

The Engine Behind AI Factories | X

https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

NVIDIA Products Solutions Industries Shop Drivers Support Q Us

Cloud & Data Center Solutions Products Data Center GPUs Software Technologies Resources Get Started

NVIDIA Blackwell Architecture

The engine behind AI factories for the age of AI reasoning—now in full production.

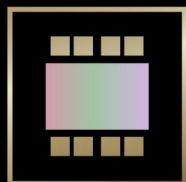
[Read Technical Brief](#)

Introduction Technological Breakthroughs Products Technical Brief [Contact Sales](#)

Breaking Barriers in Accelerated Computing and Generative AI

Explore the groundbreaking advancements the NVIDIA Blackwell architecture brings to generative AI and accelerated computing. Building upon generations of [NVIDIA technologies](#), NVIDIA Blackwell defines the next chapter in generative AI with unparalleled performance, efficiency, and scale.

Look Inside the Technological Breakthroughs



A New Class of AI Superchip

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

Second-Generation Transformer Engine

The second-generation Transformer Engine uses custom NVIDIA Blackwell Tensor Core technology combined with NVIDIA TensorRT™-LLM and NeMo™ Framework innovations to accelerate inference and training for large language models (LLMs) and Mixture-of-Experts (MoE) models. NVIDIA Blackwell Tensor Cores add new precisions, including new community-defined microscaling formats, giving high accuracy and ease of replacement for larger precisions.

NVIDIA Blackwell Ultra Tensor Cores are supercharged with 2X the attention-layer acceleration and 1.5X more AI compute FLOPS compared to NVIDIA Blackwell GPUs. The NVIDIA Blackwell Transformer Engine utilizes fine-grain scaling techniques called micro-tensor scaling, to optimize performance and accuracy enabling 4-bit floating point (FP4) AI. This doubles the performance and size of next-generation models that memory can support while maintaining high accuracy.



Secure AI

NVIDIA Blackwell includes NVIDIA Confidential Computing, which protects sensitive data and AI models from unauthorized access with strong hardware-based security. NVIDIA 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™. NVIDIA Blackwell Confidential Computing delivers nearly identical throughput performance compared to unencrypted modes. Enterprises can now secure even the largest models in a performant way, in addition to protecting AI intellectual property (IP) and securely enabling confidential AI training, inference, and federated learning.

[Learn More About NVIDIA Confidential Computing](#)

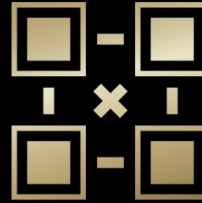


NVLink and NVLink Switch

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 NVIDIA NVLink interconnect can scale up to 576 GPUs to unleash accelerated performance for trillion- and 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. The NVIDIA NVLink Switch Chip supports clusters beyond a single server at the same impressive 1.8TB/s interconnect. Multi-server clusters with NVLink scale GPU communications in balance with the increased computing, so NVL72 can support 9X the GPU throughput than a single eight-GPU system.

[Learn More About NVIDIA NVLink and NVLink Switch](#)



Decompression Engine

Data analytics and database workflows have traditionally relied on CPUs for compute. Accelerated data science can dramatically boost the performance of end-to-end analytics, speeding up value generation while reducing cost. Databases, including Apache Spark, play critical roles in handling, processing, and analyzing large volumes of data for data analytics.

NVIDIA 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.

Reliability, Availability, and Serviceability (RAS) Engine

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



NVIDIA Blackwell Maximizes ROI in AI Inference

NVIDIA Blackwell enables the highest AI factory revenue: A \$5M investment in GB200 NVL72 generates \$75 million in token revenue—a 15x return on investment. This includes deep co-design across NVIDIA Blackwell, NVLink™, and NVLink Switch for scale-out; NVFP4 for low-precision accuracy; and NVIDIA Dynamo and TensorRT™ LLM for speed and flexibility—as well as development with community frameworks SGLang, vLLM, and more.

[Explore Key Results](#)



NVIDIA Blackwell Products



NVIDIA GB300 NVL72

The NVIDIA GB300 NVL72 delivers unparalleled AI reasoning inference performance, featuring 65X more AI compute than Hopper systems.



NVIDIA DGX SuperPOD

NVIDIA DGX SuperPOD™ is a turnkey AI data center solution that delivers leadership-class accelerated infrastructure with scalable performance for the most demanding AI training and inference workloads.



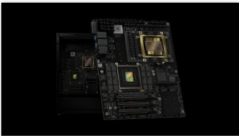
NVIDIA RTX PRO in the Data Center

Deliver powerful AI and graphics acceleration, essential enterprise features, and the flexibility to handle a wide range of workloads, from agentic and physical AI to visual computing and virtual workstations accelerated by NVIDIA RTX PRO™ data center GPUs.



NVIDIA RTX PRO Workstations

Bring the latest breakthroughs in AI, ray tracing, and neural graphics technology to power the most innovative workflows in design, engineering, and beyond with NVIDIA RTX PRO GPUs.

[Learn More >](#)

NVIDIA DGX Station

Unlike any AI desktop computer before, this system features NVIDIA Blackwell GPUs, the Grace CPU Superchip, and large coherent memory, delivering unparalleled compute performance.

[Learn More >](#)

[Learn More >](#)

NVIDIA DGX Spark

A compact, personal AI supercomputer with the NVIDIA GB10 Grace Blackwell Superchip, delivering high-performance AI capabilities and support for models up to 200 billion parameters.

[Learn More >](#)

[Learn More >](#)

NVIDIA HGX B300

NVIDIA HGX™ B300 is built for the age of AI reasoning with enhanced compute and increased memory.

[Learn More >](#)

[Learn More >](#)

NVIDIA GB200 NVL72

The NVIDIA GB200 NVL72 connects 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs in a rack-scale, liquid-cooled design.

[Learn More >](#)

NVIDIA NVFP4 Technical Blog

Learn how NVIDIA's new 4-bit NVFP4 quantization for pretraining unlocks huge improvements in training LLMs at scale and overall infrastructure efficiency.

[Read Now](#)

Products

- [Data Center GPUs](#)
- [NVIDIA DGX Platform](#)
- [NVIDIA HGX Platform](#)
- [Networking Products](#)
- [Virtual GPUs](#)

Technologies

- [NVIDIA Blackwell Architecture](#)
- [NVIDIA Hopper Architecture](#)
- [MGX](#)
- [Confidential Computing](#)
- [Multi-Instance GPU](#)
- [NVLink-C2C](#)
- [NVLink/NVSwitch](#)
- [Tensor Cores](#)

Resources

- [Accelerated Apps Catalog](#)
- [Blackwell Resources Center](#)
- [Data Center GPUs](#)
- [Data Center GPU Line Card](#)
- [Data Center GPUs Resource Center](#)
- [Data Center Product Performance](#)
- [Deep Learning Institute](#)
- [Energy Efficiency Calculator](#)
- [GPU Cloud Computing](#)
- [MLPerf Benchmarks](#)
- [NGC Catalog](#)
- [NVIDIA-Certified Systems](#)
- [NVIDIA Data Center Corporate Blogs](#)
- [NVIDIA Data Center Technical Blogs](#)
- [Qualified System Catalog](#)
- [Where to Buy](#)

Company Info

- [About Us](#)
- [Company Overview](#)
- [Investors](#)
- [Venture Capital \(NVentures\)](#)
- [NVIDIA Foundation](#)
- [Research](#)
- [Social Responsibility](#)
- [Technologies](#)
- [Careers](#)

[Follow Data Center](#) [f](#) [in](#) [X](#) [v](#)

