Designing a 2000 Chiplet Waferscale Processor

PUNEET GUPTA

ECE DEPT., UCLA

ACKNOWLEDGEMENTS: CDEN, CHIPS CENTERS AND QUALCOMM INNOVATION FELLOWSHIP FOR FUNDING

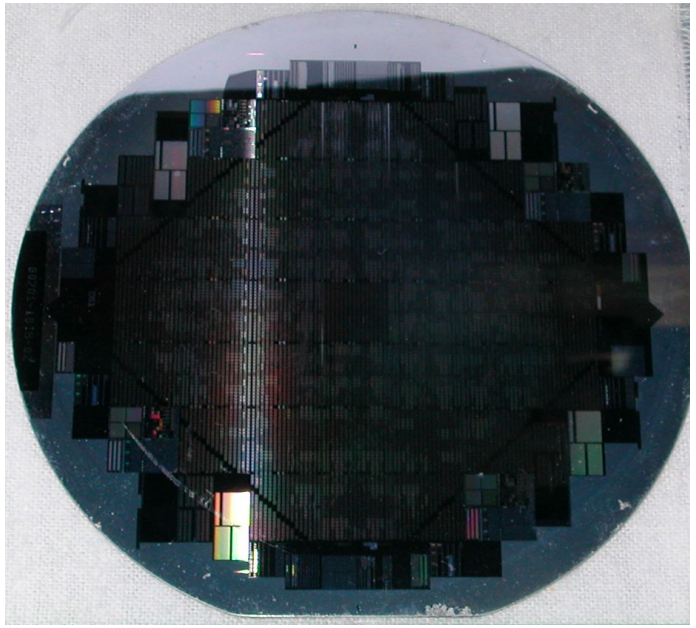
STUDENTS @UCLA: SAPTADEEP PAL, IRINA ALAM, KRUTIKESH SAHOO

COLLABORATORS: UCLA: SUBRAMANIAN S. IYER, SUDHAKAR PAMARTI. UIUC: RAKESH KUMAR

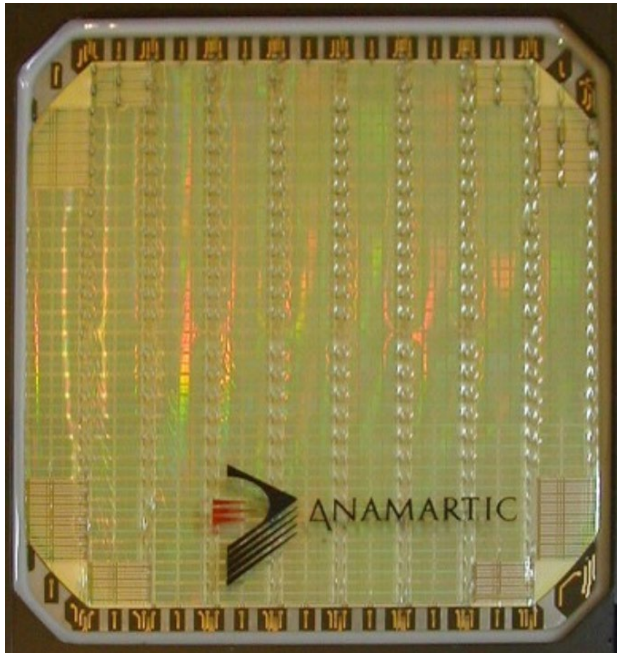
What are Waferscale Processors ?

- Processors that span a full silicon wafer
  - 100mm wafer ~ 7900mm^2
  - 200mm wafer ~ 31,400mm^2
  - 300mm wafer ~ 70,000mm^2
- Comparison: largest System on Chip ~ 800mm^2
- Challenge: fabrication, packaging, design, architecture, test *all* is tailored to serve at most 800mm^2
- This talk
  - Why even bother building waferscale systems ? → A case study of benefits
  - How do we address the myriad of daunting challenges in designing waferscale systems ? → an early attempt at solving and designing a waferscale system

A Brief History of Waferscale Computing



Gene Amdahl's Trilogy Systems



Tandem Computers, Fujitsu

Other efforts: ITT Corporation, Texas Instruments. Recent efforts: Spinnaker (Neuromorphic Chip)

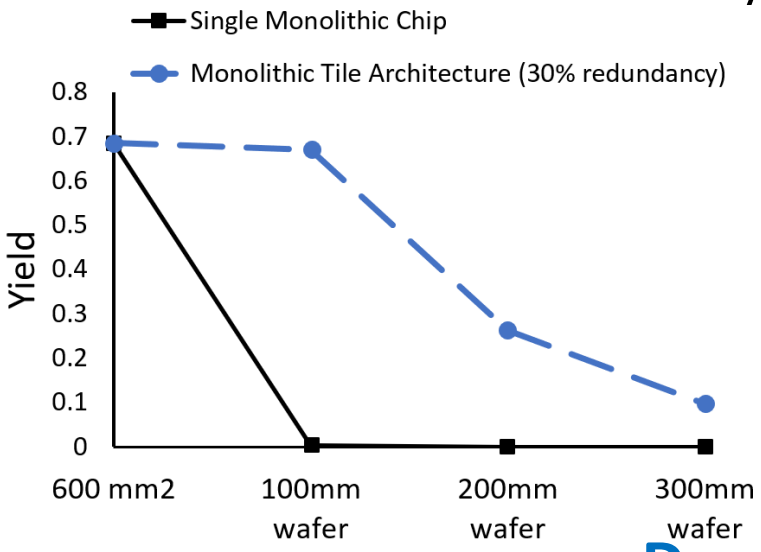
What Happened to Waferscale Integration?

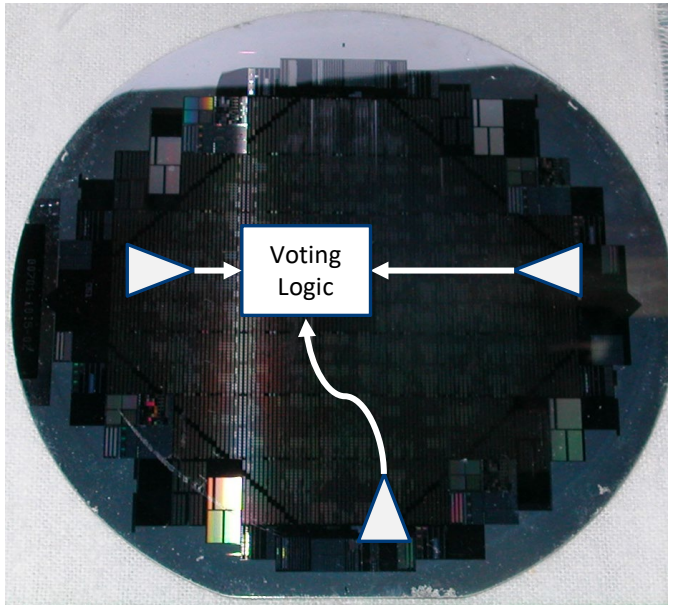
Didn't work out (e.g., Trilogy Systems was one of the biggest financial disasters in Silicon Valley before 2001)

Their Approach to Waferscale: **Monolithic**

Area of chip ↑

Probability of defects ↑



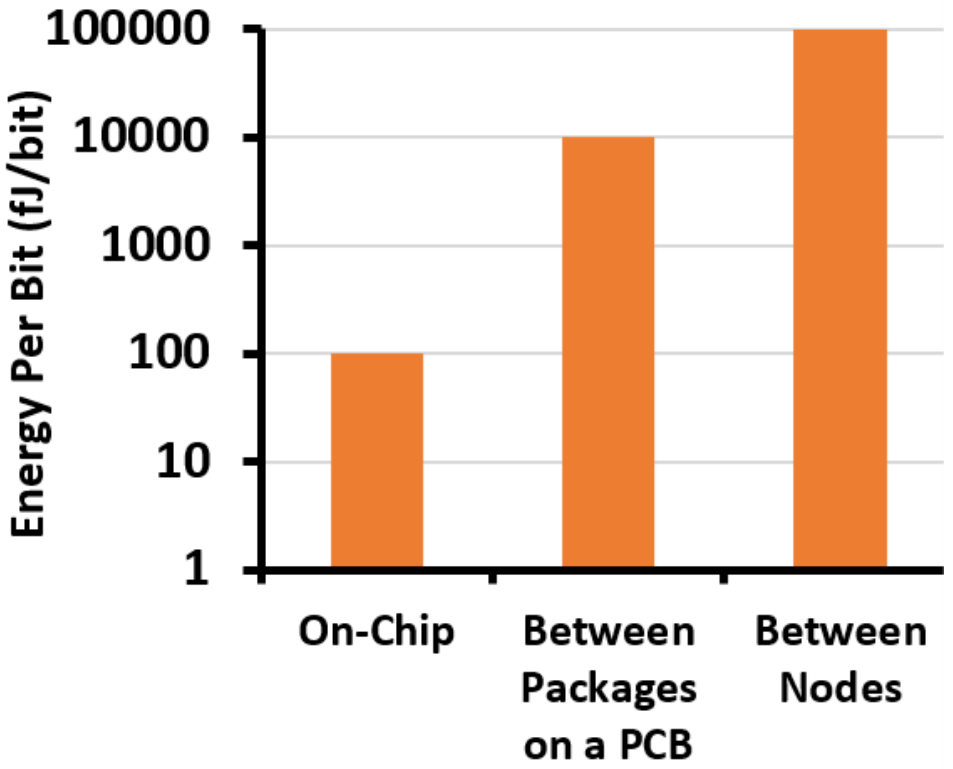


Some mitigation possible through TMR, etc. - but prohibitively expensive

Deemed commercially unviable

Time to Give Waferscale Another Go?

