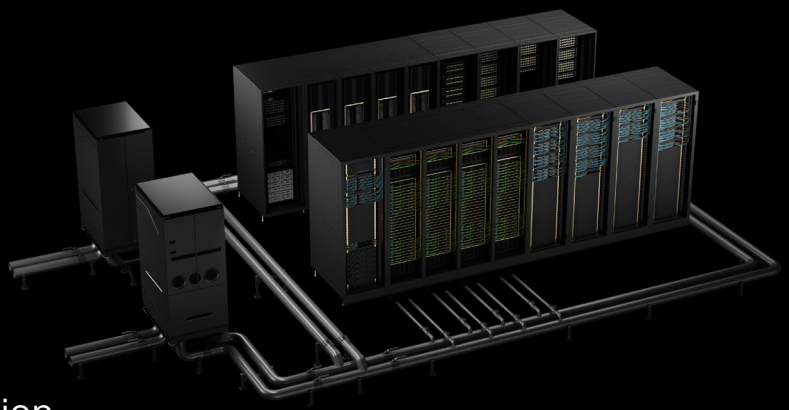
Datasheet



NVIDIA Blackwell

The engine of the new industrial revolution.

Breaking Barriers in Accelerated Computing

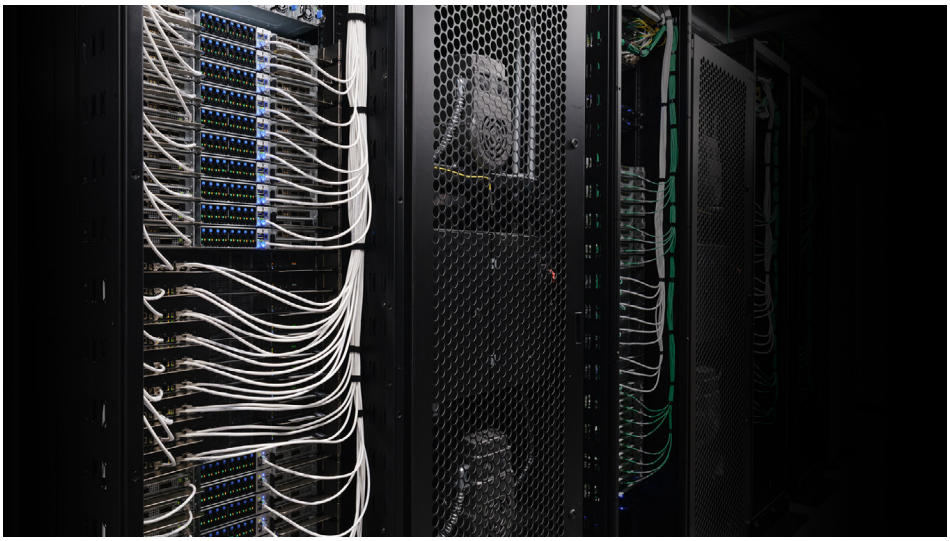
The NVIDIA Blackwell architecture introduces groundbreaking advancements for generative AI and accelerated computing. The incorporation of the second-generation Transformer Engine, alongside the faster and wider **NVIDIA NVLink™** interconnect, propels the data center into a new era, with orders of magnitude more performance compared to the previous architecture generation. Further advances in **NVIDIA Confidential Computing** technology raise the level of security for real-time LLM inference at scale without performance compromise. And Blackwell's new decompression engine combined with Spark RAPIDS™ libraries deliver unparalleled database performance to fuel data analytics applications. 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 HGX B200

