Google TPU 8i Pod

1024x TPU 8i | TPU 8 | pod | 2026

OVERVIEW

10,342	294.9 TB	8807.4 TB/s	—
FP4 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP4	10342.4	10100.0	–

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

CONFIGURATION

Accelerator	TPU 8i
Count	1024
Memory per GPU	288 GB
Bandwidth per GPU	8.6 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP4 Efficiency	—
Aggregate Bandwidth	8807.4 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: TPU 8
Form Factor: pod
Release Year: 2026
Process Node: —

INTERCONNECT

Boardfly topology, maximum seven-hop ICI network diameter

NOTES

Google's eighth-generation inference pod, built on a Boardfly topology rather than a torus, with a maximum seven-hop ICI network diameter. Google describes joining up to 1,152 TPU 8i chips together with up to 1,024 of them active; the figures here use the 1,024 active count, because crediting the pod with throughput from chips Google does not call active would overstate it. At Google's published 10.1 PFLOPS of FP4 per chip that is 10.34 EFLOPS, with 294.9 TB of HBM and 8.8 PB/s of aggregate memory bandwidth, plus 393 GB of on-chip Vmem SRAM. Google publishes no pod-level compute figure; the total is chip count times its published per-chip peak, the same convention its own v4 and Ironwood pod figures confirm. FP4 is the only precision stated and no sparse figure exists. Announced at Google Cloud Next in April 2026 and not yet generally available.