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

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

*In Proceedings of the 23rd ACM International conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)***

Scheduled Tutorial in ISCA 2018 (Jun 3, 2018)

http://synergy.ece.gatech.edu/tools/maeri

Spatial DNN Accelerators

Benefits of Spatial DNN Accelerators

- **High throughput** with massive paralleism 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

Challenges of Spatial DNN Accelerators

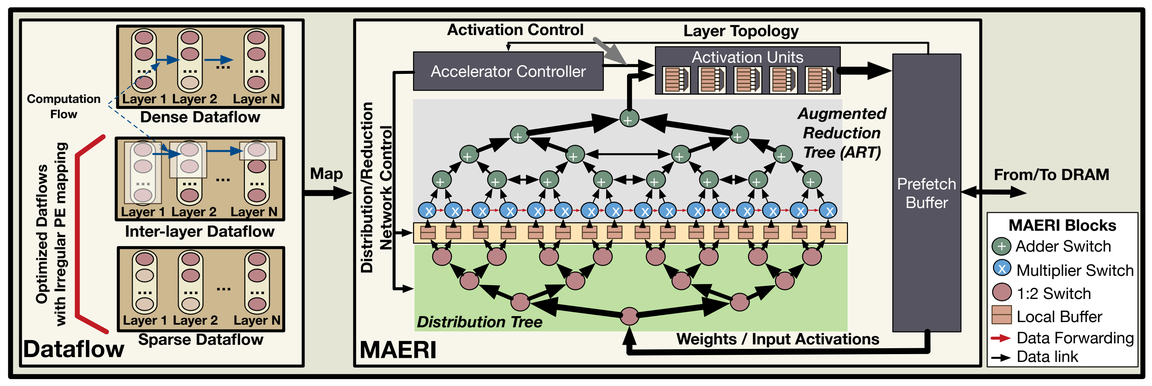
- Programmability
- Underutilization

→ **Sources of Underutilization Problem**

- **DNN topology**
  - Various DNN **layer types and dimensions**
- **Mapping Inefficiencies**
  - **Partitioning strategies** (intra- or cross-layer, tiling)
  - **Data density optimizations** (sparsity, compression, etc.)
  - **Data reuse strategies** (weight- and output-stationary, etc.)
- **Interconnects (NoCs)**
  - Insufficient **bandwidth**
  - **Inflexibility** base on rigid interconnects

→ MAERI

MAERI Overview

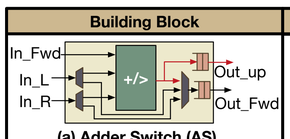
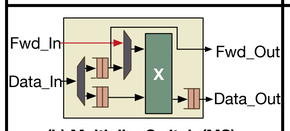
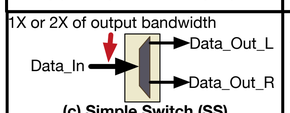
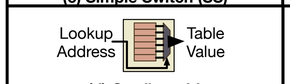


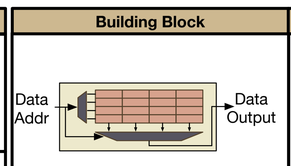
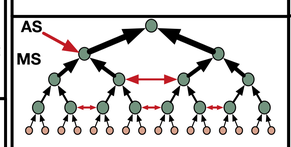
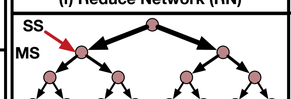
Features

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

→ **Enables high-utilization even with irregular dataflow (Sparse, inter-layer fusion, etc.)**

Building Blocks of MAERI

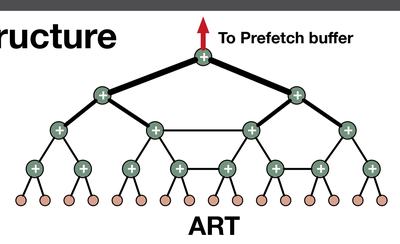
Building Block	Description
 (a) Adder Switch (AS)	The AS is an adder augmented with a tiny switch that enables us to map arbitrary adder trees over our non-blocking reduce network called ART. It also contains comparators for pooling operations.
 (b) Multiplier Switch (MS)	The MS is a multiplier augmented with a tiny switch that is used for local data forwarding. It is used by CNNs for generating partial sums from weights and input activation values, and by RNNs for generating gate values, input activations, and previous output activations.
 (c) Simple Switch (SS)	The SS provides 1:2 switching functionality in the distribute network's chubby tree nodes.
 (d) Configurable Look-up Table (LT)	LTs implement activation functions such as *sigmoid* or *tanh*. We load all the necessary functions for a neural network and change its target function in run time based on the configuration generated by *MAERI*.

