NVIDIA Blackwell Architecture Technical Brief

Built for the Age of AI Reasoning

***Updated to include the Blackwell Ultra GB300 Superchip, Blackwell Ultra GB300 NVL72 Rack-scale System, and HGX B300 Server System.**

V2.1

Table of Contents

List of Figures

List of Tables

NVIDIA GB300 NVL72 Built for the Age of AI Reasoning

For years, advancements in AI have followed a clear trajectory through pretraining scaling: larger models, more data, and greater computational resources lead to breakthrough capabilities. Building more intelligent systems is no longer just about pretraining bigger models. Instead, it's about refining them and making them able to think and reason.

By refining AI models to specialized tasks, post-training scaling improves models to deliver more conversational responses. Tuning models with domain-specific and synthetic data enhances their ability to understand nuanced contexts and provide accurate outputs. Synthetic data generation has no upper limit which translates to a significant need for compute resources in post-training scaling.

Now, a new scaling law has emerged to amplify intelligence: **test-time scaling**.

Also known as long thinking, test-time scaling dynamically increases compute during AI inference to enable deeper reasoning. AI reasoning models don't just generate responses in a single pass, they actively think, weigh multiple possibilities, and refine answers in real time.

This shift towards post-training scaling and test-time scaling demands exponentially more compute, real-time processing, and high-speed interconnects. Post-training can require 30x more compute than pretraining to develop customized derivative models, and long thinking can require 100x more compute than a single inference pass to solve for incredibly complex tasks.

Enter NVIDIA Blackwell architecture, built with the specific purpose of handling data center-scale reasoning AI workflows with up to 30X the energy efficiency of the prior NVIDIA Hopper GPU generation.

This technical brief introduces the benefits of NVIDIA Blackwell in detail, including the Blackwell Ultra GPU, the GB300 NVL72 Rack-scale System, and the HGX B300 server. GB200 NVL72 and HGX B200 system information is also included.

NVIDIA Blackwell and Blackwell Ultra Overview

The NVIDIA Blackwell and Blackwell Ultra products are designed to address the needs of ever-increasing AI complexity, including larger model sizes and AI reasoning, with a long list of new innovations.

With NVIDIA Blackwell and Blackwell Ultra products, every enterprise can use and deploy state-of-the-art LLMs with affordable economics, optimizing their business with the benefits of reasoning AI. At the same time, NVIDIA Blackwell and Blackwell Ultra products enable the next era of AI models, supporting high throughput with real-time performance, something unattainable without Blackwell's architectural innovations.

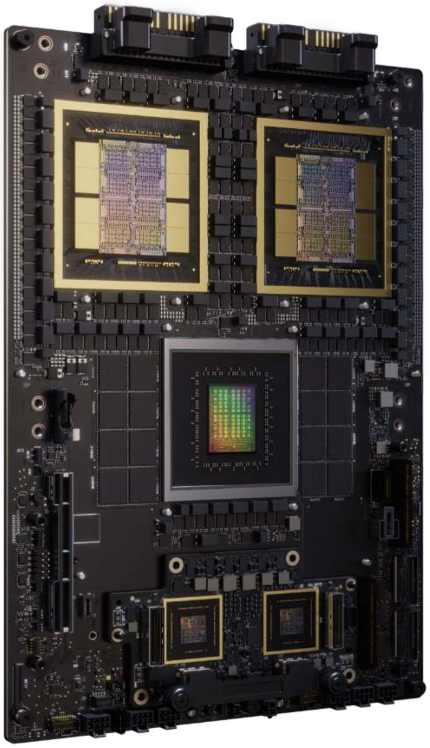


Figure 1. NVIDIA Grace Blackwell Ultra Superchip with ConnectX-8 SuperNICs

NVIDIA Blackwell Architectural Innovations

The Blackwell architecture introduces groundbreaking advancements for AI reasoning and accelerated computing with Blackwell and Blackwell Ultra products. The incorporation of a new second-generation Transformer Engine, alongside faster and wider NVIDIA® NVLink® interconnects, 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 AI inference at scale without compromising performance. And NVIDIA Blackwell's new Decompression Engine combined with Spark RAPIDS™ libraries delivers 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 reasoning AI with unparalleled performance, efficiency, and scale.

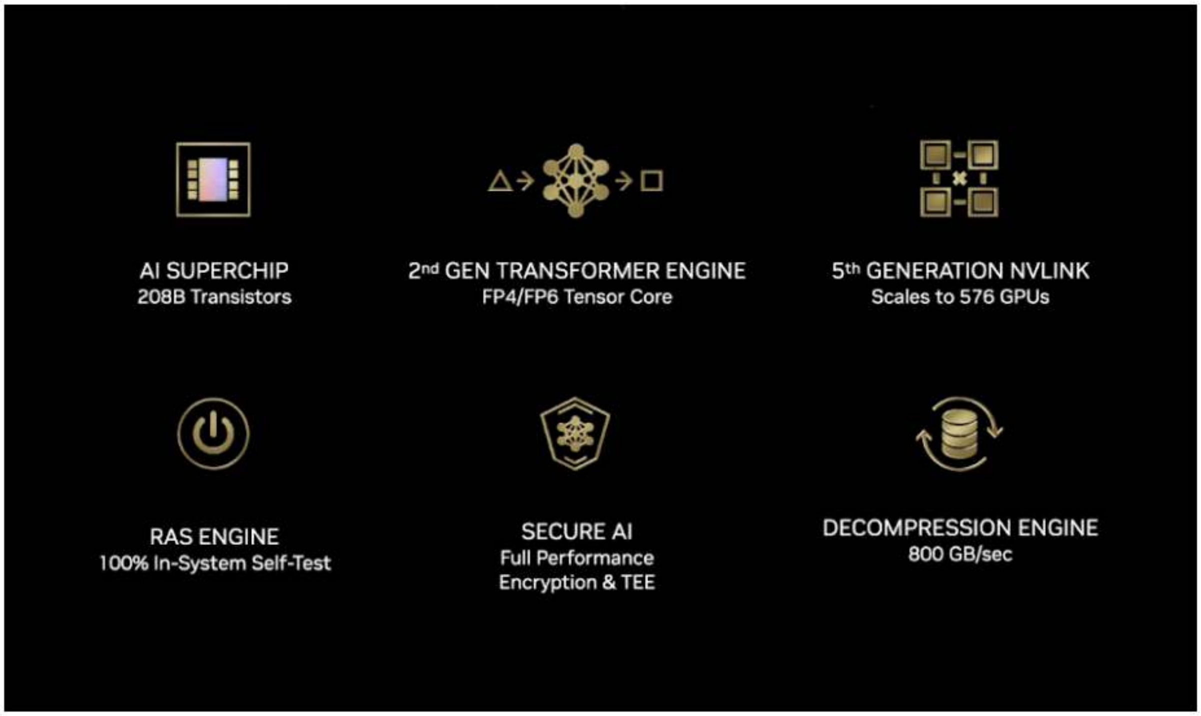


Figure 2. NVIDIA Blackwell Architecture's Technological Breakthroughs

A New Class of AI GPU

Built with 208 billion transistors, more than 2.5x the amount of transistors in NVIDIA Hopper GPUs, and using TSMC's 4NP process tailored for NVIDIA, Blackwell is the largest GPU ever built. NVIDIA Blackwell achieves the highest compute ever on a single chip, 20 petaFLOPS.

This architecture can incorporate a significant amount of computing power by merging two GPU dies. Each of the two GPU dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent chip.

The Blackwell architecture is much more than a chip with high floating-point operations per second (FLOPS) computational rates. It continues to build upon and benefit from NVIDIA's rich ecosystem of development tools, CUDA-X™ libraries, over four million developers, and over 3,000 applications scaling performance across thousands of nodes.

The new NVIDIA Blackwell Ultra GPU is built for the age of AI reasoning with enhanced compute and increased memory. The GB300 NVL72 delivers 1.5x more AI performance than the NVIDIA GB200 NVL72, as well as 50X more AI reasoning productivity, 35X faster AI reasoning inference, 30x higher energy efficiency, and 25x lower cost per token compared with NVIDIA Hopper™ systems.

