NVIDIA B200

Also known as: B200 SXM6 · HGX B200 · Blackwell B200 · GB100

The B200 is NVIDIA's first-generation Blackwell data center GPU: two reticle-limited dies joined by a 10 TB/s NV-HBI link and packaged as one SXM6 module with 180 GB of HBM3e. It sells as a bare module only inside an 8-GPU HGX B200 baseboard, never standalone, and that baseboard is the building block for DGX B200 and third-party OEM systems (Supermicro, Dell, HPE, Lenovo). In a private AI build it is the training and large-batch inference workhorse for shops that need FP4 throughput and 180 GB of HBM per GPU but do not need the 288 GB tier or the extra power headroom of B300. As of mid-2026 it sits one rung below Blackwell Ultra (B300) in NVIDIA's current lineup and OEMs are shifting new quotes toward B300, but B200 fleets already deployed remain the majority of installed Blackwell capacity.

180 GB HBM3e Memory · 7.7 TB/s bandwidth	**9 / 18 PFLOPS** FP4 Tensor (NVFP4) · dense / sparse	**4.5 / 9 PFLOPS** FP8 Tensor · dense / sparse
1000 W TGP · SXM6, air or liquid depending on chassis	**1.8 TB/s bidirectional** NVLink · NVLink 5, per GPU	

COMPUTE	01/05
Architecture	Blackwell, dual-die · two reticle-limited dies, 10 TB/s NV-HBI
Transistors	208 billion · across both dies
FP64	37 TFLOPS · no sparsity
FP32	75 TFLOPS · no sparsity
FP16 / BF16 Tensor	2.25 / 4.5 PFLOPS · dense / sparse
FP8 Tensor	4.5 / 9 PFLOPS · dense / sparse
FP4 Tensor (NVFP4)	9 / 18 PFLOPS · dense / sparse
INT8 Tensor	4.5 / 9 POPS · dense / sparse

MEMORY	02/05
Capacity	180 GB HBM3e · software-visible; some early marketing cited 192 GB physical
Bandwidth	7.7 TB/s · often rounded to 8 TB/s in marketing collateral
MIG partitions	up to 7 instances · 23 GB per instance

INTERCONNECT	03/05
NVLink	1.8 TB/s bidirectional · NVLink 5, 900 GB/s per direction
GPU-to-GPU fabric	8-GPU HGX baseboard · fully connected NVLink domain
PCIe	PCIe Gen5, 128 GB/s · host connectivity
Die-to-die (NV-HBI)	10 TB/s · internal, between the two dies

POWER & THERMAL	04/05
TGP	1000 W · per GPU, SXM6
8-GPU baseboard power	~8-10 kW · GPUs only, excludes host CPUs and NICs
Cooling	air or direct liquid · chassis-dependent; air-cooled HGX B200 exists but liquid is common at rack density

PHYSICAL & PLATFORM	05/05
Form factor	SXM6 module · not sold as a discrete PCIe card at this tier
Baseboard	HGX B200 8-GPU · 1.4 TB aggregate HBM3e, 64 TB/s aggregate bandwidth
Reference systems	DGX B200, OEM HGX servers · Supermicro, Dell, HPE, Lenovo ship qualified chassis
Networking (typical)	ConnectX-7/8, up to 400 Gb/s · platform-dependent, not a GPU-level spec

PROCUREMENT

Typical lead time

Reported 6 to 16 weeks for allocated OEM orders as of mid-2026; varies heavily by NVIDIA Partner Network tier and existing allocation relationship.

Pricing

Reported MSRP band of $30,000 to $40,000 per GPU when bought in 8-GPU HGX clusters; loose SXM modules from secondary channels have been quoted at $45,000 to $55,000. An 8-GPU HGX B200 server (GPUs, host CPUs, networking, chassis) runs a reported $400,000 to $500,000. Cloud rental ranges roughly $2.70 to $16 per GPU-hour depending on provider and commitment. All figures reported, not vendor list prices.

- B200 is not sold as a standalone part: it ships soldered to an 8-GPU HGX baseboard, so procurement is effectively a system-level (or half-system) purchase, not a card purchase.
- New-build quotes from major OEMs are increasingly steering buyers toward B300 platforms; B200 supply is now mostly secondary market, existing allocation, or residual OEM inventory.
- Blackwell as a generation has been reported sold out into mid-2026 on a multi-million-unit backlog, so quoted lead times track allocation tier more than manufacturing capacity.
- Budget for liquid cooling and power delivery separately: an 8-GPU HGX B200 baseboard alone draws on the order of 8 to 10 kW before host CPUs, NICs, or storage.

FIELD NOTES

- The 180 GB vs 192 GB memory figure is a real discrepancy across NVIDIA's own materials, not a typo you introduced. Design and quote against 180 GB software-visible HBM3e; treat 192 GB as the physical stack size, not usable capacity.
- At 1000 W per GPU and eight GPUs per baseboard, air cooling is thermally possible but tight; most new HGX B200 deployments at rack scale go direct-liquid-cooled from day one rather than retrofitting later.
- H200 still makes sense if your workload is memory-bandwidth-bound inference on 70 to 100B-parameter dense models and you want to stay on a Hopper-compatible stack with no kernel retuning; B200 wins once you can exploit FP4 or need the larger NVLink domain for training.
- FP4 throughput numbers are peak, structured-sparsity figures. Real inference gains depend on whether your quantization pipeline actually produces NVFP4-compatible weights; dense FP8 remains the safer default for production serving until FP4 accuracy is validated on your model.
- Because B200 only exists on an 8-GPU baseboard, any capacity planning in units smaller than 8 GPUs is a fiction; plan procurement, power, and cooling in whole-baseboard increments.

QUESTIONS WE GET ON THIS PART

How much does an NVIDIA B200 cost?

Reported pricing lands around $30,000 to $40,000 per GPU when purchased inside an 8-GPU HGX cluster, with an 8-GPU HGX B200 server running roughly $400,000 to $500,000 fully configured. Loose modules on secondary channels have traded higher, $45,000 to $55,000. These are reported market figures, not published NVIDIA list prices.

How much memory does the B200 actually have?

180 GB of HBM3e is the software-visible, spec-sheet number used by NVIDIA's own OEM documentation (Lenovo, Dell). Some earlier marketing referenced 192 GB, which reflects the physical HBM3e stack size before reserved capacity, not what a workload can allocate.

B200 vs B300: which one should I buy?

B300 (Blackwell Ultra) adds 288 GB of HBM3e versus B200's 180 GB, a higher 1,400 W power ceiling, and roughly 1.5x the dense NVFP4 throughput per GPU. Choose B200 for training and inference workloads that fit comfortably in 180 GB per GPU and where OEM allocation favors it; choose B300 for large-context or trillion-parameter-class reasoning workloads that need the extra memory headroom.

Can I buy a single B200 GPU?

No. B200 is an SXM6 module soldered to an 8-GPU HGX baseboard; there is no standalone PCIe B200 card. The practical purchase unit is the 8-GPU baseboard or a full DGX/OEM server built around it.

REFERENCES

[1] https://lenovopress.lenovo.com/lp2226-thinksystem-nvidia-b200-180gb-1000w-gpu

[2] https://www.tomshardware.com/pc-components/gpus/nvidias-next-gen-ai-gpu-revealed-blackwell-b200-gpu-delivers-up-to-20-petaflops-of-compute-and-massive-improvements-over-hopper-h100

[3] https://www.thundercompute.com/blog/nvidia-b200-pricing

[4] https://getdeploying.com/gpus/nvidia-b200

[5] https://www.nvidia.com/en-us/data-center/dgx-b200/

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