MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects

Hyokjung Kwon, Ananda Samajdar, and Tushar Krishna
Georgia Institute of Technology


Scheduled Tutorial in ISCA 2018 (Jun 3, 2018)
http://synergy.ece.gatech.edu/tools/maeri

Spatial DNN Accelerators

- High throughput with massive parallelism provided by spatially placed compute units (or, processing elements; PEs)
- Data reuse opportunities within PE array
- Direct communication among compute units without talking to memory
- Energy efficiency

MAERI Overview

Features
- Fine-grained compute units
- Programmable interconnections
- CNN/RNN (LSTM) support

Augmented Reduction Tree (ART)

- Enables near 100% mapping efficiency
- MAERI supports reconfigurable interconnect with flexibility and sufficient bandwidth to support irregular dataflow
- MAERI provides reconfigurable interconnect with flexibility and sufficient bandwidth to support irregular dataflow

Walk-through Examples

Evaluations

Area and Power (Accelerator)

- SA: Systolic Array
- ASC: Augmented Reduction Tree
- MAP: Mapping Algorithm
- Sun: Sunflower (a form of the sunflower)
- Fused Alexnet
- Alexnet
- VGG-16

Performance (throughput/latency)

- In average, MAERI provided 92% lower latency

Conclusion

- Diverse DNN dimensions and optimizations (sparsity, fused layer, etc.) introduce irregular dataflow in DNN Accelerators
- Most of DNN accelerators contain rigid interconnections that involve mapping inefficiency with irregular dataflow
- MAERI provides reconfigurable interconnect with flexibility and sufficient bandwidth to support irregular dataflow
- MAERI provided 8-458% better compute unit utilization across multiple dataflow mappings that results in 72.4% lower latency in average