Exploiting 2.5D/3D Heterogeneous Integration for AI Computing

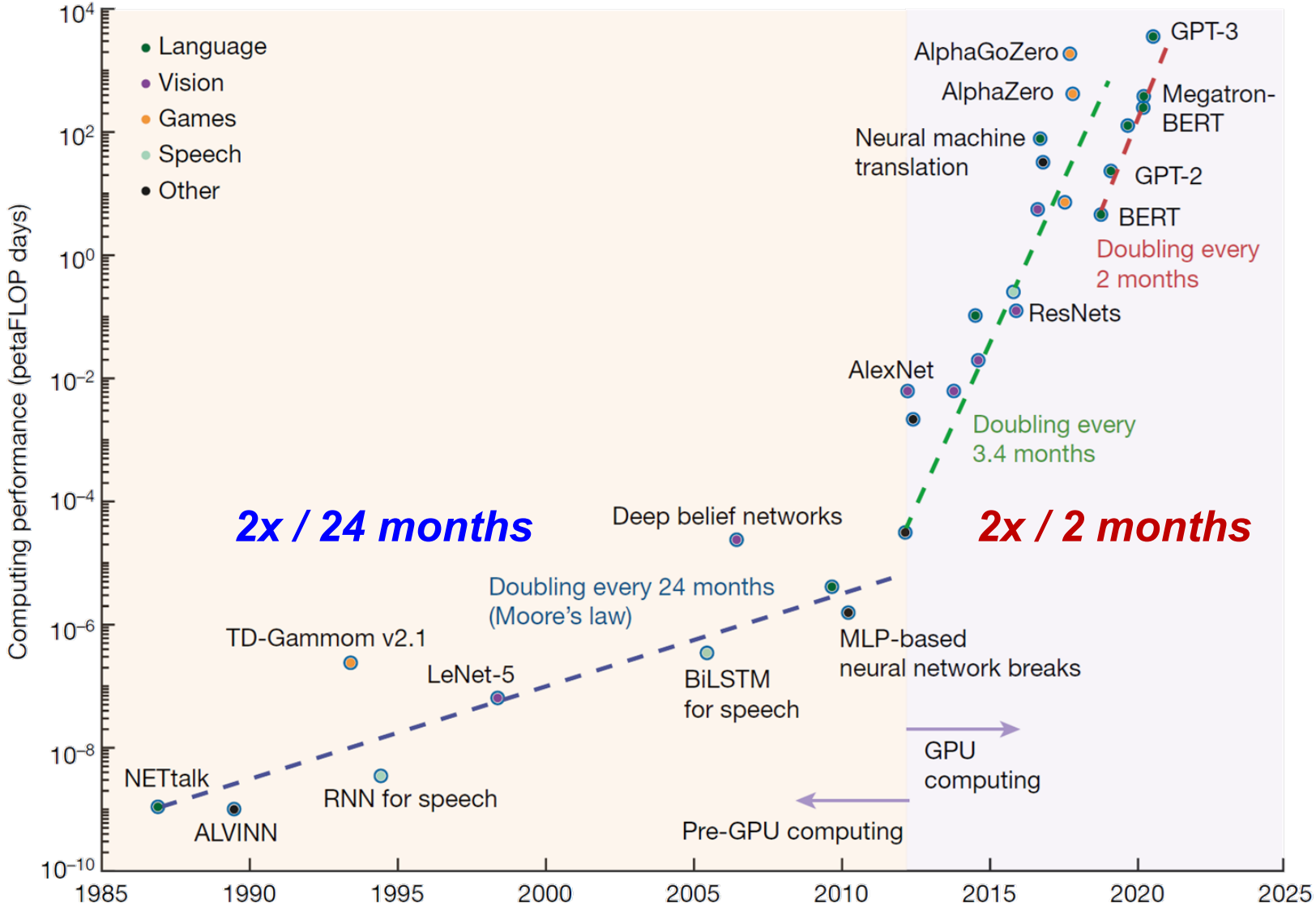
Zhenyu Wang[1], Jingbo Sun[1], Alper Goksoy[2], Sumit K. Mandal[3], Yaotian Liu[1], Jae-sun Seo[4], Chaitali Chakrabarti[1], Umit Y. Ogras[2], Vidya Chhabria[1], Jeff Zhang[1], Yu (Kevin) Cao[5]

[1]Arizona State University, [2]University of Wisconsin-Madison, [3]Indian Institute of Science, [4]Cornell Tech, [5]University of Minnesota

Toward 2.5D/3D Heterogeneous Integration

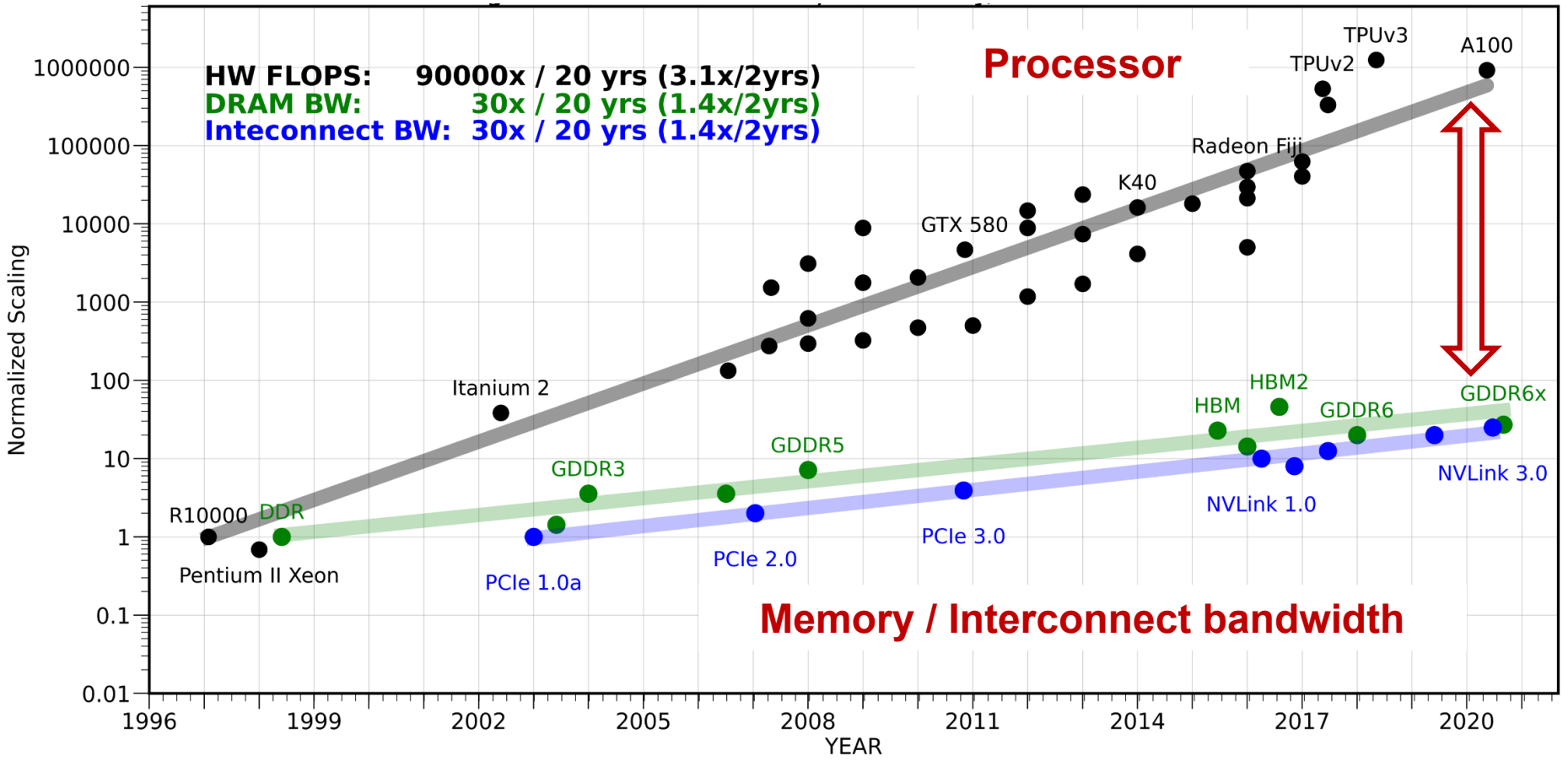
- Interconnection beyond the monolithic design
- 2.5D/3D benchmarking: HISIM
  - Analytical performance modeling
  - Thermal simulation
- Benchmark studies
- Summary

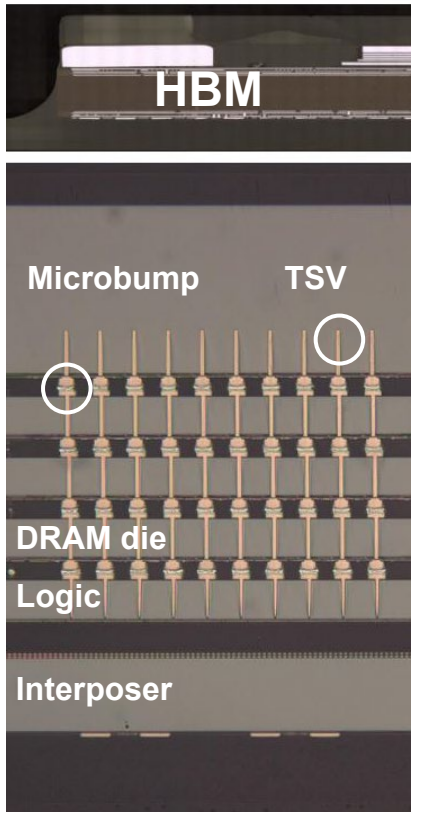
Bigger AI Models



[A. Mehonic, A. J. Kenyon, 2022]

