Google TPU v5p Pod

8960x TPU v5p | TPU v5p | pod | 2023

OVERVIEW

4,113	851.2 TB	24774.4 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	4112.6	459.0	–
FP8	4112.6	459.0	–

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

CONFIGURATION

Accelerator	TPU v5p 95GB
Count	8960
Memory per GPU	95 GB
Bandwidth per GPU	2.8 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	450 W
FP8 Efficiency	—
Aggregate Bandwidth	24774.4 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: TPU v5p
Form Factor: pod
Release Year: 2023
Process Node: 5nm

INTERCONNECT

3D torus ICI, 1,200 GBps bidirectional per chip, 50 Gbps data center network per chip

NOTES

A full TPU v5p pod is 8,960 chips in a 3D torus, the largest TPU pod Google built before Ironwood. Per chip Google publishes 459 TFLOPS bf16 and 459 TFLOPS fp8 as two separate rows carrying the same number: v5p gets no fp8 throughput advantage over bf16. Across the pod that is 4,112.6 PFLOPS at either precision. Memory is 8,960 x 95 GiB HBM at 2,765 GBps per chip, with 1,200 GBps of bidirectional inter-chip interconnect and 50 Gbps of data center network per chip, and two TensorCores per chip. The table also lists four SparseCores per chip; those are embedding-lookup dataflow processors and have nothing to do with weight sparsity, so no figure here is a sparse figure. Google publishes no pod-level compute or power figure; the compute above is chip count times published per-chip peak.