AWS EC2 trn2.48xlarge

16x Trainium2 | NeuronCore-v3 | rack | 2024

OVERVIEW

20.8	1.5 TB	46.4 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP32	2.9	181.3	–
TF32	10.7	668.8	–
FP16	10.7	668.8	–
BF16	10.7	668.8	–
FP8	20.8	1300.0	–

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

CONFIGURATION

Accelerator	Trainium2
Count	16
Memory per GPU	96 GB
Bandwidth per GPU	2.9 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	46.4 TB/s

SYSTEM DETAILS

Vendor: AWS
Family: EC2 Instance
Architecture: NeuronCore-v3
Form Factor: rack
Release Year: 2024
Process Node: —

INTERCONNECT

NeuronLink-v3 at 1,024 GB/s per chip within the instance; EFAv3 networking at 3,200 Gbps

NOTES

The trn2.48xlarge is AWS's single-node Trainium2 instance: sixteen Trainium2 chips joined by NeuronLink-v3 at 1,024 GB/s per chip, with 1,536 GiB of device memory at 46.4 TB/s, 192 vCPUs, 2 TB of host memory and 3,200 Gbps of EFAv3 networking. AWS rates it at 20.8 PFLOPS of dense FP8, 10.7 PFLOPS across dense FP16, BF16 and TF32, and 2.9 PFLOPS of dense FP32, with a separate sparse figure of 41 PFLOPS covering FP8, FP16, BF16 and TF32 but not FP32. Note that the sparse figure is larger than the dense FP8 one, so the two must not be confused: dense FP16 is 10.7 PFLOPS, not 41. AWS's sparse mode runs at the chip's low-precision rate whatever format is fed in, which is why 41 is 1.97x the dense FP8 rate but 3.84x the dense BF16 rate; it is not 2:4 structured

sparsity. AWS publishes no power figure for this instance or for the Trainium2 chip, so total system power is left blank rather than estimated. The trn2u.48xlarge variant carries the same specification and is the instance from which UltraServers are composed.