Blackwell Tensor Core Architecture

Tensor Cores are specialized high-performance compute cores for matrix multiply and accumulate (MMA) math operations that provide groundbreaking performance for AI and HPC applications. Tensor Cores operating in parallel across SMs in one NVIDIA GPU deliver massive increases in throughput and efficiency compared to standard Floating-Point (FP), Integer (INT), and FMA (Fused Multiply-Accumulate) operations. Tensor Cores were first introduced in the NVIDIA Tesla® V100 GPU, and further enhanced in each new NVIDIA GPU architecture generation.

As generative and reasoning AI models increase in size and complexity, it's critical to improve training and inference performance. To meet these compute needs, new fifth-generation Tensor Core architecture in Blackwell supports new number formats such as FP4, including community-defined microscaling (OCP) formats. The Blackwell architecture supports all the data types and numerical formats listed in Table 1.

Table 1. Blackwell Architecture Supported Data Types

Data Type
FP64
FP32
TF32
FP16
BF16
FP8
INT8
FP6
FP4

Second-Generation Transformer Engine

Blackwell introduces the new second-generation Transformer Engine. The second-generation Transformer Engine uses custom Blackwell Tensor Core technology combined with NVIDIA Dynamo, TensorRT-LLM and Nemo Framework innovations to accelerate inference and training for LLMs, AI reasoning, and Mixture-of-Experts (MoE) models.

The Blackwell Transformer Engine utilizes advanced dynamic range management algorithms and fine-grain scaling techniques, called micro-tensor scaling, to optimize inference performance, accuracy, and enable FP4 AI. This doubles the performance of Blackwell's FP4 Tensor Core, doubles the parameter bandwidth to the HBM memory, and doubles the size of models supported per GPU.

Innovations in Dynamo and TensorRT-LLM, including quantization to 4-bit precision, custom kernels with expert parallelism mapping, and disaggregation are democratizing today's MoE models for real-time inference, using less hardware and less energy, with less cost.

For training, the second-generation Transformer Engine works with Nemo Framework and Megatron-Core innovations in new expert parallelism techniques that combine with other parallelism techniques and fifth-generation NVLink for unprecedented model performance. Lower precision formats open possibilities for further acceleration of large-scale training.

With the Blackwell second-generation Transformer Engine, enterprises can use and deploy state-of-the-art AI reasoning models with affordable economics, optimizing their business with the benefits of generative AI. NVIDIA Blackwell makes the next era of AI reasoning models possible—supporting both training and real-time inference.

Attention Layer Acceleration

The Blackwell Ultra GPU provides a 2X speedup over Blackwell GPUs for attention layer compute with new instructions to improve the performance of long input sequences. Doubling the attention operations using the Blackwell Ultra GPU architecture enhances AI performance by reducing latency and enabling faster and more intelligent decision-making for AI reasoning models. This acceleration also helps lower compute costs by reducing processing time, leading to energy and infrastructure savings. Enterprises can scale more efficiently, handling larger workloads with the same resources, ultimately driving greater efficiency, cost savings, and a competitive advantage in AI-driven business operations.

Performant Confidential Computing and Secure AI

Generative AI holds tremendous potential for businesses. Optimizing revenue, providing business insights, and aiding in generative content are only a few of the benefits. But adoption of generative AI can be difficult for businesses that need to train them on private data that can be subject to privacy regulations or includes proprietary information.

NVIDIA Confidential Computing capabilities extend the Trusted Execution Environment (TEE) beyond CPUs to GPUs. Confidential Computing on NVIDIA Blackwell was architected to deliver the fastest, most secure, and attestable (evidence-based) protections for LLMs and other sensitive data. NVIDIA Blackwell introduces the first TEE-I/O capable GPU in the industry, while providing the most performant confidential compute solution with TEE-I/O capable hosts, as well as inline protection over NVLink (providing confidentiality plus integrity).

Blackwell Confidential Computing delivers nearly identical throughput performance as compared to unencrypted modes. Customers can now secure even the largest models in a performant way, in addition to protecting AI intellectual property (IP) and securely enable confidential AI training, inference, and federated learning.

Fifth-Generation NVLink and NVLink Switch

Unlocking the full potential of exascale computing and AI reasoning models hinges on the need for swift, seamless communication among every GPU within a server cluster. The fifth generation of NVLink scales up to 576 GPUs to accelerate performance for reasoning AI models thanks to the NVLink Switch chips built with it. Fifth-generation NVLink doubles the performance of fourth-generation NVLink in NVIDIA Hopper. While the new NVLink in Blackwell and Blackwell Ultra GPUs also uses two high-speed differential pairs in each direction to form a single link as in the Hopper GPU, NVIDIA Blackwell architecture doubles the effective bandwidth per link to 50 GB/sec in each direction.

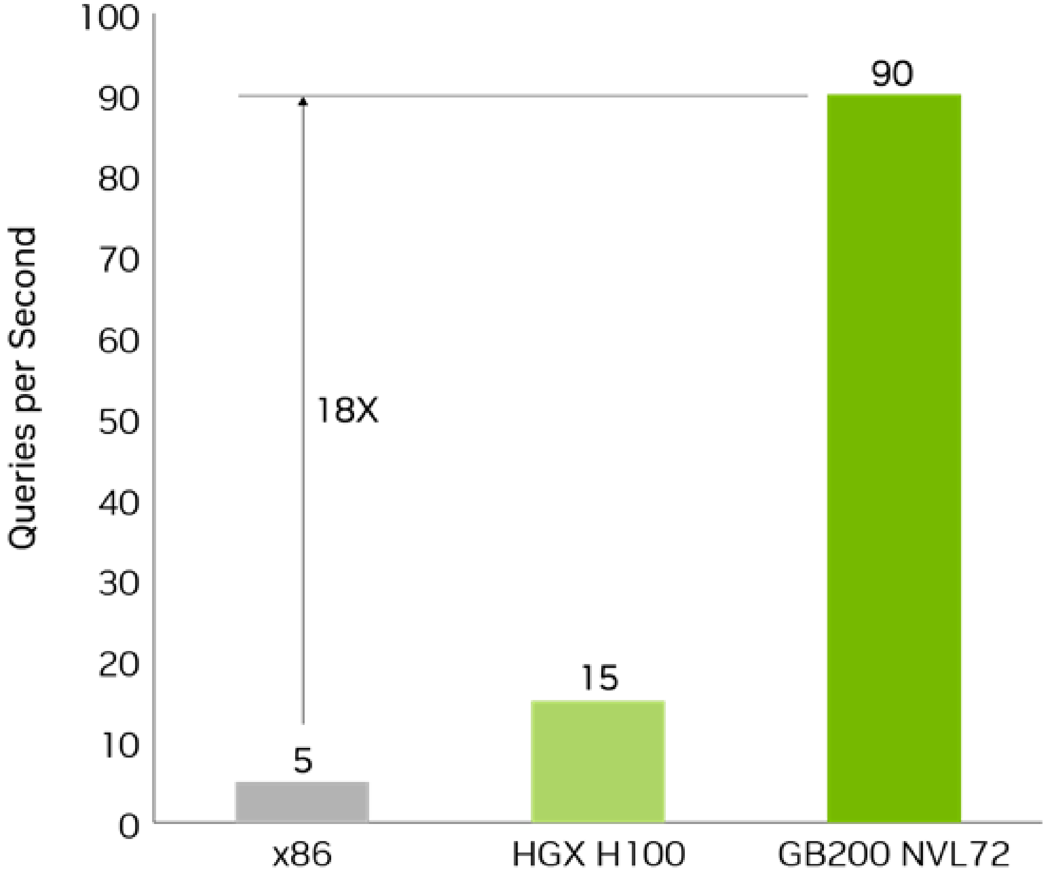
Blackwell and Blackwell Ultra GPUs include 18 fifth-generation NVLink links to provide 1.8 TB/sec total bandwidth, 900 GB/sec in each direction. 1.8TB/s of bidirectional throughput per GPU is over 14X the bandwidth of PCIe Gen5, ensuring high-speed communication for today's most complex large models. Within a 72 GPU NVLink domain, 130 TB/s of aggregate bandwidth is transferred - that's more data movement than the entire Internet.

