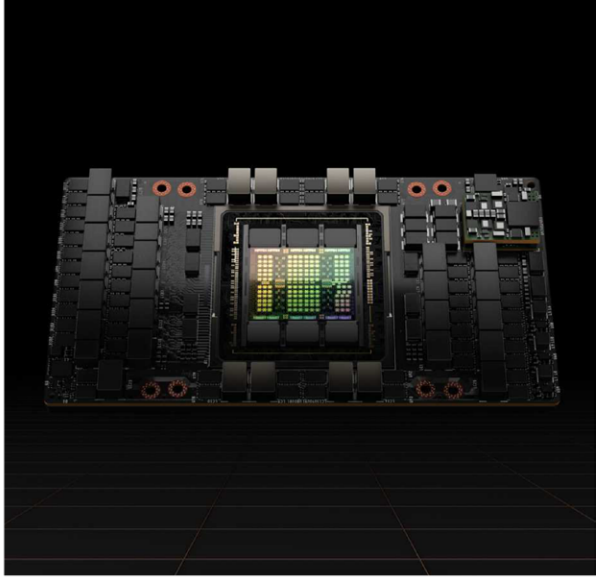
Datasheet

NVIDIA H100 Tensor Core GPU

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



Take an order-of-magnitude leap in 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 trillionparameter 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.

Ready for enterprise AI?

NVIDIA H100 Tensor Core GPUs for mainstream servers come with a five-year software subscription, including enterprise support, to the NVIDIA AI Enterprise software suite, simplifying AI adoption with the highest performance. This ensures organizations have access to the AI frameworks and tools they need to build H100-accelerated AI workflows such as AI chatbots, recommendation engines, vision AI, and more. Access the NVIDIA AI Enterprise software subscription and related support benefits for the NVIDIA H100 here.

Specifications		
	H100 SXM	**H100 PCIe**
FP64	34 TFLOPS	26 TFLOPS
FP64 Tensor Core	67 TFLOPS	51 TFLOPS
FP32	67 TFLOPS	51 TFLOPS
TF32 Tensor Core	989 TFLOPS*	756 TFLOPS*
BFLOAT16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP8 Tensor Core	3,958 TFLOPS*	3,026 TFLOPS*
INT8 Tensor Core	3,958 TOPS*	3,026 TOPS*
GPU memory	80GB	80GB
GPU memory bandwidth	3.35TB/s	2TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)
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 NVIDIACertified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs	Partner and NVIDIACertified Systems with 1–8 GPUs
NVIDIA AI Enterprise	Add-on	Included

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

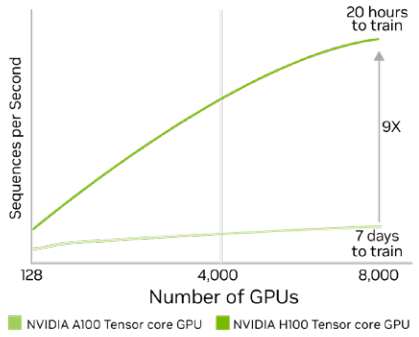
Securely accelerate workloads from enterprise to exascale.

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.

Technology Breakthroughs

Up to 9X higher AI training on largest models

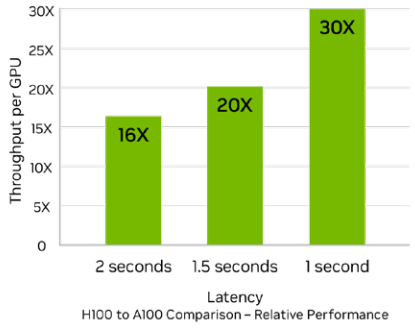
Mixture of Experts (395 Billion Parameters)



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

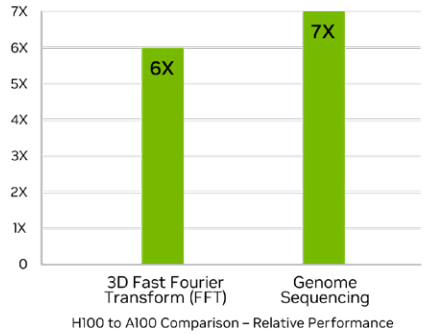
Up to 30X higher AI inference performance on largest models

Megatron Chatbot Inference (530 Billion Parameters)



Projected performance subject to change. Inference on Megatron 530B parameter model 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



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

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 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.

The convergence of GPU and SmartNIC.

