Improving Scalability of Flexible On-chip Interconnect for DL Accelerators Using Logic-on-Logic 3D ICs

Canlin Zhang[1], Gauthaman Murali[1], Sung-Kyu Lim[1], Tushar Krishna[1]

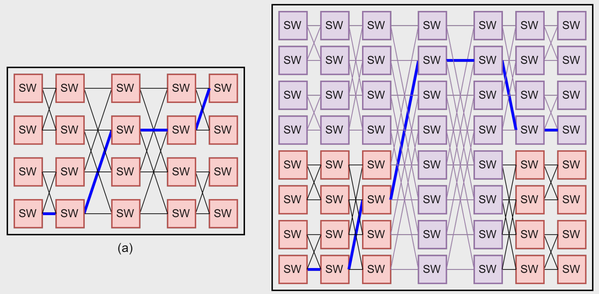
School of Electrical and Computer Engineering [1]

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

Background: Scalability Issues

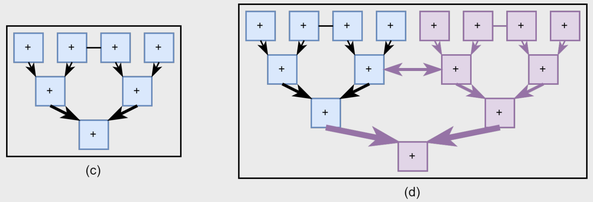
Flexible interconnects are used in DL accelerators to support flexible dataflow. However, two scalability issues prevent accelerator scale-up:

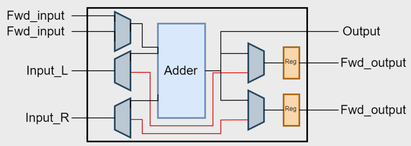
Wirelength & Critical Path



Flexible distribution networks (DN) often feature complex, non-blocking topologies.
Moreover, buffer-less switch designs further exacerbates wirelength issues, worsening timing.

Area & Power



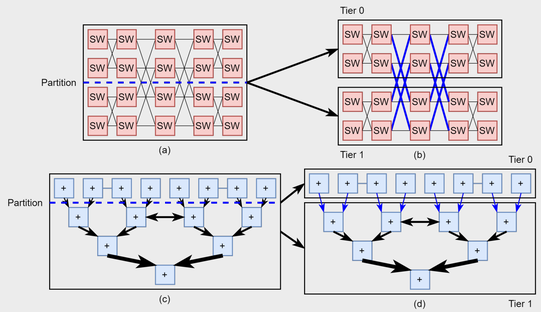


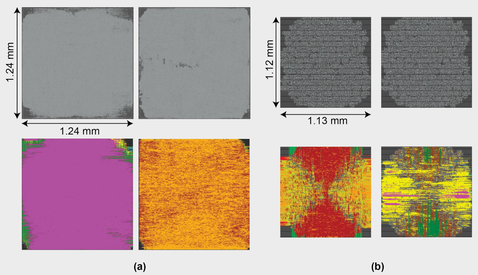
Flexible reduction networks (RN) often feature complex switch designs (Adder + datapath logic), causing area and power scalability issues.

Our Method: Logic-on-Logic 3D

Logic-on-Logic 3D features two logic tiers connected by hybrid bonding. This offers two benefits: Extra area for PNR; Short links between two tiers.

Partition Methodology





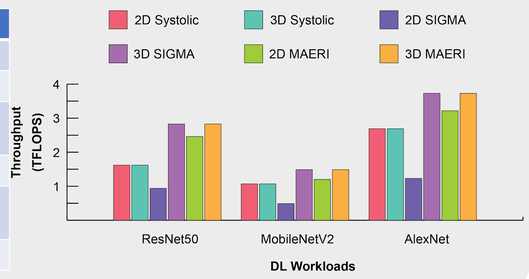
Left: (a) SIGMA DN; (b) MAERI RN Right: (a) PNR of SIGMA DN; (b) PNR of MAERI RN. PDK is TSMC 28nm (3D)

Overall, we partition SIGMA DN (a) and MAERI RN (b) by dividing their topologies.
In general, our methods try to:
1. Maximize the # of links converted to 3D hybrid bonding to improve timing 2. Maintain balance between two logic tiers to minimize footprint.

Results (SIGMA DN)

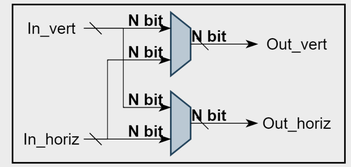
For 2D baseline, we observer poor scalability in wirelength and footprint. There is also a significant drop in design frequency.
In contrast, our 3D design achieves much better scalability in footprint. We also observe a significant drop in wirelength, leading to 3x increase of design frequency.
This increase in frequency lead to significant throughput increase.

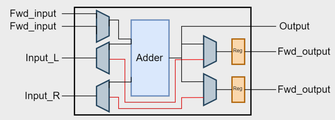
Metrics	2D		3D	
# PEs	256	1024	256	1024
Footprint	0.32 mm2	**3.04 mm2**	0.17 mm2	**1.54 mm2**
# Metal Layers	6	**8**	6 + 6	8 + 6
Wirelength	9.7 m	**132.9 m**	7.9 m	**104.8 m**
Design Freq	1.8 GHz	**0.66 GHz**	2 GHz	2 GHz
Chip Power	0.86 W	**2.88 W**	0.79 W	**5.7 W**



Implications and Future Work

We could classify flexible interconnects into two groups: Wire-heavy and Logic-heavy.
1. Wire-heavy designs (e.g., SIGMA DN) have less buffer and more connections. We observe logic-on-logic 3D to have drastic improvement in frequency and wirelength.
2. Logic-heavy designs (e.g., MAERI RN) have more buffer and combinational logic. We observe less but still significant improvement in area and wirelength.
Takeaway: Different architectures have different response to 3D IC.





Top: SIGMA DN Switch; Bottom: MAERI RN Switch

Metrics	3D MAERI RN		3D SIGMA DN	
# PEs	256	1024	256	1024
Footprint	0.32 mm2	**1.26 mm2**	0.17 mm2	**1.54 mm2**
# Metal Layers	6 + 6	6 + 6	6 + 6	**8 + 6**
Wirelength	3.63 m	**27.1 m**	7.9 m	**104.8 m**
Design Freq	2 GHz	2 GHz	2 GHz	2 GHz
Chip Power	1.06 W	4.09 W	0.79 W	5.7 W

Future Work

We are actively developing new partitioning techniques beside topology partitioning. Moreover, we are expanding the scope of our work to full-accelerator designs.