Bigger Gaps in Data Movement

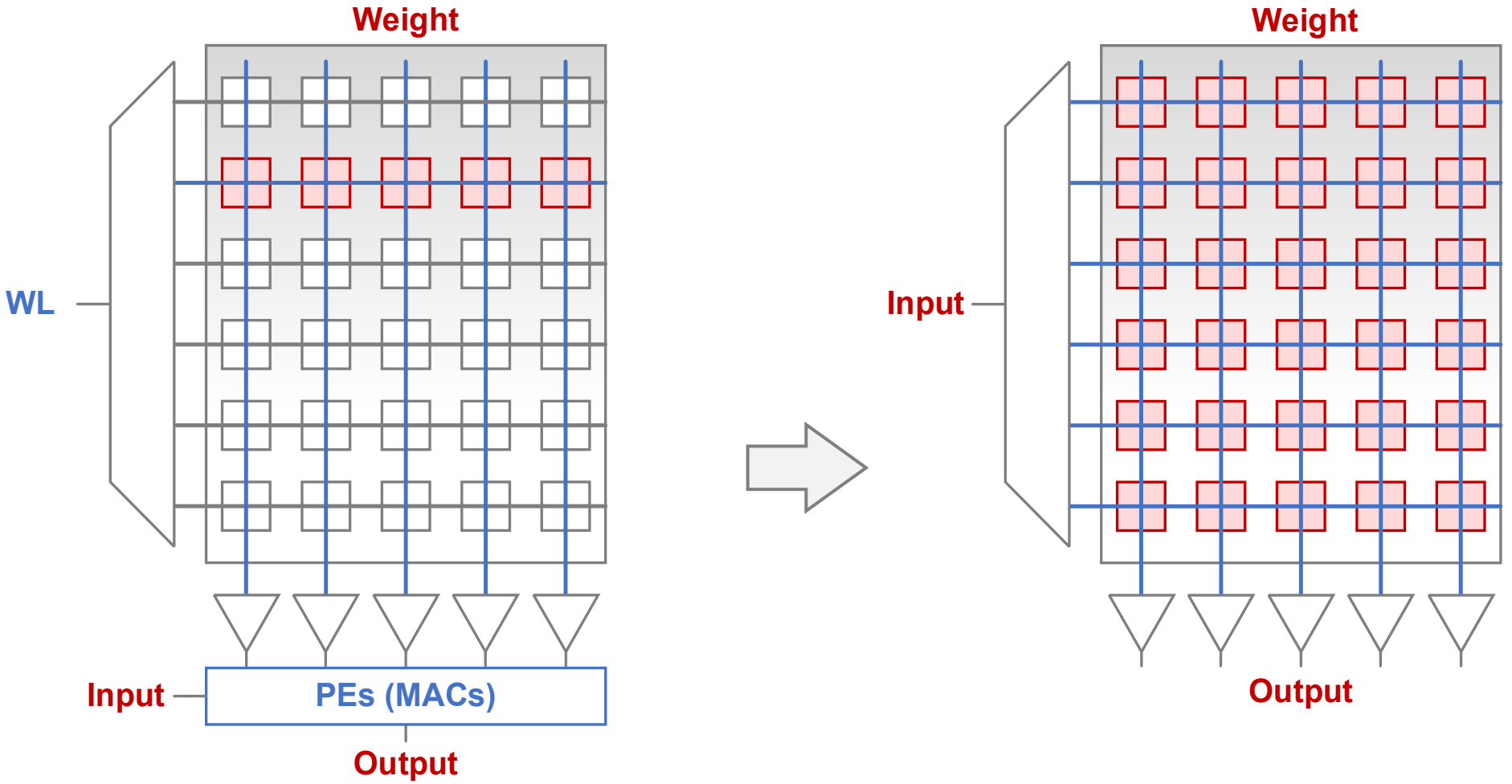




[A. Gholami, 2020; SK Hynix]

