AWS Inferentia vs AWS Trainium3

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

OVERVIEW

	AWS Inferentia	AWS Trainium3
FP32 TFLOPS	–	**183.0**
Memory	**8 GB**	**144 GB**
Memory Bandwidth	**50 GB/s**	**4.9 TB/s**
TDP	–	–

PERFORMANCE BY PRECISION

Precision	AWS Inferentia peak TFLOPS (dense)	AWS Trainium3 peak TFLOPS (dense)
FP64	–	–
FP32	–	183.0
TF32	–	671.0
BF16	64.0	671.0
FP16	64.0	671.0
FP8	–	2517.0
FP6	–	–
FP4	–	2517.0
INT8	128.0	–

SPECIFICATIONS

	AWS Inferentia	AWS Trainium3
Architecture	NeuronCore-v1	NeuronCore-v4
Process Node	–	3nm
Launch Year	2018	2025
Form Factor	–	–
Memory Type	DDR4	HBM3e
Interface Width	–	–
Interconnect	NeuronLink-v1	NeuronLink-v4

	AWS Inferentia	AWS Trainium3
TDP	–	–
FP32 Efficiency	–	–