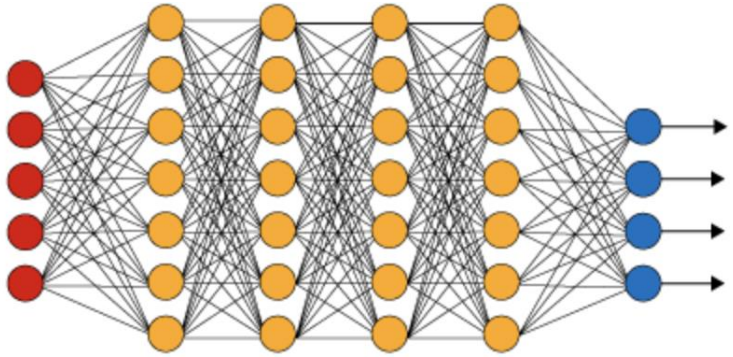
Waferscale Architectures

Saptadeep Pal, Subramanian S. Iyer and Puneet Gupta
University of California, Los Angeles

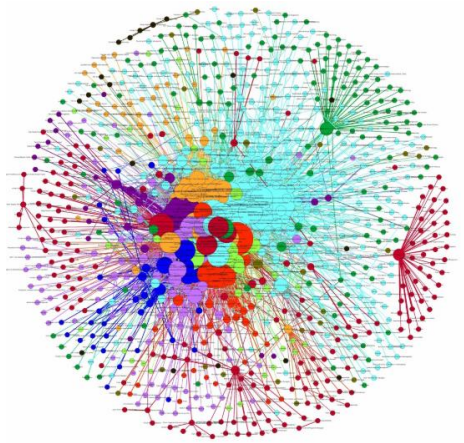
In collaboration with Rakesh Kumar
University of Illinois at Urbana-Champaign

Table of Contents

Massively Parallel Applications

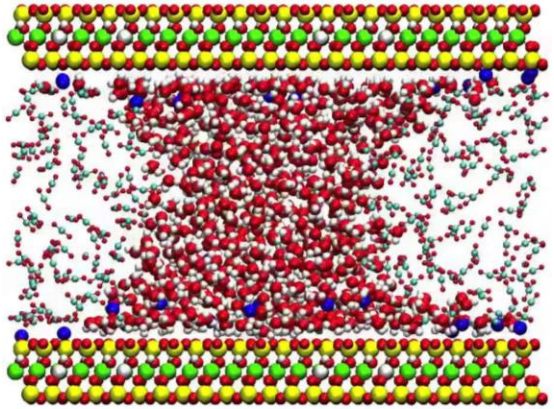


Artificial Intelligence

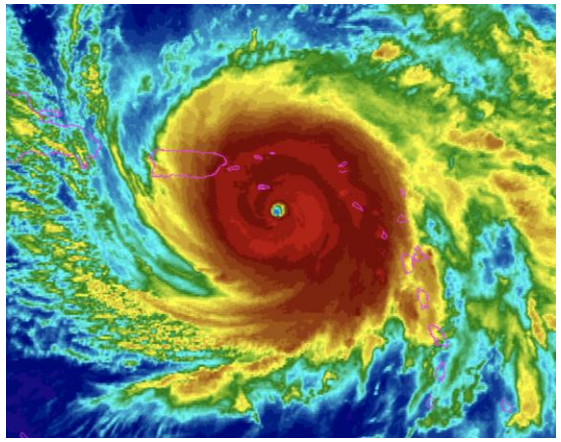


Social Networks

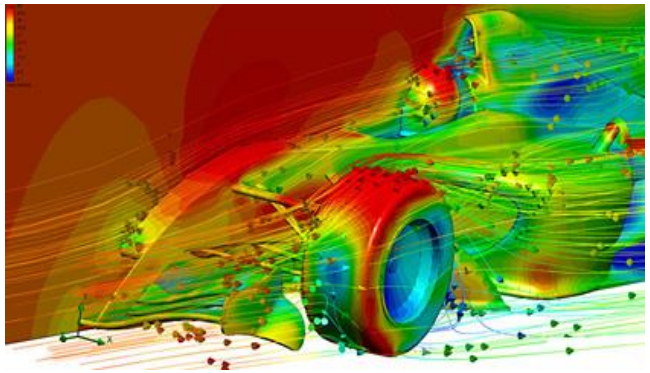
Big Data



Molecular Dynamics Simulation

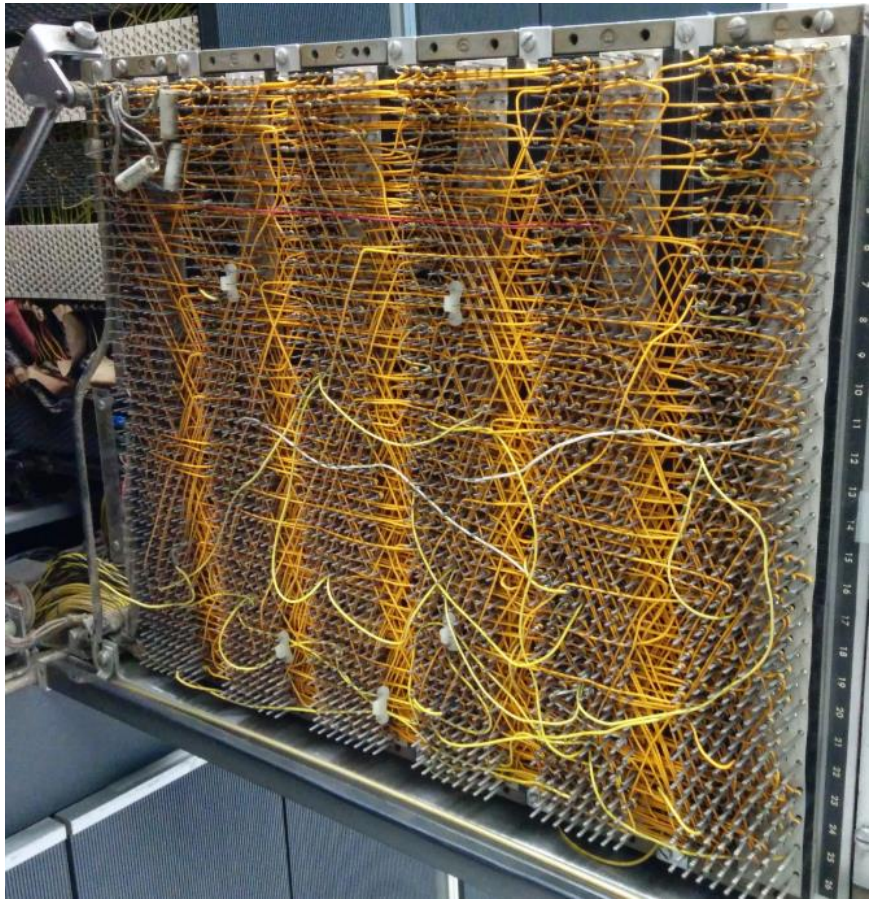


Weather Modelling

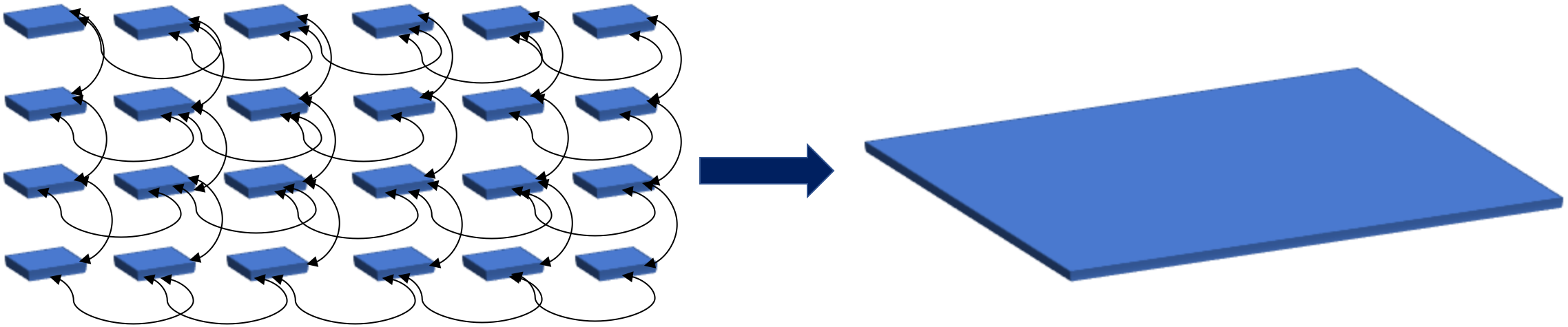


Multi-Physics Simulation

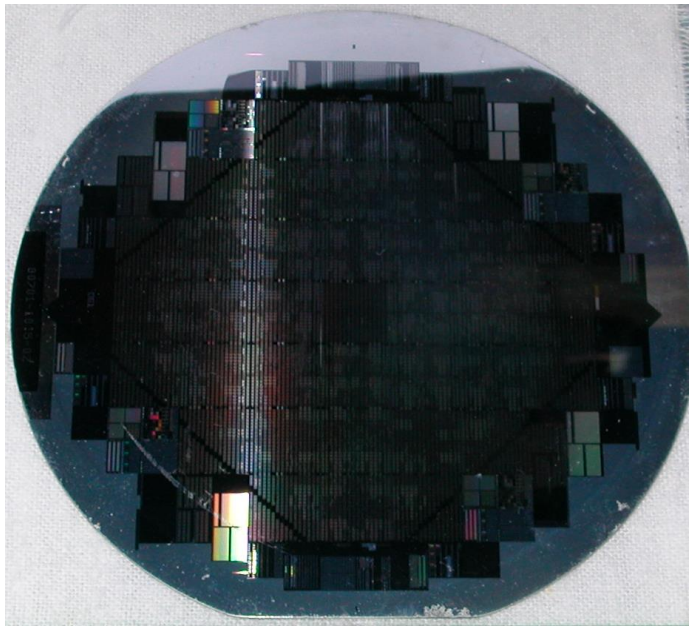
Communication is Expensive!!



