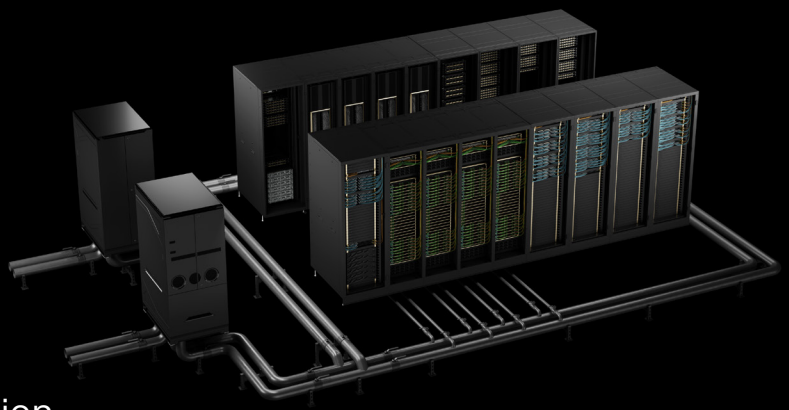
Datasheet



NVIDIA Blackwell

The engine of the new industrial revolution.

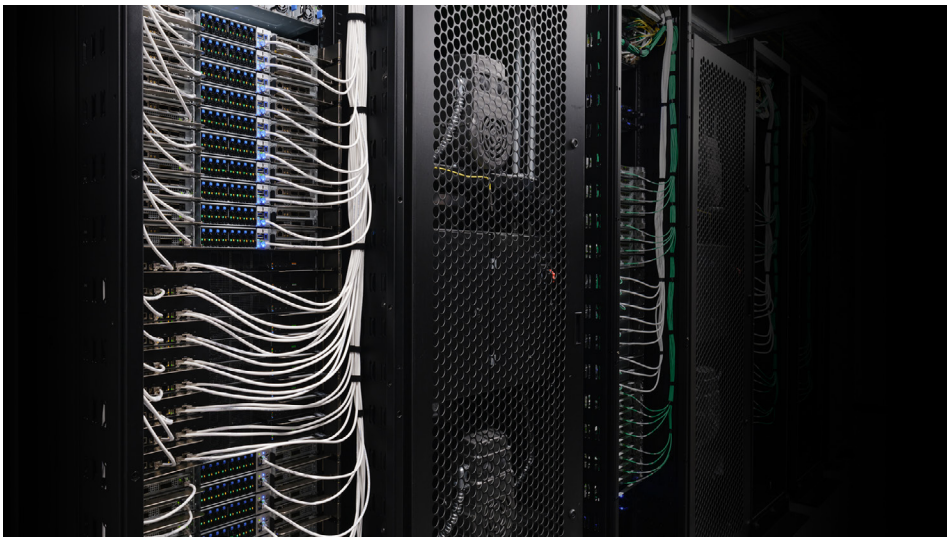
Breaking Barriers in Accelerated Computing

The NVIDIA Blackwell architecture introduces groundbreaking advancements for generative AI and accelerated computing. The incorporation of the second-generation Transformer Engine, alongside the faster and wider NVIDIA NVLink™ interconnect, propels the data center into a new era, with orders of magnitude more performance compared to the previous architecture generation. Further advances in NVIDIA Confidential Computing technology raise the level of security for real-time large language model (LLM) inference at scale without performance compromise. And NVIDIA Blackwell's new decompression engine combined with Spark RAPIDS™ libraries deliver unparalleled database performance to fuel data analytics applications. NVIDIA Blackwell's multiple advancements build upon generations of accelerated computing technologies to define the next chapter of generative AI with unparalleled performance, efficiency, and scale.

