NVIDIA Blackwell Architecture Technical Brief

Powering the New Era of Generative AI and Accelerated Computing

V1.1

Table of Contents

List of Figures

List of Tables

Pioneering AI Innovation

In the rapidly evolving landscape of AI and large language models (LLMs), the pursuit of real-time performance and scalability is paramount. From healthcare to automotive industries, organizations are diving deeper into the realms of generative AI and accelerated computing solutions. This surge in demand for generative AI solutions is catalyzing a growing need to accommodate ever-expanding model sizes and complexities across enterprises.

Enter NVIDIA Blackwell GPU architecture, the world's largest GPU, built with the specific purpose of handling data center-scale generative AI workflows with up to 25X the energy efficiency of the prior NVIDIA Hopper GPU generation.

This technical brief introduces the benefits of NVIDIA Blackwell in detail, including the next-generation superchip, Grace Blackwell GB200, and the next-generation, high-performance HGX systems, NVIDIA HGX B200 and HGX B100.

NVIDIA Blackwell GPU and Superchip Overview

Large language models (LLMs) require immense computational power for real-time performance. The computational demands of LLMs also translate into higher energy consumption as more and more memory, accelerators, and servers are required to fit, train, and infer from these models. Organizations aiming to deploy LLMs for real-time inference must grapple with these challenges.

The NVIDIA Blackwell architecture and portfolio of products are designed to address the needs of ever-increasing AI model sizes and parameters with a long list of new innovations, including a new second-generation Transformer Engine.

The NVIDIA Blackwell architecture was named to honor David H. Blackwell, an amazing and inspiring American mathematician and statistician known for the Rao-Blackwell Theorem, and many contributions and advancements in probability theory, game theory, statistics, and dynamic programming.

