# AWS Trainium2

NeuronCore-v3 2024

## OVERVIEW

| 181.0 | 96 GB | 2.9 TB/s | — |
|---|---|---|---|
| FP32 TFLOPS | VRAM | Memory Bandwidth | TDP |

## PERFORMANCE METRICS

| Precision | Peak TFLOPS | Bits | Type |
|---|---|---|---|
| FP8 | 1299.0 | 8 | Standard |
| FP16 | 667.0 | 16 | Standard |
| BF16 | 667.0 | 16 | Standard |
| TF32 | 667.0 | 32 | Standard |
| FP32 | 181.0 | 32 | Standard |

## MEMORY SPECIFICATIONS

| | |
|---|---|
| Capacity | 96 GB |
| Type | — |
| Bandwidth | 2.9 TB/s |
| Interface Width | — |

## POWER & EFFICIENCY

| | |
|---|---|
| TDP | — |
| Est. Max Power | — |
| FP32 Efficiency | — |
| FP16 Efficiency | — |

## HARDWARE DETAILS

Vendor: AWS
Architecture: NeuronCore-v3
Process Node: —
Form Factor: —
Launch Year: 2024

## COMPUTE ENGINE

NeuronCore-v3: 8

## INTERCONNECT

GPU-to-GPU: NeuronLink-v3
Bandwidth: 1.3 TB/s

## ENTERPRISE FEATURES

Sparsity: Yes