Key Offerings

- NVIDIA GB200 NVL72
- NVIDIA GB200 NVL4
- NVIDIA HGX B200

NVIDIA GB200 NVL72



Key Features

- 36 NVIDIA Grace CPUs
- 72 NVIDIA Blackwell GPUs
- Up to 17 terabytes (TB) of LPDDR5X memory with error-correction code (ECC)
- Supports up to 13.5 TB of HBM3E
- Up to 30.5 TB of fast-access memory
- NVLink domain: 130 terabytes per second (TB/s) of low-latency GPU communication

Powering the New Era of Computing

Unlocking Real-Time Trillion-Parameter Models

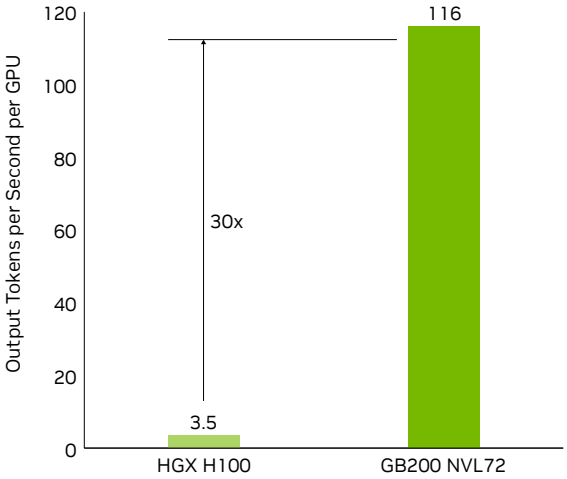
NVIDIA GB200 NVL72 connects 36 NVIDIA Grace™ CPUs and 72 NVIDIA Blackwell GPUs in an NVLink-connected, liquid-cooled, rack-scale design. Acting as a single, massive GPU, it delivers 30x faster real-time, trillion-parameter LLM inference.

The GB200 Grace Blackwell Superchip is a key component of the NVIDIA GB200 NVL72, connecting two high-performance NVIDIA Blackwell GPUs and an NVIDIA Grace CPU with the NVLink-C2C interconnect.

Real-Time LLM Inference

GB200 NVL72 introduces cutting-edge capabilities and a second-generation Transformer Engine which enables FP4. This advancement is made possible with a new generation of Tensor Cores, which introduce new microscaling formats, giving high accuracy and greater throughput. Additionally, the GB200 NVL72 uses NVLink and liquid cooling to create a single, massive 72-GPU rack that can overcome communication bottlenecks.

GPT-MoE-1.8T Real-Time Throughput

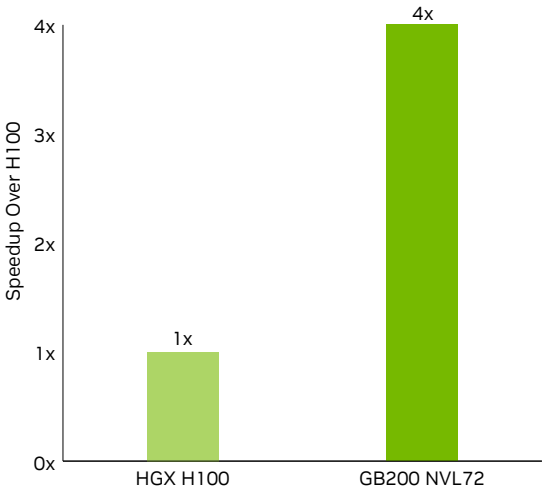


Projected performance, subject to change. LLM inference and energy efficiency: token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, 32,768 input/1,024 output, NVIDIA HGX H100 scaled over InfiniBand (IB) versus GB200 NVL72.

Massive-Scale Training

GB200 NVL72 includes a faster second-generation Transformer Engine featuring 8-bit floating point (FP8) precision, which enables a remarkable 4x faster training for large language models at scale. This breakthrough is complemented by the fifth-generation NVLink, which provides 1.8 TB/s of GPU-to-GPU interconnect, InfiniBand networking, and NVIDIA Magnum IO™ software.

GPT-MoE-1.8T Model Training Speedup

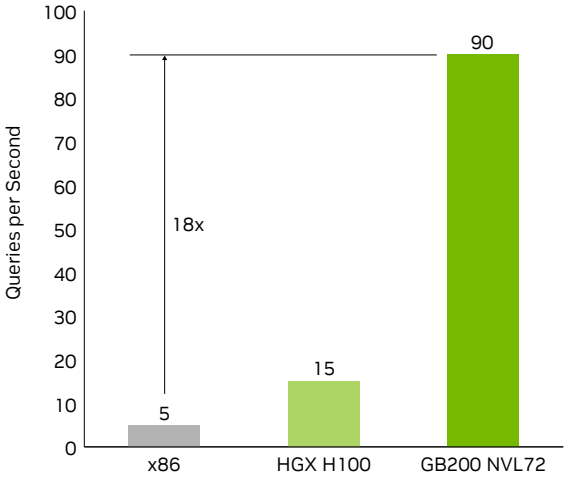


Projected performance, subject to change. Training GPT-MoE-1.8T - 4096x HGX H100 scaled over IB vs. 456x GB200 NVL72 scaled over IB. Cluster size: 32,768.