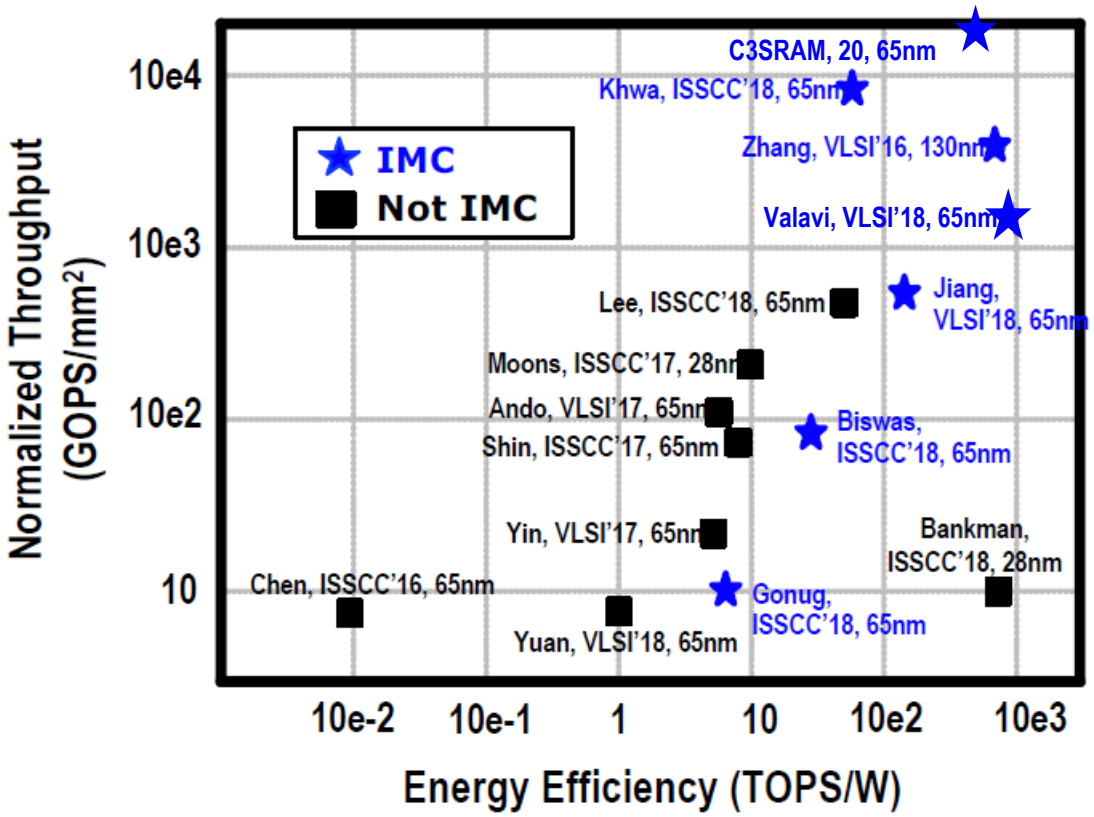
In-Memory Computing

- IMC combines memory access and computation into a single unit

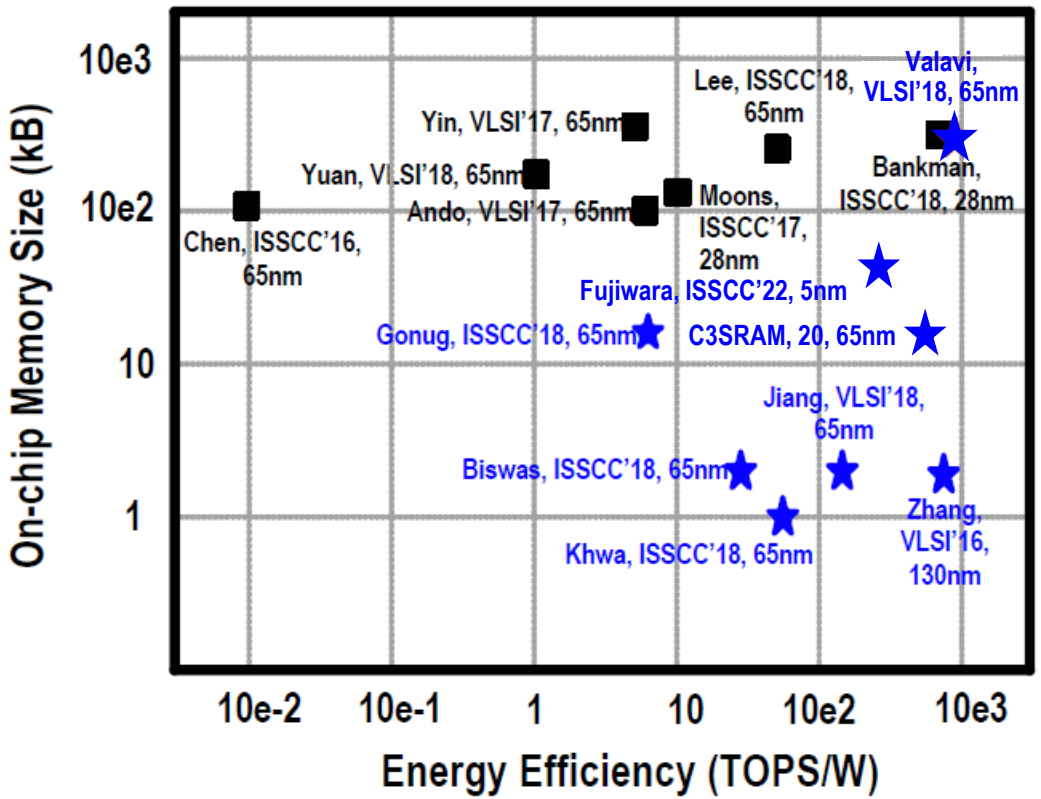


Promises and Challenges

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



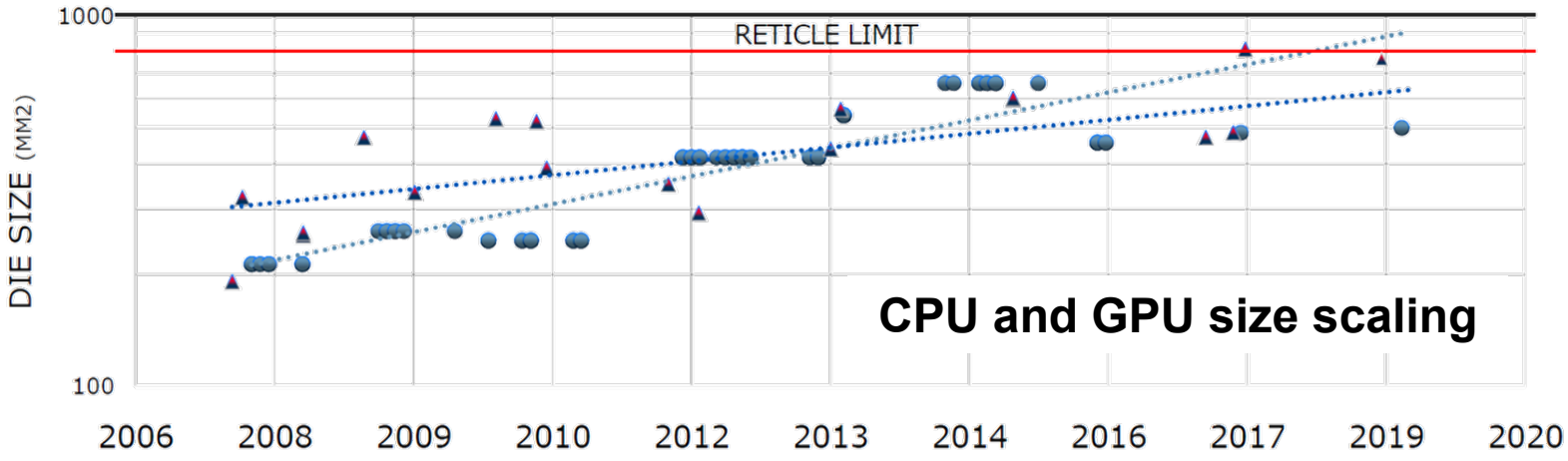
- **Limited scale** due to robustness and peripheral circuits

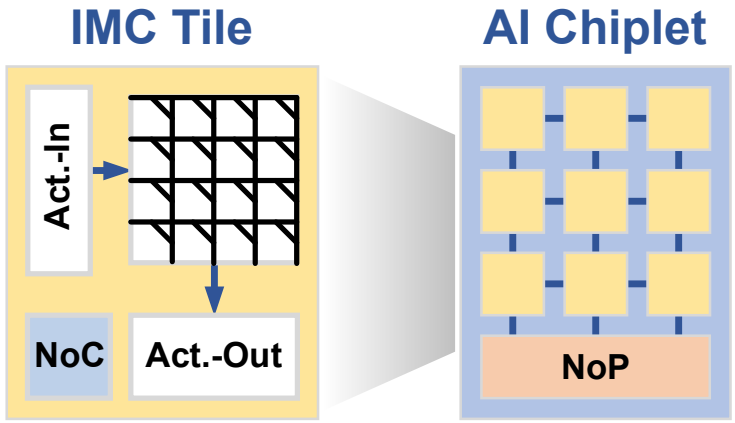


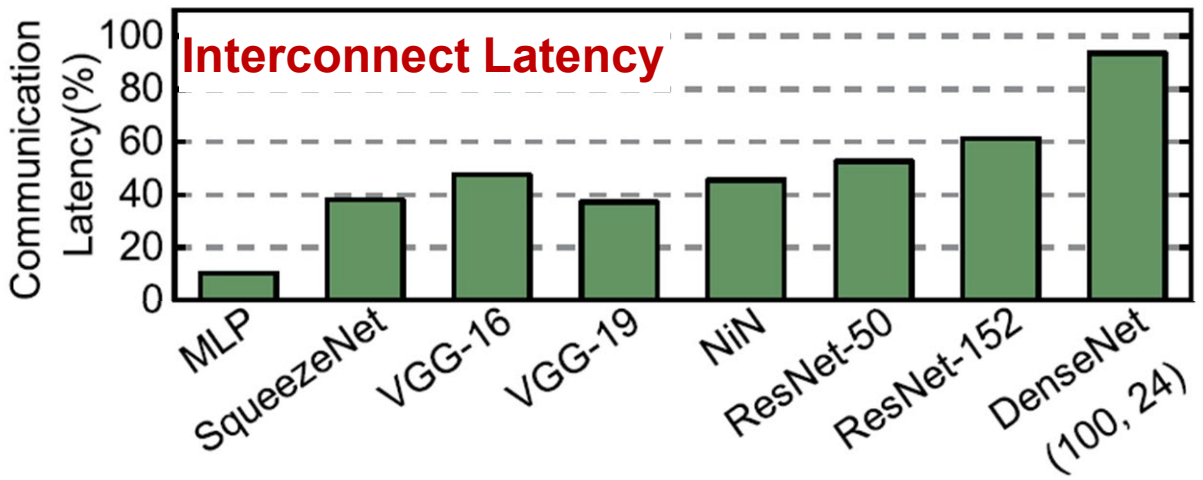
[N. Verma, ISSCC 2019]

System Scaling of IMC

- **Die area/cost** and **Interconnection** limiting a monolithic design for large-scale AI computing



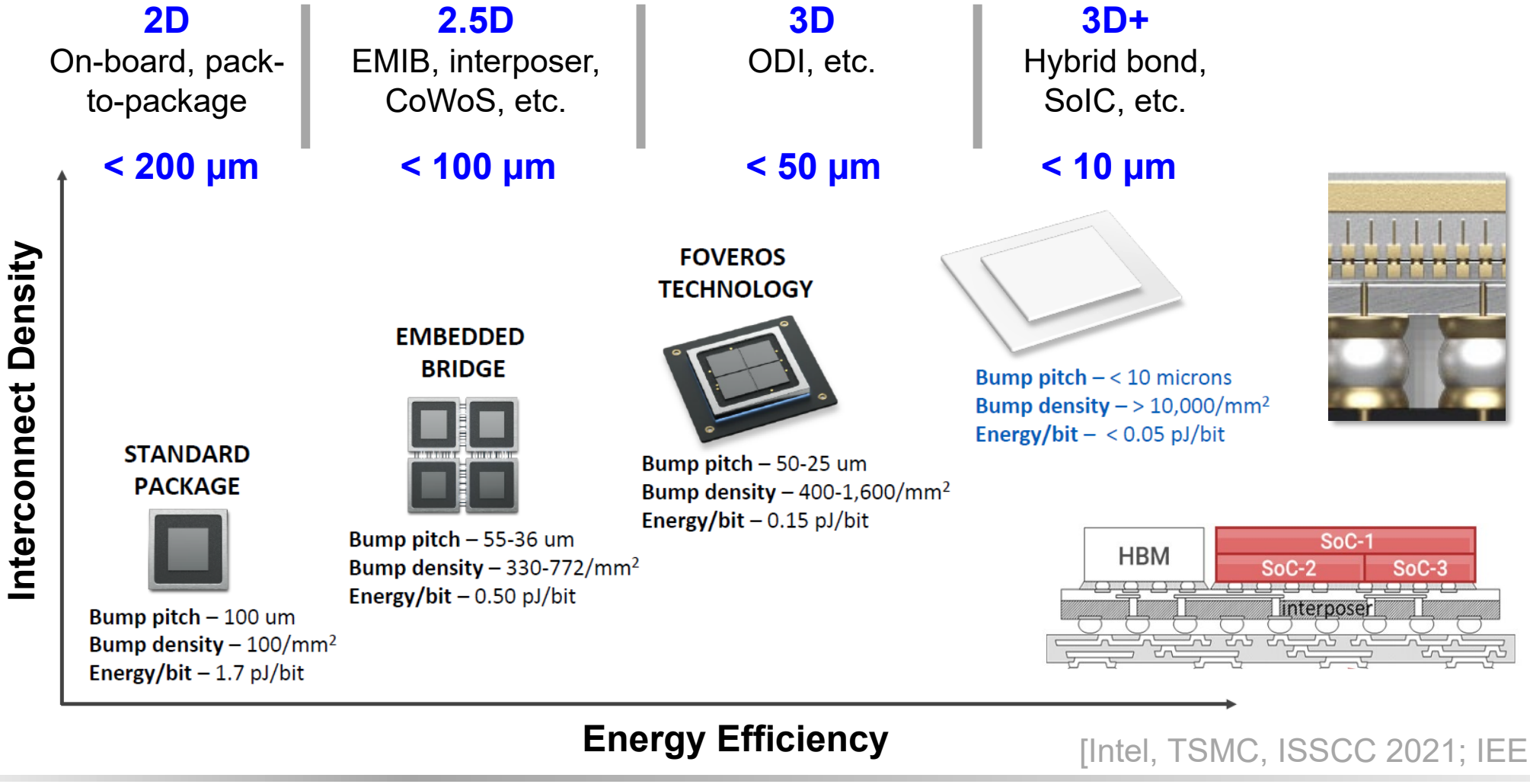




[AMD, ISSCC 2021; G. Krishnan, et al., IEEE D&T, 2020 and JETCAS, 2020]

