A New Heterogeneous Chiplet-Based Architecture for AI Computing

Yu (Kevin) Cao

School of Electrical, Energy and Computer Engineering

Arizona State University

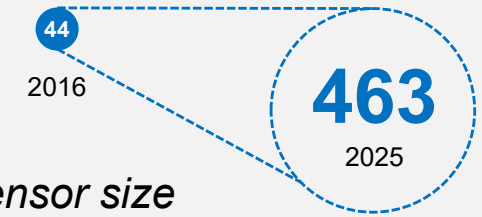
Heterogeneous In-Memory Computing (IMC)

- Challenges and needs of large-scale IMC
  - Robustness, peripheral circuits and **interconnection**
- Chiplet-based benchmark tool: SIAM
- Heterogeneous IMC with 2.5D/3D chiplet
  - Big-little chiplets for efficiency
  - IMC chiplets for 3D sensing
- Summary and future perspectives

Everything Goes UP

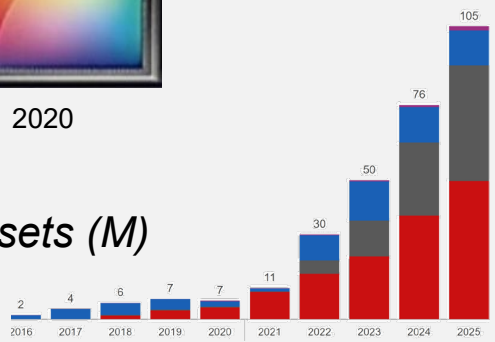
- From data volume to information processing algorithms

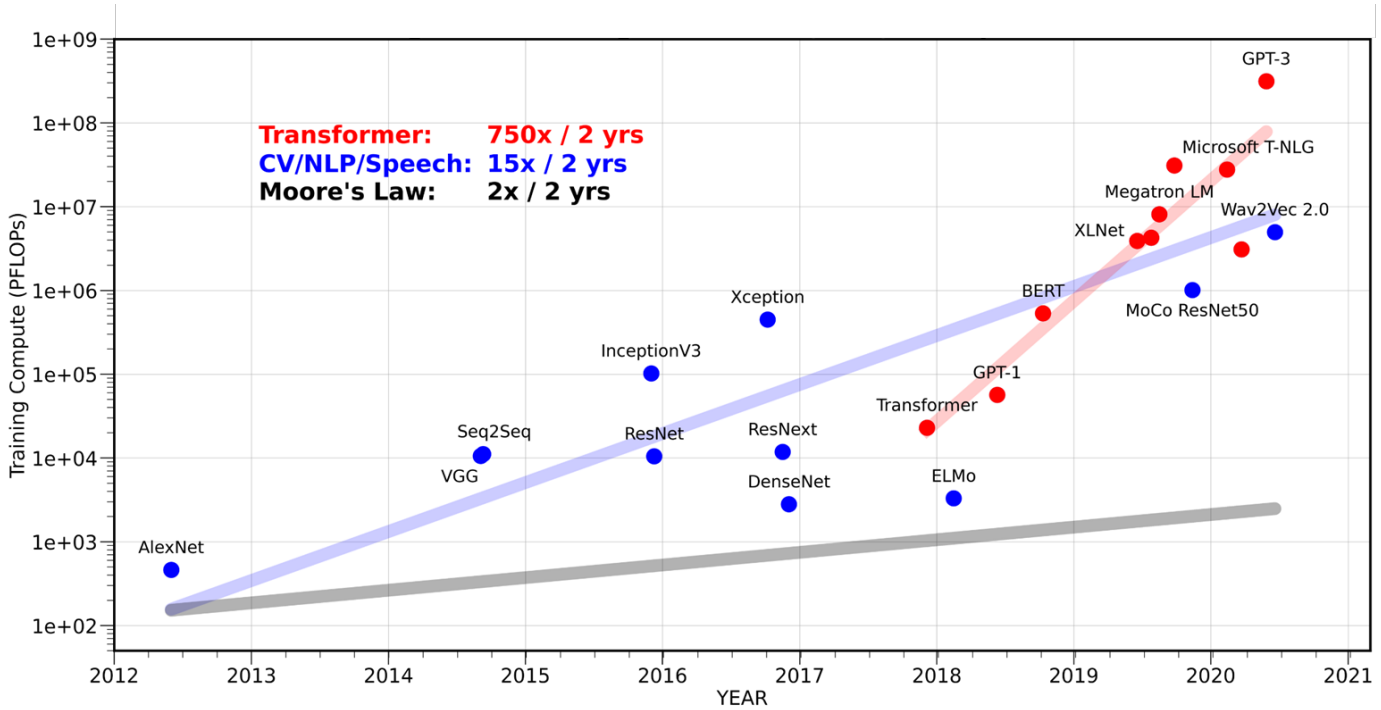
Digital data created: Billion GB/Day



Phone image sensor size

Global XR headsets (M)

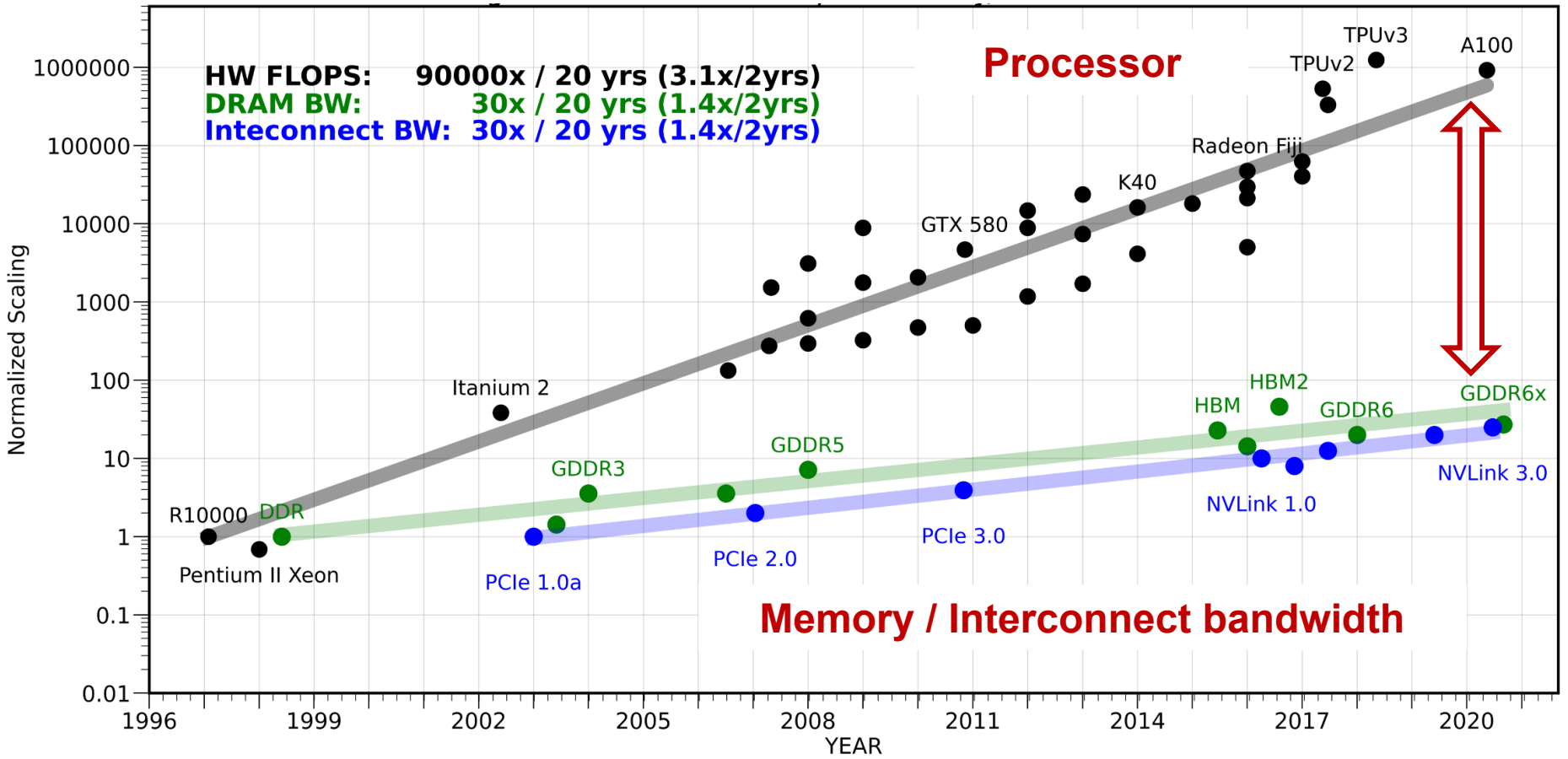


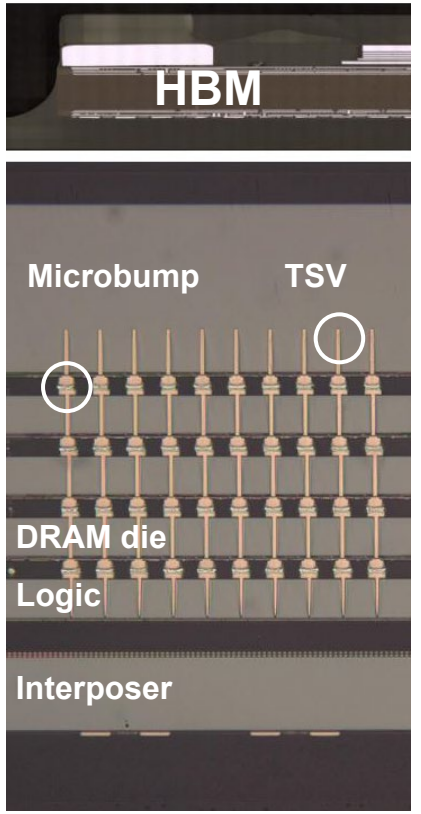


[Micro Focus; Counterpoint, 2021; A. Gholami, 2020]