The NVIDIA NVLink Switch enables 130TB/s GPU bandwidth in one 72 GPU NVLink domain (NVL72) for model parallelism, and delivers 4X bandwidth efficiency with new NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ FP8 support. NVLink and NVLink Switch used together support clusters beyond a single server at the same impressive 1.8 TB/s interconnect. Multi-server clusters using NVLink Switch can scale GPU communications in balance with the increased computing, enabling the GB300 NVL72 to support 9X the GPU throughput as compared to a single eight-GPU system.

Decompression Engine

Data analytics and database workflows have traditionally been slow and cumbersome, relying on CPUs for compute. Accelerated data science can dramatically boost the performance of end-to-end analytics, speeding up value generation and time to insights while reducing cost. Databases play critical roles in handling, processing, and analyzing large volumes of data for data analytics. Blackwell architecture's new dedicated

Decompression Engine can decompress data at a rate of up to 800GB/s. In combination with 8TB/s of HBM3e (High Bandwidth Memory) using one GPU in GB200 and the Grace CPU's high-speed NVLink-C2C (Chip-to-Chip) interconnect, Blackwell and Blackwell Ultra 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, NVIDIA Blackwell is18X faster than CPUs and 6X faster than NVIDIA H100 GPUs for query benchmarks.



Projected performance subject to change. Database join and aggregation workload with Snappy / Deflate compression derived from TPC-H Q4 query. Custom query implementations for x86, HGX H100 single GPU, and single GPU from GB200 Grace Blackwell Superchip

Figure 3. GB200 Grace Blackwell Database Join Query Using Decompression Engine

RAS Engine

Blackwell architecture 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. This builds intelligent resilience that saves time, energy, and computing costs.

NVIDIA's RAS engine provides in-depth diagnostic information that can identify areas of concern and plan for maintenance. The RAS engine reduces turnaround time by quickly localizing the source of issues and minimizes downtime by facilitating effective remediation. Administrators can flexibly adjust compute resources and optimal checkpoint strategies to facilitate uninterrupted large-scale training jobs. If the RAS engine identifies that a replacement component is needed, stand-by capacity is activated to ensure work finishes on time with the least performance degradation. Any hardware replacements that are required can be scheduled to avoid unplanned downtime.

NVIDIA Grace Blackwell Ultra / Blackwell NVL72 Rack-scale Systems

Table 2. System Specifications for GB300 NVL72 and GB200 NVL72

System Spec	GB300 NVL72	GB200 NVL72
NVL72 Rack Configuration	72 Grace Blackwell Ultra GPUs	72 Grace Blackwell GPUs
Compute Trays	18	18
NVLink Switch Trays	9	9
GPUs | CPUs Per Compute Node	4 Blackwell GPUs | 2 Grace CPUs	4 Blackwell GPUs | 2 Grace CPUs
Total Rack GPUs | CPUs	72 | 36	72 | 36
FP4 Tensor Core Dense/Sparse	1,080 / 1,440 petaFLOPS	720 / 1,440 petaFLOPS
FP8/FP6 Tensor Core Dense/Sparse	360 / 720 petaFLOPS	360 / 720 petaFLOPS
INT8 Tensor Core Dense/Sparse	12/ 24 petaOPS	360 / 720 petaOPS
FP16/BF16 Tensor Core Dense/Sparse	180 / 360 petaFLOPS	180 / 360 petaFLOPS
TF32 Tensor Core Dense/Sparse	90 / 180 petaFLOPS	90 / 180 petaFLOPS
FP32	5,760 teraFLOPS	5,760 teraFLOPS
FP64/FP64 Tensor Core	100 teraFLOPS	2,880 teraFLOPS
HBM Memory Architecture	HBM3e	HBM3e
HBM Memory Size	20 TB	13.5 TB
HBM Memory Bandwidth	576 TB/s	Up to 576 TB/s
Fast Memory	37 TB	Up to 31 TB
NVLink Switch	9 NVLink Switches	9 NVLink Switches
NVLink Bandwidth Bidirectional	130 TB/s	130 TB/s
CPU Cores	2592 Arm Neoverse V2 cores	2592 Arm Neoverse V2 cores

