AWS Inferentia2 vs AWS Trainium2

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

OVERVIEW

	AWS Inferentia2	AWS Trainium2
FP32 TFLOPS	**47.5**	**181.0**
Memory	**32 GB**	**96 GB**
Memory Bandwidth	**820 GB/s**	**2.9 TB/s**
TDP	—	—

PERFORMANCE BY PRECISION

Precision	AWS Inferentia2 peak TFLOPS (dense)	AWS Trainium2 peak TFLOPS (dense)
FP64	—	—
FP32	47.5	181.0
TF32	190.0	667.0
BF16	190.0	667.0
FP16	190.0	667.0
FP8	190.0	1299.0
FP6	—	—
FP4	—	—
INT8	380.0	—

SPECIFICATIONS

	AWS Inferentia2	AWS Trainium2
Architecture	NeuronCore-v2	NeuronCore-v3
Process Node	—	—
Launch Year	2022	2024
Form Factor	—	—
Memory Type	HBM	—
Interface Width	—	—
Interconnect	NeuronLink-v2	NeuronLink-v3

	AWS Inferentia2	AWS Trainium2
TDP	–	–
FP32 Efficiency	–	–