Google TPU 8t Superpod

9600x TPU 8t | TPU 8 | pod | 2026

OVERVIEW

120,960	2.1 PB	62668.8 TB/s	—
FP4 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP4	120960.0	12600.0	–

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

CONFIGURATION

Accelerator	TPU 8t
Count	9600
Memory per GPU	216 GB
Bandwidth per GPU	6.5 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP4 Efficiency	—
Aggregate Bandwidth	62668.8 TB/s

SYSTEM DETAILS

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

INTERCONNECT

3D torus ICI with 47 Pb/s non-blocking bisectional bandwidth; Virgo network links over 134,000 chips

NOTES

Google's eighth-generation training superpod: 9,600 TPU 8t chips on a 3D torus, with 47 petabits per second of non-blocking bisectional bandwidth, and a Virgo network able to link over 134,000 chips across superpods. At Google's published 12.6 PFLOPS of FP4 per chip that is 120.96 EFLOPS across the pod, with 2.07 PB of HBM3e and 62.7 PB/s of aggregate memory bandwidth. Google publishes no pod-level compute figure for this generation; the total here is chip count times its published per-chip peak, which is Google's own arithmetic — on the two generations where it publishes both, v4 and Ironwood, the pod figure equals exactly that product. FP4 is the only precision Google states, and it publishes no sparse figure for any TPU, so this is dense. Announced at Google Cloud Next in April 2026 and not yet generally available.