Contribution ID: **94** Type: **Poster**

Evaluation of Versal AI Engines for Simple Neural Network Inference

AI Engines (AIEs) are a component of the AMD Versal Adaptive Compute Acceleration Platform (ACAP). It is an innovative subsystem that offers extensive parallelism and enhanced compute density. Each AIE is a VLIW processor equipped with a powerful multiply-accumulate (MAC) unit that can perform multiple MAC operations in the same cycle. These processors are grouped together in a 2-D grid of AIEs to form the AIE array. Each AIE can work independently and communicate with the rest of the array via a versatile interconnection network. This particular architecture sparked our interest for an evaluation, since it is machine learning-focused and could possibly have better power efficiency and performance over our existing neural network-to-FPGA flow.

In this evaluation, we mapped one dense neural network (1-D model) and one convolutional neural network (2-D model) to the AIE section of the Versal VCK190. The 1-D model consists of two Dense, one ReLU and one Sigmoid layer and the 2-D model consists of one Convolution, one Dense, one ReLU and one Softmax layer. These models represent generic neural network operations while keeping the simplicity to implement. We explored the best coding practices and characteristics of the AIE.

Additionally, we mapped these models to the FPGA fabric portion of the VCK190 and compared the cost and performance with our AIE implementation. Based on six metrics, we found that while the AIE's efficiency is slightly better than the FPGA fabric in terms of power and silicon area utilization, it is worse than the FPGA in terms of performance, resource utilization and price. This difference is due to limitations in interconnection and the inefficiency of hardware units when the vector data path cannot adapt to certain operations of the input data.

Focus areas

Author: SHEN, Yilin

Co-authors: JOHNSON, Caroline; HAUCK, Scott

Presenter: SHEN, Yilin