Tenstorrent Galaxy Blackhole

32x Blackhole | Blackhole | rack

OVERVIEW

23.0	1.0 TB	16.0 TB/s	12.0 kW
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP8	23.0	718.8	–

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

CONFIGURATION

Accelerator	Blackhole p150a
Count	32
Memory per GPU	32 GB
Bandwidth per GPU	512 GB/s

POWER & EFFICIENCY

Aggregate Power	12.0 kW
TDP per GPU	300 W
FP8 Efficiency	1.92 TFLOPS/W
Aggregate Bandwidth	16.0 TB/s

SYSTEM DETAILS

Vendor: Tenstorrent
Family: Galaxy
Architecture: Blackhole
Form Factor: rack
Release Year: —
Process Node: —

INTERCONNECT

10 x 400 GbE per ASIC for 32 TB/s aggregate; up to 56 x 800 GbE QSFP-DD ports for 11.2 TB/s

NOTES

A 6U air-cooled chassis holding 32 Blackhole ASICs and rated by Tenstorrent at 23 PFLOPS of Block FP8, with 6.2 GB of on-die SRAM at 2.9 PB/s, 1 TB of GDDR6 at 16 TB/s, and 32 TB/s of aggregate Ethernet from ten 400 GbE links per ASIC plus up to 56 external 800 GbE QSFP-DD ports carrying 11.2 TB/s. It draws 8 to 10 kW on average and 12 kW at maximum, and is hosted by a single AMD EPYC 9004 with up to 32 cores and up to 576 GB of DDR5. Note that the Galaxy does not use the retail p150 bin: its figures work out at 718.75 TFLOPS and 193.75 MB of SRAM per ASIC, against 664 TFLOPS and 180 MB on the p150 card, so the ASICs here are a higher bin of the same silicon and Tenstorrent does not publish their Tensix core count. The compute figure recorded is Tenstorrent's own 23 PFLOPS for the machine, not 32 times the card, which would understate it by eight percent. Tenstorrent labels nothing dense or sparse and publishes no sparse rate.