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

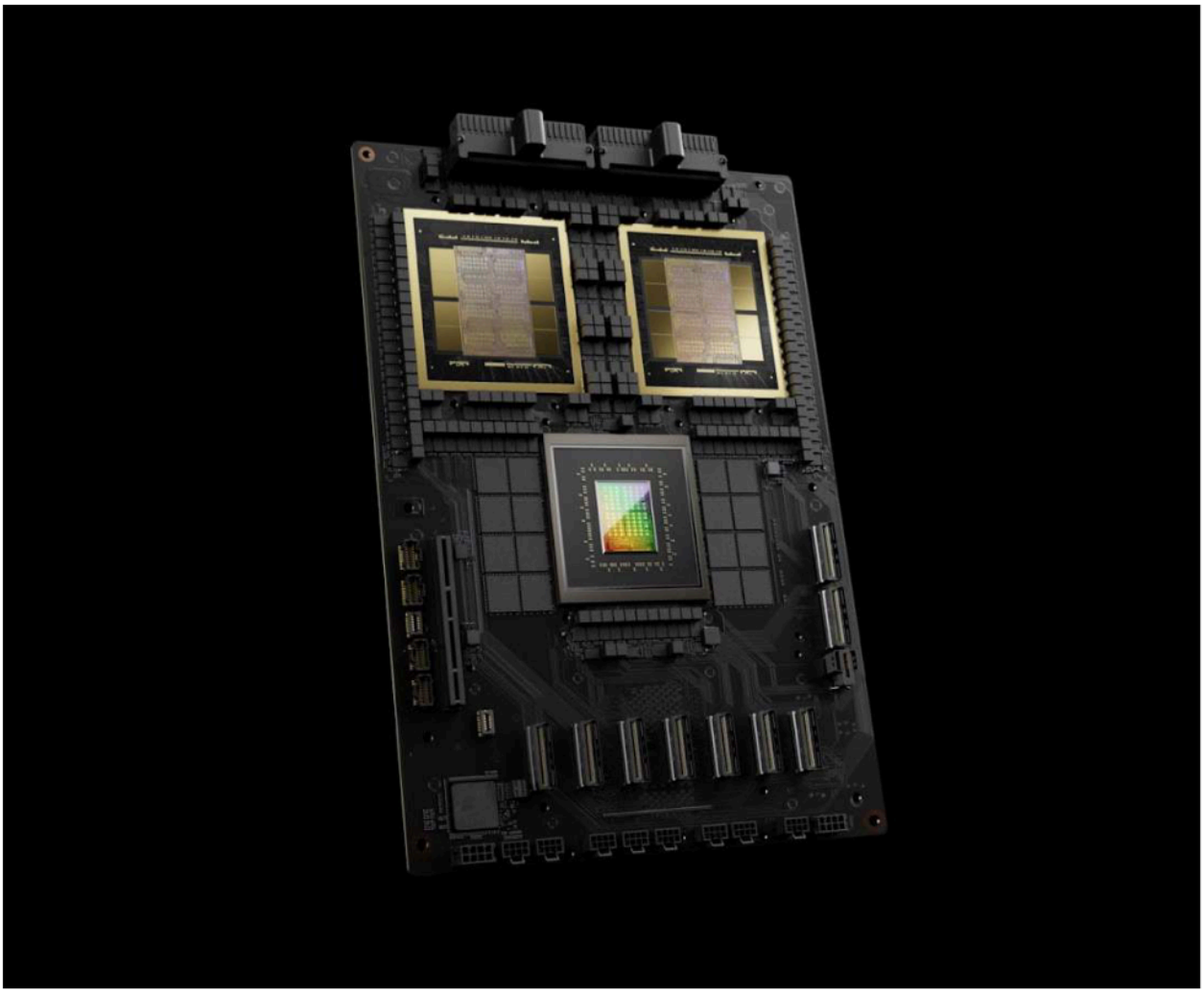


Figure 1. NVIDIA GB200 Superchip Incl. Two Blackwell GPUs and One Grace CPU

NVIDIA Blackwell Architectural Innovations

The Blackwell architecture introduces groundbreaking advancements for generative AI and accelerated computing. 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 real-time generative 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 generative AI with unparalleled performance, efficiency, and scale.

