NVIDIA H100 SXM

Also known as: H100 SXM5 · GH100 · HGX H100 GPU

The H100 SXM is NVIDIA's Hopper-generation compute GPU in the socketed SXM5 form factor, sold only as a component of an HGX baseboard (4 or 8 GPU) inside partner or DGX systems, never as a standalone card. It is the workhorse part behind most of the Hopper-era training and inference fleet still in production: 80 GB of HBM3 at 3.35 TB/s, a fourth-generation Transformer Engine, and 900 GB/s of NVLink to the other GPUs on the baseboard. As of mid-2026 it is a prior-generation part: NVIDIA's shipping frontier is Blackwell Ultra (B300) and Rubin has entered production, so H100 buys now are almost always driven by compatibility with an existing Hopper fleet, software stack, or by price rather than by peak performance. It remains a rational choice when a workload fits in 80 GB and the buyer can get it cheaper per FLOP than a Blackwell or H200 seat.

80 GB HBM3 GPU memory · 3.35 TB/s bandwidth	**3,958 TFLOPS** FP8 Tensor Core · with sparsity; 1,979 TFLOPS dense	**1,979 TFLOPS** FP16 / BF16 Tensor Core · with sparsity; 989 TFLOPS dense
900 GB/s NVLink bandwidth · 4th-generation NVLink, per GPU	**700 W** Max TDP · configurable, SXM5	**SXM5** Form factor · HGX baseboard only, not sold as a standalone card

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 · 16,896 CUDA cores, 528 4th-gen Tensor Cores

MEMORY	02/05
Capacity	80 GB HBM3
Bandwidth	3.35 TB/s
Multi-Instance GPU	Up to 7 MIGs · 10 GB each

INTERCONNECT	03/05
NVLink	900 GB/s · 4th-generation, per GPU
PCIe	Gen5, 128 GB/s · host link, module itself is SXM not PCIe
HGX 8-GPU baseboard fabric	4x NVSwitch · 3rd-gen, all-to-all 900 GB/s per GPU
Decoders	7 NVDEC, 7 JPEG · no NVENC

POWER & THERMAL	04/05
Max TDP	700 W · configurable down to lower profiles by OEM
Power delivery	Via SXM5 socket · no external power cable on the module
Cooling	Air or liquid · OEM-dependent; liquid increasingly preferred at 700 W in dense racks

PHYSICAL & PLATFORM	05/05
Silicon	GH100, TSMC 4N · ~80 billion transistors
Form factor	SXM5 module · not field-serviceable as a discrete card
Platform	HGX H100 (4 or 8 GPU) or DGX H100 · sold only inside a qualified system

PROCUREMENT

Typical lead time

Highly channel-dependent: reported 2 to 12 weeks for in-stock allocations through OEM and distributor channels in Q1 to Q2 2026, versus 36 to 52 weeks reported by some resellers for large or custom-networked orders. Asia-Pacific channels report the shortest waits, US and EU resellers the longest.

Pricing

Reported street price $24,000 to $40,000 per GPU depending on volume and channel; an 8-GPU HGX H100 server reported in the $216,000 to $300,000 range depending on OEM and networking. Used/secondary-market SXM5 cards reported at $6,000 to $15,000, down sharply from the 2023 peak.

- Rarely sold as a bare module: buyers procure an HGX baseboard or a full DGX/OEM server, not a discrete card.
- Secondary market has grown as hyperscalers and neoclouds refresh into Blackwell; inspect hours, warranty status, and NVSwitch baseboard compatibility before buying used.
- Cloud on-demand H100 capacity is often easier to get than owned hardware right now; reserved pools at the major clouds skew toward existing customers.
- Most H100 SXM systems ship as 8-GPU HGX nodes with 4x NVSwitch; 4-GPU baseboards exist but are less common in new builds.

FIELD NOTES

- 80 GB per GPU is the real ceiling on this part: a 70B-parameter model at FP16 with KV cache headroom fits, but large MoE or long-context serving workloads will push you toward H200 or B200 for memory capacity, not raw compute.
- Sparsity numbers on the datasheet (3,958 TFLOPS FP8) assume structured 2:4 sparsity that most production models do not hit; budget compute against the dense figures (1,979 TFLOPS FP8) for realistic throughput planning.
- At 700 W per GPU and 8 GPUs per node, air cooling is workable but liquid cooling meaningfully lowers fan power and lets you run denser racks; factor this into total cost of ownership, not just the GPU price.
- If the workload needs more than 80 GB per GPU or you are buying new today, price out H200 SXM before committing to H100: the memory bandwidth and capacity jump often beats the price delta once you count fewer GPUs needed per model shard.
- Because H100 SXM is not sold standalone, the actual constraint on lead time is usually the HGX baseboard and chassis, not the silicon itself; ask vendors for baseboard and NVSwitch stock separately from GPU die allocation.

QUESTIONS WE GET ON THIS PART

Is the H100 still worth buying in 2026?

Only if it is cheaper per FLOP than H200 or a Blackwell part for your workload, or you need it to match an existing Hopper fleet. NVIDIA's current shipping frontier is Blackwell Ultra with Rubin now in production, so H100 is a value or compatibility buy, not a performance one.

How much does an H100 SXM cost?

Reported street prices run roughly $24,000 to $40,000 per GPU depending on channel and volume, with an 8-GPU HGX server in the $216,000 to $300,000 range. Used SXM5 cards on the secondary market have been reported as low as $6,000 to $15,000.

Can I buy a single H100 SXM card?

Not really. SXM5 is a socketed module that ships as part of an HGX baseboard or a complete DGX/OEM server; if you need a single discrete card, look at the PCIe H100 instead.

H100 vs H200: which should I pick?

Pick H100 when budget or fleet compatibility with existing Hopper systems dominates and 80 GB is enough for the model. Pick H200 when you need more memory capacity or bandwidth per GPU, since compute throughput on the two parts is otherwise identical.

REFERENCES

[1] https://www.nvidia.com/en-us/data-center/h100/

[2] https://www.thundercompute.com/blog/nvidia-h100-specs-full-guide

[3] https://www.hyperstack.cloud/technical-resources/performance-benchmarks/comparing-nvidia-h100-pcie-vs-sxm-performance-use-cases-and-more

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

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

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

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