Google TPU v6e Pod

256x TPU v6e | TPU v6e | pod | 2024

OVERVIEW

235	8.2 TB	419.3 TB/s	—
BF16 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	235.0	918.0	–
INT8	470.0	1836.0	–

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

CONFIGURATION

Accelerator	TPU v6e 32GB
Count	256
Memory per GPU	32 GB
Bandwidth per GPU	1.6 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
BF16 Efficiency	—
Aggregate Bandwidth	419.3 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: TPU v6e
Form Factor: pod
Release Year: 2024
Process Node: —

INTERCONNECT

2D torus ICI, 800 GBps bidirectional per chip

NOTES

A full TPU v6e (Trillium) pod is 256 chips in a 2D torus, the largest slice being a 16x16 topology across 64 VMs. Per chip Google publishes 918 TFLOPS bf16 and 1,836 TOPS int8, giving 235.0 PFLOPS and 470.0 POPS across the pod: a 4.7x per-chip gain over v5e at the same pod size. Memory is 256 x 32 GB HBM at 1,638 GBps per chip, with 800 GBps of bidirectional inter-chip interconnect per chip. v6e is the first generation to widen the matrix unit to 256x256 multiply-accumulators, from 128x128 on every prior TPU. Google publishes no pod-level compute or power figure; the compute above is chip count times published per-chip peak. No sparse figures are published, so all figures are dense.