NVIDIA DGX B300

Also known as: DGX B300 · D0B3-G2304 · DGX Blackwell Ultra

DGX B300 is NVIDIA's factory-built Blackwell Ultra appliance: eight B300 GPUs with 2.3 TB of HBM3e, dual Xeon 6776P hosts, eight ConnectX-8 NICs at 800 Gb/s each, and a single NVIDIA support contract in a 10U air-cooled chassis. It replaces DGX B200 at the top of the DGX line for inference and reasoning workloads, trading B200's stronger FP64 for 50% more GPU memory and about 1.5x the dense FP4 throughput. It is also the first DGX deployable in MGX racks, with front-facing I/O and a 54 VDC busbar variant alongside the classic 200-240 V PSU version. Same buyer logic as every DGX: pay the premium for NVIDIA-validated integration and support, or build the equivalent HGX B300 OEM box and keep BOM control.

8x B300 SXM GPUs · Blackwell Ultra, NVLink 5	**2.3 TB HBM3e** Total GPU memory · 8x 288 GB	**144 PFLOPS** FP4 inference · NVIDIA marketing figure
14.5 kW System max power · air-cooled	**10U rackmount** Form factor · MGX-rack deployable	**8x ConnectX-8** Networking · up to 800 Gb/s InfiniBand or Ethernet per port

COMPUTE	01/05
GPUs	8x NVIDIA B300 SXM · Blackwell Ultra architecture
FP4 inference	144 PFLOPS · NVIDIA figure; dense per-GPU NVFP4 is 15 PFLOPS
FP8 training	72 PFLOPS · NVIDIA figure
GPU-to-GPU fabric	NVLink 5 · 1.8 TB/s per GPU bidirectional
CPUs	2x Intel Xeon Platinum 6776P · 64 cores each, 2.3 GHz
vs. DGX B200	1.5x dense FP4, 2x attention throughput · B200 keeps the FP64 edge

MEMORY & STORAGE	02/05
Total GPU memory	2.3 TB HBM3e · 8x 288 GB
System memory	2 TB default · upgradable to 4 TB with 128 GB DIMMs
OS storage	2x 1.92 TB NVMe M.2 · boot
Data cache storage	8x 3.84 TB E1.S NVMe SED · ~30.7 TB raw

NETWORKING	03/05
Compute fabric NICs	8x ConnectX-8 single-port OSFP · up to 800 Gb/s InfiniBand/Ethernet each
NIC-to-GPU path	PCIe Gen6, direct to GPU · NICs live in the GPU tray, all I/O front-facing
In-band management/storage	2x BlueField-3 DPU dual-port QSFP112 · up to 400 Gb/s each
Host management	1 GbE RJ-45 on-board + 1 GbE BMC

POWER & COOLING	04/05
System max power	14.5 kW · NVIDIA user guide
Power supplies	12x 3.2 kW, N+N redundant · 200-240 VAC PSU version
DC variant	54 VDC busbar version · for MGX rack deployment
Cooling	Air-cooled · front-to-back
Airflow	1,500 CFM (PSU) / 1,350 CFM (busbar) · at 70% PWM
Operating temperature	10 C to 30 C (50 F to 86 F) · inlet; tighter than DGX B200's 35 C ceiling

PHYSICAL	05/05
Form factor	10U rackmount · first MGX-rack-deployable DGX
Height	17.4 in (442 mm)
Width	19 in (482.6 mm)
Depth	35.74 in (904.2 mm)
Weight	370.4 lb (168 kg) PSU version · 271.2 lb (123 kg) busbar version

PROCUREMENT

Typical lead time

Shipping since January 2026. Reported 8 to 14 weeks for single-system orders through NVIDIA partners as of mid-2026; the 288 GB memory tier is in especially heavy demand relative to supply, and large committed-volume orders get allocation priority.

Pricing

$300,000 to $350,000 baseline reported as of Q1 2026, with fully configured reseller quotes reaching $400,000 to $500,000. NVIDIA does not publish a list price; every quote runs through a partner (Exxact lists the system as SKU D0B3-G2304+P2CMI36 with 3-year Business Standard Support, quote-only). That is roughly $37,500 to $43,750 per GPU at system level.

