Google TPU v5e Pod

256x TPU v5e | TPU v5e | pod | 2023

OVERVIEW

50.4	4.1 TB	204.8 TB/s	—
BF16 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	50.4	197.0	–
INT8	100.6	393.0	–

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

CONFIGURATION

Accelerator	TPU v5e 16GB
Count	256
Memory per GPU	16 GB
Bandwidth per GPU	800 GB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
BF16 Efficiency	—
Aggregate Bandwidth	204.8 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: TPU v5e
Form Factor: pod
Release Year: 2023
Process Node: —

INTERCONNECT

2D torus ICI, 400 GBps bidirectional per chip

NOTES

A full TPU v5e pod is 256 chips in a 2D torus. Per chip Google publishes 197 TFLOPS bf16 and 393 TOPS int8, giving 50.4 PFLOPS and 100.6 POPS across the pod. Memory is 256 x 16 GB HBM at 800 GiBps per chip, with 400 GBps of bidirectional inter-chip interconnect per chip. v5e is Google's cost-and-efficiency oriented generation, which is why its pod is a thirty-fifth the size of the v5p pod launched alongside it. Google publishes no pod-level compute or power figure; the compute above is chip count times published per-chip peak, the convention Google's own published v4 and Ironwood pod figures both confirm. No sparse figures are published, so all figures are dense.