NEUCHIPS Recommendation Accelerator RecAccel™

Accelerating Recommendation Inference for data center servers

Overview

The NEUCHIPS RecAccel™ is world's first recommendation engine for Deep Learning Recommendation Model (DLRM). It can perform 500,000 inferences per second. It is equipped with an ultra-high capacity, high bandwidth memory subsystem for embedding table lookup, and a massively parallel compute FPGA for neural network inference. Running open-source PyTorch DLRM, RecAccel™ outperforms server class CPU and inference GPU by 28X and 65X, respectively.

Characteristics

- Embedding-specific memory architecture, allocation and access scheme.
- Application-specific processing pipeline.
- Scalable Multiply-And-Accumulator (MAC) array.

Features

- NEUCHIPS Asymmetric Quantization
- NEUCHIPS Advanced Symmetric Quantization
- Patented memory architecture
- Frequency: 100MHz
- Integrated Interface: PCIe Gen3
- FPGA: Intel Stratix
- Minimizing TCO

Accelerator Instructions

- Read weight
- Read activation
- Write activation without charge
- Configuration

FPGA Prototype