Chiplet-based Waferscale Computing

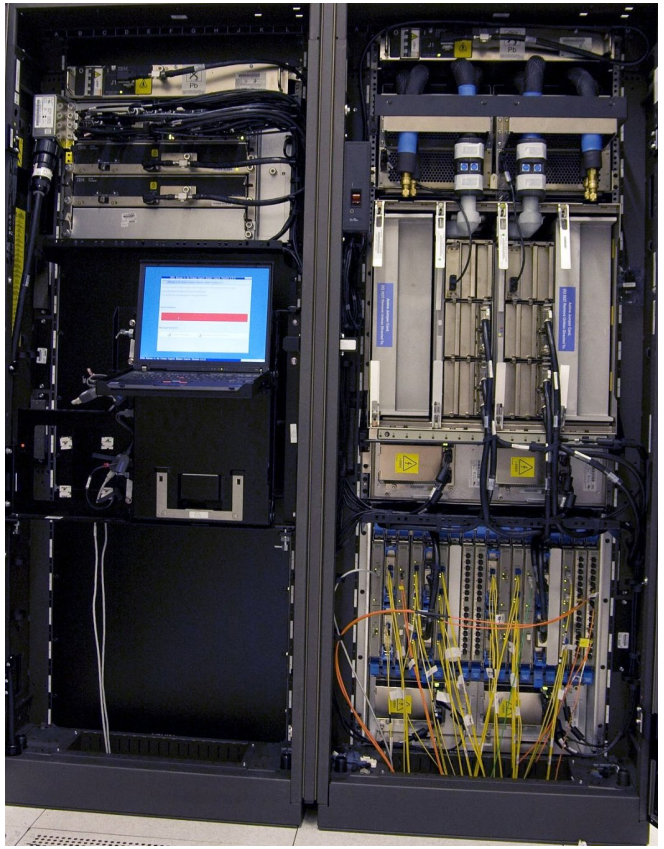
Rakesh Kumar

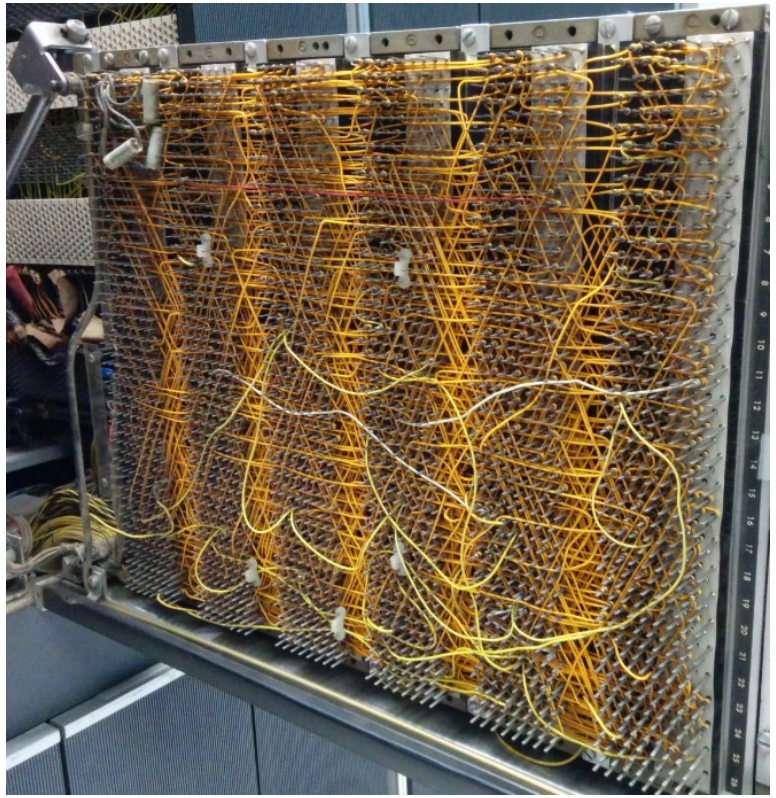
University of Illinois

Urbana-Champaign

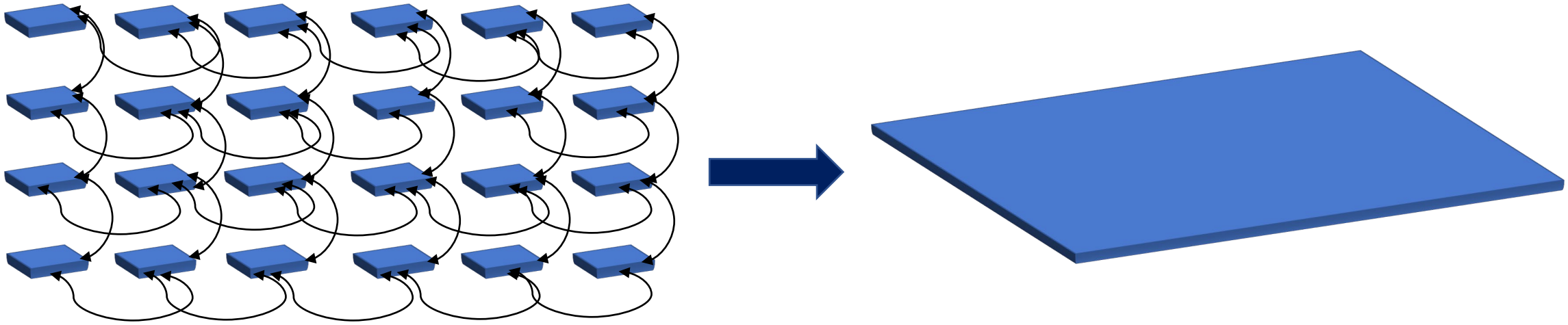
(collaboration with UCLA – afternoon talk will cover the design and implementation aspects of this work)