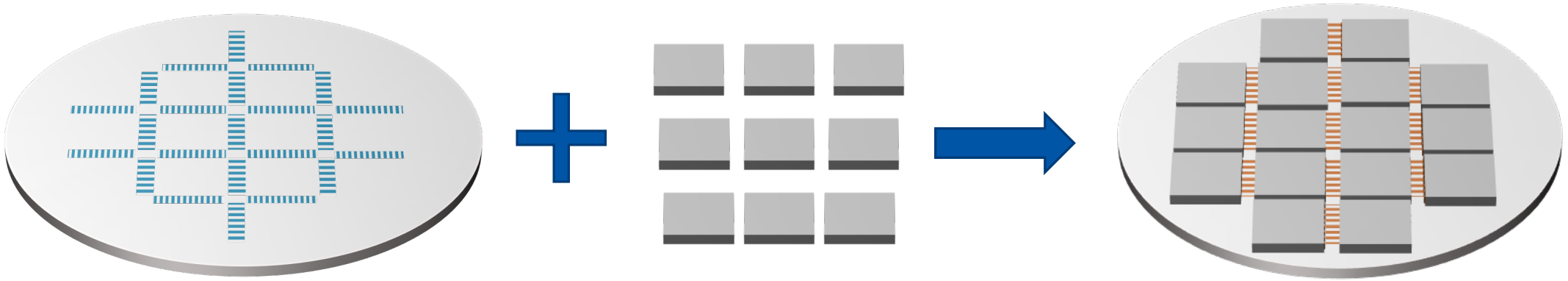
- Highly parallel applications are spread across many processors
- Communication between the processors is still a big bottleneck
  - Low Bandwidth (a few 100s of GBps)
  - High energy per bit (10s of pJ/bit)
  - Real estate on chip (15-25% of the chip is devoted to SERDES I/Os)



Re-imagining Waferscale Integration

Q: What do we need from waferscale integration?