Data Processing

Databases play critical roles in handling, processing, and analyzing large volumes of data for enterprises. GB200 NVL72 takes advantage of the high-bandwidth-memory performance, NVLink-C2C, and dedicated decompression engines in the NVIDIA Blackwell architecture to speed up key database queries by 18x compared to CPU, delivering a 5x better TCO.

Database Join Query

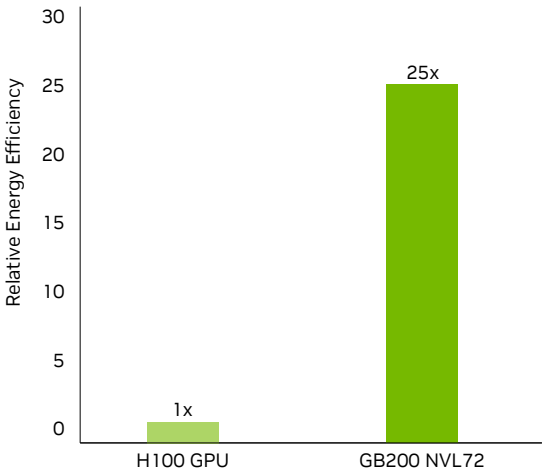


Projected performance, subject to change. Database join query throughput comparing GB200 NVL72, 72x H100, and 72 x86 CPUs

Energy-Efficient Infrastructure

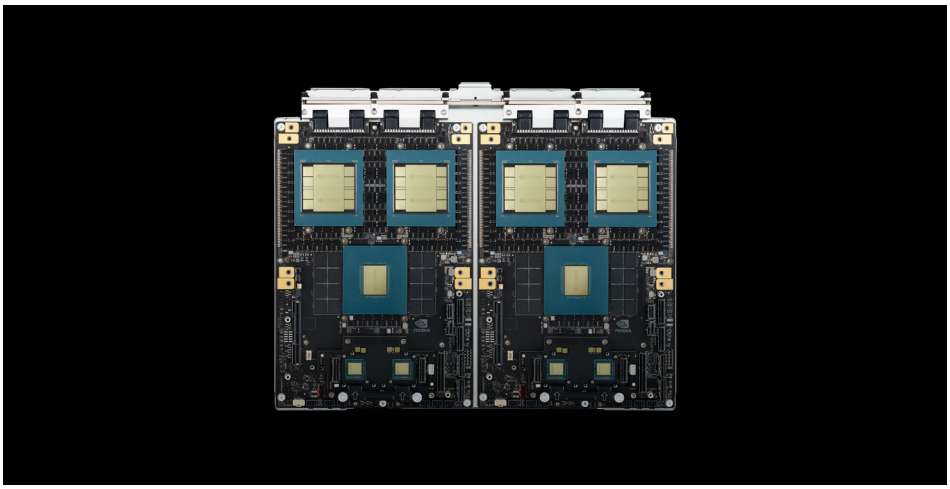
Liquid-cooled GB200 NVL72 racks reduce a data center's carbon footprint and energy consumption. Liquid cooling increases compute density, reduces the amount of floor space used, and facilitates high-bandwidth, low-latency GPU communication with large NVLink domain architectures. Compared to NVIDIA H100 air-cooled infrastructure, GB200 NVL72 delivers 25x more performance at the same power while reducing water consumption.

Energy Efficiency



Projected performance, subject to change. Energy savings for 65 racks eight-way HGX H100 air-cooled versus one rack GB200 NLV72 liquid-cooled with equivalent performance on GPT MoE 1.8T real-time inference throughput.

