

NVIDIA DGX B200

8x B200

Blackwell

rack

2024

OVERVIEW

36.0 FP8 PFLOPS Dense	1.4 TB Aggregate Memory	64.0 TB/s Aggregate Bandwidth	14.3 kW Aggregate Power
---	---	---	---

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.3	37.0	—
FP32	0.6	75.0	—
TF32	8.8	1100.0	—
FP16	18.0	2250.0	—
BF16	18.0	2250.0	—
FP8	36.0	4500.0	—
FP4	72.0	9000.0	—
INT8	36.0	4500.0	—

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

CONFIGURATION

Accelerator	B200 SXM 180GB
Count	8
Memory per GPU	180 GB
Bandwidth per GPU	7.7 TB/s

POWER & EFFICIENCY

Aggregate Power	14.3 kW
TDP per GPU	1000 W
FP8 Efficiency	2.52 TFLOPS/W
Aggregate Bandwidth	64.0 TB/s

SYSTEM DETAILS

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

INTERCONNECT

NVLink 5.0 / 2x NVSwitch, 14.4 TB/s aggregate

NOTES

Unveiled at GTC March 2024. 10 RU air-cooled chassis with 2x Intel Xeon Platinum 8570 (112 cores) and 2 TB system memory, configurable to 4 TB. Networking is 4x OSFP serving 8x single-port ConnectX-7 plus 2x dual-port QSFP112 BlueField-3 DPUs, each up to 400 Gb/s. Same eight B200

SXM parts as HGX B200; the 14.3 kW figure is the whole system, against 10.2 kW for the bare baseboard.