A: High density interconnection



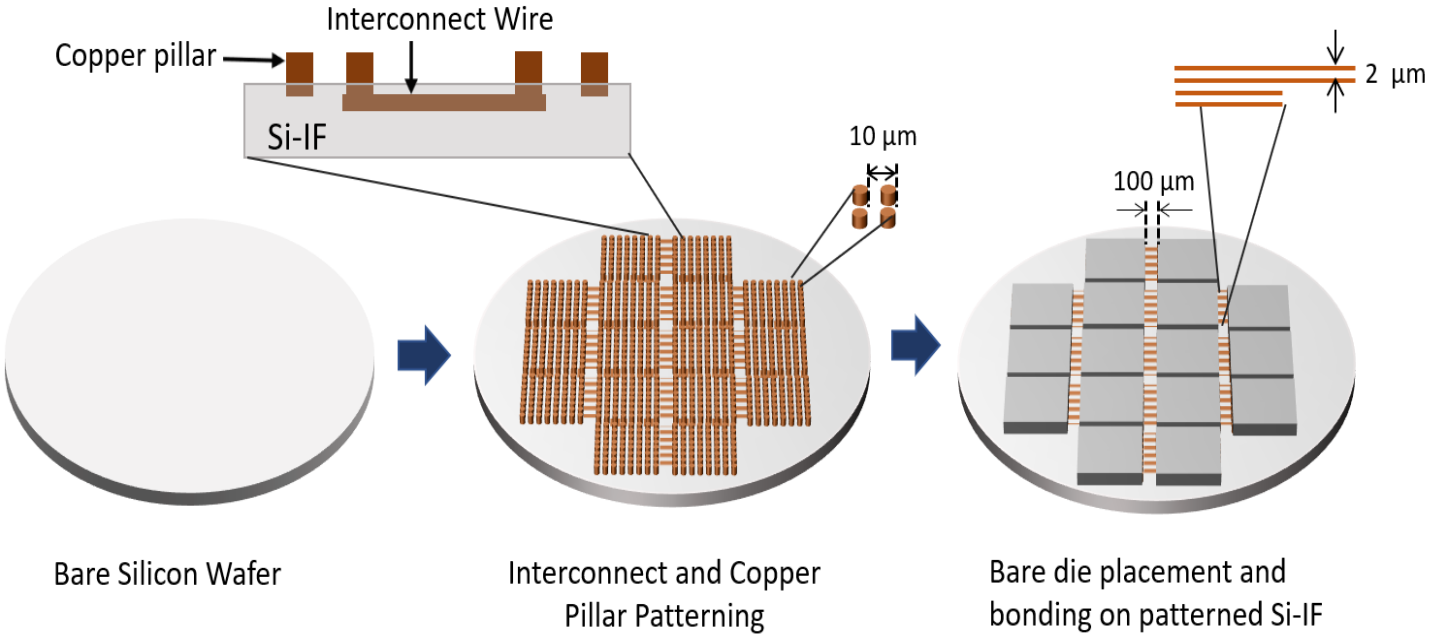
A wafer with interconnect wiring only

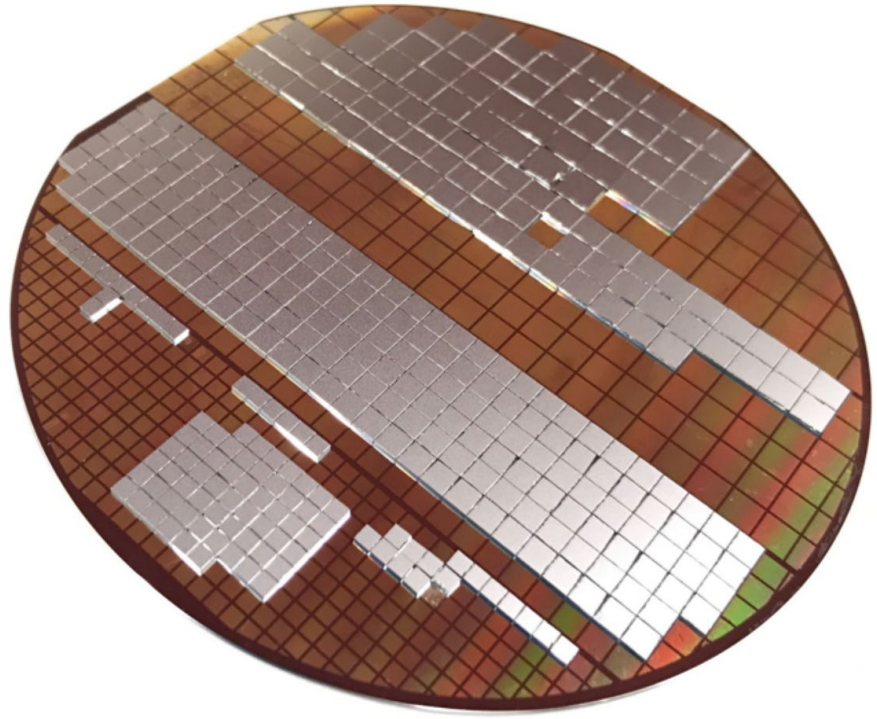
Small known good dies

Bond the dies on to the interconnect wafer

Enabling WSI Technology

UCLA Silicon Interconnect Fabric (Si-IF)*





Measured Bond Yield >99%

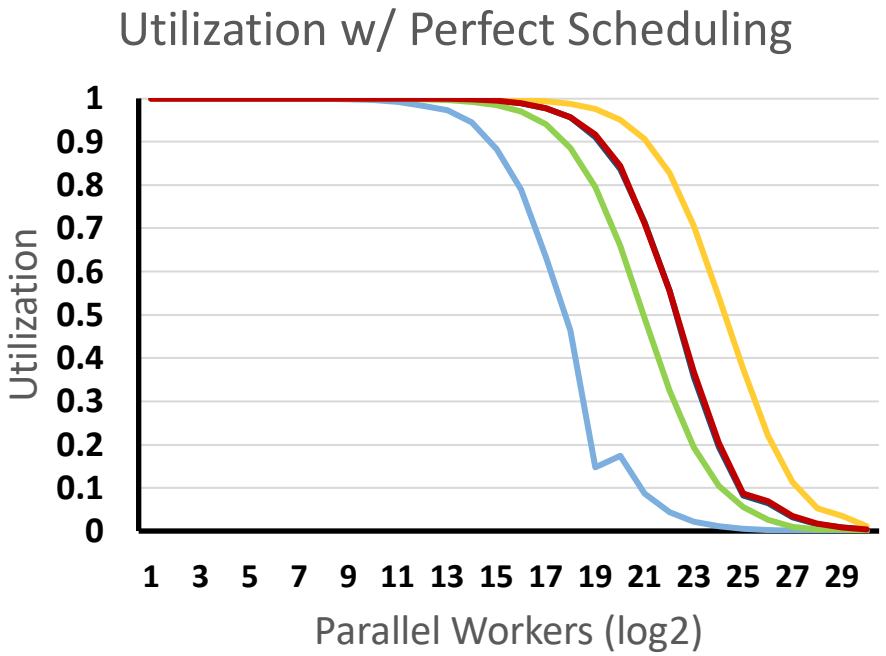
Allows waferscale integration with high yield

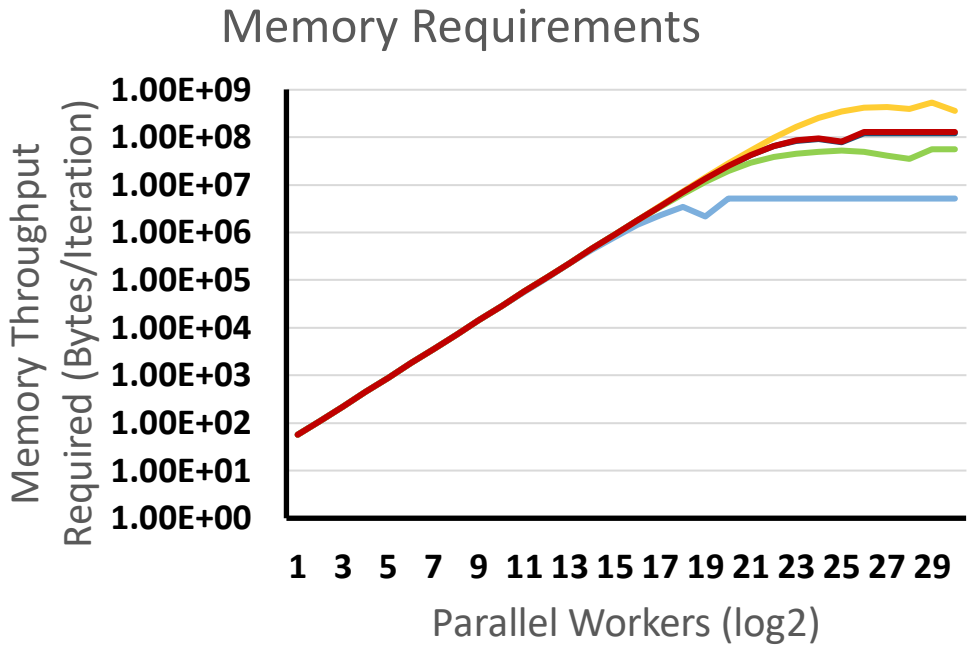
*UCLA CHIPS Programme: https://www.chips.ucla.edu/research/project/4

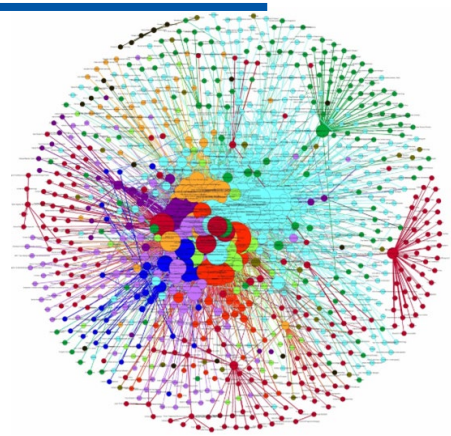
Designing a Waferscale Graph Processor Prototype: Challenges and Solutions

[Appeared in DAC'21, ECTC'21]

Graph Applications Have Unique Characteristics





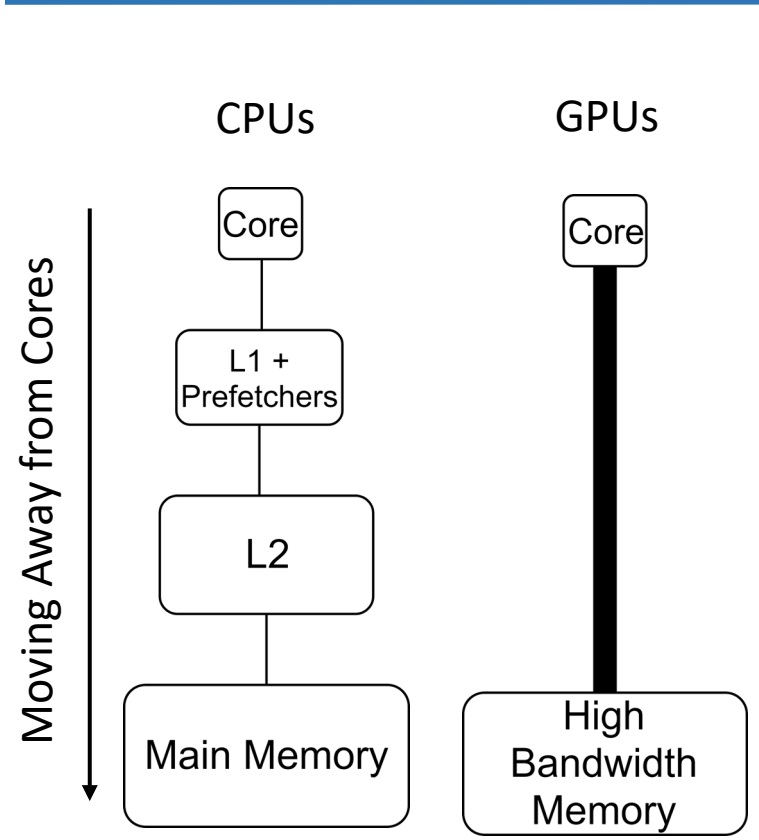


Massive number of processing cores needed

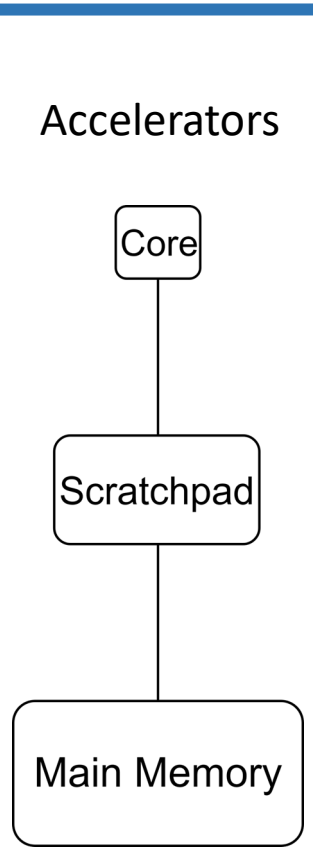
Bandwidth beyond 50 TBps @ 10 ns/iteration

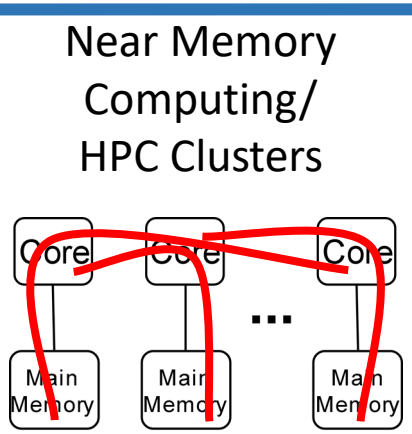
—hollywood-2009 —sk-2005 —soc-Slashdot0902 —webbase-2001 —wikipedia-20070206