Memory Access

- Compute is only part of the performance picture

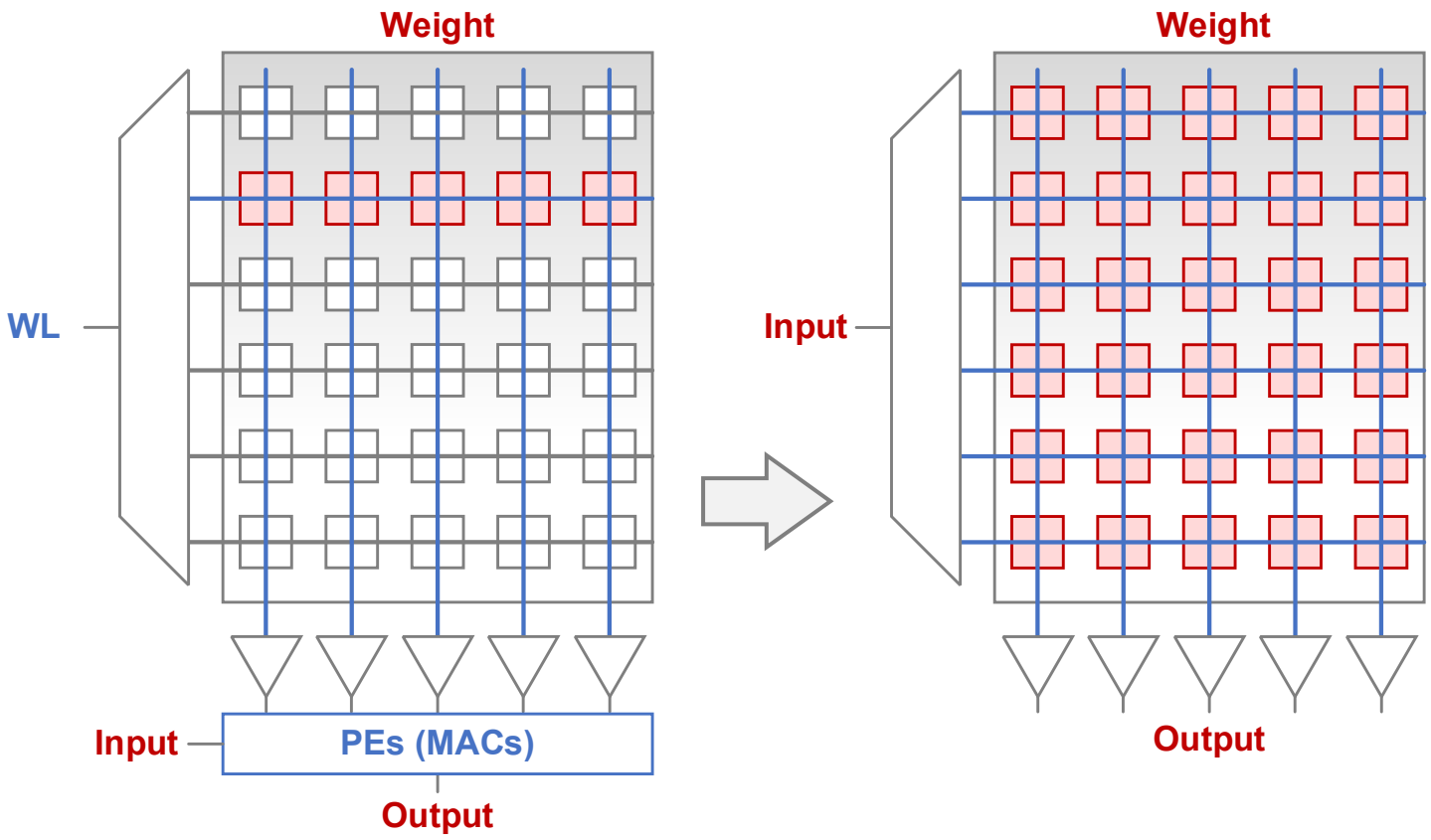




[A. Gholami, 2020; Hynix]

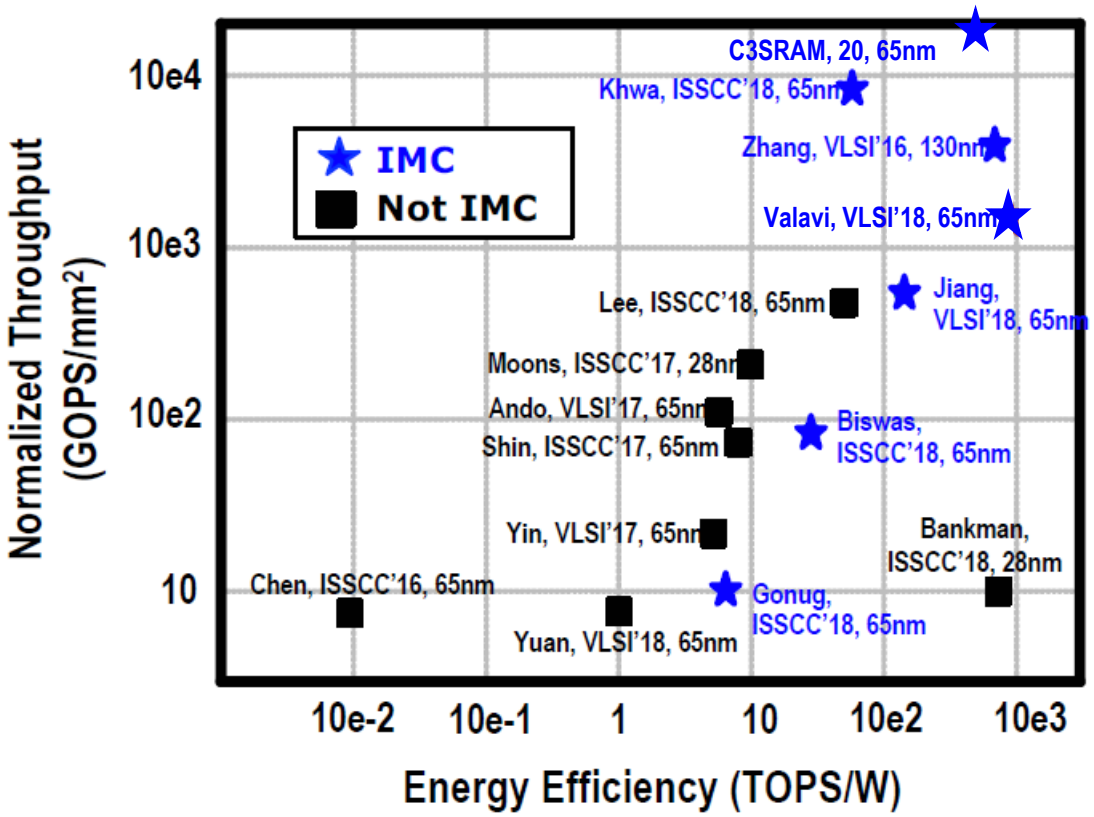
In-Memory Computing

- IMC combines memory access and computation into a single unit
  - 70-90% AI computing is Multiply-Accumulate (MAC)
- Analog computing: resistive (current) or capacitive (charge)
  - Digital IMC is also under research
- Diverse cells: CMOS, NVM (e.g., RRAM) and others
- Ideally, all weights stored on chip

