



Supermicro GPU SuperServer AS -8125GS-TNHR (HGX H200)

8x H200

Hopper

rack

2024

OVERVIEW

15.8 FP8 PFLOPS Dense	1.1 TB Aggregate Memory	38.4 TB/s Aggregate Bandwidth	— Aggregate Power
---------------------------------	-----------------------------------	---	-----------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.3	34.0	—
FP32	0.5	67.0	—
TF32	4.0	494.5	—
FP16	7.9	989.5	—
BF16	7.9	989.5	—
FP8	15.8	1979.0	—
INT8	15.8	1979.0	—

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	H200 SXM 141GB
Count	8
Memory per GPU	141 GB
Bandwidth per GPU	4.8 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	700 W
FP8 Efficiency	—
Aggregate Bandwidth	38.4 TB/s

SYSTEM DETAILS

Vendor: Supermicro	Family: GPU SuperServer	Architecture: Hopper
Form Factor: rack	Release Year: 2024	Process Node: 4nm

INTERCONNECT

NVIDIA NVLink with NVSwitch between GPUs; PCIe 5.0 x16 CPU-to-GPU

NOTES

Supermicro's 8U AMD EPYC platform for the NVIDIA HGX H200 8-GPU baseboard, and the EPYC sibling of the Intel-based SYS-821GE-TNHR. Two SP5 sockets take AMD EPYC 9004 processors up to 128 cores and 256 threads at up to 400 W each, air cooled, with 24 DIMM slots supporting up to 6

TB of DDR5. GPUs connect to the CPUs over PCIe 5.0 x16 and to each other over NVLink with NVSwitch, and the chassis provides eight NICs for GPU-direct RDMA at a 1:1 ratio with the GPUs. The CSE-GP801TS chassis is 8U, weighs 102.1 kg gross, takes up to ten heavy-duty fans and six 3000 W Titanium supplies in 3+3 redundancy. Compute is identical to the H100 configuration because H100 and H200 are the same silicon at the same clocks; the difference is memory, 1.128 TB of HBM3e here against 0.64 TB. Supermicro publishes no FLOPS figure for this system; the compute shown is NVIDIA's for the HGX H200 baseboard it carries. Supermicro publishes no system power draw either, only the supply configuration, so total power is left blank.