Graph Processing Requires a New Architecture



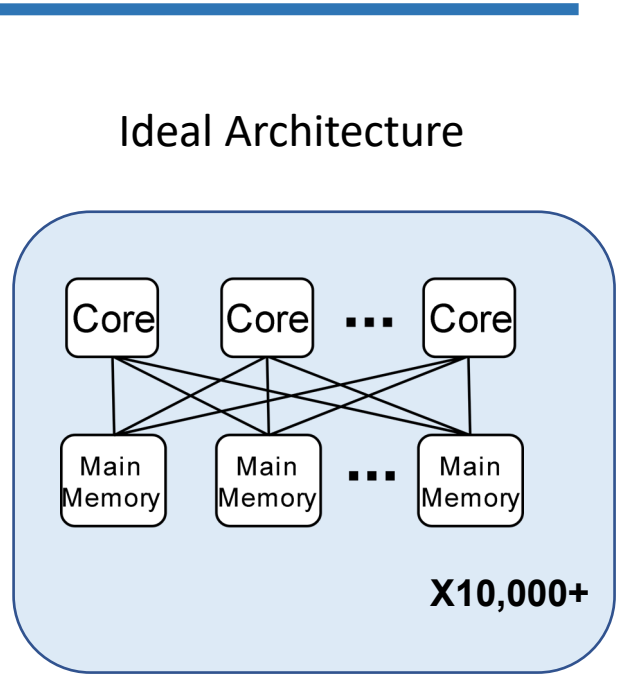
Requires Locality





Requires Predictable Layout

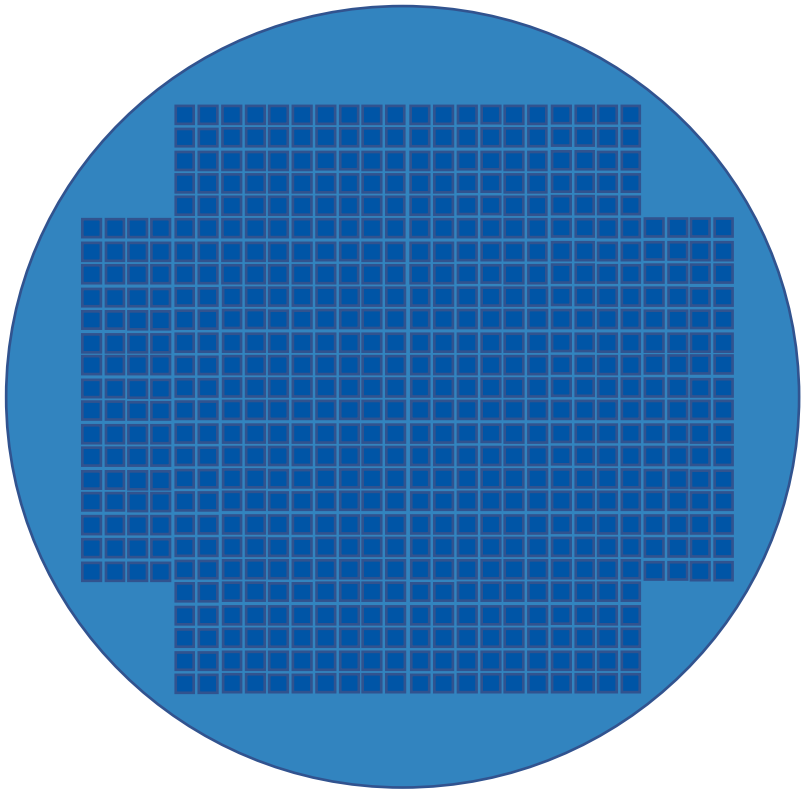
Requires Expensive Pre-Processing

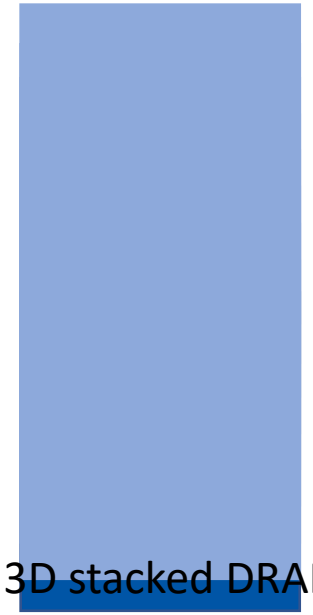


Requires Nothing

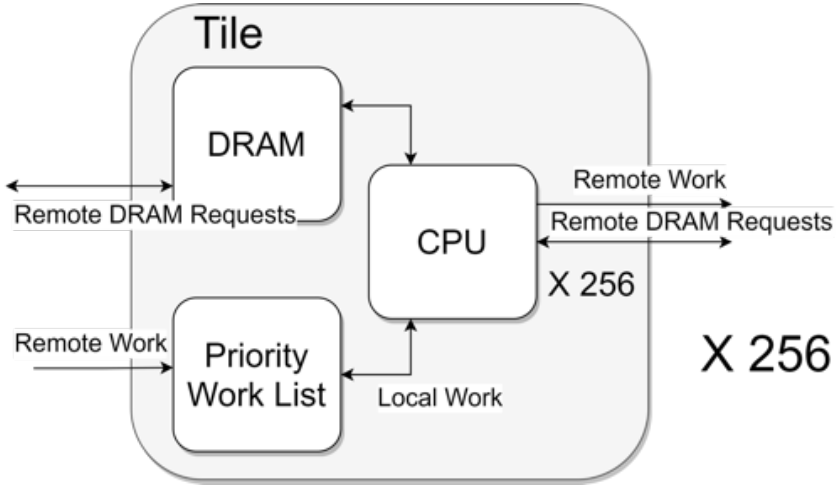
Provides High Random Access Bandwidth

Waferscale Graph Engine Overview





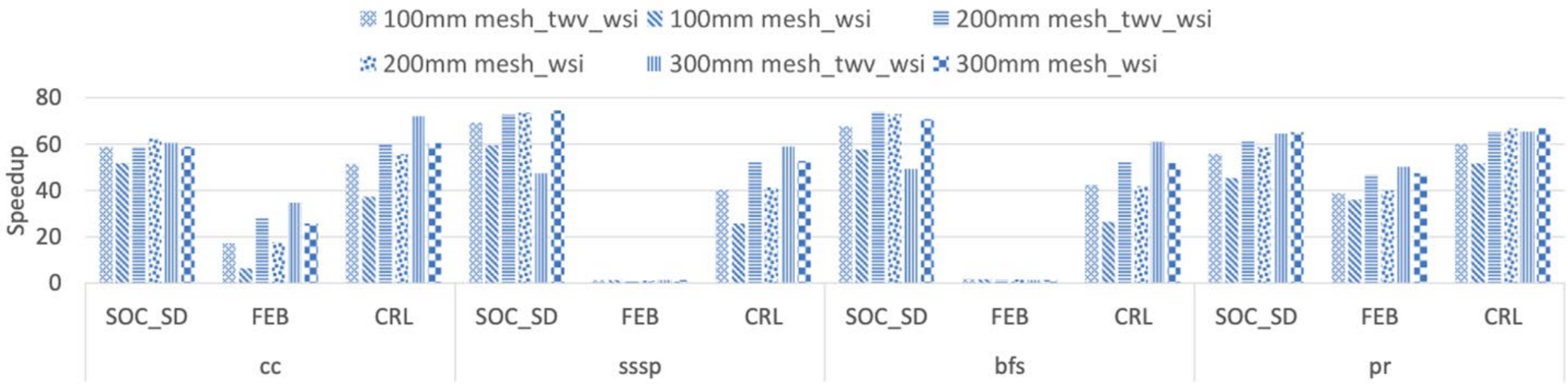
Node: 3D stacked DRAM with Compute in the logic layer



