

NVIDIA GB200 NVL32

32x B200

Blackwell

rack

2025

OVERVIEW

144 FP8 PFLOPS Dense	— Aggregate Memory	— Aggregate Bandwidth	53.3 kW Aggregate Power
--------------------------------	-----------------------	--------------------------	-----------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	1.2	37.0	27.00
FP32	2.4	75.0	54.00
TF32	35.2	1100.0	1501.00
FP16	72.0	2250.0	2701.00
BF16	72.0	2250.0	2701.00
FP8	144.0	4500.0	5403.00
FP4	288.0	9000.0	10806.00
INT8	144.0	4500.0	5403.00

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

CONFIGURATION

Accelerator	B200 SXM 180GB
Count	32
Memory per GPU	180 GB
Bandwidth per GPU	7.7 TB/s

POWER & EFFICIENCY

Aggregate Power	53.3 kW
TDP per GPU	1000 W
FP8 Efficiency	5403.00 TFLOPS/W
Aggregate Bandwidth	—

SYSTEM DETAILS

Vendor: NVIDIA	Family: NVL	Architecture: Blackwell
Form Factor: rack	Release Year: 2025	Process Node: 4nm

INTERCONNECT

NVLink

NOTES

16 GB200 superchips (32 B200 GPUs + 16 Grace CPUs) in a rack-scale system

