

NVIDIA GeForce RTX 5080 16GB vs NVIDIA GB200 NVL72 GPU 186GB

Side by side datacenter GPU specifications from the Flopper.io database.

OVERVIEW

	NVIDIA GeForce RTX 5080 16GB	NVIDIA GB200 NVL72 GPU 186GB
FP32 TFLOPS	56.3	80.0
VRAM	16 GB	186 GB
Memory Bandwidth	960 GB/s	8.0 TB/s
TDP	360 W	—

PERFORMANCE BY PRECISION

Precision	NVIDIA GeForce RTX 5080 16GB peak TFLOPS (dense)	NVIDIA GB200 NVL72 GPU 186GB peak TFLOPS (dense)
FP64	0.9	40.0
FP32	56.3	80.0
TF32	112.6	1250.0
BF16	225.1	2500.0
FP16	225.1	2500.0
FP8	450.2	5000.0
FP6	—	5000.0
FP4	900.5	10000.0
INT8	450.2	5000.0

SPECIFICATIONS

	NVIDIA GeForce RTX 5080 16GB	NVIDIA GB200 NVL72 GPU 186GB
Architecture	Blackwell	Blackwell
Process Node	4nm	4nm
Launch Year	2025	2024
Form Factor	PCIe	SXM
Memory Type	GDDR7	HBM3e

	NVIDIA GeForce RTX 5080 16GB	NVIDIA GB200 NVL72 GPU 186GB
Interface Width	256-bit	—
Interconnect	—	NVLink
TDP	360 W	—
FP32 Efficiency	0.16 TFLOPS/W	—