Gigabyte G593-SD1-LAX3

8x H200 | Hopper | rack | 2024

OVERVIEW

15.8	1.1 TB	38.4 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	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: Gigabyte
Family: G593
Architecture: Hopper
Form Factor: rack
Release Year: 2024
Process Node: 4nm

INTERCONNECT

NVIDIA NVLink with NVSwitch, 900 GB/s GPU-to-GPU

NOTES

A 5U direct-liquid-cooled server carrying the NVIDIA HGX H200 baseboard with eight SXM GPUs, connected at 900 GB/s by NVLink and NVSwitch. Both CPUs and GPUs are liquid cooled, which is what lets Gigabyte fit the platform into 5U rather than the 8U an air-cooled equivalent needs. Two LGA 4677 sockets take 5th or 4th generation Intel Xeon Scalable processors or the Xeon CPU Max series, with 32 DIMM slots across eight DDR5 channels, eight Gen5 NVMe bays, twelve PCIe Gen5

x16 slots and BlueField-3 DPU compatibility. Power comes from four plus two redundant 3000 W Titanium supplies. It weighs 91 kg net and orders as 6NG593SD1DR000LBX3. Note that Gigabyte's page quotes "over 32 petaflops of FP8 deep learning compute" for an eight-way HGX H200; that is NVIDIA's own marketing figure for the baseboard rather than a Gigabyte measurement, and it is the sparse number. The dense figures shown here are NVIDIA's for the same baseboard, and they are what this site publishes throughout.