Chiplet-Based Integration Scale-Down and Scale-Out

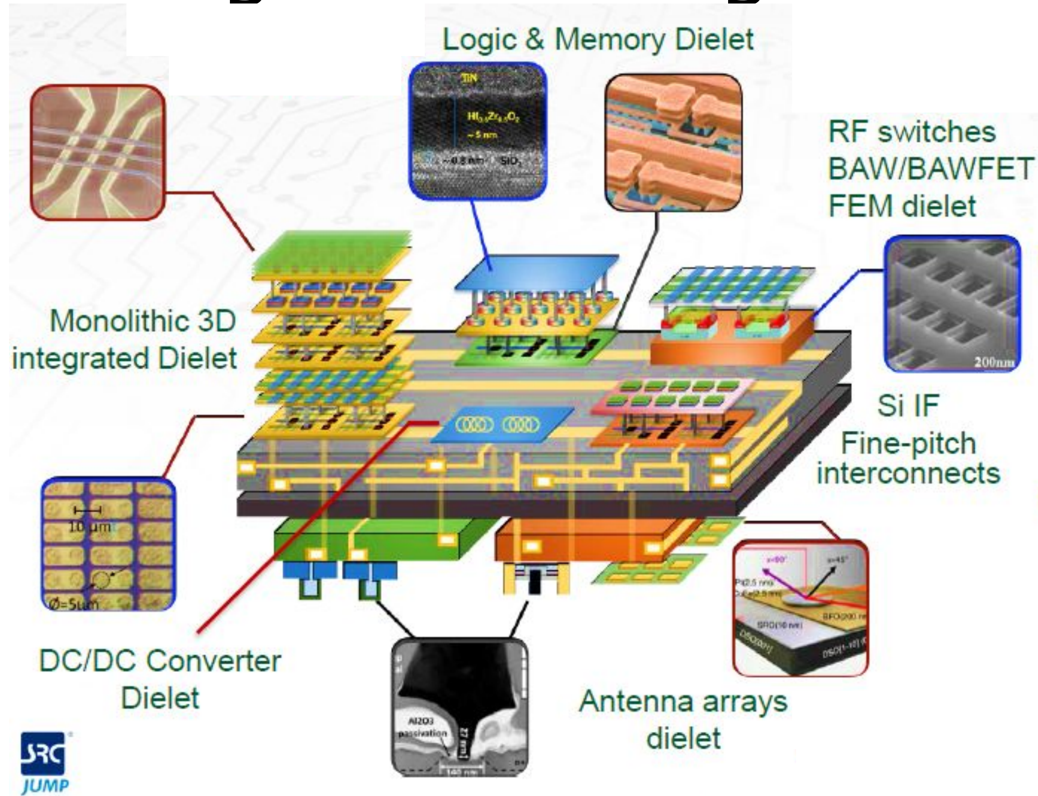
Boris Vaisband

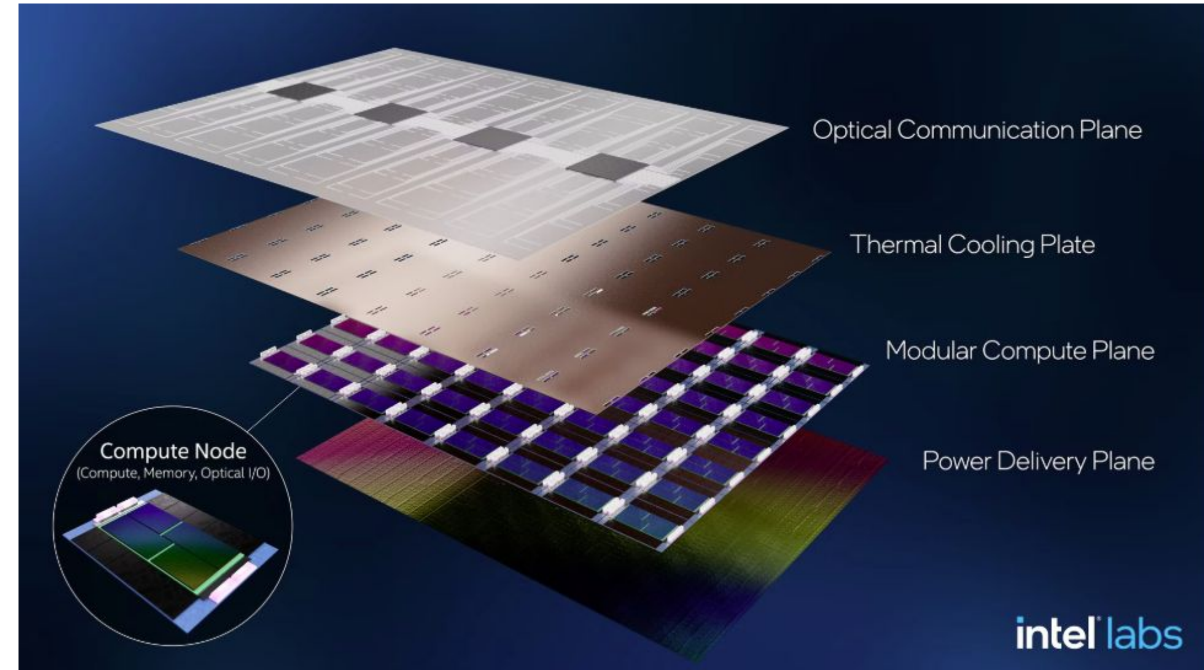
THInK Team

The Heterogeneous Integration Knowledge Team

UCI Samueli School of Engineering

Heterogeneous Integration – "More than Moore"



