UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Georgia Tech

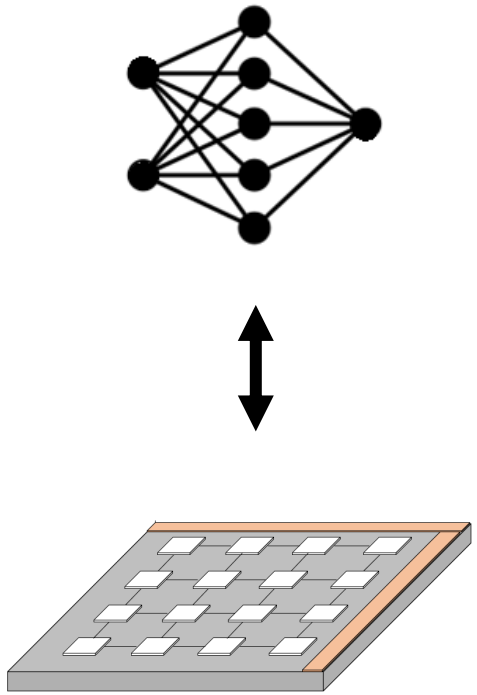
Dataflow-Architecture Co-Design for 2.5D DNN Accelerators using Wireless Network-on-Package

Robert Guirado (rguirado@ac.upc.edu)

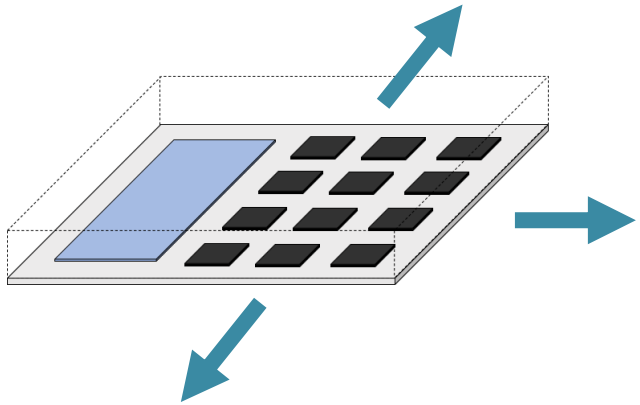
Hyoukjun Kwon, Sergi Abadal, Eduard Alarcón, Tushar Krishna