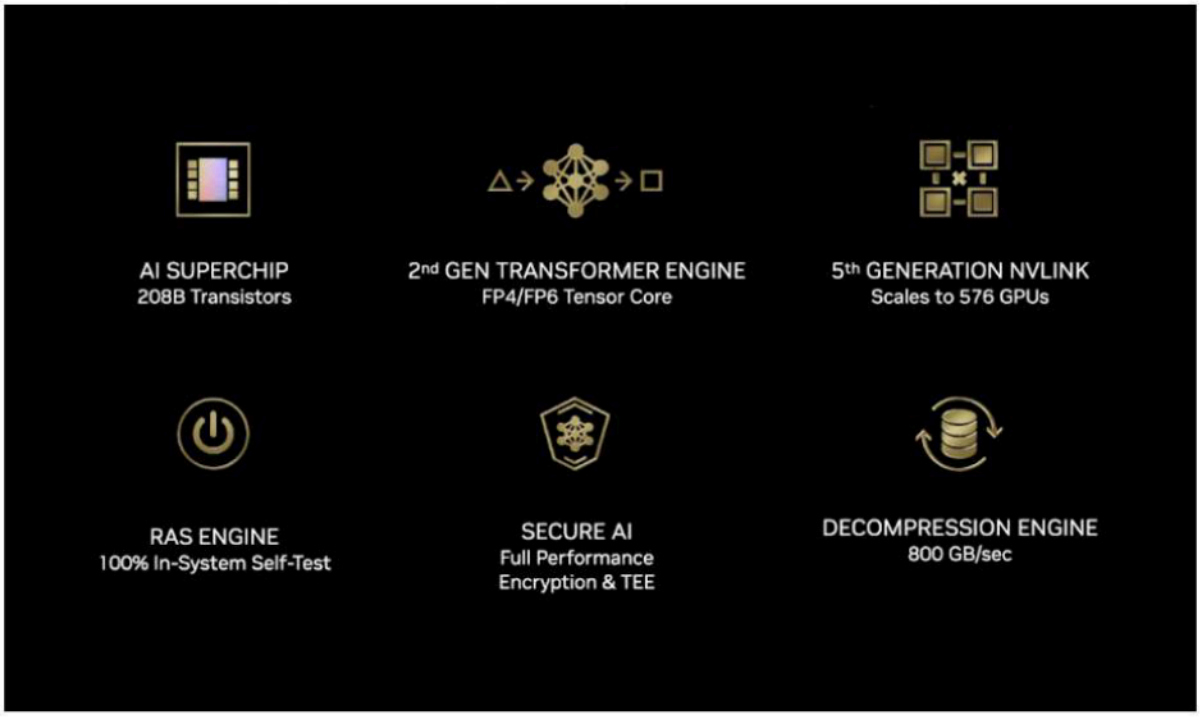


Figure 2. NVIDIA Blackwell Architecture's Technological Breakthroughs

A New Class of AI Superchip

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 is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two 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, unified GPU.

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.

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 TensorRT-LLM and Nemo Framework innovations to accelerate inference and training for LLMs and Mixture-of-Experts (MoE) models.

To supercharge inference of large MoE models, Blackwell Tensor Cores add new precisions, including new community-defined microscaling formats, giving high accuracy and greater throughput. The Blackwell Transformer Engine utilizes advanced dynamic range management algorithms and fine-grain scaling techniques, called micro-tensor scaling, to optimize performance and accuracy and enable FP4 AI. This doubles the performance with Blackwell's FP4 Tensor Core, doubles the parameter bandwidth to the HBM memory, and doubles the size of next-generation models per GPU.

Innovations in TensorRT-LLM, including quantization to 4-bit precision, and custom kernels with expert parallelism mapping, 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 MoE models with affordable economics, optimizing their business with the benefits of generative AI. NVIDIA Blackwell makes the next era of MoE models possible—supporting training and real-time inference on models over 10-trillion-parameters in size.

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 trillion-parameter AI models hinges on the need for swift, seamless communication among every GPU within a server cluster. The fifth generation of NVLink can scale up to 576 GPUs to accelerate performance for trillion- and multi-trillion-parameter AI models thanks to the NVLink Switch ASIC and switches built with it. Fifth-generation NVLink doubles the performance of fourth-generation NVLink in NVIDIA Hopper. While the new NVLink in Blackwell GPUs also uses two high-speed differential pairs in each direction to form a single link as in the Hopper GPU, NVIDIA Blackwell doubles the effective bandwidth per link to 50 GB/sec in each direction.

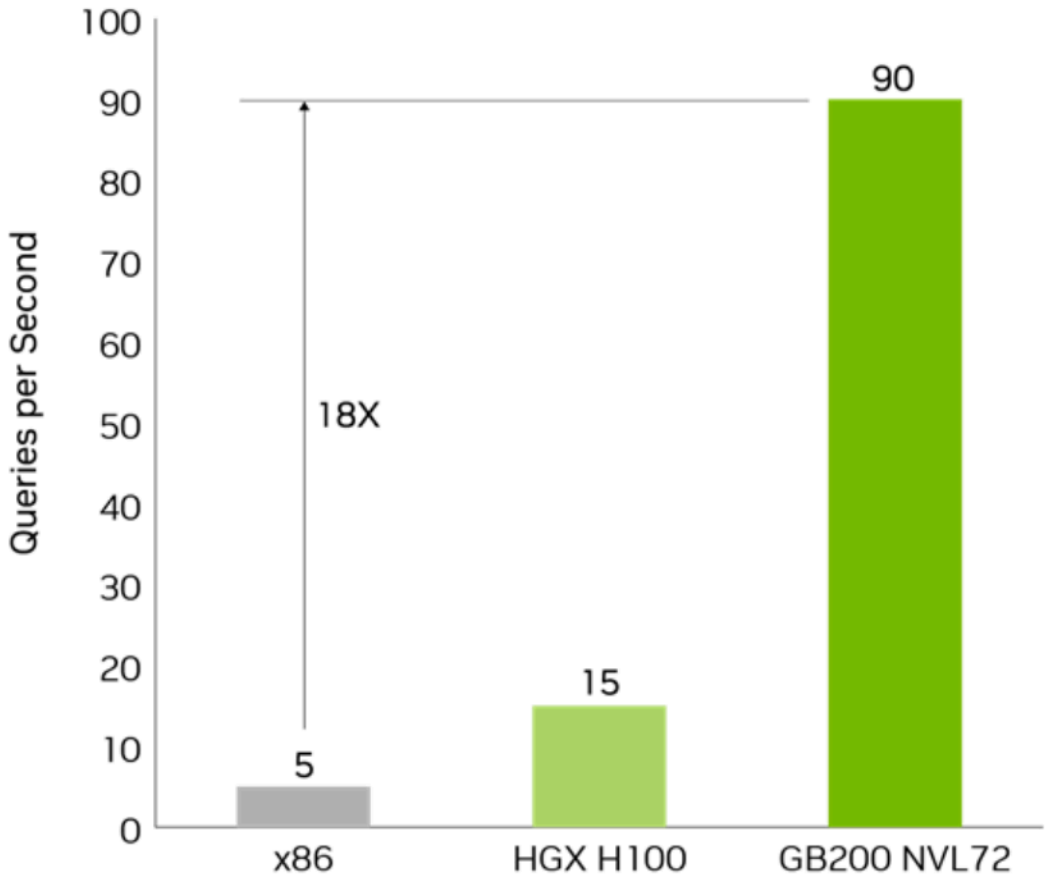
Blackwell GPUs include 18 fifth-generation NVLink links to provide 1.8 TB/sec total bandwidth, 900 GB/sec in each direction. 1.8TB/s of bidirectional throughput per GPU is over 14X the bandwidth of PCIe Gen5, ensuring high-speed communication for today's most complex large models. That's nearly seven petabytes of data transfer in an hour from one GPU, or more data than 18 years of streaming 4K movies, or the entire Internet bandwidth processed by just 11 Blackwell GPUs.