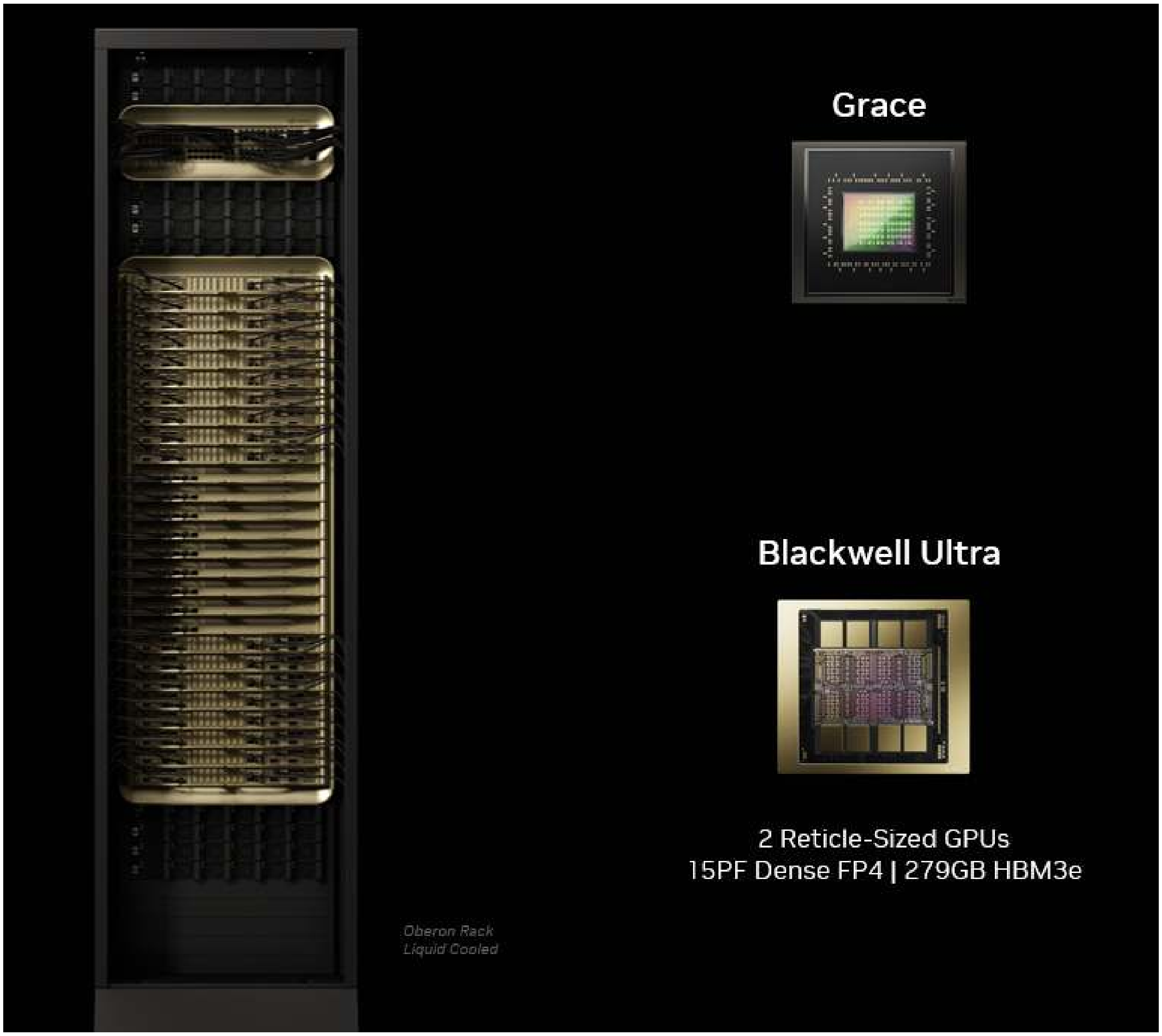


Figure 4. NVIDIA GB300 NVL72

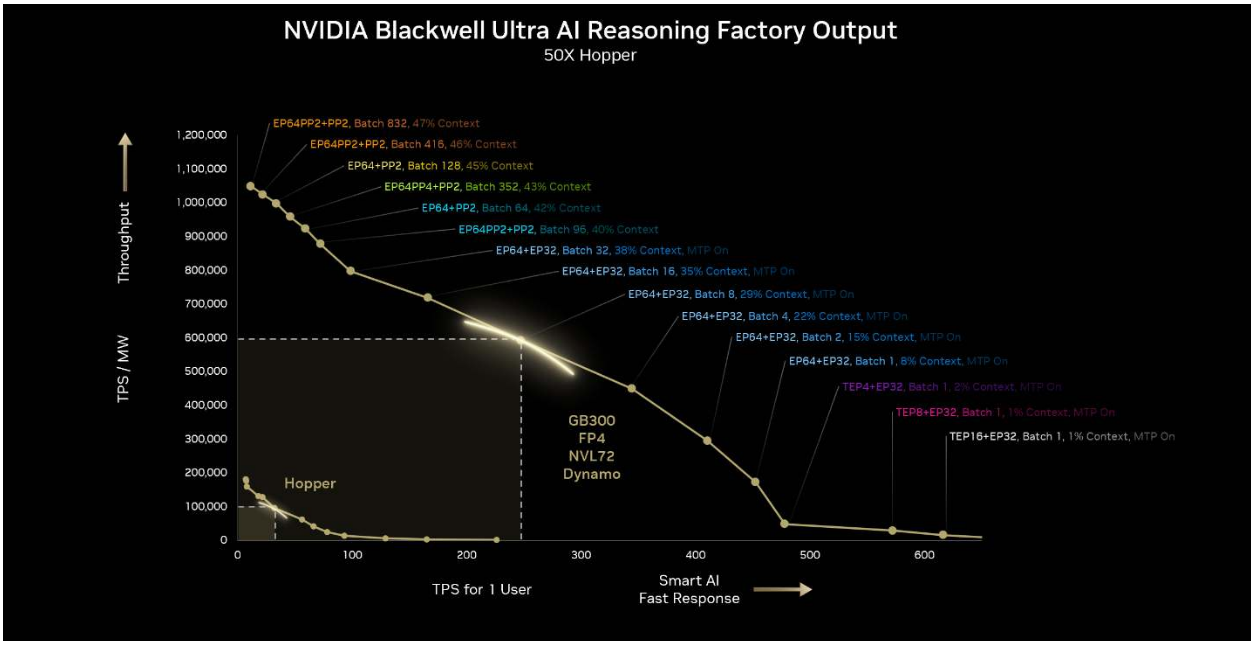
Blackwell Ultra GB300 NLV72

The NVIDIA GB300 NVL72 connects 36 Grace CPUs and 72 Blackwell Ultra GPUs in a rack-scale design, supercharging inference, training, and data processing. The GB300 NVL72 is a liquid-cooled, rack-scale solution that boasts a 72-GPU NVLink domain that acts as a single massive GPU and delivers optimized connectivity built for AI reasoning.

Maximizing AI Factory Performance and Revenue

Boosting AI Factory output by 50X

Jensen Huang framed the Pareto Frontier curves as an optimization framework for AI factory large language model inference, balancing throughput per megawatt (TPS / MW on the y-axis) against per-user latency experience (TPS for 1 User on the x-axis). AI factory efficiency is a balance of the raw throughput and the fast response with the optimal operating point being at the corner of the curves. The factory output can be viewed as the area under the curve or approximated by the box area created by the balance point. NVIDIA Dynamo acts as a real-time orchestrator that dynamically partitions GPU resources across GPUs, GPU memory, and NVLink, enabling systems to fluidly navigate the Pareto curve rather than being locked to fixed operational points. By optimizing parallelization strategies (expert/tensor/pipeline) and management across Blackwell Ultra GPUs, Dynamo achieves 50x more AI factory output or productivity compared to Hopper systems, while maintaining low latency, maximizing revenue per token through efficient token manufacturing.



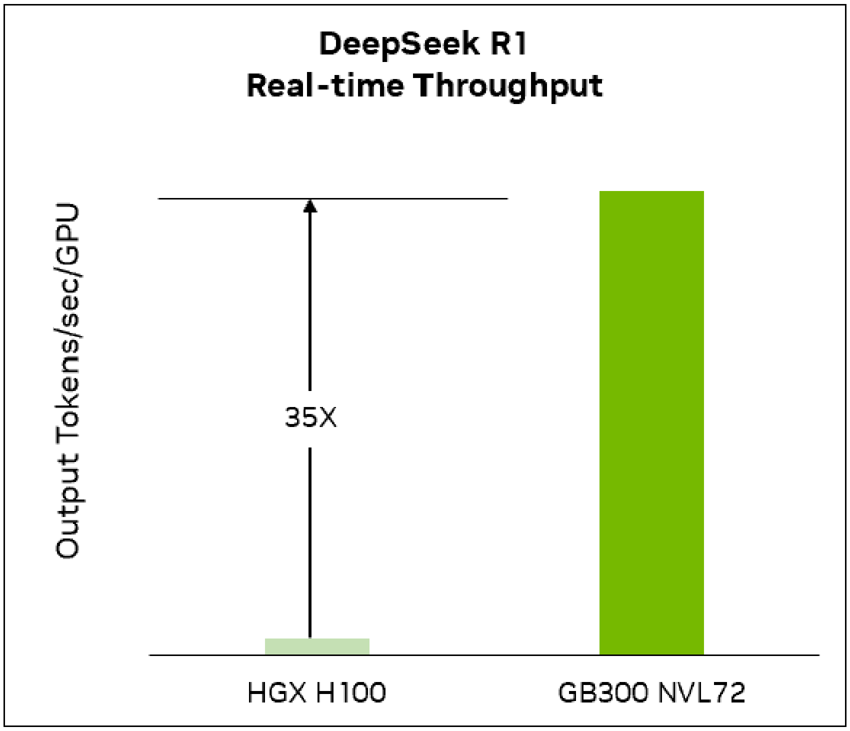
Projected DeepSeek R1 inference, 1 MW, FP4, NVL72, Dynamo, and TRT-LLM Continuous Optimization
32K ISL / 8K OSL

Figure 5. GB300 Delivers 50X AI Factory Output Increase for AI Reasoning

Lowering TCO through Performance and Energy Efficiency

Boosting Performance by 35X

The NVIDIA GB300 NVL72 has a 35X inference performance boost that transforms the speed and scalability of AI applications when compared with H100 for AI reasoning (DeepSeek-R1), enabling more complex strategic problem-solving by AI chatbots.



Projected DeepSeek R1 Inference, Disaggregated, ISL=32K, OSL=8K @ 200 TPS/user, FTL=2s

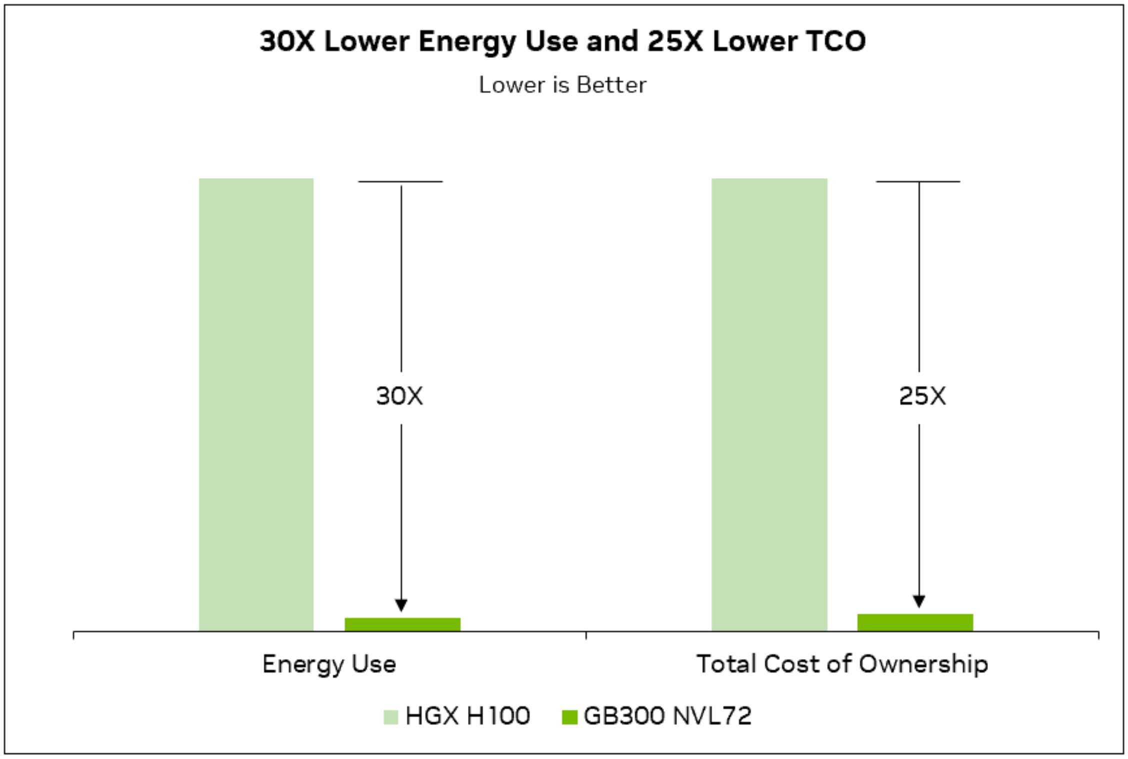
Figure 6. GB300 Delivers 35X Performance Improvement for AI Reasoning