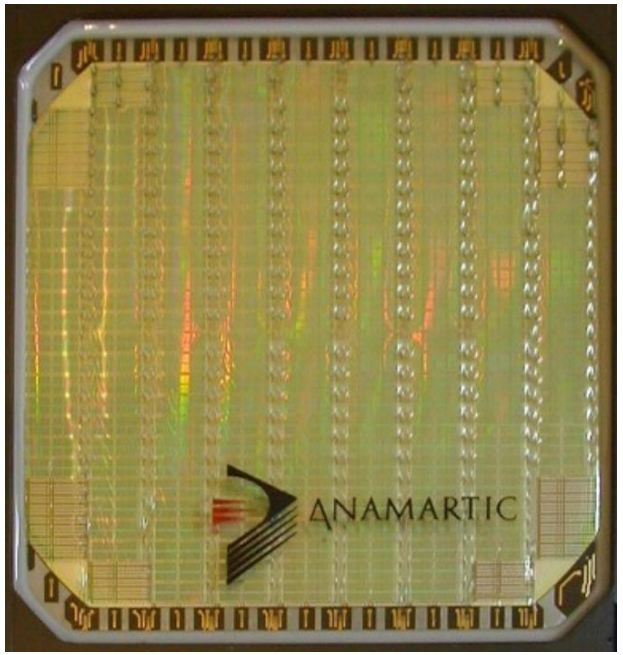
One Large Chip



A Brief History of Waferscale Computing



Gene Amdahl's Trilogy Systems



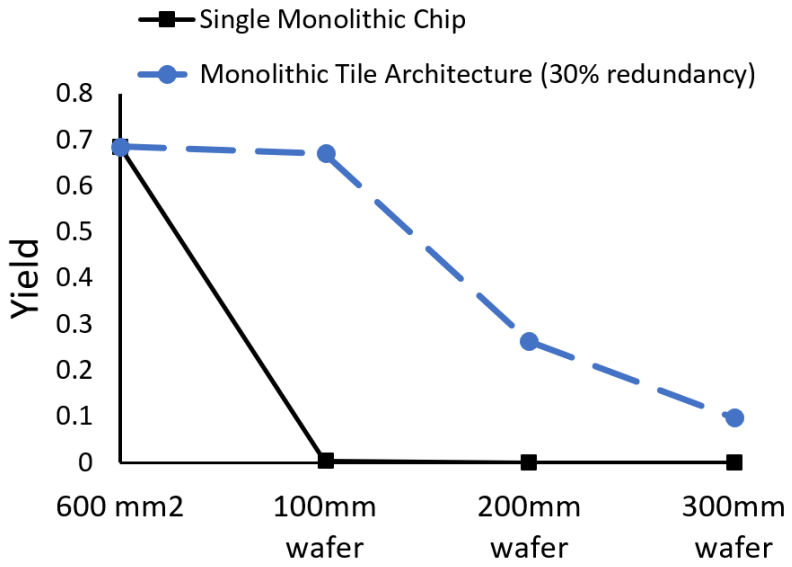
Tandem Computers, Fujitsu

Other efforts: ITT Corporation, Texas Instruments. Recent efforts: Spinnaker

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)

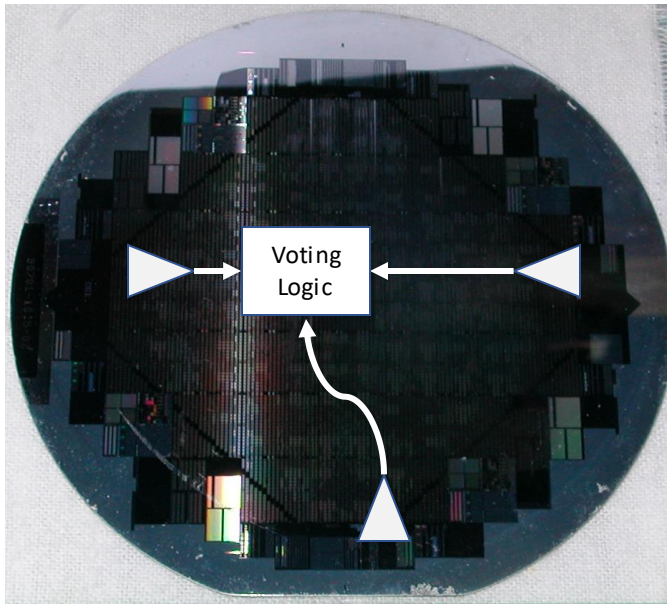
Their Approach to Waferscale: **Monolithic**



Area of chip ↑

Probability of defects ↑

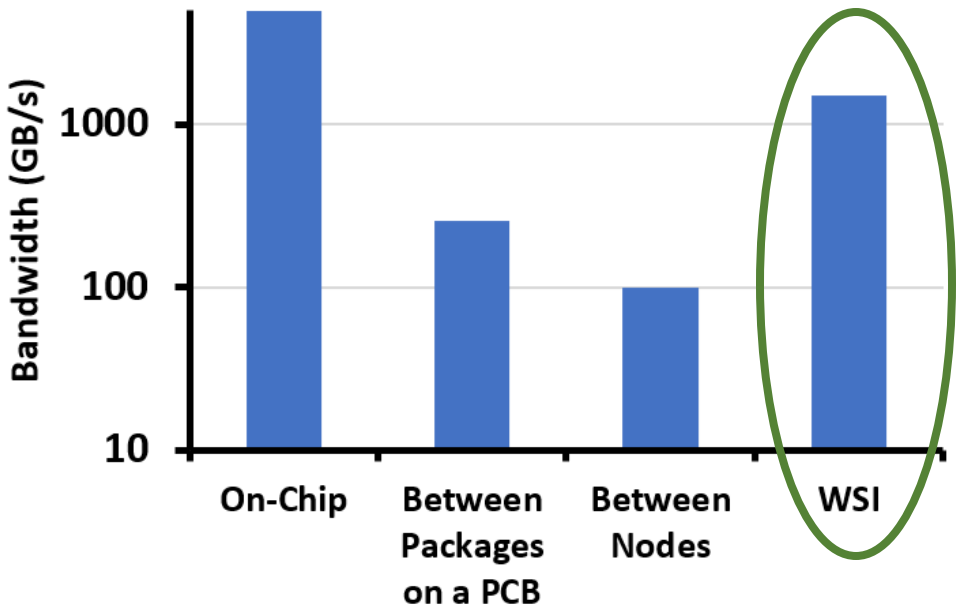
Variability ↑

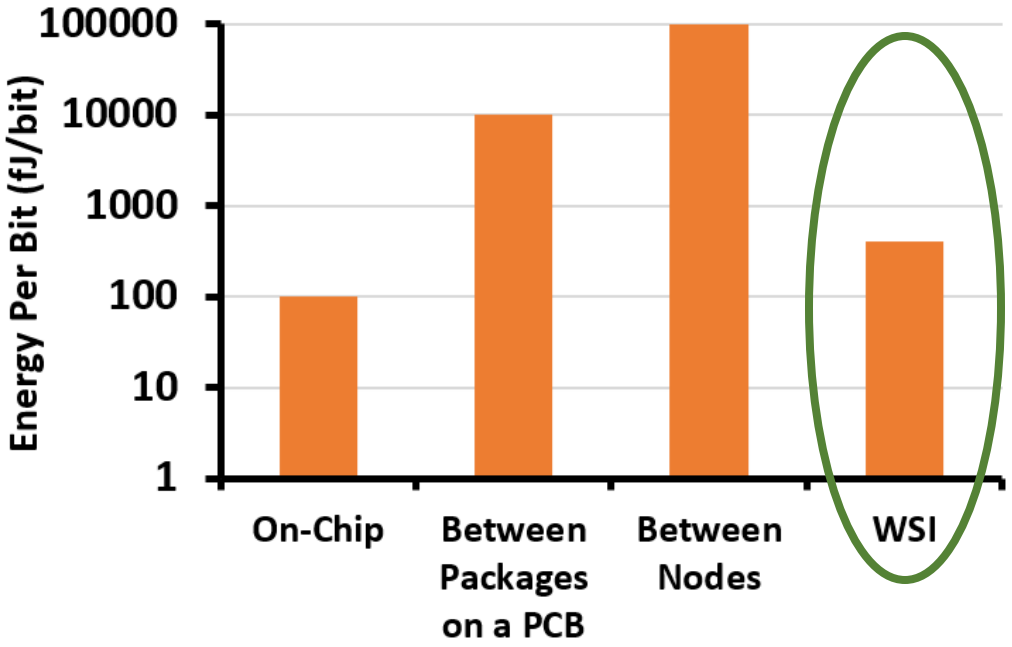


Some mitigation possible through TMR, etc. - but prohibitively expensive

Deemed commercially unviable

Time to Give Waferscale Another Go?





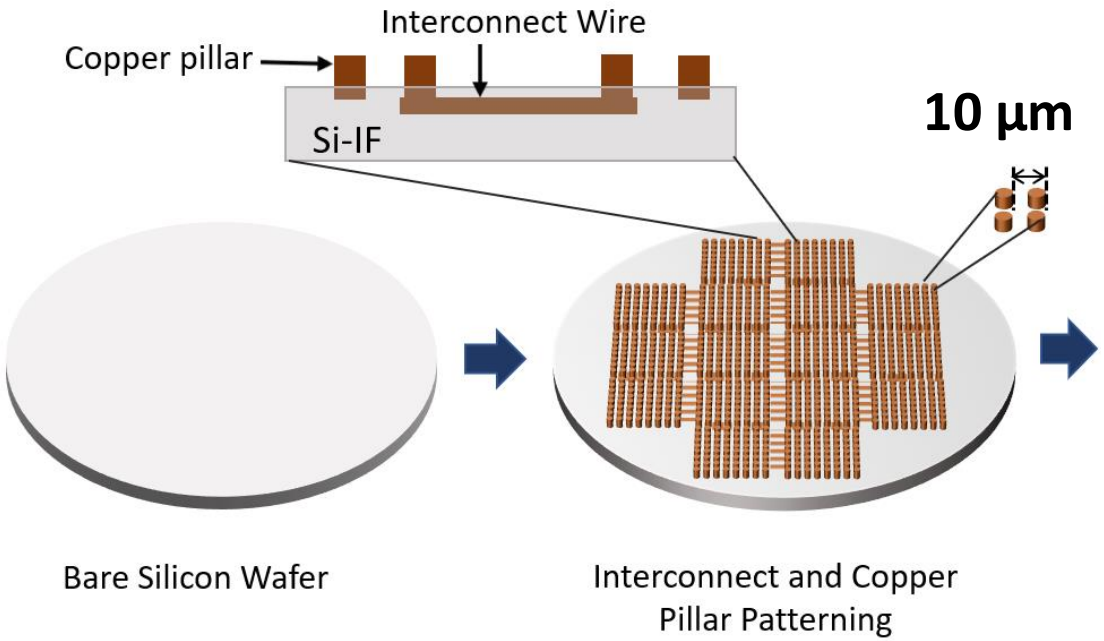
However, to achieve waferscale integration, we need to solve the **<u>yield problem</u>**