NVIDIA GB200 NVL4



Key Features

- Four NVIDIA Blackwell GPUs
- Two NVIDIA Grace CPUs
- 32 TB/s of bandwidth
- 1.8 TB of fast memory
- 2x faster high-performance compute (HPC) generation over generation

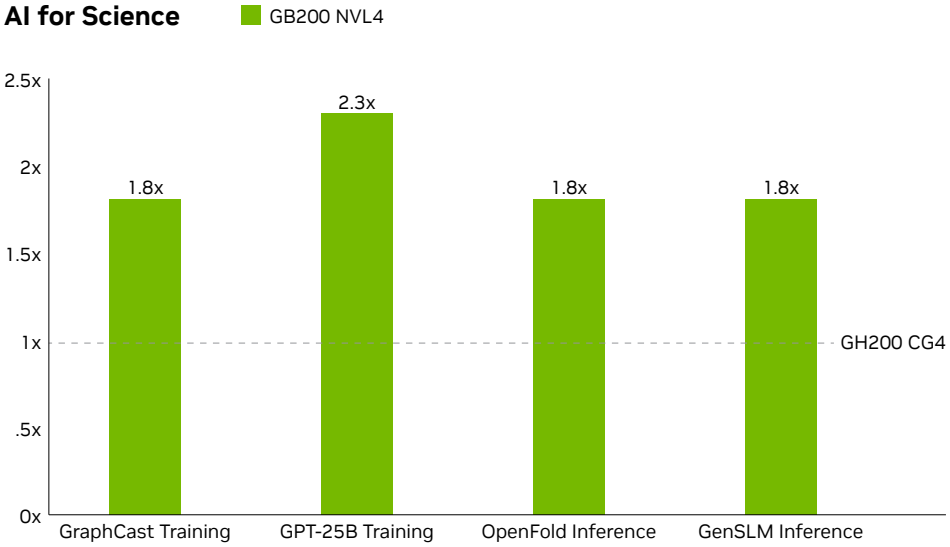
High-Performance Accelerator for Converged HPC and AI

NVIDIA GB200 NVL4 delivers revolutionary performance through four NVIDIA Blackwell GPUs connected with an NVLink bridge and two NVIDIA Grace CPUs connected over NVLink-C2C. Compatible with liquid-cooled, NVIDIA MGX™ modular servers, it provides up to 2x performance for scientific computing, AI model training, and inference over the prior generation.

Advancing Scientific Breakthroughs

Purpose-built for scientific computing, this platform unlocks the future of converged HPC and AI with 1.8 TB of coherent memory to enable breakthroughs like Nobel Prize-winning AlphaFold 2's protein structure prediction and advanced weather forecasting with data-driven AI models. These advances deliver transformative ROI by accelerating drug discovery and climate prediction timelines, bringing safer, more effective drugs to market faster and making accurate, timely forecasts that benefit everything from crisis management to agriculture.

AI for Science

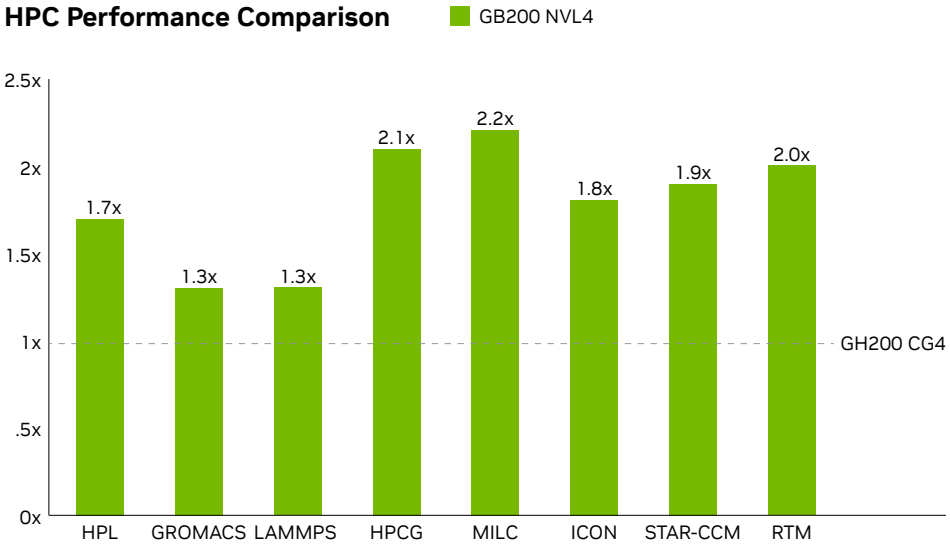


Projected performance on GB200 NVL4. Subject to change. GB200 NVL4 compared to GH200 CG4 (4x Grace Hopper Superchips). Assumes CG4 network at 200G/GPU and GB200 NVL4 network at 400G/GPU. Power assumptions CG4 = 700 W/GPU; NVL4 = 1200 W/GPU.