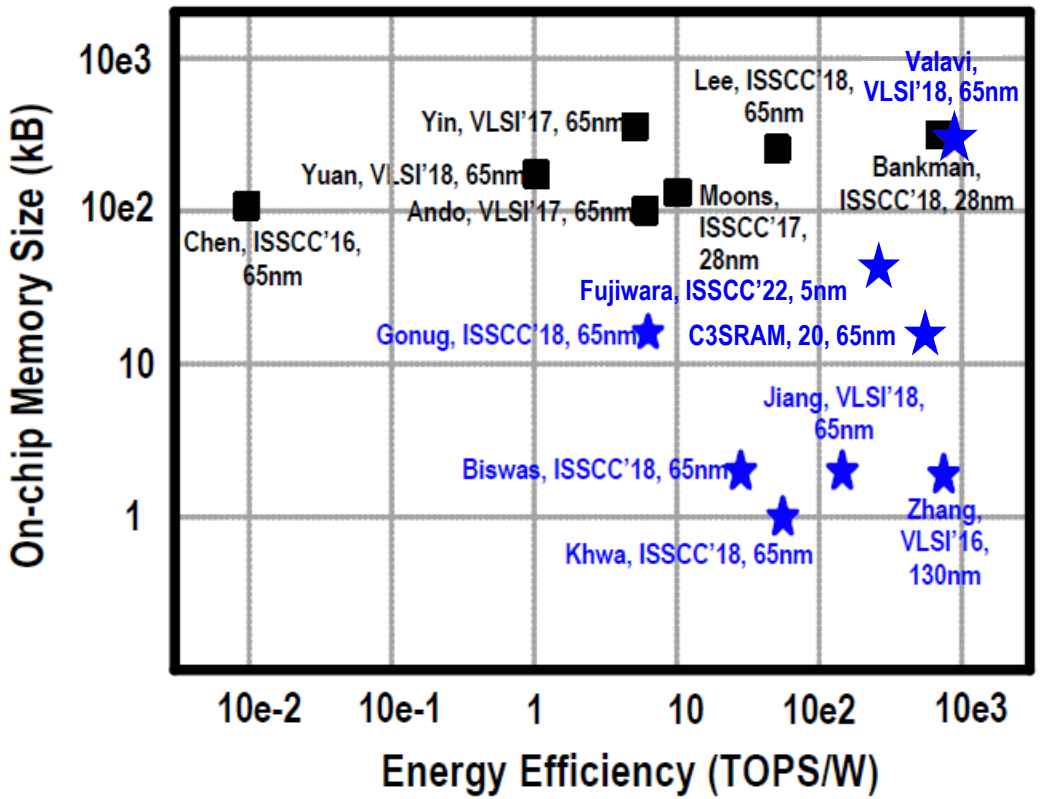


Promises and Challenges

- Potential for **10-100X** higher energy efficiency and throughput



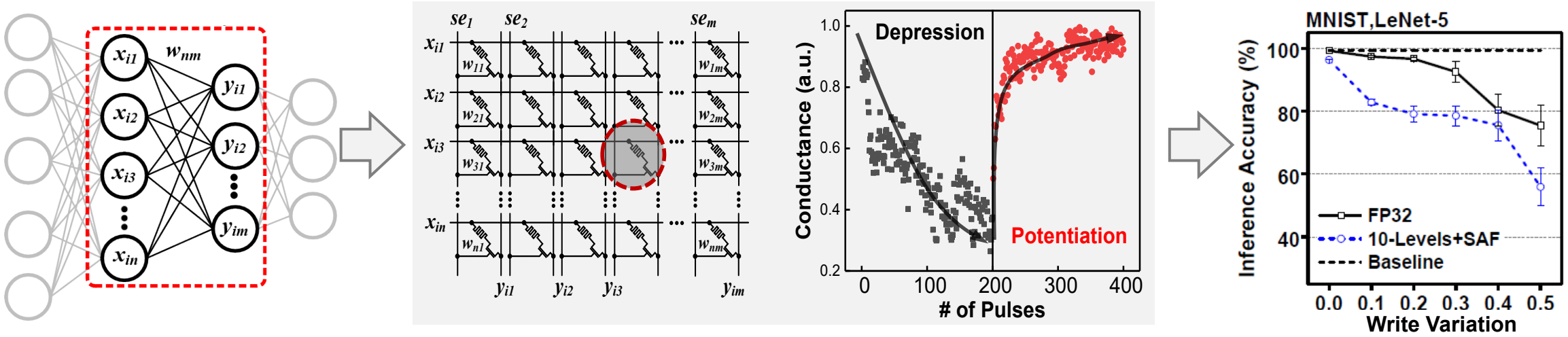
- **Limited scale**: robustness, peripheral circuits, and interconnection



[N. Verma, ISSCC 2019]

Robustness in the IMC Tile

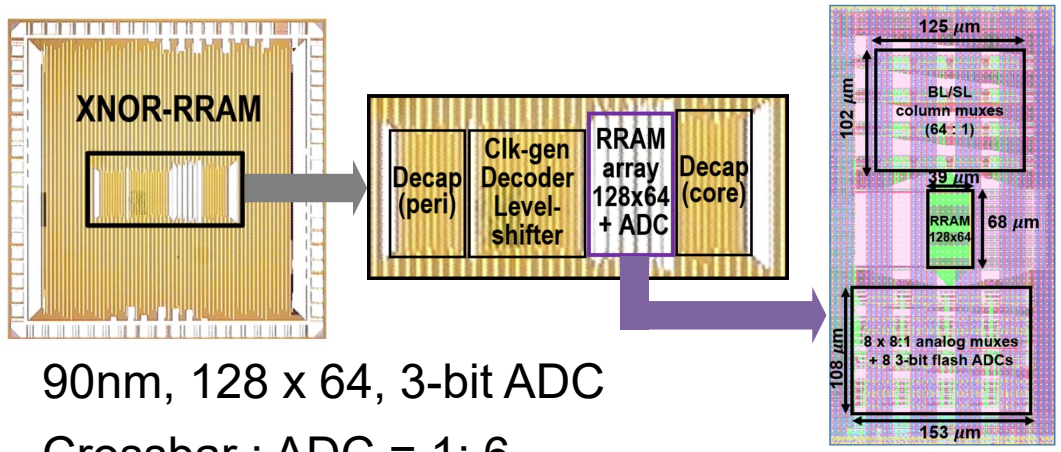
- A fundamental issue in analog computing



- SRAM based:
  - Nonlinearity
  - Device variations
  - Circuit mismatch and parasitics
  - Temperature dependence
- RRAM based:
  - Limited R_{off}/R_{on} (levels)
  - Process variations
  - Stuck-at faults, retention and endurance
  - Parasitics in the crossbar

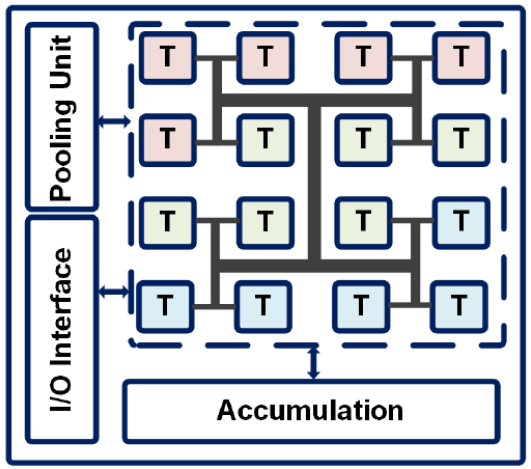
Circuits and Interconnection

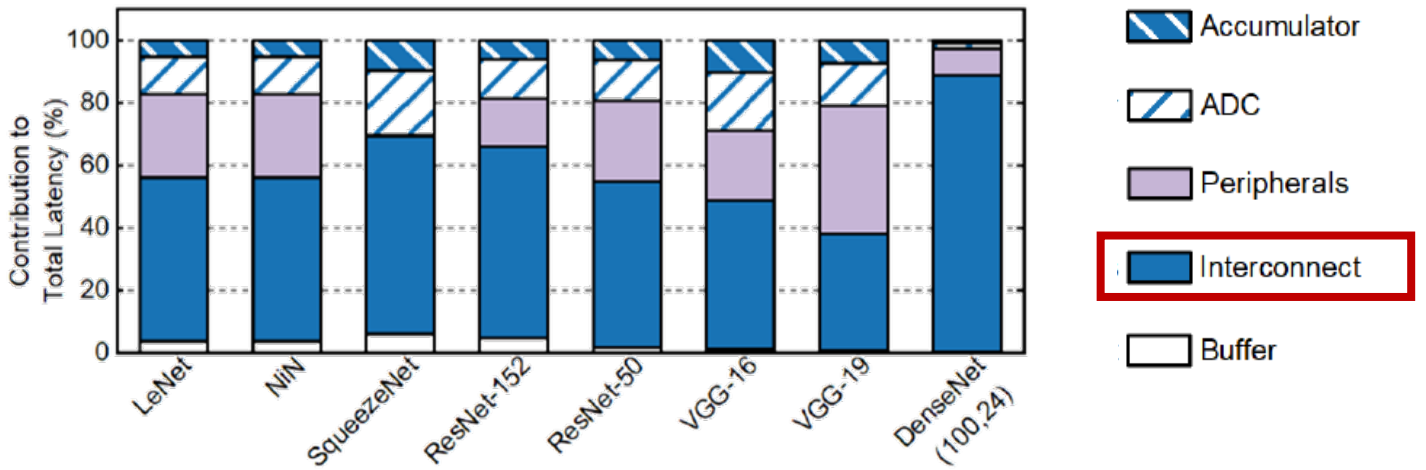
- **Peripheral** circuits (ADC, buffer, adder, scaler, etc.) dominate the area and power consumption
- **Interconnection** and **die cost**: Limiting factors to a monolithic design for large-scale AI computing



