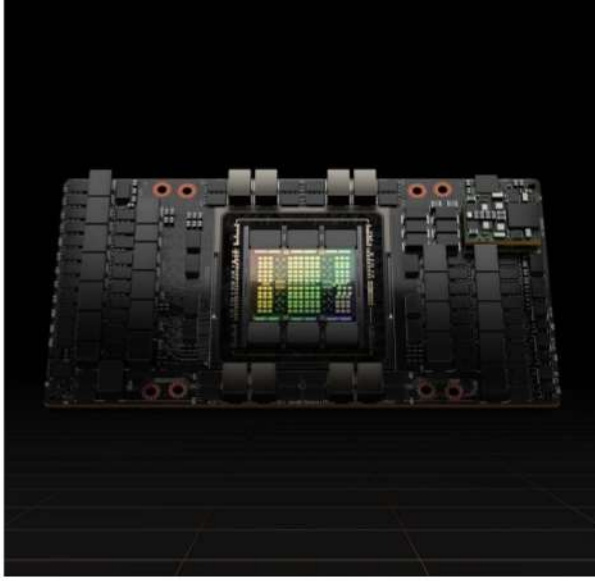
Datasheet

NVIDIA H100 Tensor Core GPU

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



Take 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.

Ready for Enterprise AI?

NVIDIA H100 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.

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 4X 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.

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.

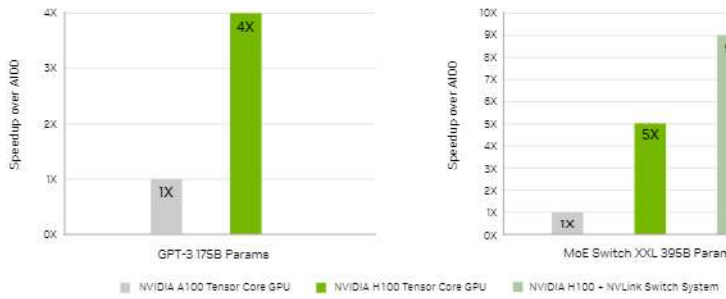
Technical Specifications			
	H100 SXM	**H100 PCIe**	**H100 NVL**[1]
FP64	34 teraFLOPS	26 teraFLOPS	68 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS	134 teraFLOPS
FP32	67 teraFLOPS	51 teraFLOPS	134 teraFLOPS
TF32 Tensor Core	989 teraFLOPS[2]	756 teraFLOPS[2]	1,979 teraFLOPS[2]
BFLOAT16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPS[2]
FP16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPS[2]
FP8 Tensor Core	3,958 teraFLOPS[2]	3,026 teraFLOPS[2]	7,916 teraFLOPS[2]
INT8 Tensor Core	3,958 TOPS[2]	3,026 TOPS[2]	7,916 TOPS[2]
GPU memory	80GB	80GB	188GB
GPU memory bandwidth	3.35TB/s	2TB/s	7.8TB/s[3]
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	14 NVDEC 14 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)	2x 350-400W (configurable)
Multi-instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 10GB each	Up to 14 MIGs @ 12GB each
Form factor	SXM	PCIe > dual-slot > air-cooled	2x PCIe > dual-slot > air-cooled
Interconnect	NVLink: > 900GB/s PCIe > Gen5: 128GB/s	NVLink: > 600GB/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	Partner and NVIDIA-Certified Systems with 2-4 pairs
NVIDIA Enterprise	Add-on	Included	Included

[1] Preliminary specifications. May be subject to change. Specifications shown for 2x H100 NVL PCIe cards paired with NVLink Bridge.

[2] With sparsity.

[3] Aggregate HBM bandwidth.

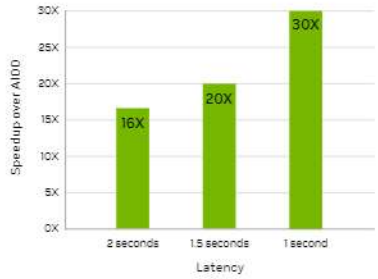
Up to 4X Higher AI Training on GPT-3



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 largest models

Megatron Chatbot Inference (530 Billion Parameters)

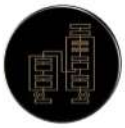


Inference on Megatron 530B parameter model chatbot for input sequence length=128, output sequence length =20 , A100 cluster: HDR IB network, H100 cluster: NDR IB network for 16 H100 configurations, 32 A100 vs 16 H100 for 1 and 1.5 sec, 16 A100 vs 8 H100 for 2 sec.

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.

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