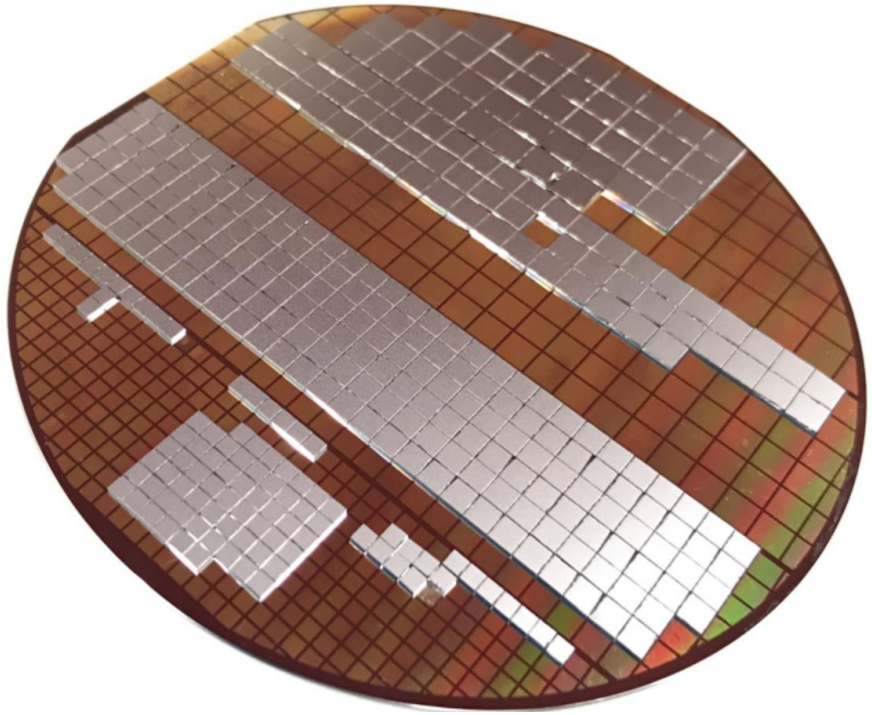
Re-imagining WSI Technology

Q: What do we need from waferscale integration?

A: High density interconnection

UCLA Silicon Interconnect Fabric (Si-IF)*





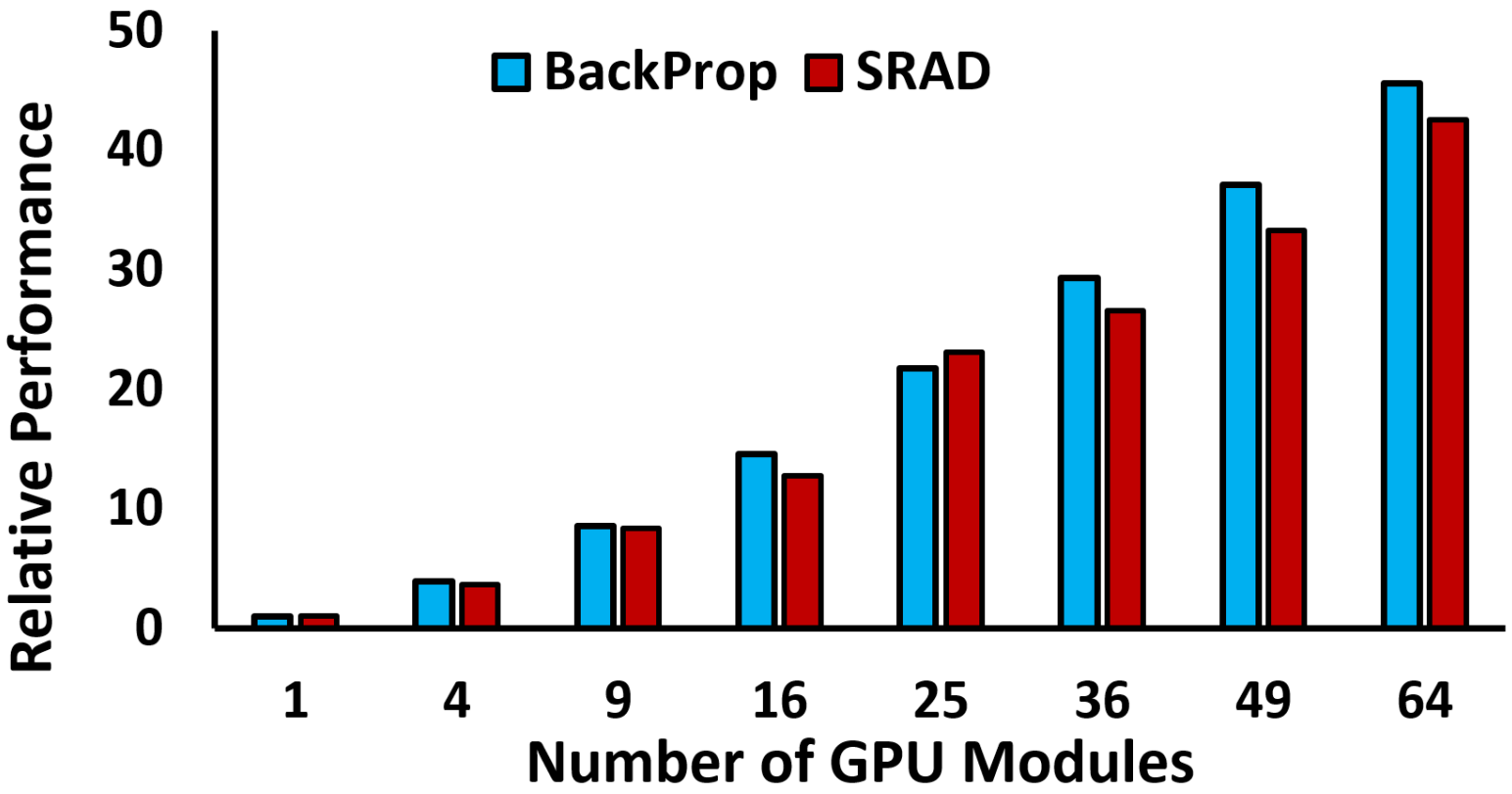
Measured Bond Yield >99%

Allows heterogeneous waferscale integration with high yield

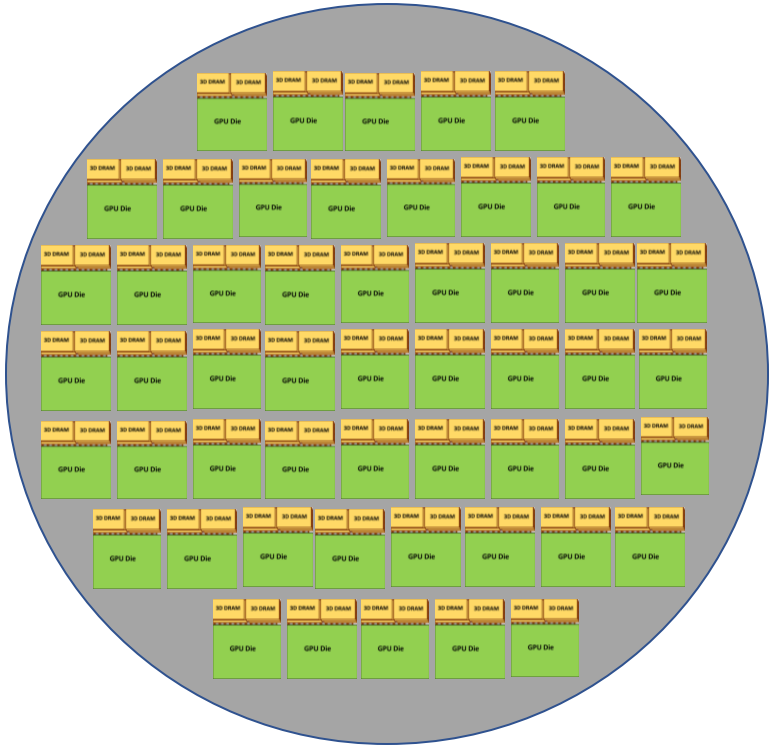
*UCLA CHIPS Programme: https://www.chips.ucla.edu/research/project/4

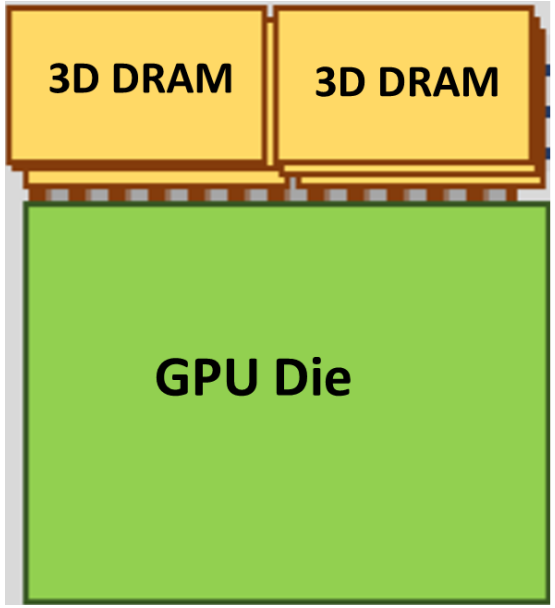
A Case for Waferscale GPU

GPU applications scale well with compute and memory resources



Waferscale GPU Overview





A GPU Module: GPM

- GPU die = 500 mm^2
- 3D DRAM die = 100 mm^2
- **Total Area = 700 mm^2**

- GPU Die power = 200 W
- DRAM Die power = 35 W
- **Total Power = 270 W**

300 mm wafer has enough area for about **72 GPU modules (GPM)**.

Architecting a Waferscale GPU

Q: Can we build a 72-GPM waferscale GPU ?