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 GB200 NVL72 to support 9X the GPU throughput as compared to a single eight-GPU system.

The NVLink Switch works with the NVIDIA Unified Fabric Manager (UFM®) to offer production-proven management for the NVLink compute fabric.

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, including Apache Spark, play critical roles in handling, processing, and analyzing large volumes of data for data analytics. Blackwell's new dedicated Decompression Engine can decompress data at a rate of up to 800GB/s, and 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, 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 performs 18X faster than CPUs and 6X faster than NVIDIA H100 Tensor Core 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 Superchip.

Figure 3. GB200 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 GB200 Superchip and GB200 NVL72

The NVIDIA GB200 Grace Blackwell Superchip connects two high-performance NVIDIA Blackwell Tensor Core GPUs and an NVIDIA Grace CPU using the NVIDIA® NVLink®-C2C interconnect that delivers 900 gigabytes per second (GB/s) of bidirectional bandwidth to the two GPUs.

Table 1. NVIDIA Blackwell GB200 Specifications

GPU Spec	GB200 Superchip
Configuration	1 Grace CPU : 2 Blackwell GPUs
FP4 Tensor Core Dense/Sparse	20 / 40 petaFLOPS
FP8/FP6 Tensor Core Dense/Sparse	10 / 20 petaFLOPS
INT8 Tensor Core Dense/Sparse	10 / 20 petaOPS
FP16/BF16 Tensor Core Dense/Sparse	5 / 10 petaFLOPS
TF32 Tensor Core Dense/Sparse	2.5 / 5 petaFLOPS
FP32	180 teraFLOPS
FP64 Tensor Core Dense	90 teraFLOPS
FP64	90 teraFLOPS
HBM Memory Architecture	HBM3e 8x2-sites
HBM Memory Size	Up to 384 GB
HBM Memory Bandwidth	Up to 16 TB/s
Decompression Engine	Yes
Decoders	2x 7 NVDEC 2x 7 NVJPEG
CPU core count	72 Arm® Neoverse V2 cores

CPU L1 cache	64 KB i-cache + 64 KB d-cache
CPU L2 cache	1 MB per core
CPU L3 cache	114 MB
LPDDR5X | Bandwidth	Up to 480 GB | Up to 512 GB/s
Multi-Instance GPU (MIG) instances	2x 7
Form factor	Superchip module
NVLink Support	NVLink v5
NVLink Bandwidth Bidirectional	2x 1.8 TB/s
PCIe Gen 6 Bandwidth Bidirectional	2x 256 GB/s Gen6
TDP	Configurable up to 2700 W
Server options	NVIDIA GB200 NVL72 Scales to 576 GPUs

Preliminary specifications subject to change.