A Brief History of Waferscale Computing

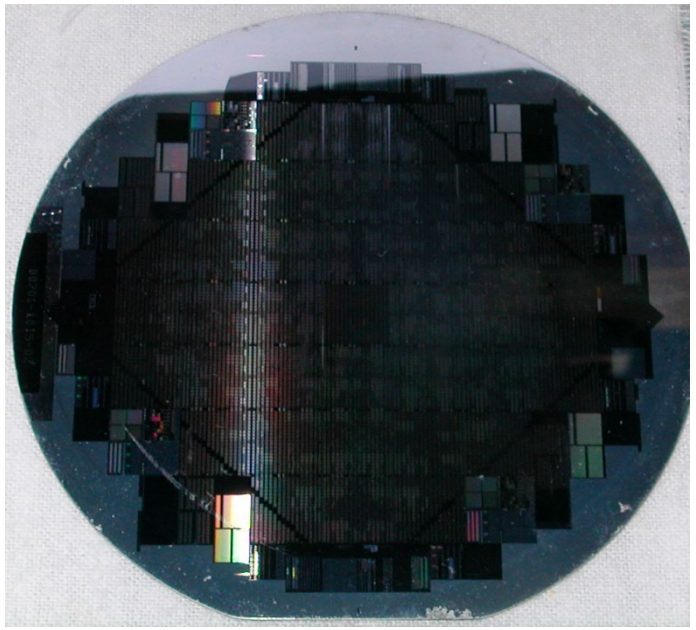




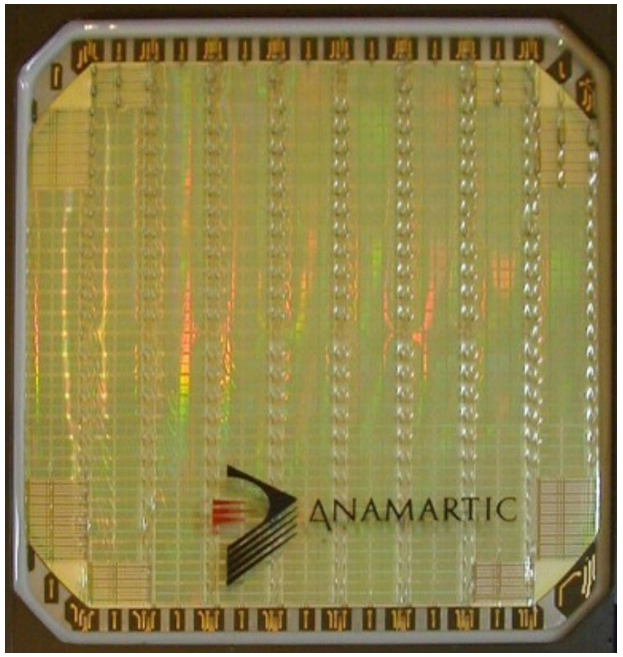
A Brief History of Waferscale Computing



A Brief History of Waferscale Computing