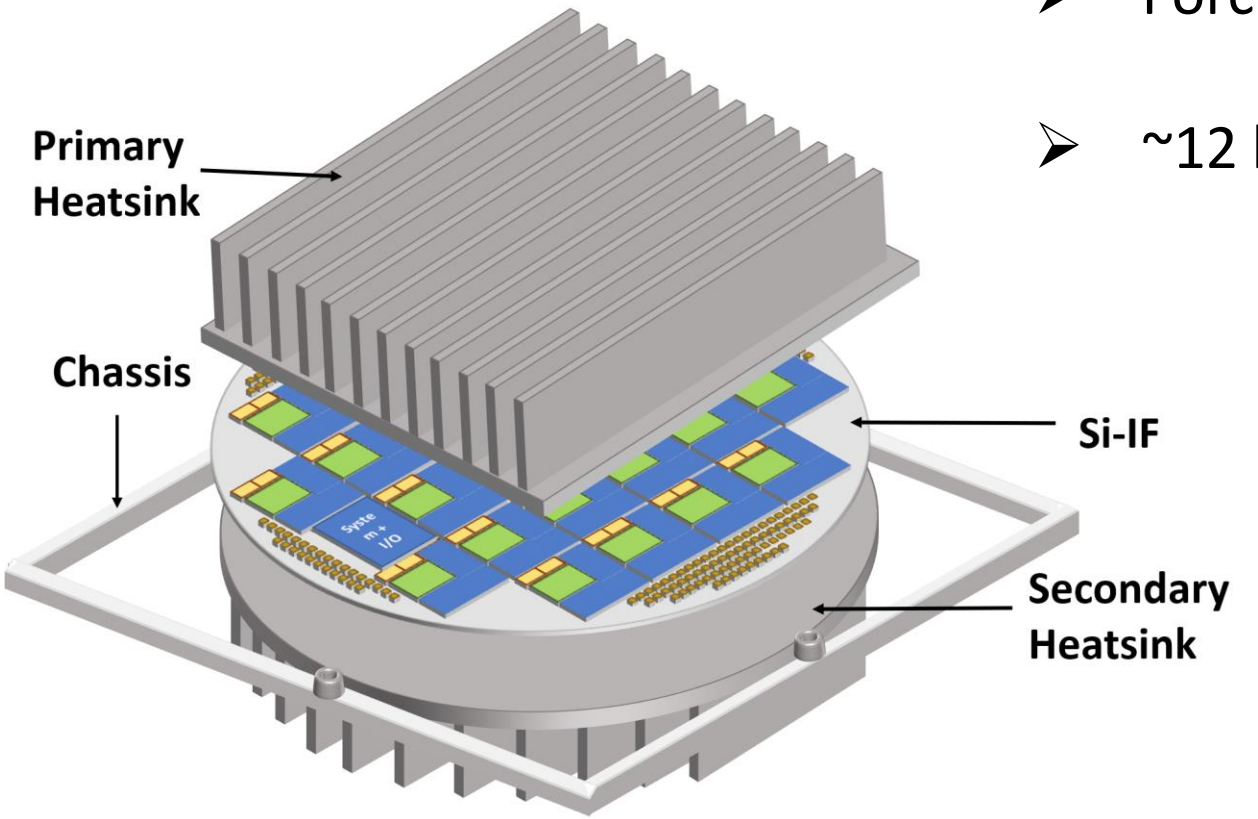
Three major physical constraints:

1. **<u>Thermal</u>**
   - Waferscale GPU would dissipate kWs of power

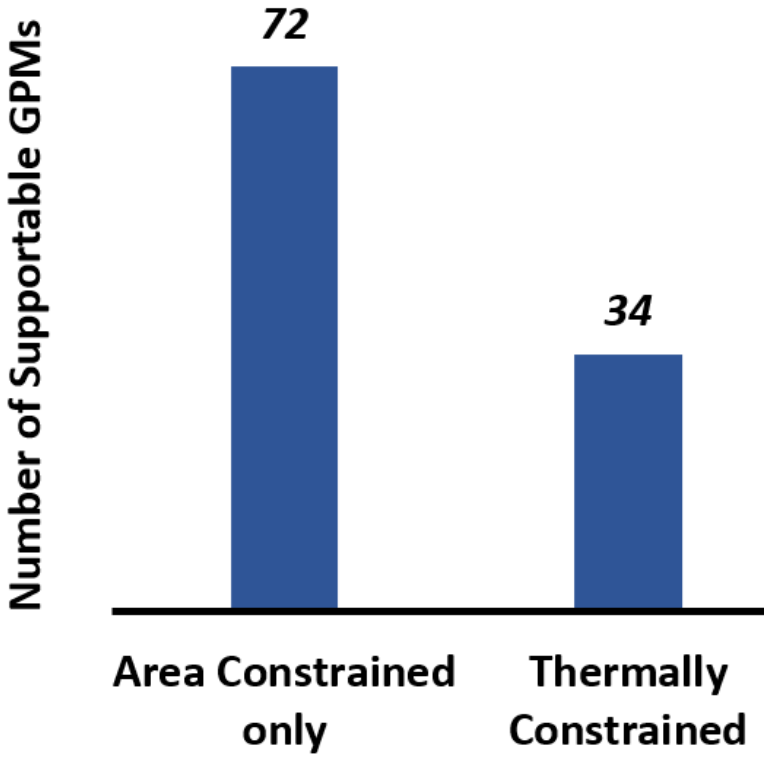
2. **<u>Power Delivery</u>**
   - How to supply kWs of power to the GPU modules?
   - Voltage Regulator Module (VRM) overhead?

3. **<u>Network of GPMs</u>**
   - Si-IF has up to 4 metal layers, what network topology to build?

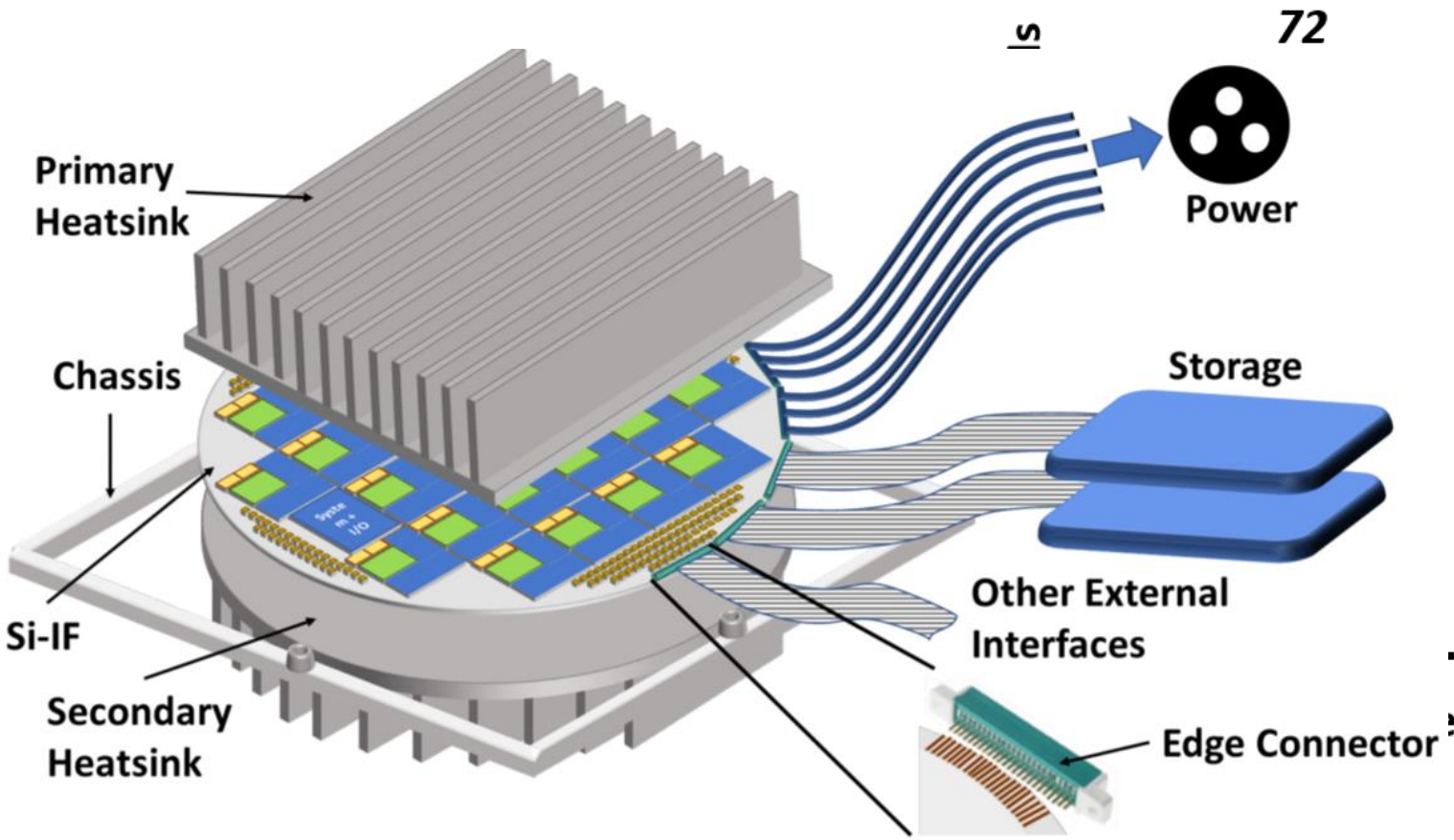
Thermal Design Power

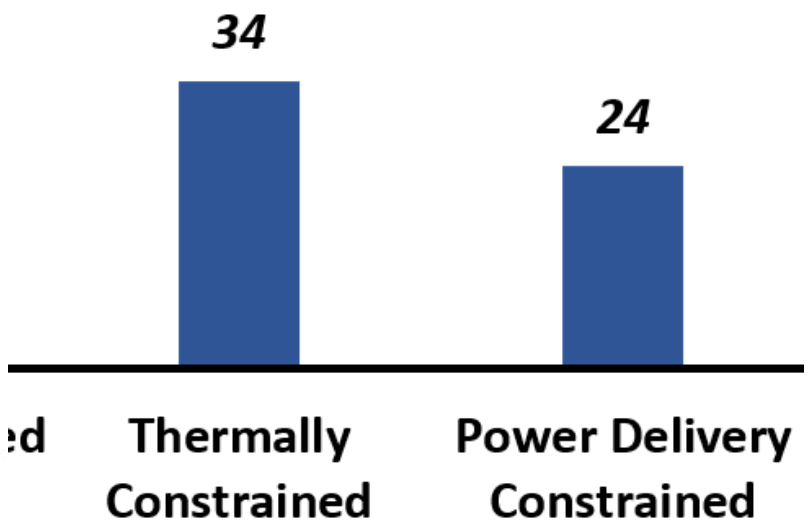


- Forced air-cooling with two heat sinks
- ~12 kW TDP i.e., **34 GPMs** can be supported

