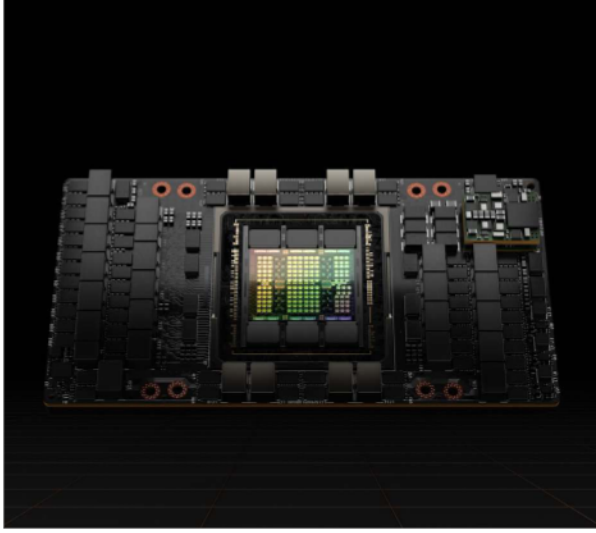
Datasheet

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.

PNY Part Number: NVH100NVLTCGPU-KIT

Technical Specifications	
	H100 NVL
FP64	30 teraFLOPS
FP64 Tensor Core	60 teraFLOPS
FP32	60 teraFLOPS
TF32 Tensor Core*	835 teraFLOPS
BFLOAT16 Tensor Core*	1,671 teraFLOPS
FP16 Tensor Core*	1,671 teraFLOPS
FP8 Tensor Core*	3,341 teraFLOPS
INT8 Tensor Core*	3,341 teraFLOPS
GPU Memory	94GB
GPU Memory Bandwidth	3.9TB/s
Decoders	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	350-400W (configurable)
Multi-Instance GPUs	Up to 7 MIGS @ 12GB each
Form Factor	PCIe dual-slot air-cooled
Interconnect	**NVIDIA NVLINK:** 600GB/s **PCIe Gen5:** 128GB/s
Server Options	Partner and NVIDIA Certified Systems with 1-8 GPUs
NVIDIA Enterprise	Included