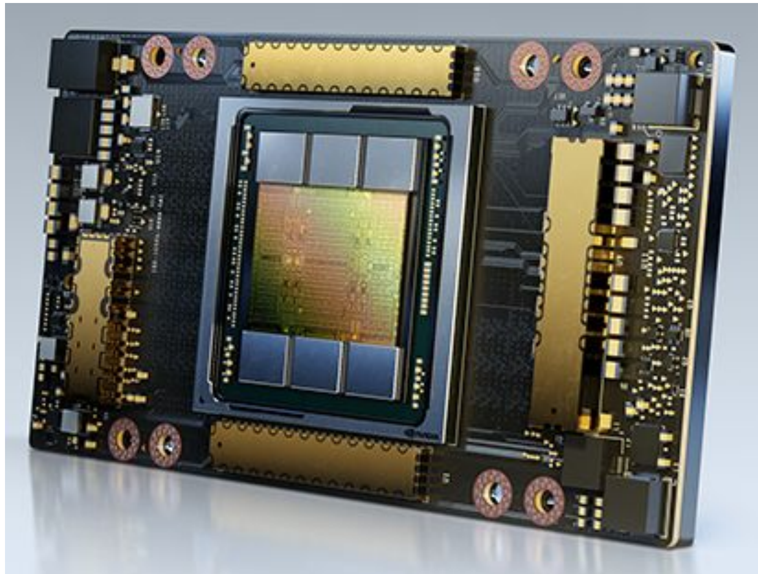


Tremendous investment

Latest and Greatest (and Largest) SoCs – NVIDIA

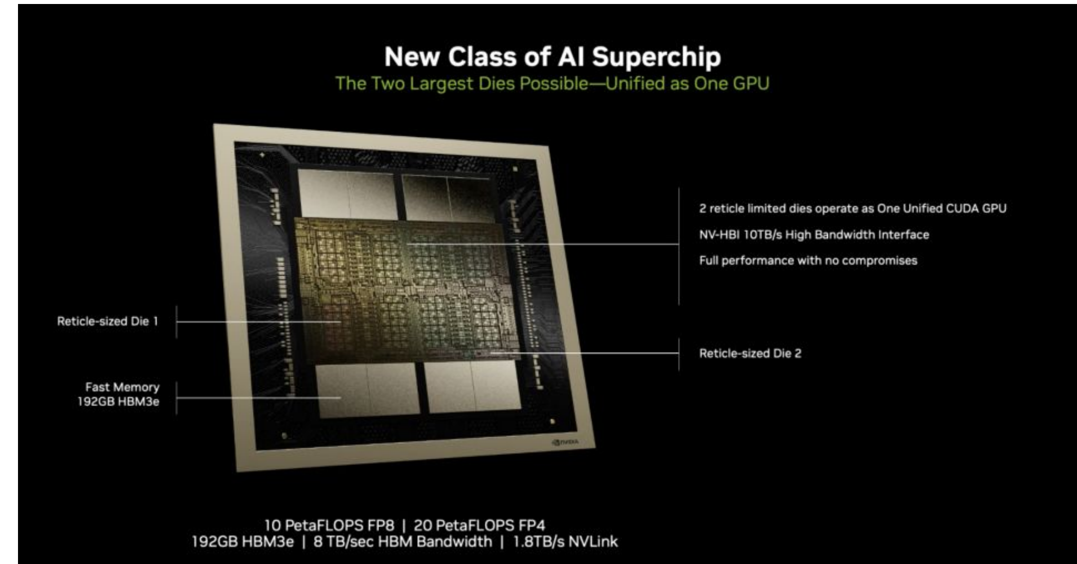
NVIDIA A100 GPU (2020)

- 54.2B transistors
- Area – 836 mm^2
- Full reticle field in 7 nm

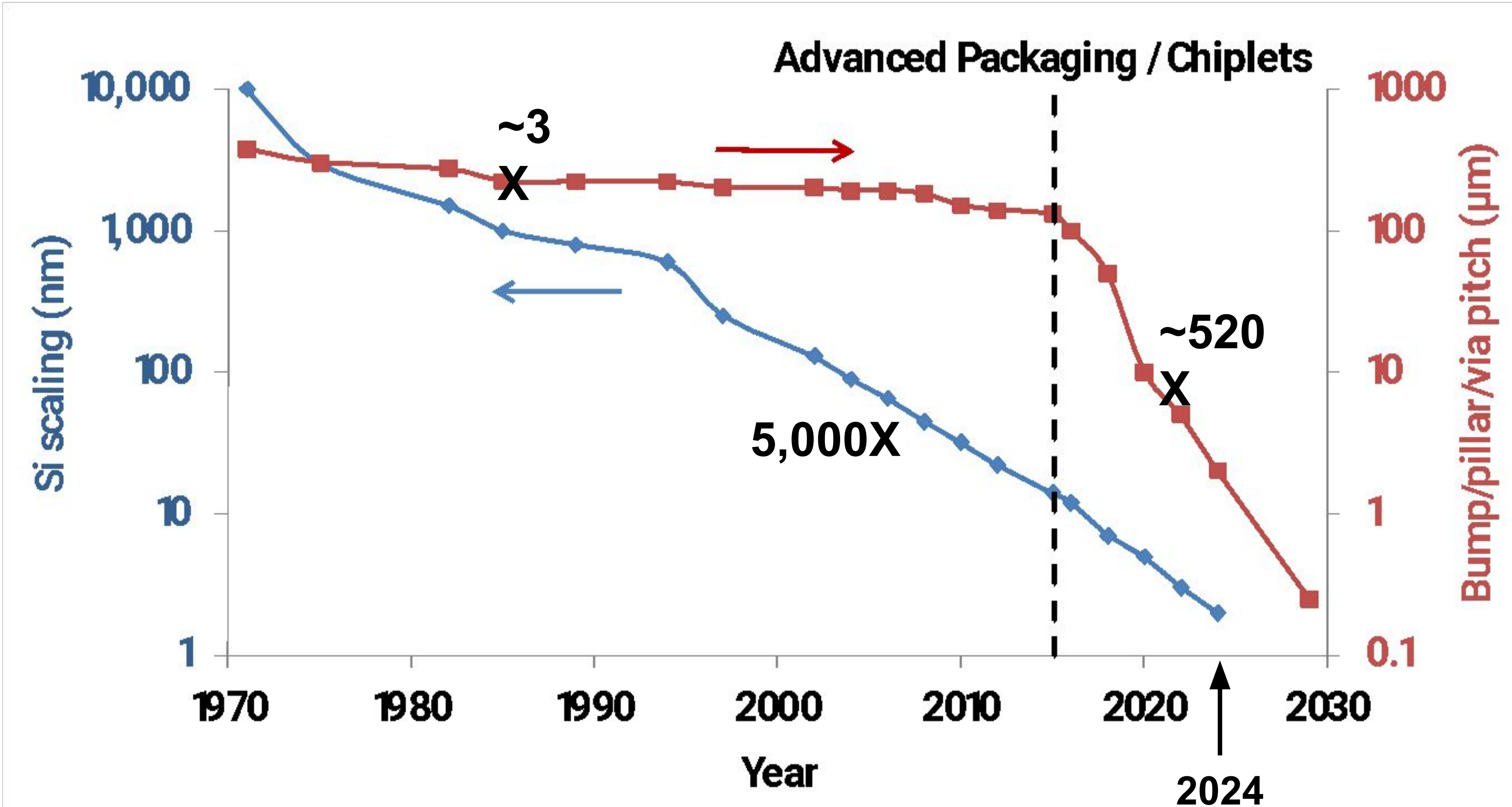


NVIDIA Blackwell GPU (2025)

- 92.2B transistors
- Area – 750 mm^2
- Two full-reticle dies stitched on interposer