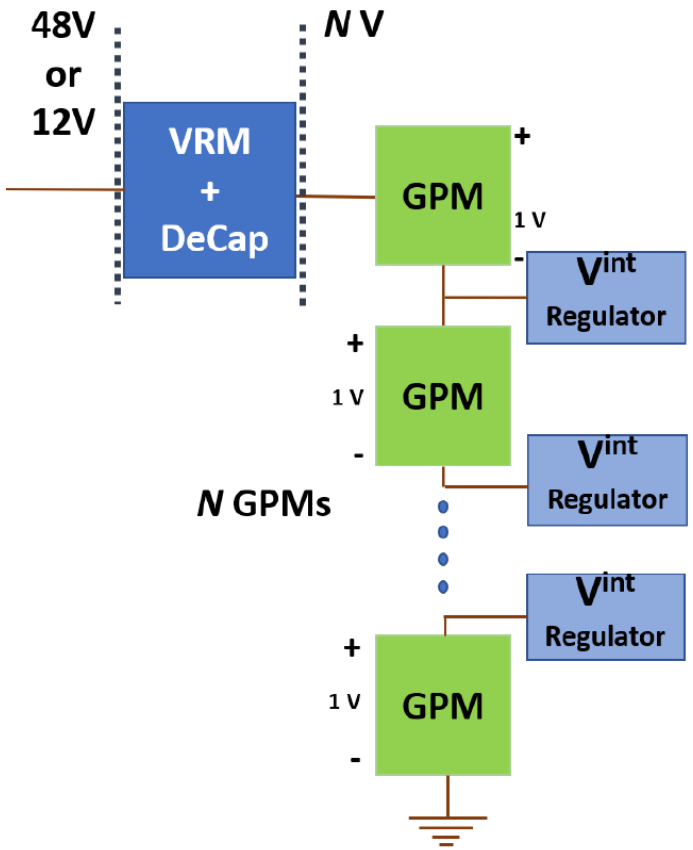


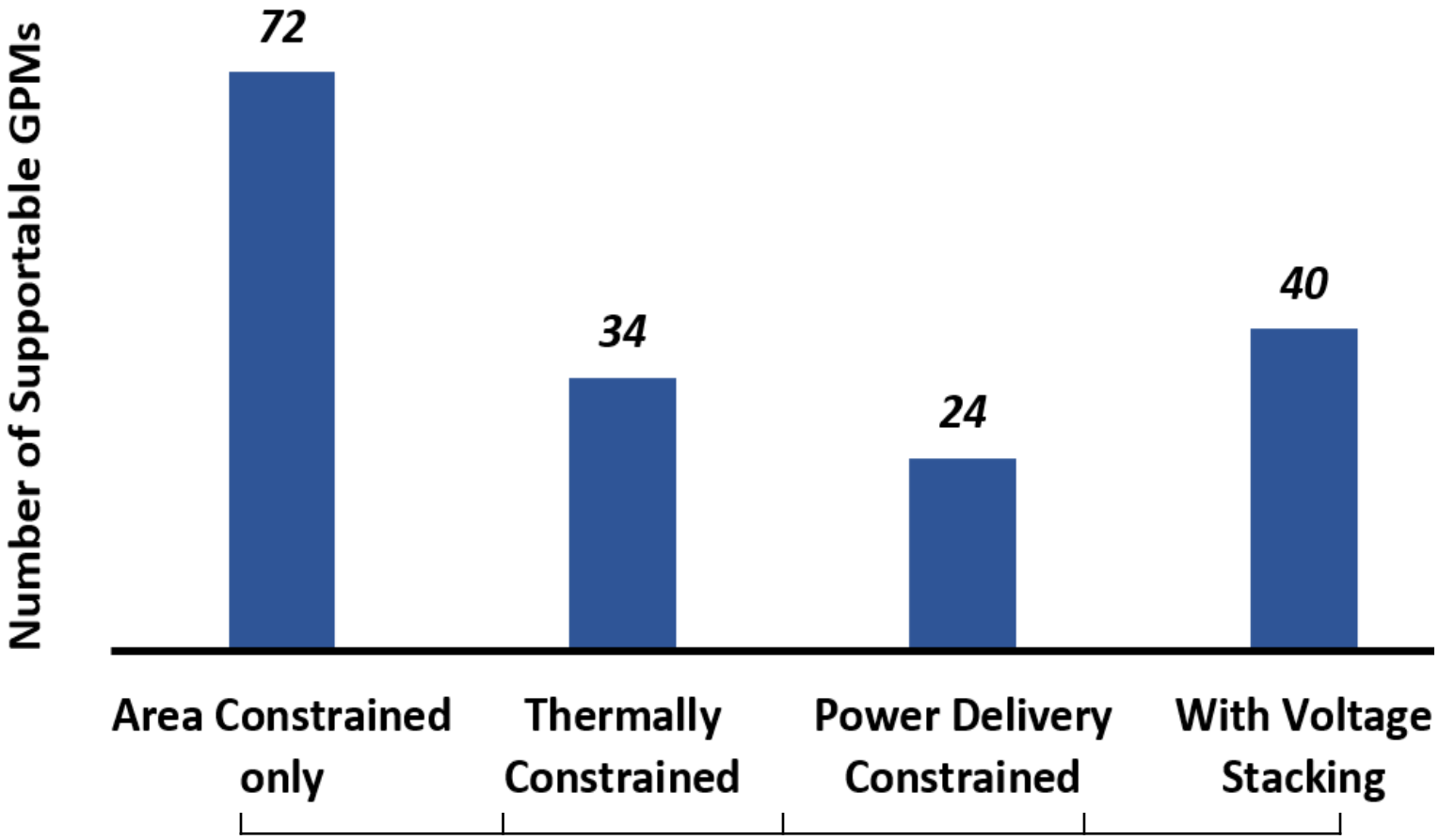
Power Delivery



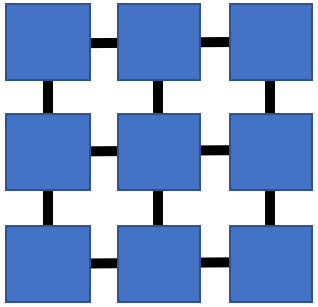


Stacked Power Delivery

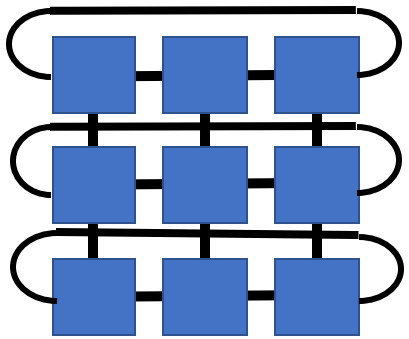




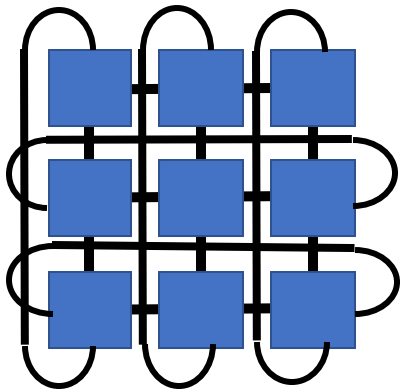
Waferscale Inter-GPM Network



Mesh



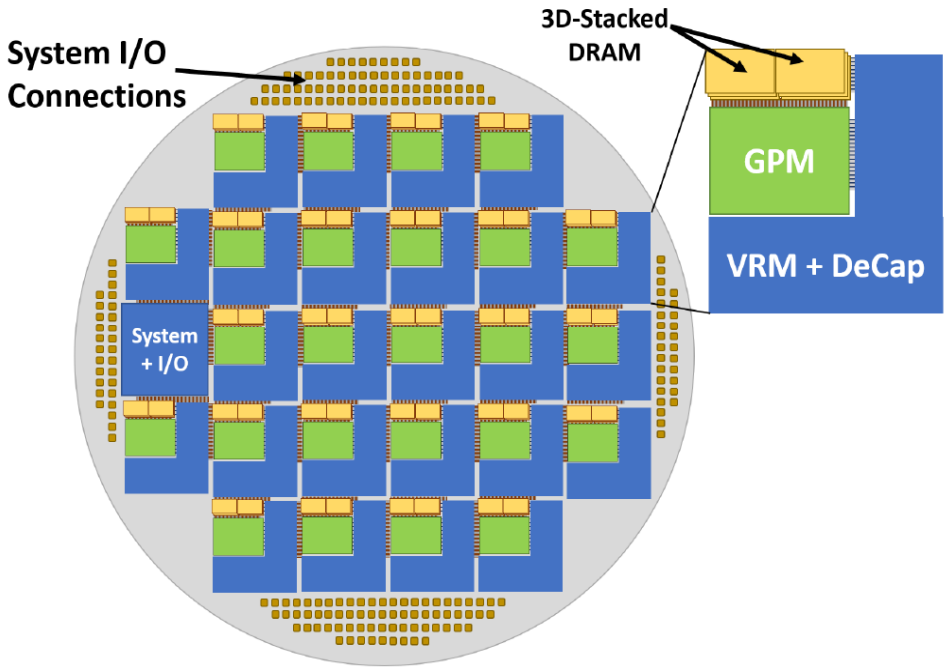
1D-Torus

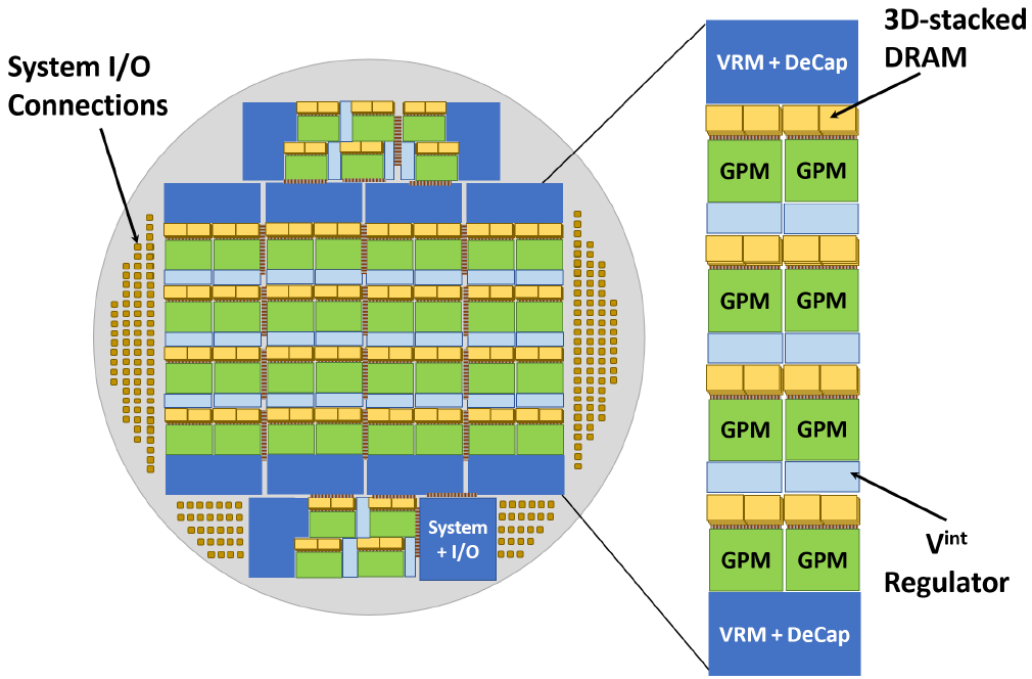


2D-Torus

Num. Layers	Topology	Inter-GPM BW (TBps)	Si-IF Yield
1	Mesh	0.75	95.9%
2	**Mesh**	**1.5**	**91.9%**
	1-D Torus	1.5	84.3%
3	2D Torus	1.9	74%

Final WS-GPU Architectures