Increasing Energy Efficiency by 30X

Data centers running AI workloads are often constrained by energy and cooling limitations, making efficiency a critical factor, with the goal of maximizing the performance for every unit of energy consumed. Blackwell Ultra delivers 30X higher energy efficiency than the Hopper generation.

Lowering TCO by 25X

Total cost of ownership (TCO) continues to be a focus of enterprises seeking to achieve significantly higher AI performance at a fraction of the cost. With GB300 NVL72, customers will benefit from reduced hardware expenses, in addition to minimized cooling and maintenance costs over the lifecycle of AI deployments. This enables organizations to allocate resources more effectively, accelerating AI-driven innovation without excessive capital expenditures.



Projected performance subject to change. Disaggregated inference, ISL=32K, OSL=8K @ 100 TPS/user, FTL=5s
For trillion parameter AI models, compared to H100 air-cooled infrastructure, GB200 delivers 25X lower TCO and energy at the same performance.

Figure 7. Reduction in Energy Use and Total Cost of Ownership with GB300

End-to-End AI Acceleration at Rack-Scale

With up to 279 GB of HBM3e memory per GPU and 37 TB of high-speed memory per rack, coupled with over an exaFLOP of FP4 compute and a 72-GPU unified NVLink domain, Blackwell Ultra supports much larger models and can scale up with fewer nodes, opening the door to breakthroughs in AI. Combined with CUDA-X libraries for accelerated computing, NVIDIA accelerates the entire hardware and software computing stack.

Optimized for Every Data Center

Investing in optimized data centers is not just a performance advantage—it's a strategic necessity for organizations looking to stay competitive in the AI-driven future. With 65X more AI FLOPS than HGX H100, GB300 NVL72 enables significantly more inference for AI models. By combining the CPU with the GPU and high-speed interconnects via NVLink, data transfers across multiple systems have never been more efficient.

Accelerated Networking Platforms for Scalable, Secure, and High-Throughput AI Performance

GB300 NVL72, acting as a single, extremely powerful unit of computing, requires robust networking to achieve optimal application performance. Paired with NVIDIA Quantum-X800 InfiniBand, Spectrum-X Ethernet, Connect-X SuperNICs, and BlueField-3 DPUs, GB300 NVL72 delivers unprecedented levels of scalable performance, efficiency, and security in massive-scale AI data centers.

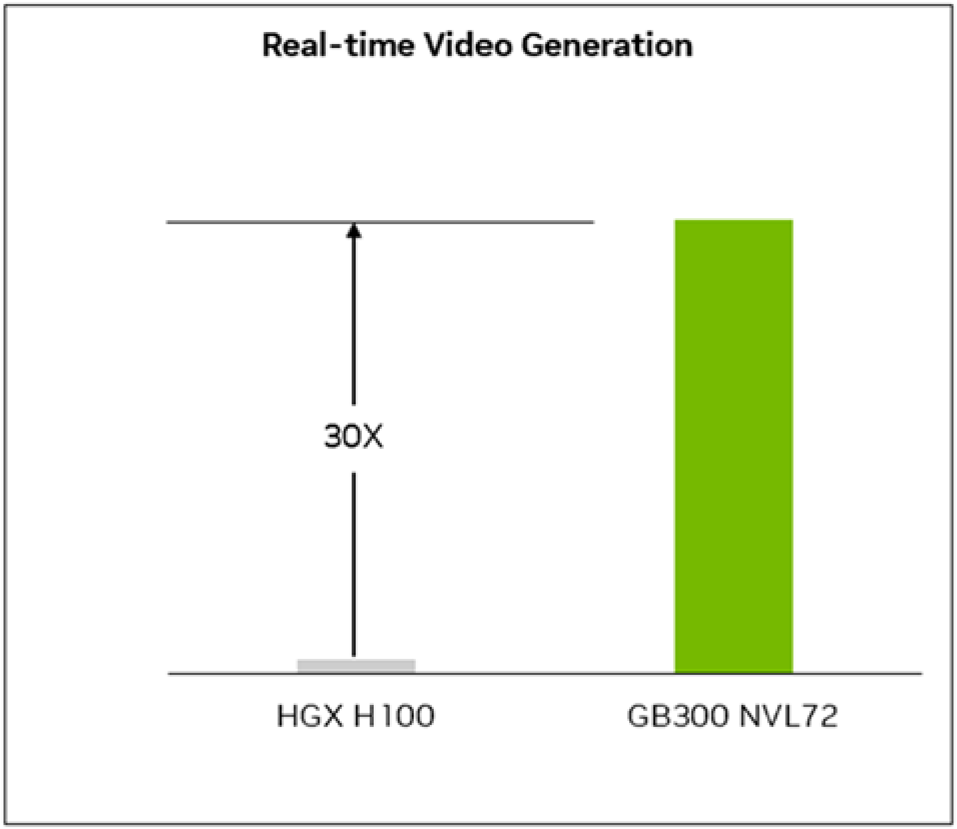
With 800 Gb/s of total data throughput available for each GPU in the system, GB300 NVL72 seamlessly integrates with NVIDIA networking platforms, enabling AI factories and cloud data centers to handle trillion-parameter models without bottlenecks. The GB300 NVL72 architecture is the first to introduce PCIe Gen6 connectivity between GPU and ConnectX-8 SuperNIC, thus eliminating the need for a standalone PCIe switch interface.

ConnectX-8 SuperNICs Enable Exceptional Inference Performance

The new NVIDIA ConnectX®-8 SuperNIC™ provides a full 800 gigabits per second (Gb/s) of network connectivity for each GPU in GB300 NVL72 systems. 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. Additionally, the ConnectX-8 SuperNIC supports line-speed network encryption for Internet Protocol Security (IPsec) and PSP Protocol Security (PSP), bolstering the platform’s security with GPU-to-GPU connection encryption.

Accelerating Real-Time Video Generation and Multi-modal AI Capabilities by 30X

Unlike state-of-the-art LLMs which operate within a context window of only 128,000 tokens, a single five second video generation processes four million tokens, requiring nearly 90 seconds to generate on today's state-of-the-art NVIDIA Hopper GPUs. The Blackwell Ultra platform enables real-time video generation from world foundation models like NVIDIA Cosmos, providing a 30X performance improvement vs Hopper Generation and allows customers to create customized, photo-realistic, and temporally and spatially stable video and 3D world simulators for their Physical AI applications.