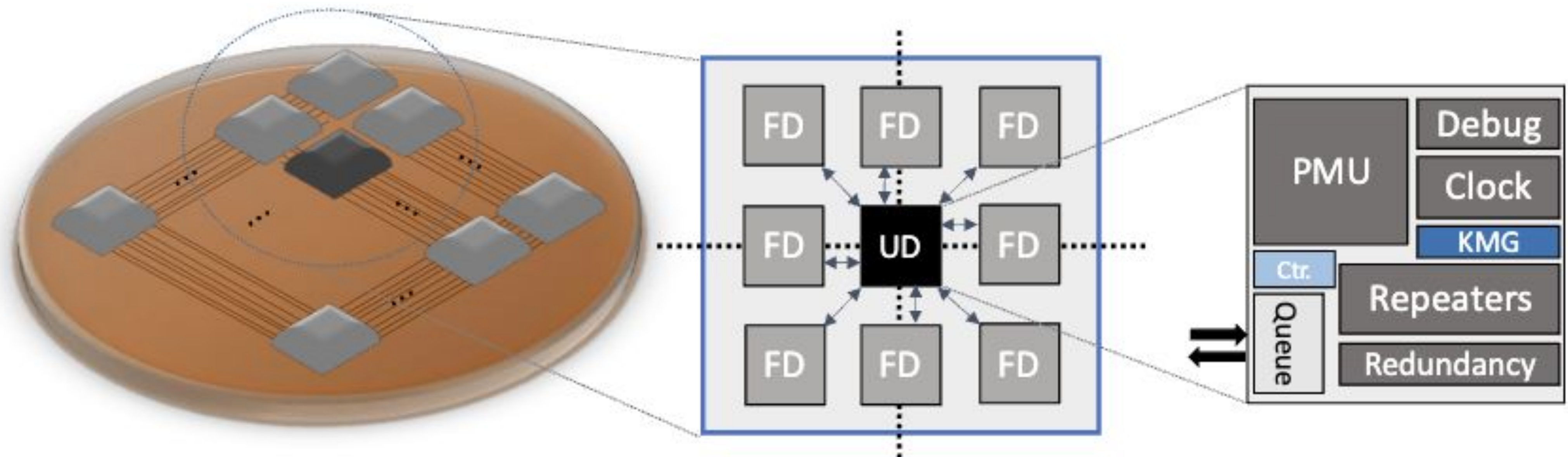
Heterogeneous Integration – System-Level Scaling



THInK Team
The Heterogeneous Integration Knowledge Team
UCI Samueli School of Engineering

Agenda

- The chiplets paradigm shift
- Silicon interconnect fabric
- AEPeX America
- Future outlook



The Chiplets Paradigm Shift

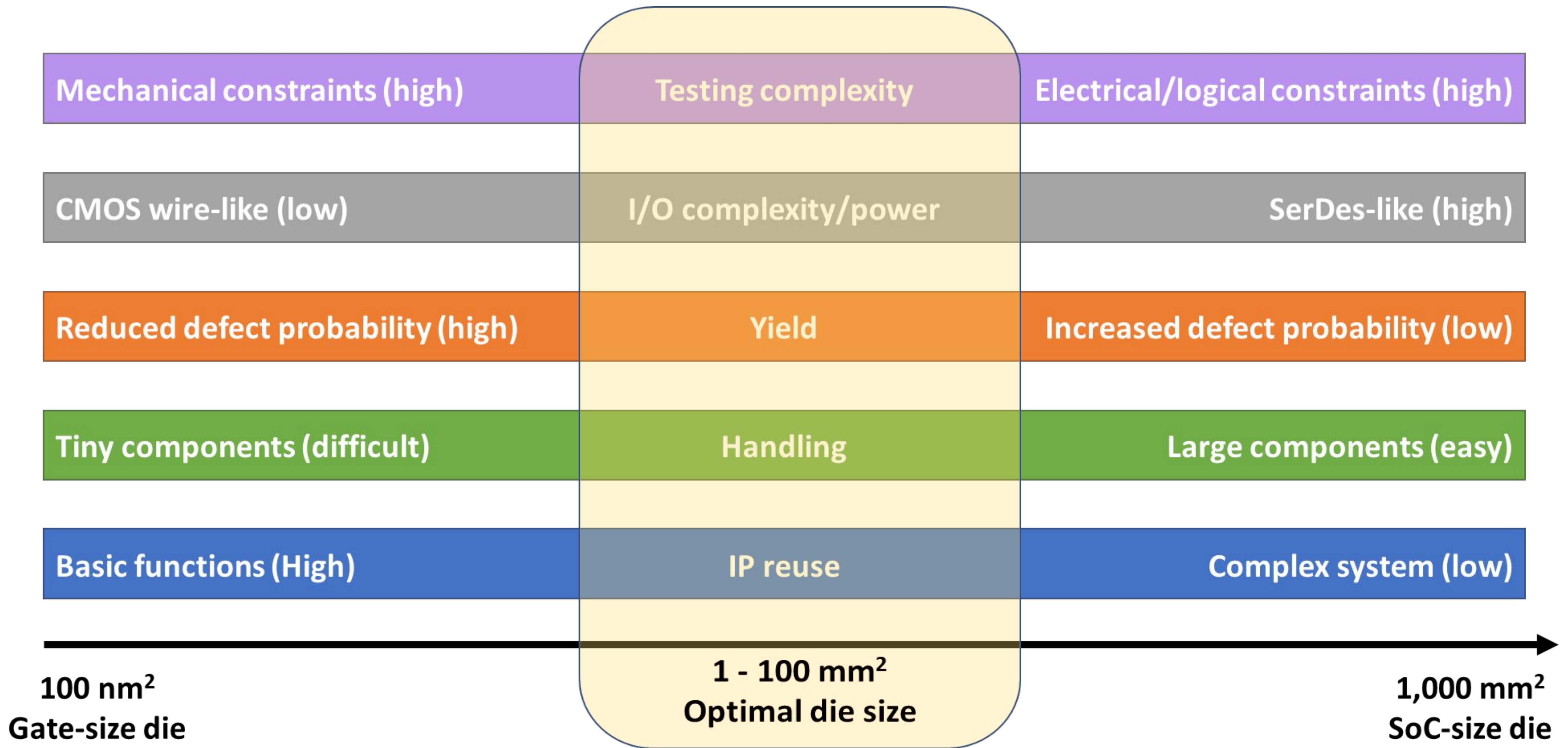
Stop scaling out single chips

- SoCs are limited
- Homogeneous
- High power
- Low TDP
- Complex SerDes
- High NRE
- Low yield

Let's play Lego

- Assemble small chiplets
- Heterogeneous
- IP reuse
- Low NRE
- High yield (not so simple)
- Shift complexity to system

What is the Right Chiplet Size?



S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon Interconnect Fabric: A Versatile Heterogeneous Integration Platform for AI Systems," *IBM Journal of Research and Development*, Vol. 63, No. 6, pp. 5:1-5:16, November/December 2019.

Challenges in Chiplet-Based Heterogeneous Integration

Classical VLSI revisited

- Not a comprehensive list

Technology

- Platform fabrication
- Chiplet integration
- TSV process
- Disparate processes
- CTE mismatches
- Cost
- Fabrication tools

Methodologies

- Partitioning
- Power delivery
- Synchronization
- Floorplanning and routing
- CAD tools
- Communication
- Built-in self-test

Circuits

- Multi-voltage/clock domains
- High efficiency and density power conversion
- Interconnect design
- Interface circuits
- IP reuse – eco system
- "Electronic storm"

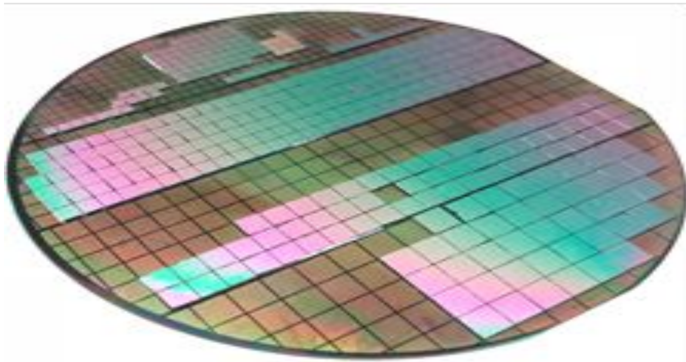
Architecture

- Choice and arrangement of chiplets
- Network design (à la NoC)
- Neuromorphic systems
- Thermal management

