Google TPU7x (Ironwood) Pod

9216x TPU v7 | Ironwood | pod | 2025

OVERVIEW

42,523	1.8 PB	68014.1 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	21261.3	2307.0	–
FP8	42522.6	4614.0	–

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

CONFIGURATION

Accelerator	TPU v7 192GB
Count	9216
Memory per GPU	192 GB
Bandwidth per GPU	7.4 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	68014.1 TB/s

SYSTEM DETAILS

Vendor: Google
Family: TPU
Architecture: Ironwood
Form Factor: pod
Release Year: 2025
Process Node: —

INTERCONNECT

3D torus ICI, 1,200 GBps bidirectional per chip

NOTES

A full TPU7x (Ironwood) pod is 9,216 chips in a 3D torus, the largest TPU pod Google has built. Per chip Google publishes 2,307 TFLOPS bf16 and 4,614 TFLOPS fp8, giving 21.26 EFLOPS and 42.52 EFLOPS across the pod; the fp8 total matches the 42.5 exaflops Google has stated publicly for an Ironwood pod, which independently confirms that chip count times per-chip peak is Google's own arithmetic. Memory is 9,216 x 192 GiB HBM at 7,380 GBps per chip, for 1.77 PB of HBM and 68 PB/s of aggregate memory bandwidth. Supported slice shapes go up to 8x16x16, or 2,048 chips, within the pod. No int8 figure is published. The table lists four SparseCores per chip; those are embedding-lookup dataflow processors, not weight sparsity. Google publishes no sparse figures and no pod-level power.