



# NVIDIA GB200 NVL72

Also known as: GB200 NVL72 · DGX GB200 NVL72 · GB200 Grace Blackwell NVL72

The GB200 NVL72 is a liquid cooled, 48U rack that wires 72 Blackwell B200 GPUs and 36 Grace CPUs into a single NVLink domain, so software addresses one accelerator with a 13.4 TB HBM3e pool instead of 36 separate servers stitched together over InfiniBand. It targets teams training or serving frontier scale models that need that pool at NVLink speed rather than sharded across PCIe or network hops. Hosting one draws 125 kW to 135 kW of continuous IT load through direct to chip liquid cooling with zero rack fans, a facility class that only purpose built AI data centers and a small number of retrofitted colocation halls can deliver today.

|                                                                                         |                                                                              |                                                                   |
|-----------------------------------------------------------------------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------|
| <b>72x Blackwell B200</b><br>GPU count · one NVLink domain, no PCIe fabric between GPUs | <b>13.4 TB HBM3e</b><br>NVLink domain memory · plus 17 TB LPDDR5X CPU memory | <b>125 kW to 135 kW</b><br>Rack power · operating power, reported |
| <b>direct to chip liquid cooling</b><br>Cooling · zero rack fans                        | <b>130 TB/s aggregate</b><br>NVLink bandwidth · 1.8 TB/s per GPU             |                                                                   |

|                         |                                                                       |       |
|-------------------------|-----------------------------------------------------------------------|-------|
| COMPUTE                 |                                                                       | 01/05 |
| GPU                     | 72x NVIDIA Blackwell B200 · one NVLink domain                         |       |
| CPU                     | 36x NVIDIA Grace · 72 core Arm Neoverse V2, 2,592 cores total         |       |
| FP4 Tensor Core, sparse | 1,440 PFLOPS · rack aggregate                                         |       |
| FP4 Tensor Core, dense  | 720 PFLOPS · rack aggregate                                           |       |
| FP8/FP6 Tensor Core     | 720 PFLOPS · rack aggregate                                           |       |
| Compute trays           | 18x 1U trays · 4 GPU plus 2 Grace CPU per tray, 2 Superchips per tray |       |
| MEMORY                  |                                                                       | 02/05 |
| GPU memory (HBM3e)      | 13.4 TB total · roughly 186 GB per GPU, 372 GB per Superchip          |       |
| GPU memory bandwidth    | 576 TB/s aggregate · rack total                                       |       |
| CPU memory (LPDDR5X)    | 17 TB total · 480 GB per Grace CPU                                    |       |
| CPU memory bandwidth    | 14 TB/s aggregate · rack total                                        |       |

|                               |                                                                                                                                           |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| NVLink Domain03/05            |                                                                                                                                           |
| 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-7, up to 400 Gb/s per GPU · InfiniBand NDR or Spectrum-X Ethernet                                                                |
|                               |                                                                                                                                           |
| Power and Liquid Cooling04/05 |                                                                                                                                           |
| Operating power               | 125 kW to 135 kW · Supermicro rack spec, reported                                                                                         |
| Peak/transient power          | up to 192 kW · HPE EDPp figure, roughly 1.5x TDP, reported                                                                                |
| Power shelves                 | 8x 1U, 132 kW total · 6x 5.5 kW PSUs per shelf, N+N redundant                                                                             |
| Cooling architecture          | 100 percent direct to chip liquid cooling · no air cooled components in the compute path                                                  |
| Facility coolant loop         | ASHRAE W-series chilled water · reported supply range 18C to 25C, return up to 45C depending on CDU config                                |
| In-rack CDU capacity          | up to 250 kW · or 1.3 MW in-row CDU shared across multiple racks                                                                          |
|                               |                                                                                                                                           |
| Physical and Weight05/05      |                                                                                                                                           |
| Rack dimensions               | 2,236 mm x 600 mm x 1,068 mm · 48U, reference chassis                                                                                     |
| Rack weight                   | approximately 1.36 metric tons · reported, varies by OEM chassis and options                                                              |
| Floor loading                 | concentrated point load · reported around 1,875 kg per square meter, exceeds most 1,000 to 1,500 kg per square meter 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. Site prep (power, chilled water, floor reinforcement) and NVLink cabling commissioning typically add 2 to 4 more months before the rack is production ready.