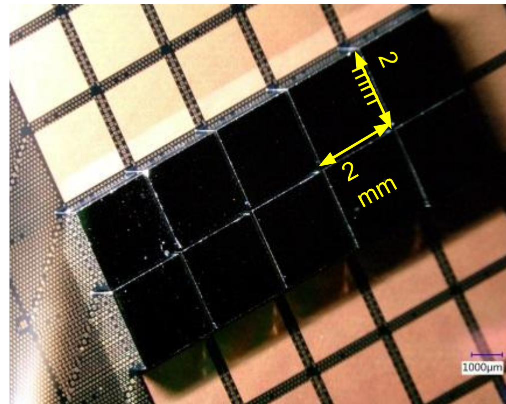
Silicon Interconnect Fabric (Si-IF)

- Replace traditional packaging
  - No package
  - No PCB
- Enable heterogeneous integration
  - Materials (Si, III-V, etc.)
  - Technology nodes
  - Size and aspect ratio

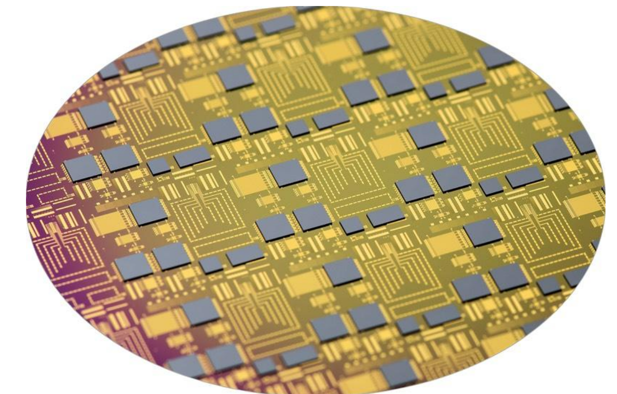
- Interconnect technology
  - Scaling interconnect is key 2 to 10 µm
  - Bare dies attached directly to Si-IF at <100 µm spacing using TCB
- **SoC-like SoW**



371 bare dies integrated on Si-IF using 10 µm pitch interconnect at 100 µm inter-die spacing.



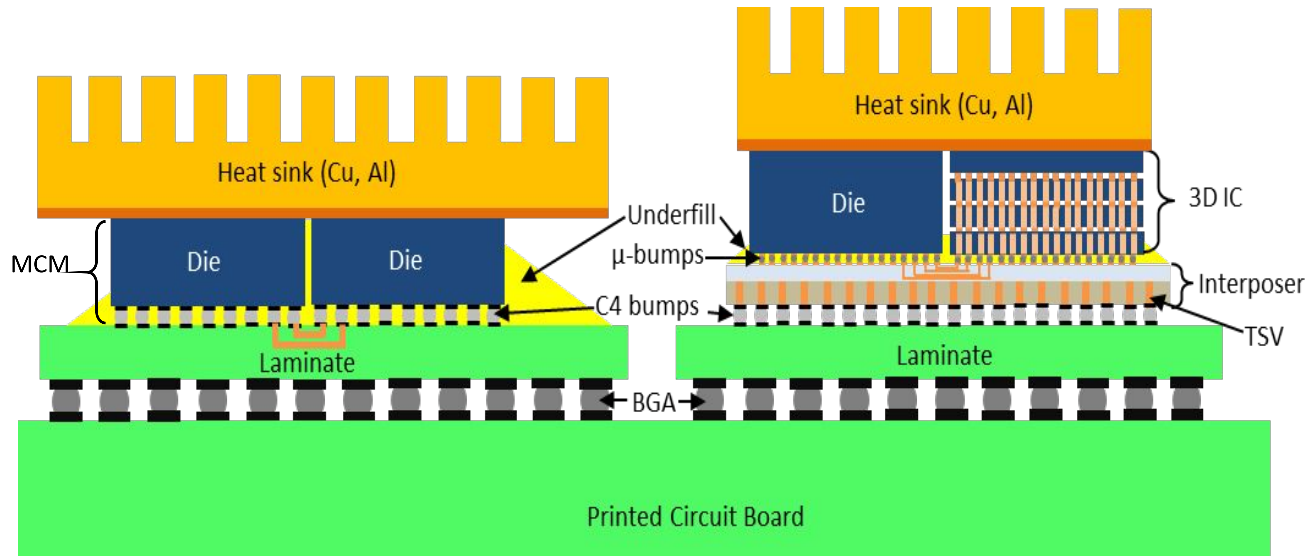
Ten daisy chain dielets (4 mm^2) interconnected on the Si-IF.

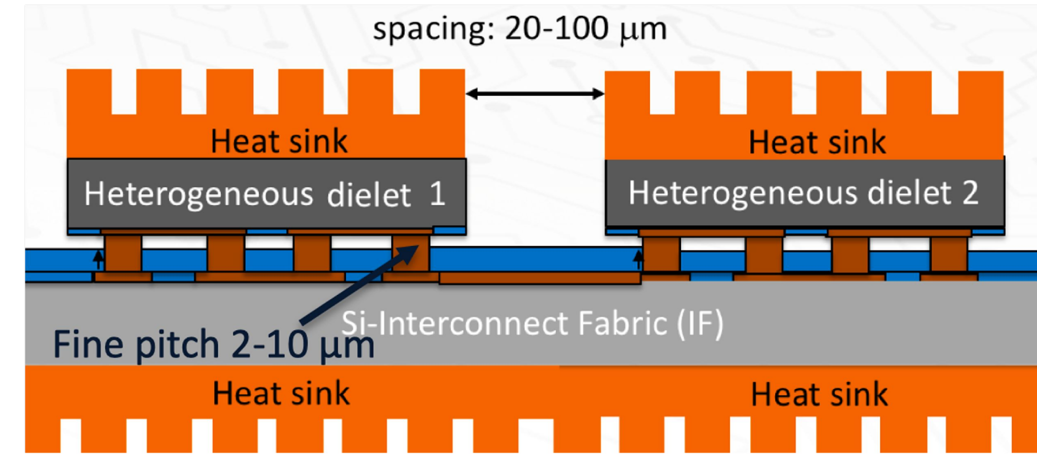


InP daisy chain dielets integrated on the Si-IF.

A. Bajwa *et al.*, *ECTC* 2017. E. Sorensen *et al.*, *ECTC* 2019. S. S. Iyer *et al.*, *IBM Journal of R&D* 2019. S. Jangam *et al.*, *TCPMT* 2021.

