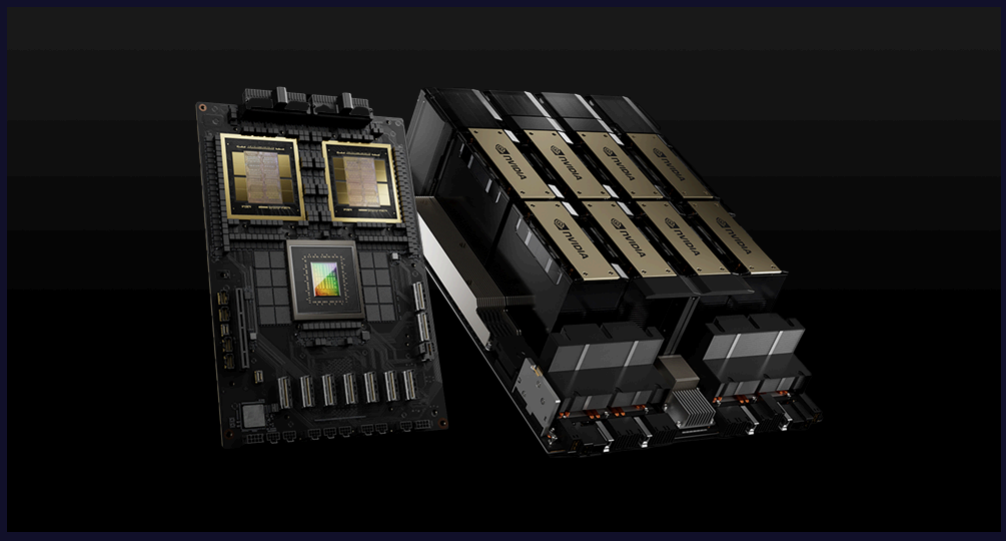
October 1, 2024 • 5 min read Updated on 25 Feb 2025

NVIDIA Blackwell vs NVIDIA Hopper: A Detailed Comparison

Innovation AI LLM Gen AI Deep Learning ...



Written by

Damanpreet Kaur Vohra
Technical Copywriter, NexGen cloud

Share this post

In our blog, we will discuss the key differences between NVIDIA Hopper and Blackwell architectures, comparing their performance, features, and suitability for AI and HPC workloads. NVIDIA Hopper, launched in 2022, introduced powerful AI capabilities with 80 billion transistors and HBM3 memory, excelling in LLMs and scientific computing. Blackwell, released in 2024, delivers a massive leap with 208 billion transistors, second-gen Transformer Engine, and 10 TB/s interconnect, making it ideal for generative AI and large-scale simulations. We'll explore how these architectures impact AI acceleration and real-world applications.

"We created a processor for the generative AI era," said NVIDIA's CEO Jensen Huang at GTC 2024, as he announced the highly anticipated NVIDIA Blackwell chip. This announcement broke records in the AI space. Designed to meet the demands of the most complex AI models and data-intensive workloads, Blackwell delivers a 2.5x performance boost over its previous-gen architecture NVIDIA Hopper.

While NVIDIA's Hopper architecture set a new standard for AI and accelerated computing in 2022, Blackwell takes things to another level. But what makes each of these architectures unique and how do they stack up against each other? In this blog, we'll explore the key differences between NVIDIA Hopper and Blackwell, and which is best suited for your AI and HPC projects.

NVIDIA Hopper vs Blackwell: Key Features

The table below shows the NVIDIA Hopper vs NVIDIA Blackwell specs:

Feature	NVIDIA Hopper Architecture	NVIDIA Blackwell Architecture
Transistor Count	80 billion transistors	208 billion transistors

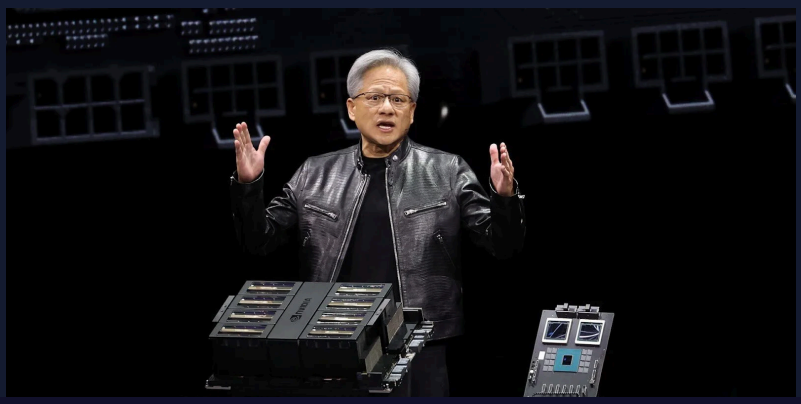
Transformer Engine	First-generation	Second-generation
Decompression Engine	No	Yes
Energy Efficiency	Improved over the previous generation	25x more energy-efficient than Hopper
Interconnect Technology	Fourth-generation NVLink	Fifth-generation NVLink
Chip-to-Chip Interconnect	900 GB/s	10 TB/s
Applications	Generative AI, LLMs, Data processing, Quantum computing	Accelerated Computing, AI, LLM

About NVIDIA Hopper

NVIDIA Hopper was launched in 2022 and named after Grace Hopper, a pioneering computer scientist and U.S. Navy rear admiral who was instrumental in the development of computer programming languages. The NVIDIA Hopper architecture excels in tasks like transformer-based AI models, large-scale language models (LLMs), and scientific computing. With 80 billion transistors and HBM3 memory, the H100 GPU offers up to 4 petaflops of AI performance, making it ideal for enterprises and research institutions working on AI, large-scale data centres, and scientific simulations. Hopper's Transformer Engine is key for accelerating tensor operations, essential for deep learning tasks.