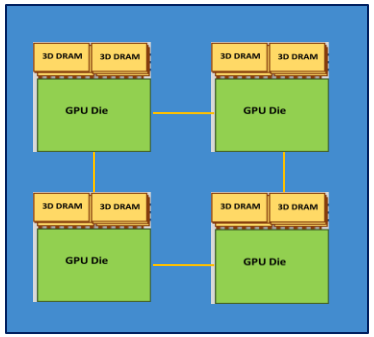


24 GPM Floorplan without voltage stacking

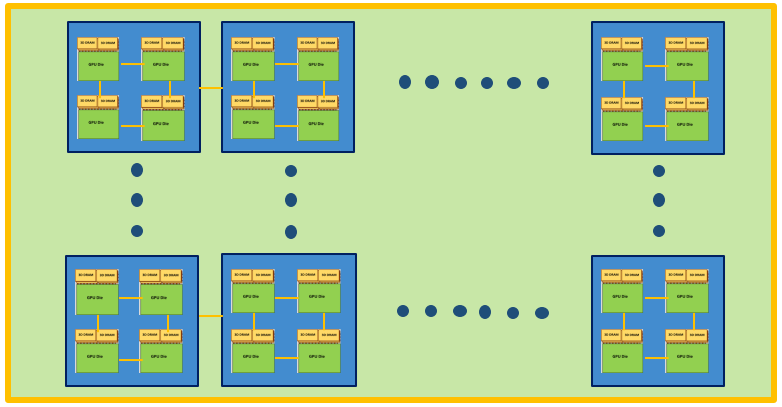
40 GPM Floorplan with voltage stacking

Inter-GPM Network: **Mesh**

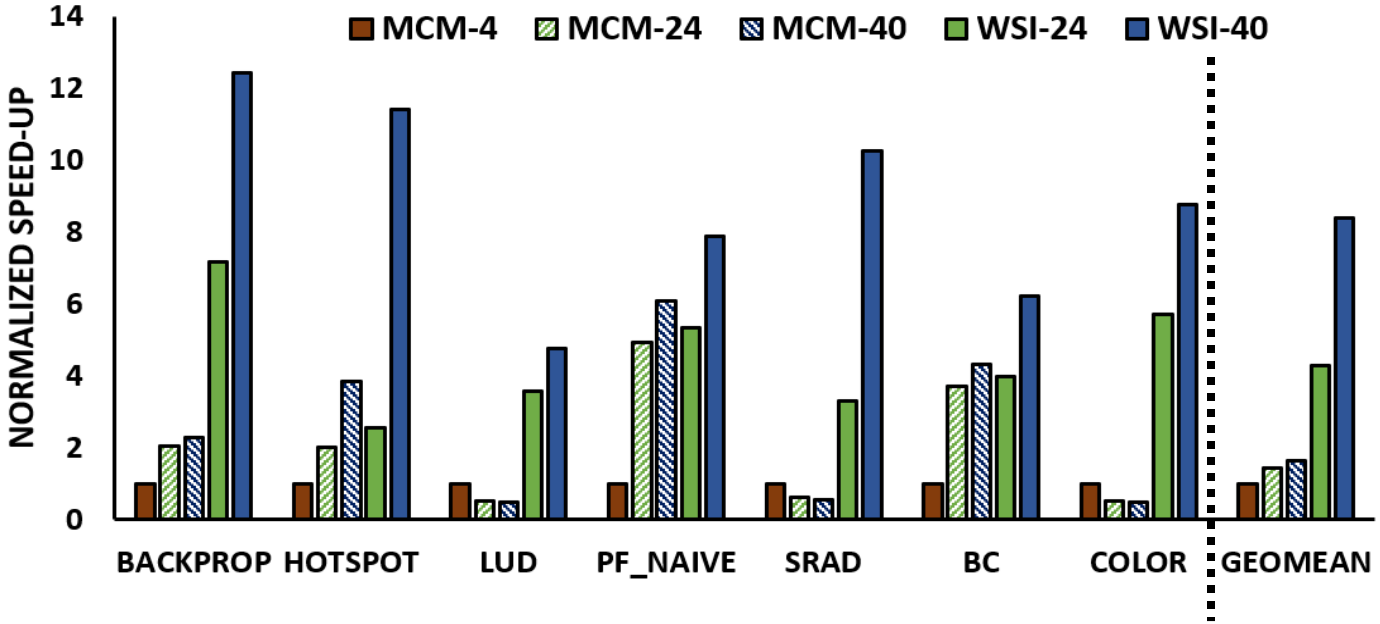
Results – WS-GPU Performance Improvement



MCM Package



PCB

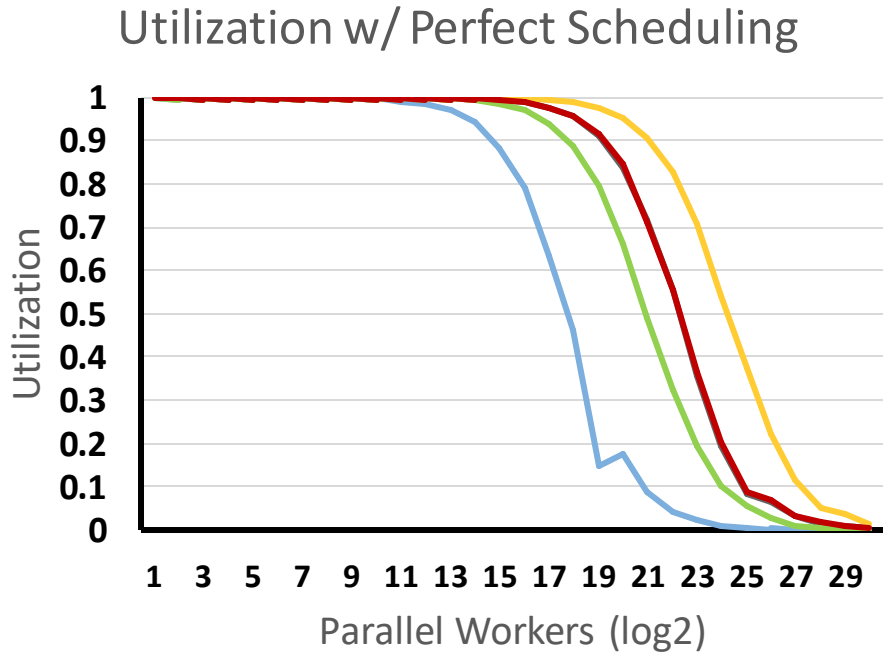


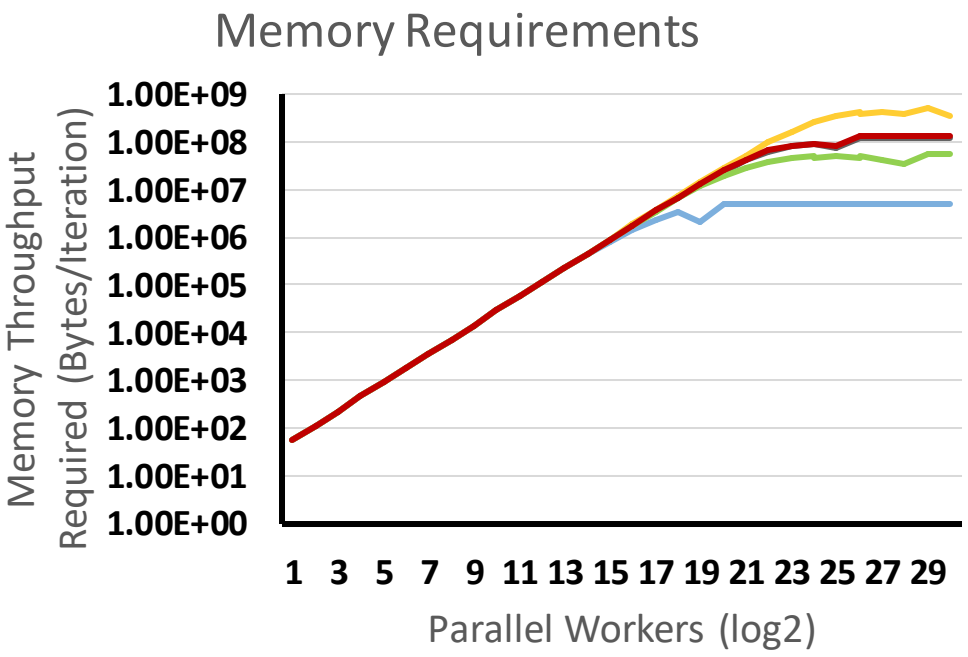
- WSI with 24 GPMs performs **2.97x** better than multi-MCM configuration (**EDP: 9.3x**)
- WSI with 40 GPMs performs **5.2x** better than multi-MCM configuration (**EDP: 22.5x**)

