AWS Inferentia2 vs NVIDIA L40 48GB

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

OVERVIEW

	AWS Inferentia2	NVIDIA L40 48GB
FP32 TFLOPS	47.5	90.5
Memory	32 GB	48 GB
Memory Bandwidth	820 GB/s	864 GB/s
TDP	–	300 W

PERFORMANCE BY PRECISION

Precision	AWS Inferentia2 peak TFLOPS (dense)	NVIDIA L40 48GB peak TFLOPS (dense)
FP64	–	–
FP32	47.5	90.5
TF32	190.0	90.5
BF16	190.0	181.1
FP16	190.0	181.1
FP8	190.0	362.0
FP6	–	–
FP4	–	–
INT8	380.0	362.0
INT4	–	724.0

SPECIFICATIONS

	AWS Inferentia2	NVIDIA L40 48GB
Architecture	NeuronCore-v2	Ada Lovelace
Process Node	–	4nm
Launch Year	2022	2023
Form Factor	–	PCIe
Memory Type	HBM	GDDR6
Interface Width	–	–

	AWS Inferentia2	NVIDIA L40 48GB
Interconnect	NeuronLink-v2	–
TDP	–	300 W
FP32 Efficiency	–	0.30 TFLOPS/W