Traditional Packaging vs. Si-IF

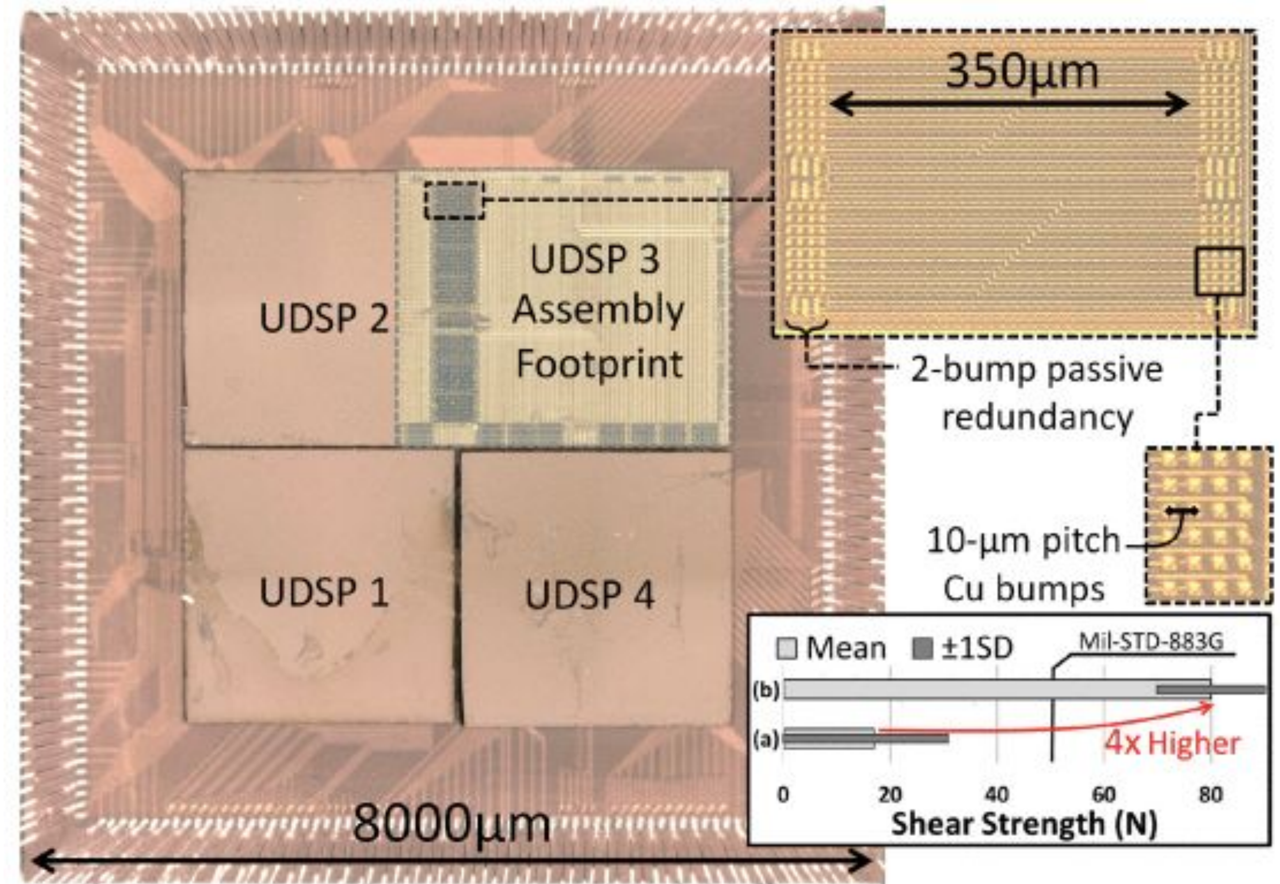




Parameter	Traditional packaging	Si-IF
Interconnect pitch	50 µm (interposer) - 1 mm (PCB)	2 - 10 µm
Inter-die spacing	Hundreds of µm – tens of mm	<100 µm
Hierarchy	Several packaging levels	Single packaging level
Materials	Disparate (Si, Cu, FR4, molding compound, etc.)	Mainly three (Si, Cu, and oxide)
Heat sinking	Limited	Excellent
Interconnect material	Solder-based	Metal-metal

System-Level Design Challenges – From Technology to Enablement

- Communication
- Power delivery and thermal management
- Testing
- Architectures
- Synchronization
- External communication
- Floorplanning
- Hardware security
- More…



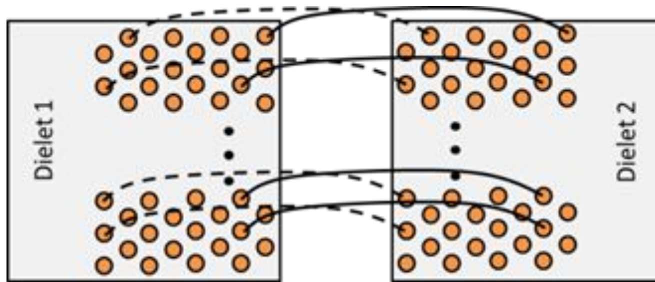
Micrograph of 2x2 UDSP assembly on the Si-IF, with a 10 µm channel footprint.

S. S. Nagi *et al.*, “A 16-nm 784-Core Digital Signal Processor Array, Assembled as a 2 × 2 Dielet With 10-µm Pitch Interdielet I/O for Runtime Multiprogram Reconfiguration,” *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 111-123, 2023.

Communication on Si-IF

Short range

- SuperCHIPS
- Other short-range protocols

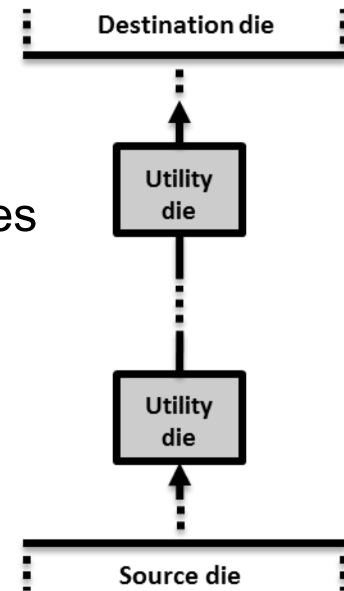


Mid range

- "One over" hops
- Exploit repeaters
  - On UCs
  - On FCs
- No load on global network

Long range

- UC-to-UC communication
  - Optical interconnect
  - RF communication
  - SerDes
- UC manages
  - Optimal routes
  - Alternative routes
  - Priorities



UC – utility chiplet, interchangeably with utility dielets (UDs)

FC – functional chiplet, interchangeably with functional dielets (FDs)

B. Vaisband, A. Bajwa, and S. S. Iyer, "Network on Interconnect Fabric," *Proceedings of the IEEE International Symposium on Quality Electronic Design (ISQED)*, 2018.