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

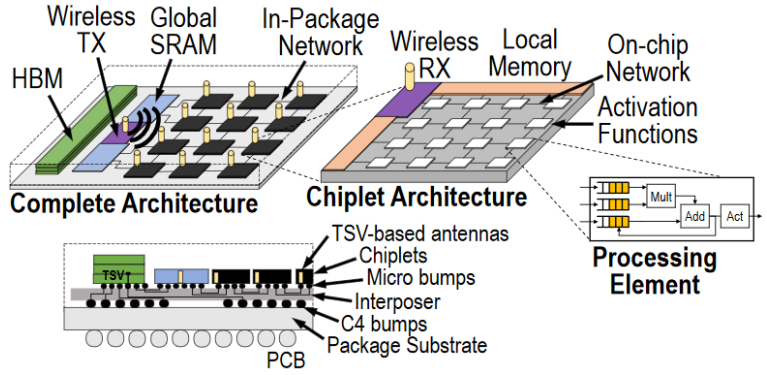
Overview



Problem: Electrical scale-out is bandwidth-limited

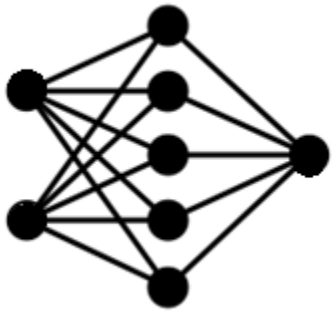


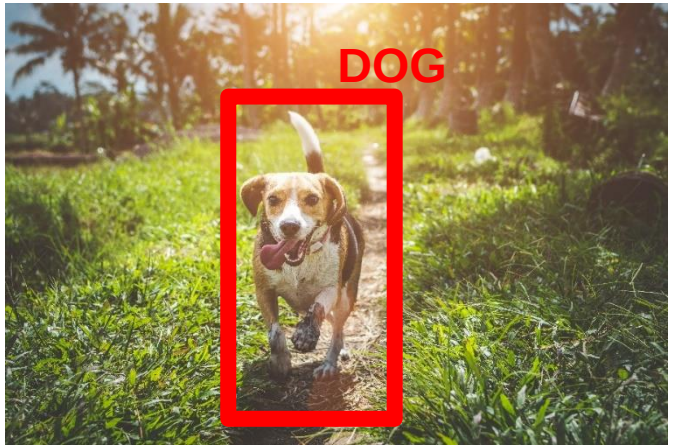
WIENNA solves this via WNoP technology

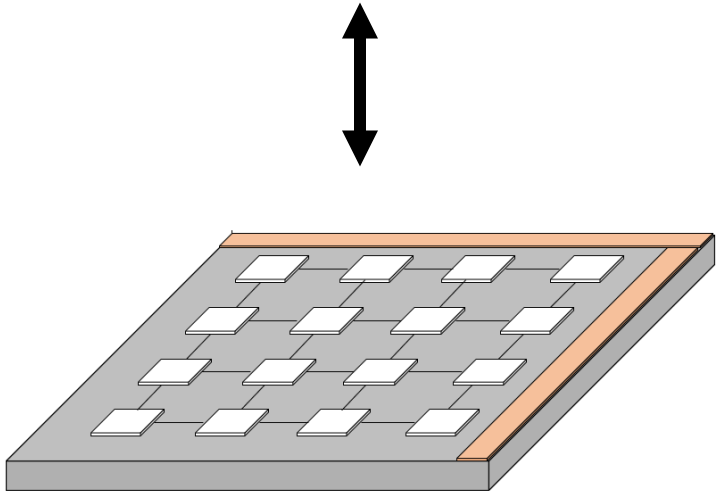


WIENNA: Up to 5.1X speedup and 38% energy savings

Background: DNN



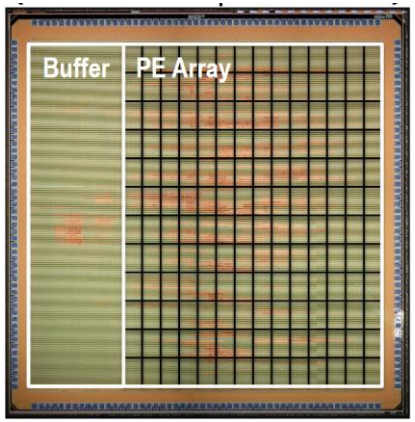




"DeepPose" A.Toshev et al. CVPR 2014

Background: DNN Acceleration

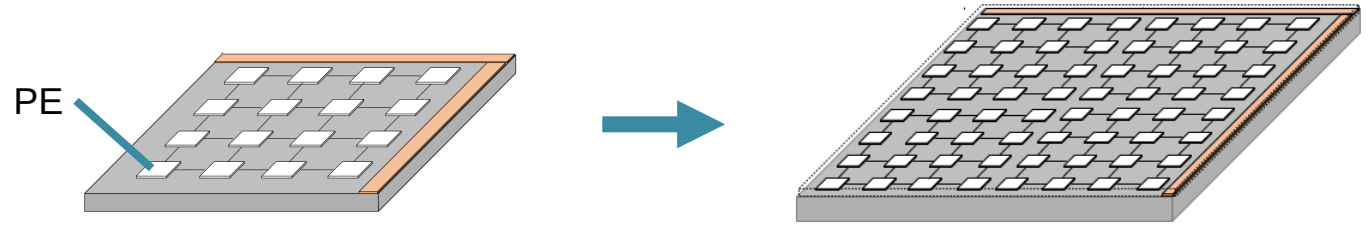
- Higher throughput and energy-efficiency than GPUs and CPUs via parallelism and dedicated circuitry.
- Off-chip memory + global shared memory + array of PEs connected via a Network-on-Chip (NoC)
- Specific dataflows will define:
  - Data partitioning
  - Data movement
  - Data reuse



"Eyeriss" Chen, Y et al.