- **Area = 110 mm²**
- **Power = 35 W**

300 mm wafer has enough area for about **480** 3D-stacked Node

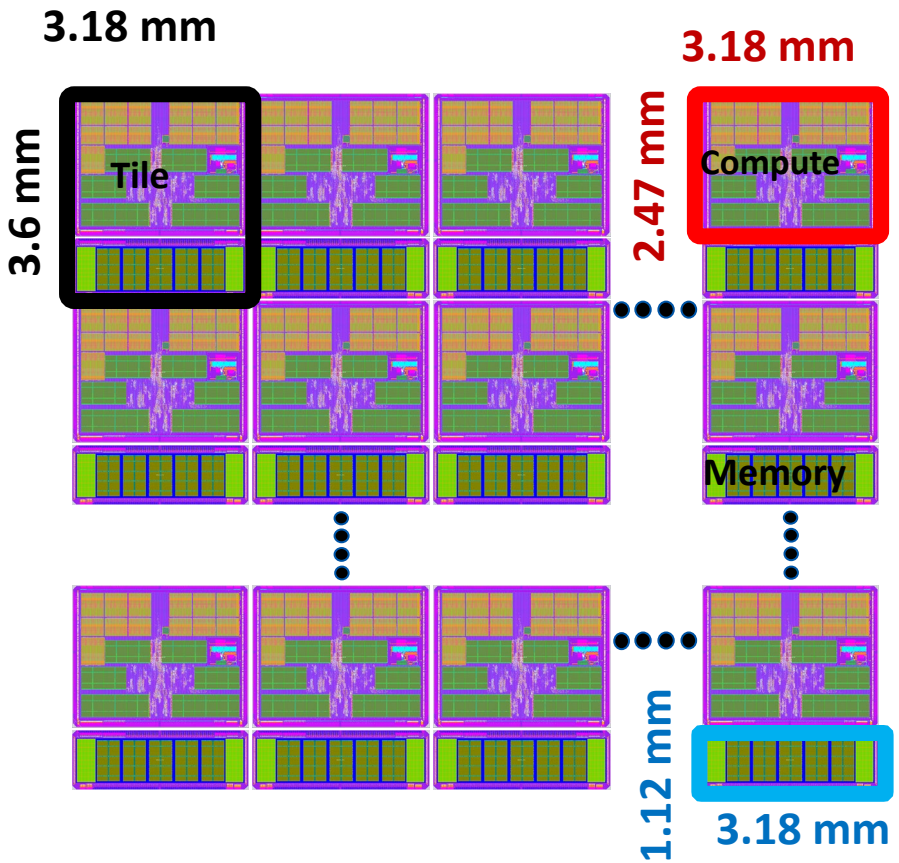
Speedup compared to a Multi-Chip Interposer Baseline



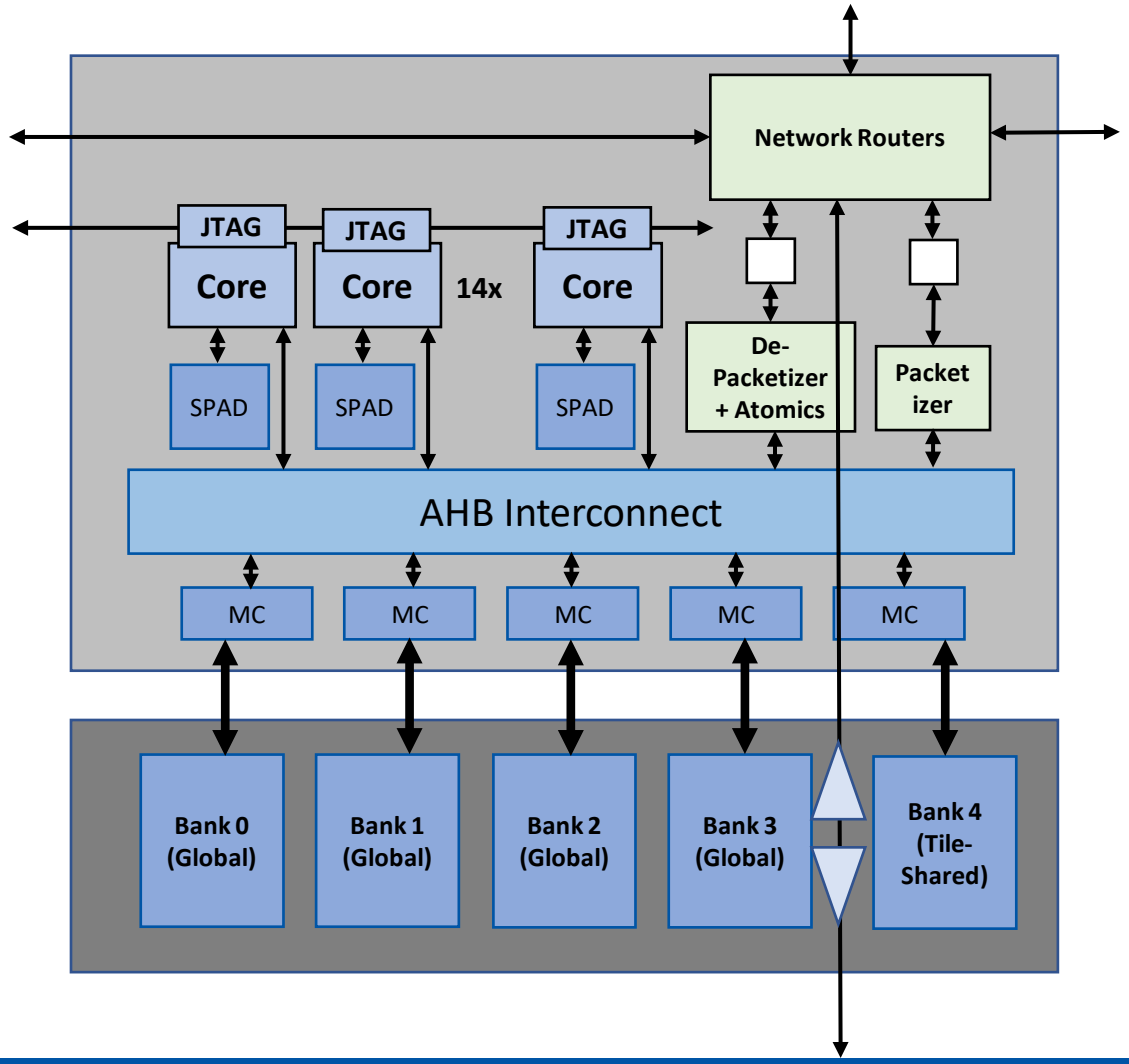
- Up to **60-70x** speedup for 300 mm architecture compared to an multi-MCM baseline

Building a (Simplified) 1024-Tile Architecture

1. Two dies per tile:
   - ***Compute die*** – 7.86 mm^2
   - ***Memory die*** – 3.6 mm^2
2. Implemented in **TSMC N40-LP**
1. Tiles : **1024 (Total 14,336 Cores) → 2048 Chiplets**
1. Total Memory Bandwidth (Data only) : **23.35 TB/**s
1. Total Network Bandwidth (Data only) : **9.83 TB/s**
1. Total Compute : **4.3 TOPs**
1. Power : 300 mW (Per Tile), **700 W** (total including losses)
   Peripheral Power and Signal Delivery



Tile Micro-Architecture



1. Architecture:
   - ***Compute die*** – **14x** ARM CORTEX-M3 core
     64KB Private SRAM per cores
     Custom Network Infrastructure
     Clock Management
     JTAG Infrastructure
   - ***Memory die*** – **5x** 128KB Globally Shared SRAM
     Feedthrough network interface

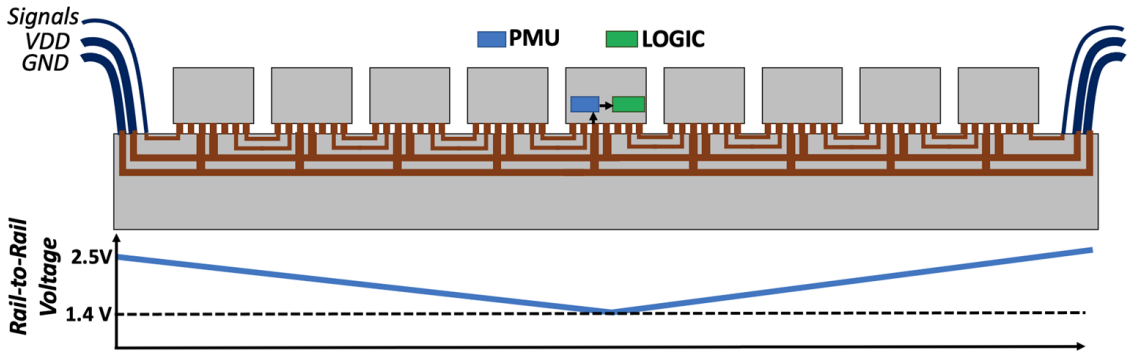
2. Unique Features:
   - Support for compare-and-swap atomic operation
   - Packet priority schemes to avoid network deadlocks
   - Dual network for fault tolerance