Building Block	Description
 (e) Prefetch Buffer (PB)	The PB works like a cache memory between DRAM and the computation units. We implement this as a private scratchpad. Because the characteristics of a prefetch buffer would differ based on SRAM technology library and they are usually commercial libraries, we provide a default multi-banked implementation using flip-flops.
 (f) Reduce Network (RN)	The RN is a network structure for reduce and collection operations. It is based on a new adder tree structure, augmented reduction tree (ART) we propose in this paper. ART facilitates mapping multiple configurable non-blocking adder trees and minimizing inactive multiplier switches.
 (g) Distribute Network (DN)	The DN is based on a chubby-tree structure, which is a tree-based network with wider link bandwidth in higher levels of a tree. We exploit abandunt bandwidth at high level links to enable multicast functionality, which is one of the most common traffic patterns in DNN accelerators.

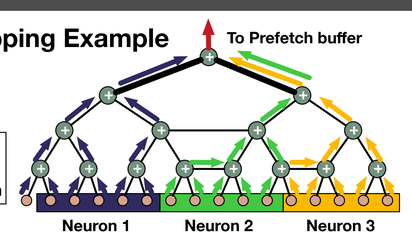
- Separate multiplier/adder integrated with network switch
- Non-blocking Networks supporting efficient mapping

Augmented Reduction Tree (ART)

Structure

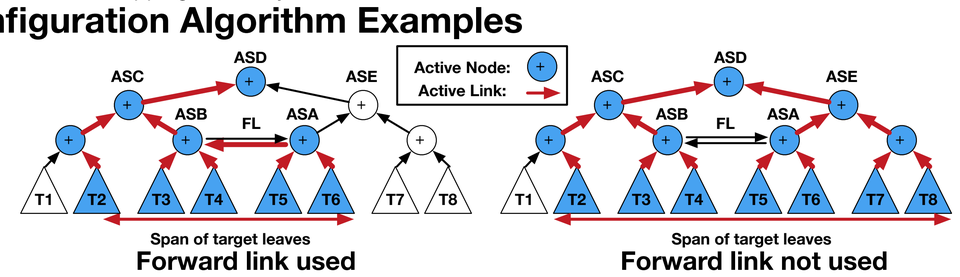


Mapping Example



- Augments binary adder tree with extra forwarding links and extra bandwidth for flexibility
- Enables near 100% mapping efficiency

Configuration Algorithm Examples

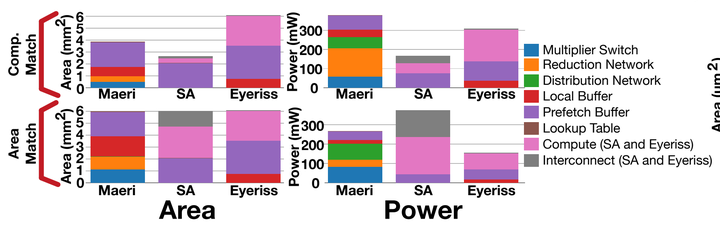


- Identify the usage of each forwarding links by checking usage of subtrees
- Forward link is used only if less than equal to half of subtrees of the parent node (ASE) of the node with forward link (ASA)

Walk-through Examples

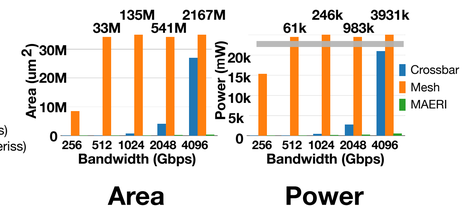
Evaluations

Area and Power (Accelerator)



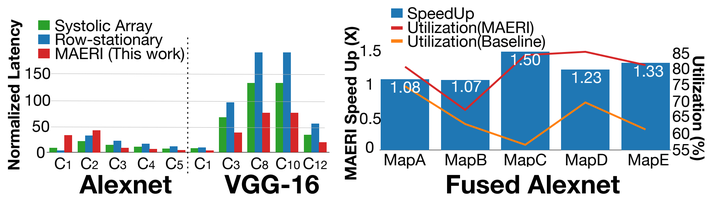
* SA: Systolic Array
*Comp match: The same number of compute units

Area and Power (NoCs)



- NoCs (or interconnects) in MAERI provides significantly better scalability in area and power

Performance (throughput/latency)



- In average, MAERI provided 72.4% lower latency

Utilization and NoCs

- Compared to rigid NoCs (interconnects), MAERI provides better utilization via efficient mapping

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