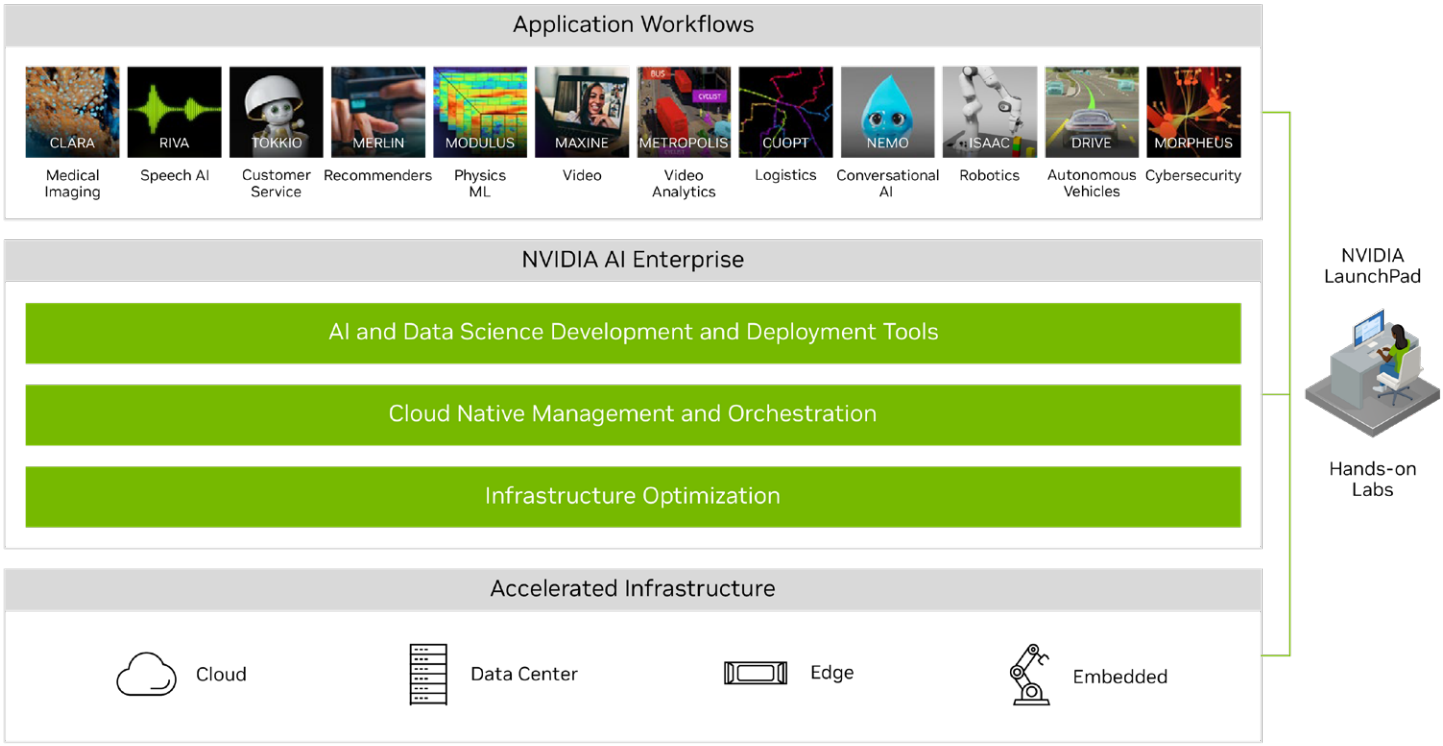
NVIDIA H100 CNX combines the power of the NVIDIA H100 with the advanced networking capabilities of the **NVIDIA ConnectX®-7** smart network interface card (SmartNIC) in a single, unique platform. This convergence delivers unparalleled performance for GPU-powered IO-intensive workloads, such as distributed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

Accelerate every workload, everywhere.

The NVIDIA H100 is an integral part of the NVIDIA data center platform. Built for AI, HPC, and data analytics, the platform accelerates over 3,000 applications, and is available everywhere from data center to edge, delivering both dramatic performance gains and cost-saving opportunities.

Deploy H100 with the NVIDIA AI platform.

NVIDIA AI is the end-to-end open platform for production AI built on NVIDIA H100 GPUs. It includes NVIDIA accelerated computing infrastructure, a software stack for infrastructure optimization and AI development and deployment, and application workflows to speed time to market. Experience NVIDIA AI and **NVIDIA H100 on NVIDIA LaunchPad** through free hands-on labs.



Ready to Get Started?

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