From 2.5D to 3D and 3D+

- 10-100X improvement / generation in data speed and bandwidth density



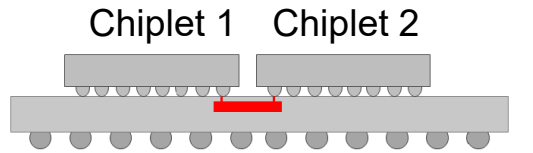
[Intel, TSMC, ISSCC 2021; IEEE HIR, 2021]

Toward 2.5D/3D Heterogeneous Integration

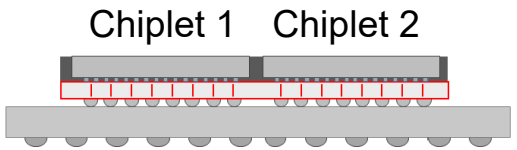
- Interconnection beyond the monolithic design
- 2.5D/3D benchmarking: HISIM
  - Analytical performance modeling
  - Thermal simulation
- Benchmark studies
- Summary

2.5D Integration of Chiplets

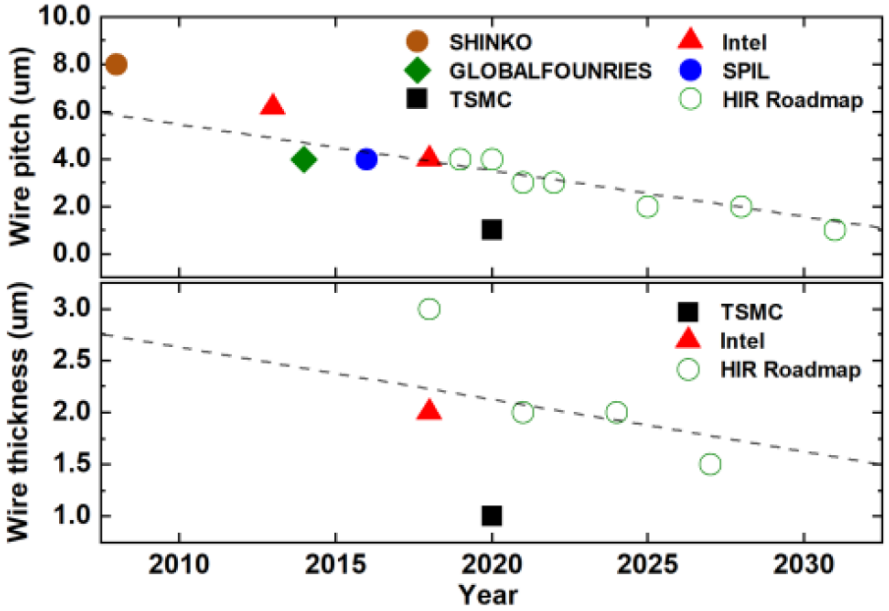
- Leveraging monolithic fabrication to prepare fine-pitch, high-density interconnections, which interface the PCB to connect multiple chiplets

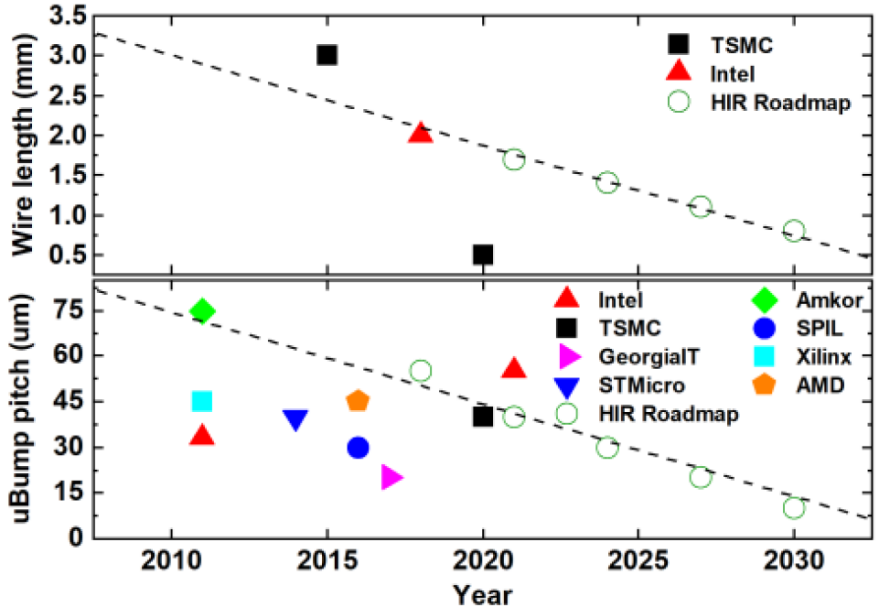


Embedded Multiple Interconnect Bridge (EMIB)



Silicon Interposer

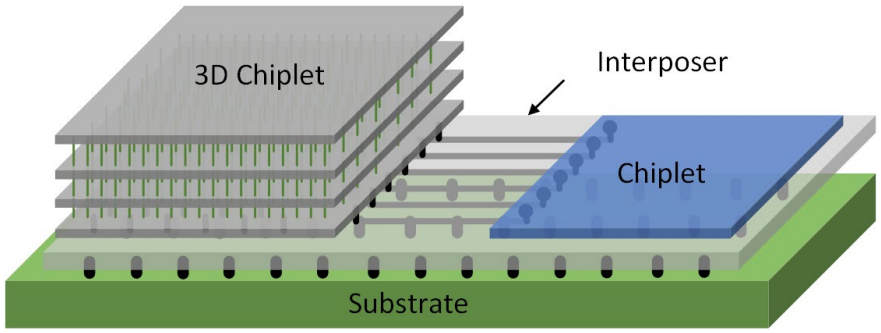


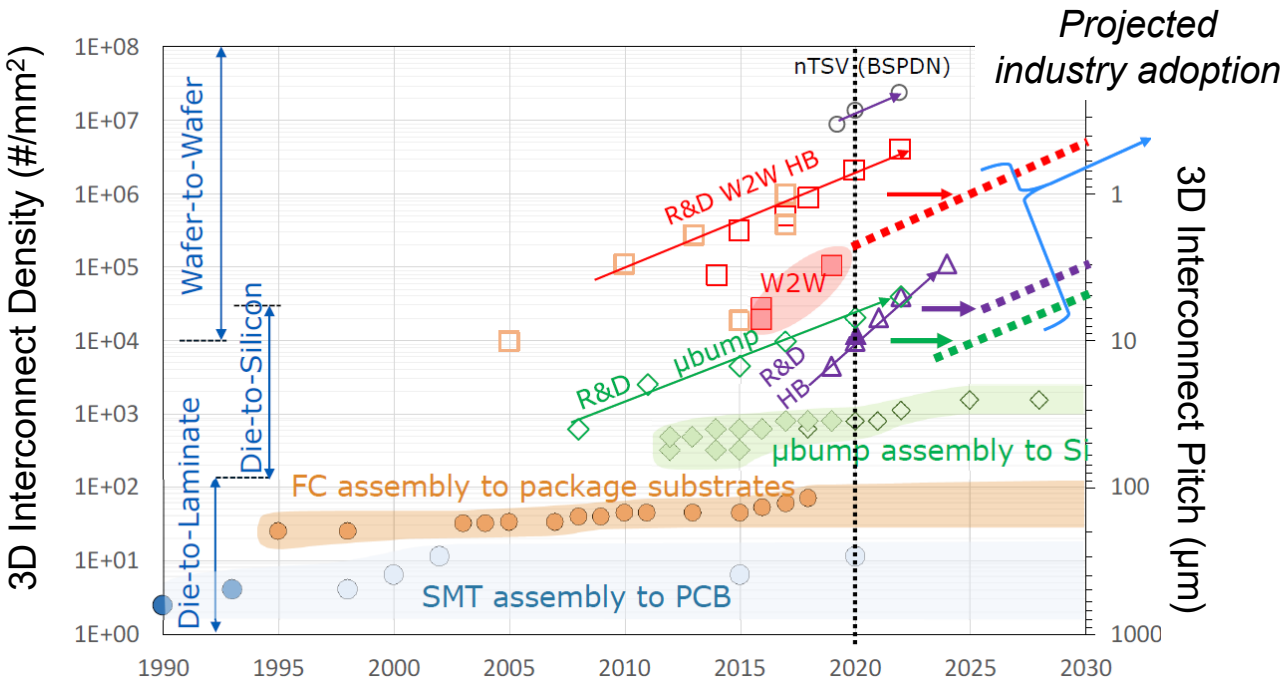


[Intel, 2019; Z. Wang, IEDM 2022]

Roadmap of 3D Packaging

- "Another direction of improvement of computing power is to make physical machines three-dimensional." – Richard P. Feynman, 1985
- From 2010 to 2030: bandwidth density (Gbps/mm-3) ***from <10 to 10^9***, energy efficiency (pJ/bit) ***from >1 to 0.01***