NVIDIA GB200 NVL72

Powering the New Era of Computing



Unlocking Real-Time Trillion-Parameter Models

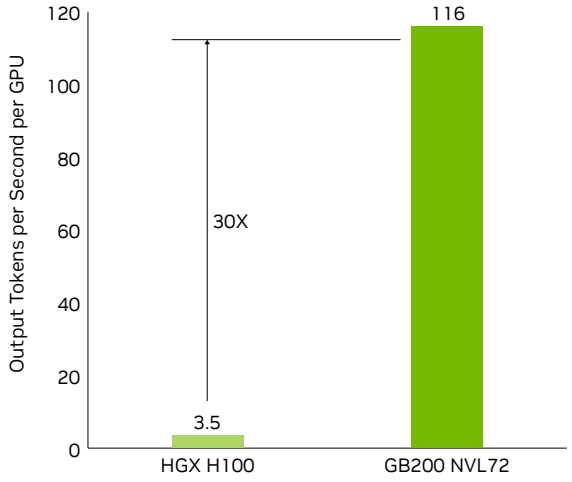
NVIDIA GB200 NVL72 connects 36 Grace CPUs and 72 Blackwell GPUs in an NVIDIA NVLink-connected, liquid-cooled, rack-scale design. Acting as a single, massive GPU, it delivers 30X faster real-time trillion-parameter large language model (LLM) inference.

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 AI. 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



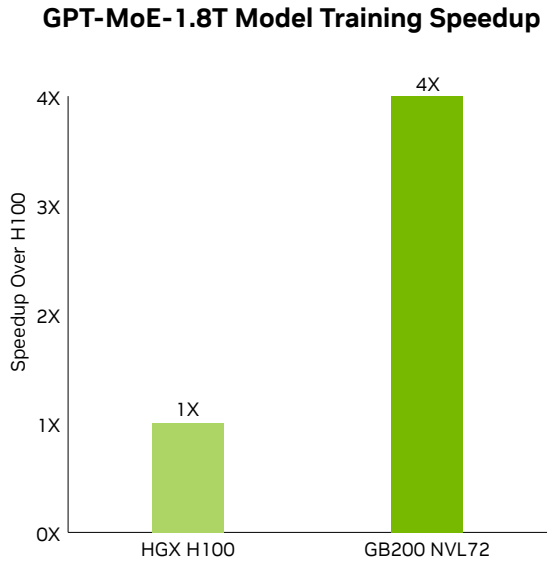
Projected performance, subject to change. LLM inference and energy efficiency: token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, 32,768 input/1,024 output, NVIDIA HGX™ H100 scaled over InfiniBand (IB) versus GB200 NVL72.

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 13.5TB of HBM3E
- Up to 30.5TB of fast-access memory
- NVLink domain: 130 terabytes per second (TB/s) of low latency GPU communication

Massive-Scale Training

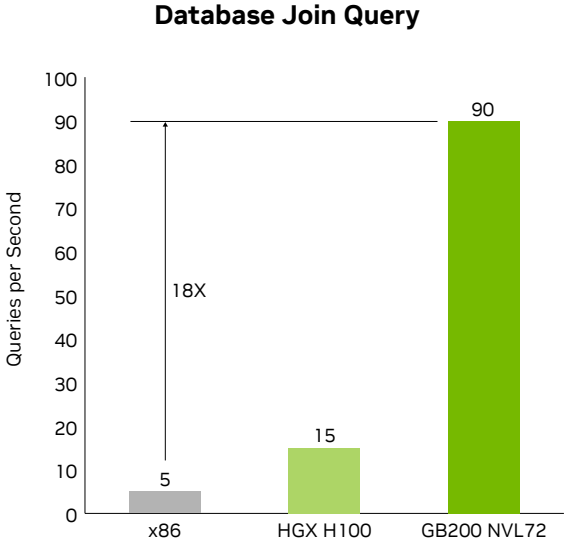
GB200 NVL72 includes a faster second-generation Transformer Engine featuring 8-bit floating point (FP8) precision, which enables a remarkable 4X faster training for large language models at scale. This breakthrough is complemented by the fifth-generation NVLink, which provides 1.8 terabytes per second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