Short-Range Communication – SuperCHIPS

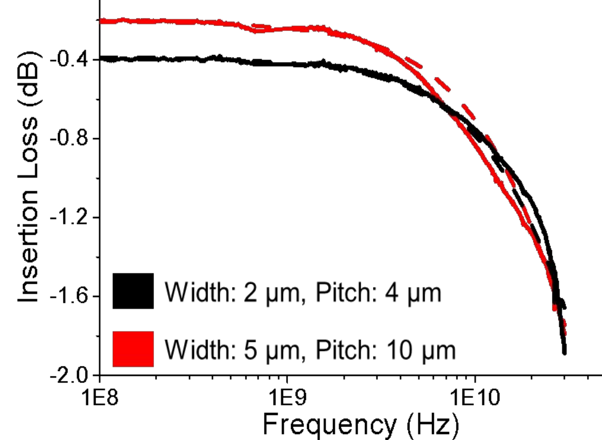
On-chip-like parasitics

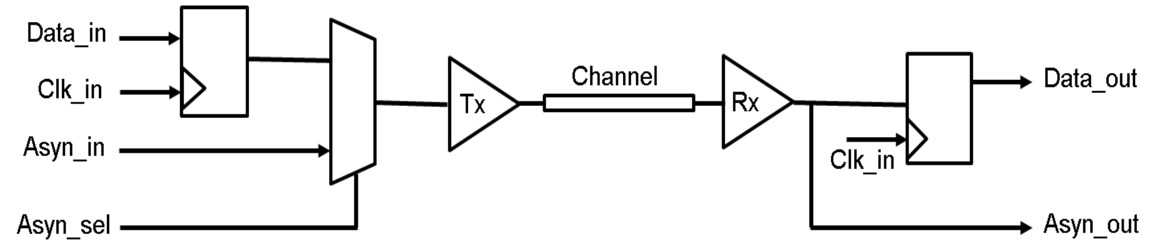
Horizontal interconnect	
Resistance (mΩ/µm)	9.1
Capacitance (fF/µm)	0.1
Inductance (pH/µm)	0.85
Conductance (Ω^{-1}/µm)	10^{-6}

Vertical pillars	
Resistance (mΩ)	50 to 70
Capacitance (fF)	3 to 4

Insertion loss: < –2 dB up to 30 GHz

Crosstalk: < –15 dB up to 20 GHz



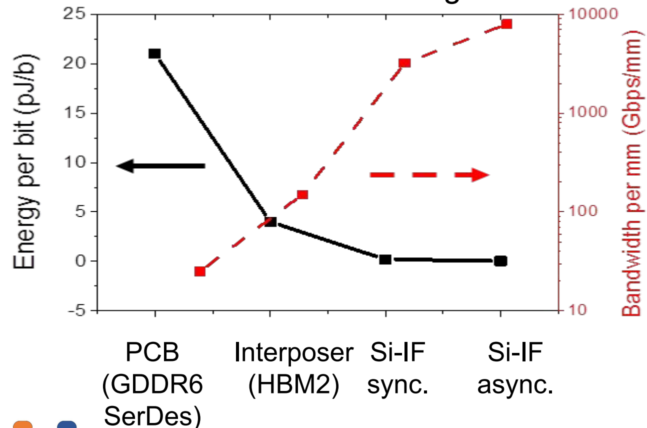


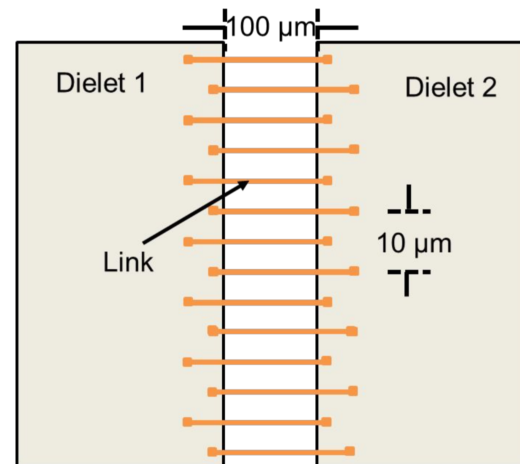
Energy per bit: 0.04 pJ/bit

- 20 to 100 times lower

Data bandwidth: 8 Tbps/mm

- 20 to 120 times higher





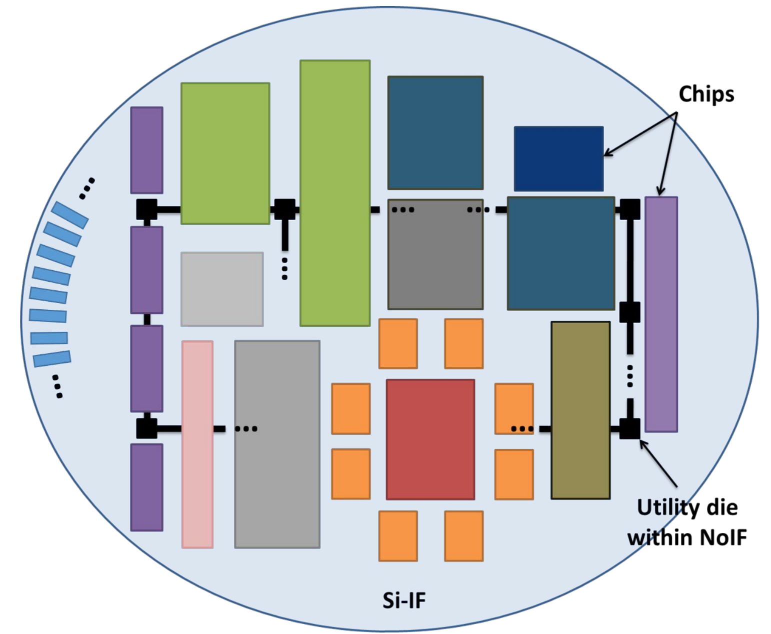
Interconnect pitch/protocol	10 µm on Si IF SuperCHIPS Async	10 µm on Si IF SuperCHIPS Sync	50 µm on Si Interposer HBM2	400 µm on FR4 PCB/ SerDes
No of signal links per mm	200		20	2.5
Inter-die distance (µm)	<100		<5,000	10,000
Overall latency (ps)	35	1 clock cycle	300	~1,000
Max data-rate/link (Gbps)	10	4	2	40
Energy per bit (pJ/b)	<0.04	<0.2	6	23.2
Max Bandwidth per mm* (Gbps/mm)	8,000	1,600	150	25

Four layers of wiring on the Si-IF

S. Jangam *et al.*, "Demonstration of a Low Latency (<20 ps) Fine-pitch (≤10 µm) Assembly on the Silicon Interconnect Fabric," *Proceedings of the IEEE Electronic Components and Technology Conference (ECTC)*, pp. 1801-1805, 2020.

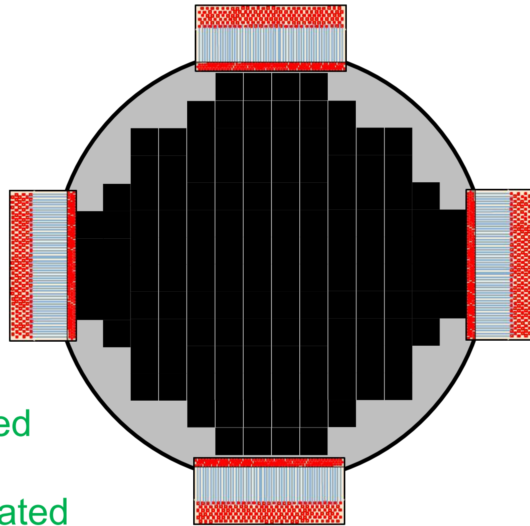
Network on Interconnect Fabric (NoIF)

- Borrows concepts from NoC
  - Communication
  - Queuing
  - Prioritization
  - Quality of service
- Supports system-level integration
  - Power management
  - Global interconnect
  - Test
  - Redundancy allocation
  - Rerouting
  - Synchronization
  - More…
- Based on intelligent utility chiplets (UCs)
  - UCs are nodes of the NoIF
- Si-IF is passive – everything is a chiplet
- Hierarchical approach
  - A UC consists of multiple dies
  - Legacy IP can be reused (*e.g.*, SerDes)



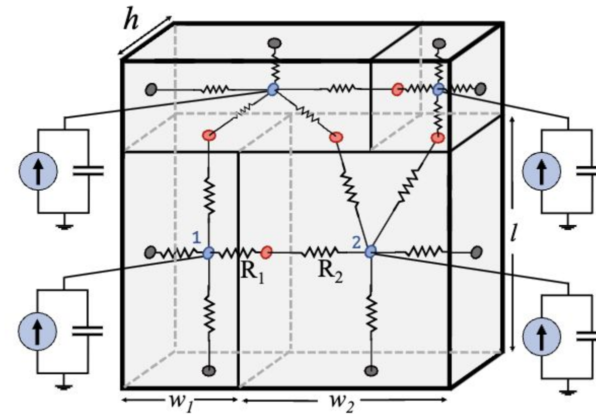
B. Vaisband, A. Bajwa, and S. S. Iyer, "Network on Interconnect Fabric," *Proceedings of the IEEE International Symposium on Quality Electronic Design (ISQED)*, 2018.