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

Figure 8. 30X Video Generation Performance Improvement for Physical AI Applications using GB300 NVL72

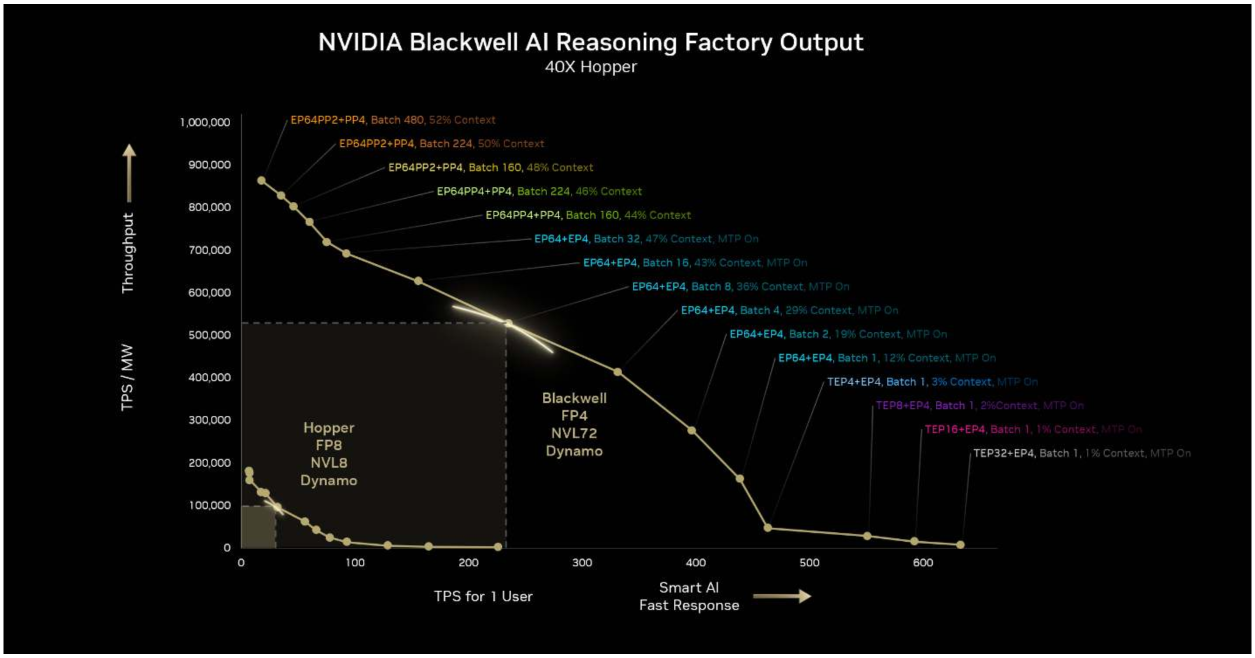
Blackwell GB200 NVL72

Maximizing AI Factory Performance and Revenue

Boosting AI Factory output by 40X

The Frontier Pareto curves illustrate the AI factory inference productivity balancing throughput and single user response time. NVIDIA Dynamo enhances productivity by up to 40x on GB200 NVL72 compared to Hopper systems with real-time GPU resource orchestration, parallelization, and management of Blackwell GPUs, enabling maximum efficiency in token manufacturing in a fixed power condition of 1 megawatt.

https://docs.nvidia.com/multi-node-nvlink-systems/partition-guide.pdf
https://www.theregister.com/2025/03/23/nvidia_dynamo/



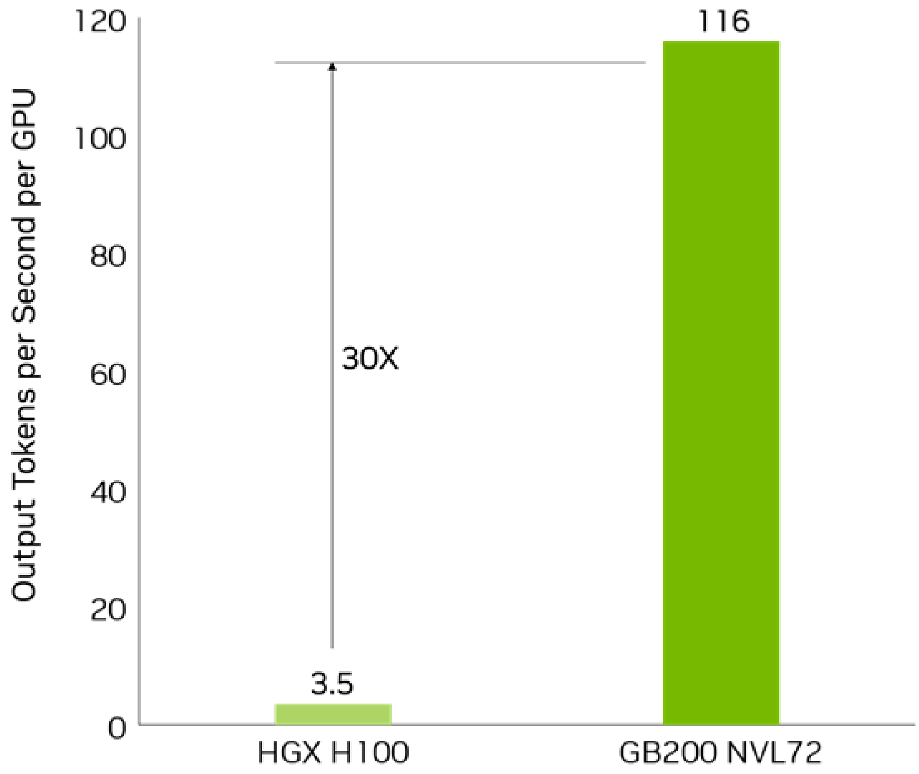
DeepSeek R1, 1 MW, FP4, NVL72, Dynamo, and TRT-LLM Continuous Optimization
32K ISL / 8K OSL

Figure 9. GB200 Delivers 40X AI Factory Output Increase for AI Reasoning

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

The GB200 NVL72 introduces cutting-edge capabilities and a second-generation Transformer Engine that significantly accelerates LLM inference workloads, enabling real-time performance for resource-intensive applications like multi-trillion parameter language models. GB200 NVL72 delivers a 30X speedup compared to H100 with 25X lower TCO and 25X less energy with the same number of GPUs for massive models such as a GPT-MoE-1.8. This advancement is made possible with a new generation of Tensor Cores, which introduce new precisions including FP4. Additionally, the GB200 utilizes NVLink and liquid cooling to create a single massive 72-GPU rack that can overcome communication bottlenecks.

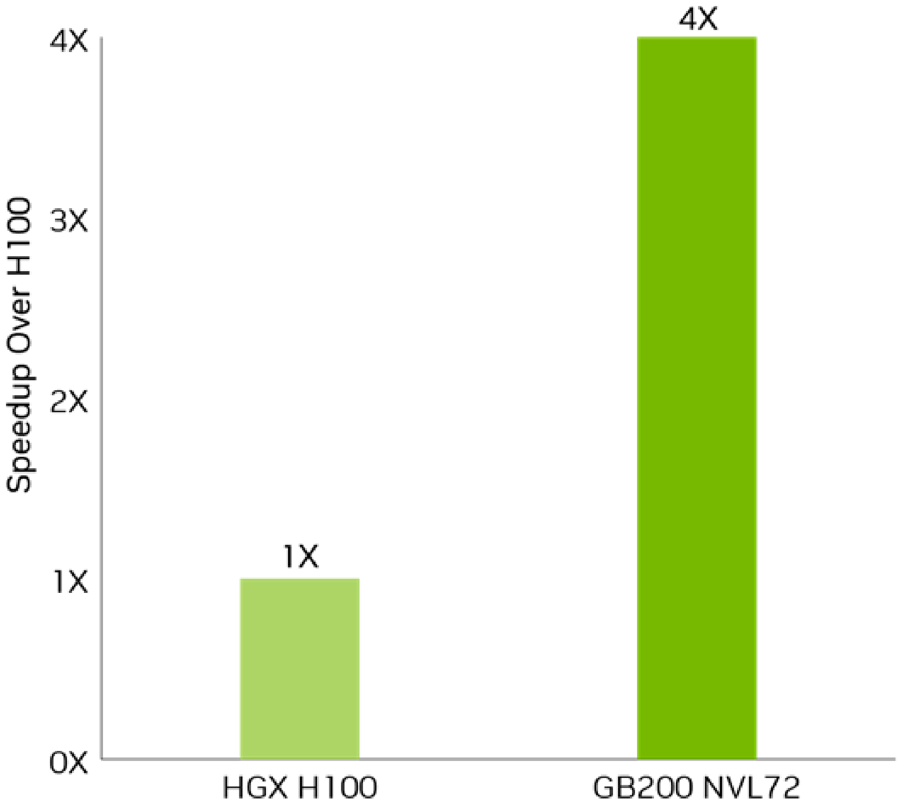
The GB200 is a revolutionary solution for high-performance inference tasks, showcasing NVIDIA's commitment to pushing the boundaries of AI.



