GIGABYTE™ | NVIDIA®

DLB2-CB1

NVIDIA® GB200 NVL72: The Pinnacle of Rack-Scale Design

As the flagship product in the Blackwell lineup, the NVIDIA GB200 NVL72 boasts a fully liquid-cooled design and it uses an Arm-based NVIDIA Grace™ CPUs. This rack-scale configuration interconnects all nodes using the latest NVIDIA NVLink™ technology, delivering the performance of "one big GPU."

This cutting-edge solution outperforms the previous-generation NVIDIA HGX™ H100 GPU by 30x in inference and 4x in training, all while achieving a 25x reduction in TCO. With unmatched interconnect speeds and energy efficiency, the GB200 NVL72 sets a new benchmark for AI and HPC workloads.

LLM Inference	LLM Training	Energy Efficiency	Data Processing
30X	**4X**	**25X**	**18X**
vs. NVIDIA H100 GPU	vs. NVIDIA H100 GPU	vs. NVIDIA H100 GPU	vs. CPU

Key Features

- 36 NVIDIA Grace™ CPUs
- 72 NVIDIA Blackwell GPUs
- Up to 17 TB of LPDDR5X memory with error-correction code (ECC)
- Supports up to 13.5 TB of HBM3E
- Up to 30.5 TB of fast-access memory
- NVIDIA NVLink™ domain: 130 TB/s of low-latency GPU communication

Management Switches
- 2 x Out-of-band management switches
- 1 x Optional OS switch for rack-to-rack connection

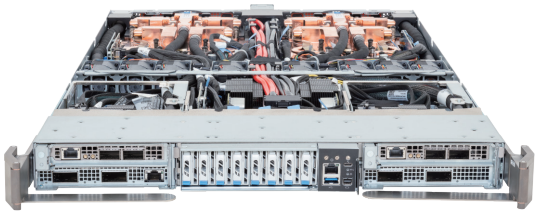
3 x 1RU 33kW Power Shelves

10 x Compute Trays
- 1RU XN14-CB0-LA01

9 x NVIDIA NVLink™ Switch Trays
- 1RU NVLink Switch tray
- 144 x NVLink ports per tray
- Fifth-generation NVLink with 1.8TB/s GPU-GPU interconnect

8 x Compute Trays
- 1RU XN14-CB0-LA01

3 x 1RU 33kW Power Shelves



XN14-CB0-LA01 Compute Tray

- 2 x NVIDIA GB200 Grace™ Blackwell Superchip
- 2 x 372GB HBM3e GPU memory with 16TB/s bandwidth
- 2 x 480GB LPDDR5X CPU memory with 1,024GB/s bandwidth
- 8 x E1.S Gen5 NVMe drive bays

Specifications

NVIDIA GB200 NVL72 Specs

	GB200 NVL72	GB200 Grace Blackwell Superchip
Configuration	36 Grace CPUs, 72 Blackwell GPUs	1 Grace CPU, 2 Blackwell GPUs
FP4 Tensor Core	1,440 PFLOPS	40 PFLOPS
FP8/FP6 Tensor Core	720 PFLOPS	20 PFLOPS
INT8 Tensor Core	720 POPS	20 POPS
FP16/BF16 Tensor Core	360 PFLOPS	10 PFLOPS
TF32 Tensor Core	180 PFLOPS	5 PFLOPS
FP32	5,760 TFLOPS	160 TFLOPS
FP64	2,880 TFLOPS	80 TFLOPS
FP64 Tensor Core	2,880 TFLOPS	80 TFLOPS
GPU Memory | Bandwidth	Up to 13.4 TB HBM3e | 576 TB/s	Up to 372 GB HBM3e | 16 TB/s
NVLink Bandwidth	130 TB/s	3.6 TB/s
CPU Core Count	2,592 Arm Neoverse V2 cores	72 Arm Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17 TB LPDDR5X | Up to 18.4 TB/s	Up to 480GB LPDDR5X | Up to 512 GB/s

DLB2-CB1 Rack Specs

Dimensions	1,068L x 600W x 2,299H mm
GPUs	72 x NVIDIA Blackwell GPUs
CPUs	36 x NVIDIA Grace™ CPUs
Compute Trays	18 x GIGABYTE XN14-CB0-LA01
NVLink Switch Trays	9 x NVIDIA NVLink™ Switch
Management Switch	2 x OOB management switches + 1 x Optional OS switch for rack-to-rack connection
Power Shelves	6 x 1RU 33kW power shelves
Bus Bars	1 x 54V DC 1400A Bus Bar
Cable Cartridges	4 x NVIDIA NVLink™ cable cartridges
CDU	Compatible with in-row or in-rack CDU
Part Numbers	9NDLB2CB1MR000

XN14-CB0-LA01 Specs

Form Factor	1RU Liquid-cooled server node
Superchip	2 x NVIDIA GB200 Grace™ Blackwell Superchip
Storage	8 x E1.S Gen5 NVMe drive bays 1 x M.2 (PCIe 5.0 x4)
Networking	2 x NVIDIA ConnectX®-7 NICs 2 x NVIDIA BlueField®-3 DPUs 1 x 1Gb/s LAN (Intel® I210-AT)
Front IO	1 x USB 3.2 Gen1, 1 x Mini-DP, 1 x RJ45, 1 x MLAN
Rear IO	4 x NVIDIA NVLink™ Switch connectors
OS Support	Ubuntu 22.04.3 arm64 Red Hat Enterprise Linux Server 9.3 aarch64 SUSE Linux Enterprise Server 15 SP5 aarch64
System Cooling	8 x 40x40x56mm Fans 1 x Superchip cold plate loops

Learn more at **https://www.GigaComputing.com/en**

Designed by GIGA COMPUTING