Datasheet

NVIDIA Blackwell Ultra

Built for the age of AI reasoning.

NVIDIA GB300 NVL72

Designed for AI Reasoning Performance

AI has evolved around three fundamental scaling dimensions: pretraining, post-training, and inference-time scaling—also known as long thinking or reasoning. This third dimension is critical for enabling agentic AI, where models must dynamically reason through complex queries during inference. Unlike traditional one-shot inference, test-time scaling can demand up to 100x more compute, as models evaluate multiple potential responses before selecting the most accurate outcome.

The NVIDIA Blackwell Ultra GPU is designed for this new era of AI reasoning. It delivers up to 20 petaFLOPS of FP4 sparse inference performance, offering exceptional efficiency for large-scale deployments. With 279 GB of HBM3E memory, it supports expansive KV caching and long-context inference without offloading. Blackwell Ultra also features 800 Gbps NVIDIA® ConnectX®-8 networking, doubling interconnect bandwidth compared to NVIDIA Blackwell to enable seamless scaling across data centers. A newly optimized attention engine delivers 2.5x faster attention performance compared to NVIDIA Hopper™, significantly accelerating throughput for reasoning.

Key Offerings

- NVIDIA GB300 NVL72
- NVIDIA HGX B300

NVIDIA GB300 NVL72

Powering the New Era of AI Reasoning

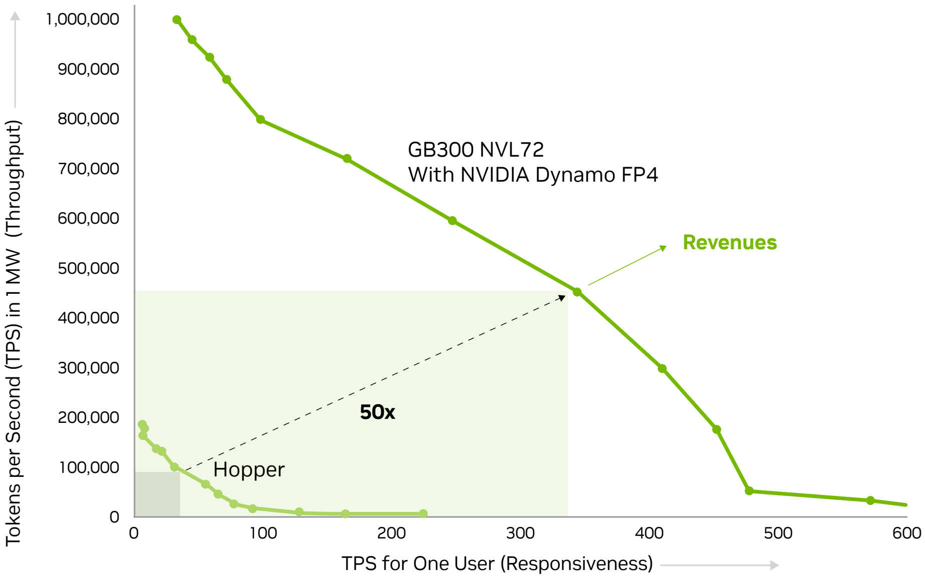
The NVIDIA GB300 NVL72 features a fully liquid-cooled, rack-scale architecture that integrates 72 NVIDIA Blackwell Ultra GPUs and 36 Arm®-based NVIDIA Grace™ CPUs into a single platform, purpose-built for test-time scaling inference or AI reasoning tasks. AI factories accelerated by the GB300 NVL72—leveraging NVIDIA Quantum-X800 InfiniBand or Spectrum-X™ Ethernet, ConnectX-8 SuperNIC™, and NVIDIA Mission Control™ management—deliver up to a 50x overall increase in AI factory output performance compared to Hopper-based platforms.

End-to-End AI Acceleration at Rack Scale

With 279 GB of HBM3E memory per Blackwell Ultra chip and up to 37 TB of high-speed memory per rack, coupled with 1.44 exaFLOPS of compute, and a 72-GPU unified NVIDIA NVLink™ domain, Blackwell Ultra provides unprecedented speed and scale to support larger models while giving rise to breakthroughs in AI. Combined with CUDA-X™ libraries for accelerated computing, NVIDIA accelerates the entire hardware and software computing stack.

Increase AI Factory Output Performance by 50x

The frontier curve illustrates key parameters that determine AI factory token revenue output. The vertical axis represents GPU tokens per second (TPS) throughput in one megawatt (MW) AI factory, while the horizontal axis quantifies user interactivity and responsiveness as TPS for a single user. At the optimal intersection of throughput and responsiveness, GB300 NVL72 yields a 50x overall increase in AI factory output performance compared to the Hopper architecture for maximum token revenue.



DeepSeek-R1 ISL = 32K, OSL = 8K, GB300 NVL72 with FP4 Dynamo disaggregation. H100 with FP8 In-flight batching. Projected performance subject to change.

