

AI Server under GB200 Architecture

Article Overview

The NVIDIA GB200 architecture combines Grace CPUs and Blackwell GPUs into high-performance AI servers, enabling rack-scale AI workloads with unprecedented throughput and efficiency.

Overview of GB200 Architecture

The NVIDIA GB200 is a unified AI computing module that integrates one Grace CPU and two Blackwell GPUs into a single superchip, interconnected via NVLink-C2C for high-bandwidth, coherent memory access at 900 GB/s. This design allows the CPU and GPUs to share memory efficiently, reducing PCIe overhead and enabling large-scale AI and HPC workloads. Each GB200 superchip uses liquid cooling for the most power-intensive components, while other components are air-cooled to maintain optimal performance.

Rack-Scale Deployment

GB200 modules are deployed in rack-scale systems such as the DGX SuperPOD and Dell PowerEdge XE8712:

-

DGX SuperPOD: Built from Scalable Units (SU), each containing 8 DGX GB200 systems. A single DGX GB200 rack includes 18 compute trays, each with two GB200 superchips, totaling 72 GPUs per rack. The system leverages NVSwitch and NVLink networks for high-speed interconnects and supports large AI/ML workloads with hybrid liquid-air cooling

Article Content

Jun 30, 2026

Nvidia GB200 NVL2: Rack server for large AI models

Below the GB200 NVL72 server rack presented in March 2024, Nvidia is offering the GB200 NVL2 rack plug-in unit

Jan 28, 2026

Running AI Workloads on Rack-Scale Supercomputers: From

NVIDIA GB200 NVL72 and GB300 NVL72 leverage Blackwell architecture to provide rack-scale, high-density GPU

Apr 10, 2026

PowerEdge XE8712 with NVIDIA GB200: A High-Density

To meet that demand, Dell Technologies has introduced a new class of AI optimized servers: the Dell PowerEdge

Jan 15, 2026

100 MW Hyperscale AI Blueprint | Siemens

This technical paper outlines a reference blueprint for a 100 MW hyperscale AI data center designed to support NVIDIA GB200

Jul 18, 2026

AWS AI infrastructure with NVIDIA Blackwell: Two powerful compute ...

As AI capabilities evolve rapidly, you need infrastructure built not just for today's demands but for all the possibilities that

Oct 09, 2025

New Amazon EC2 P6e-GB200 UltraServers accelerated by NVIDIA

Amazon announces the general availability of EC2 P6e-GB200 UltraServers, powered by NVIDIA Grace Blackwell

Feb 14, 2026

ESC NM2N721-E1 | ASUS Servers

NVIDIA GB200 NVL72 delivers a computing power of 1,440 PFLOPS in FP4 precision, leveraging its advanced Tensor Cores and

Mar 26, 2026

Key Components of the DGX SuperPOD

NVIDIA Mission Control # The DGX GB200 SuperPOD Reference Architecture represents the best practices for

Sep 20, 2025

ASUS Announces Advanced AI POD Design Built with NVIDIA at

ASUS L11/L12-validated solutions empower enterprises to deploy AI at scale with confidence through world-class

Jul 22, 2026

Introducing the NVIDIA GB200 GPU

Recently, they released GB200 GPU, which is purpose-built for AI and high-performance computing. In this article, we

Sep 18, 2025

GB200 Hardware Architecture

We share estimates of the year-over-year change in BMC demands for general servers and AI servers based on the

Jan 23, 2026

NVIDIA GPU Cluster Interconnect: B200/B300/GB200/GB300

This system analyzes the cluster interconnection architecture of NVIDIA B200/B300/GB200/GB300, covering DGX,

Mar 19, 2026

NVIDIA GB200 AI Chip : Architecture, Working & Its Applications

This Article Elaborates on a Powerful Chip Built on the Blackwell Architecture, NVL72 System, AI Performance, Cooling System, &

Nov 09, 2025

NVIDIA B200 and GB200 AI GPUs Technical Overview: Unveiled at

At the 2024 GTC conference, NVIDIA introduced its new AI GPU models, the B200 and GB200, under the Blackwell

Apr 13, 2026

NVIDIA Blackwell 2026 — GB200, B200 AI Chips & Data Center

NVIDIA Blackwell The GPU architecture powering the AI revolution. GB200, B200, NVL72 — how NVIDIA's Blackwell chips became

Oct 14, 2025

NVIDIA GB200 User Guide: Specs, Features and Use Cases

Learn about the NVIDIA Blackwell GB200. Enterprise-scale AI with superior performance, advanced networking and

Sep 16, 2025

DGX SuperPOD Architecture

Each DGX GB200 leverages hybrid cooling with both direct liquid cooling and air cooling to manage the amount of

Jun 05, 2026

GB200 NVL72 | NVIDIA

GB200 NVL72 introduces cutting-edge capabilities and a second-generation Transformer Engine, which enables FP4 AI. When

Feb 16, 2026

Supercharge AI Tasks with the NVIDIA GB200

NVIDIA GB200's Blackwell architecture boosts AI performance, improving efficiency in healthcare, finance, and

Nov 15, 2025

GB200 NVL72 | NVIDIA

Compatible with liquid-cooled NVIDIA MGX™ modular servers, it provides up to 2x performance for scientific computing, AI for

Jul 01, 2026

Amazon P6e-GB200 UltraServers now available for the highest GPU ...

Amazon EC2 P6e-GB200 UltraServers offer the highest GPU-based AI training and inference performance in EC2.

Jan 29, 2026

Top 12 NVIDIA GPUs for AI Training & Inference in 2026

Compare the top 12 NVIDIA GPUs for AI in 2026, including H100, H200, B200, GB200, and RTX cards for training,

Sep 16, 2025

NVIDIA DGX SuperPOD: Next Generation Scalable Infrastructure for AI ...

NVIDIA DGX SuperPOD with NVIDIA DGX GB200 systems is the next generation of data center scale architecture to meet the

Aug 11, 2025

ASUS unveils its new AI POD: a complete rack of liquid

ASUS unveils its new ASUS AI POD at SC24: a new complete rack solution powered by

Jul 07, 2026

NVIDIA Blackwell Platform Arrives to Power a New Era of Computing

Powering a new era of computing, NVIDIA today announced that the NVIDIA Blackwell platform has arrived —

Nov 04, 2025

Unlocking the Future of AI with ASUS AI POD, featuring

The ASUS AI POD with NVIDIA GB200 NVL72 integrates cutting-edge hardware,

Feb 07, 2026

NVIDIA GB200 Supply Chain: The Global Ecosystem

Learn about the NVIDIA GB200 supply chain. We analyze the massive global ecosystem of hundreds of

Oct 22, 2025

NVIDIA Data Center GPU Specs: A Complete Comparison Guide

At OCP 2024, it published the GB200 NVL72 architecture, enabling a single rack to interconnect up to 72 GPUs via NVLink at 1.8

Nov 03, 2025

NVIDIA B200 vs B300 vs GB200 vs GB300: AI Cluster Interconnect ...

The evolution from B200 to B300 and from GB200 to GB300 reflects a broader shift in AI infrastructure design.

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: <https://www.iiss.co.za>

Email: sales@iiss.co.za

Phone: +27 63 174 2895

Address: Unit 8, Innovation Park, 15 Rivonia Road, Sandton, Johannesburg,
2196, South Africa

This document is for informational purposes only. Specifications subject to
change without notice.