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

Data Processing

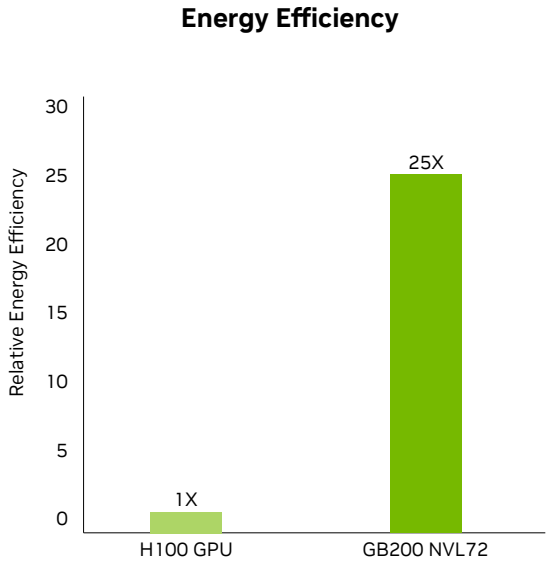
Databases play critical roles in handling, processing, and analyzing large volumes of data for enterprises. GB200 NVL72 takes advantage of the high-bandwidth-memory performance, NVLink-C2C, and dedicated decompression engines in the NVIDIA Blackwell architecture to speed up key database queries by 18X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



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 the **NVIDIA H100** air-cooled infrastructure, GB200 NVL72 delivers 25X more performance at the same power while reducing water consumption.



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 HGX B200

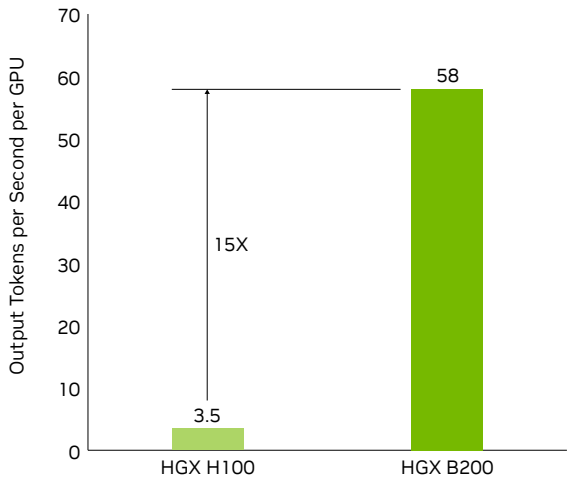
Propelling the Data Center Into a New Era of Accelerated Computing

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 high-performance computing (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 **Blackwell Tensor Core** technology combined with 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

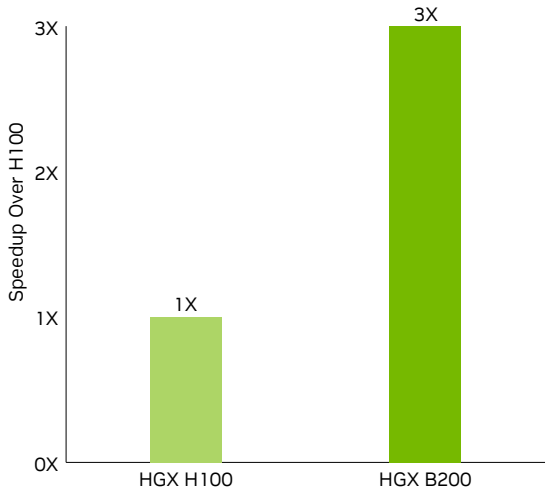
HGX B200 Key Features

- 8 NVIDIA Blackwell GPUs
- Up to 1.4 terabytes (TB) of HBM3E memory
- 1800GB/s NVLink between GPUs via NVSwitch™ chip
- 15X faster real-time LLM inference
- 3X faster training performance

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 1.8TB/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

