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

http://synergy.ece.gatech.edu/tools/maeri/maeri-tutorial-hpca-2019

Tushar Krishna
Georgia Tech

Tutorial @ HPCA 2019
Feb 16 2019
## Schedule: Afternoon

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 - 2:20</td>
<td>Overview of MAERI</td>
<td>Tushar</td>
</tr>
<tr>
<td>2:20 – 3:00</td>
<td>MAERI Mapper</td>
<td>Zhongyuan</td>
</tr>
<tr>
<td>3:00 – 3:20</td>
<td>MAERI RTL</td>
<td>Hyoukjun</td>
</tr>
<tr>
<td>3:20 – 3:40</td>
<td>MAERI Demo + Hands-on Exercises</td>
<td>Hyoukjun</td>
</tr>
<tr>
<td>3:40 – 4:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>4:00 – 4:30</td>
<td>Extensions</td>
<td>Michael</td>
</tr>
<tr>
<td>4:30 – 4:40</td>
<td>Wrap-Up</td>
<td>Tushar</td>
</tr>
</tbody>
</table>
Dataflows within DNN Accelerators

• **DNN Topologies**
  - Layer *size / shape*
  - Layer *types*: Convolution / Pool / FC / LSTM
  - New *sub-structure*: e.g., Inception in Googlenet

• **Compiler/Mapper (e.g., MAESTRO)**
  - Loop *reordering*
  - Loop *tiling size*
  - Cross-layer mapping

• **Algorithmic Optimization**
  - Weight pruning: *Sparse* workload
The current trend for supporting multiple dataflows

- New Dataflow $\rightarrow$ New Accelerator
  - **Data reuse:** FlexFlow (2017), Eyeriss (2016), ...
  - **Cross-layer:** Fused CNN (2016)
  - **Sparse CNN:** SCNN (2017), EIE (2016), ...
  - **LSTM:** ESE (2017), ...

Can we have one architectural solution that can handle arbitrary dataflows and provides $\sim$100% utilization?
What is the computation in a DNN?

Compute weighted sum

\[ \sum (W_i X_i) \]

Independent multiplication
Accumulation of partial products

Our Key insight: Each dataflow translates into neurons of different sizes
The MAERI Abstraction

How to enable flexible grouping?

Reconfigurable Interconnects!
Virtual Neuron (VN): Temporary grouping of compute units for an output

How to enable flexible grouping?

Reconfigurable Interconnects!
Traffic Patterns in DNN Accelerators*

Distribution
- **e.g.** weight distribution to PEs in weight-stationary accelerator

Collection
- **e.g.** output collection in output-stationary accelerator

Local Forwarding
- **e.g.** Input forwarding in row-stationary accelerator

* PB: Prefetch buffer (Global buffer)
* NoC: Network-on-Chip (Interconnection network)
* PE: Processing element (Compute units)
The MAERI Implementation

- **Distribution Network**
  - Spatial Reuse via Multicasts
  - High Bandwidth via fat links

- **Linear Local Network**
  - Forwarding of weights
  - Spatio-Temporal Reuse

- **Reduction Network**
  - High Bandwidth via fat links
  - Provably Non-blocking Reductions via forwarding links
Why 1D Configurable Design?

Systolic Array (TPU): Map kernels on fixed 2D array

<table>
<thead>
<tr>
<th>Dimension</th>
<th>GEMM1</th>
<th>GEMM2</th>
<th>GEMM3</th>
<th>GEMM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>3</td>
<td>128</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>N</td>
<td>128</td>
<td>256</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>K</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Why 1D Configurable Design?

MAERI: Constructs 2D array at run time

<table>
<thead>
<tr>
<th>Dimension</th>
<th>GEMM1</th>
<th>GEMM2</th>
<th>GEMM3</th>
<th>GEMM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>3</td>
<td>128</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>N</td>
<td>128</td>
<td>256</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>K</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
MAERI Operation Example

Dataflow (from CPU)

Accelerator Controller

Controller configures switches

Distribution Tree

Multiplier Switch

Adder Switch

Lookup Table

Lookup Table

Legend

Lookup Table

Inputs

Weights

Virtual Neuron Construction

<Step 1>

Sparse weight filter

VN size = 5

Output Row 0

Output Row 1

Output Row 2

Activation Units

From/To DRAM

Simple Switch

Virtual Neuron Construction

VN size = 5

Inputs

Weights

Outputs

February 16, 2019
MAERI Operation Example

Accelarator Controller

Dataflow (from CPU)