90nm, 128 x 64, 3-bit ADC

Crossbar : ADC = 1: 6

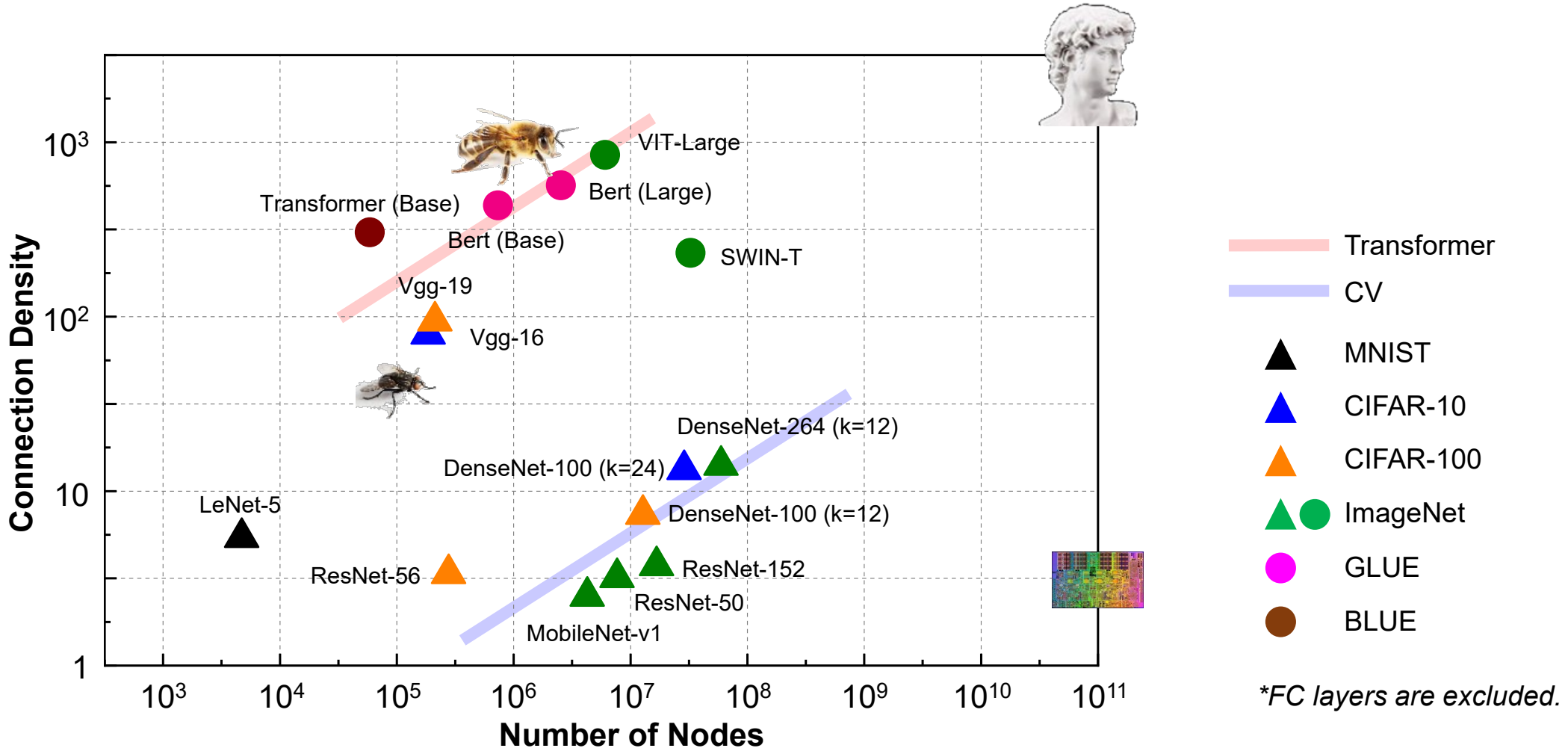




[S. Yin, Micro 2019; G. Krishnan, JxCDC 2020]

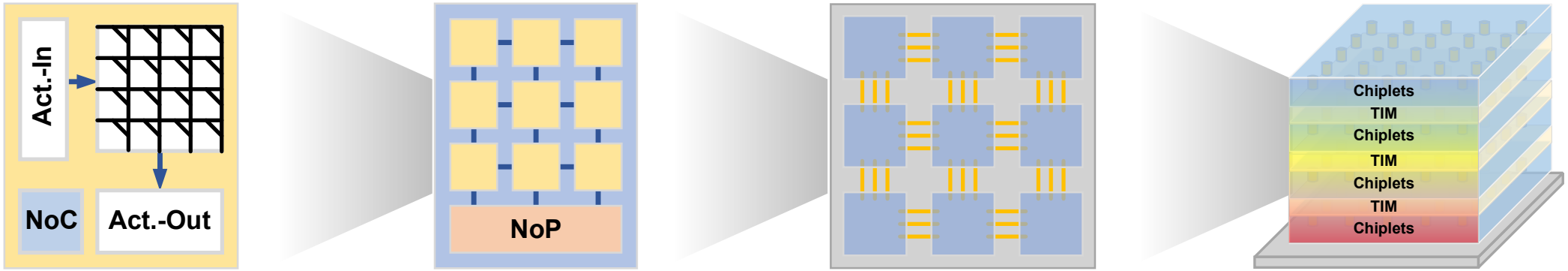
Interconnection in AI Algorithms

- A higher connection density improves the efficiency and learning capability



Toward 2.5D/3D Heterogeneous Integration (HI)

- "Another direction of improvement of computing power is to make physical machines three-dimensional." – Richard P. Feynman, 1985



Computing Tile **Chiplet** **2.5D/3D HI System**

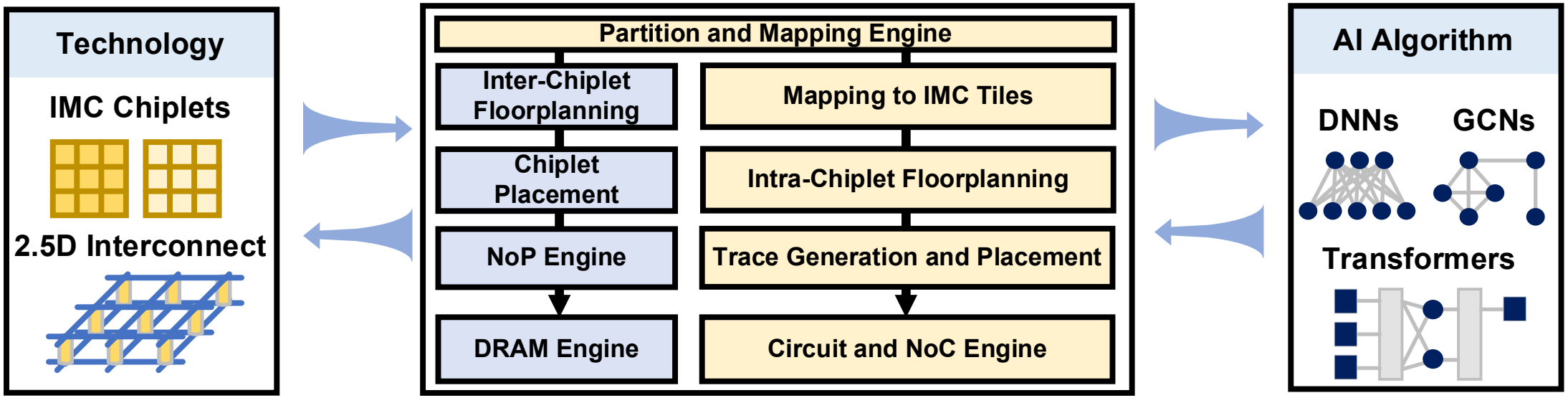
Tile	Digital	Analog
Robustness	✓	
Density		✓
Periphery	✓	
Throughput		✓

Chiplet	Big	Little
Model size	✓	
NoP channels	✓	
Utilization		✓
Energy		✓

Interconnect	NoC	NoP
Bandwidth	✓	
Latency	✓	
Energy	✓	
Scalability		✓

Chiplet-based Performance Benchmarking

- Scalable In-memory Acceleration with Mesh (**SIAM**)
- Cross-layer: Device, circuits, chiplet-based architecture and algorithms
- Interconnect-centric: On-chip NoC, DDR/HBM with memory, NoP for chiplets

