# NVIDIA B300

Also known as: B300 SXM · Blackwell Ultra · HGX B300 · GB300

The B300 is NVIDIA's Blackwell Ultra data center GPU, a binned and re-engineered Blackwell die that trades the B200's 8-Hi HBM3e stacks for 12-Hi stacks to land 288 GB per GPU, and raises the power ceiling to 1,400 W to push dense NVFP4 throughput about 1.5x past B200. It ships on the same 8-GPU HGX baseboard pattern as B200 and is also the GPU building block of the GB300 NVL72 rack (72 GPUs, 36 Grace CPUs, fully liquid-cooled). In a private AI build it is the part to reach for when a single GPU's memory ceiling, not raw FLOPS, is what is blocking a large-context or reasoning-model deployment. As of mid-2026 it is NVIDIA's current top-of-stack Blackwell part, shipping since January 2026, and new OEM quotes are increasingly steering toward it over B200.

| | | |
|---|---|---|
| **288 GB HBM3e**<br>Memory · 8 TB/s bandwidth, 12-Hi stacks | **15 / 20 PFLOPS**<br>FP4 Tensor (NVFP4) · dense / sparse, per GPU | **5 / 10 PFLOPS**<br>FP8 Tensor · dense / sparse, per GPU |
| **up to 1400 W**<br>TGP · 1,100 W air bin exists; top bin needs liquid | **1.8 TB/s bidirectional**<br>NVLink · NVLink 5, per GPU | |

| COMPUTE | 01/05 |
|---|---|
| Architecture | Blackwell Ultra, dual-die · 208 billion transistors across both dies |
| Streaming Multiprocessors | 160 SMs · 8 graphics processing clusters |
| Tensor Cores | 640 · 5th generation |
| FP8 Tensor | 5 / 10 PFLOPS · dense / sparse |
| FP4 Tensor (NVFP4) | 15 / 20 PFLOPS · dense / sparse, 1.5x B200 dense |
| Attention throughput | ~2x B200 · doubled SFU throughput for attention layers |

| MEMORY | 02/05 |
|---|---|
| Capacity | 288 GB HBM3e · 50% more than B200's 180 GB |
| Stack configuration | 8 stacks, 12-Hi · vs B200's 8-Hi stacks |
| Bandwidth | 8 TB/s · per GPU |
| Bus width | 8,192-bit · 16 x 512-bit memory controllers |

| INTERCONNECT | 03/05 |
|---|---|
| NVLink | 1.8 TB/s bidirectional · NVLink 5, same generation as B200 |
| Die-to-die (NV-HBI) | 10 TB/s · internal, between the two dies |
| PCIe | PCIe Gen6, 256 GB/s · double B200's Gen5 host link |
| NVLink-C2C (Grace pairing) | 900 GB/s · coherent CPU-GPU link in GB300 Superchip |
| GB300 NVL72 rack fabric | 130 TB/s aggregate · 72-GPU NVLink domain |

| POWER & THERMAL | 04/05 |
|---|---|
| TGP | up to 1400 W · per GPU, ~40% above B200; liquid-cooled bin |
| Air-cooled bin | 1,100 W · per GPU, e.g. Supermicro 8U air-cooled HGX B300 and NVIDIA's air-cooled DGX B300 |
| 8-GPU baseboard power | 8.8 to 11.2 kW · GPUs only (1,100 W air to 1,400 W liquid bins), before CPUs and networking |
| Cooling | air (1,100 W bin) or direct liquid (up to 1,400 W) · chassis-dependent; DGX B300 is air-cooled at 1,500 CFM |
| DGX B300 system power | 14.5 kW max · 8-GPU system including CPUs, DPUs, storage, per NVIDIA user guide |

| PHYSICAL & PLATFORM | 05/05 |
|---|---|
| Form factor | SXM module · not sold as a discrete PCIe card |
| Baseboard | HGX B300 8-GPU · 2.3 TB aggregate HBM3e |
| GB300 NVL72 | 72 GPUs + 36 Grace CPUs · 1 rack, fully liquid-cooled |
| GB300 NVL72 FP4 (rack) | ~1.1 exaFLOPS dense · aggregate, sparse figure runs higher |
| Reference systems | DGX B300, GB300 NVL72, OEM HGX B300 · Supermicro, Dell, HPE, Lenovo qualifying |

## PROCUREMENT

Typical lead time

Reported 8 to 14 weeks for DGX B300 and HGX B300 orders as of mid-2026, longer for GB300 NVL72 rack-scale allocations; the 288 GB memory tier is described as in especially heavy demand relative to supply.

