

NVIDIA A2 16GB

Ampere

PCIe

2021

OVERVIEW

4.5 FP32 TFLOPS	16 GB Memory GDDR6	200 GB/s Memory Bandwidth	60 W 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	18.0	36.0	16	Standard
FP16	18.0	36.0	16	Standard
FP8	—	not published	—	Standard
FP6	—	not published	—	Standard
FP4	—	not published	—	Standard
INT8	36.0	72.0	8	Standard
INT4	72.0	144.0	4	Standard

MEMORY SPECIFICATIONS

Capacity	16 GB
Type	GDDR6
Bandwidth	200 GB/s
Interface Width	128-bit

POWER & EFFICIENCY

TDP	60 W
Max power	60 W
FP32 dense TFLOPS per W of TDP	0.07 TFLOPS/W
FP16 dense TFLOPS per W of TDP	0.30 TFLOPS/W

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

Base Clock: 1.44 GHz

Boost Clock: 1.77 GHz

INTERCONNECT

PCIe: 4.0 x8

ENTERPRISE FEATURES

ECC: Yes

Sparsity: Yes