Physical Synthesis for Advanced Neural Network Processors

Zhuolun He\textsuperscript{1}, Peiyu Liao\textsuperscript{1}, Siting Liu\textsuperscript{1}, Yuzhe Ma\textsuperscript{1}, Yibo Lin\textsuperscript{2}, Bei Yu\textsuperscript{1},

\textsuperscript{1}The Chinese University of Hong Kong
\textsuperscript{2}Peking University
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Deep Learning: an Important Workload

- Deep learning has emerged as an important workload
- Yet, how to efficiently execute them?
Dataflow Optimization and Physical Synthesis

- Dataflow optimization becomes essential
  - Schedules operation by data availability
  - Exposes opportunity for parallelism and data reuse
- Dataflow regularity gives rise to new physical synthesis approaches
  - Directly determines system performance!
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Datapath Driven Standard Cell Placement

- Classical idea: bit-sliced DSP datapaths [Cai, DAC’90]
  - Decide ordering of linearly placed blocks
  - Solved by A* in the search space

- Standard cell placement [Tsay, TCAD’95]
  - Strongly connected subcircuits (cones) are extracted
  - BFS + heuristics
  - Placed as macro cells
Datapath Driven Placement: Abstract Physical Model

- Placer has little control of exact locations if datapath is generated separately
- Abstract physical model [Ye, ICCAD’00]
  - Bit-sliced abstraction
  - Compiled from HDL
  - Blocks are placed abutted
- Two-step heuristic for linear placement
  - Quadratic assignment
  - Sliding window optimization

APM of a booth multiplier [Ye, ICCAD’00]
Datapath Driven Systolic Array Placement

- Systolic arrays are a popular choice to support neural network computations
- Current FPGA CAD tools cannot synthesize them in high quality
- One solution: restrict fixed locations for PEs [Zhang, ISCAS’19]
  - Sufficient DSPs, close to used I/O banks

PE placement with floorplan constraints [Zhang, ISCAS’19]
Datapath Driven Placement: Many More

- Detailed placement [Serdar, DATE’01]
- SOC placement [Tong, JOS’02] [Jing, ICCAS’02]
- Parallel multiplier design [Bae, ISPD’15]
- General ASIC design [Ye, ISCAS’02] [Chou, DAC’12] [Wang, IETCDS’17]
- ...
Regularity Extraction

- Consider cells with the same bit-slice are lined up horizontally [Nijssen IWLS’96]
  - geometric regularity: circuit is fitted onto a matrix of rectangular buckets
  - interconnect regularity: most nets are within one slice/one stage
- Typical methods for regularity extraction
  - Search-wave expansion [Nijssen IWLS’96]
  - Signature-based [Arikati, ICCAD’97]
  - Template covering [Chowdhary, TCAD’99]
  - Network flow [Xiang, ISPD’13]
  - Bipartite graph vertex covering [Huang, DAC’17]
Regularity Extraction: Network Flow Approach

- **Datapath main frame** (DMF) [Xiang, ISPD’13]
  - a set of $n$ disjoint paths from input to output
  - maximize the number of datapath gates on these paths
  - Can be optimally identified by the min-cost max-flow algorithm

DMF identification can be transformed to a network flow problem [Xiang, ISPD’13]
Machine Learning Techniques for Datapath Driven Placement

- Machine learning techniques are involved [Ward, DAC’12]
  - Graph-based (e.g., automorphism) and physical features (e.g., cell area) are analyzed and extracted from the netlist
  - Features are fed to SVMs and NNs to classify and evaluate datapath patterns
  - Maximize the evaluation accuracies of datapath and non-datapath pattern
  - Proposed new placer: PADE

PADE effectively handles datapath in placement. [Ward, DAC’12]
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Problem Formulation

- Floorplan computation kernels of a neural network [Michael, ISPD’20]
  - Blocks are soft: kernel sizing
  - Floorplanning
  - Optimize performance, wirelength, etc.

