

NVIDIA GB200 NVL72

72x B200

Blackwell

rack

2025

OVERVIEW

360 FP8 PFLOPS Dense	— Aggregate Memory	— Aggregate Bandwidth	120.0 kW Aggregate Power
--------------------------------	-----------------------	--------------------------	------------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	2.7	37.0	27.00
FP32	5.4	75.0	54.00
TF32	79.2	1100.0	1500.00
FP16	180.0	2500.0	2700.00
BF16	180.0	2500.0	2700.00
FP8	360.0	5000.0	5400.00
FP4	720.0	10000.0	10800.00
INT8	324.0	4500.0	5400.00

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

CONFIGURATION

Accelerator	B200 SXM 180GB
Count	72
Memory per GPU	180 GB
Bandwidth per GPU	7.7 TB/s

POWER & EFFICIENCY

Aggregate Power	120.0 kW
TDP per GPU	1000 W
FP8 Efficiency	5400.00 TFLOPS/W
Aggregate Bandwidth	—

SYSTEM DETAILS

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

INTERCONNECT

NVLink

NOTES

36 GB200 superchips (72 B200 GPUs + 36 Grace CPUs) in a liquid-cooled rack-scale system