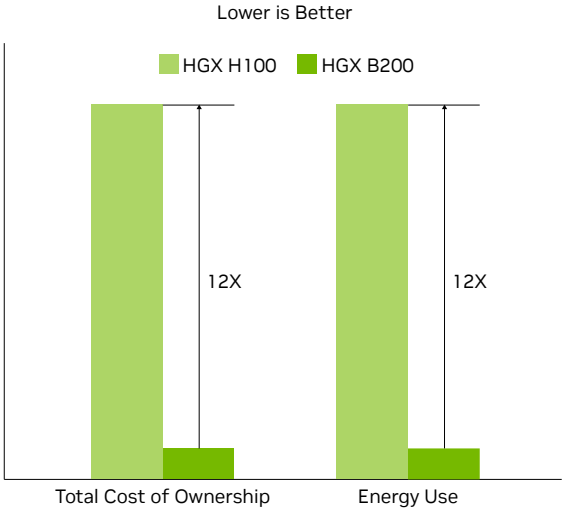


Projected performance, subject to change. 32,768 GPU scale, 4,096x eight-way HGX H100 air-cooled cluster: 400G IB network, 4,096x 8-way HGX B200 air-cooled cluster: 400G IB network.

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 **Hopper generation**.

12X Lower Energy Use and TCO



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	**HGX B200**
Blackwell GPUs | Grace CPUs	72 | 36	8 | 0
CPU Cores	2,592 Arm Neoverse V2 Cores	-
Total FP4 Tensor Core	1,440 petaFLOPS	144 petaFLOPS
Total FP8/FP6 Tensor Core	720 petaFLOPS/petaOPS	72 petaFLOPS/petaOPS
Total Fast Memory	Up to 30TB	Up to 1.4TB
Total Memory Bandwidth	Up to 576TB/s	Up to 62TB/s
Total NVLink Bandwidth	130TB/s	14.4TB/s
	Individual Blackwell GPU Specifications	
FP4 Tensor Core	20 petaFLOPS	18 petaFLOPS
FP8/FP6 Tensor Core	10 petaFLOPS	9 petaFLOPS
INT8 Tensor Core	10 petaOPS	9 petaOPS
FP16/BF16 Tensor Core	5 petaFLOPS	4.5 petaFLOPS
TF32 Tensor Core	2.5 petaFLOPS	2.2 petaFLOPS
FP32	80 teraFLOPS	75 teraFLOPS
FP64/FP64 Tensor Core	40 teraFLOPS	37 teraFLOPS
GPU Memory | Bandwidth	186GB HBM3e | 8 TB/s	180GB HBM3e | 7.7 TB/s
Multi-Instance GPU (MIG)	7	
Decompression Engine	Yes	
Decoders	7 NVDEC[2] 7 NVJPG	
Max Thermal Design Power (TDP)	Configurable up to 1,200W	Configurable up to 1,000W
Interconnect	5th Generation NVLink: 1.8TB/s PCIe Gen5: 128GB/s	
Server Options	NVIDIA GB200 NVL72 partner and NVIDIA-Certified Systems™ with 72 GPUs	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs

1. Preliminary specifications, subject to change. All Tensor Core numbers except FP64 with sparsity.
2. Supported formats provide these speed-ups over H100 Tensor Core GPUs: 2X H.264, 1.25X HEVC, 1.25X VP9. AV1 support is new to Blackwell GPUs.

AI Superchip

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

2nd Generation Transformer Engine

The second-generation Transformer Engine uses custom **Blackwell Tensor Core** technology combined with NVIDIA TensorRT-LLM and NeMo Framework innovations to accelerate inference and training for large language models (LLMs) and mixture-of-experts (MoE) models.

NVLink and NVLink Switch

The fifth-generation of NVIDIA NVLink interconnect can scale up to 576 GPUs to unleash accelerated performance for multi-trillion-parameter AI models. The NVIDIA NVLink Switch chip enables 130TB/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

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

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

Blackwell's decompression engine and ability to access massive amounts of memory in the **NVIDIA Grace CPU** over a high-speed link—900 gigabytes per second (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.

Sustainable Computing

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.

Learn about AI workload workflows with **NVIDIA AI Enterprise via NVIDIA Launchpad's hands-on labs**.

Ready to Get Started?

To learn more about the NVIDIA Blackwell, visit:
www.openzeka.com/blackwell

OPEN ZEKA

NVIDIA