NVIDIA GB300 NVL72

Also known as: GB300 NVL72 · DGX GB300 NVL72 · Blackwell Ultra NVL72 · GB300 Grace Blackwell Ultra NVL72

The GB300 NVL72 is the Blackwell Ultra revision of NVIDIA's rack scale platform: the same 48U, 72 GPU, single NVLink domain layout as GB200 NVL72, but each B300 GPU carries 288 GB of HBM3e instead of roughly 186 GB, lifting the rack's NVLink-domain memory pool to 20 TB and its dense FP4 throughput to about 1.1 EFLOPS. It is aimed at teams running large scale reasoning and long context inference alongside training, where the extra memory per GPU and doubled 800 Gb/s scale out networking matter more than raw FLOPS growth. Facility demands go up along with the compute: reported operating power sits at 132 kW to 140 kW continuous, again 100 percent direct to chip liquid cooled with no rack fans.

72x Blackwell Ultra B300 GPU count · one NVLink domain, no PCIe fabric between GPUs	**20 TB HBM3e** NVLink domain memory · 288 GB per GPU, plus 17 TB LPDDR5X CPU memory	**132 kW to 140 kW** Rack power · operating power, reported
direct to chip liquid cooling Cooling · zero rack fans	**130 TB/s aggregate** NVLink bandwidth · 1.8 TB/s per GPU	

COMPUTE	01/05
GPU	72x NVIDIA Blackwell Ultra B300 · one NVLink domain
CPU	36x NVIDIA Grace · 72 core Arm Neoverse V2, 2,592 cores total
FP4 Tensor Core, dense	approximately 1.1 EFLOPS · reported, 72 x 15 PFLOPS per GPU, about 1.5x GB200 NVL72 dense FP4
FP8/FP6 Tensor Core	approximately 0.5 EFLOPS class · scaled from per-GPU Blackwell Ultra figures, reported
Compute trays	18x 1U trays · 4 GPU plus 2 Grace CPU per tray, 2 Superchips per tray

MEMORY	02/05
GPU memory (HBM3e)	**20 TB total** · 288 GB per GPU, up from roughly 186 GB on GB200 NVL72
GPU memory bandwidth	**up to 576 TB/s aggregate** · rack total, reported unchanged from GB200 NVL72
CPU memory (LPDDR5X)	**17 TB total** · 480 GB per Grace CPU
CPU memory bandwidth	**14 TB/s aggregate** · rack total
Combined fast memory	**approximately 37 TB** · HBM3e plus LPDDR5X, per NVIDIA

NVLINK DOMAIN	03/05
GPUs per domain	**72** · single non blocking NVLink domain
NVLink switch trays	**9x** · 4 ports per compute tray
Aggregate NVLink bandwidth	**130 TB/s** · GPU to GPU, rack total
Per GPU NVLink bandwidth	**1.8 TB/s** · NVLink 5
External scale out fabric	**ConnectX-8, up to 800 Gb/s per GPU** · Quantum-X800 InfiniBand or Spectrum-X Ethernet, double GB200 NVL72's 400 Gb/s

POWER AND LIQUID COOLING	04/05
Operating power	**132 kW to 140 kW** · Supermicro rack spec, reported
Peak/transient power	**up to 155 kW** · Lenovo/HPE class EDPp figure, reported
Power shelves	**8x 1U, 132 kW total** · 6x 5.5 kW PSUs with built in capacitor per shelf, N+N redundant
Cooling architecture	**100 percent direct to chip liquid cooling** · no air cooled components in the compute path
In-rack CDU capacity	**up to 250 kW, N+1 pumps** · or 1.8 MW in-row CDU supporting up to 8 racks
Liquid to air option	**up to 200 kW sidecar CDU** · no facility water required, for retrofit sites

PHYSICAL AND WEIGHT

05/05

Rack dimensions	2,236 mm x 600 mm x 1,068 mm · 48U, reference chassis, same footprint as GB200 NVL72
Rack weight	approximately 1.36 metric tons · reported, varies by OEM chassis and options
Floor loading	concentrated point load · reported in the same 1,800 to 2,000 kg per square meter class as GB200 NVL72, exceeds most raised floor ratings
Rack footprint	approximately 1.34 sq m · 600 mm x 1,068 mm base

PROCUREMENT

Typical lead time

6 to 12 months from order to delivery, reported. Industry wide NVL72 cabinet order backlog was estimated at 10,000 to 11,000 units entering 2026; full rack shipments began ramping in Q3 2025 per TrendForce, with first units delivered to CoreWeave via Dell that July.