Legend
- Simple Switch
- Multiplier Switch
- Adder Switch
- Lookup Table

Legend
- Simple Switch
- Multiplier Switch
- Adder Switch
- Lookup Table

Activation Units

Distribution Tree

Augmented Reduction Tree

Weights

<Step 2>
Weight Distribution

Distribution bandwidth is tunable

Inputs

Outputs

Distribution bandwidth is tunable

February 16, 2019

Tushar Krishna | Georgia Institute of Technology

MAERI Tutorial @ HPCA 2019
MAERI Operation Example

Dataflow (from CPU)

Legend
- Simple Switch
- Multiplier Switch
- Adder Switch
- Lookup Table

Activation Units

<Step 2>
Weight Distribution
MAERI Operation Example

Accelerator Controller

Dataflow (from CPU)

Legend
- Simple Switch
- Multiplier Switch
- Adder Switch
- Lookup Table

Augmented Reduction Tree

Distribution Tree

Activation Units

<Step 3>
Input Distribution

Utilize multicast to reduce latency and energy

February 16, 2019
MAERI Operation Example

Dataflow (from CPU)

Legend
- Simple Switch
- Multiplier Switch
- Adder Switch
- Lookup Table

Activation Units

Simultaneous reduction

<Step 4>
Partial sum reduction

Accelerator Controller

Distribution Tree

Augmented Reduction Tree

Inputs

Weights

Outputs

MAERI Tutorial @ HPCA 2019  Tushar Krishna | Georgia Institute of Technology  February 16, 2019
MAERI Operation Example

Accelerator Controller

Dataflow (from CPU)

Legend
- Simple Switch
- Multiplier Switch

Legend
- Simple Switch
- Multiplier Switch

Activation Units

Lookup Table

Distribution Tree

Augmented Reduction Tree

Inputs

Weights

Outputs

Weights: Stationary

Inputs: Partially reused via forwarding

<Step 5>

Sliding Window

Repeat Step 4 - 5

MAERI Tutorial @ HPCA 2019

Tushar Krishna | Georgia Institute of Technology

February 16, 2019
Our Key insight: Each dataflow translates into neurons of different sizes.
Example Mapping – Sparse CNN

Our Key insight: Each dataflow translates into neurons of different sizes
Our Key insight: Each dataflow translates into neurons of different sizes.
Performance with Dense Workload

• Total Latency (Runtime) for Convolution

* Normalized to ideal case (100% utilization, Infinite bandwidth)

MAERI reduces runtime upto 65%, 42% in avg.

MAERI reduces runtime upto 63%, 57% in avg.

LSTMs from Yonghui We, et. al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.", Arxiv Preprint, 2016
MAERI reduces energy upto 57% and 28% in average compared to Row-Stationary (dense dataflow) and 7.1% in average compared to Systolic Array (sparse dataflow)
• **MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  *In Proc of the 23rd ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS),* Mar 2018
  - *Selected as Honorable Mention in IEEE Micro Top Picks in Computer Architecture 2019*

• **A Communication-Centric Approach for Designing Flexible DNN Accelerators**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  *In IEEE Micro Special Issue on Hardware Acceleration, Nov/Dec 2018*

• **Enabling Efficient Mapping Space Exploration on a Reconfigurable Neural Accelerator**
  Zhongyuan Zhao, Hyoukjun Kwon, Sachit Kuhar, Weiguang Sheng, Zhigang Mao, and Tushar Krishna
  *In Proc of the IEEE Inter. Symp. on Performance Analysis of Systems and Software (ISPASS),* Mar 2019

• **Rethinking NoCs for Spatial Neural Network Accelerators**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  *In Proc of 11th International Symposium on Networks-on-Chip (NOCS),* Oct 2017
## Schedule: Afternoon

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 - 2:20</td>
<td>Overview of MAERI</td>
<td>Tushar</td>
</tr>
<tr>
<td>2:20 – 3:00</td>
<td>MAERI Mapper</td>
<td>Zhongyuan</td>
</tr>
<tr>
<td>3:00 – 3:20</td>
<td>MAERI RTL</td>
<td>Hyoukjun</td>
</tr>
<tr>
<td>3:20 – 3:40</td>
<td>MAERI Demo + Hands-on Exercises</td>
<td>Hyoukjun</td>
</tr>
<tr>
<td>3:40 – 4:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>4:00 – 4:30</td>
<td>Extensions</td>
<td>Michael</td>
</tr>
<tr>
<td>4:30 – 4:40</td>
<td>Wrap-Up</td>
<td>Tushar</td>
</tr>
</tbody>
</table>