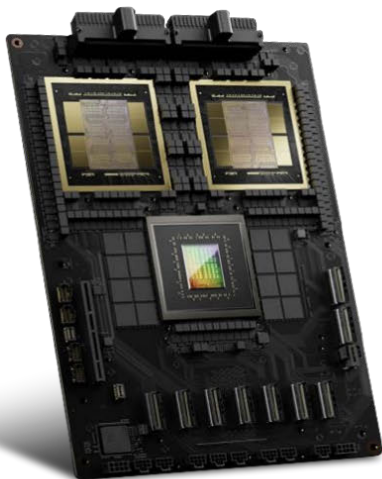


GB300



NVL72

BUILT FOR THE AGE OF AI INFERENCE



Key Features

- 36 NVIDIA Grace™ CPUs
- 72 NVIDIA Blackwell Ultra GPUs
- Up to 18TB of LPDDR5X memory with ECC
- Up to 21TB of HMB3e
- Up to 40TB of fast-access memory
- NVIDIA NVLink™ domain: 130 TB/s of low-latency GPU communication

Target Workloads

- AI Reasoning Inference
- Agentic and Physical AI

Next-Gen AI Supercomputer

- Powered by NVIDIA Blackwell Ultra GPUs and Grace CPUs
- Designed for trillion-parameter LLMs, real-time inference, and scientific computing

Unprecedented Performance

- 50 x overall increase in AI factory output performance vs. HGX H100
- 70 x more AI Flops (FP4 dense) vs. HGX H100
- Supports FP4, FP8, INT8, FP16, TF32, FP32, FP64 precision modes

Memory & Bandwidth

- 21 TB HBM3e memory across 72 GPUs
- Total aggregate 130 TB/s NVLink™ bandwidth via NVSwitch™

Networking & Scalability

- ConnectX-8 NICs: 800 GB/s networking bandwidth
- Supports multi-node GPU clusters with NVLink™ interconnect
- Built for hyperscale AI data centers

Enterprise & Cloud Ready

- Built for multi-tenant cloud and zero-trust environments
- Supports liquid cooling and composable infrastructure

Key Applications

- LLM Inference
- AI Training at Scale
- Data Analytics & Reasoning
- Scientific Computing & Simulation
- Enterprise & Cloud AI Infrastructure

NVIDIA GPU COMPARISON

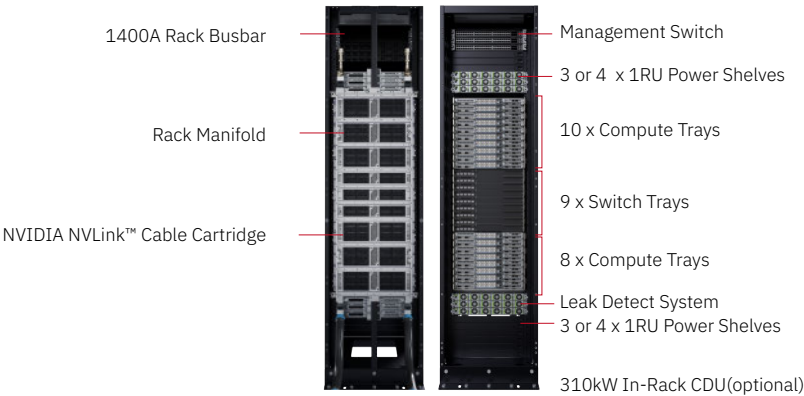
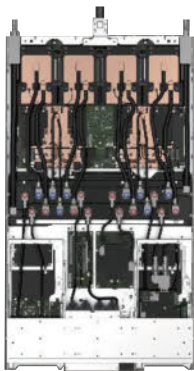
	GB300 NVL72	HGX B300 8-GPU	GB200 NVL72	HGX B200 8-GPU	HGX H200 8-GPU	HGX H200 4-GPU	HGX H100 8-GPU	HGX H100 4-GPU
Blackwell Ultra			Blackwell			Hopper		
GPUs	72	8	72	8	8	4	8	4
Total FP4*	1,100 1,440 PFLOPS	105 144 PFLOPS	720 1,440 PFLOPS	72 144 PFLOPS	- -	- -	- -	- -
Total FP8**	720 PFLOPS	72 PFLOPS	720 PFLOPS	72 PFLOPS	32 PFLOPS	16 PFLOPS	32 PFLOPS	16 PFLOPS
Total Memory Size	21TB HBM3e	2.3TB HBM3e	13.4TB HBM3e	1.4TB HBM3e	1.1TB HBM3e	564GB HBM3e	640GB HBM3e	320GB HBM3e
Total Memory Bandwidth	576TB/s	64TB/s	576TB/s	62TB/s	38.4TB/s	19.2TB/s	27GB/s	13GB/s
NVLink		5 th Generation				4 th Generation		
GPU-to-GPU Bandwidth		1.8TB/s				900GB/s		900GB/s
Total Aggregate Bandwidth	130TB/s	14.4TB/s	130TB/s	14.4TB/s	7.2TB/s	3.6TB/s	7.2TB/s	3.6TB/s

