

NVIDIA® H200

SPECIFICATIONS

Product	NVIDIA H200 Tensor Core GPU
Architecture	Hopper
Form Factor	H200 SXM
Tensor Cores Gen 3	432
FP64	34 TFLOPS
Peak FP64 Tensor Core	67 TFLOPS
Peak FP32	67 TFLOPS
TF32 Tensor Core	989 TFLOPS
Peak FP16 Tensor Core	1979 TFLOPS
Peak INT8 Tensor Core	3958 TFLOPS
INT4 Tensor Core	1248 TOPS 2496 TOPS Sparsity
NVLink Interconnect	900GB/s
GPU Memory	141GB
Multi-Instance GPUs	Up to 7 MIGs @16.5GB each
Memory Interface	5120-bit
Memory Bandwidth	4.8TB/s
System Interface	PCIe 4.0 x16
Max Thermal Design Power	Up to 700W (configurable)
vGPU Support	NVIDIA AI Enterprise