Enabling Rapid Design Space Exploration and Prototyping of DNN Accelerators

Tushar Krishna
Georgia Tech

Tutorial @ HPCA 2019
Feb 16 2019
Spatial (or Dataflow) Accelerators

- Millions of Parameters (i.e., weights)
  - Billions of computations
  - Heavy data movement

  Spread computations across hundreds of ALUs

  Reuse data within the array via local memories and direct communication

  Examples: MIT Eyeriss, Google TPU, …
Two Key HW Design Challenges

• How do we map billions of computations over *limited* compute and memory resources (aka *Dataflow*)?

• How do we design the accelerator to efficiently map arbitrary layer types and dataflows?
MAESTRO: Analytical cost model for DNN dataflows

https://arxiv.org/abs/1805.02566
<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 to 9:00</td>
<td>Introduction and Background</td>
<td>Tushar</td>
</tr>
<tr>
<td>9:00 to 10:00</td>
<td>A primer on DNN Dataflows</td>
<td>Michael</td>
</tr>
<tr>
<td>10:00 – 10:30</td>
<td>MAESTRO Data Directives</td>
<td>Prasanth</td>
</tr>
<tr>
<td>10:30 – 10:50</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>10:50- 11:10</td>
<td>MAESTRO Data Directives [contd]</td>
<td>Prasanth</td>
</tr>
<tr>
<td>11:10 – 11:45</td>
<td>MAESTRO Analytical Model</td>
<td>Hyoukjun</td>
</tr>
<tr>
<td>11:45 – 12:30</td>
<td>MAESTRO Hands-on Exercises</td>
<td>Hyoukjun &amp; Prasanth</td>
</tr>
<tr>
<td>12:30 – 2:00</td>
<td>Lunch</td>
<td></td>
</tr>
</tbody>
</table>
MAERI – DNN Accelerator for Flexible Dataflows

Find Optimal Dataflow

Deep Neural Network

MAERI Mapper

Weight, Input, Output SRAM

To/From DRAM

Dataflow Configs

MAERI RTL

Verilog

Cycle-Accurate Sims

Virtual Neurons

MAERI

Zhao et al, ISPASS 2019

Kwon et al., ASPLOS 2018,

Kwon et al., ASPLOS 2018,
## 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</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>Hands-on Exercises</td>
<td>Hyoukjun &amp; Zhongyuan</td>
</tr>
<tr>
<td>4:30 – 5:00</td>
<td>Extensions</td>
<td>Michael</td>
</tr>
<tr>
<td>5:00 – 5:10</td>
<td>Wrap-Up</td>
<td>Tushar</td>
</tr>
</tbody>
</table>
Tool Release and Resources

• Slides and Video will be posted on the tutorial page
  • http://synergy.ece.gatech.edu/tools/maeri/maeri-tutorial-hpca2019/

• All codebases will be added to github by tomorrow evening
  • Link will be added on tutorial website

• Feedback
  • Please add your name to the sign-up if you have not
    • For statistics
    • We will send out a feedback form

• This is all work in progress
  • Please reach out to us if you find a bug
    • Better still – fix it and contribute back on github!
Future Extensions

• MAESTRO
  • validation
  • support for sparsity
  • support for other layer-types

• MAERI
  • Testbenches for other layer-types and networks
  • Mapper to Testbench auto-generator
  • Code-optimization for FPGAs
Presenters

Tushar Krishna
Assistant Professor, School of ECE, Georgia Tech
PhD (MIT) in 2014
tushar@ece.gatech.edu

Michael Pellauer
Sr. Research Scientist, NVIDIA
Intel VSSAD (2010-2015)
PhD (MIT) in 2010
mpellauer@nvidia.com

Hyoukjun Kwon
PhD Candidate
School of CS, Georgia Tech
hyoukjun@gatech.edu

Prasanth Chatarasi
PhD Candidate
School of CS, Georgia Tech
cprasanth@gatech.edu

Zhongyuan Zhao
PhD Candidate
School of CS, Shanghai Jiaotong University
zyzhao.sjtu@gmail.com
Enabling Rapid Design Space Exploration and Prototyping of DNN Accelerators

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