Kernel floorplanning. Figure adopted from ISPD Contest.
Datapath Driven Floorplan

- Our strategy: datapath-driven floorplan
  - Neural network is usually hierarchical: a stack of layers
  - Arrange the floorplan according to the datapath

Datapath driven floorplans. (a) Single-row; (b) Multi-row.
Floorplan Compacting

- The floorplan can be further compacted
  - Vertically: push to center
  - Horizontally: a linear programming problem

(a) Mamba floorplan; (b) Horizontally compacted Mamba floorplan.
Result: Datapath Driven Floorplan Outperforms Benchmark Statistics

<table>
<thead>
<tr>
<th>Case</th>
<th>#Kernels</th>
<th>λ₁</th>
<th>λ₂</th>
<th>Max Time</th>
<th>WL</th>
<th>Protocol</th>
<th>Cost</th>
<th>Max Time</th>
<th>WL</th>
<th>Protocol</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>37044</td>
<td>3611.5</td>
<td>11</td>
<td>40655.5</td>
<td>35280</td>
<td>2047</td>
<td>13</td>
<td>37327</td>
</tr>
<tr>
<td>B</td>
<td>34</td>
<td>1</td>
<td>0</td>
<td>70560</td>
<td>6657</td>
<td>20</td>
<td>77217</td>
<td>65856</td>
<td>4905</td>
<td>17</td>
<td>70761</td>
</tr>
<tr>
<td>C</td>
<td>102</td>
<td>1</td>
<td>0</td>
<td>76608</td>
<td>15696</td>
<td>69</td>
<td>92034</td>
<td>65772</td>
<td>4278</td>
<td>281</td>
<td>70050</td>
</tr>
<tr>
<td>D</td>
<td>54</td>
<td>1</td>
<td>0</td>
<td>38304</td>
<td>9327.5</td>
<td>44</td>
<td>47631.5</td>
<td>34944</td>
<td>3071.5</td>
<td>89</td>
<td>38015.5</td>
</tr>
<tr>
<td>E</td>
<td>17 10 100</td>
<td>1</td>
<td>0</td>
<td>36288</td>
<td>2080.5</td>
<td>7</td>
<td>57793</td>
<td>39690</td>
<td>590</td>
<td>16</td>
<td>47190</td>
</tr>
<tr>
<td>F</td>
<td>34 10 100</td>
<td>1</td>
<td>0</td>
<td>76608</td>
<td>3237</td>
<td>15</td>
<td>110478</td>
<td>70560</td>
<td>1475</td>
<td>14</td>
<td>86710</td>
</tr>
<tr>
<td>G</td>
<td>102 10 100</td>
<td>1</td>
<td>0</td>
<td>91728</td>
<td>7784</td>
<td>29</td>
<td>172468</td>
<td>69888</td>
<td>2508</td>
<td>141</td>
<td>109068</td>
</tr>
<tr>
<td>H</td>
<td>54 10 100</td>
<td>1</td>
<td>0</td>
<td>47040</td>
<td>4450</td>
<td>21</td>
<td>93640</td>
<td>43008</td>
<td>893</td>
<td>115</td>
<td>63438</td>
</tr>
<tr>
<td>I</td>
<td>27 4 0</td>
<td>1</td>
<td>0</td>
<td>56448</td>
<td>3790</td>
<td>16</td>
<td>71608</td>
<td>52920</td>
<td>612</td>
<td>13</td>
<td>55368</td>
</tr>
<tr>
<td>J</td>
<td>81 4 0</td>
<td>1</td>
<td>0</td>
<td>63504</td>
<td>8009.5</td>
<td>52</td>
<td>95542</td>
<td>57792</td>
<td>1117.5</td>
<td>286</td>
<td>62262</td>
</tr>
<tr>
<td>K</td>
<td>18 4 0</td>
<td>1</td>
<td>0</td>
<td>576</td>
<td>236</td>
<td>3</td>
<td>1520</td>
<td>504</td>
<td>400</td>
<td>14</td>
<td>2104</td>
</tr>
<tr>
<td>L</td>
<td>54 4 0</td>
<td>1</td>
<td>0</td>
<td>1280</td>
<td>910.5</td>
<td>60</td>
<td>4922</td>
<td>504</td>
<td>774</td>
<td>114</td>
<td>3600</td>
</tr>
<tr>
<td>M</td>
<td>25 4 0</td>
<td>1</td>
<td>0</td>
<td>2359296</td>
<td>9359</td>
<td>24</td>
<td>2396732</td>
<td>2336256</td>
<td>5100</td>
<td>67</td>
<td>2356656</td>
</tr>
<tr>
<td>N</td>
<td>28 4 0</td>
<td>1</td>
<td>0</td>
<td>2268</td>
<td>707.5</td>
<td>0</td>
<td>5098</td>
<td>1599</td>
<td>448.5</td>
<td>9</td>
<td>3393</td>
</tr>
<tr>
<td>O</td>
<td>27 40 400</td>
<td>1</td>
<td>0</td>
<td>63504</td>
<td>1202</td>
<td>6</td>
<td>113984</td>
<td>52920</td>
<td>612</td>
<td>13</td>
<td>82600</td>
</tr>
<tr>
<td>P</td>
<td>81 40 400</td>
<td>1</td>
<td>0</td>
<td>115101</td>
<td>4015</td>
<td>24</td>
<td>285301</td>
<td>66528</td>
<td>2273</td>
<td>102</td>
<td>198248</td>
</tr>
<tr>
<td>Q</td>
<td>18 40 400</td>
<td>1</td>
<td>0</td>
<td>1152</td>
<td>178</td>
<td>1</td>
<td>8672</td>
<td>504</td>
<td>400</td>
<td>14</td>
<td>22104</td>
</tr>
<tr>
<td>R</td>
<td>54 40 400</td>
<td>1</td>
<td>0</td>
<td>1372</td>
<td>1443</td>
<td>30</td>
<td>71092</td>
<td>504</td>
<td>774</td>
<td>114</td>
<td>77064</td>
</tr>
<tr>
<td>S</td>
<td>25 40 400</td>
<td>1</td>
<td>0</td>
<td>2495376</td>
<td>3551</td>
<td>25</td>
<td>2647416</td>
<td>2396160</td>
<td>1899.5</td>
<td>65</td>
<td>2498140</td>
</tr>
<tr>
<td>T</td>
<td>28 40 400</td>
<td>1</td>
<td>0</td>
<td>5720</td>
<td>555.5</td>
<td>0</td>
<td>27940</td>
<td>2015</td>
<td>367.5</td>
<td>9</td>
<td>20315</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.17</td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Processing-in-Memory Technologies