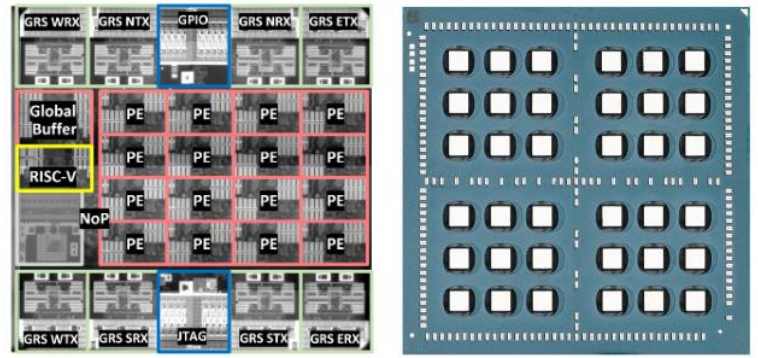
Background: DNN Acceleration

- Two approaches to increase computing power:
  - Scale-up → more PEs



  - Scale-out → more chiplets

- 2.5D chiplet integration enables efficient scale-out

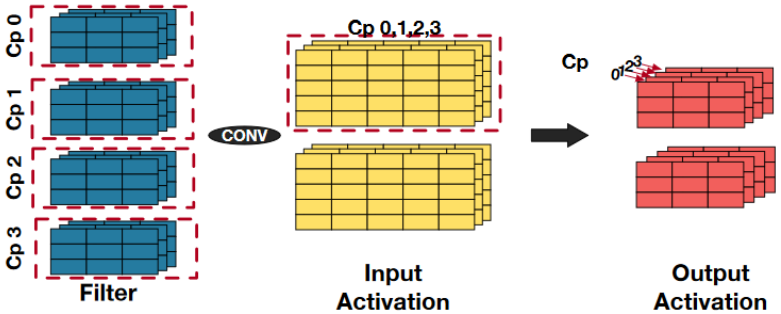


(a) Simba chiplet (b) Simba package

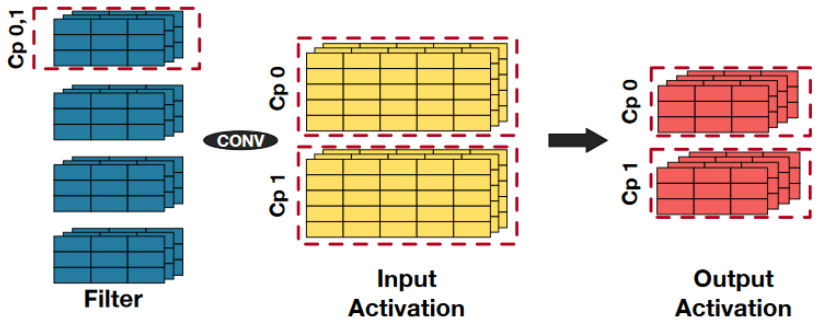
"Simba" Shao, YS et al.

Baseline dataflows

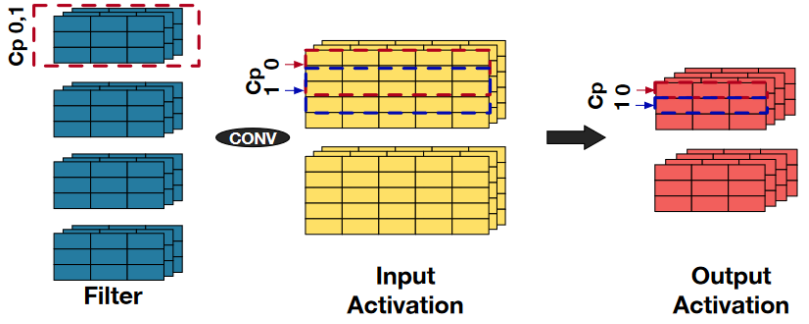
Dataflows: DNN mapping strategies for leveraging data movement



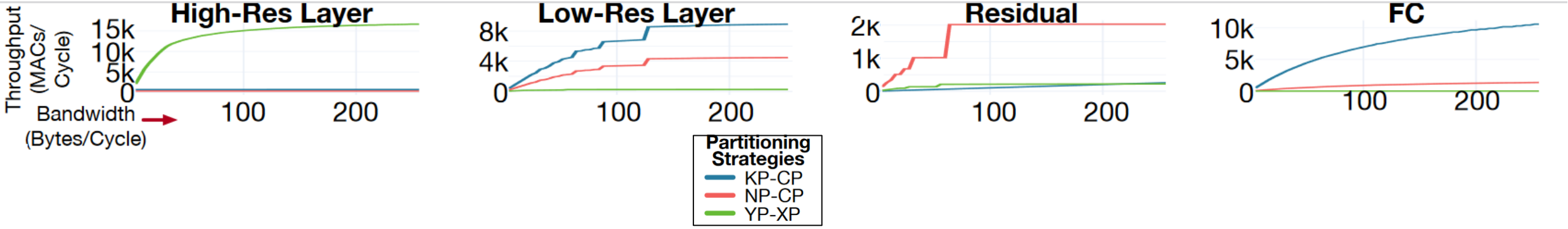
Filter-partitioning (KP-CP)



Batch-partitioning (NP-CP)

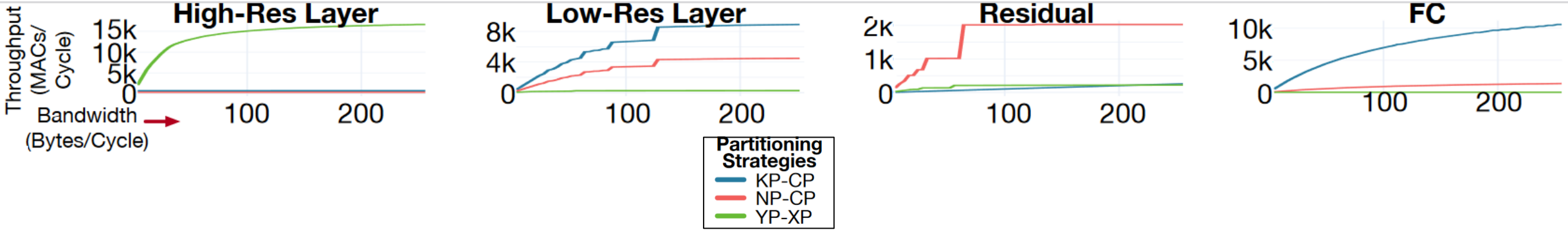


Activation-partitioning (YP-XP)