Table 2. System Specifications for GB200 NVL72

System Spec	GB200 NVL72
Configuration	36 GB200 Superchips
Compute	18 GB200 Superchip Nodes
FP4 Tensor Core Dense/Sparse	720 / 1,440 petaFLOPS
FP8/FP6 Tensor Core Dense/Sparse	360 / 720 petaFLOPS
INT8 Tensor Core Dense/Sparse	360 / 720 petaOPS
FP16/BF16 Tensor Core Dense/Sparse	180 / 360 petaFLOPS

TF32 Tensor Core Dense/Sparse	90 / 180 petaFLOPS
FP32	6,480 teraFLOPS
FP64 Tensor Core Dense	3,240 teraFLOPs
FP64	3,240 teraFLOPs
HBM Memory Architecture	HBM3e
HBM Memory Size	Up to 13.5 TB
HBM Memory Bandwidth	Up to 576 TB/s
Fast Memory	Up to 30 TB
NVLink Switch	9 NVLink Switches
NVLink Bandwidth Bidirectional	130 TB/s
CPU Cores	2592 Arm Neoverse V2 cores

Preliminary specifications subject to change.

Figure 4. NVIDIA GB200 NVL72

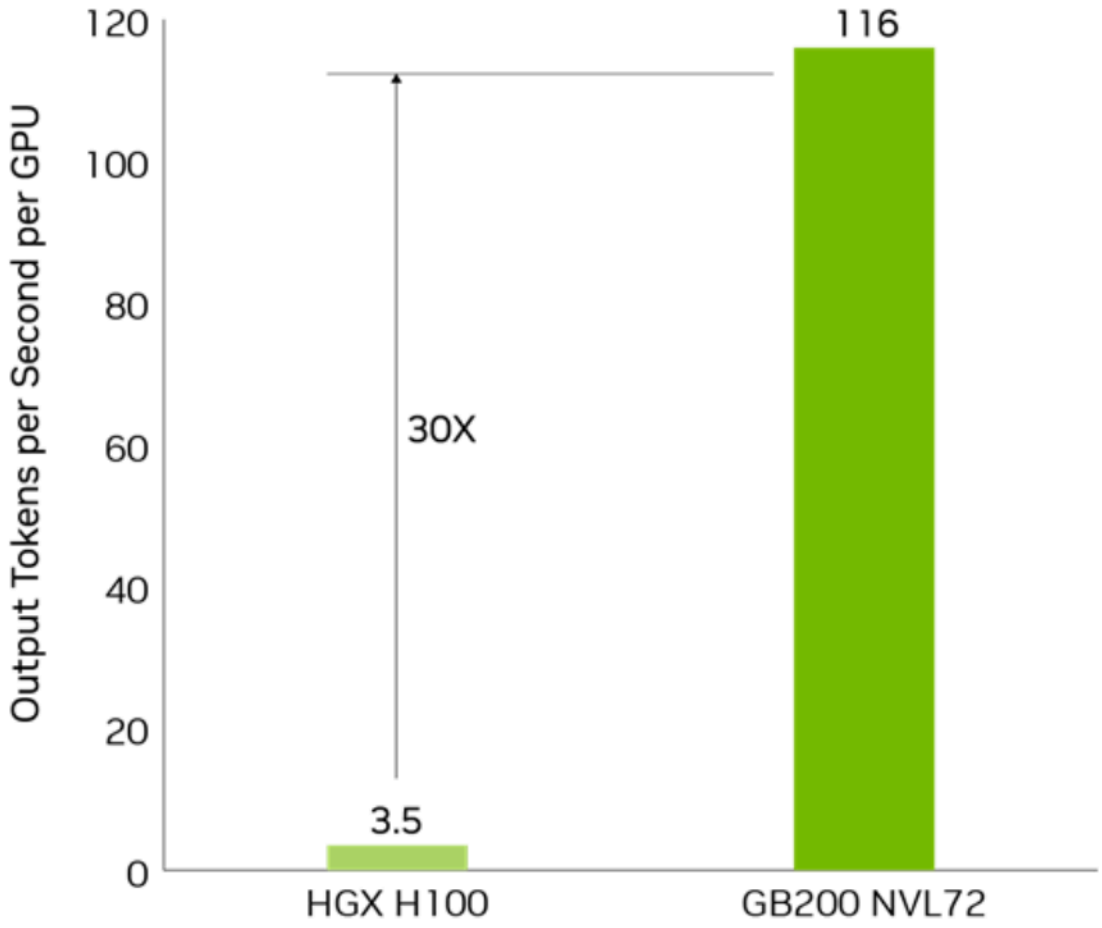
The NVIDIA GB200 NVL72 cluster connects 36 GB200 Superchips (36 Grace CPUs and 72 Blackwell GPUs) in a rack-scale design. The GB200 NVL72 is a liquid cooled, rack-scale 72-GPU NVLink domain, that can act as a single massive GPU to deliver 30X faster real-time trillion parameter LLM inference than the prior generation (see Figure 5).

