



NVIDIA HGX B200

Also known as: HGX B200 8-GPU baseboard · B200 SXM6 platform

HGX B200 is the 8-GPU Blackwell baseboard NVIDIA sells to OEMs, not a server you can rack yourself. Supermicro, Dell, Lenovo, HPE, ASUS and ASRock Rack each bolt their own CPUs, chassis, power delivery, storage and networking around the same baseboard and NVSwitch fabric, then sell the result as a complete server. Most private HGX builds go through one of those OEM boxes rather than NVIDIA's own DGX B200, because the OEM path lets a buyer choose CPU vendor, memory footprint, NIC mix and air vs liquid cooling to fit an existing rack and budget, while DGX B200 ships as a fixed configuration at a materially higher price for the same GPU count.

8x B200 SXM6 GPUs · Blackwell, 4 nm, fully NVLinked	1.4 TB HBM3e total GPU memory · 180 GB per GPU	~62 TB/s aggregate Memory bandwidth · 7.7 TB/s per GPU
14.4 TB/s aggregate NVLink · 5th-gen NVLink, 1.8 TB/s per GPU	1,000 W air / 1,200 W liquid GPU TDP · per GPU, configurable by OEM	~10.2 kW Baseboard power · 8-GPU baseboard only, air-cooled config

COMPUTE01/05	
GPU	NVIDIA B200 (Blackwell) · SXM6 form factor, 8 per baseboard
Process node	4 nm · TSMC 4NP
FP4 Tensor (dense)	144 PFLOPS · 8-GPU aggregate
FP8/FP6 Tensor	72 PFLOPS · 8-GPU aggregate
FP64 Tensor	296 TFLOPS · 8-GPU aggregate
GPU TDP	1,000 W (air) / 1,200 W (liquid) · configurable per GPU
MEMORY02/05	
Memory type	HBM3e · per GPU
Memory per GPU	180 GB
Total baseboard memory	1.4 TB · 8x 180 GB
Bandwidth per GPU	7.7 TB/s
Aggregate bandwidth	~62 TB/s · reported, sums per-GPU figures

INTERCONNECT		03/05
NVLink generation	5th generation	
GPU-to-GPU bandwidth	1.8 TB/s per GPU · bidirectional	
Aggregate NVLink bandwidth	14.4 TB/s · across all 8 GPUs	
Switch fabric	NVSwitch (5th-gen) · full any-to-any GPU mesh	
Scale-out networking	0.8 TB/s · baseboard-level figure; actual NICs are OEM-selected	
Typical OEM NICs	Dual/octal ConnectX-7 · e.g. 8x 400GbE QSFP + 2x 200GbE QSFP112 on Supermicro's 10U SKU	
POWER & COOLING		04/05
Baseboard power (air)	~10.2 kW · 8 GPUs at 1,000 W plus NVSwitch overhead	
GPU-only draw (air)	8 kW · 8x 1,000 W	
GPU-only draw (liquid)	9.6 kW · 8x 1,200 W	
Full server draw	10 to 15 kW · reported, adds dual CPUs, DRAM, NVMe, fans, PSUs	
Cooling options	Air or direct-to-chip liquid · OEM-selected; liquid unlocks the 1,200 W TDP bin	
OEM PLATFORMS		05/05
Ships from	Supermicro, Dell, Lenovo, HPE, ASUS, ASRock Rack · each builds its own chassis around the HGX baseboard	
Example SKUs	Supermicro AS-A126GS-TNBR, Dell PowerEdge XE9680L, Lenovo ThinkSystem SR780a V3	
CPU choice	AMD EPYC or Intel Xeon · OEM- and SKU-dependent	
vs. DGX B200	Same baseboard, OEM picks the rest · DGX is NVIDIA's own fixed-config, higher-priced equivalent	

PROCUREMENT

Typical lead time

8 to 16 weeks for standard ('Golden SKU') configurations through an OEM with an established NVIDIA allocation, reported. Custom BOMs (non-standard CPU, memory, or networking mix) have been reported beyond 20 weeks. Buyers without existing OEM/NVIDIA Partner Network relationships can face significantly longer waits given 2026 HBM3e supply constraints.

Pricing

\$300,000 to \$500,000+ for a fully configured 8-GPU server, reported. Aivres quotes an HGX B200 platform starting near \$340,000 depending on config; heavier storage, networking, and liquid-cooled builds push past \$500,000. NVIDIA's own DGX B200 (same baseboard, fixed configuration) is separately estimated at \$600,000 to \$750,000, reported, which is why most private builds go through an OEM instead.