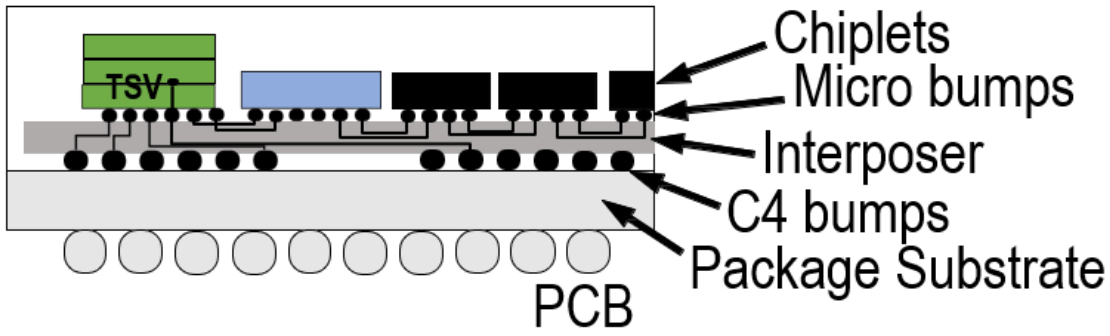
Observations

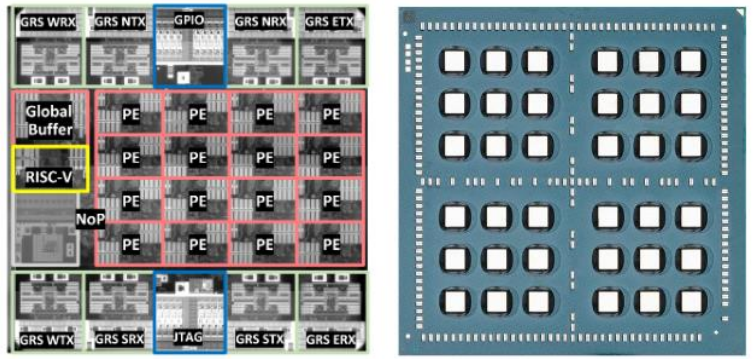
- Different layer types favour different partition strategies.
- Different layer types saturate to peak throughput at different bandwidth values.
- The communication fabric for data distribution plays a key role in performance.
- Broadcast support and high-bandwidth are critical for scalability.
- Supporting 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.



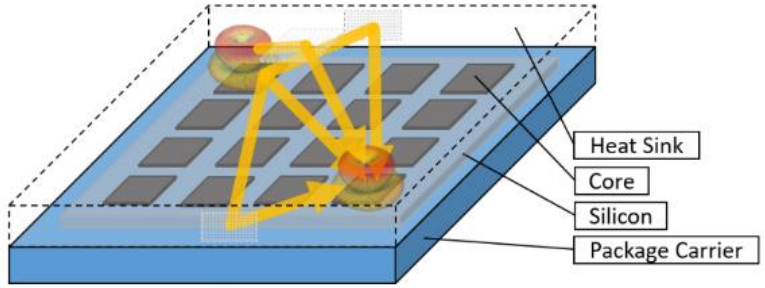


(a) Simba chiplet (b) Simba package

"Simba" Shao, YS et al.