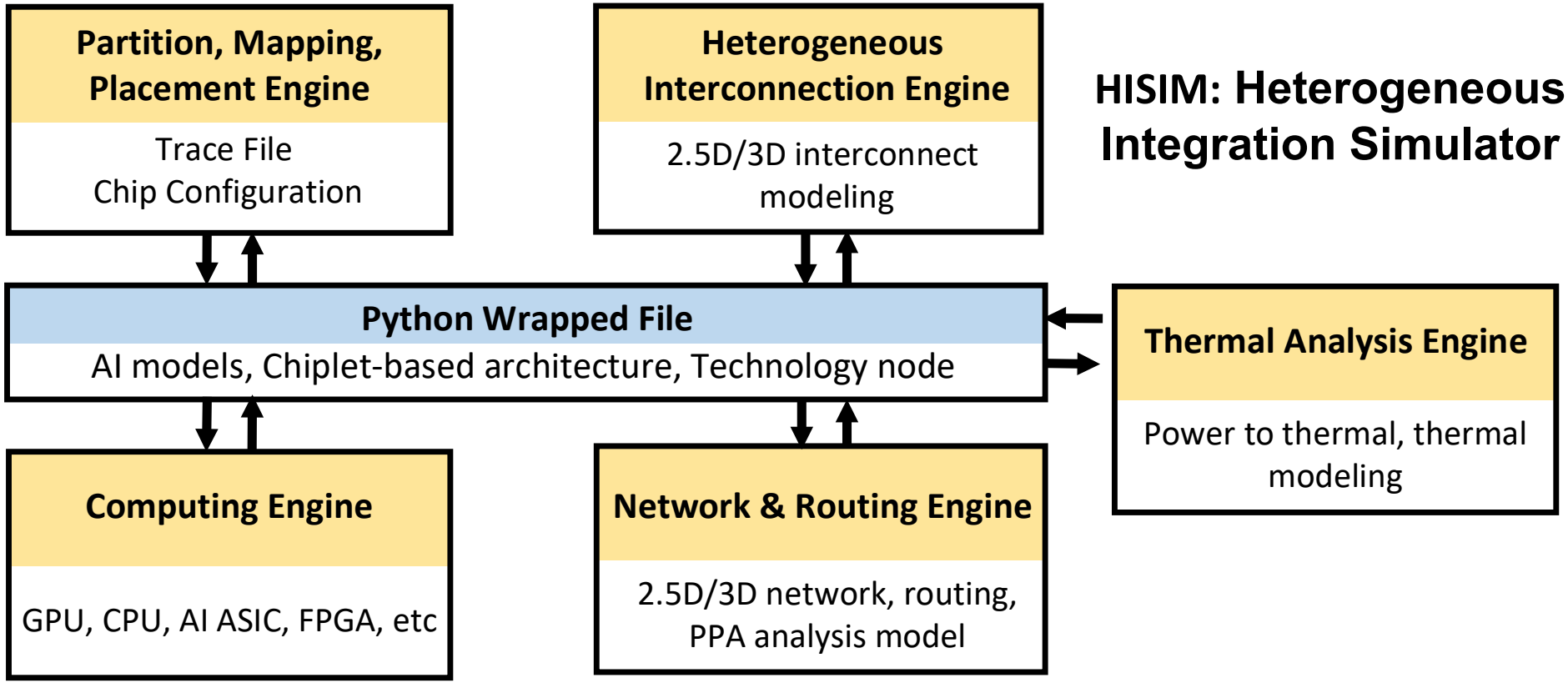




[IEEE HIR, 2021; IMEC, 2021]

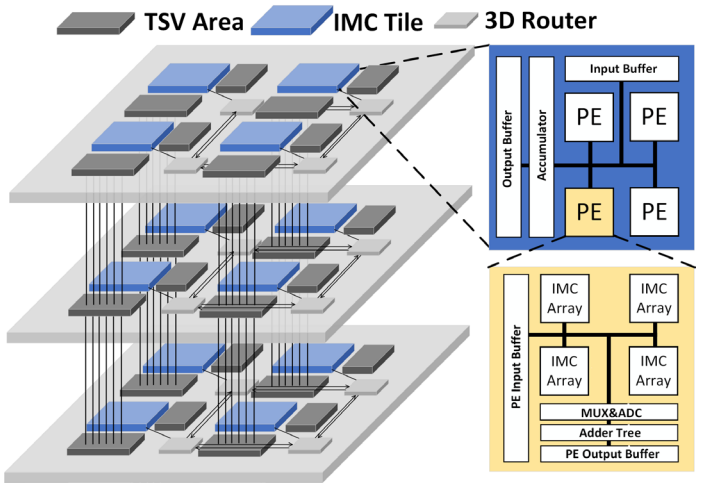
Simulation Engines in HISIM

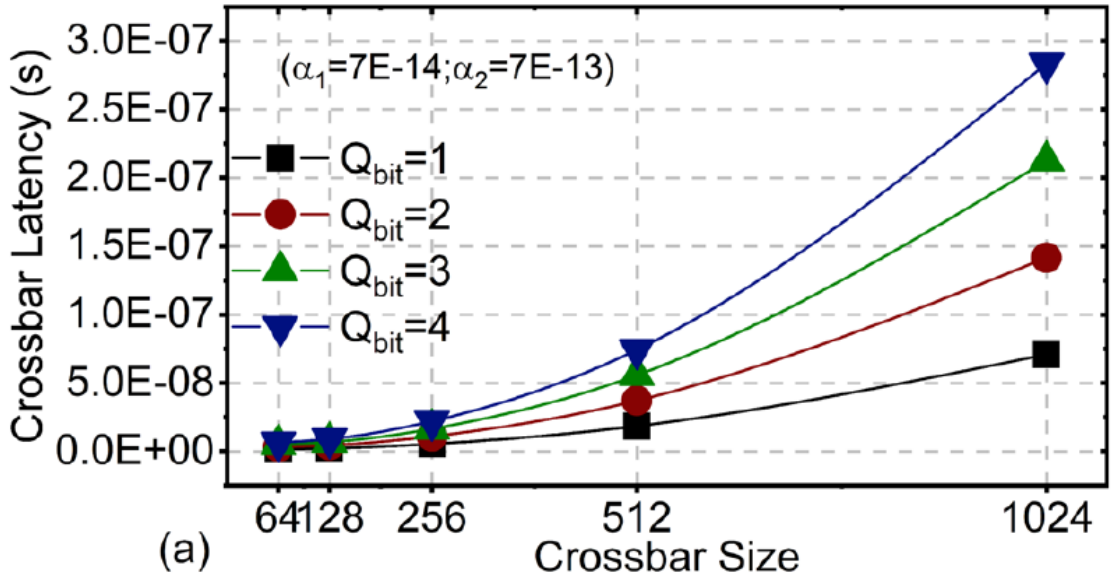
- Heterogeneous Integration Simulator with Interconnect Modeling (**HISIM**)
- **10^4-10^6x** faster than previous simulators in performance benchmarking

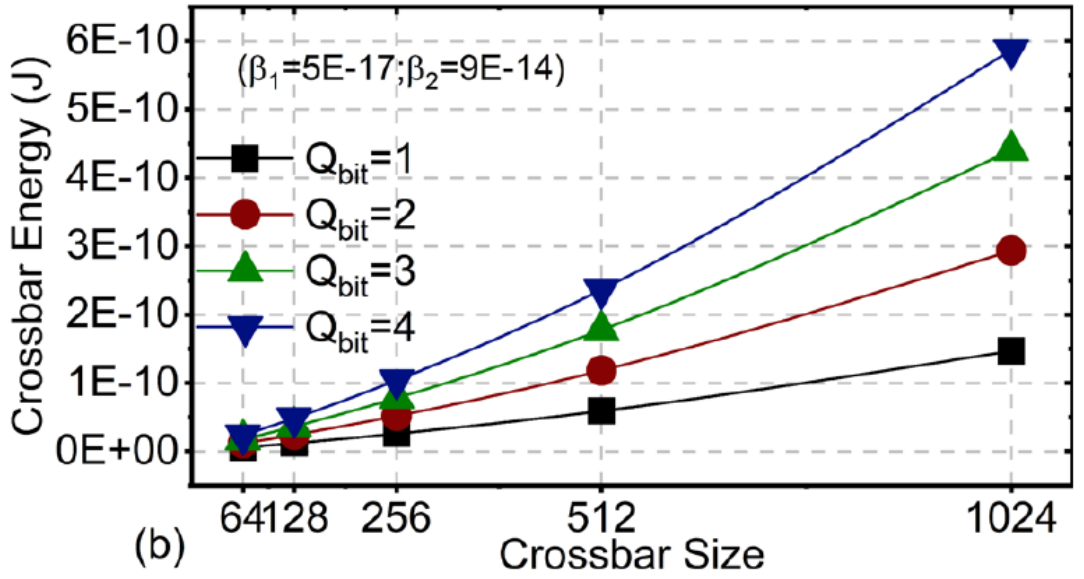


Compute Unit

- IMC: Model size impacts the area. Activation volume impacts data movement. Model sparsity impacts power consumption
- Analytical PPA modeling for IMC chiplets