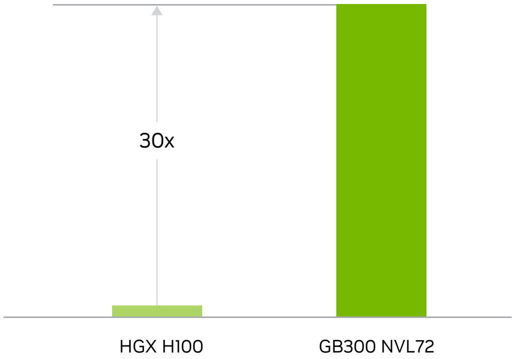
Accelerating Real-Time Video Generation by 30x

GB300 NVL72 introduces cutting-edge capabilities for diffusion-based video generation models. A single five-second video generation sequence processes 4 million tokens, requiring nearly 90 seconds to generate on NVIDIA Hopper GPUs. The Blackwell Ultra platform enables real-time video generation from world foundation models, such as NVIDIA Cosmos™, providing a 30x performance improvement versus Hopper. This allows the creation of customized, photo-realistic, temporally and spatially stable video for physical AI applications.

NVIDIA GB300 NVL72 Key Features

- 36 NVIDIA Grace CPUs
- 72 NVIDIA Blackwell Ultra GPUs
- 17 TB of LPDDR5X memory with error-correction code (ECC)
- 20 TB of HBM3E
- Up to 37 TB of fast-access memory
- NVLink domain: 130 terabytes per second (TB/s) of low-latency GPU communication

Real-Time Video Generation



Projected performance subject to change. Relative performance represented as frames per second per GPU. 5-second video generation using Cosmos-1.0-Diffusion-7B-Video2World 720p 60 FPS.

NVIDIA HGX B300

Purpose-Built for AI Reasoning

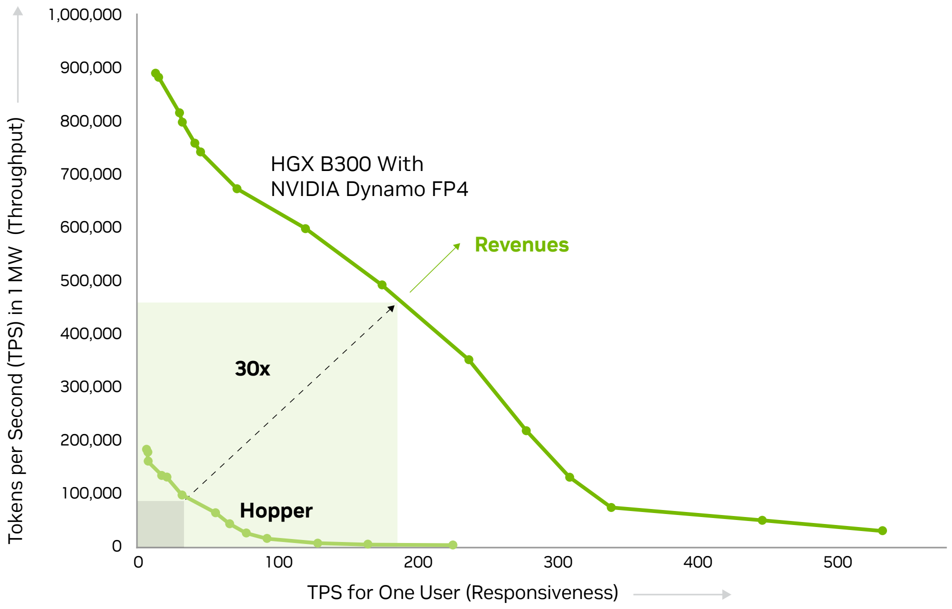
Key Features

- 8 NVIDIA Blackwell Ultra GPUs
- Over 2 TB of HBM3E memory
- 1,800 GBps NVLink between GPUs via NVSwitch™ chip
- 2.6x faster training performance (vs. H100)

NVIDIA HGX™ B300 is built for the age of AI reasoning with enhanced compute and increased memory. Featuring 7x more AI compute than the Hopper platform, over 2 TB of HBM3E memory, and high-performance networking integration with NVIDIA ConnectX-8 SuperNICs, HGX B300 delivers breakthrough performance on the most complex workloads from training, agentic systems, and reasoning, to real-time video generation for every data center.

Boost Revenue With HGX B300 AI Factory Output

The frontier curve illustrates key parameters that determine AI factory token revenue output. The vertical axis represents GPU tokens per second (TPS) throughput in one megawatt (MW) AI factory, while the horizontal axis quantifies user interactivity and responsiveness as TPS for a single user. At the optimal intersection of throughput and responsiveness, HGX B300 yields a 30x overall increase in AI factory output performance compared to the Hopper architecture for maximum token revenue.