Per GPU

FP4*	15 20 PFLOPS	13 18 PFLOPS	10 20 PFLOPS	9 18 PFLOPS	- -	- -	- -	- -
FP8**	10 PFLOPS	9 PFLOPS	10 PFLOPS	9 PFLOPS	4 PFLOPS		4 PFLOPS	
FP16**	5 PFLOPS	4.4 PFLOPS	5 PFLOPS	4.5 PFLOPS	2 PFLOPS		2 PFLOPS	
Memory Size	288GB HBM3e	288GB HBM3e	186GB HBM3e	180GB HBM3e	141GB HBM3e		80GB HBM3e	
Memory Bandwidth	8TB/s	8TB/s	8TB/s	7.7TB/s	4.8TB/s		3.4TB/s	
Max TDP	1,400w	1,200w	1,200w	1,000w	700w		700w	

* Specifications shown are dense | sparse ** Specifications shown is sparse

Built with NVIDIA Blackwell Ultra, GB300 NVL72 Leads The NewEra of AI and The Next Wave of Accelerated Computing



1U with 2 x NVIDIA GB300 Grace™ Blackwell Ultra Superchips

CPU & GPU	2 x NVIDIA Grace™ Arm® Neoverse V2 CPUs(72-core) 4 x NVIDIA Blackwell Ultra Tensor Core GPUs
CPU Memory	960GB LPDDR5X Per Tray (2 x 480GB)
GPU Memory	1,152GB HBM3e Per Tray(4x288GB)
Networking	4 x NVIDIA NVLink™ Connectors(1.8TB/s) 4 x NVIDIA ConnectX®-8 OSFP Ports (800Gb/s) 1 x NVIDIA BlueField®-3 DPUs(400Gb/s)
Storage	4 x E1.S NVMe
BMC	Aspeed AST2600 BMC Controller
Security	TPM 2.0
Cooling System	Liquid-Cooled, 8 x4056 Fans
Form Factor/Dimension	1U MGX Chassis (WxHxD):438 x43.6x 766 mm (172.4x 17.2 x 301.6 inch)

NVIDIA GB300 NVL72 Rack

GPU	72 x NVIDIA Blackwell Ultra B300 GPUs
CPU	36 x NVIDIA Grace™ Arm® Neoverse V2 CPUs
Compute Tray	18 x NVIDIA GB300 Compute Trays
Networking	18 x NVIDIA BlueField®-3 DPUs, 72 x ConnectX®-8 SuperNICs
Switch Tray	9 x 5th Gen NVIDIA NVLink™ Switches Total NVLink™ Bandwidth-130TB/s
Management	2 x NVIDIA SN2201 Management Switches
Power Shelf	6 or 8 x 1U 33kW (6 x 5.5KW PSUs)
Rack Dimension (WxHxD)	48U NVIDIA MGX™ Rack 600 x 2296 x 1200 mm 23.6 x 90.4 x 47.2 inch
Weight	1480KG