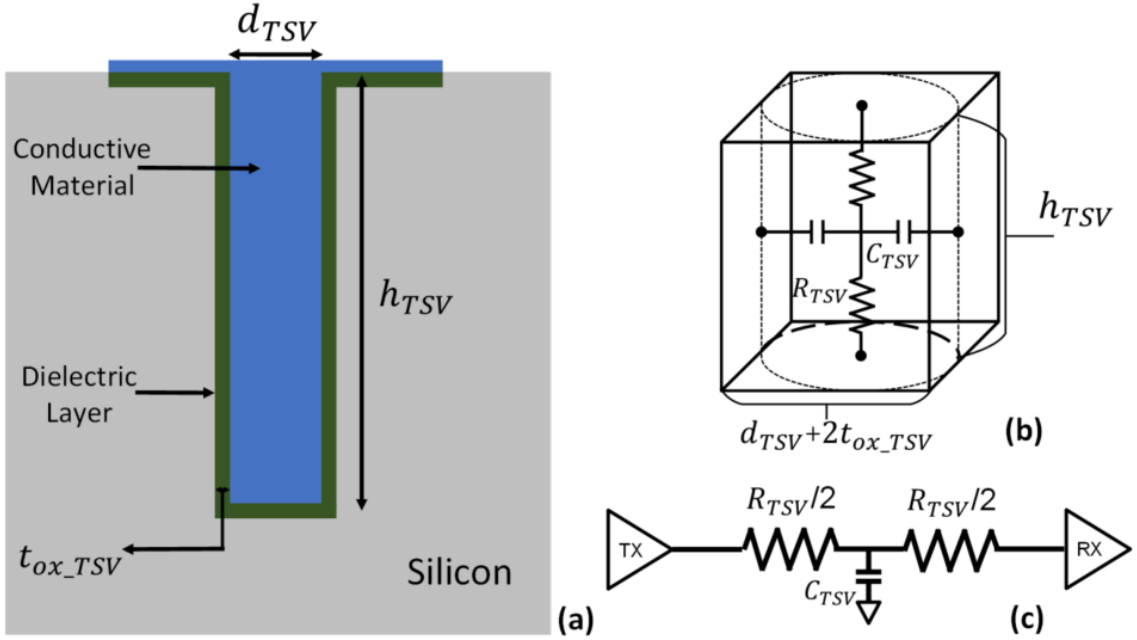


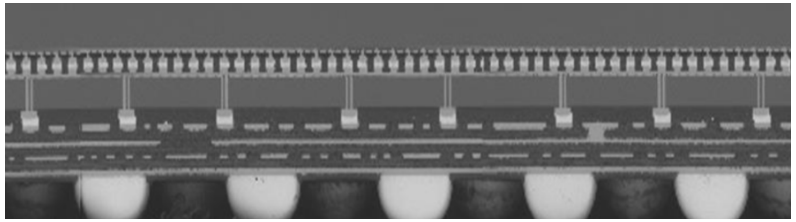




2.5D/3D Interconnect Modeling

- Analytical models of TSVs
  - Convert the TSV geometry into RC parameters
  - RC product for bandwidth calculation
- Analytical parasitic models for micro bumps and hybrid bonding



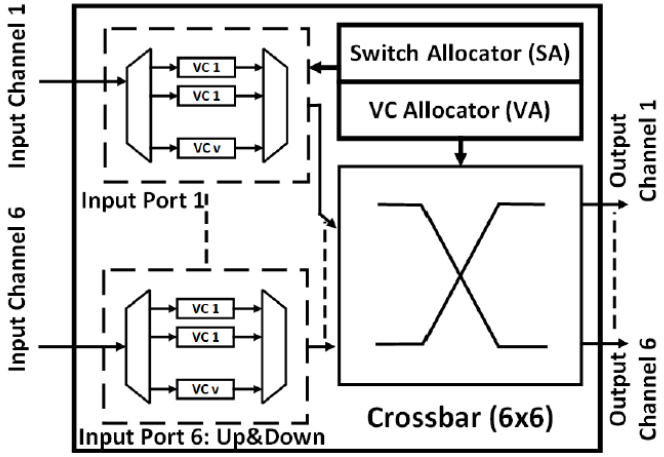


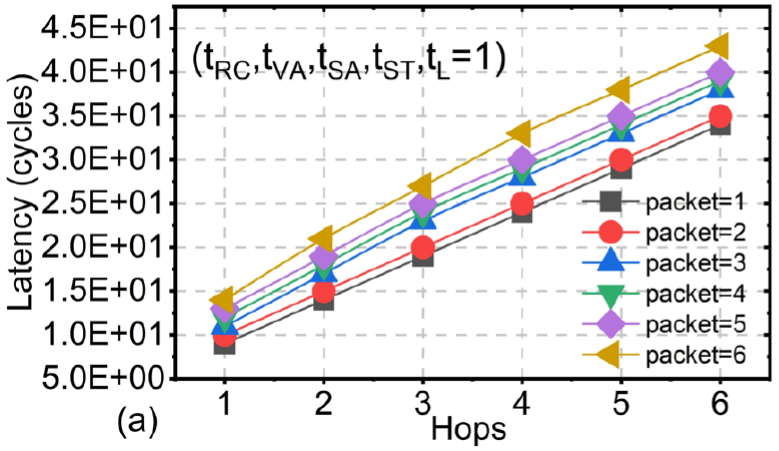
Intel Foveros

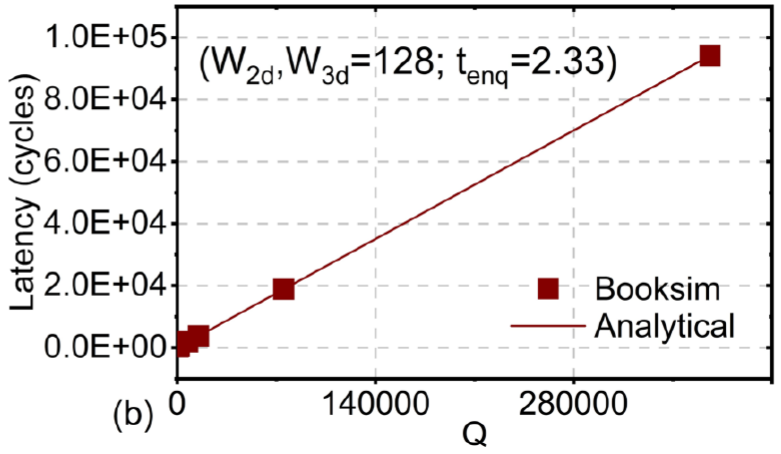
r_{TSV} (um)	d_{TSV} (um)	h_{TSV} (um)	R_{TSV} (mOhm)	C_{TSV} (fF)
1.25	2.5	25	87.12	4.098974014
2.5	5	50	43.56	15.0902274
5	**10**	**100**	**21.78**	**57.64186203**
10	20	200	10.89	225.0042397
15	30	300	7.26	502.0420495
20	40	400	5.445	888.7547285

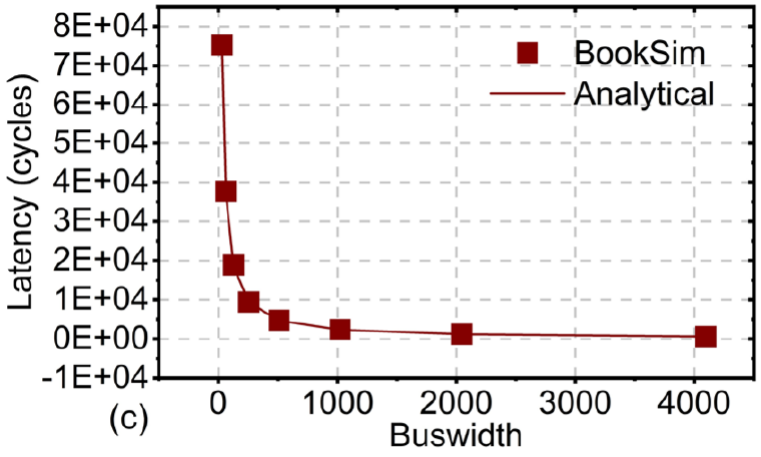
2.5D/3D Network Modeling

- 2D and 3D network routers calibrated with ORION 3.0
- Custom Booksim for 2D and 3D traffic calculation
- Analytical PPA modeling of NoC and NoP, scalable with data volume, bandwidth and routing schemes



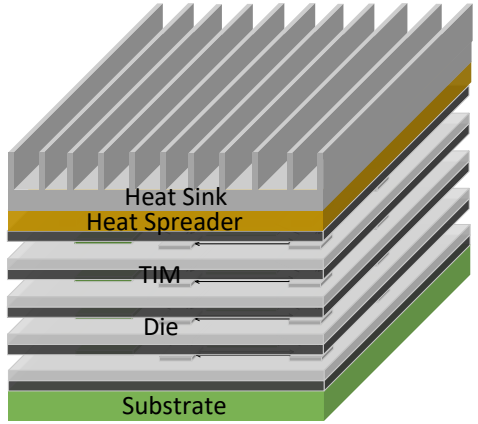




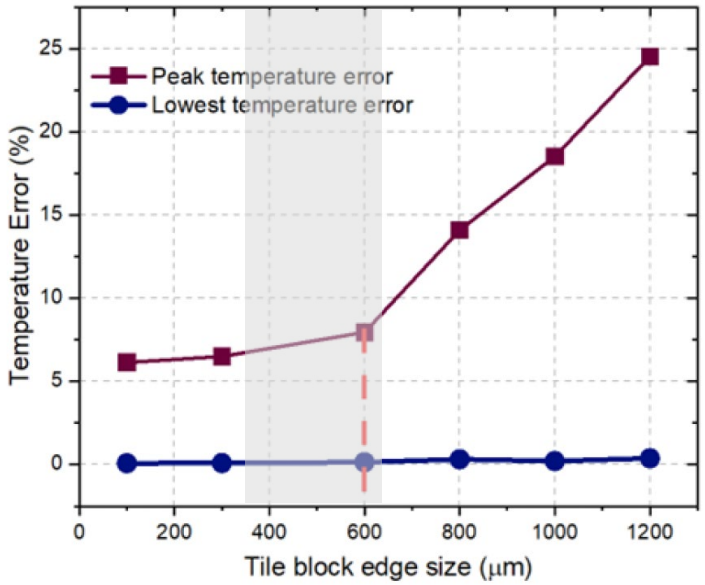


Thermal Analysis

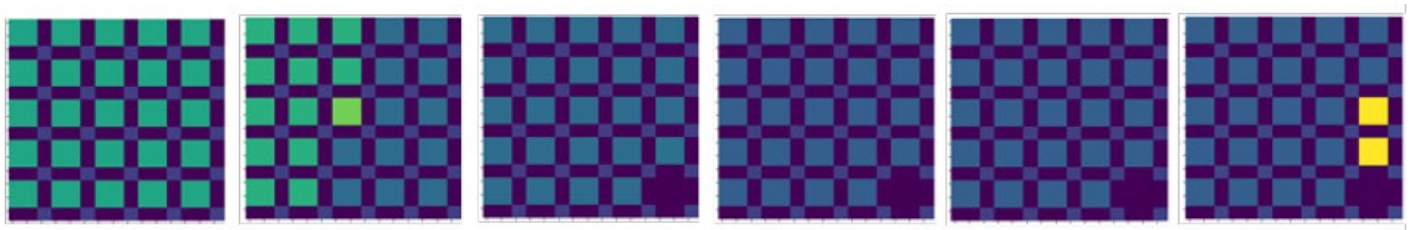
- Efficient thermal prediction
  - Static thermal modeling for 3D tiers
  - Physics-informed GNN for full 3D thermal analysis under packaging variations



