

AWS Inferentia

NeuronCore-v12018

OVERVIEW

— FP32 TFLOPS	8 GB Memory DDR4	50 GB/s Memory Bandwidth	— TDP
------------------	---------------------	-----------------------------	----------

PERFORMANCE METRICS

Precision	Peak (dense)	Bits	Type
FP64	—	—	Standard
FP32	—	—	Standard
TF32	—	—	Standard
BF16	64.0	16	Standard
FP16	64.0	16	Standard
FP8	—	—	Standard
FP6	—	—	Standard
FP4	—	—	Standard
INT8	128.0	8	Standard

MEMORY SPECIFICATIONS

Capacity	8 GB
Type	DDR4
Bandwidth	50 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-v1	Process Node: —
Form Factor: —	Launch Year: 2018	

COMPUTE ENGINE

NeuronCore-v1: 4

INTERCONNECT

GPU-to-GPU: NeuronLink-v1

Bandwidth: 32 GB/s
direction and scope not stated by vendor