NVIDIA H200 SXM

Also known as: H200 SXM5 · GH100 (HBM3e) · HGX H200 GPU

The H200 SXM is the memory-upgraded revision of Hopper: the same GH100 compute die as the H100, but with 141 GB of HBM3e across six memory stacks instead of 80 GB across five, and 4.8 TB/s of bandwidth instead of 3.35 TB/s. Compute throughput per GPU is identical to H100, so the H200 exists to fix the memory ceiling, not to add FLOPS: bigger KV caches, larger single-GPU model shards, and fewer GPUs needed to hold a given model. NVIDIA markets it as drop-in compatible with H100-qualified systems at the same 700 W envelope, which is why it is the common upgrade path for operators with an existing Hopper fleet who are not ready to requalify for Blackwell. It sits above H100 and below Blackwell (B200/B300) in NVIDIA's current lineup, positioned as the memory-bound-workload answer within Hopper rather than a new architecture.

141 GB HBM3e GPU memory · 4.8 TB/s bandwidth, six memory stacks	**3,958 TFLOPS** FP8 Tensor Core · with sparsity; same compute as H100	**+43%** Memory bandwidth vs H100 · 4.8 TB/s vs 3.35 TB/s
900 GB/s NVLink bandwidth · 4th-generation NVLink, per GPU	**700 W** Max TDP · configurable, same envelope as H100 SXM5	**SXM** Form factor · drop-in compatible with H100-qualified HGX baseboards

COMPUTE	01/05
FP64	34 TFLOPS
FP64 Tensor Core	67 TFLOPS
FP32	67 TFLOPS
TF32 Tensor Core	989 TFLOPS · with sparsity; 494.5 TFLOPS dense
BF16 / FP16 Tensor Core	1,979 TFLOPS · with sparsity; 989 TFLOPS dense
FP8 Tensor Core	3,958 TFLOPS · with sparsity; 1,979 TFLOPS dense
INT8 Tensor Core	3,958 TOPS · with sparsity
Streaming multiprocessors	132 SM · same GH100 die as H100: 16,896 CUDA cores, 528 4th-gen Tensor Cores

MEMORY	02/05
Capacity	141 GB HBM3e · nearly 2x H100's 80 GB
Bandwidth	4.8 TB/s · 1.4x H100's 3.35 TB/s
Stack configuration	6x 24 GB HBM3e · vs 5x 16 GB HBM3 on H100
Multi-Instance GPU	Up to 7 MIGs · 18 GB each

INTERCONNECT	03/05
NVLink	900 GB/s · 4th-generation, per GPU, unchanged from H100
PCIe	Gen5, 128 GB/s · host link
HGX 8-GPU baseboard fabric	4x NVSwitch · all-to-all 900 GB/s per GPU
Decoders	7 NVDEC, 7 JPEG · no NVENC

POWER & THERMAL	04/05
Max TDP	700 W · same envelope as H100 SXM5; NVIDIA markets drop-in compatibility
Power delivery	Via SXM socket · no external power cable on the module
Cooling	Air or liquid · HBM3e substrate routing is tighter than H100, but OEM-qualified boards handle either

PHYSICAL & PLATFORM	05/05
Silicon	GH100, TSMC 4N · identical compute die to H100, memory subsystem revised
Form factor	SXM module · not field-serviceable as a discrete card
Platform	HGX H200 (4 or 8 GPU) or DGX H200 · sold only inside a qualified system

PROCUREMENT

Typical lead time

Reported as short as 14 to 21 days per-card and 21 to 45 days for an 8-GPU node through some channels in Q1 to Q2 2026, but other resellers report 36 to 52+ weeks amid HBM3e supply constraints. Treat any single quoted lead time as channel-specific; confirm against current HBM3e allocation before committing to a delivery date.

Pricing

Reported street price $30,000 to $45,000 per GPU; an 8-GPU HGX H200 server reported in the $300,000 to $500,000 range, with several sources clustering around $308,000 to $315,000 for the node. Typically a 10 to 30 percent premium over H100 per GPU for the memory upgrade.

- HBM3e is the binding supply constraint, not the compute die: expect H200 allocation to track memory availability more than GPU wafer output.
- Sold only as part of an HGX baseboard or complete DGX/OEM server, same as H100; no standalone card purchase.
- Drop-in compatibility with existing H100-qualified racks (same 700 W envelope, same NVLink topology) is the main reason operators choose H200 over jumping straight to Blackwell.
- Secondary market for H200 is thin in 2026 compared with H100; most units in circulation are still under first-owner deployment.

FIELD NOTES

- Compute per GPU is identical to H100: the entire case for H200 is memory capacity and bandwidth. If your bottleneck is FLOPS, not memory, H200 will not move your throughput.
- 141 GB per GPU changes sharding math for large models: workloads that needed 2 H100s per shard for memory reasons often fit on 1 H200, which can net out cheaper per unit of served capacity despite the higher unit price.
- The bandwidth jump (4.8 TB/s vs 3.35 TB/s) matters most for memory-bound inference (large batch, long context, high KV-cache pressure) and less for compute-bound training steps; benchmark your actual workload before assuming H200 is worth the premium.
- 'Drop-in compatible' is a vendor claim about power and socket, not a guarantee: verify your specific HGX baseboard and cooling loop are qualified for HBM3e's tighter PCB routing tolerances before assuming a straight swap.
- Given the tighter and more volatile lead times reported for H200 versus H100, lock allocation early if the design depends on 141 GB per GPU; do not assume H200 supply behaves like the now-easing H100 market.

QUESTIONS WE GET ON THIS PART

How much does an H200 cost?

Reported street prices run roughly $30,000 to $45,000 per GPU, with an 8-GPU HGX H200 server commonly quoted in the $300,000 to $500,000 range. Prices vary significantly by channel, volume, and how tight HBM3e supply is at the time of quote.

H100 vs H200 for inference?

H200 wins for inference workloads with large KV caches, long context windows, or big batch sizes, because the extra 61 GB and 43 percent more bandwidth reduce the number of GPUs needed per model instance. For short-context, compute-bound inference, the two parts perform the same and H100 is usually the cheaper choice.

Is H200 a new architecture or just more memory on H100?

It is the same GH100 Hopper die and the same compute throughput as H100; the change is entirely in the memory subsystem, 141 GB of HBM3e at 4.8 TB/s instead of 80 GB of HBM3 at 3.35 TB/s.

Can I upgrade an existing H100 HGX server to H200?

NVIDIA markets H200 as drop-in compatible with H100-qualified systems at the same 700 W envelope, but this needs to be verified against your specific baseboard and cooling loop, since HBM3e imposes tighter PCB substrate routing requirements than HBM3.

REFERENCES

[1] https://www.nvidia.com/en-us/data-center/h200/

[2] https://www.gpu.fm/buying-guide

[3] https://intuitionlabs.ai/articles/nvidia-ai-gpu-pricing-guide

[4] https://www.thundercompute.com/blog/nvidia-h200-pricing

[5] https://www.spheron.network/blog/gpu-shortage-2026/

[6] https://www.spheron.network/blog/nvidia-h200-specs/

[7] https://www.spheron.network/blog/nvidia-rubin-vs-blackwell-vs-hopper/

[8] https://www.runpod.io/articles/guides/nvidia-h200-gpu

Latest version of this datasheet: https://gpusmith.com/hardware/gpus/nvidia-h200-sxm. Specifications compiled from the references above; verify against vendor documentation before purchase.