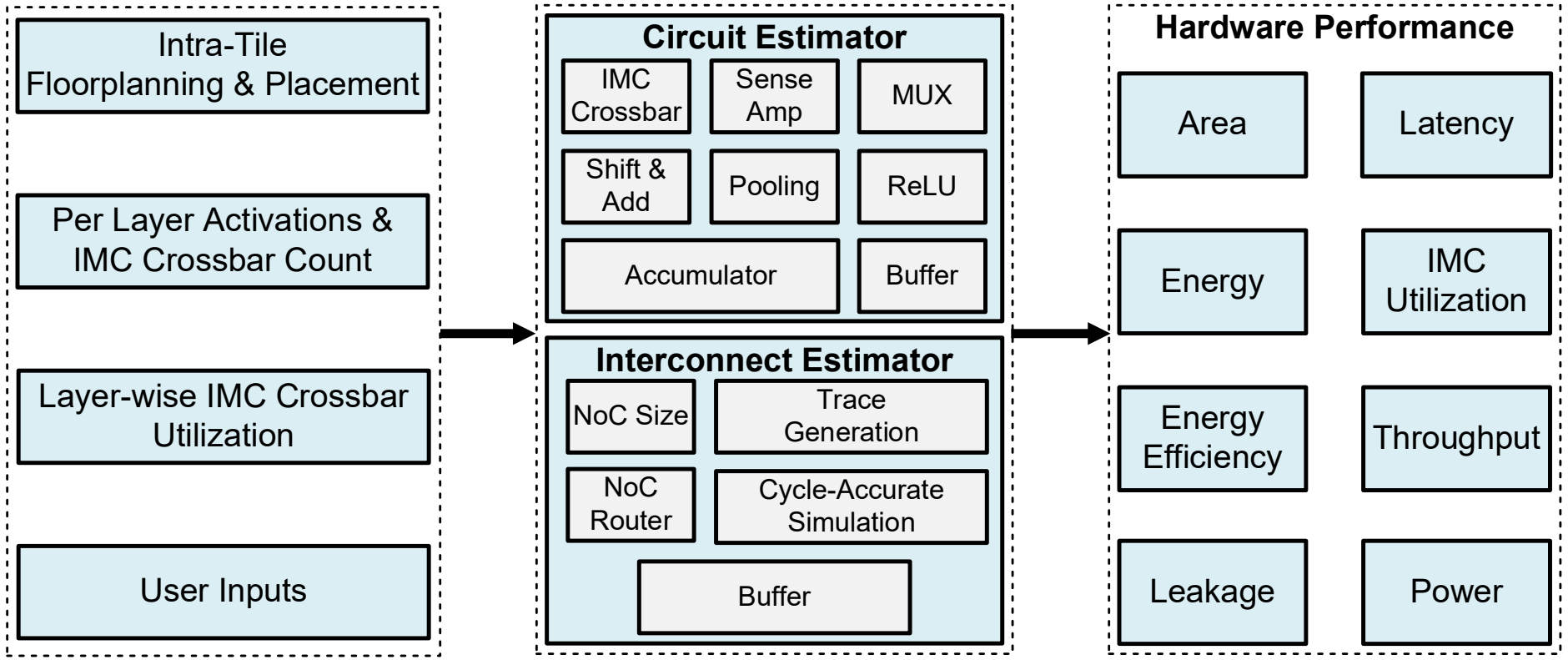


github.com/gkrish19/SIAM-Chiplet-based-Scalable-In-Memory-Acceleration-with-Mesh-for-Deep-Neural-Networks

[G. Krishnan, ESWeek 2021]

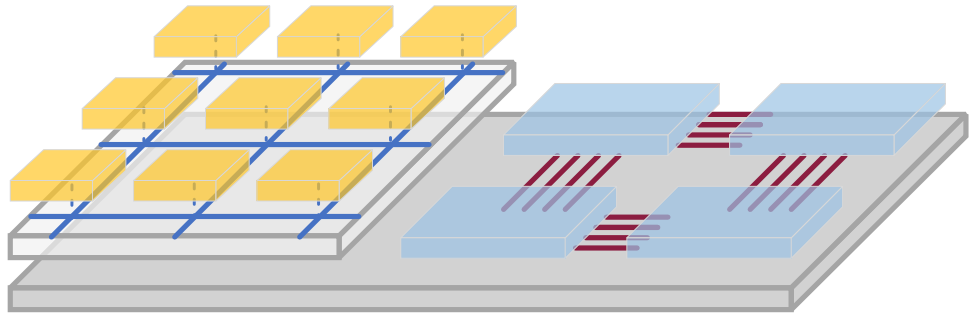
On-chip Circuit and NoC Engine

- SPICE and behavioral models for on-chip tiles and NoC circuits
  - Tiles can be homogeneous or customized
  - Network topologies include NoC-mesh, NoC-tree and H-tree

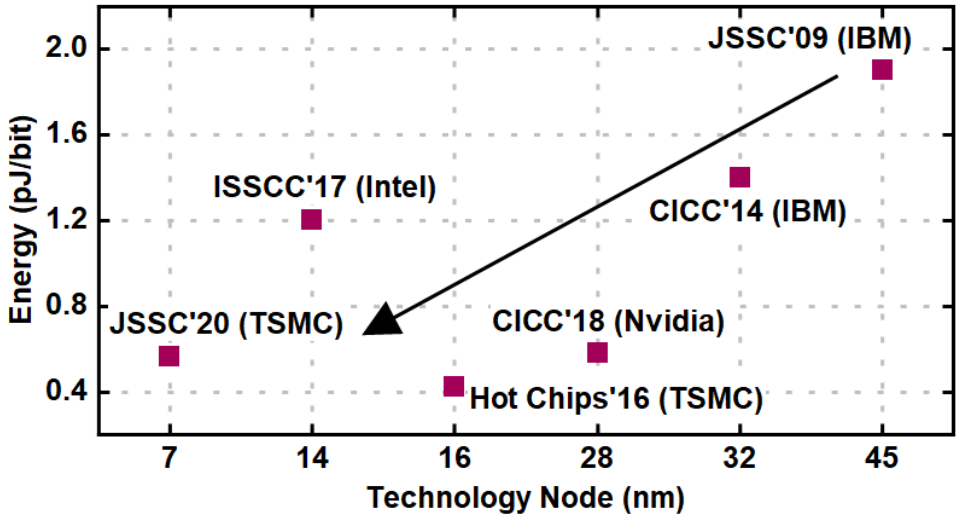


Network-on-Package and DRAM Engines

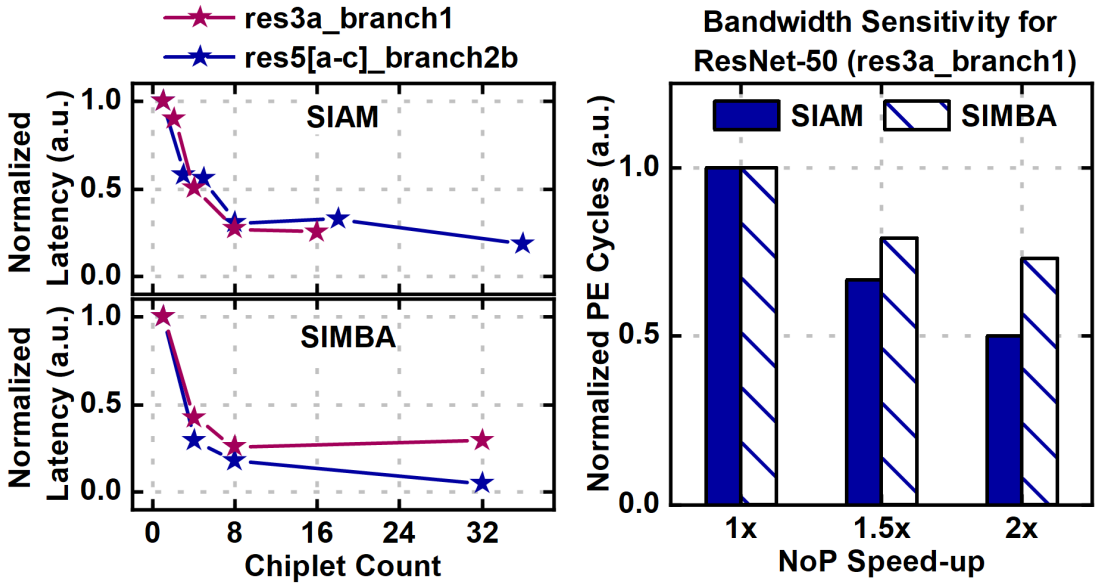
- Channel RLC parasitics calculated with PTM
- Driver cost extracted from AIB or Si data
- DRAM engine: DDR3, DDR4 and HBM
- On-going: SIAM for 3D integration