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-

1.8T[1] (see Figure 5). 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 as a revolutionary solution for high-performance inference tasks, showcasing NVIDIA's commitment to pushing the boundaries of AI.



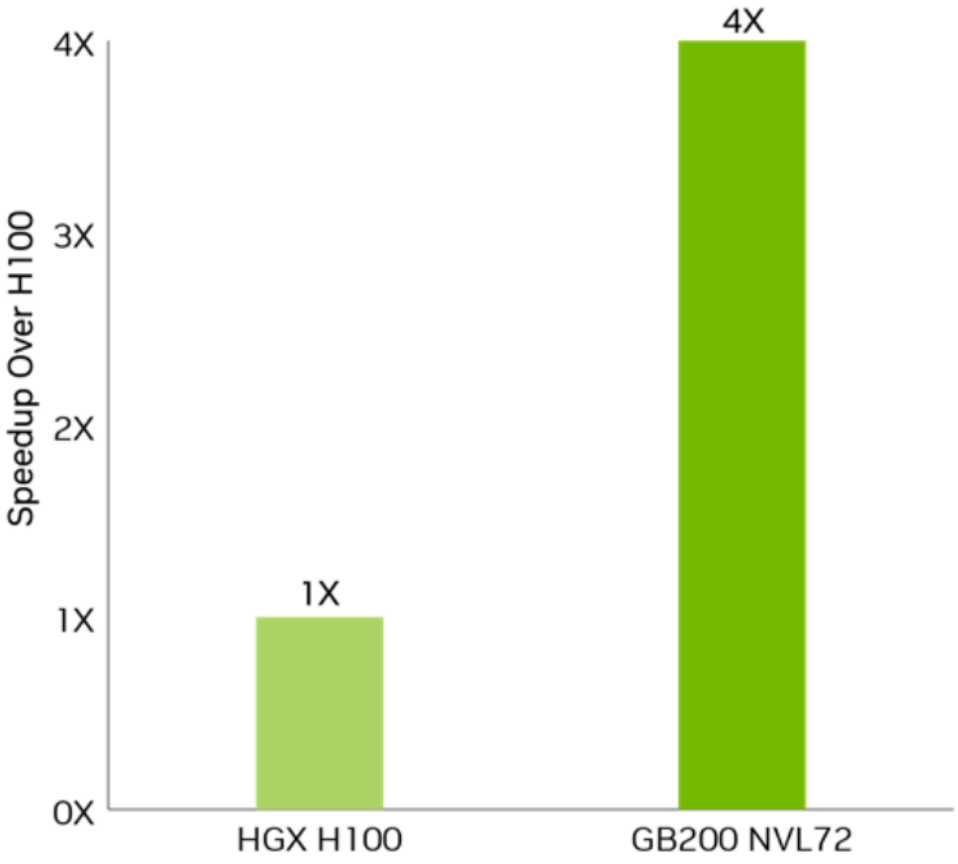
Projected performance subject to change. Token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,024 output, 8x eight-way HGX H100 air-cooled: 400GB IB Network vs 18 GB200 Superchip liquid-cooled: NVL36, per GPU performance comparison

