AWS Inferentia2 vs AWS Trainium3

Side by side datacenter GPU specifications from the Flopper.io database.

OVERVIEW

	AWS Inferentia2	AWS Trainium3
FP32 TFLOPS	**47.5**	**183.0**
Memory	**32 GB**	**144 GB**
Memory Bandwidth	**820 GB/s**	**4.9 TB/s**
TDP	—	—

PERFORMANCE BY PRECISION

Precision	AWS Inferentia2 peak TFLOPS (dense)	AWS Trainium3 peak TFLOPS (dense)
FP64	—	—
FP32	47.5	183.0
TF32	190.0	671.0
BF16	190.0	671.0
FP16	190.0	671.0
FP8	190.0	2517.0
FP6	—	—
FP4	—	2517.0
INT8	380.0	—

SPECIFICATIONS

	AWS Inferentia2	AWS Trainium3
Architecture	NeuronCore-v2	NeuronCore-v4
Process Node	—	3nm
Launch Year	2022	2025
Form Factor	—	—
Memory Type	HBM	HBM3e
Interface Width	—	—
Interconnect	NeuronLink-v2	NeuronLink-v4

	AWS Inferentia2	AWS Trainium3
TDP	–	–
FP32 Efficiency	–	–