Dataflow-Architecture Co-Design for 2.5D DNN Accelerators using Wireless Network-on-Package

Robert Guirado (rguirado@ac.upc.edu)
Hyoukjun Kwon, Sergi Abadal, Eduard Alarcón, Tushar Krishna

26th Asia and South Pacific Design Automation Conference (ASP-DAC)
Virtual – January 18-21, 2021
Overview

Problem: Electrical scale-out is bandwidth-limited

WIENNA solves this via WNoP technology

WIENNA: Up to 5.1X speedup and 38% energy savings
Background: DNN

“DeepPose” A.Toshev et al.
CVPR 2014
Background: DNN Acceleration

- Higher throughput and energy-efficiency than GPUs and CPUs via parallelism and dedicated circuitry.
- Off-chip memory + global shared memory + array of PEs connected via a Network-on-Chip (NoC)
- Specific dataflows will define:
  - Data partitioning
  - Data movement
  - Data reuse

“Eyeriss” Chen, Y et al.
Background: DNN Acceleration

- Two approaches to increase computing power:
  - Scale-up → more PEs
  - Scale-out → more chiplets
- 2.5D chiplet integration enables efficient scale-out

"Simba" Shao, YS et al.
Baseline dataflows

Dataflows: DNN mapping strategies for leveraging data movement

- Filter-partitioning (KP-CP)
- Batch-partitioning (NP-CP)
- Activation-partitioning (YP-XP)
Observations

- Different layer types favour different partition strategies.
- Different layer types saturate to peak throughput at different bandwidth values.
- The communication fabric for data distribution plays a key role in performance.
- Broadcast support and high-bandwidth are critical for scalability.
- Supporting adaptive partitioning strategies for each layer, rather than picking a fixed one for all layers, is crucial for performance.
Motivation

- Electrical scale-out via 2.5D interposers is bandwidth-limited
  - Large chiplet microbumps compared to pitch wires
- This allows only neighbour-to-neighbour connections.
- Collection latency can be hidden, distribution is in critical path.

“Simba” Shao, YS et al.
Wireless Network-on-Package

- WNoP: short-range wireless transceivers
  - Processor or memory chiplets can be augmented with antennas and TRXs to communicate within chiplet or to other chiplets
  - Package as a propagation medium

- WNoP advantages:
  - Natural broadcast capability
  - Low latency and linear energy dependence
  - High-bandwidth
  - Dynamic topology
State-of-the-art transceivers reach up to 100 Gbps

We extrapolate power and area trends based on 70+ short-range TRXs with different modulations and technologies

WIENNA Architecture

- Two-level hierarchy:
  - HBM, SRAM, chiplet array, NoP
  - Within-chiplet architecture

- Wireless distribution
  - Broadcast enabled
  - SRAM (Tx) → Chiplets (Rx)

- Wired collection through interposer
  - 256 chiplets and 64 PEs per chiplet.
WIENNA Architecture

Filter-partitioning dataflow:

→ Filters unicast
WIENNA Architecture

Filter-partitioning dataflow:

- Filters unicast
- Inputs broadcast
WIENNA Architecture

Filter-partitioning dataflow:

→ Filters unicast
→ Inputs broadcast
→ Output computation
WIENNA Architecture

Filter-partitioning dataflow:

- Filters unicast
- Inputs broadcast
- Output computation
- Reduction
Methodology

- The accelerator cost model MAESTRO\(^1\) has been used, which considers latency, bandwidth and multicast characteristics of NoP to estimate performance metrics.
- DNN models: Resnet50 and UNet.
- We consider both conservative (C) and aggressive (A) design points for both the electric baseline and WIENNA.

\(^1\) H. Kwon et al., Understanding Reuse, Performance, and Hardware Cost of DNN Dataflows: A Data-Centric Approach, MICRO 2019
Results: Throughput

- WIENNA improves the end-to-end throughput by 2.7-5.1X on Resnet50 and 2.2-3.8X on UNet.
- WIENNA can achieve better results than interposer with the same relative bandwidth, due to single-cycle broadcast in distribution.
- Different layers require different partitioning and bandwidths → Adaptive
**Results: Energy**

- Average reduction of 38% in energy consumption due to broadcasts and single hop transmissions.
- Multicast opportunities given by partitioning strategies are leveraged.

- More results in the paper…

![Energy Comparison Chart](chart.png)
Results: Overheads

- Memory overhead:
  - Area: 4%
  - Power: 1%

- Chiplet overhead:
  - Area: 16%
  - Power: 25%

<table>
<thead>
<tr>
<th>Component Sub-element</th>
<th>Area (mm²)</th>
<th>Area (%)</th>
<th>Power (mW)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chiplets (256×)</td>
<td>1646</td>
<td>97</td>
<td>89600</td>
<td>89</td>
</tr>
<tr>
<td>PEs (64×) + Mem</td>
<td>5</td>
<td>78</td>
<td>90</td>
<td>26</td>
</tr>
<tr>
<td>Wireless RX</td>
<td>1</td>
<td>16</td>
<td>90</td>
<td>25</td>
</tr>
<tr>
<td>Collection NoP Router</td>
<td>0.43</td>
<td>6</td>
<td>170</td>
<td>49</td>
</tr>
<tr>
<td>Memory (1×)</td>
<td>53</td>
<td>3</td>
<td>10167</td>
<td>11</td>
</tr>
<tr>
<td>Global SRAM</td>
<td>51</td>
<td>96</td>
<td>10000</td>
<td>99</td>
</tr>
<tr>
<td>Wireless TX</td>
<td>2</td>
<td>4</td>
<td>167</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>1699</td>
<td>100</td>
<td>99767</td>
<td>100</td>
</tr>
</tbody>
</table>
Conclusion

- New scalable design methodology of 2.5D DNN accelerators based on wireless NoP technology.

- Dataflow requirements and architecture capabilities considered to reduce energy (up to 75%) and improve throughput (up to 5.1X).

- Reduced area and power overheads.
Acknowledgments

Architecting More Than Moore
Wireless Plasticity for Massive Heterogeneous Computer Architecture

www.wiplash.eu

@Wiplash_Eu

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863337
Dataflow-Architecture Co-Design for 2.5D DNN Accelerators using Wireless Network-on-Package

Robert Guirado (rguirado@ac.upc.edu)
Hyoukjun Kwon, Sergi Abadal, Eduard Alarcón, Tushar Krishna

26th Asia and South Pacific Design Automation Conference (ASP-DAC)
Virtual – January 18-21, 2021