Gene Amdahl's Trilogy Systems

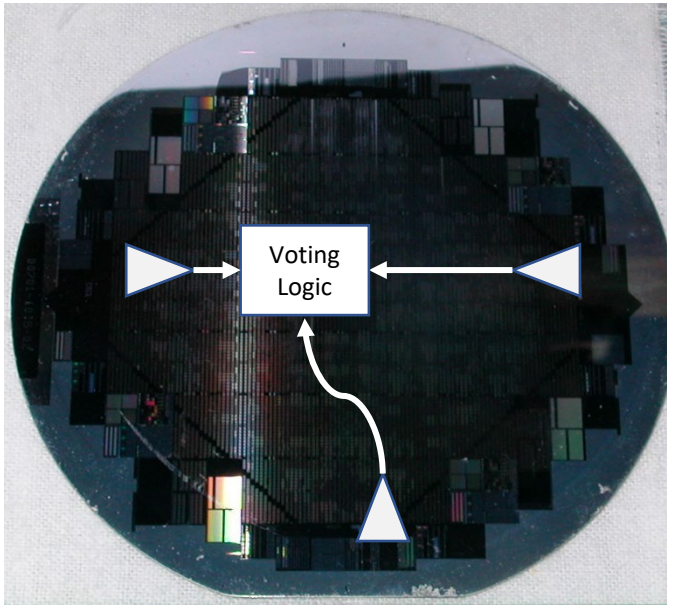


Tandem Computers, Fujitsu

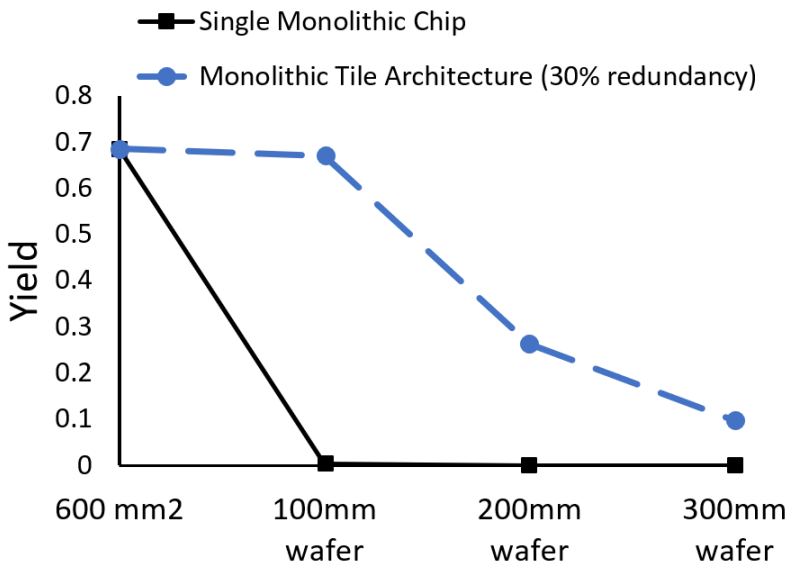
Other efforts: ITT Corporation, Texas Instruments. Recent efforts: Spinnaker (Neuromorphic Chip)