Pricing

Reported single-GPU pricing around $50,000; a fully configured 8-GPU DGX B300 system has been reported in the $300,000 to $500,000 range. A single GB300 NVL72 rack has been estimated at roughly $3.7 to $4.0 million by analyst commentary on large hyperscaler orders. Cloud rental has run roughly $2.50 to $18 per GPU-hour depending on provider and commitment. All figures reported and should be treated as directional, not vendor list prices.

- Like B200, B300 does not exist as a standalone card: it is an SXM module on an 8-GPU HGX baseboard or a GB300 Superchip tray, so the practical purchase unit is a baseboard, system, or rack.
- GB300 NVL72 and the 1,400 W GPU bin require direct liquid cooling, but air-cooled B300 paths exist at the 1,100 W bin: NVIDIA's DGX B300 is air-cooled (14.5 kW, 1,500 CFM) and Supermicro ships an 8U air-cooled HGX B300. Air still demands ~15 kW-per-node rack power, so the facility conversation happens either way.
- Reported hardware prices exclude liquid cooling infrastructure, power delivery upgrades, and facility work, which add materially to total deployment cost at these power densities.
- New Blackwell allocation from major OEMs is skewing toward B300 and GB300 over B200; expect B300 to be the default quote even when a workload would fit in B200's 180 GB.

## FIELD NOTES

- The FP4 sparse-to-dense ratio on B300 is only about 1.33x (15 to 20 PFLOPS), not the 2x you'd expect from structured sparsity and that B200 actually delivers (9 to 18 PFLOPS). Do not assume Blackwell Ultra scales sparsity the same way standard Blackwell does when you're modeling throughput.
- The 1,400 W bin is liquid-only, but B300 is not liquid-only: air-cooled chassis exist at the 1,100 W bin (DGX B300, Supermicro 8U HGX B300). The real facility constraint is 14 to 15 kW per node however you cool it; decide air 1,100 W vs liquid 1,400 W before the GPU order, because it moves both performance and the CDU/loop question.
- The jump to 288 GB matters most when a single model replica, its KV cache, or its optimizer state doesn't fit in 180 GB. If your working set fits comfortably on B200, the memory headroom on B300 buys you little beyond the ~1.5x dense FP4 uplift.
- H200 (141 GB HBM3e, ~700 W, Hopper software stack) is still the pragmatic choice for teams that need memory-bandwidth-bound inference today and cannot deliver 14 to 15 kW per node; even air-cooled B300 roughly doubles H200-node power density.
- Treat published NVFP4 throughput as a ceiling, not a delivered number: it assumes a quantization pipeline that actually produces valid NVFP4 weights and a kernel stack tuned for it. Benchmark your own model before sizing a cluster on the PFLOPS line.

## QUESTIONS WE GET ON THIS PART

### How much does an NVIDIA B300 cost?

Reported figures put a single B300 GPU around $50,000, with a fully configured 8-GPU DGX B300 system in the $300,000 to $500,000 range. A full GB300 NVL72 rack has been estimated near $3.7 to $4.0 million based on large hyperscaler order commentary. These are reported market figures, not published NVIDIA list prices.

### B200 vs B300: what's the actual difference?

B300 has 288 GB of HBM3e versus B200's 180 GB, a 1,400 W power ceiling versus B200's 1,000 W, and roughly 1.5x B200's dense NVFP4 throughput per GPU. B300 also requires direct liquid cooling in every deployment path, while B200 has air-cooled HGX options.

**Does B300 require liquid cooling?**

Only at the top power bin. The 1,400 W TGP bin and the GB300 NVL72 rack are direct-liquid-cooled, but air-cooled B300 systems ship at the 1,100 W bin: NVIDIA's own DGX B300 is air-cooled (14.5 kW max, 1,500 CFM), and Supermicro offers an 8U air-cooled HGX B300 (SYS-822GS-NBR).

**What is GB300 NVL72?**

GB300 NVL72 is NVIDIA's rack-scale system built from B300: 72 Blackwell Ultra GPUs and 36 Grace CPUs in one liquid-cooled rack, connected by a 130 TB/s aggregate NVLink fabric, delivering roughly 1.1 exaFLOPS of dense FP4 throughput at the rack level.

REFERENCES


[1] https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/

[2] https://www.nvidia.com/en-us/data-center/dgx-b300/

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

[4] 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

[5] https://vast.ai/article/everything-you-need-to-know-about-the-nvidia-blackwell-ultra-b300

[6] https://www.spheron.network/blog/nvidia-b300-blackwell-ultra-guide/

[7] https://getdeploying.com/gpus/nvidia-b300

[8] https://docs.nvidia.com/dgx/dgxb300-user-guide/introduction-to-dgxb300.html

[9] https://www.supermicro.com/en/accelerators/nvidia