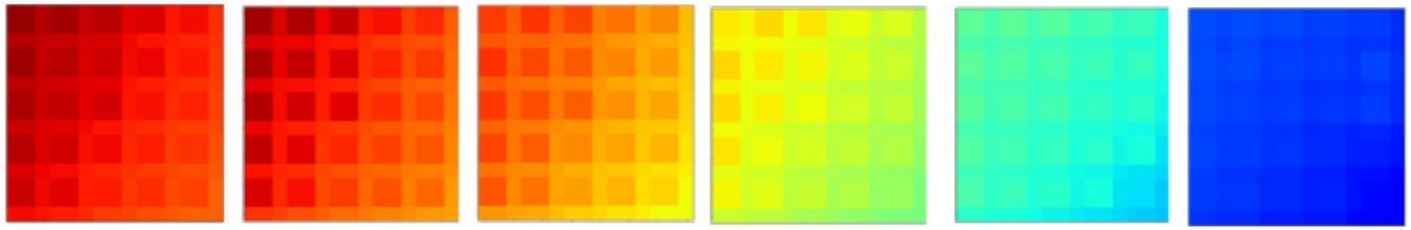
Si thermal diffusion distance



Power Map (6 tiers, ResNet-110 on CIFAR-100)



Thermal Map

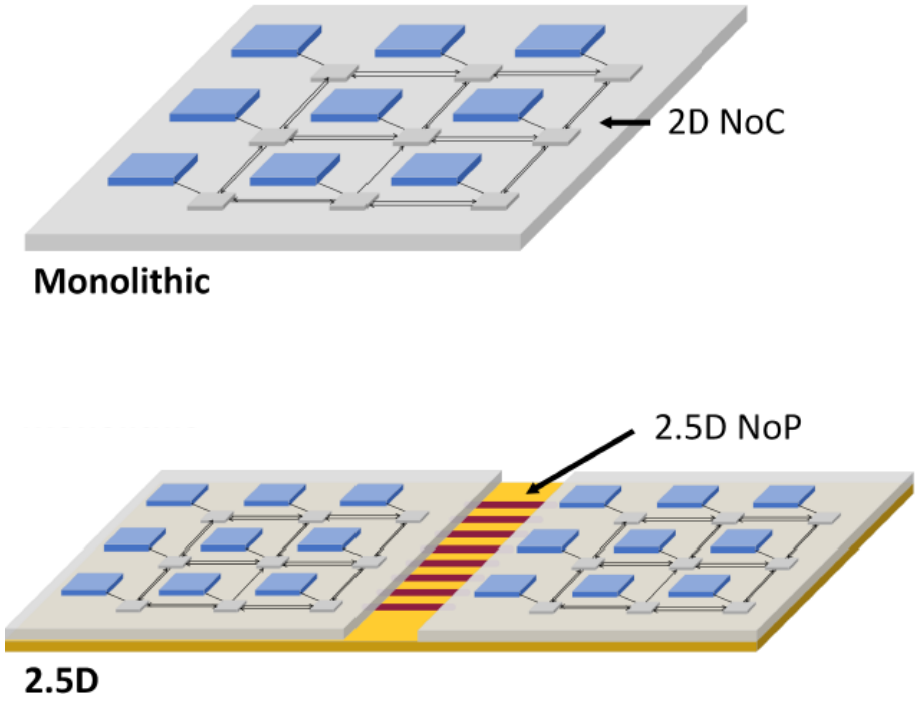


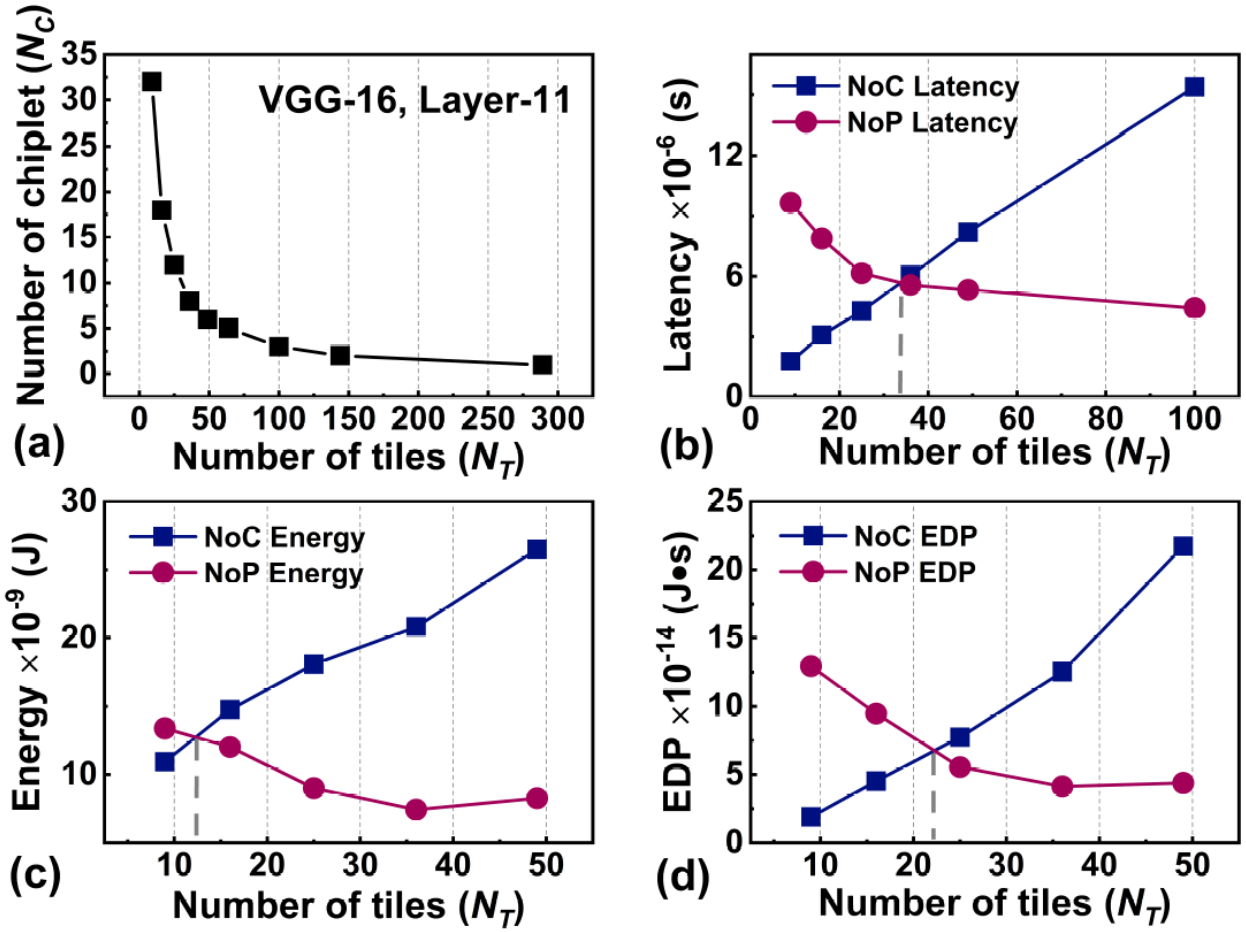
Toward 2.5D/3D Heterogeneous Integration

- Interconnection beyond the monolithic design
- 2.5D/3D benchmarking: HISIM
  - Analytical performance modeling
  - Thermal simulation
- Benchmark studies
- Summary

2D NoC vs. 2.5D NoP

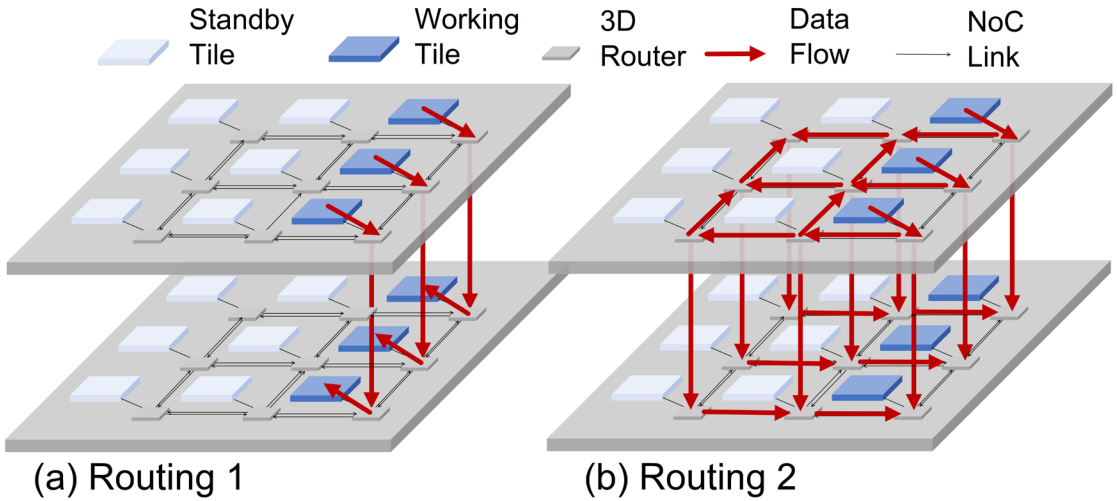
- NoC cost increases fast with chiplet size
  - AIB used in NoP simulation