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 ↑

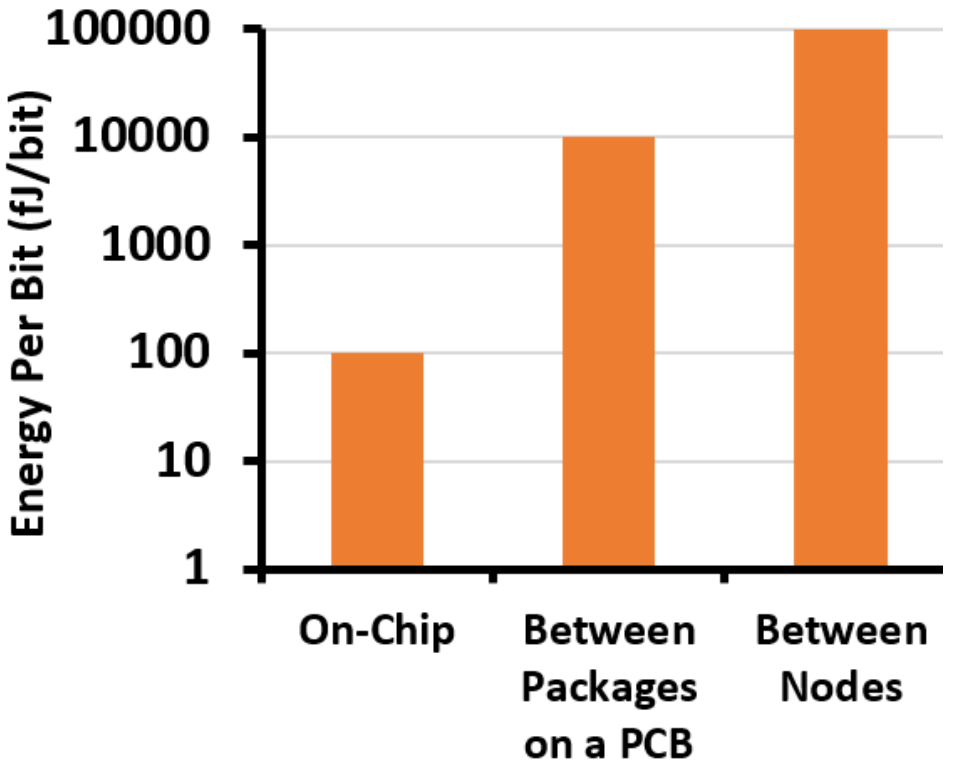
Some mitigation possible through TMR, etc. - but prohibitively expensive

Deemed commercially unviable

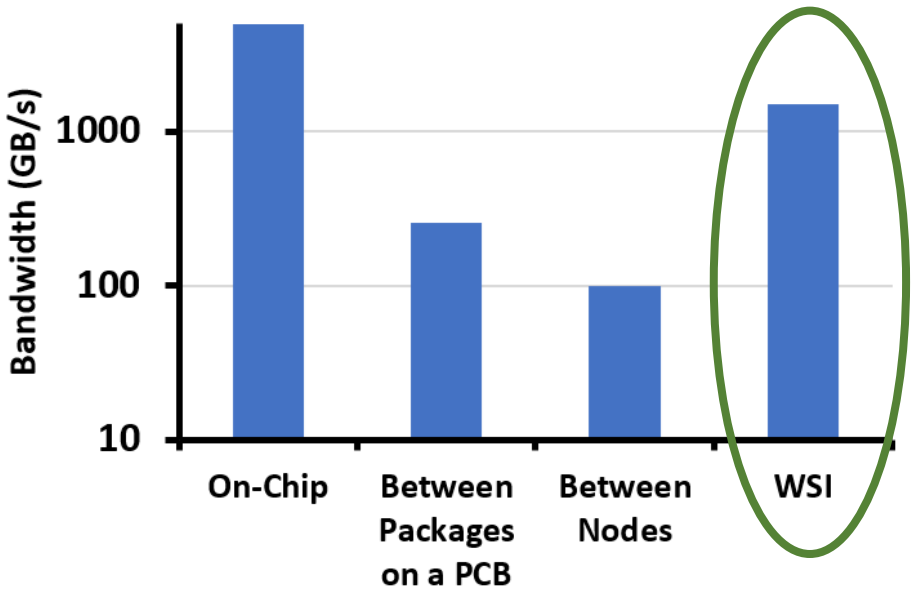
Time to Give Waferscale Another Go?