Other Enablement Aspects

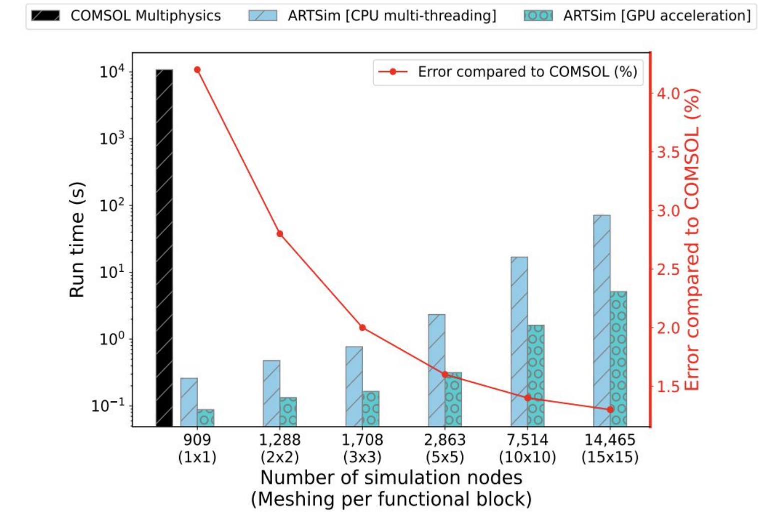


FlexCon (Copper links)

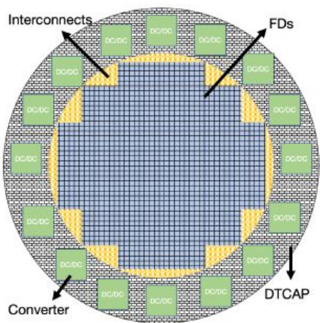
- Minimal space occupied
- Components (*e.g.*, SerDes) can be integrated into FlexCon
- Limited reach



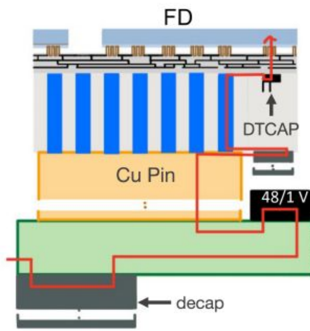
Node reduction scheme to reduce run time.



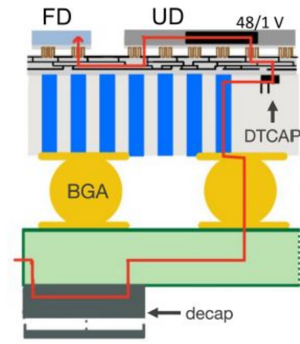
Steady-state simulation run time as a function of the number of nodes within the thermal netlist



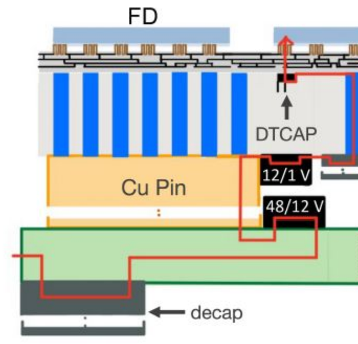
(a)



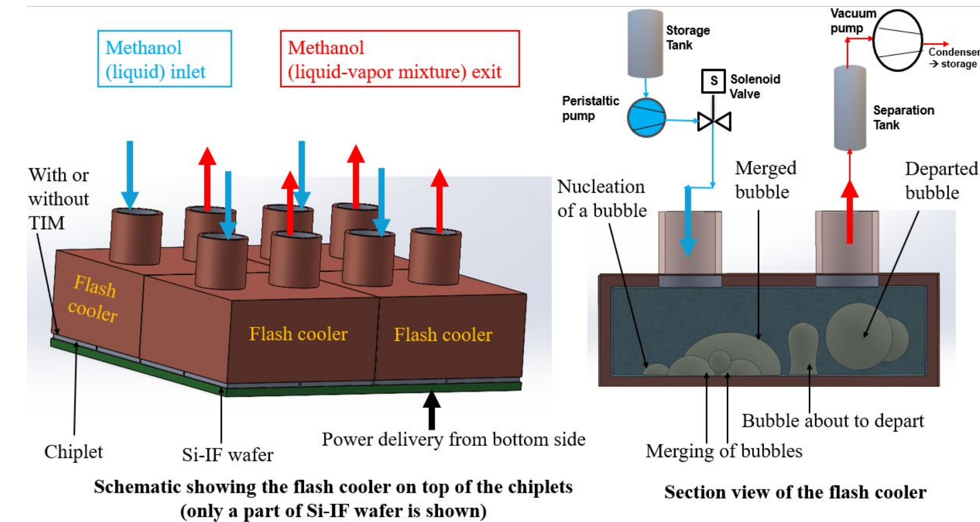
(b)



(c)



(d)