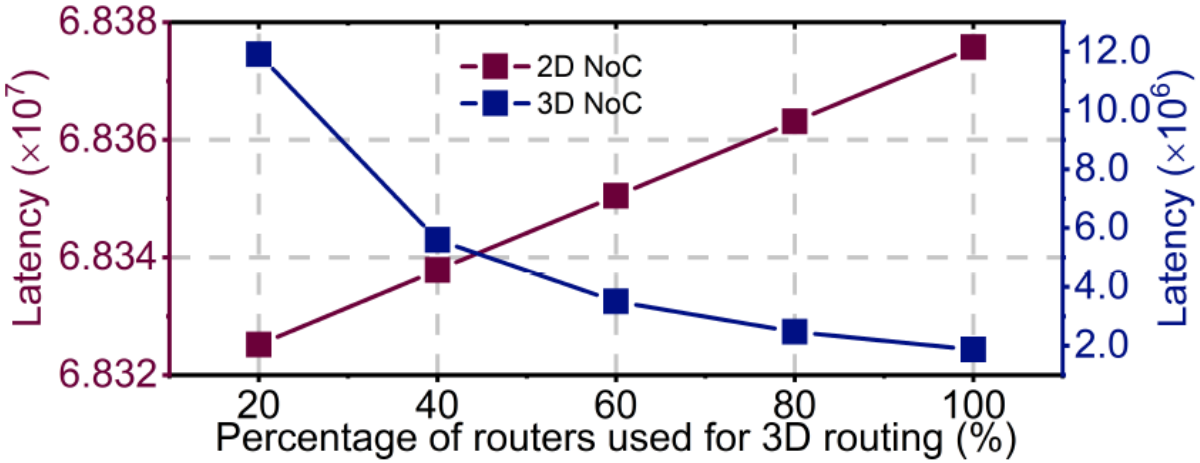


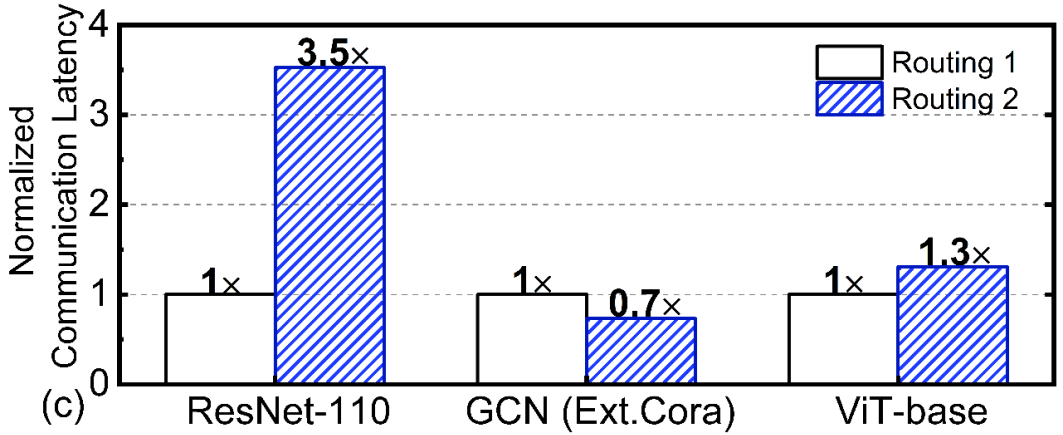


3D Routing

- Tradeoff between 2D NoC and 3D NoP
- 3D TSVs are increasingly efficient

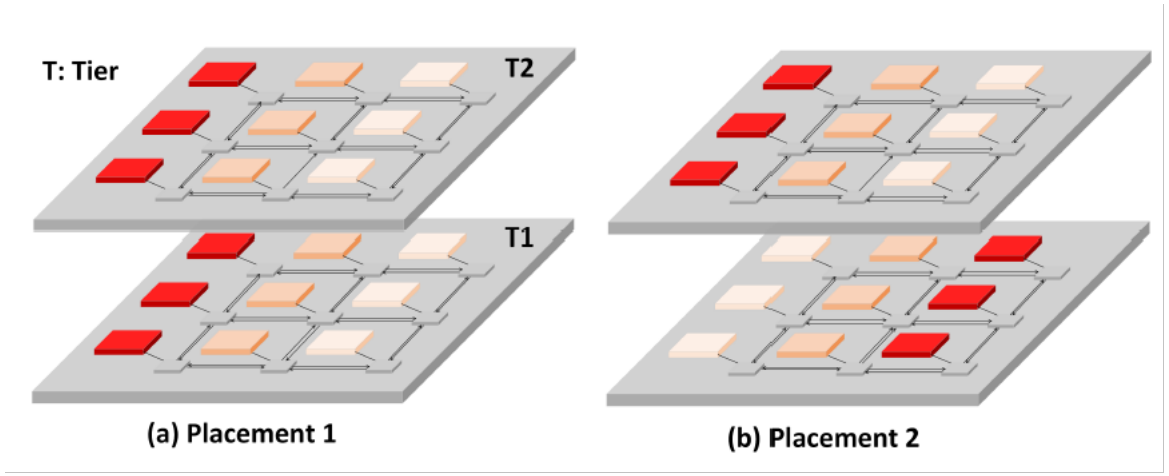


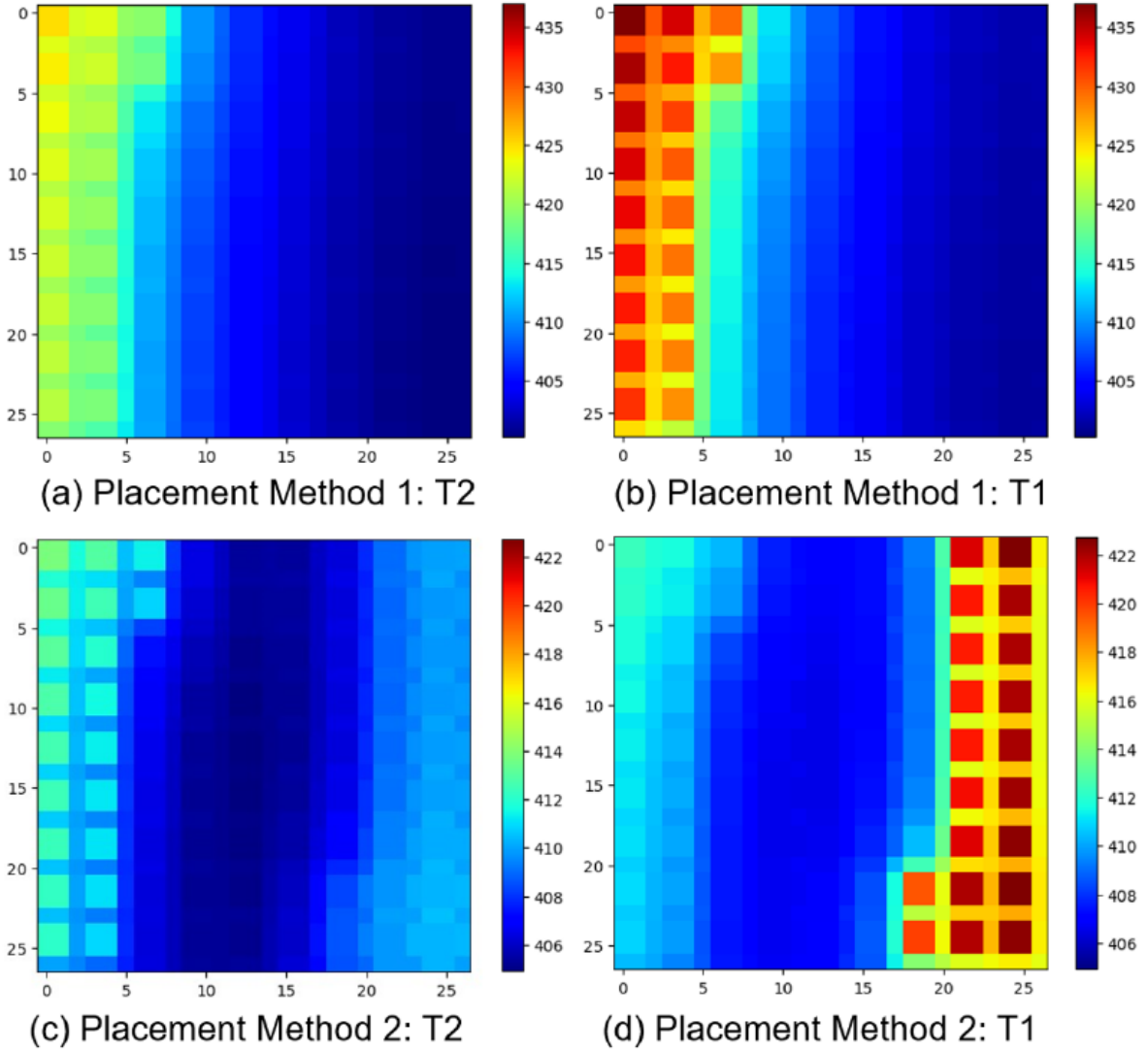




3D Placement

- Power map of AI computing is non-uniform, leading to a non-uniform thermal map
- >10^{o}C cooling is achieved in this example of 2 tiers





Challenges Ahead

- Thermal management: Workload assignment, thermal-aware control, etc.
- Power delivery and integrity: On-chip and on-package PDN
- Reliability and testing: Robust computing and networks
- Architecture: System partition in 3D HI
- Device-chiplet-system-algorithm co-design!

The HI roadmap is solid for 10+ years, with >2x / 2 years!