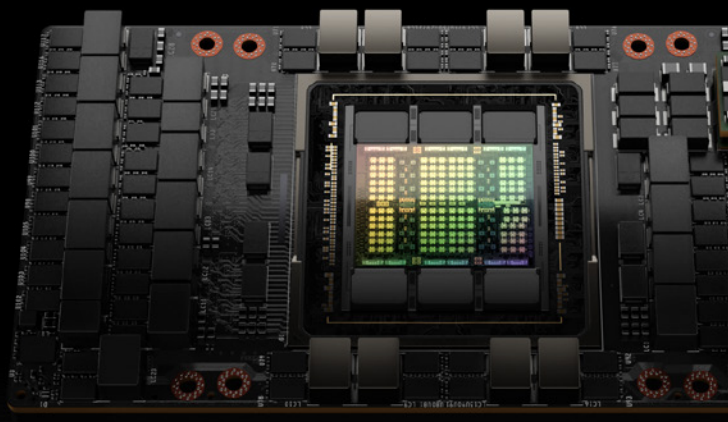
DATASHEET



NVIDIA H100 TENSOR CORE GPU

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

An Order-of-Magnitude Leap for Accelerated Computing

The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With NVIDIA® NVLink® Switch System, up to 256 H100 GPUs can be connected to accelerate exascale workloads, while the dedicated Transformer Engine supports trillion-parameter language models. H100 uses breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous generation.

Securely Accelerate Workloads from Enterprise to Exascale

NVIDIA H100 GPUs feature fourth-generation Tensor Cores and the Transformer Engine with FP8 precision, further extending NVIDIA's market-leading AI leadership with up to 9X faster training and an incredible 30X inference speedup on large language models. For high-performance computing (HPC) applications, H100 triples the floating-point operations per second (FLOPS) of FP64 and adds 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.

H100 is part of the complete NVIDIA data center solution that incorporates building blocks across hardware, networking, software, libraries, and optimized AI models and applications from the NVIDIA NGC™ catalog. Representing the most powerful end-to-end AI and HPC platform for data centers, it allows researchers to deliver real-world results and deploy solutions into production at scale.

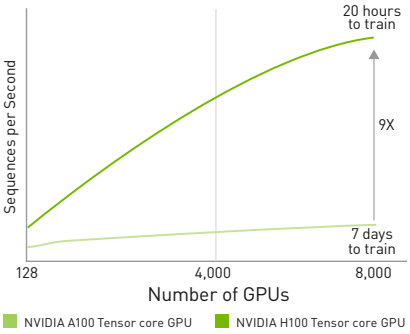
SPECIFICATIONS

	H100 SXM	H100 PCIe
FP64	**30 TFLOPS**	**24 TFLOPS**
FP64 Tensor Core	**60 TFLOPS**	**48 TFLOPS**
FP32	**60 TFLOPS**	**48 TFLOPS**
TF32 Tensor Core	**1,000 TFLOPS***	**800 TFLOPS***
BFLOAT16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP8 Tensor Core	**4,000 TFLOPS***	**3,200 TFLOPS***
INT8 Tensor Core	**4,000 TOPS***	**3,200 TOPS***
GPU memory	**80GB**	**80GB**
GPU memory bandwidth	**3TB/s**	**2TB/s**
Decoders	**7 NVDEC** **7 JPEG**	**7 NVDEC** **7 JPEG**
Max thermal design power (TDP)	**700W**	**350W**
Multi-Instance GPUs	**Up to 7 MIGS @ 10GB each**	
Form factor	**SXM**	**PCIe** **dual-slot air-cooled**
Interconnect	**NVLink: 900GB/s PCIe Gen5: 128GB/s**	**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**

* Shown with sparsity. Specifications 1/2 lower without sparsity.

Up to 9X Higher AI Training on Largest Models

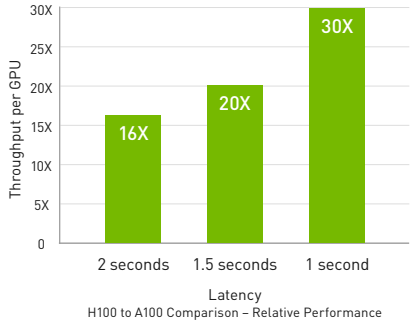
Mixture of Experts (395 Billion Parameters)



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

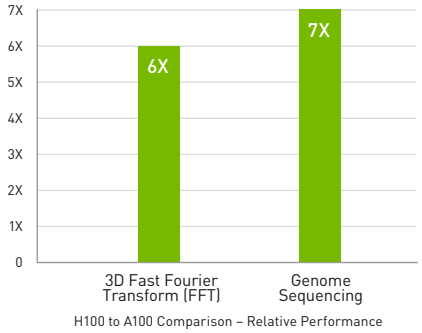
Up to 30X Higher AI Inference Performance on Largest Models

Megatron Chatbot Inference (530 Billion Parameters)



Projected performance subject to change. Inference on Megatron 530B parameter model chatbot for input sequence length=128, output sequence length =20 | A100 cluster: HDR IB network | H100 cluster: NDR IB network for 16 H100 configurations | 32 A100 vs 16 H100 for 1 and 1.5 sec | 16 A100 vs 8 H100 for 2 sec

Up to 7X Higher Performance for HPC Applications



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

The Technology Breakthroughs of NVIDIA Hopper

WORLD'S MOST ADVANCED CHIP

Built with 80 billion transistors using a cutting-edge TSMC 4N process custom tailored for NVIDIA's accelerated compute needs, H100 is the world's most advanced chip ever built. It 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 supports clusters of up to 256 H100s and delivers 9X higher bandwidth than InfiniBand HDR on the NVIDIA Ampere architecture.

NVIDIA CONFIDENTIAL COMPUTING

NVIDIA Confidential Computing is a built-in security feature of Hopper that makes NVIDIA H100 the world's first accelerator with confidential computing capabilities. Users can protect the confidentiality and integrity of their data and applications in use while accessing the unsurpassed acceleration of H100 GPUs.

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.

NVIDIA H100 CNX Converged Accelerator

NVIDIA H100 CNX combines the power of the NVIDIA H100 with the advanced networking capabilities of the **NVIDIA ConnectX®-7** smart network interface card (SmartNIC) in a single, unique platform. This convergence delivers unparalleled performance for GPU-powered IO-intensive workloads, such as distributed AI training in the enterprise data center and 5G processing at the edge. **Learn more about NVIDIA H100 CNX**.

Enterprise-Ready

The NVIDIA H100 Tensor Core GPU— powered by the NVIDIA Hopper architecture, the new engine for the world's AI infrastructure—is an integral part of the NVIDIA data center platform. Built for deep learning, HPC, and data analytics, the platform accelerates over 2,700 applications, including every major deep learning framework. Additionally, NVIDIA AI Enterprise, an end-to-end, cloud-native suite of AI and data analytics software, is certified to run on H100 in hypervisor-based virtual infrastructure with VMware vSphere. This enables management and scaling of AI workloads in a hybrid cloud environment. The complete NVIDIA platform is available everywhere, from data center to edge, delivering both dramatic performance gains and cost-saving opportunities.

OPTIMIZED SOFTWARE AND SERVICES FOR ENTERPRISE

NVIDIA CERTIFIED

EVERY DEEP LEARNING FRAMEWORK

mxnet PYTORCH APACHE Spark TensorFlow

2,000+ GPU-ACCELERATED APPLICATIONS

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF
- Simcenter STAR-CCM+

Ready to Get Started?

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

NVIDIA