



NVIDIA DGX A100 640GB

8x A100 Ampere rack 2020

OVERVIEW

2.50 BF16 PFLOPS Dense	0.6 TB Aggregate Memory	16.3 TB/s Aggregate Bandwidth	6.5 kW Aggregate Power
----------------------------------	-----------------------------------	---	----------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.1	9.7	11.94
FP32	0.2	19.5	24.00
TF32	1.2	156.0	192.00
FP16	2.5	312.0	384.00
BF16	2.5	312.0	384.00
INT8	5.0	624.0	768.00

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	A100 SXM4 80GB
Count	8
Memory per GPU	80 GB
Bandwidth per GPU	2.0 TB/s

POWER & EFFICIENCY

Aggregate Power	6.5 kW
TDP per GPU	400 W
BF16 Efficiency	384.00 TFLOPS/W
Aggregate Bandwidth	16.3 TB/s

SYSTEM DETAILS

Vendor: NVIDIA	Family: DGX	Architecture: Ampere
Form Factor: rack	Release Year: 2020	Process Node: 7nm

INTERCONNECT

NVSwitch + NVLink

NOTES

DGX A100 with 640GB total GPU memory