High-Performance Computing

High-performance computing is fueling the advancement of scientific computing. From weather forecasting and energy exploration to computational fluid dynamics and computer-aided engineering simulations, researchers are fusing traditional science with AI, machine learning, big data analytics, and edge computing to solve the mysteries of the world around us.

HPC Performance Comparison



Projected performance on GB200 NVL4. Subject to change. GB200 NVL4 compared to GH200 CG4 (4x Grace Hopper Superchips). Assumes CG4 network at 200G/GPU and GB200 NVL4 network at 400G/GPU. Power assumptions CG4 = 700 W/GPU; NVL4 = 1200 W/GPU.

NVIDIA HGX B200

Key Features

- Eight NVIDIA Blackwell GPUs
- 1.4 TB of HBM3E memory
- 1,800 GB/s NVLink between GPUs via NVIDIA NVSwitch™
- 15x faster real-time LLM inference
- 3x faster training performance

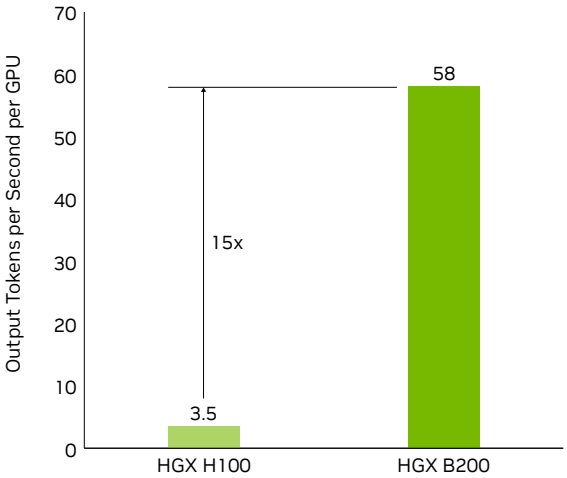
Propelling the Data Center Into a New Era of Accelerated Computing

The NVIDIA HGX B200 propels the data center into a new era of accelerating computing and generative AI, integrating NVIDIA Blackwell GPUs with a high-speed interconnect to accelerate AI performance at scale. As a premier accelerated scale-up x86 platform with up to 15x faster real-time inference performance, 12x lower cost, and 12x less energy use, HGX B200 is designed for the most demanding AI, data analytics, and HPC workloads.

Real-Time Inference for the Next Generation of Large Language Models

HGX B200 achieves up to 15x higher inference performance over the previous NVIDIA Hopper™ generation for massive models such as GPT MoE 1.8T. The second-generation Transformer Engine uses custom NVIDIA Blackwell Tensor Core technology combined with NVIDIA TensorRT™-LLM and NVIDIA NeMo™ Framework innovations to accelerate inference for LLMs and mixture-of-experts (MoE) models.

GPT-MoE-1.8T Real-Time Throughput

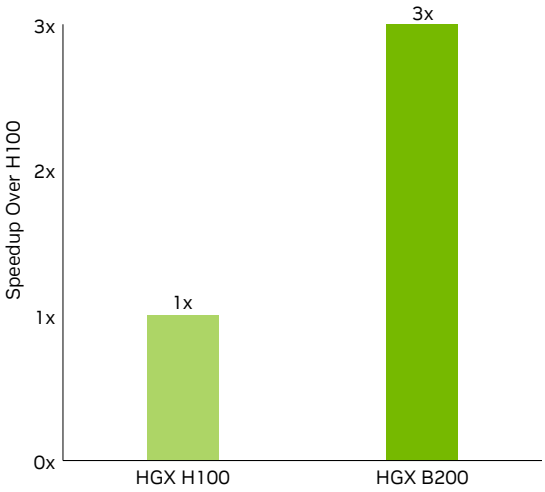


Projected performance, subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way HGX H100 GPUs air-cooled versus 1x eight-way HGX B200 air-cooled, per GPU performance comparison

Next-Level Training Performance

The second-generation Transformer Engine, featuring FP8 and new precisions, enables a remarkable 3x faster training for large language models like GPT MoE 1.8T. This breakthrough is complemented by fifth-generation NVLink with 1.8 TB/s of GPU-to-GPU interconnect, NVSwitch chip, InfiniBand networking, and NVIDIA Magnum IO software. Together, these ensure efficient scalability for enterprises and extensive GPU computing clusters.