- PIM provides massive parallelism with high energy efficiency [Wang, ASAP’19]

- However, lots of systems rely on external control

- Neural networks can be implemented with NVMs
  - RRAM [Chi, SIGARCH’16] [Sun, DATE’18]
  - STT-MRAM [Yan, SC’18] [Pan, TMAGN’18]
  - PCM [Kim, DATE’20]
  - memristor [Yao, Nature’20]

- In-memory analog simulation
  - memristor crossbar [Li, Nat. Commun’18]
  - FTJ [Li, Adv.Mater’20]
Nanophotonic and Optical Neural Networks

- The nanophotonic circuit is an alternative neuromorphic computing system
  - ultra-high bandwidth, speed and ultra-low energy consumption
- Signals are encoded in the amplitude of optical pulses
- Implements a multi-layer optical neural network (ONNs)
- Recent advances
  - Exploring nonlinear functions with ONNs [Zuo, OPTICA’19]
  - Reducing area overhead of ONNs [Gu, ASPDAC’20]
3D Technologies and Beyond-CMOS devices

- 3D technologies offer a wide spectrum of integration schemes
  - TSV-based Tetris [Gao, ASPLOS’17]
  - TCI-based QUEST [Ueyoshi, ISSCC’18]
  - M3D-based accelerators [Chang, ISLPED’17] [Chang, JETC’18]

- Beyond-CMOS devices give rise to new solutions for low-power design
  - HyperFET for spiking neural networks [Tsai, TMSCS’16]
  - TFET for cellular neural networks [Palit, ISLPED’13]
Outline

Introduction

Survey: Datapath Driven Placement

Case Study: Datapath Driven Floorplan for Neural Network Processors

Discussion: Advanced Technologies for Neural Network Processors

Conclusion
Conclusion

► Review on datapath driven placement
  ► Placement with datapath constraints
  ► Regularity extraction methods
  ► Machine learning techniques

► Case study: floorplan for advanced neural network processors
  ► Datapath driven floorplan greatly outperforms standard methods

► Advanced technologies for neural network processor design
  ► Processing-in-Memory
  ► Nanophotonic
  ► 3D technologies
  ► Beyond-CMOS devices
Thank You