Router cost from Si data



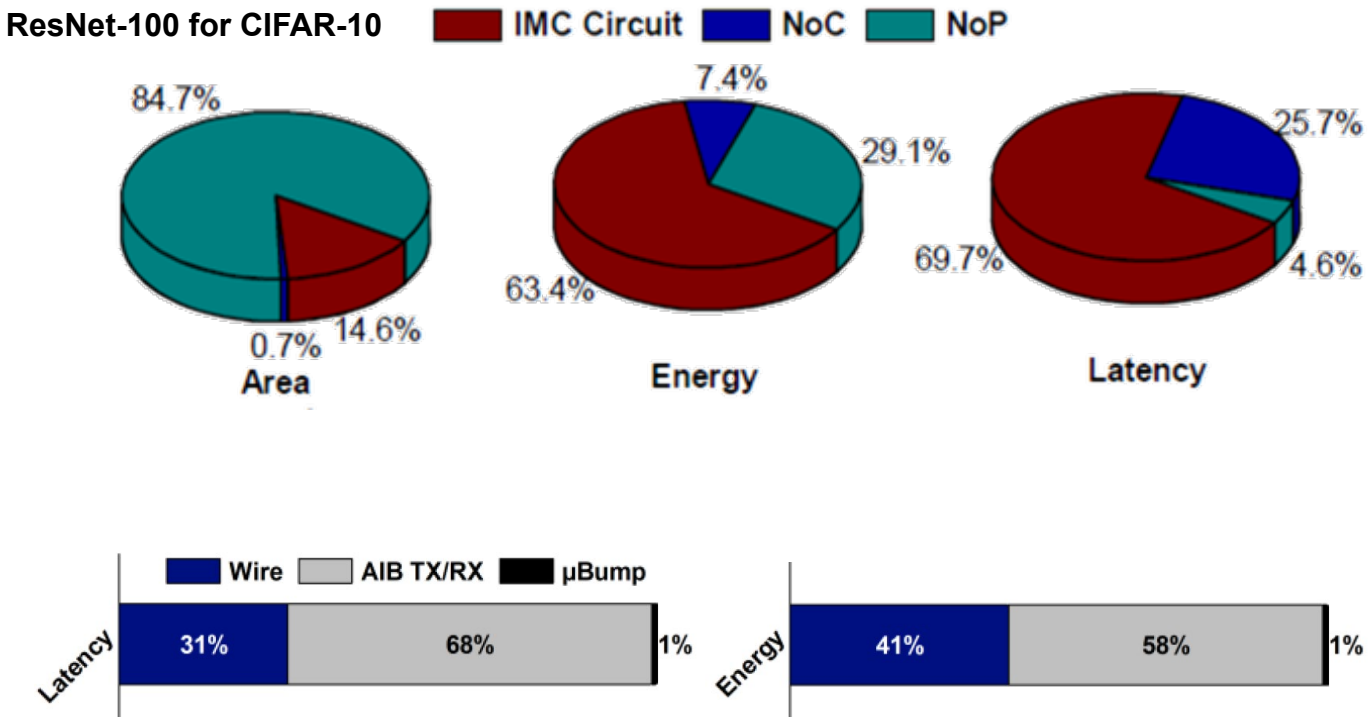
Calibration with SIMBA

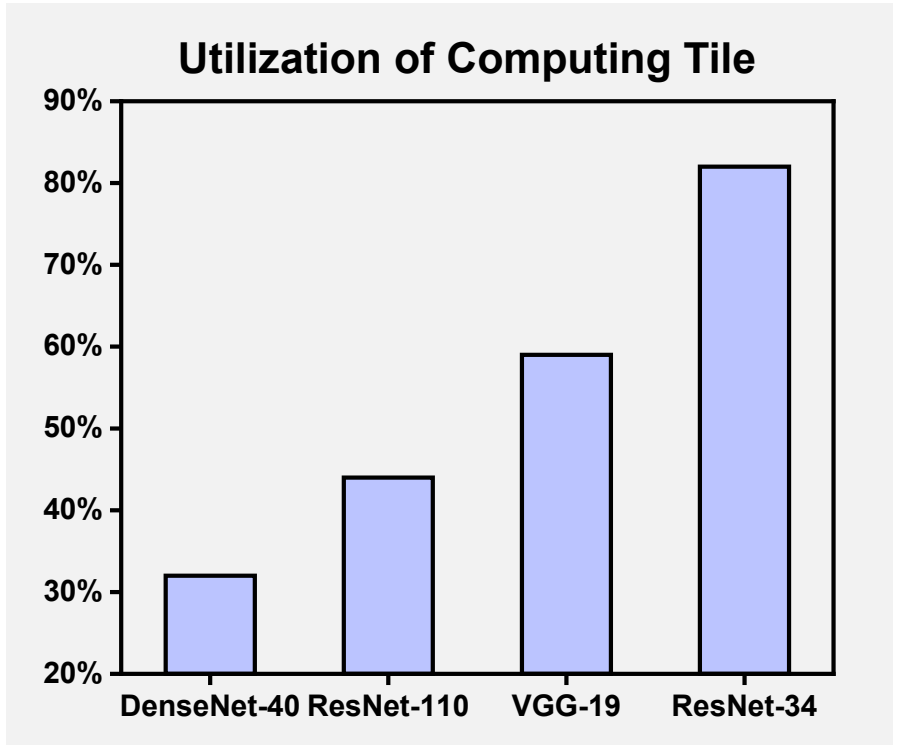


[Y. Kim 2015; S. Ghose, 2018; Y. S. Shao, et al., 2019]

Homogeneous Mapping of IMC

- Network-on-Package (NoP) is a main contributor to area and energy consumption
  - AIB transceiver/receiver dominates NoP cost
- The utilization rate significantly varies on a fixed tile size (crossbar dimension)



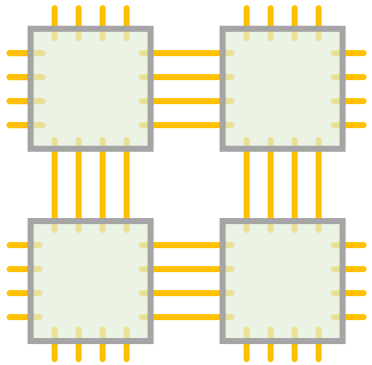


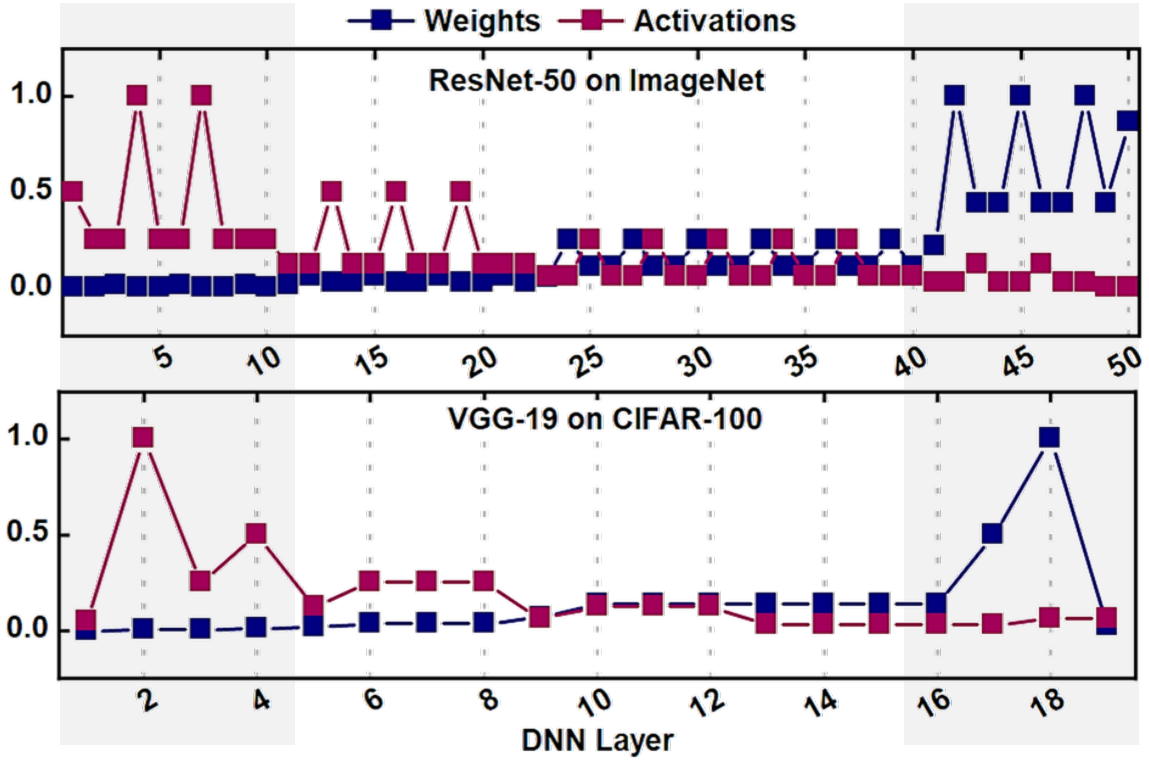
[G. Krishnan, ESWeek 2021; G. Krishnan, ICCAD 2022; Z. Wang, IEDM 2022]