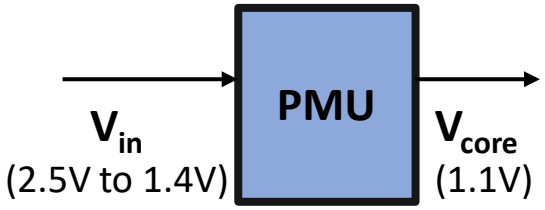
Challenges Faced While Designing the System

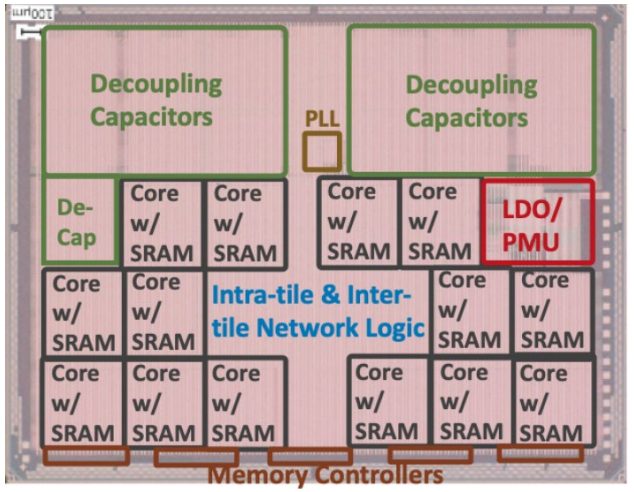
1. How should we **deliver power** to all the flip-chip bonded chiplets across the wafer?
2. How can we reliably **distribute clock** across such a large area?
3. What is the **testing strategy** for such a large system?
4. What is the **inter-chip network architecture** and how do we achieve resiliency if a few chiplets fail?
5. How to **design the waferscale Si-IF substrate**?

Power Delivery

- Edge Power Delivery at 2.5V
- LDO based power management at each node
- On-chip decoupling capacitance (20nF per tile)
- DeCap consumes 30% of the chip area
  - *Deep Trench Capacitors in Si-IF would help*

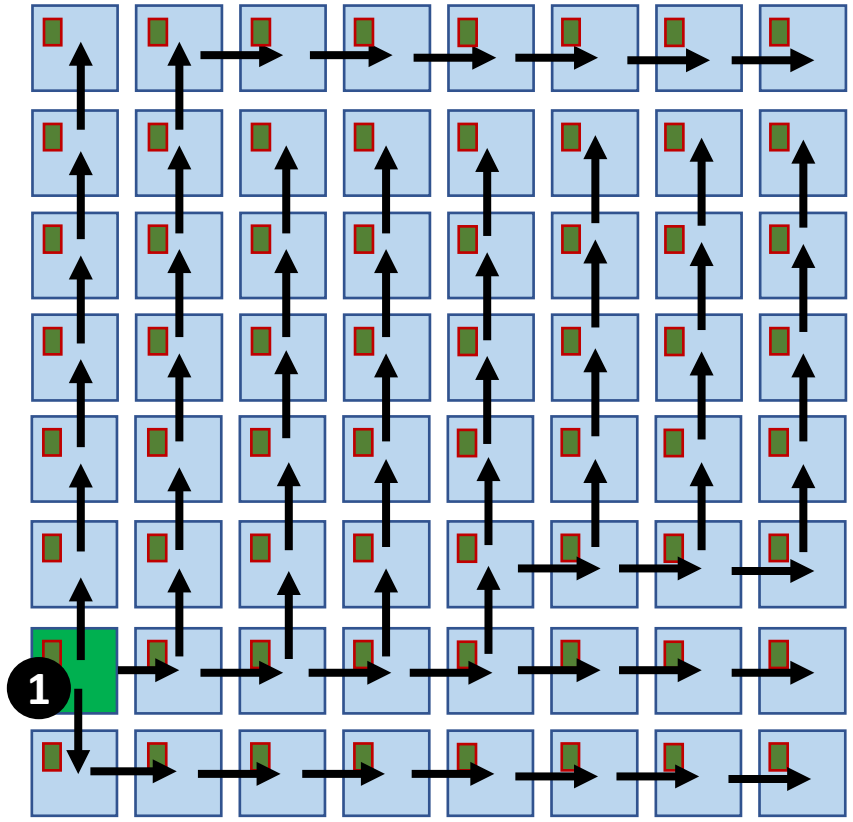




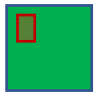


Waferscale Clocking – Clock Generation

- PLL in each die for clock generation
- However, stable reference voltage needed by PLL not present away from center
- Only PLLs in edge dies can be used
- Generate fast clock at the edge and distribute



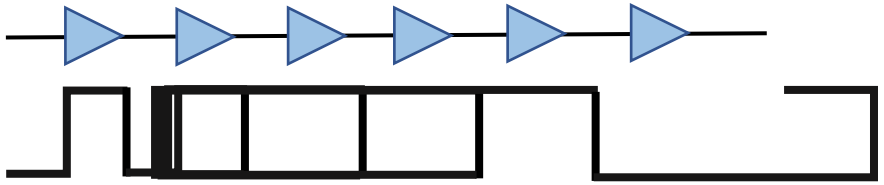
- PLL

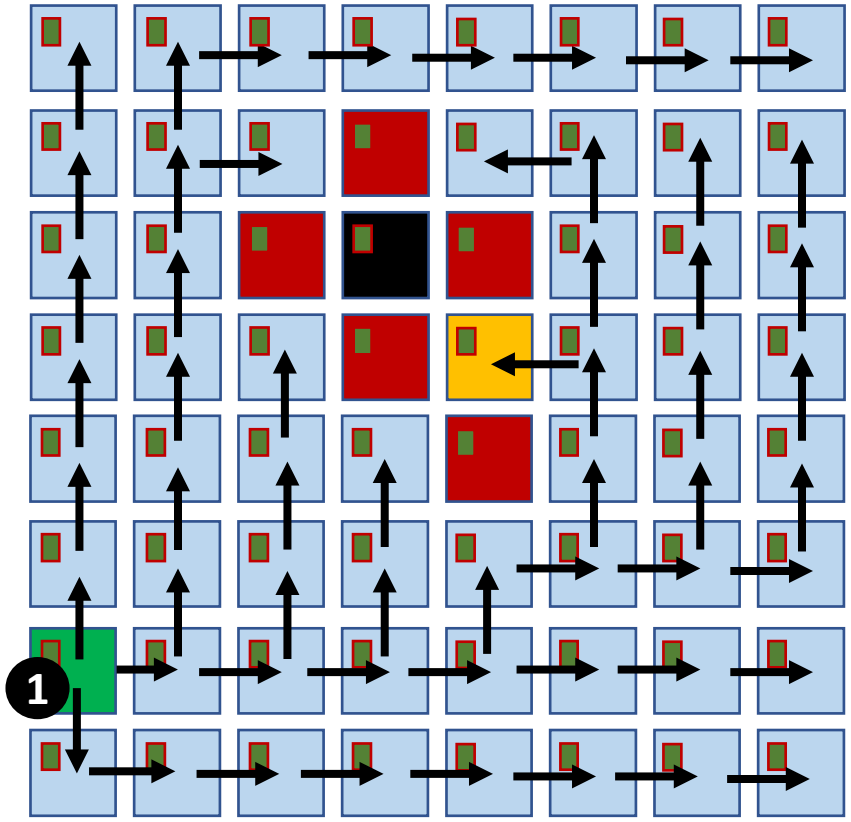


- Clock generating edge tile

Waferscale Clocking – Clock Distribution

- Fast clock will be forwarded
- Clock inverted at each hop to avoid duty cycle distortion accumulation
- Communication between dies using asynchronous interfaces
- Fault tolerance in clock distribution network

