

NVIDIA HGX H100

4x H100

Hopper

baseboard

2022

OVERVIEW

7.92 FP8 PFLOPS Dense	0.3 TB Aggregate Memory	13.4 TB/s Aggregate Bandwidth	4.2 kW Aggregate Power
---	---	---	--

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.1	34.0	32.38
FP32	0.3	67.0	63.81
TF32	2.0	494.5	941.90
FP16	4.0	989.5	1884.76
BF16	4.0	989.5	1884.76
FP8	7.9	1979.0	3769.52
INT8	7.9	1979.0	3769.52

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

CONFIGURATION

Accelerator	H100 SXM5 80GB
Count	4
Memory per GPU	80 GB
Bandwidth per GPU	3.4 TB/s

POWER & EFFICIENCY

Aggregate Power	4.2 kW
TDP per GPU	700 W
FP8 Efficiency	3769.52 TFLOPS/W
Aggregate Bandwidth	13.4 TB/s

SYSTEM DETAILS

Vendor: NVIDIA	Family: HGX	Architecture: Hopper
Form Factor: baseboard	Release Year: 2022	Process Node: 4nm

INTERCONNECT

NVLink

NOTES

HGX H100 4-GPU configuration for smaller deployments