

HPE Cray XD670

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: HPE	Family: Cray	Architecture: Hopper
Form Factor: rack	Release Year: 2024	Process Node: 4nm

INTERCONNECT

NVIDIA NVLink with NVSwitch across the eight SXM modules

NOTES

A 5U server carrying eight NVIDIA H200 SXM GPUs at 700 W each with 141 GB of HBM3e apiece, driven by a single two-socket node of 5th generation Intel Xeon Scalable processors at up to 400 W, with 32 DDR5 DIMMs across eight channels per socket at up to 5,600 MHz. It is air cooled with a direct liquid cooling option. HPE has taken six first-place results with it in MLPerf Inference v5.1, including computer vision and LLM chat and text generation. Two things this row deliberately does

not claim: HPE describes the GPUs only as "8x NVIDIA H200 SXM" and never names the HGX baseboard, though eight H200 SXM at 700 W is precisely the HGX H200 8-GPU configuration, and the compute figures shown are NVIDIA's for that baseboard rather than any HPE claim; and HPE publishes component TDPs but no system power figure, so total power is left blank rather than summing 8 x 700 W and 2 x 400 W, which would count silicon only and understate the machine. Part numbers are also not recorded, since none appears on HPE's page.