### Pricing

\$2.8M to \$3.4M per rack, reported (an HSBC analyst estimate puts a training class NVL72 near \$3M). NVIDIA does not publish list pricing; final cost depends on OEM, networking configuration, storage, and support tier.

- Sold exclusively through NVIDIA OEM and ODM partners (Dell, HPE, Lenovo, Supermicro, Quanta, and others), never direct from NVIDIA.
- Hyperscalers and large neoclouds (Microsoft, Oracle, CoreWeave, Google, Amazon, Meta) hold the bulk of early allocation; enterprise orders queue behind those commitments.
- Deployment is normally bundled with a systems integrator or the OEM's own field services team for liquid cooling commissioning; few buyers self install a first rack.
- Available today as GB200 NVL72 cloud instances at CoreWeave and Oracle Cloud Infrastructure for teams that want the NVLink domain without owning the facility.

## FIELD NOTES

- 125 to 135 kW of continuous, liquid cooled load per rack is well outside what most enterprise data centers can deliver; a large majority of existing halls top out around 40 to 60 kW per rack on air cooling alone.
- The rack's point load (reported near 1,875 kg per square meter) exceeds standard raised floor ratings, so slab on grade installation or floor reinforcement is a real line item, not a footnote, in any facility budget.
- Zero rack fans means every watt of GPU, CPU, NVLink switch, and PSU heat lands on the facility water loop; a chiller or cooling tower undersized for this delta will throttle the whole rack, not just the hot node.
- For teams that cannot host 120 kW plus liquid cooled racks, a fleet of air cooled 8-GPU HGX B200 servers remains the realistic near term alternative, at the cost of losing the single NVLink domain across all 72 GPUs.
- Treat the 13.4 TB NVLink-domain memory figure as the actual reason to buy this system over a cluster of smaller nodes: it is what lets very large models live in one coherent address space instead of being sharded across a slower network fabric.

## QUESTIONS WE GET ON THIS PART

### How much does a GB200 NVL72 cost?

Reported figures cluster around \$2.8M to \$3.4M per rack, with an HSBC analyst estimate near \$3M for a training class configuration. NVIDIA does not publish an official price; actual cost depends on the OEM, networking, storage, and support contract.

### What power does a GB200 NVL72 rack need?

Operating power is reported at 125 kW to 135 kW continuous, with peak or transient draw reported as high as 192 kW by at least one OEM. That is delivered as 480V three phase into 8 power shelves feeding a fully liquid cooled compute path.

### Can a normal data center host a GB200 NVL72?

Not without significant upgrades. Most colocation and enterprise facilities are built for 40 to 60 kW air cooled racks; this system needs a facility chilled water loop, reinforced or slab on grade flooring for its concentrated weight, and 480V distribution most legacy halls do not have.

### Is GB200 NVL72 still current, or has GB300 replaced it?

Both are shipping. GB200 NVL72 remains in production and is deployed at CoreWeave, Oracle, Microsoft, and other hyperscalers; GB300 NVL72 is the newer Blackwell Ultra generation with more memory per GPU and higher FP4 throughput, sold alongside GB200 rather than as an immediate replacement.

#### REFERENCES

- [1] <https://www.nvidia.com/en-us/data-center/gb200-nvl72/>
- [2] [https://www.supermicro.com/datasheet/datasheet\\_SuperCluster\\_GB200\\_NVL72.pdf](https://www.supermicro.com/datasheet/datasheet_SuperCluster_GB200_NVL72.pdf)
- [3] <https://www.hpe.com/us/en/collaterals/collateral.a50009224enw.html>
- [4] <https://www.servethehome.com/this-is-the-nvidia-dgx-gb200-nvl72/>
- [5] <https://intro1.com/blog/gb200-nvl72-deployment-72-gpu-liquid-cooled>
- [6] <https://x.com/firstadopter/status/1940792115124228482>
- [7] <https://www.coreweave.com/news/coreweave-first-cloud-provider-to-announce-general-availability-of-nvidia-gb200-nvl72-instances>
- [8] <https://www.datacenterdynamics.com/en/news/nvidia-gb200-nvl72-now-available-via-oracle-cloud/>