Projected performance subject to change. Token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,024 output, 8x eight-way HGX H100 air-cooled: 400GB IB Network vs 18 GB200 Superchip liquid-cooled: NVL72, per GPU performance comparison

Figure 10. GB200 1.8T GPT-MoE Real-Time Inference Performance Using Second-Generation Transformer Engine Next-Level AI Training Performance

GB200 includes a faster Transformer Engine featuring FP8 precision and delivers 4X faster training performance for large language models like GPT-MoE-1.8T compared to the NVIDIA Hopper GPU generation. The performance boost provides a 9X reduction in rack space and a 3.5X reduction in TCO and energy usage. This breakthrough is complemented by the fifth-generation NVLink (which enables 1.8 TB/s of GPU-to-GPU interconnect and a larger 72-GPU NVLink domain), InfiniBand networking, and NVIDIA Magnum IO™ software. Together, these ensure efficient scalability for enterprises and facilitate the implementation of extensive GPU computing clusters.



Projected performance subject to change. 32,768 GPU scale, 4,096x HGX H100 air-cooled cluster: 400G IB network, 456x GB200 NVL72 liquid-cooled cluster: 800G IB network

Figure 11. GB200 1.8T GPT-MoE Model Training Speed-Up Using Transformer Engine

Accelerating Data Processing and Physics-Based Simulation

GB200, with its tightly-coupled CPU and GPU, brings new opportunities in accelerated computing for data processing, and engineering design and simulation.

Databases play critical roles in handling, processing, and analyzing large volumes of data for enterprises. GB200 takes advantage of the high-bandwidth NVLink-C2C and dedicated Decompression Engine in Blackwell to speed up key database queries by 18X compared to CPU, resulting in 7X less energy, and 5X lower TCO.

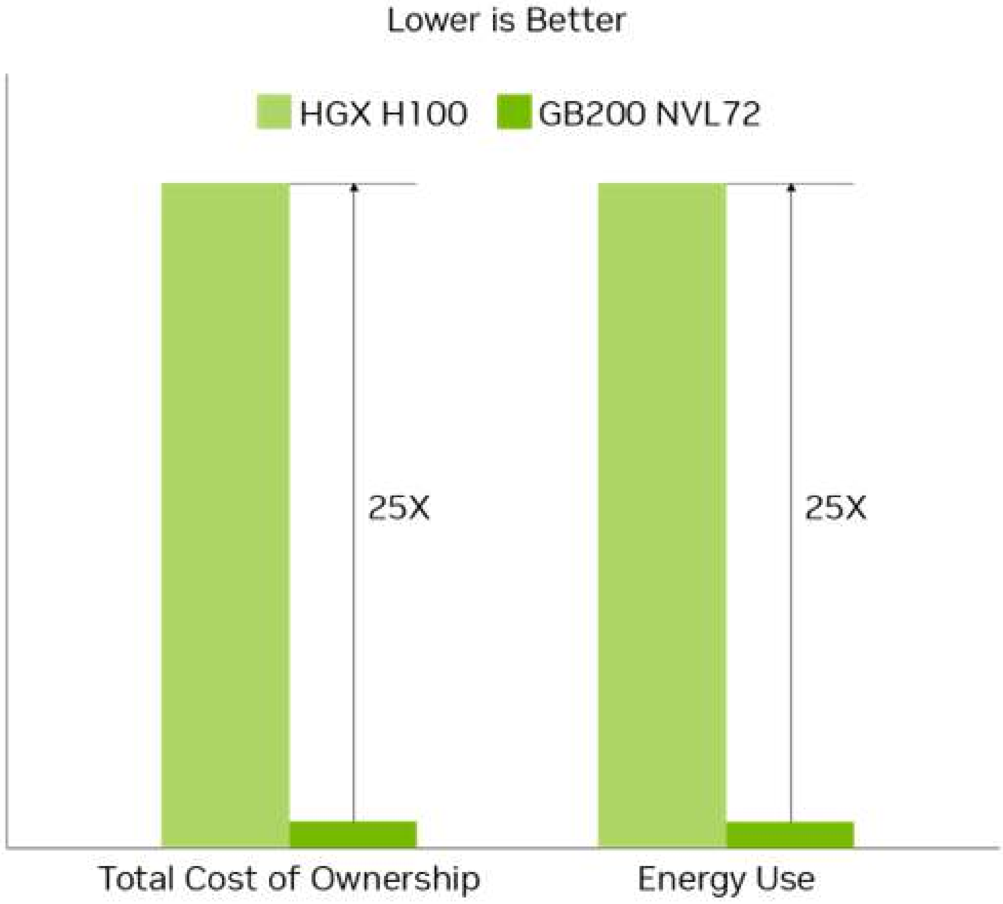
Physics-based simulations are still the mainstay of product design and development. From silicon chips to pharmaceuticals, testing and improving products by simulating them rather than performing physical testing saves billions of dollars every year.

Application-specific integrated circuits are designed almost exclusively on CPUs in a long and complex workflow, which often includes an analog analysis to identify voltages and currents throughout. The Cadence SpectreX simulator is one example of a solver and runs 13x faster on a GB200 than on an x86 CPU.

GPU-accelerated computational fluid dynamics (CFD) is a critical tool for engineers and equipment designers to study or predict the behavior of their designs. Cadence Fidelity, a large eddy simulator (LES), runs simulations up to 22x faster on GB200 than x86 CPU.

Sustainable Computing

Compute density and compute power are driving a transition from air cooling to liquid cooling. Using liquid instead of air has many positive impacts inside and outside the data center including higher performance per rack, reduced water consumption for cooling, and allowing data centers to run at higher ambient air temperatures, which further reduces energy consumption.



TCO and energy savings for 65 racks eight-way HGX H100 air-cooled versus 1 rack GB200 NLV72 liquid-cooled with equivalent performance on GPT-MoE-1.8T real-time inference throughput.

Figure 12. 25X Lower Energy Use and TCO

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™ with NVIDIA GPUs. 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 NVIDIA Blackwell Ultra accelerated computing, 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 build.nvidia.com.

NVIDIA Blackwell HGX

The NVIDIA Blackwell HGX B300 and HGX B200 systems include groundbreaking advancements for generative AI, data analytics, and high-performance computing,

NVIDIA HGX™ B300:
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.

HGX B200: A Blackwell x86 platform based on an eight-Blackwell GPU baseboard, delivering 144 petaFLOPS of AI performance. HGX B200 delivers the best performance (15X more than HGX H100) and TCO (12X more than HGX H100) for x86 scale-up platforms and infrastructure. Each GPU is configurable up to 1000 Watts per GPU.

Table 3. System Specifications for HGX B300 and HGX B200

	HGX B300	**HGX B200**
	Per Server Specs Below	
Blackwell GPUs	16 Blackwell Ultra Die for 8 GPUs	16 Blackwell Die for 8 GPUs
FP4 Tensor Core Dense/Sparse	108/144 petaFLOPS	72/144 petaFLOPS
FP8/FP6 Tensor Core Dense/Sparse	36/72 petaFLOPS	36/72 petaFLOPS
Fast Memory	2.1 TB	Up to 1.4 TB
Aggregate Memory Bandwidth	62 TB/s	Up to 62 TB/s
Aggregate NVLink Switch Bandwidth	14.4 TB/s	14.4 TB/s
	Per GPU Specs Below	
FP4 Tensor Core Dense/Sparse	14/18 petaFLOPS	9/18 petaFLOPS
FP8/FP6 Tensor Core Dense/Sparse	4.5/9 petaFLOPS	4.5/9 petaFLOPS
INT8 Tensor Core Dense/Sparse	0.15/ 0.30 petaOPS	4.5/9 petaOPS

FP16/BF16 Tensor Core Dense/Sparse	2.2/4.5 petaFLOPS	2.2/4.5 petaFLOPS
TF32 Tensor Core Dense/Sparse	1.1/2.2 petaFLOPS	1.1/2.2 petaFLOPS
FP32	75 teraFLOPS	75 teraFLOPS
FP64 Tensor Core | FP64	1.2 teraFLOPS	37 teraFLOPS
GPU memory | Bandwidth	270 GB HBM3e | 7.7 TB/s	Up to 192 GB HBM3e | 7.7 TB/s
Max thermal design power (TDP)	1100 W	1000W
Interconnect	NVLink 5 PCIe Gen6	NVLink 5 PCIe Gen5
Server options	NVIDIA HGX B300 partner and NVIDIA-Certified Systems	NVIDIA HGX B200 partner and NVIDIA-Certified Systems

NVIDIA Blackwell Architecture’s Role in the Age of AI Reasoning

AI has evolved to need three distinct scaling laws that describe how applying compute resources in different ways impacts model performance.