Pricing

$3.7M to $4.0M per training class rack, reported (a Loop Capital analyst estimate derived from Apple's GB300 order). Inference optimized configurations have been reported as high as $6M to $6.5M. NVIDIA does not publish list pricing; figures vary by OEM, networking, and support tier.

- Sold exclusively through NVIDIA OEM and ODM partners (Dell was first to ship, followed by HPE, Lenovo, Supermicro, and Quanta), never direct from NVIDIA.
- CoreWeave, Microsoft Azure, and AWS all reached general availability on GB300 NVL72 instances within months of launch; enterprise buyers largely queue behind hyperscaler and neocloud allocation.
- TrendForce flags GB300 as the 2026 mainstream Blackwell Ultra SKU, pushing liquid cooling from a minority to a majority standard across new AI capacity that year.
- As with GB200 NVL72, deployment is normally bundled with the OEM's field services or a systems integrator for liquid cooling commissioning; this is not a self install product.

FIELD NOTES

- The jump from GB200 to GB300 NVL72 is a memory and networking upgrade more than a raw compute upgrade: 288 GB per GPU and 800 Gb/s scale out per GPU matter most for long context inference and multi rack training, not for single rack FLOPS.
- 132 to 140 kW continuous per rack pushes further past what most enterprise data centers can deliver than GB200 NVL72 already did; treat any facility not already qualified for 120 kW plus liquid cooled racks as needing a fresh power and cooling study, not a minor upgrade.
- Floor loading and slab requirements carry over unchanged from GB200 NVL72 since the reference chassis and rough weight are the same, but combine that with the higher power draw and the facility bar for GB300 is strictly higher, not just different.
- For most enterprise buyers who cannot host a 130 kW plus liquid cooled rack, an air cooled 8-GPU HGX B300 or B200 fleet remains the practical fallback, at the cost of losing the single 72-GPU NVLink domain and the larger per GPU memory pool.
- Because GB200 and GB300 NVL72 share a chassis footprint and power shelf design, a facility built out for GB200 is close to GB300 ready; the delta is mostly a higher cooling and power budget, not a redesign.

QUESTIONS WE GET ON THIS PART

How much does a GB300 NVL72 cost?

Reported estimates put a training class rack at $3.7M to $4.0M, based on an analyst's read of Apple's GB300 order; inference optimized configurations have been reported as high as $6M to $6.5M. NVIDIA does not publish an official price.

GB200 NVL72 vs GB300 NVL72: what actually changed?

Same 72 GPU, 48U, single NVLink domain layout, but GB300 uses Blackwell Ultra B300 GPUs with 288 GB HBM3e each (versus roughly 186 GB), pushing rack memory to 20 TB, dense FP4 throughput to about 1.1 EFLOPS, and scale out networking to 800 Gb/s per GPU via ConnectX-8, double GB200's 400 Gb/s.

What power does a GB300 NVL72 rack need?

Operating power is reported at 132 kW to 140 kW continuous, with peak or transient draw reported up to 155 kW by at least one OEM. That runs through the same 8-shelf, 480V three phase power architecture as GB200 NVL72.

Is GB300 NVL72 actually shipping, or still announced?

It is shipping. Dell delivered the first units to CoreWeave in July 2025, CoreWeave reached general availability that August, and Microsoft Azure and AWS both stood up GB300 NVL72 capacity within the following months.

REFERENCES

[1] https://www.nvidia.com/en-us/data-center/gb300-nvl72/

[2] https://www.supermicro.com/datasheet/datasheet_SuperCluster_GB300_NVL72.pdf

[3] https://www.tomshardware.com/pc-components/gpus/nvidia-announces-blackwell-ultra-b300-1-5x-faster-than-b200-with-288gb-hbm3e-and-15-pflops-dense-fp4

[4] https://introl.com/blog/why-nvidia-gb300-nvl72-blackwell-ultra-matters

[5] https://www.coreweave.com/blog/coreweave-leads-the-way-with-first-nvidia-gb300-nvl72-deployment

[6] https://www.dell.com/en-us/blog/dell-delivers-market-s-first-nvidia-gb300-nvl72-to-coreweave/

[7] https://www.trendforce.com/presscenter/news/20250318-12522.html

[8] https://www.aitooldiscovery.com/ai-infra/nvidia-gb300-explained

Latest version of this datasheet: https://gpusmith.com/hardware/systems/nvidia-gb300-nvl72. Specifications compiled from the references above; verify against vendor documentation before purchase.