

AWS Inferentia2

NeuronCore-v2

2022

OVERVIEW

47.5 FP32 TFLOPS	32 GB Memory HBM	820 GB/s Memory Bandwidth	— TDP
---------------------	---------------------	------------------------------	----------

PERFORMANCE METRICS

Precision	Peak (dense)	Bits	Type
FP64	—	—	Standard
FP32	47.5	32	Standard
TF32	190.0	32	Standard
BF16	190.0	16	Standard
FP16	190.0	16	Standard
FP8	190.0	8	Standard
FP6	—	—	Standard
FP4	—	—	Standard
INT8	380.0	8	Standard

MEMORY SPECIFICATIONS

Capacity	32 GB
Type	HBM
Bandwidth	820 GB/s
Interface Width	—

POWER & EFFICIENCY

TDP	—
Max power	—
FP32 dense TFLOPS per W of TDP	—
FP16 dense TFLOPS per W of TDP	—
Per W of TDP figures divide dense peak by accelerator TDP. Board power only: excludes host CPUs, networking, cooling and facility overhead.	

HARDWARE DETAILS

Vendor: AWS	Architecture: NeuronCore-v2	Process Node: —
Form Factor: —	Launch Year: 2022	

COMPUTE ENGINE

NeuronCore-v2: 2

INTERCONNECT

GPU-to-GPU: NeuronLink-v2	Bandwidth: 192 GB/s direction and scope not stated by vendor
---------------------------	---