AI Algorithms on IMC Chiplets

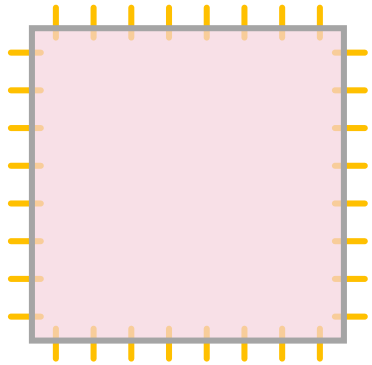
- Inherent non-uniform distribution of weights and activations across layers

Low Ws, high Xs:
Little chiplet
with high bandwidth



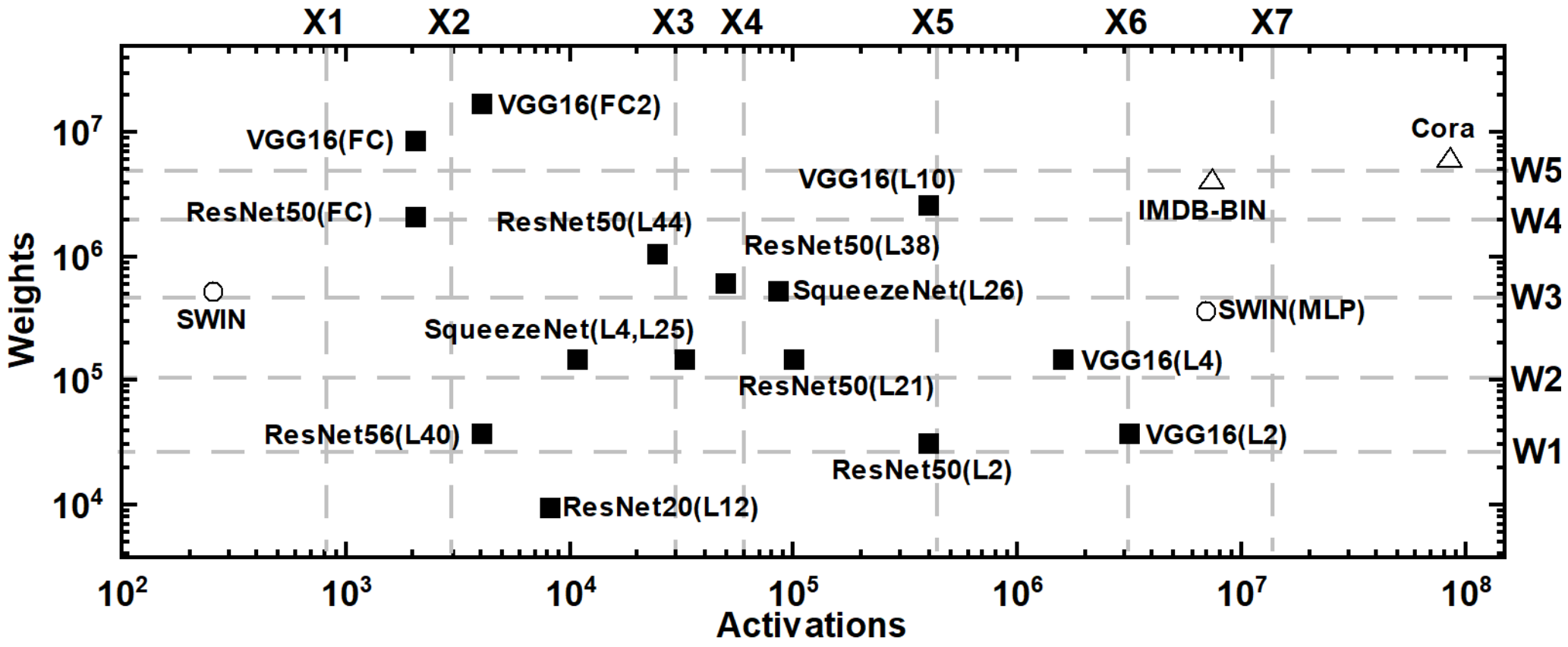


High Ws, low Xs:
Big chiplet
with low bandwidth



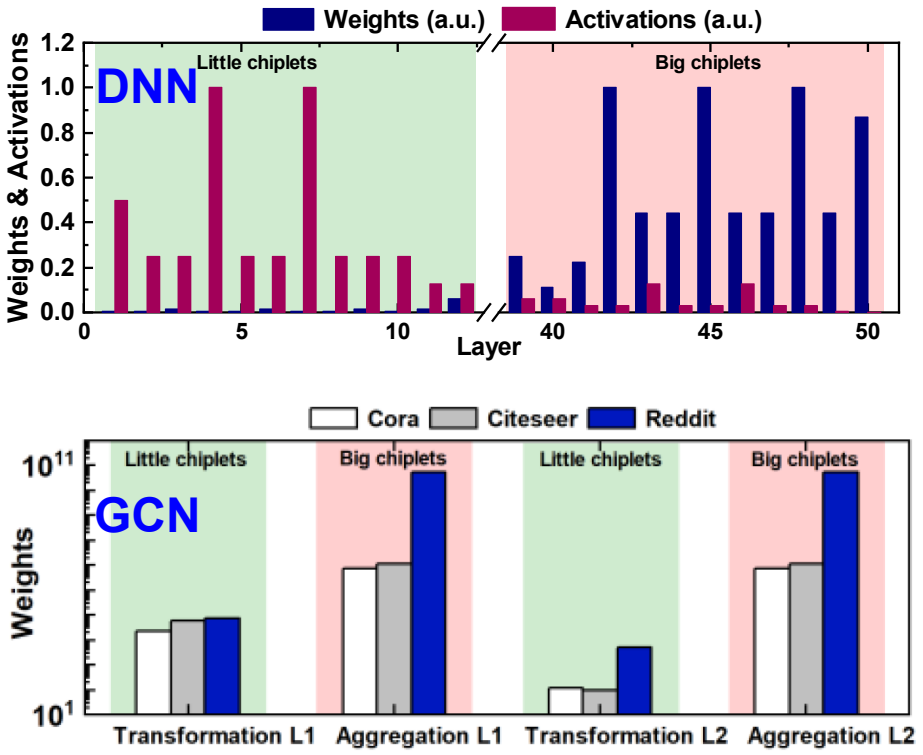
Heterogeneity in AI Algorithms

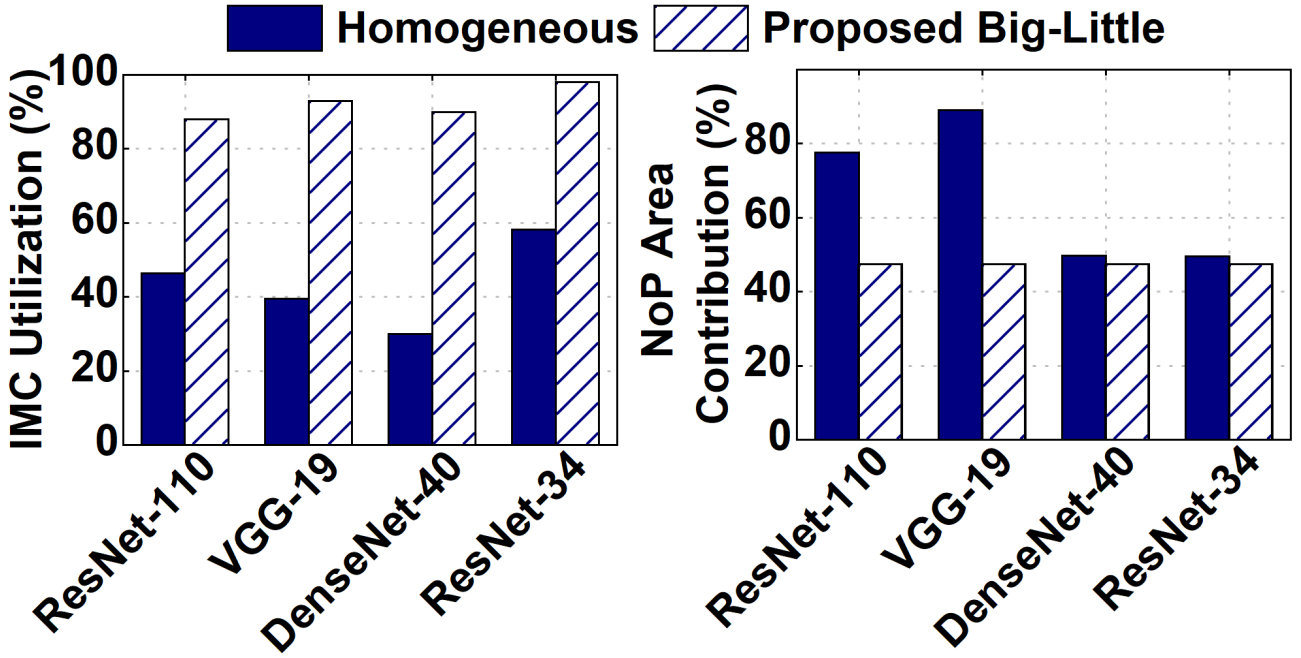
- Weight (**W**): defines how many IMC tiles and chiplets for W storage (**computing**)
- Activation (**X**): defines how much data movement intra- and inter-chiplets (**communication**)



