AWS Trn2 UltraServer

64x Trainium2 | NeuronCore-v3 | rack | 2024

OVERVIEW

83.2	6.1 TB	185.6 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP32	11.6	181.3	–
TF32	42.8	668.8	–
FP16	42.8	668.8	–
BF16	42.8	668.8	–
FP8	83.2	1300.0	–

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

CONFIGURATION

Accelerator	Trainium2
Count	64
Memory per GPU	96 GB
Bandwidth per GPU	2.9 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	185.6 TB/s

SYSTEM DETAILS

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

INTERCONNECT

NeuronLink-v3 at 1,024 GB/s per chip within each instance and 256 GB/s per chip between instances; EFAv3 networking at 3,200 Gbps

NOTES

The Trn2 UltraServer joins four trn2u.48xlarge instances into a single 64-chip Trainium2 machine, connected by NeuronLink-v3 at 1,024 GB/s per chip within an instance and 256 GB/s per chip between them. It carries 6,144 GiB of device memory at 185.6 TB/s, 768 vCPUs and 8 TB of host memory. AWS rates it at 83.2 PFLOPS of dense FP8, 42.8 PFLOPS across dense FP16, BF16 and TF32, and 11.6 PFLOPS of dense FP32, with a separate sparse figure of 164 PFLOPS covering FP8, FP16, BF16 and TF32 but not FP32. The sparse figure exceeds the dense FP8 one, so the two must not be conflated: dense FP16 is 42.8 PFLOPS, not 164. AWS's sparse mode runs at the chip's low-precision

rate regardless of input format, making the sparse figure 1.97x dense FP8 but 3.84x dense BF16, so it is not 2:4 structured sparsity. Every figure here is AWS's own published aggregate and each one reconciles exactly against the Trainium2 chip specification. AWS publishes no power figure for the UltraServer or the chip.