- Reported pricing has DGX B300 launching cheaper than DGX B200's street range, which resellers attribute to the MGX-standard form factor cutting NVIDIA's build cost.
- The DGX premium buys NVIDIA-validated integration, firmware/driver stack and support, not more silicon than an HGX B300 OEM build.
- Choose the PSU version for classic 19-inch racks, the busbar version for MGX racks with 54 VDC distribution; the busbar unit is ~100 lb lighter but assumes rack-level power shelves.
- Networking BOM (switch ports, optics, cables for 8x 800 Gb/s) is a materially larger line item than on ConnectX-7 systems; budget the fabric alongside the node.

FIELD NOTES

- DGX B300 is air-cooled at 14.5 kW in a 10U chassis. That is a legacy-colo-killer, not a liquid-cooling problem: no coolant loop needed, but very few air-cooled racks deliver 14.5 kW to one node, let alone several, and 1,500 CFM per node makes hot-aisle containment mandatory.
- The inlet ceiling dropped to 30 C (vs 35 C on DGX B200). Facilities running warm-aisle setpoints for efficiency need to re-check their supply-air temperature before swapping B300 nodes into B200 slots.
- 144 PFLOPS FP4 is a marketing figure; dense per-GPU NVFP4 is 15 PFLOPS (120 PFLOPS per node). Size clusters on dense numbers and your own model benchmarks, not the headline.
- The 2.3 TB of HBM3e is the real purchase driver: a single node now holds a ~600B-parameter model in FP8 with KV-cache headroom. If your working set fits in DGX B200's 1.44 TB, the B300 premium buys you little for today's workload.
- ConnectX-8 at 800 Gb/s means Quantum-X800 InfiniBand or Spectrum-X800 Ethernet switch fabric to exploit it; pairing with a 400 Gb/s fabric halves the scale-out bandwidth you paid for.
- FP64 regressed versus B200. Teams with HPC-adjacent workloads (CFD, simulation) mixed into their AI cluster should keep B200 or H200 nodes for those jobs rather than assuming newer is faster everywhere.

QUESTIONS WE GET ON THIS PART

How much does a DGX B300 cost?

NVIDIA does not publish a list price. Reported baseline pricing is $300,000 to $350,000 as of Q1 2026, with fully configured reseller quotes running $400,000 to $500,000 depending on support term and region. Quotes go through NVIDIA partners.

Is the DGX B300 air-cooled or liquid-cooled?

Air-cooled. NVIDIA's user guide specifies front-to-back airflow of roughly 1,500 CFM and a 14.5 kW maximum draw in a 10U chassis, with a 10 C to 30 C inlet window. No liquid loop is required, but the per-rack power density usually forces high-density colo space anyway.

DGX B300 vs DGX B200: which one?

DGX B300 carries 2.3 TB of GPU memory (vs 1.44 TB), about 1.5x the dense FP4 throughput and 2x the attention throughput, plus 800 Gb/s ConnectX-8 networking. DGX B200 keeps a significant FP64 advantage and a wider 35 C inlet window. Reasoning/inference at scale favors B300; mixed HPC workloads can still favor B200.

DGX B300 vs HGX B300: what is the difference?

Same 8-GPU Blackwell Ultra platform either way. DGX B300 is NVIDIA's fixed-configuration appliance with NVIDIA-direct support; HGX B300 is the baseboard OEMs (Supermicro, Dell, HPE, Lenovo, Aivres) build their own servers around, including liquid-cooled and denser variants NVIDIA does not sell as DGX.

REFERENCES

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

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

[3] https://www.exxactcorp.com/NVIDIA-D0B3-G2304-P2CMI36-E9032838

[4] https://tech-insider.org/nvidia-blackwell-gpu-pricing/

[5] https://www.naddod.com/ai-insights/deep-dive-into-nvidia-dgx-b300

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