Beyond performance, Hopper also introduced Confidential Computing capabilities, which offer enhanced data security, allowing enterprises to protect sensitive information while performing complex computations.

About NVIDIA Blackwell



NVIDIA Blackwell was released in 2024 and named after David Harold Blackwell, a renowned statistician and mathematician whose work in probability theory and dynamic programming has had a lasting impact on computational sciences. Blackwell's contributions to the field of mathematical statistics align with the architecture's focus on generative AI, LLMs, and data-centric workloads.

The NVIDIA GB200 GPU from the Blackwell series boasts 208 billion transistors and HBM3e memory, offering a massive leap in performance compared to the Hopper series. With up to 20 petaflops of AI performance, Blackwell is designed to handle the most demanding computational tasks, such as training large AI models, running complex simulations, and accelerating generative AI applications.

One of Blackwell's standout features is its dedicated decompression engine, which speeds up data processing by up to 800 GB/s—making it 6x faster than Hopper when dealing with large datasets. Blackwell also enhances Confidential Computing, ensuring a secure and efficient environment for sensitive AI workloads, while introducing breakthroughs in quantum simulations and scientific research, making it the next frontier in AI and HPC computing.

NVIDIA Hopper vs Blackwell: Performance Comparison

The performance of NVIDIA Hopper architecture and Blackwell benchmarks shows the significant advancements each architecture brings to cater to different aspects of AI and high-performance computing (HPC).

NVIDIA Hopper Series H100

The NVIDIA H100 Tensor Core GPU is a powerhouse for AI inference, training, and accelerated computing. Leveraging innovations from the Hopper architecture, the H100 is equipped with 80 billion transistors and HBM3 memory, delivering breakthrough performance and scalability for AI models, HPC tasks, and enterprise data centres. The H100 features a Transformer Engine with FP8 precision, offering up to 4x faster training compared to previous generations for large models like Llama 3.

Key capabilities of the NVIDIA Hopper Series H100 GPU include:

- **Up to 60 teraflops** of FP64 computing for HPC applications.
- **188GB HBM3 memory** with high bandwidth for efficient processing of large datasets.
- **NVIDIA NVLink** provides 900 GB/s bidirectional throughput, making multi-GPU scaling across servers seamless.

The NVIDIA H100 GPUs are designed to supercharge large language models (LLMs) like Llama 3, offering up to 5x faster inference compared to the NVIDIA A100 and with enterprise-ready features like NVIDIA Confidential Computing and NVIDIA AI Enterprise software, it simplifies AI deployment with high levels of security, scalability, and manageability.

NVIDIA Blackwell Series GB200

The NVIDIA Blackwell GB200 NVL/72 36 GPU shows a major leap in generative AI and accelerated computing. Built using 208 billion transistors and featuring HBM3e memory, the NVIDIA GB200 NVL/72 36 provides massive computational power for AI training, inference, and large-scale simulations.

Key NVIDIA Blackwell features of NVIDIA GB200 NVL/72 36 include:

- **Second-generation Transformer Engine** with new micro-tensor scaling techniques for **FP4 precision**, doubling the performance of next-gen AI models while maintaining high accuracy.
- **10 TB/s chip-to-chip interconnect**, enabling faster data communication within multi-die GPUs for enhanced generative AI processing.
- **NVIDIA Decompression Engine**, delivering up to **900 GB/s** bandwidth, significantly accelerating data processing for large datasets and analytics workloads.
- **Confidential Computing** capabilities, protecting sensitive AI models with hardware-based security and **TEE-I/O** integration.

With **up to 20 petaflops** of AI performance, Blackwell is ideal for generative AI models, LLMs, and

NVIDIA Hopper vs Blackwell GPU: Products

As an official NVIDIA NCP Partner, we integrate the latest NVIDIA technologies with flexible configuration options so you can scale your AI and HPC projects. At **AI Supercloud**, we offer the following cutting-edge hardware:

Hopper Series

- **NVIDIA HGX H100**
- **NVIDIA HGX H200**

How to Access NVIDIA Hopper GPUs

The NVIDIA Hopper series, including the NVIDIA HGX H100 and NVIDIA HGX H200, are available on the AI Supercloud for deployment. These GPUs are designed for high-performance AI and accelerated computing workloads, offering industry-leading capabilities for tasks like large language model (LLM) inference, HPC, and enterprise AI solutions.