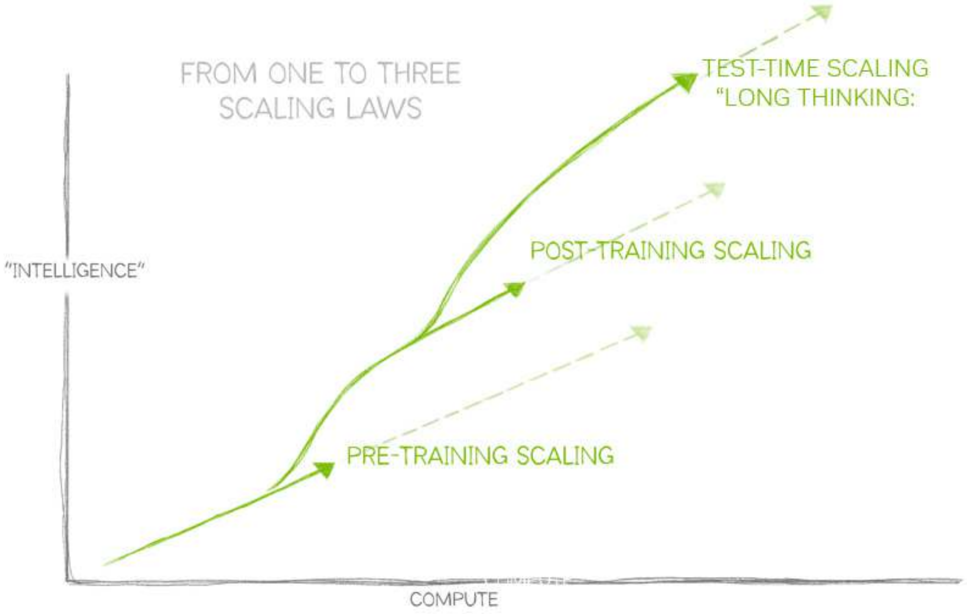


Figure 13. Three AI Scaling Laws

Pre-training Scaling: the original law of AI development. It demonstrated that by increasing training dataset size, model parameter count and computational resources, developers could expect predictable improvements in model intelligence and accuracy. Pre-training scaling led to models that achieved groundbreaking capabilities. It spurred major innovations in model architecture, including the rise of billion- and trillion-parameter transformer models, growing 50 million-X in five years in compute requirements.

Post-training scaling: While pre-training scaling teaches a model the knowledge of the Internet, post-training teaches a model how to think and further improves a model’s specificity and relevance for an organization’s desired use case. While pretraining is like sending an AI model to school to learn foundational skills, post-training enhances the model with skills applicable to its intended job. To support post-training, developers can use synthetic data to augment or complement their fine-tuning dataset. The total compute required for post-training is 30X more than the compute required for pre-training.

Test time scaling (also known as long thinking or reasoning): LLMs generate quick responses to input prompts and can provide the right answers to simple questions, but it may not be as effective for complex queries. An essential capability for agentic AI workloads requires the LLM to reason through the question before coming up with an answer, which takes place during inference. Models using test time scaling are estimated to require 100X more compute when compared to traditional inference, allowing them to reason through multiple potential responses before arriving at the best answer.

NVIDIA Blackwell is the once-in-a-generation platform with the compute capacity and energy efficiency needed to effectively serve the three scaling laws as the foundation for the age of AI reasoning.

Appendix A

Advanced Parallelism Techniques in AI Inference for Trillion-Parameter Models

The deployment of trillion-parameter models like the GPT 1.8T MoE (Mixture of Experts) presents unique challenges in AI inference, particularly in managing computational resources effectively while ensuring optimal user experience. This appendix explores the various parallelism techniques that can be employed to address these challenges, focusing on data, tensor, pipeline, and expert parallelism.

Parallelism Techniques in AI Inference

1. **Data Parallelism (DP) Data parallelism** involves hosting multiple copies of the entire model across different GPUs or clusters, processing independent user requests simultaneously. This approach scales linearly with the number of GPUs, enhancing throughput without impacting user interactivity. However, it requires significant memory as each GPU holds a complete model copy.
2. **Tensor Parallelism (TP)** Tensor parallelism splits each layer of the model across multiple GPUs, allowing different parts of a user request to be processed in parallel. This method can improve user interactivity by allocating more resources per request, thus reducing processing time. However, it relies heavily on high-bandwidth inter-GPU communication, which can become a bottleneck at large scales.
3. **Pipeline Parallelism (PP)** In pipeline parallelism, different groups of model layers are distributed across GPUs, with each part of a user request processed sequentially across the pipeline. This technique helps manage large models by distributing weights, but may lead to inefficiencies in processing and does not significantly enhance user interactivity.
4. **Expert Parallelism (EP)** Expert parallelism routes requests to specific experts within the model to different GPUs, reducing the interaction with unnecessary parameters. The results, after expert processing, require all-to-all communication over high bandwidth GPU interconnect. It requires complex management of data routing and reassembly, and its effectiveness is limited by the number of available experts.

Combining Parallelism Techniques

Combining different parallelism methods can mitigate the limitations of individual techniques. Using both expert and pipeline parallelism can double user interactivity with minimal loss in throughput. Similarly, integrating tensor, expert, and pipeline parallelism can triple GPU throughput without sacrificing user interactivity. Combining different parallelisms for the right deployment scenario is an exhaustive solution space exploration and requires a large set of compute resources.

Maximizing Throughput and Managing Operational Phases

Efficient management of the prefill and decode phases i.e., context processing and generation phase is crucial for maximizing throughput. Techniques like inflight batching and chunking can optimize GPU utilization by allowing dynamic management of request processing, preventing bottlenecks during these phases.

Inflight Batching and Chunking

Inflight batching and chunking are critical techniques for optimizing GPU utilization and enhancing user experience in the deployment of LLMs. These methods address the operational phases of AI inference—prefill and decode—by managing how data is processed across GPU resources.

- **Chunk Size Considerations:** The size of chunks plays a pivotal role in balancing GPU throughput and user interactivity. Larger chunk sizes reduce the number of iterations needed during the prefill phase, leading to a quicker time to first token (TTFT). However, this also extends the duration of the decode phase, lowering the tokens per second (TPS) rate. Conversely, smaller chunk sizes facilitate faster token output, enhancing TPS but increasing TTFT. This trade-off is crucial in determining the optimal chunk size for specific deployment scenarios.

Impact of Chunk Size on GPT 1.8T MoE Model

Using the GPT 1.8T MoE model as an example, the effect of varying chunk sizes from 128 to 8,192 tokens was analyzed across more than 2,700 combinations of parallelism and chunk-length configurations. This extensive analysis helps in understanding how different settings impact the balance between throughput and interactivity.

Conclusion

The deployment of trillion-parameter models requires sophisticated parallelism strategies to balance throughput and user interactivity effectively. By understanding and implementing a combination of data, tensor, pipeline, and expert parallelism, enterprises can optimize their AI inference deployments to meet both computational demands and user expectations.

For further insights into optimizing AI inference for large-scale models and a deeper dive into the different types of parallelisms, read the technical walkthrough, Demystifying AI Inference Deployments for Trillion Parameter Large Language Models.

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA CUDA, NVIDIA Omniverse, NVIDIA RTX, NVIDIA Tesla, NVIDIA Turing, NVIDIA Volta, NVIDIA Jetson AGX Xavier, NVIDIA DGX, NVIDIA HGX, NVIDIA EGX, NVIDIA CUDA-X, NVIDIA GPU Cloud, GeForce, Quadro, CUDA, GeForce RTX, NVIDIA NVLink, NVIDIA NVSwitch, NVIDIA DGX POD, NVIDIA DGX SuperPOD, and NVIDIA TensorRT, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.