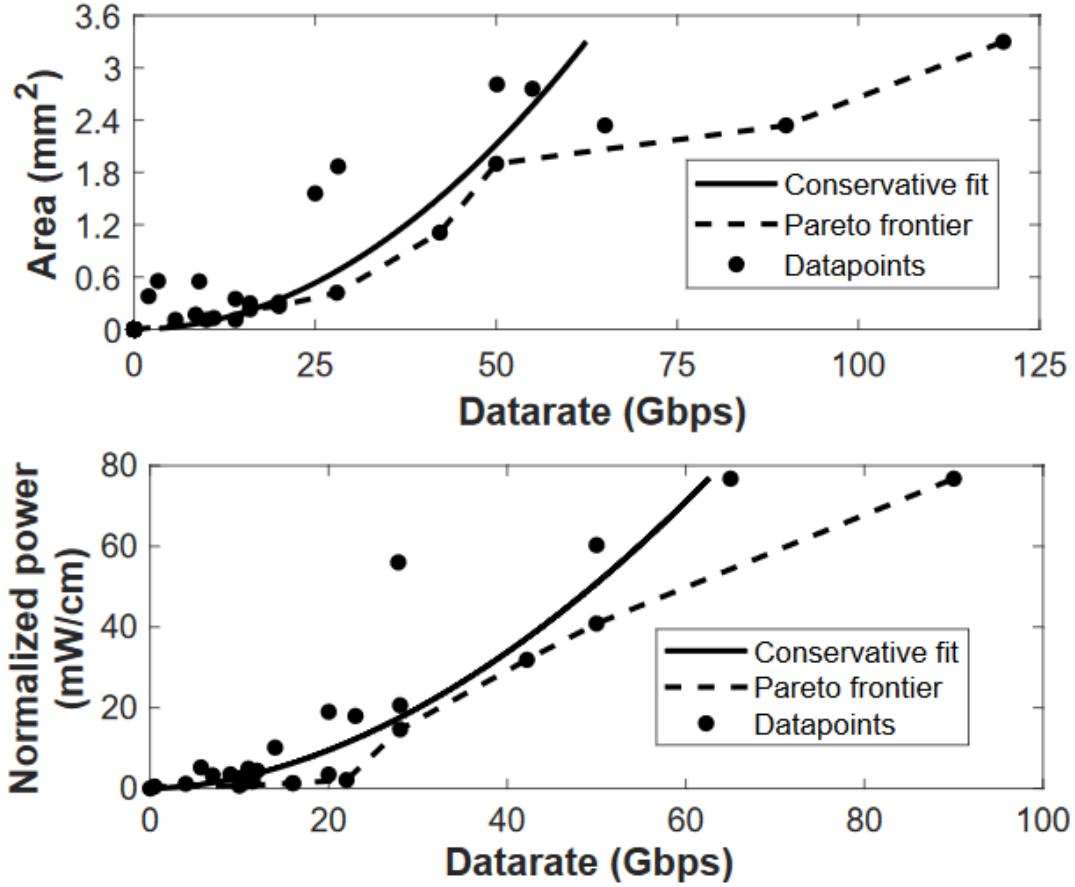
Wireless Network-on-Package

- 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



Wireless Network-on-Package

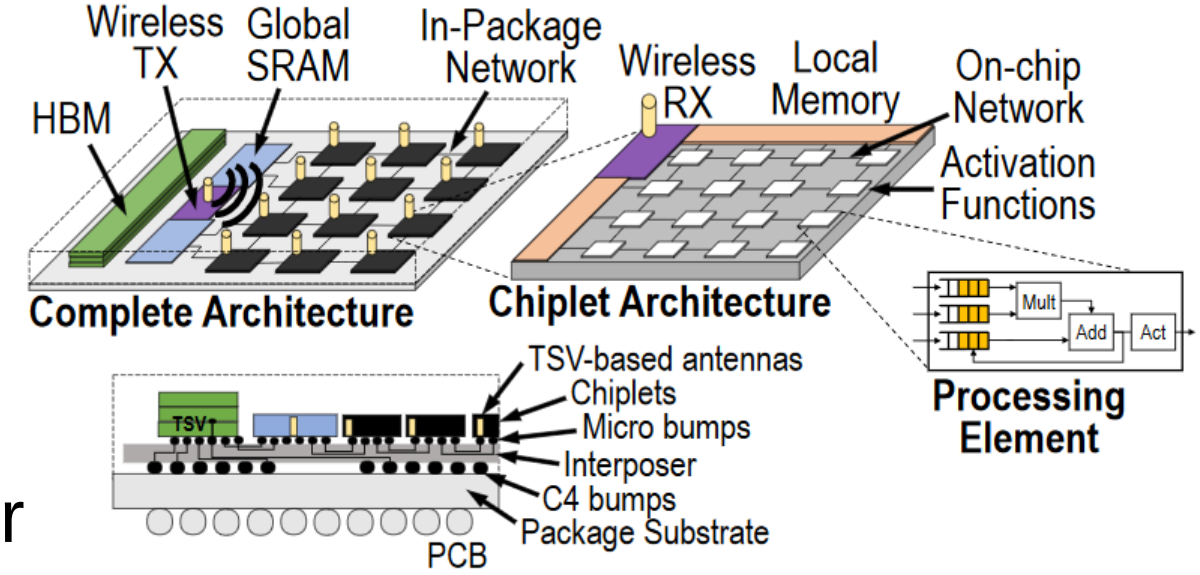
- State-of-the-art transceivers reach up to 100 Gbps
- We extrapolate power and area trends based on 70+ short-range TRXs with different modulations and technologies



X. Yu, J. Baylon, P. Wettin, et al., "Architecture and Design of Multichannel Millimeter-Wave Wireless NoC," in IEEE Design & Test, vol. 31, no. 6, pp. 19-28, Dec. 2014.