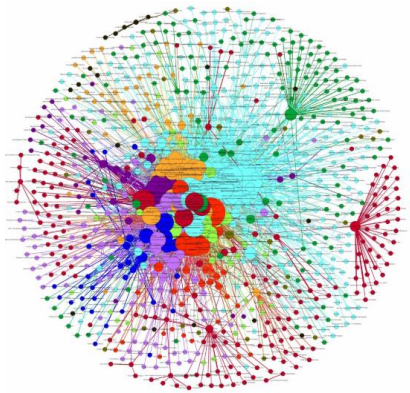
Waferscale integration can provide large performance and energy benefits

Waferscale Graph Accelerator Prototype

Graph Acceleration using Waferscale Architectures





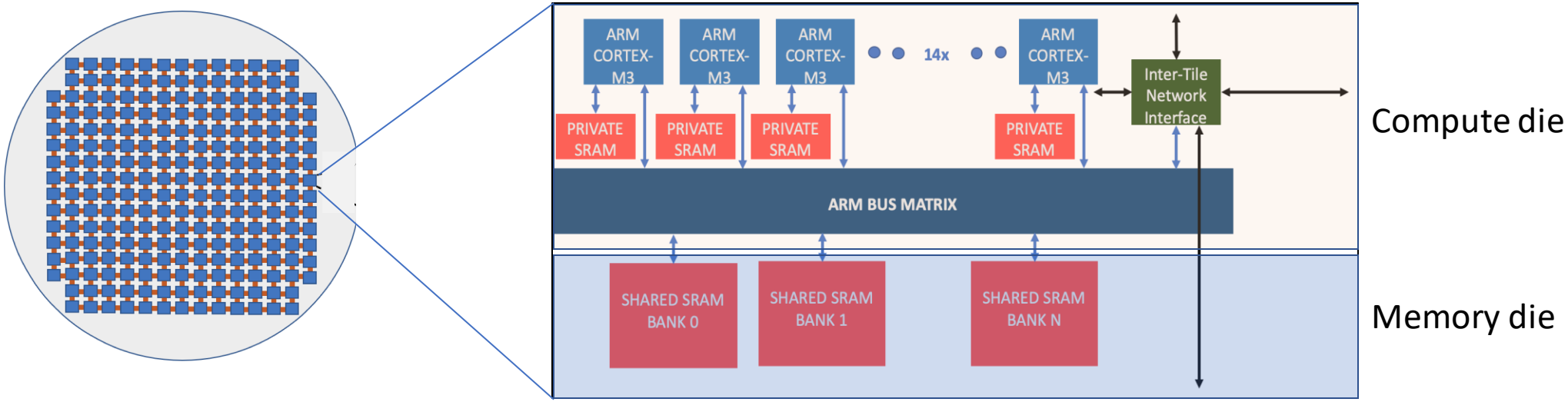


Massive number of processing cores needed **Bandwidth beyond 50 TBps @ 10 ns/iteration**

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

Graph workloads exhibit irregular memory access pattern

ARM Cortex-M3 based Prototype



(1) Scalable up to **64x64** tile array (4096 tiles). Total 57,344 cores

(2) Unified memory architecture and >100 TB/s total memory bandwidth

(3) Custom implementation of waferscale read, write and synchronization primitives.

(4) Fault-tolerant **dual** network scheme.

(5) GALS clocking using the waferscale network.

(6) Simple software libraries implemented for easy and intuitive programming
-- Programmers can treat the entire wafer as a single multi-processor

Planned Tape-out: Late-2019
Bring-up: Q2 2020.

Summary and Conclusion

- Communication between packaged processors is a major bottleneck
- Si-IF technology enables waferscale integration at high yield
- Waferscale GPU versus multi-MCM system:
  - **5.2x** performance improvement
  - **22.5x** EDP improvement
- Graph applications are good fit for waferscale processing
  - ARM Cortex-M3 based waferscale prototype under active development

Acknowledgement

- This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) through ONR grant N00014-16-1-263, UCOP through grant MRP-17-454999, and the UCLA CHIPS Consortium.

- Collaborators:

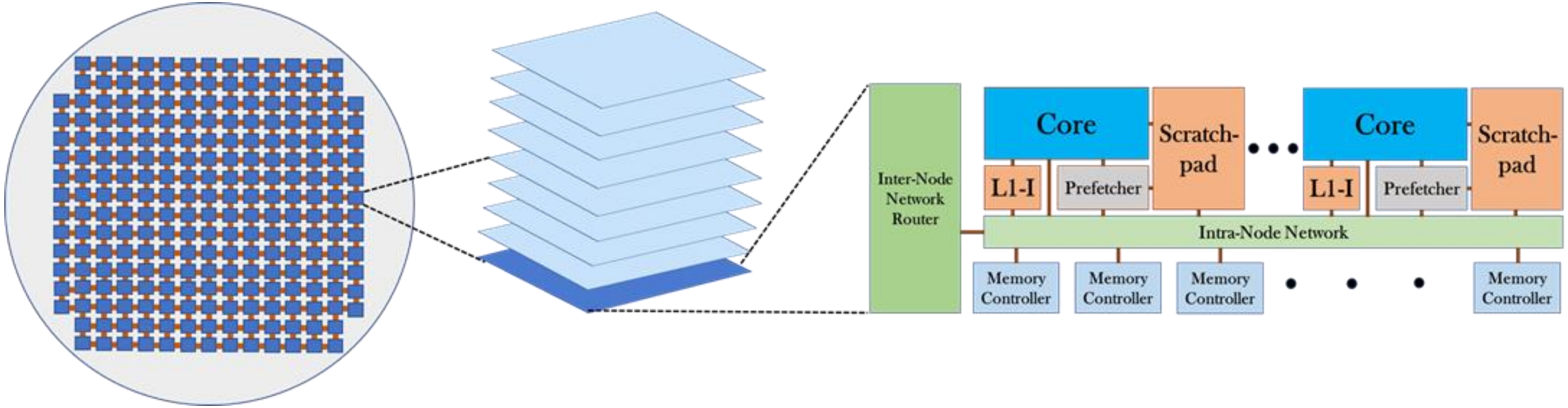
  UCLA: SivaChandra Jangam, Adeel A. Bajwa, Shi Bu, Prof. Sudhakar Pamarti

  UIUC: Matthew Tomei, Daniel Petrisko, Nick Cebry, Jinyang Liu

Thank You

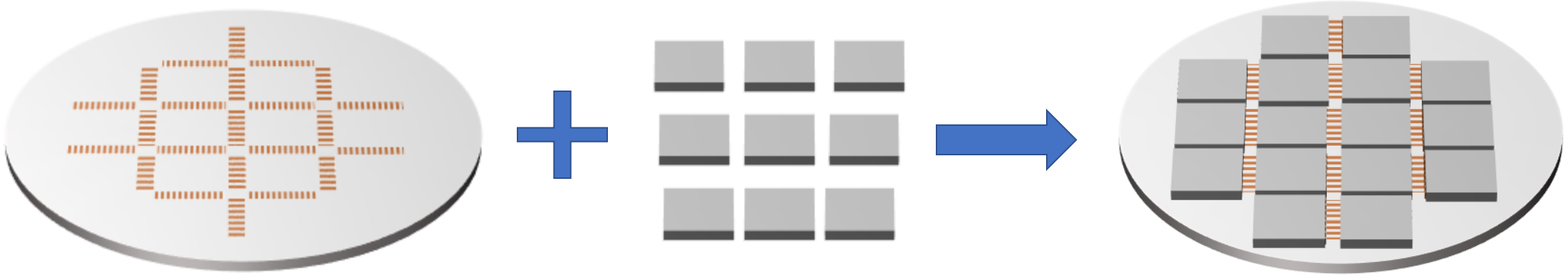
Backup

Waferscale Graph Processor Architecture



- 256 3D stacked nodes with *simple cores* in the logic layer (2 TB memory)
- *Large intra node bandwidth* in 3D memories (up to 1 TBps)
- Mesh based inter-node network. *1.5TBps link bandwidth using Si-IF*
- *Unified memory architecture.* Any core can directly access any memory

