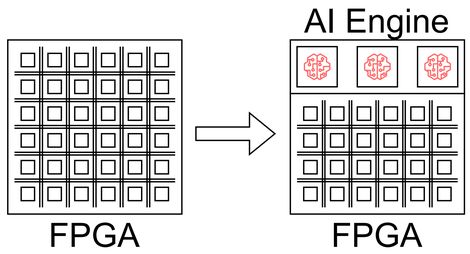
Neural Network Efficiency Evaluation on the AMD Versal AI Engine

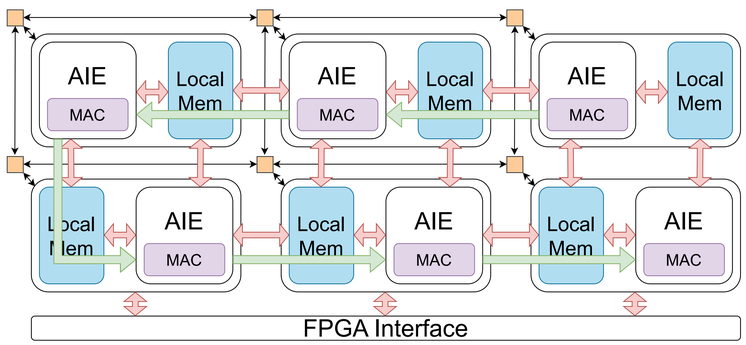
Yilin Shen, Caroline Johnson, Scott Hauck (University of Washington)

Introduction

- AI deployment requires hardware assistance (CPU/GPU/FPGA, etc.)
- FPGA is a popular solution for its low latency
- Next-gen FPGA (AMD Versal) introduced AI Engine(AIE)



The AIE looks promising:



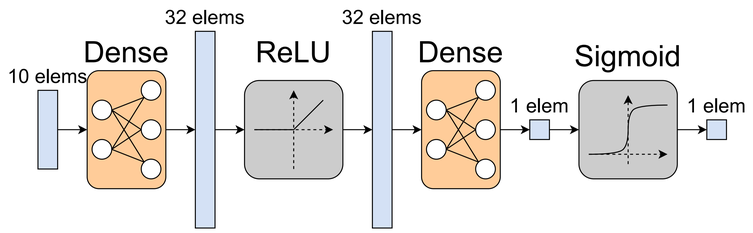
Should we use it? Let's compare AIE to FPGA!

Metrics for Evaluation

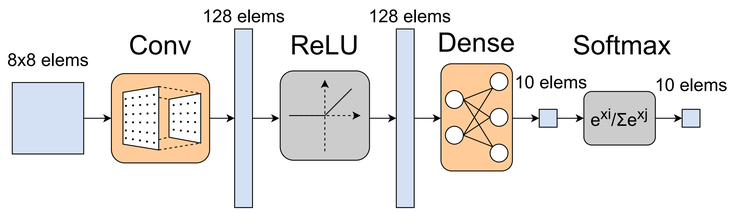
- Initiation interval (1/throughput)
- Latency
- Power
- Price
- Resource utilization
- Silicon Area utilization

The smaller, the better!

Models for Evaluation

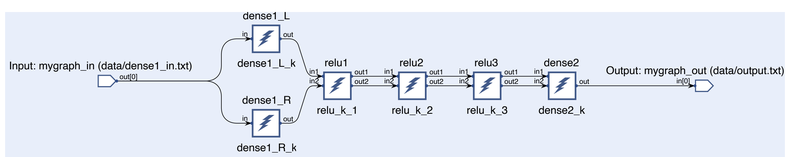


1-D model structure

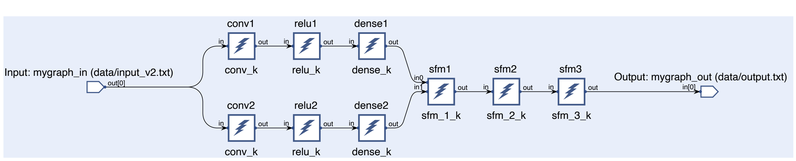


2-D model structure

AI Engine Implementation



1-D model AIE mapping

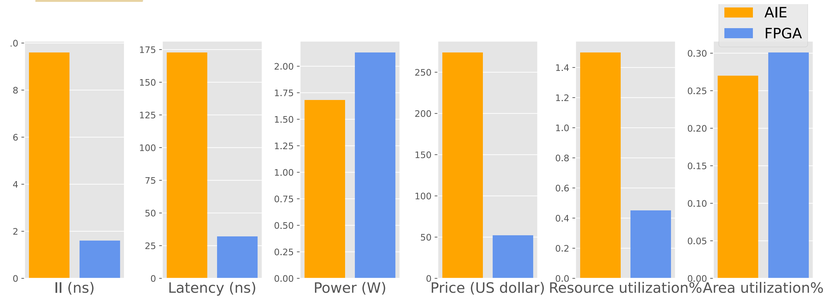


2-D model AIE mapping

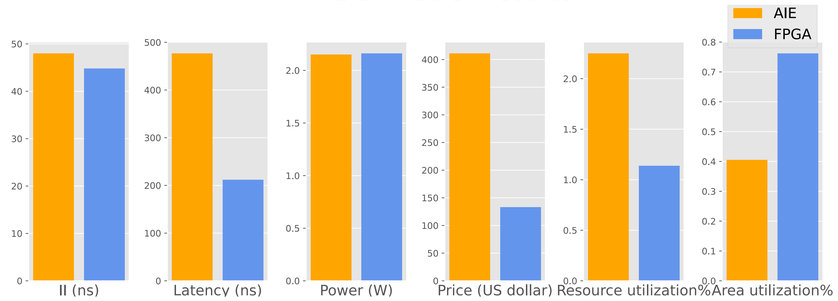
* FPGA Implementation: see [1]

Comparison Results

The smaller, the better!



1-D model Results



2-D model Results

Conclusion

- If throughput and latency is the priority, go for FPGA
- If cost or replication is the priority, go for FPGA
- AIE saves silicon area and power for basic ML workload

Future Work

- Map more complex models
  - Convolution with stride
  - Multi-channel convolution
- Evaluate AIE-ML: next-gen AIE
  - Optimized interconnection and enhanced MAC unit

Reference

[1] Johnson, Caroline. Evaluating the Quality of HLS4ML's Basic Neural Network Implementations on FPGAs. MS thesis. University of Washington, 2023.

NSF Grant 2117997