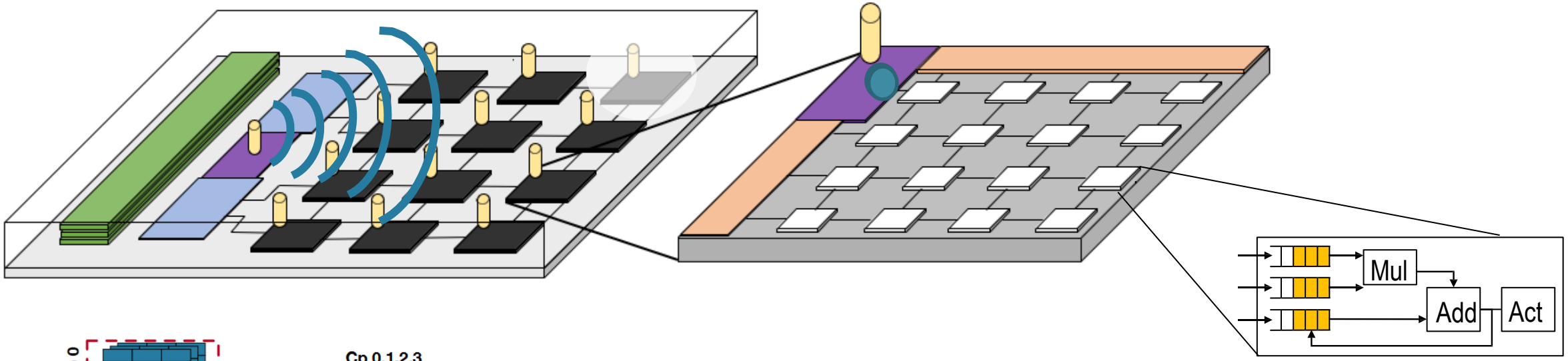
WIENNA Architecture

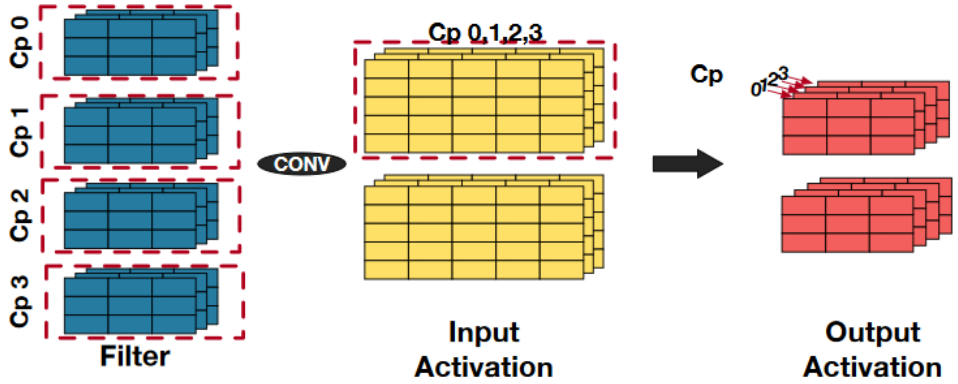
- Two-level hierarchy:
  - HBM, SRAM, chiplet array, NoP
  - Within-chiplet architecture
- Wireless distribution
  - Broadcast enabled
  - SRAM (Tx) → Chiplets (Rx)
- Wired collection through interposer
- 256 chiplets and 64 PEs per chiplet.



WIENNA Architecture

Filter-partitioning dataflow:

