Google TPU v4 Pod

4096x TPU v4 | TPU v4 | pod | 2020

OVERVIEW

1,126	131.1 TB	4915.2 TB/s	—
BF16 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	1126.4	275.0	–
INT8	1126.4	275.0	–

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

CONFIGURATION

Accelerator	TPU v4 32GB
Count	4096
Memory per GPU	32 GB
Bandwidth per GPU	1.2 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
BF16 Efficiency	—
Aggregate Bandwidth	4915.2 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: TPU v4
Form Factor: pod
Release Year: 2020
Process Node: —

INTERCONNECT

3D mesh ICI, 1.1 PB/s all-reduce bandwidth per pod, 24 TB/s bisection bandwidth

NOTES

A full TPU v4 pod is 4,096 chips in a 3D mesh. Google publishes a single peak compute figure of 275 teraflops per chip covering both bf16 and int8 rather than separate rows, giving 1,126.4 PFLOPS across the pod; Google prints this rounded as "1.1 exaflops (bf16 or int8)". Also published per pod: 1.1 PB/s all-reduce bandwidth and 24 TB/s bisection bandwidth. Memory is 4,096 x 32 GiB HBM2 at 1,200 GBps per chip. Google measures per-chip power at 90 W minimum, 170 W mean and 192 W maximum, but publishes no pod-level power, so total system power is left blank here rather than extrapolated from silicon alone. Google publishes no sparse figures for any TPU, so all figures are dense.