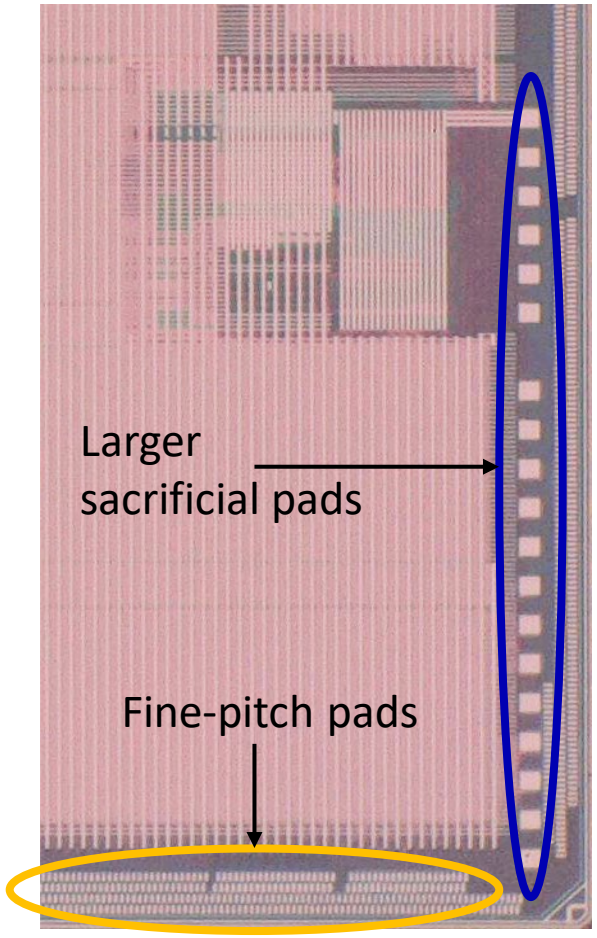




- PLL - Clock generating edge tile

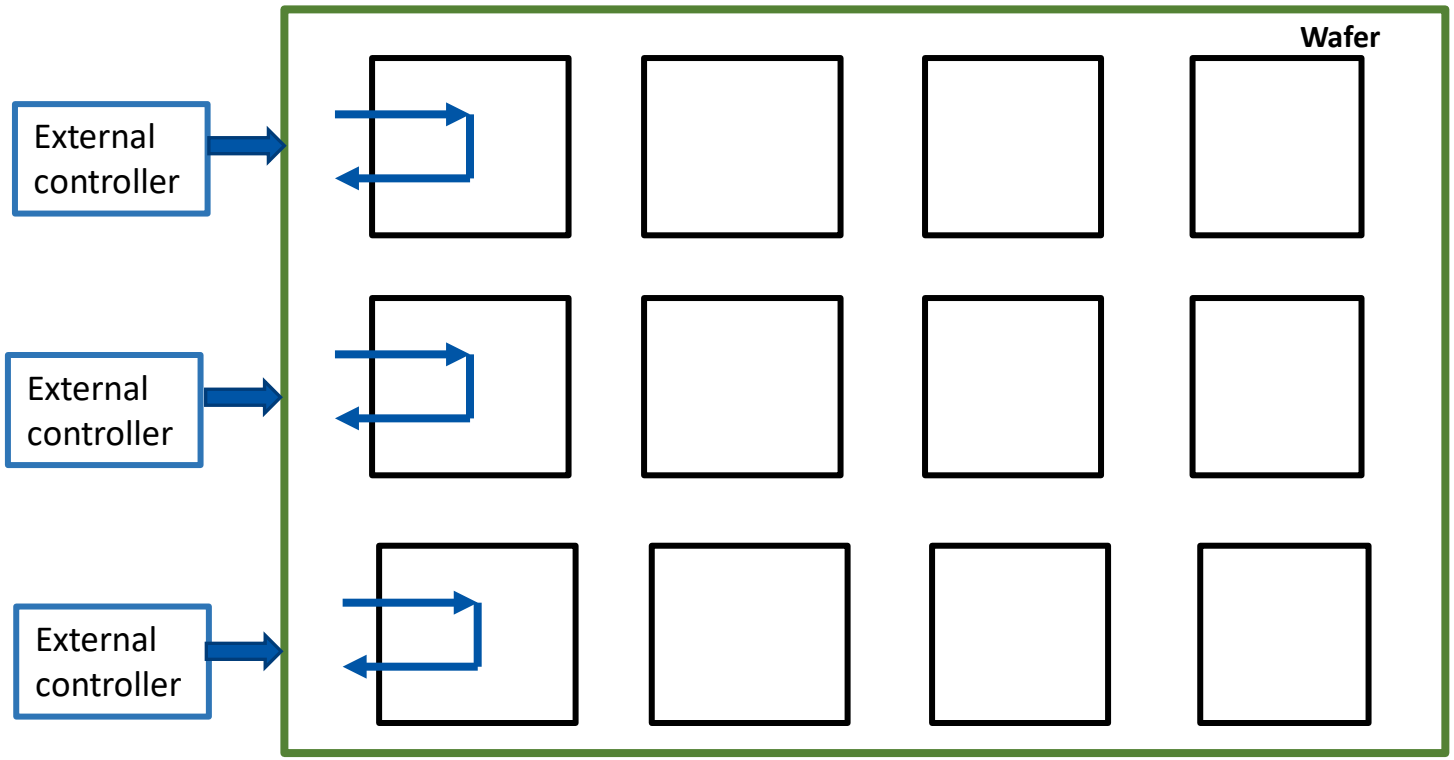
Pre-bond Die Testing

- Fine pitch pads cannot be probed
- Larger pads for probe test
- These pads are sacrificed and not used for bonding
- Only smaller pads attach to the Si-IF using fine pitch pillars

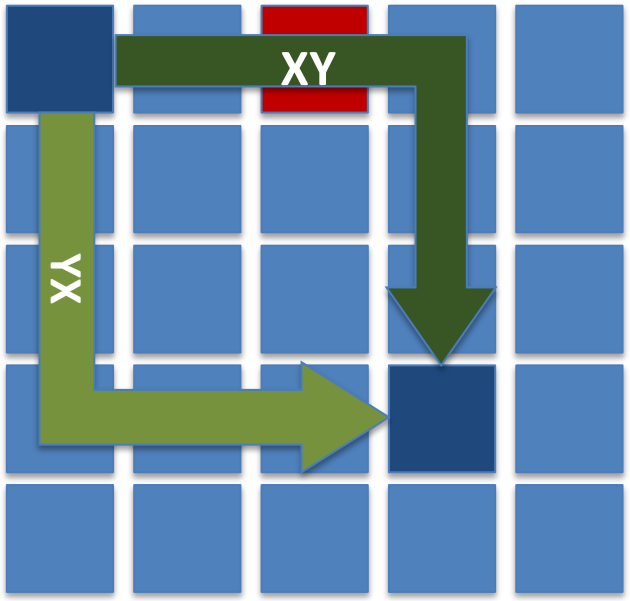


Post-bonding JTAG Test Scheme

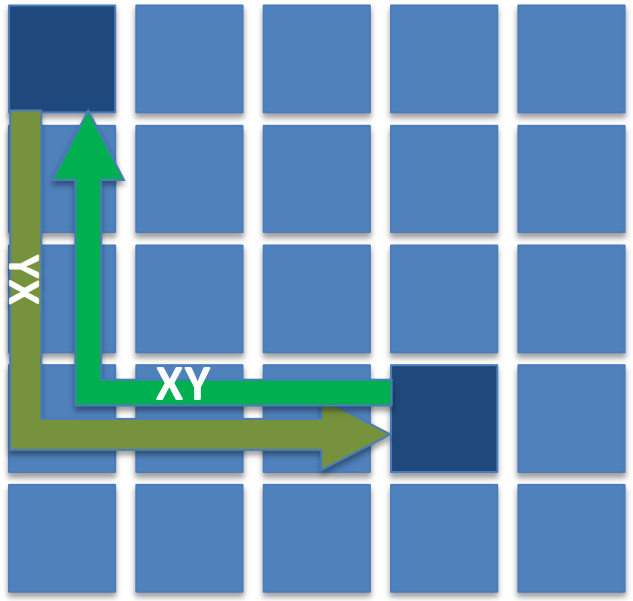
- (1) Multiple chains
  - One JTAG chain results in single point of failure vulnerability
  - Throughput is an issue:
    - 2.5 hours to load the memories using one chain
    - 5 minutes to load with 32 chains
- (2) Progressive unrolling
  - Helps identify post-bonding faulty dies
  - Similar to IEEE 1838 proposal