Mapping to Big-Little Chiplets

- (W, X) provides a rule-of-thumb for algorithm mapping, layer by layer
- **> 300×** reduction in the product of Energy-Delay-Area

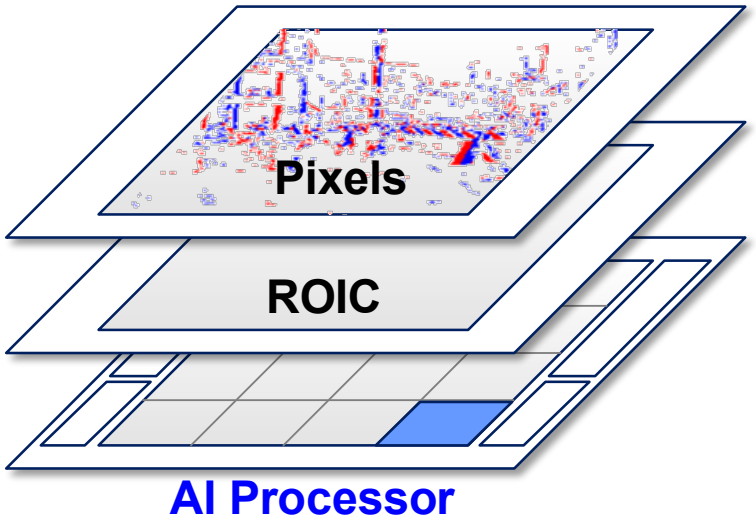


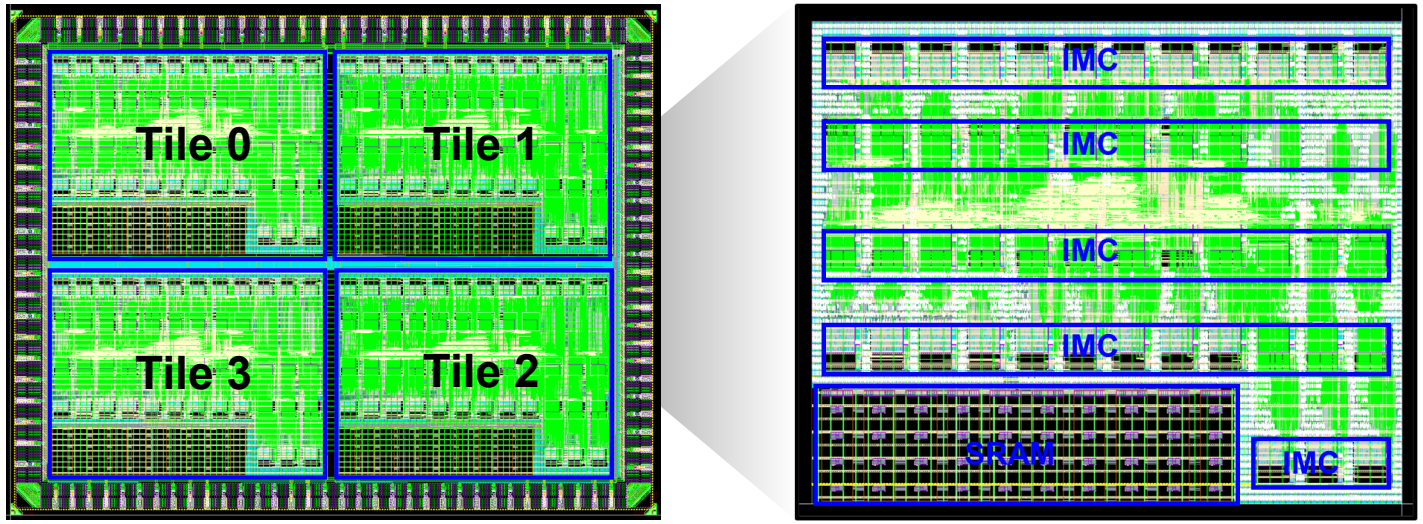


[G. Krishnan, ICCAD 2022; Z. Wang, IEDM 2022]

3D In-Sensor Computing with SIAM

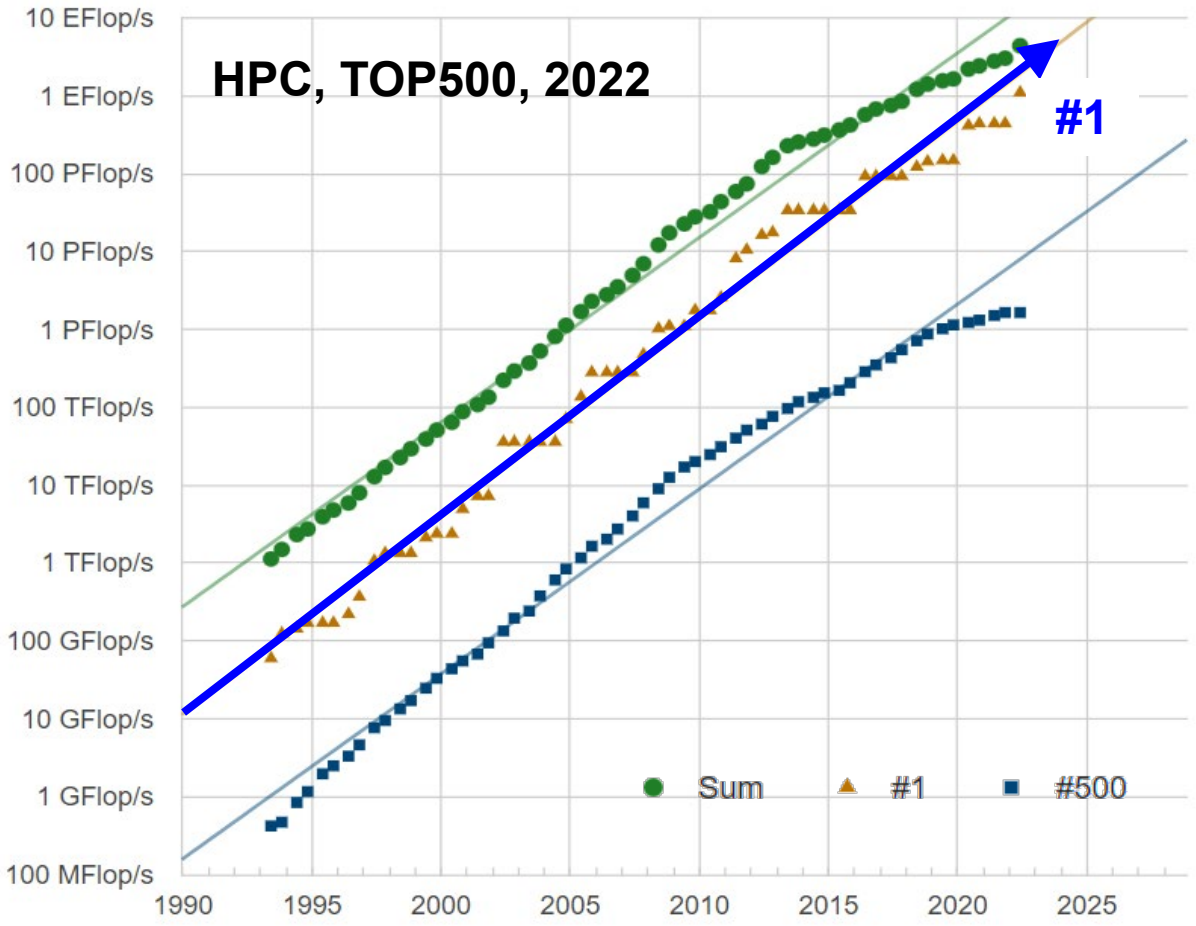
- Heterogenous integration is critical for data movement in high-definition sensors
  - 12 GB/s for 40 MP RGB at 100 fps; 1 TB/s for 4 MP DVS at 1 MHz
- 3D integration + AI computing for early detection and data compression
  - A 288 kb 65nm test chip delivered, with SIAM on architecture definition





Summary

- The demand on large-scale computing is ever increasing, driven by big algorithms and big data
- Heterogenous integration enables robust and efficient IMC systems
  - Diverse memory **devices**
  - Hybrid **circuit** structure
  - Heterogeneous chiplet **architecture**
  - Hierarchical **interconnection**
- **Co-design** across multiple layers is key to future success



[TOP500, 2022]

Acknowledgement

- Collaborators: Jae-sun Seo, Chaitali Chakrabarti (ASU), Nathaniel Cady (SUNY Poly), Umit Ogras (U. Wisconsin), Suman Datta, Arijit Raychowdhury, Shimeng Yu (GaTech), Jeff Vetter, Frank Liu (ORNL), Mahantesh Halappanavar (PNNL), Rajiv Joshi (IBM)
- Graduate students: Gokul Krishnan, Sumit K. Mandal, and much more
- JUMP-CBRIC, DARPA, DOE, NSF