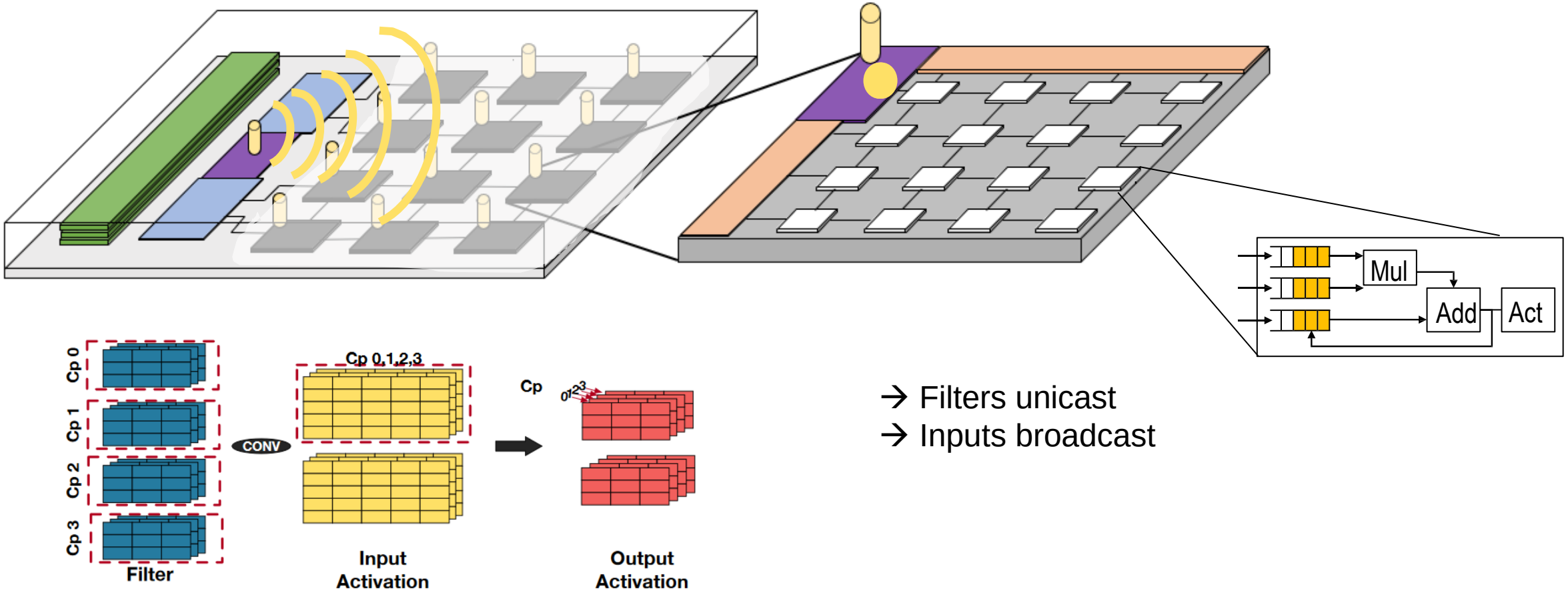




→ Filters unicast

WIENNA Architecture

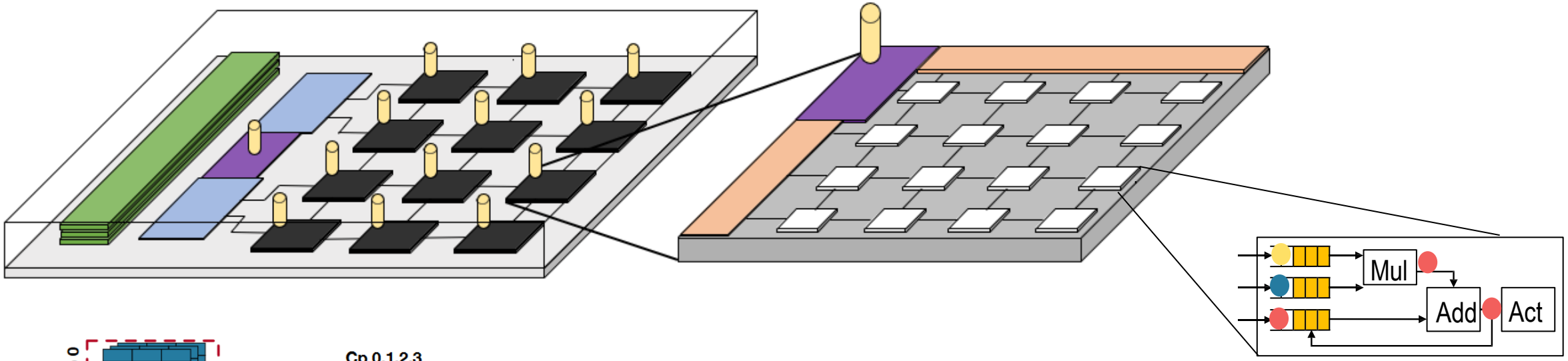
Filter-partitioning dataflow:

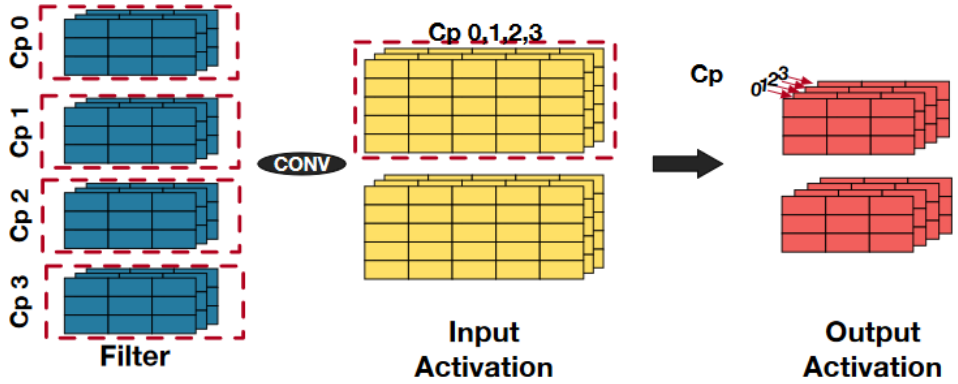


→ Filters unicast
→ Inputs broadcast

WIENNA Architecture

Filter-partitioning dataflow:

