

NVIDIA HGX A100

8x A100

Ampere

baseboard

2020

OVERVIEW

2.50 BF16 PFLOPS Dense	0.6 TB Aggregate Memory	16.3 TB/s Aggregate Bandwidth	6.0 kW Aggregate Power
----------------------------------	-----------------------------------	---	----------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.1	9.7	12.93
FP32	0.2	19.5	26.00
TF32	1.2	156.0	208.00
FP16	2.5	312.0	416.00
BF16	2.5	312.0	416.00
INT8	5.0	624.0	832.00

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	6.0 kW
TDP per GPU	400 W
BF16 Efficiency	416.00 TFLOPS/W
Aggregate Bandwidth	16.3 TB/s

SYSTEM DETAILS

Vendor: NVIDIA	Family: HGX	Architecture: Ampere
Form Factor: baseboard	Release Year: 2020	Process Node: 7nm

INTERCONNECT

NVSwitch + NVLink

NOTES

HGX A100 platform for cloud providers