GPT-MoE-1.8T Model Training Speedup



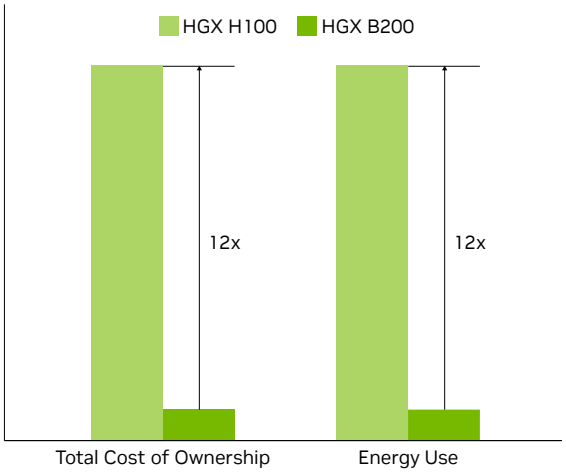
Projected performance, subject to change. 32,768 GPU scale, 4,096x eight-way HGX H100 air-cooled cluster: 400G IB network, 4,096x 8-way HGX B200 air-cooled cluster: 400G IB network.

Sustainable Computing

By adopting sustainable computing practices, data centers can lower their carbon footprints and energy consumption while improving their bottom line. The goal of sustainable computing can be realized with efficiency gains using accelerated computing with HGX. For LLM inference performance, HGX B200 improves energy efficiency by 12x and lowers costs by 12x compared to the NVIDIA Hopper generation.

12x Lower Energy Use and TCO

Lower is Better



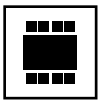
Projected performance, subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way HGX H100 GPUs air-cooled versus 1x eight-way HGX B200 air-cooled, per GPU performance comparison. TCO and energy savings for 100 racks eight-way HGX H100 air-cooled versus 8 racks eight-way HGX B200 air-cooled with equivalent performance.

Technical Specifications[1]

	GB200 NVL72	GB200 NVL4	HGX B200
NVIDIA Blackwell GPUs | Grace CPUs	72 | 36	4 | 2	8 | 0
CPU Cores	2,592 Arm® Neoverse V2 Cores	144 Arm Neoverse V2 Cores	-
Total NVFP4 Tensor Core[2]	1,440 | 720 PFLOPS	80 | 40 PFLOPS	144 | 72 PFLOPS
Total FP8/FP6 Tensor Core[2]	720 PFLOPS	40 PFLOPS	72 PFLOPS
Total Fast Memory	31 TB	1.8 TB	1.4 TB
Total Memory Bandwidth	576 TB/s	32 TB/s	62 TB/s
Total NVLink Bandwidth	130 TB/s	7.2 TB/s	14.4 TB/s
	Individual Blackwell GPU Specifications		
FP4 Tensor Core[2]	20 PFLOPS	20 PFLOPS	18 PFLOPS
FP8/FP6 Tensor Core[2]	10 PFLOPS	10 PFLOPS	9 PFLOPS
INT8 Tensor Core[2]	10 POPS	10 POPS	9 POPS
FP16/BF16 Tensor Core[2]	5 PFLOPS	5 PFLOPS	4.5 PFLOPS
TF32 Tensor Core[2]	2.5 PFLOPS	2.5 PFLOPS	2.2 PFLOPS
FP32	80 TFLOPS	80 TFLOPS	75 TFLOPS
FP64/FP64 Tensor Core	40 TFLOPS	40 TFLOPS	37 TFLOPS
GPU Memory | Bandwidth	186 GB HBM3E | 8 TB/s	186 GB HBM3E | 8 TB/s	180 GB HBM3E | 7.7 TB/s
Multi-Instance GPU (MIG)	7		
Decompression Engine	Yes		
Decoders	7 NVDEC[3] 7 nvJPEG		
Max Thermal Design Power (TDP)	Configurable up to 1,200 W	Configurable up to 1,200 W	Configurable up to 1,000 W
Interconnect	Fifth-generation NVLink: 1.8 TB/s PCIe Gen5: 128 GB/s		
Server Options	NVIDIA GB200 NVL72 partner and NVIDIA-Certified Systems™ with 72 GPUs	NVIDIA MGX partner and NVIDIA-Certified Systems	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs

1. Specifications in sparse | dense.
2. Specifications in sparse. Dense is one-half of the sparse spec shown.
3. Supported formats provide these speedups over NVIDIA H100 GPUs: 2x H.264, 1.25x HEVC, 1.25x VP9. AV1 support is new to NVIDIA Blackwell GPUs.

Explore the Technological Breakthroughs of NVIDIA Blackwell

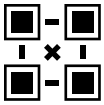


AI Superchip

NVIDIA Blackwell architecture GPUs pack 208 billion transistors and are manufactured using a custom-built TSMC 4NP process. All NVIDIA Blackwell products feature two reticle-limited dies connected by a 10 TB/s chip-to-chip interconnect in a unified single GPU.

Second-Generation Transformer Engine

The second-generation Transformer Engine uses custom NVIDIA Blackwell Tensor Core technology combined with NVIDIA TensorRT-LLM and NeMo Framework innovations to accelerate inference and training for LLMs and MoE models.



NVLink and NVLink Switch

The fifth-generation of NVIDIA NVLink can scale up to 576 GPUs to unleash accelerated performance for multi-trillion-parameter AI models. The NVIDIA NVLink Switch chip enables 130 TB/s of GPU bandwidth in one 72-GPU NVLink domain (NVL72) and delivers 4x bandwidth efficiency with NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ FP8 support.

RAS Engine

NVIDIA Blackwell adds intelligent resiliency with a dedicated reliability, availability, and serviceability (RAS) engine to identify potential faults that may occur early on to minimize downtime. NVIDIA's AI-powered predictive-management capabilities continuously monitor thousands of data points across hardware and software for overall health to predict and intercept sources of downtime and inefficiency.

Secure AI

NVIDIA Blackwell includes NVIDIA Confidential Computing, which protects sensitive data and AI models from unauthorized access with strong hardware-based security. Blackwell is the first TEE-I/O capable GPU in the industry, while providing the most performant confidential compute solution with TEE-I/O capable hosts and inline protection over NVIDIA NVLink.

Decompression Engine

NVIDIA Blackwell's Decompression Engine and ability to access massive amounts of memory in the NVIDIA Grace CPU over a high-speed link—900 GB/s of bidirectional bandwidth—accelerate the full pipeline of database queries for the highest performance in data analytics and data science with support for the latest compression formats such as LZ4, Snappy, and Deflate.

Automate the Essentials

NVIDIA Mission Control™ powers every aspect of NVIDIA GB200 NVL72 AI factory operations, from orchestrating workloads across the 72-GPU NVLink domain to integration with facilities. It brings instant agility for inference and training while providing full-stack intelligence for infrastructure resilience. Mission Control lets every enterprise run AI with hyperscale-grade efficiency, accelerating AI experimentation.

AI-Ready Enterprise Platform

NVIDIA AI Enterprise is the end-to-end software platform that brings generative AI into reach for every enterprise, providing the fastest and most efficient runtime for generative AI foundation models. It includes NVIDIA NIM™ inference microservices, AI frameworks, libraries, and tools that are certified to run on common data center platforms and mainstream NVIDIA-Certified Systems integrated with NVIDIA GPUs. Part of NVIDIA AI Enterprise, NVIDIA NIM™ is a set of easy-to-use inference microservices for accelerating the deployment of foundation models on any cloud or data center and helping to keep your data secure. Enterprises that run their businesses on AI rely on the security, support, manageability, and stability provided by NVIDIA AI Enterprise to ensure a smooth transition from pilot to production.

Together with the NVIDIA Blackwell GPUs, NVIDIA AI Enterprise not only simplifies the building of an AI-ready platform but also accelerates time to value.

Ready to Get Started?

To learn more about NVIDIA Blackwell, visit nvidia.com/blackwell

NVIDIA