NVIDIA H100 GPU

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

An Order-of-Magnitude Leap for Accelerated Computing

The NVIDIA H100 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 (LLMs) by 30X. H100 also includes a dedicated Transformer Engine to solve trillion-parameter language models.

Securely Accelerate Workloads From Enterprise to Exascale

Transformational AI Training

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. The combination of fourth-generation NVLink, which offers 900 gigabytes per second (GB/s) of GPU-to-GPU interconnect; NDR Quantum-2 InfiniBand networking, which accelerates communication by every GPU across nodes; PCIe Gen5; and NVIDIA Magnum IO™ software delivers efficient scalability from small enterprise systems to massive, unified GPU clusters.

Deploying H100 GPUs at data center scale delivers outstanding performance and brings the next generation of exascale high-performance computing (HPC) and trillion-parameter AI within the reach of all researchers.

Real-Time Deep Learning Inference

AI solves a wide array of business challenges, using an equally wide array of neural networks. A great AI inference accelerator has to not only deliver the highest performance but also the versatility to accelerate these networks.

H100 extends NVIDIA's market-leading inference leadership with several advancements that accelerate inference by up to 30X and deliver the lowest latency. Fourth-generation Tensor Cores speed up all precisions, including FP64, TF32, FP32, FP16, INT8, and now FP8, to reduce memory usage and increase performance while still maintaining accuracy for LLMs.

Exascale High-Performance Computing

The NVIDIA data center platform consistently delivers performance gains beyond Moore's law. And H100's new breakthrough AI capabilities further amplify the power of HPC+AI to accelerate time to discovery for scientists and researchers working on solving the world's most important challenges.

H100 triples the floating-point operations per second (FLOPS) of double-precision Tensor Cores, delivering 60 teraflops of FP64 computing for HPC. AI-fused HPC applications can also leverage H100's TF32 precision to achieve one petaflop of throughput for single-precision matrix-multiply operations, with zero code changes.

H100 also features new DPX instructions that deliver 7X higher performance over A100 and 40X speedups over CPUs on dynamic programming algorithms such as Smith-Waterman for DNA sequence alignment and protein alignment for protein structure prediction.

Accelerated Data Analytics

Data analytics often consumes the majority of time in AI application development. Since large datasets are scattered across multiple servers, scale-out solutions with commodity CPU-only servers get bogged down by a lack of scalable computing performance.

Accelerated servers with H100 deliver the compute power—along with 3 terabytes per second (TB/s) of memory bandwidth per GPU and scalability with NVLink and NVSwitch™—to tackle data analytics with high performance and scale to support massive datasets. Combined with NVIDIA Quantum-2 InfiniBand, Magnum IO software, GPU-accelerated Spark 3.0, and NVIDIA RAPIDS™, the NVIDIA data center platform is uniquely able to accelerate these huge workloads with higher performance and efficiency.

Enterprise-Ready Utilization

IT managers seek to maximize utilization (both peak and average) of compute resources in the data center. They often employ dynamic reconfiguration of compute to right-size resources for the workloads in use.

H100 with MIG lets infrastructure managers standardize their GPU-accelerated infrastructure while having the flexibility to provision GPU resources with greater granularity to securely provide developers the right amount of accelerated compute and optimize usage of all their GPU resources.

Built-In Confidential Computing

Traditional Confidential Computing solutions are CPU-based, which is too limited for compute-intensive workloads such as AI at scale. NVIDIA Confidential Computing is a built-in security feature of the NVIDIA Hopper architecture that made H100 the world’s first accelerator with these capabilities. With NVIDIA Blackwell, the opportunity to exponentially increase performance while protecting the confidentiality and integrity of data and applications in use has the ability to unlock data insights like never before. Customers can now use a hardware-based trusted execution environment (TEE) that secures and isolates the entire workload in the most performant way.

Built-In Confidential Computing

Traditional Confidential Computing solutions are CPU-based, which is too limited for compute-intensive workloads such as AI at scale. NVIDIA Confidential Computing is a built-in security feature of the NVIDIA Hopper architecture that made H100 the world’s first accelerator with these capabilities. With NVIDIA Blackwell, the opportunity to exponentially increase performance while protecting the confidentiality and integrity of data and applications in use has the ability to unlock data insights like never before. Customers can now use a hardware-based trusted execution environment (TEE) that secures and isolates the entire workload in the most performant way.

Exceptional Performance for Large-Scale AI and HPC

The Hopper GPU will power the NVIDIA Grace Hopper™ CPU+GPU architecture, purpose-built for terabyte-scale accelerated computing and providing 10X higher performance on large-model AI and HPC. The NVIDIA Grace CPU leverages the flexibility of the Arm® architecture to create a CPU and server architecture designed from the ground up for accelerated computing. The Hopper GPU is paired with the Grace CPU using NVIDIA’s ultra-fast chip-to-chip interconnect, delivering 900GB/s of bandwidth, 7X faster than PCIe Gen5. This innovative design will deliver up to 30X higher aggregate system memory bandwidth to the GPU compared to today's fastest servers and up to 10X higher performance for applications running terabytes of data.

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 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 Enterprise subscription. This subscription includes NVIDIA AI Enterprise to simplify 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 NIMTM, 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.

Product 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	80 GB	94 GB
GPU Memory Bandwidth	3.35 TB/s	3.9 TB/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 (MIG)	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 12GB each
Form Factor	SXM	PCIe dual-slot air-cooled
Interconnect	NVLink: 900 GB/s; PCIe Gen5: 128 GB/s	NVLink: 600 GB/s; PCIe Gen5: 128 GB/s
Server Options	HGX H100 (4 or 8 GPUs); DGX H100 (8 GPUs)	Partner & NVIDIA-Certified Systems (1–8 GPUs)
NVIDIA AI Enterprise	Add-on	Included