You can reserve the NVIDIA HGX H100 or NVIDIA HGX H200 and take advantage of our customised solutions tailored to your workload requirements. With fully managed services, on-demand scalability, and MLOps support, the AI Supercloud is optimized to help you deploy your AI projects faster and more efficiently.

Book a call today to explore how NVIDIA Hopper GPUs can accelerate your AI and HPC tasks, and to discuss custom configurations for your specific needs.

Book a Discovery Call

Blackwell Series

- **NVIDIA GB200 NVL72/36**

When are NVIDIA Blackwell GPUs Coming Out?

The highly anticipated NVIDIA Blackwell chip GB200 NVL72/36 is expected to be available by the end of 2025. These cutting-edge GPUs are designed to deliver exceptional performance for AI and high-performance computing (HPC) workloads. **We are one of the first Elite Cloud Partners in the NVIDIA Partner Network to offer NVIDIA Blackwell platform-powered compute services.** You can reserve the NVIDIA GB200 NVL72/36 in advance on the AI Supercloud and secure early access to the fastest performance for your AI projects. By reserving now, you'll be among the first to leverage Blackwell's unmatched capabilities in generative AI and large-scale model training.

Book a call today to discuss how our bespoke solutions and managed services for NVIDIA GB200 NVL72/36 can help take your AI initiatives to the next level.

Book a Discovery Call

Is Blackwell faster than Hopper?

Yes, NVIDIA Blackwell is up to 2.5 times faster than Hopper, offering a performance boost through advancements like the second-generation Transformer Engine, a decompression engine and a much faster chip-to-chip interconnect speed.

What Hopper GPUs does the AI Supercloud offer?

The NVIDIA Hopper series, including the NVIDIA HGX H100 and NVIDIA HGX H200, are available on the AI Supercloud for deployment. **Schedule a call with our solutions engineer** today to explore how NVIDIA Hopper GPUs can accelerate your AI and HPC tasks and to discuss custom configurations for your specific needs.

What is the difference between NVIDIA H100 and NVIDIA GB200?

The NVIDIA H100 features 80 billion transistors, HBM3 memory and excels in AI inference and HPC with up to 4 petaflops of AI performance. While the NVIDIA GB200 NVL72 offers 208 billion transistors, HBM3e memory and a 2.5x performance boost with up to 20 petaflops, ideal for generative AI and large-scale model training.

What is special about NVIDIA Blackwell?

NVIDIA Blackwell features 208 billion transistors, HBM3e memory, a second-generation Transformer Engine, and a decompression engine for ultra-fast data processing. Its 10 TB/s chip-to-chip interconnect and enhanced Confidential Computing make it ideal for generative AI, large-scale models, and HPC workloads.

When are NVIDIA Blackwell GPUs Coming Out?

NVIDIA Blackwell GPUs such as the NVIDIA GB200 NVL72/36, are expected to be available by the end of 2025. Early access reservations are available through AI Supercloud for those wanting to leverage its cutting-edge capabilities in generative AI and HPC.

Share this post

Stay Updated with NexGen Cloud

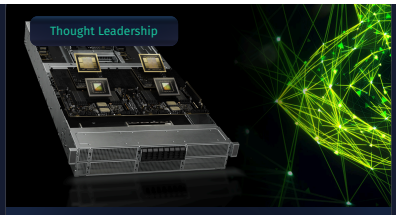
Subscribe to our newsletter for the latest updates and insights.

Enter your email to join our news | Subscribe

By subscribing, you agree to our Privacy Policy

Discover the Best

N3XGEN CLOUD



NexGen Cloud Part of First Wave to Offer ...

AI Supercloud will use NVIDIA Blackwell platform to drive enhanced efficiency, reduced costs and ...

March 19, 2024 • 5 min read

Announcements ...

NexGen Cloud and AQ Compute Advance Towards ...

AI Net Zero Collaboration to Power European AI London, United Kingdom – 26th February 2024; NexGen ...

February 27, 2024 • 5 min read

News Partnerships Announcements ...

WEKA Partners With NexGen Cloud to ...

NexGen Cloud's Hyperstack Platform and AI Supercloud Are Leveraging WEKA's Data Platform Software To ...

January 31, 2024 • 5 min read

News Partnerships Announcements ...

X (Twitter) LinkedIn Youtube

Facebook

N3XGEN CLOUD

NVIDIA Partner

NexGen Cloud is the AI Factory – accelerating the future with the industry's best GPUs in large-scale sovereign AI Cloud environments.

+44 (0) 203 475 3402

info@nexgencloud.com

Stay informed. Join our newsletter

Get the latest updates and exclusive offers.

Enter your email to join our newsletter*

Subscribe

By subscribing, you agree to our Privacy Policy

United Kingdom Address (Head office)

#1.07, 1st Floor
24 Greville St
London EC1N 8SS

Registered Office

Salisbury House, London Wall,
London EC2M 5PS,
United Kingdom

Spain Address

Ctra NACIONAL 340, KM 18.3
Local C-12,
Marbella 29600 Malaga

Products

- AI Supercloud
- Hyperstack
- NexGen Labs

Quicklinks

- Company
- Leadership Team
- Sustainability
- Security
- Partnerships
- Blog

AICPA SOC

N3XGEN CLOUD