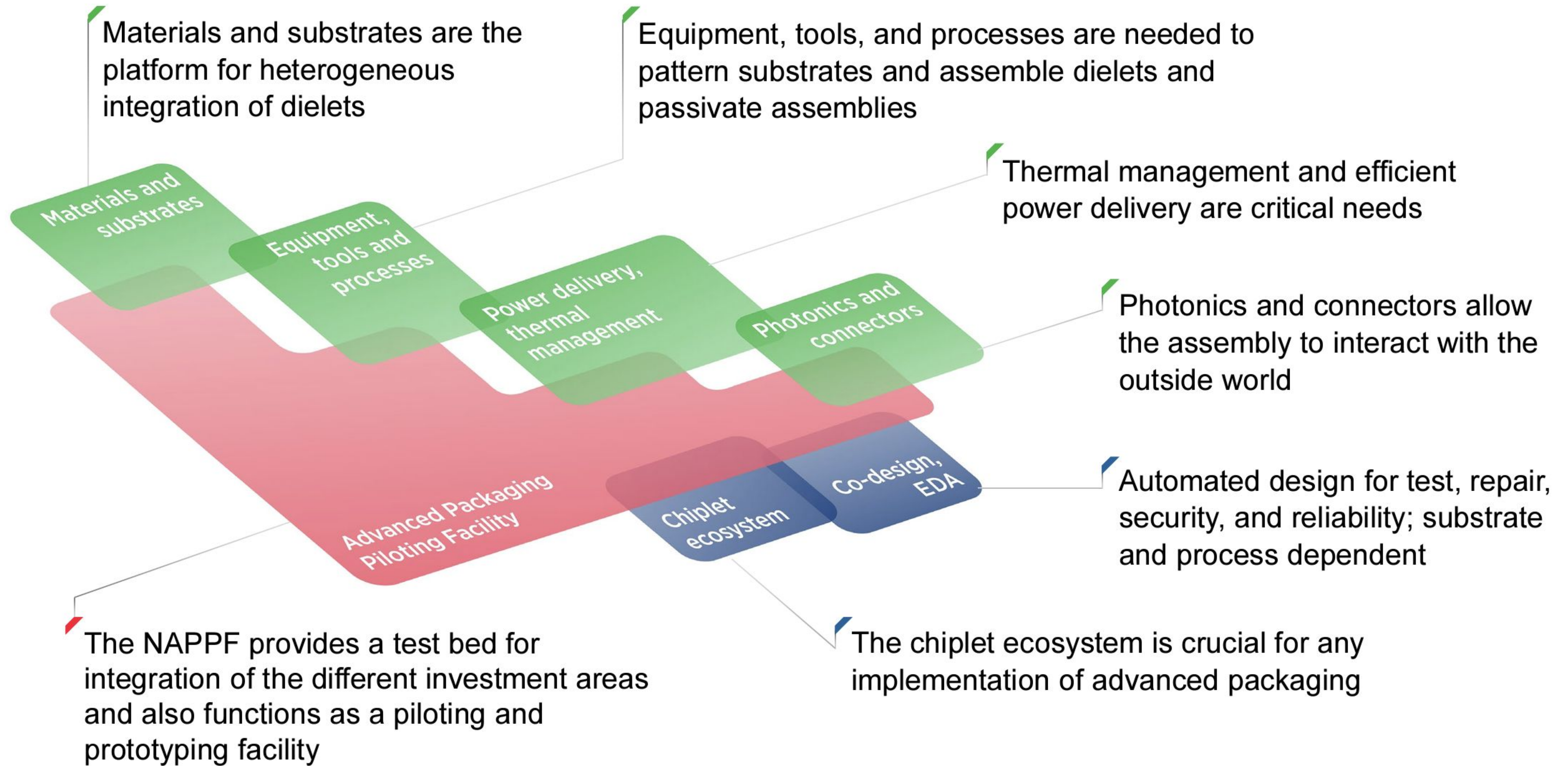
Schematic showing the flash cooler on top of the chiplets (only a part of Si-IF wafer is shown)

Section view of the flash cooler

Agenda

- The chiplets paradigm shift
- Silicon interconnect fabric
- AEPeX America
- Future outlook

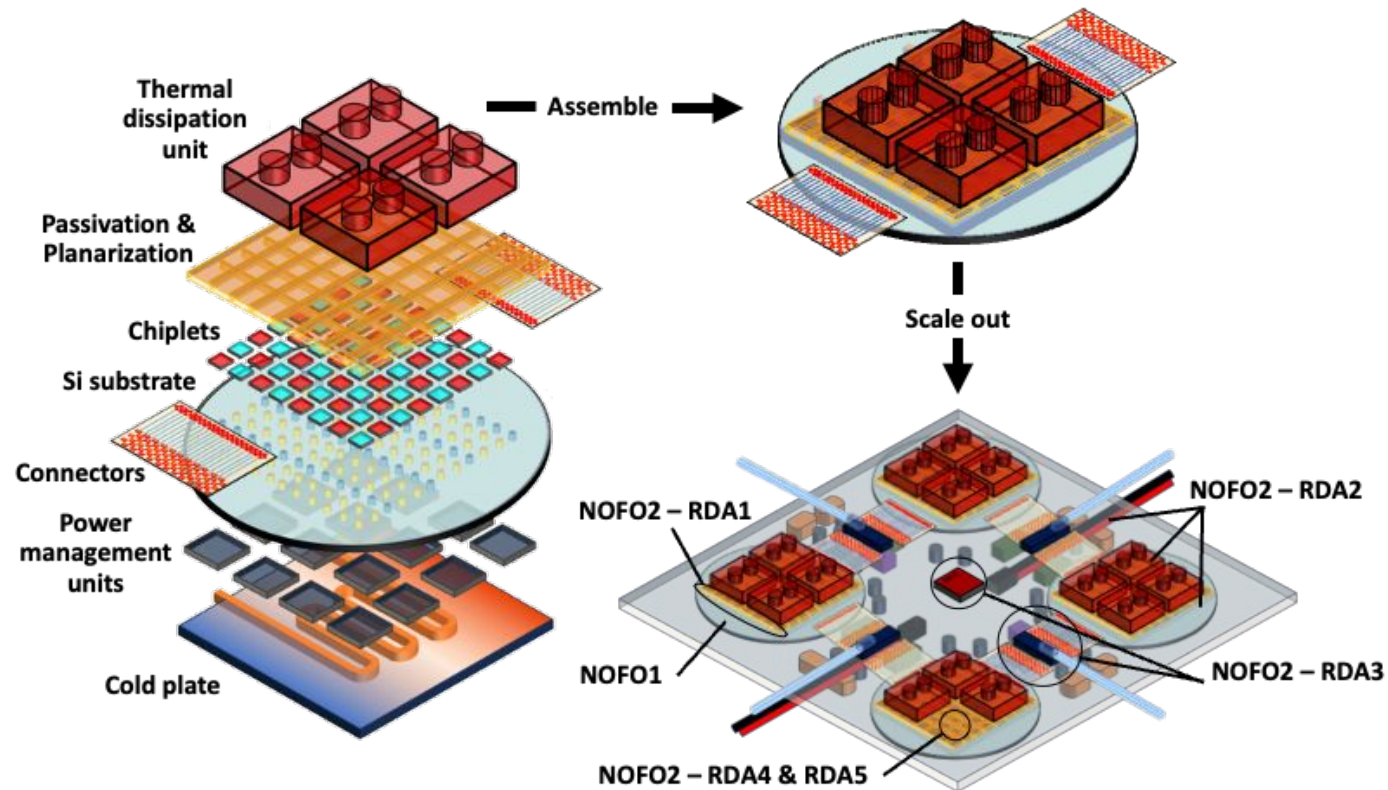
National Advanced Packaging and Manufacturing Program (NAPMP)



AEPeX America

Advanced Electronic Packaging eXchange for America

- An organic coalition of over 60 industrial and academic partners
- Access to critical infrastructure, vendors, and technical expertise
- A vision to bring the national strategy for advanced packaging to life



Cross-RDAs NAPMP vision by AEPeX America.

Scale-down, scale-out, and wire abundance.

Future Outlook

- Excellent time for chiplets and heterogeneous integration!
- Interdisciplinary research is a necessity
  - Driven by applications
- Silicon interconnect fabric
  - Increase system performance by 2-3 orders of magnitude
  - Higher bandwidth, lower latency, lower power, more memory
  - Several system-level design challenges to address
- Reduce NRE with IP reuse shorten time-to-market
- Main challenge is the eco system

THANK YOU