Figure 5. GB200 1.8T GPT-MoE Real-Time Inference Performance Using Second-Generation Transformer Engine

[1] 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, 64 H100 GPUs air-cooled vs. 18 GB200 Superchips with NVL36 liquid-cooled, per GPU performance comparison. TCO, energy savings for 100 racks eight-way HGX H100 air-cooled vs. 1 rack GB200 NVL72 liquid-cooled with equivalent performance.

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 6. 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.

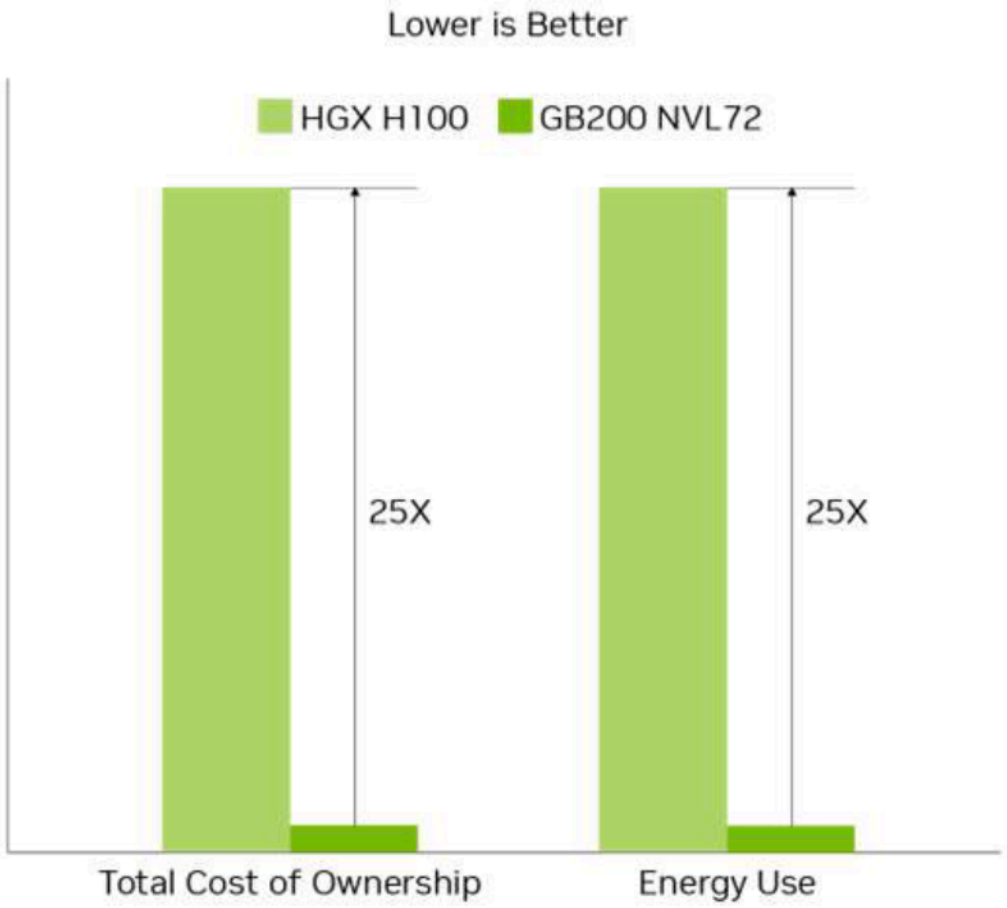
Application-specific integrated circuits are designed almost exclusively on CPUs in a long and complex workflow, which often includes an analog analysis to identify voltages and currents throughout. The Cadence SpectreX simulator is one example of a solver and runs 13x quicker on a GB200 than on an x86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

For trillion parameter AI models, compared to H100 air-cooled infrastructure, GB200 delivers 25X lower TCO and energy at the same performance[1].