- Pricing is per-server and excludes rack power, cooling infrastructure, and InfiniBand/Ethernet fabric switches, which the industry commonly estimates can rival the GPU cost itself over a build.
- Liquid-cooled SKUs carry a price premium over air-cooled but unlock the 1,200 W GPU TDP bin and denser rack packing.
- Individual B200 SXM6 modules have been quoted around \$30,000 to \$40,000 each when sold outside a full baseboard, reported.
- Confirm CPU vendor (AMD vs Intel) and NIC count early: these are the two variables that most affect both price and lead time across OEM SKUs.

FIELD NOTES

- HGX is a baseboard, not a server. The OEM decides chassis, power delivery, cooling method and network configuration; NVIDIA only fixes the GPU-to-GPU fabric. Two 'HGX B200' servers from different vendors can differ meaningfully in CPU, storage and total power draw.
- Air-cooled B200s cap at 1,000 W per GPU; liquid-cooled bins run to 1,200 W. If a build spec assumes the higher TDP figure, verify the OEM SKU actually ships with direct-to-chip liquid, not just air with a liquid-ready chassis.
- Budget rack power and cooling as a separate line item, not a rounding error. A single 8-GPU HGX B200 server can pull 10 to 15 kW, meaning a handful of nodes will exceed the power density of a legacy air-cooled rack.
- DGX B200 and HGX-OEM servers share the identical GPU baseboard and NVLink/NVSwitch fabric. The price gap between them buys NVIDIA's fixed configuration, software stack and direct support, not more silicon. Teams comfortable managing their own OS and driver stack save real money going OEM.
- Lead time, not price, is usually the binding constraint in 2026. Confirm actual allocation with the OEM's sales team before quoting a delivery date to a client; published 'ships in 4 to 6 weeks' figures generally assume an existing NVIDIA Partner Network relationship.

QUESTIONS WE GET ON THIS PART

How much does an HGX B200 server cost?

Reported street pricing for a fully configured 8-GPU HGX B200 server runs roughly \$300,000 to \$500,000 or more, depending on CPU, memory, storage, networking and cooling choices. This excludes rack power and cooling infrastructure.

What is the difference between HGX B200 and DGX B200?

They use the same 8-GPU Blackwell baseboard and NVLink/NVSwitch fabric. HGX B200 is the baseboard NVIDIA sells to OEMs like Supermicro, Dell, Lenovo, HPE, ASUS and ASRock Rack, who each build their own server around it. DGX B200 is NVIDIA's own fixed-configuration server built on that same baseboard, priced higher and sold with NVIDIA support.

How much power does an HGX B200 8-GPU server draw?

The GPUs alone draw 8 kW air-cooled (1,000 W each) or 9.6 kW liquid-cooled (1,200 W each). A fully configured server with dual CPUs, memory, storage and networking is reported to draw roughly 10 to 15 kW.

What is the lead time to get an HGX B200 server?

Reported lead times run 8 to 16 weeks for standard configurations through an OEM with an existing NVIDIA allocation, and beyond 20 weeks for custom BOMs. Buyers without an established OEM relationship should expect longer waits given ongoing HBM3e supply constraints.

REFERENCES

- [1] <https://docs.nvidia.com/enterprise-reference-architectures/hgx-ai-factory/latest/components.html>
- [2] <https://lenovopress.lenovo.com/lp2226-thinksystem-nvidia-b200-180gb-1000w-gpu>
- [3] <https://flopper.io/system/nvidia-hgx-b200-8-gpu>
- [4] <https://www.fibermall.com/blog/nvidia-hgx-b200-cooling-solution.htm>
- [5] <https://www.civo.com/blog/nvidia-dgx-vs-hgx>
- [6] <https://itctshop.com/nvidia-dgx-comparison-guide/>
- [7] <https://www.arccompute.io/arc-blog/aivres-nvidia-hgx-b200-b300-gpu-servers-air-and-liquid-cooled-performance-at-scale>
- [8] <https://www.servethehome.com/asrock-rack-8u8x-gnr2-syn-b200-review-nvidia-hgx-b200-8-gpu-server-intel-xeon/>
- [9] <https://ai-newspaper.com/articles/order-blackwell-b200-server-racks/>