

NVIDIA GB200 NVL576

576x B200

2025

OVERVIEW

2,880 FP8 PFLOPS Dense	— Aggregate Memory	— Aggregate Bandwidth	960.0 kW Aggregate Power
----------------------------------	-----------------------	--------------------------	------------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	21.3	37.0	27.00
FP32	43.2	75.0	54.00
TF32	633.6	1100.0	1500.00
FP16	1440.0	2500.0	2700.00
BF16	1440.0	2500.0	2700.00
FP8	2880.0	5000.0	5400.00
FP4	5760.0	10000.0	10800.00
INT8	2592.0	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	576
Memory per GPU	180 GB
Bandwidth per GPU	7.7 TB/s

POWER & EFFICIENCY

Aggregate Power	960.0 kW
TDP per GPU	1000 W
FP8 Efficiency	5400.00 TFLOPS/W
Aggregate Bandwidth	—

SYSTEM DETAILS

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