



NVIDIA A16 16GB

Ampere PCIe 2021

OVERVIEW

4.5 FP32 TFLOPS	16 GB Memory GDDR6	200 GB/s Memory Bandwidth	— TDP
--------------------	-----------------------	------------------------------	----------

PERFORMANCE METRICS

Precision	Peak (dense)	2:4 Sparse	Bits	Type
FP64	—	N/A	—	Standard
FP32	4.5	N/A	32	Standard
TF32	9.0	18.0	32	Standard
BF16	—	not published	—	Standard
FP16	17.9	35.9	16	Standard
FP8	—	not published	—	Standard
FP6	—	not published	—	Standard
FP4	—	not published	—	Standard
INT8	35.9	71.8	8	Standard

MEMORY SPECIFICATIONS

Capacity	16 GB
Type	GDDR6
Bandwidth	200 GB/s
Interface Width	—

POWER & EFFICIENCY

TDP	—
Max power	—
FP32 dense TFLOPS per W of TDP	—
FP16 dense TFLOPS per W of TDP	—
Per W of TDP figures divide dense peak by accelerator TDP. Board power only: excludes host CPUs, networking, cooling and facility overhead.	

HARDWARE DETAILS

Vendor: NVIDIA	Architecture: Ampere	Process Node: —
Form Factor: PCIe	Launch Year: 2021	Variant: 16GB

COMPUTE ENGINE

CUDA Cores: 1,280	Tensor Cores: 40 (3rd Gen)	SMs: 10
-------------------	----------------------------	---------

INTERCONNECT

PCIe: 4.0 x16

ENTERPRISE FEATURES

ECC: Yes

MIG: No

Sparsity: Yes