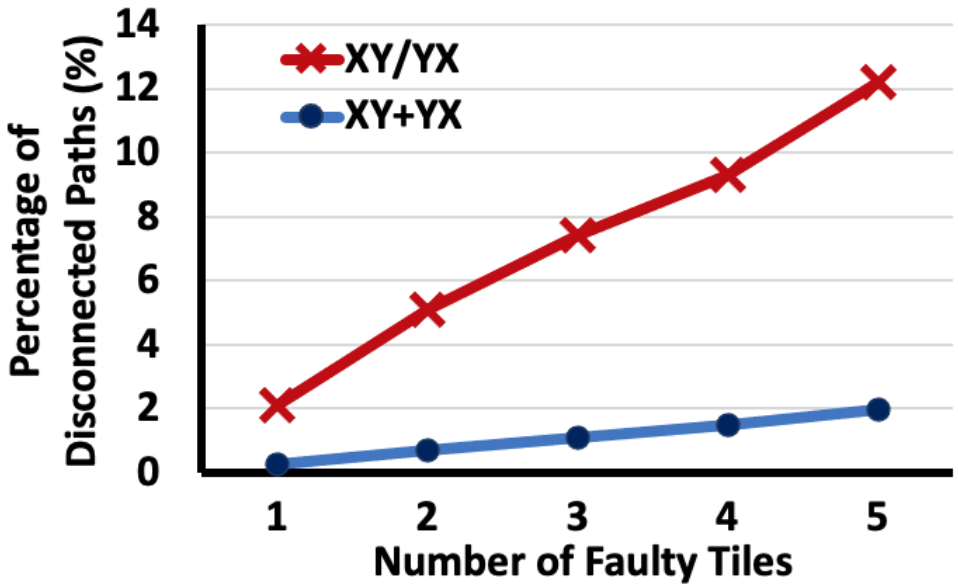
Network Resiliency



Two Separate Networks

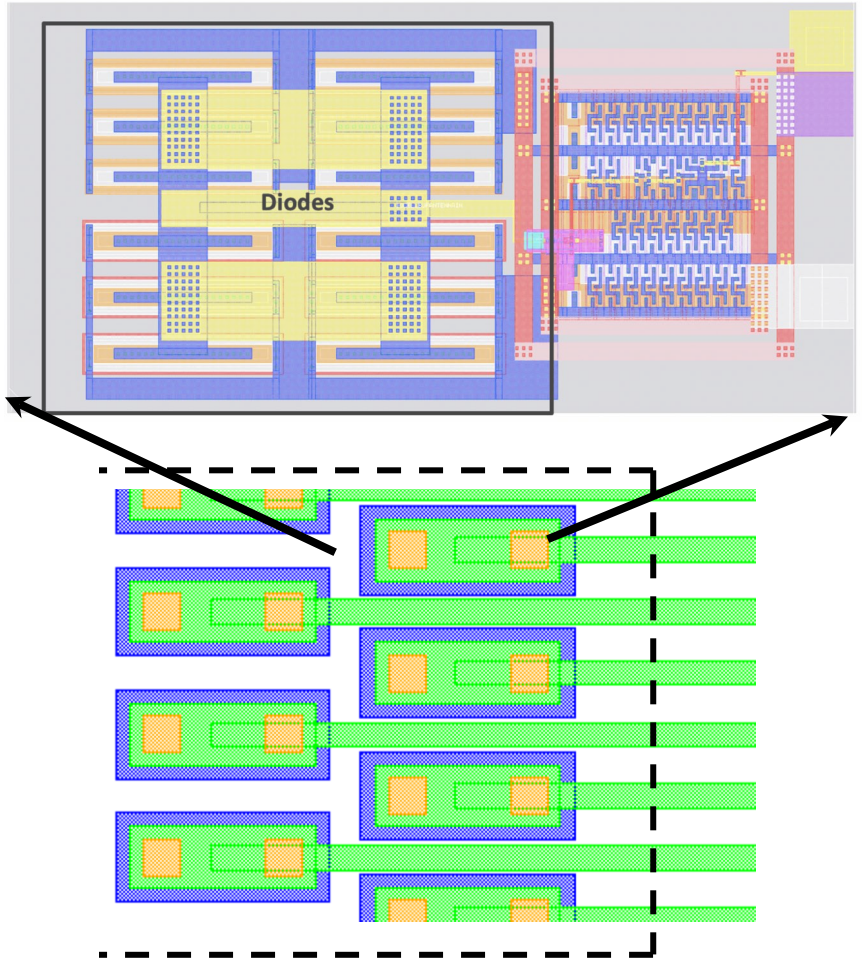


Request-Response in Complimentary Networks



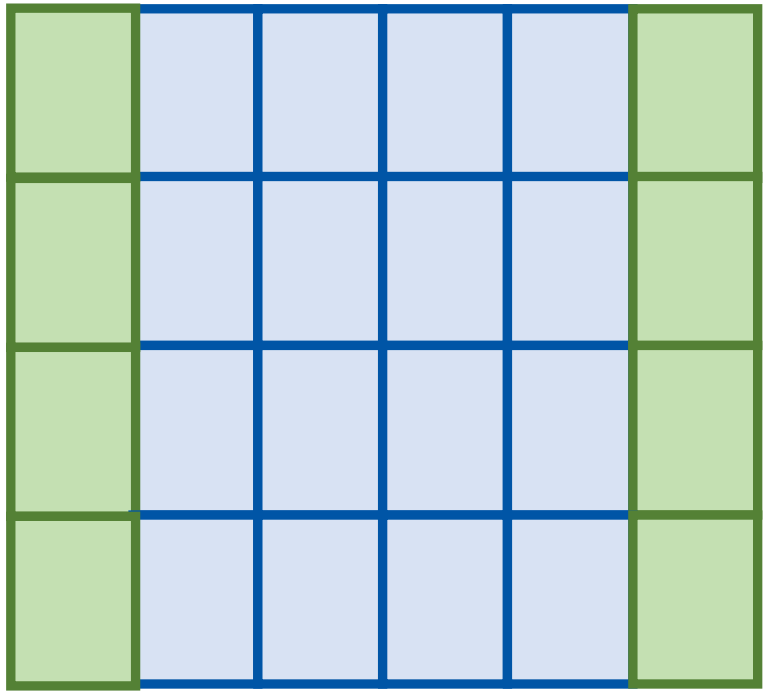
I/O Architecture

- I/O pitch of 10 um and depth of 20 um
- Simple cascaded buffer architecture
- 0.07 - 0.18 pJ/bit
- Two pillars per IO for redundancy
- *ESD diodes* and buffers need to fit within the I/O footprint

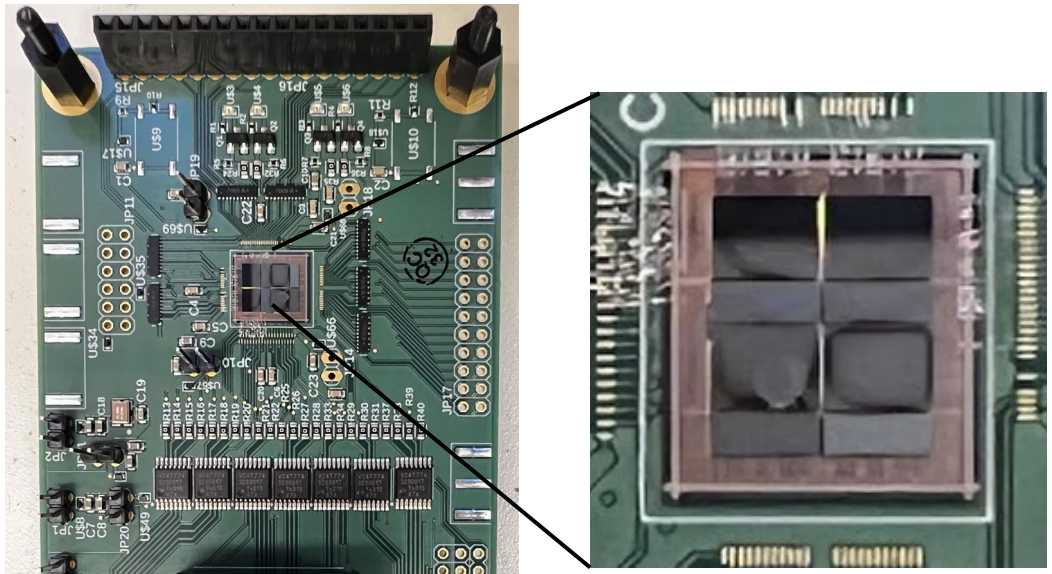


Waferscale Substrate Design – Custom Router

- Silicon Interconnect Fabric (Si-IF): 4 metal layers, >15,000mm^2
- OpenAccess C++ based efficient waferscale custom router
  - Signal routing layers are sparse, used space-based routing methodology
- Si-IF wafer much larger than maximum reticle size – designed to make it step-and-repeatable



Smaller Prototype Bring-up was Successful



4-tile Prototype

Test board with
Host Controller

Current Status:

- Small-scale prototype
- Full-tile functionality fully verified
- Runs at 300 MHz
- First demonstration of tightly coupled dis-aggregated chiplet-based system
- Custom high-density I/O PHY and protocol verified
- Full applications using multi-core communication over shared memory was verified

Future Plans:

- Wafer scale substrate manufacturing is being done in collaboration with external foundry partners