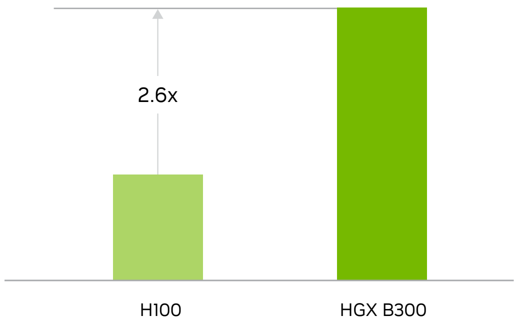


Projected performance subject to change. First token latency (FTL) = 2,000 ms, input sequence length = 32K, output sequence length = 8K.

Next-Level AI Training Performance

The HGX B300 platform delivers up to 2.6x higher training performance for large language models such as DeepSeek-R1. With over 2 TB of high-speed memory and 14.4 TB/s of NVLink Switch bandwidth, it enables massive-scale model training and high-throughput inter-GPU communication.

AI Training Performance – DeepSeek-R1



Projected performance subject to change. Perf per GPU, FP8, 16K BS, 16K sequence length.

Technical Specifications

	GB300 NVL72	HGX B300
Blackwell Ultra GPUs | Grace CPUs	72 | 36	8 | 0
CPU Cores	2,592 Arm Neoverse V2 Cores	-
Total FP4 Tensor Core[1]	1,440 PFLOPS | 1,080 PFLOPS	144 PFLOPS | 108 PFLOPS
Total FP8/FP6 Tensor Core[2]	720 PFLOPS	72 PFLOPS
Total Fast Memory	37 TB	2.1 TB
Total Memory Bandwidth	576 TB/s	62 TB/s
Total NVLink Switch Bandwidth	130 TB/s	14.4 TB/s
	Individual Blackwell Ultra GPU Specifications	
FP4 Tensor Core[1]	20 PFLOPS | 15 PFLOPS	18 PFLOPS | 14 PFLOPS
FP8/FP6 Tensor Core[2]	10 PFLOPS	9 PFLOPS
INT8 Tensor Core[2]	330 TOPS	307 TOPS
FP16/BF16 Tensor Core[2]	5 PFLOPS	4.5 PLFOPS
TF32 Tensor Core[2]	2.5 PFLOPS	2.2 PFLOPS
FP32	80 TFLOPS	75 TFLOPS
FP64/FP64 Tensor Core	1.3 TFLOPS	1.2 TFLOPS
GPU Memory | Bandwidth	279 GB HBM3E | 8 TB/s	270 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 1,400 W	Configurable up to 1,100 W
Interconnect	**Fifth-Generation NVLink:** 1.8 TB/s **PCIe Gen6:** 256 GB/s	
Server Options	NVIDIA GB300 NVL72 partner and NVIDIA-Certified Systems™	NVIDIA HGX B300 partner and NVIDIA-Certified Systems

1. Specification in Sparse | Dense
2. Specification in sparse. Dense is ½ sparse spec shown.
3. Supported formats provide these speed-ups over H100 GPUs: 2x H.264, 1.25x HEVC, 1.25x VP9. AV1

Technological Breakthroughs

AI Reasoning Inference

Test-time scaling and AI reasoning increase the compute necessary to achieve quality of service and maximum throughput. NVIDIA Blackwell Ultra’s Tensor Cores are supercharged with 2x the attention-layer acceleration and 1.5x more AI compute floating-point operations per second (FLOPS) compared to NVIDIA Blackwell GPUs.

NVIDIA Blackwell Architecture

The NVIDIA Blackwell architecture delivers groundbreaking advancements in accelerated computing, powering a new era of unparalleled performance, efficiency, and scale.

NVIDIA Grace CPU

The NVIDIA Grace CPU is a breakthrough processor designed for modern data center workloads. It provides outstanding performance and memory bandwidth with 2x the energy efficiency of today’s leading server processors.

High-Capacity HBM3E Architecture

Larger memory capacity allows for larger batch sizing and maximum throughput performance. NVIDIA Blackwell Ultra GPUs offer 1.5x larger HBM3E memory in combination with added AI compute compared to its predecessor, boosting AI reasoning throughput for the largest context lengths.

NVIDIA ConnectX-8 SuperNIC

The NVIDIA ConnectX-8 SuperNIC’s input/output (IO) module hosts two ConnectX-8 devices, providing 800 gigabits per second (Gb/s) of network connectivity for each GPU in the NVIDIA GB300 NVL72. This delivers best-in-class remote direct-memory access (RDMA) capabilities with either NVIDIA Quantum-X800 InfiniBand or Spectrum-X Ethernet networking platforms, enabling peak AI workload efficiency.

Fifth-Generation NVIDIA NVLink

Unlocking the full potential of accelerated computing requires seamless communication between every GPU. The fifth generation of NVIDIA NVLink is a scale–up interconnect that unleashes accelerated performance for AI reasoning models.

Automate the Essentials With NVIDIA Mission Control

NVIDIA Mission Control powers every aspect of NVIDIA GB300 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™ 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 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 NVIDIA Blackwell Ultra, visit
www.nvidia.com/blackwell