AWS Trn3 UltraServer

144x Trainium3 | NeuronCore-v4 | rack | 2025

OVERVIEW

362	20.7 TB	705.6 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP32	26.4	183.0	–
TF32	96.6	671.0	–
FP16	96.6	671.0	–
BF16	96.6	671.0	–
FP8	362.4	2517.0	–
FP4	362.4	2517.0	–

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

CONFIGURATION

Accelerator	Trainium3
Count	144
Memory per GPU	144 GB
Bandwidth per GPU	4.9 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	705.6 TB/s

SYSTEM DETAILS

Vendor: AWS
Family: UltraServer
Architecture: NeuronCore-v4
Form Factor: rack
Release Year: 2025
Process Node: 3nm

INTERCONNECT

NeuronSwitch-v1 all-to-all fabric over NeuronLink-v4, 2 TB/s per chip

NOTES

The Trn3 UltraServer is AWS's rack-scale Trainium3 machine, generally available since 2 December 2025 and reaching 144 chips joined by NeuronSwitch-v1, an all-to-all fabric built on NeuronLink-v4. AWS publishes up to 362 FP8 PFLOPS, up to 20.7 TB of HBM3e and 706 TB/s of aggregate memory bandwidth, and states over twice the chip density per rack of Trn2. The figures stored here are chip count times AWS's published per-chip specifications, which is AWS's own arithmetic: all three of its published aggregates reconcile exactly against the Trainium3 chip at 2,517 MXFP8 TFLOPS, 144 GB

of HBM3e and 4.9 TB/s. AWS labels dense and sparse itself; the sparse figures apply to FP16, BF16 and TF32 but explicitly not to FP8, and they are 3.75x the dense rate rather than twice it, because AWS's sparse mode runs at the chip's low-precision rate whatever format is fed in. AWS publishes no power figure for the UltraServer or for the chip, so total system power is left blank rather than estimated. One inconsistency in AWS's own material is worth noting: this machine's page gives NeuronLink-v4 as 2 TB/s per chip while the Trainium3 architecture documentation gives 2.56 TB/s per device. AWS describes only this one Trn3 UltraServer configuration; there is no 64-chip variant and no Gen1/Gen2 naming in its documentation.