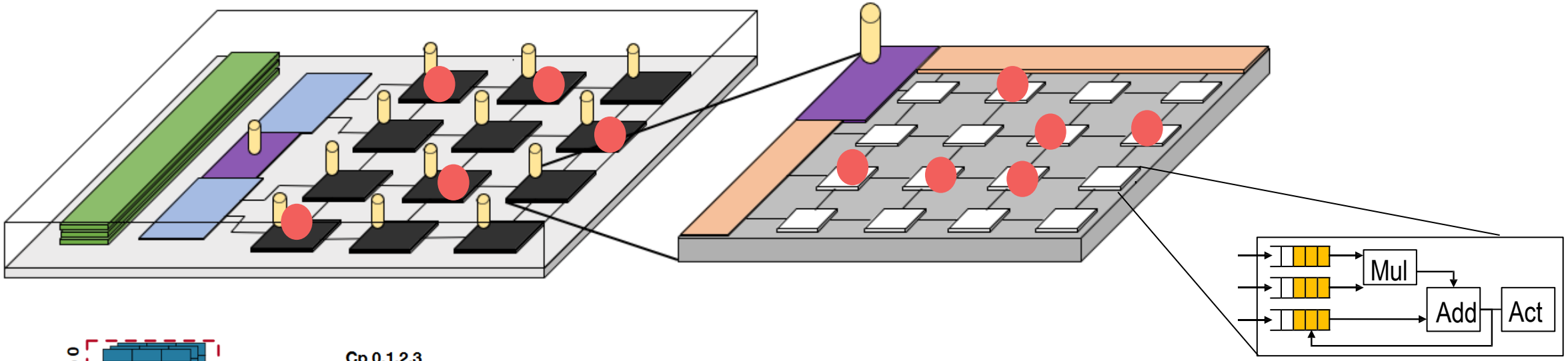


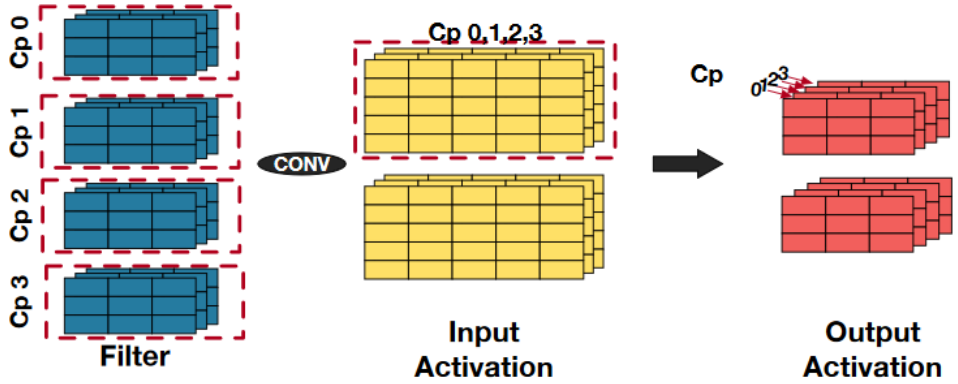


→ Filters unicast
→ Inputs broadcast
→ Output computation

WIENNA Architecture

Filter-partitioning dataflow:





→ Filters unicast
→ Inputs broadcast
→ Output computation
→ Reduction

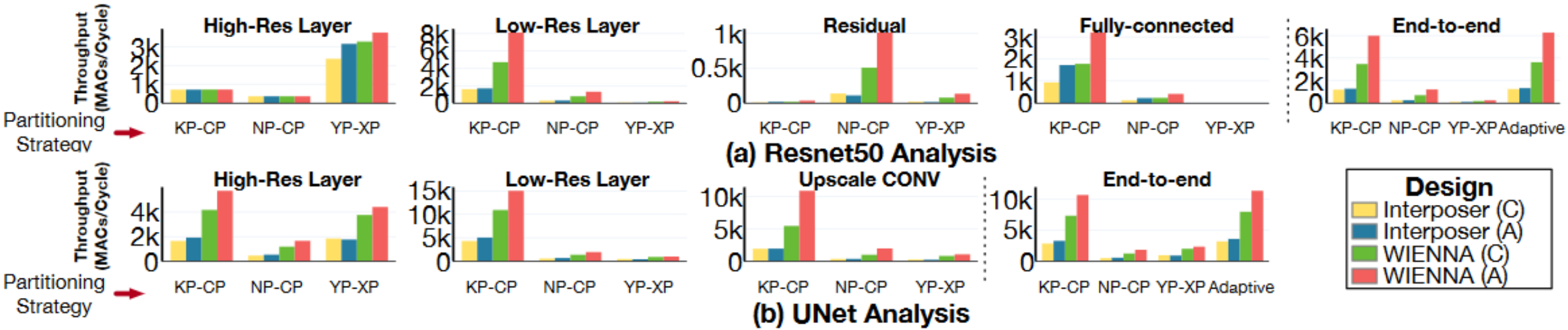
Methodology

- 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

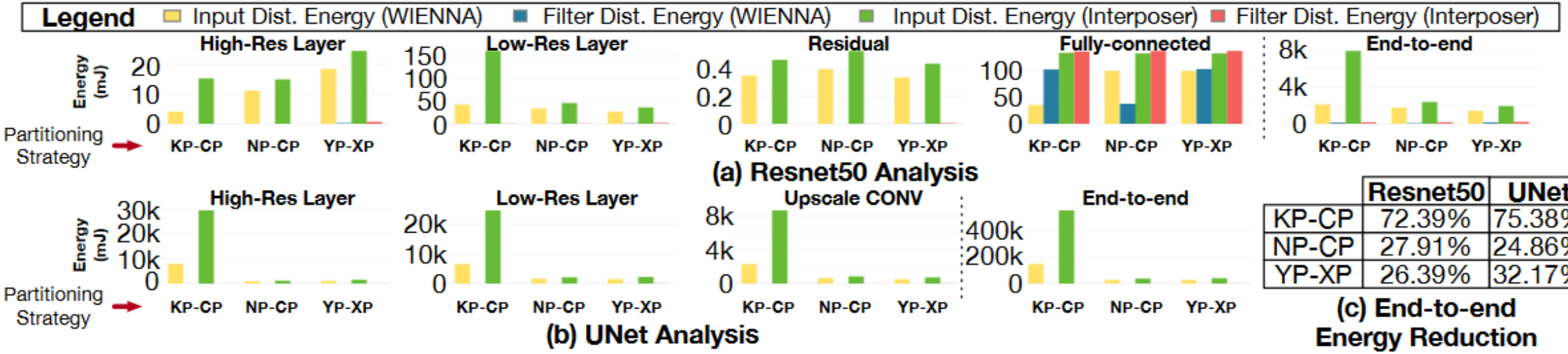
Results: Throughput

- 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.



	Resnet50	UNet
KP-CP	72.39%	75.38%
NP-CP	27.91%	24.86%
YP-XP	26.39%	32.17%

- More results in the paper…

Results: Overheads

- Memory overhead:
  - Area: 4%
  - Power: 1%
- Chiplet overhead:
  - Area: 16%
  - Power: 25%

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

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

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Georgia Tech

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