

AWS Inferentia2 vs AWS Trainium

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

OVERVIEW

	AWS Inferentia2	AWS Trainium
FP32 TFLOPS	47.5	47.5
Memory	32 GB	32 GB
Memory Bandwidth	820 GB/s	820 GB/s
TDP	—	—

PERFORMANCE BY PRECISION

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

SPECIFICATIONS

	AWS Inferentia2	AWS Trainium
Architecture	NeuronCore-v2	NeuronCore-v2
Process Node	—	—
Launch Year	2022	2020
Form Factor	—	—
Memory Type	HBM	HBM
Interface Width	—	—
Interconnect	NeuronLink-v2	NeuronLink-v2

	AWS Inferentia2	AWS Trainium
TDP	—	—
FP32 Efficiency	—	—