Re-imagining Waferscale Integration

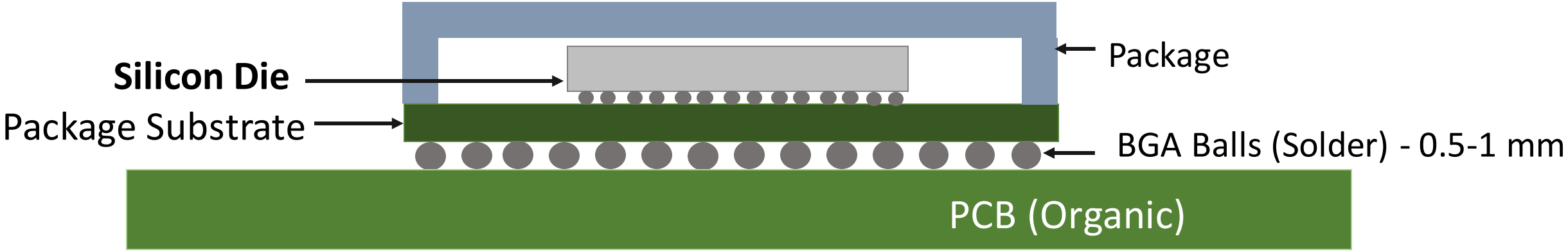


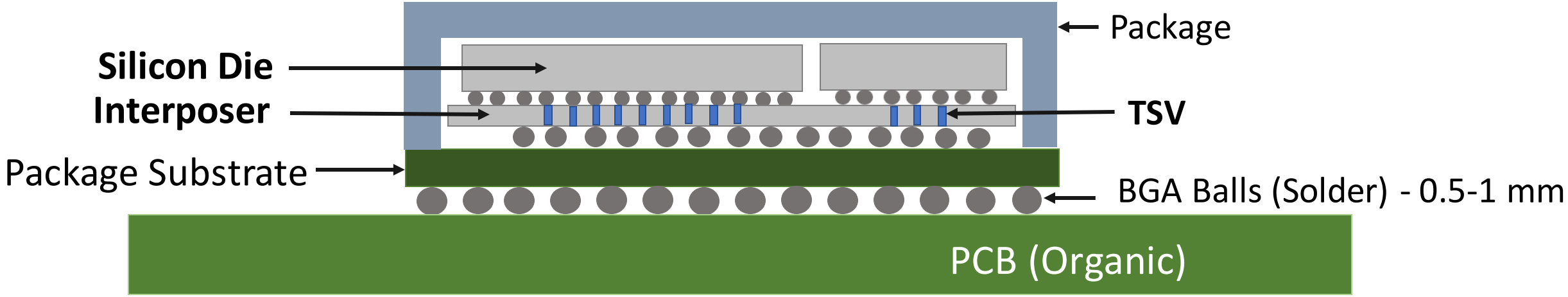
A wafer with interconnect wiring only

Small known good dies

Bond the dies on to the interconnect wafer

Why Interposers or Packages Do not Scale



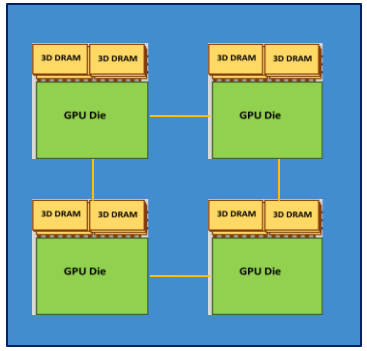


Experimental Methodology

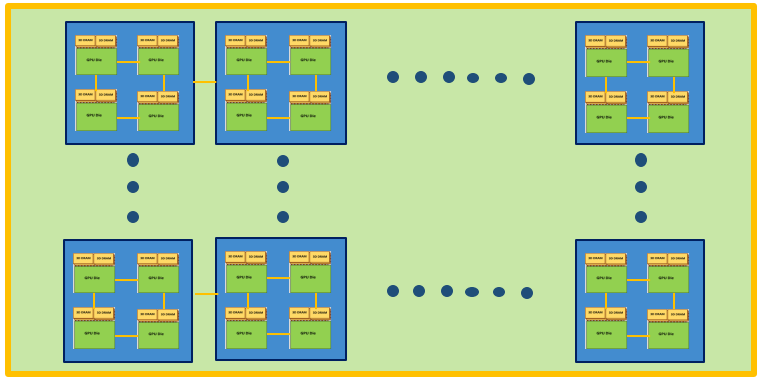
Simulator: In-house Trace-based GPU Simulator (Validated against Gem5-GPU)

Baselines: MCM-GPU, iso-GPM multi-MCM GPU integrated on PCB

Benchmark	Suite	Domain
backprop	Rodinia	Machine Learning
hotspot	Rodinia	Physics Simulation
lud	Rodinia	Linear Algebra
particlefilter naive	Rodinia	Medical Imaging
srad	Rodinia	Medical Imaging
color	Pannotia	Graph Coloring
bc	Pannotia	Social Media



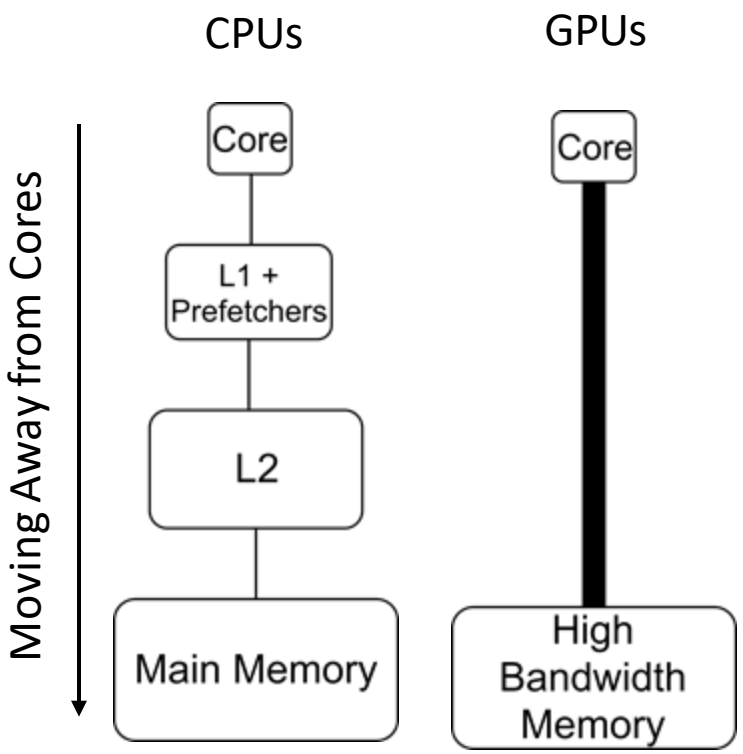
MCM Package



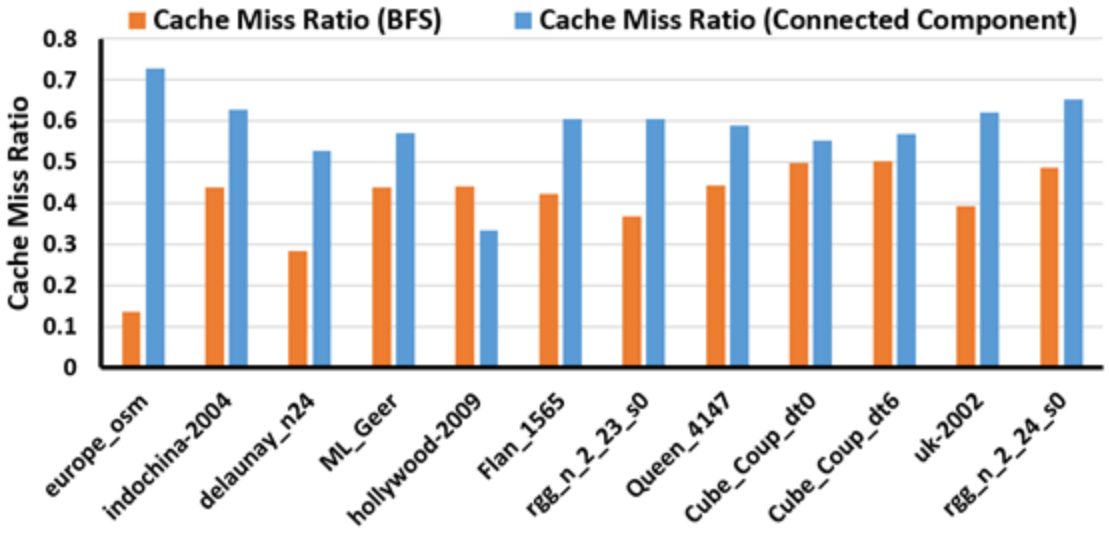
PCB

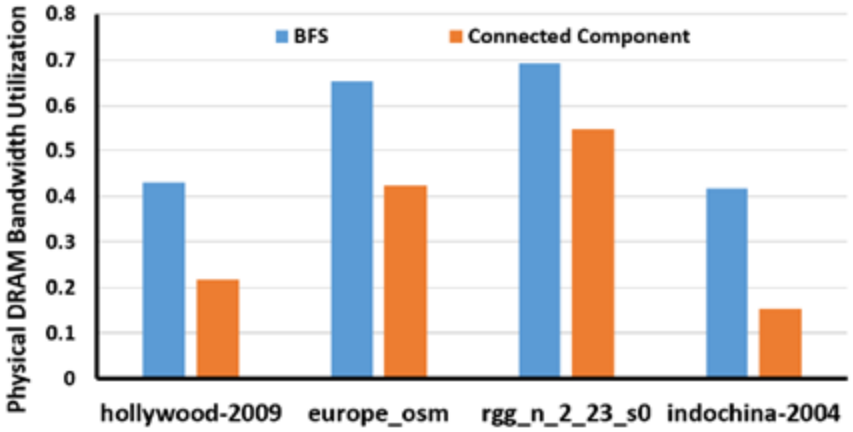
	PCB	Package	WSI
Bandwidth	256GBps	1.5TBps	1.5TBps
Energy	10pJ/b	0.54pJ/b	1pJ/b

Graph Processing Requires a New Architecture



Assumes Locality



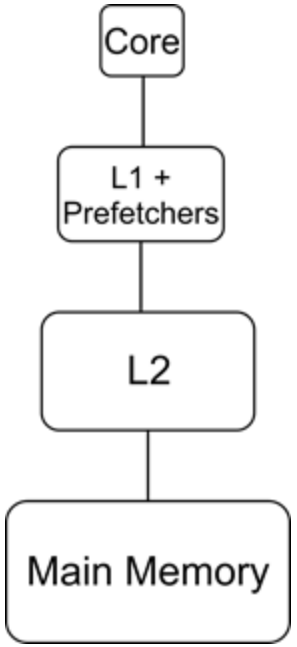


Graph Applications Exhibit Poor Memory Locality

Graph Processing Requires a New Architecture

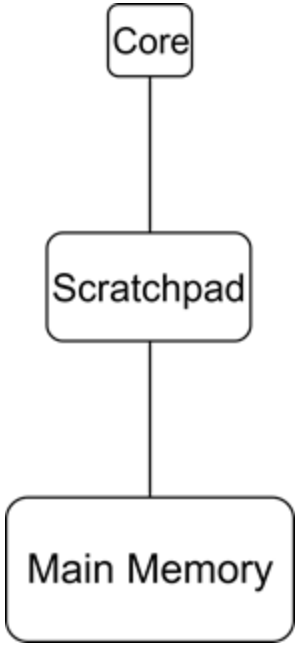
Moving Away from Cores

CPUs

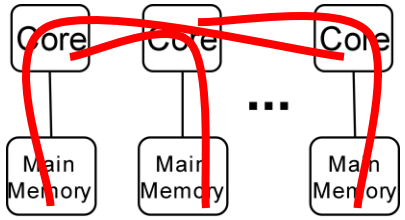


GPUs

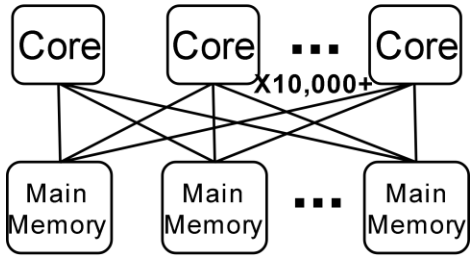
Accelerators



Near Memory Computing/ HPC Clusters



Ideal Architecture



Requires Locality

Requires Predictable Layout

Requires Nothing

Provides High Random Access Bandwidth

Requires Expensive Pre-Processing