AWS Inferentia vs AWS Inferentia2

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

OVERVIEW

	AWS Inferentia	AWS Inferentia2
FP32 TFLOPS	–	47.5
Memory	8 GB	32 GB
Memory Bandwidth	50 GB/s	820 GB/s
TDP	–	–

PERFORMANCE BY PRECISION

Precision	AWS Inferentia peak TFLOPS (dense)	AWS Inferentia2 peak TFLOPS (dense)
FP64	–	–
FP32	–	47.5
TF32	–	190.0
BF16	64.0	190.0
FP16	64.0	190.0
FP8	–	190.0
FP6	–	–
FP4	–	–
INT8	128.0	380.0

SPECIFICATIONS

	AWS Inferentia	AWS Inferentia2
Architecture	NeuronCore-v1	NeuronCore-v2
Process Node	–	–
Launch Year	2018	2022
Form Factor	–	–
Memory Type	DDR4	HBM
Interface Width	–	–
Interconnect	NeuronLink-v1	NeuronLink-v2

	AWS Inferentia	AWS Inferentia2
TDP	–	–
FP32 Efficiency	–	–