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 7. 25X Lower Energy Use and TCO

Accelerated Networking Platforms for Generative AI

GB200 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-X800 Ethernet, and BlueField-3 DPUs, GB200 delivers unprecedented levels of performance, efficiency, and security in massive-scale AI data centers.

Quantum-X800 InfiniBand forms the foundation of the AI compute fabric, capable of scaling beyond 10,000 GPU in a two-level fat tree topology, which is 5X higher than the previous NVIDIA Quantum-2 generation. NVIDIA Spectrum-X800 and BlueField-3 DPU platforms are used to scale across the data center, providing accelerated GPU access to data, secure cloud multi-tenancy, and efficient data center operations.

NVIDIA Blackwell HGX

The NVIDIA Blackwell HGX B200 and HGX B100 include the same groundbreaking advancements for generative AI, data analytics, and high-performance computing and extend the HGX to include Blackwell GPUs.

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.

HGX B100: A Blackwell x86 platform based on an eight-Blackwell GPU B100 baseboard, delivering 112 petaFLOPs of AI performance. HGX B100 is a premier accelerated x86 scale-up platform designed for the fastest time to deployment with drop-in replacement compatibility for existing HGX H100 infrastructure. Each GPU is configurable up to 700 Watts per GPU.

Table 3. System Specifications for HGX B200 and HGX B100

	HGX B200	HGX B100
Blackwell GPUs	8	8
FP4 Tensor Core	144 PetaFLOPS	112 PetaFLOPS
FP8/FP6/INT8	72 PetaFLOPS	56 PetaFLOPS
Fast Memory	Up to 1.5 TB	Up to 1.5TB
Aggregate Memory Bandwidth	Up to 64 TB/s	Up to 64 TB/s
Aggregate NVLink Bandwidth	14.4 TB/s	14.4 TB/s
Per Blackwell GPU Specifications		
FP4 Tensor Core	18 petaFLOPS	14 petaFLOPS
FP8/FP6 Tensor Core	9 petaFLOPS	7 petaFLOPS
INT8 Tensor Core	9 petaOPS	7 petaOPs
FP16/BF16 Tensor Core	4.5 petaFLOPS	3.5 petaFLOPS

TF32 Tensor Core	2.2 petaFLOPS	1.8 petaFLOPS
FP32	80 teraFLOPS	60 teraFLOPS
FP64 Tensor Core	40 teraFLOPS	30 teraFLOPS
FP64	40 teraFLOPS	30 teraFLOPS
GPU memory | Bandwidth	Up to 192 GB HBM3e | Up to 8 TB/s	
Max thermal design power (TDP)	1,000W	700W
Interconnect	NVLink: 1.8TB/s PCIe Gen6: 256GB/s	NVLink: 1.8TB/s PCIe Gen6: 256GB/s
Server options	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs	NVIDIA HGX B100 partner and NVIDIA-Certified Systems with 8 GPUs

Preliminary specifications subject to change.
All petaFLOPS and petaOPS are with Sparsity except FP64 which is dense.

NVIDIA Blackwell Architecture's Role in the Age of Generative AI

Generative AI has elevated computing to a new era, characterized by the staggering capabilities of AI models boasting 10 trillion or more parameters. When AlexNet kicked off the AI boom in 2012, it used 60 million parameters. Just over a decade later, today's complexity has surged over 160,000 fold.

These new models can now find cures for cancer, predict extreme weather events, automate robots to perform industrial inspections, and unlock new economic opportunities in every industry. Yet, the journey to harness their full potential presents challenges, notably the vast computational resources and time required for model training. The new extremely large LLMs combined with the need for real-time inference reveals more challenges and complexities of scale, deployment, and operations.

NVIDIA Blackwell is the once-in-a-generation platform with the power and energy efficiency needed to effectively train and infer from these models and serve as the foundation for the age of generative AI. Blackwell architecture will be deployed into trillion dollar markets and democratize real time usage of the new gargantuan models. Training these models needs NVIDIA Blackwell's exaFLOPs of compute. Deploying them requires dozens of Blackwell GPUs to work as a single unified GPU.

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.