NVIDIA DGX H200

8x H200 | Hopper | rack | 2024

OVERVIEW

15.8	1.1 TB	38.4 TB/s	8.5 kW
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.3	34.0	64.00
FP32	0.5	67.0	126.12
TF32	4.0	494.5	930.82
FP16	7.9	989.5	1862.59
BF16	7.9	989.5	1862.59
FP8	15.8	1979.0	3725.18
INT8	15.8	1979.0	3725.18

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	8.5 kW
TDP per GPU	700 W
FP8 Efficiency	3725.18 TFLOPS/W
Aggregate Bandwidth	38.4 TB/s

SYSTEM DETAILS

Vendor: NVIDIA
Family: DGX
Architecture: Hopper
Form Factor: rack
Release Year: 2024
Process Node: 4nm

INTERCONNECT

NVSwitch + NVLink

NOTES

Latest DGX system with H200 GPUs and 141GB HBM3e memory per GPU