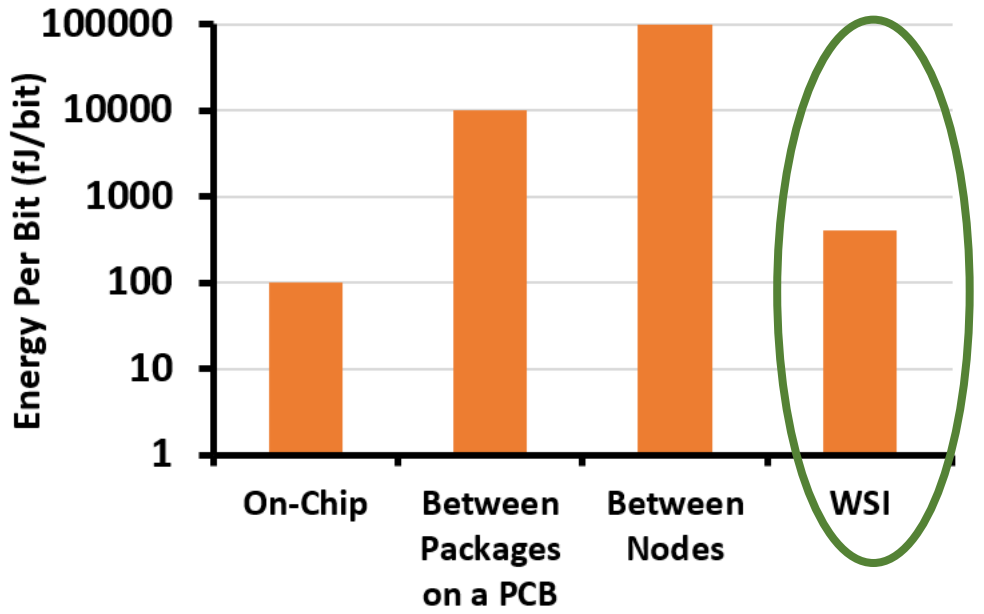
- Highly parallel applications are spread across many processors

- 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)



Time to Give Waferscale Another Go? (2)



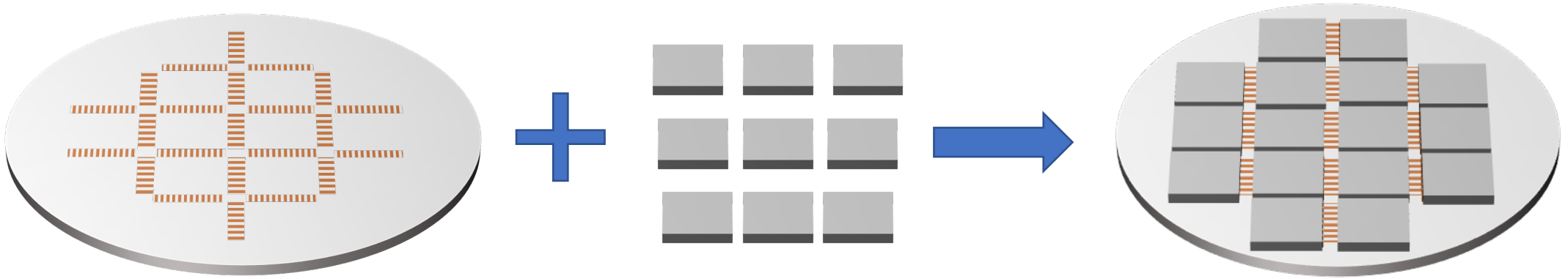


However, to achieve waferscale integration, we need to solve the **yield problem**

Re-imagining Waferscale Integration

Q: What do we need from waferscale integration?

A: High density interconnection



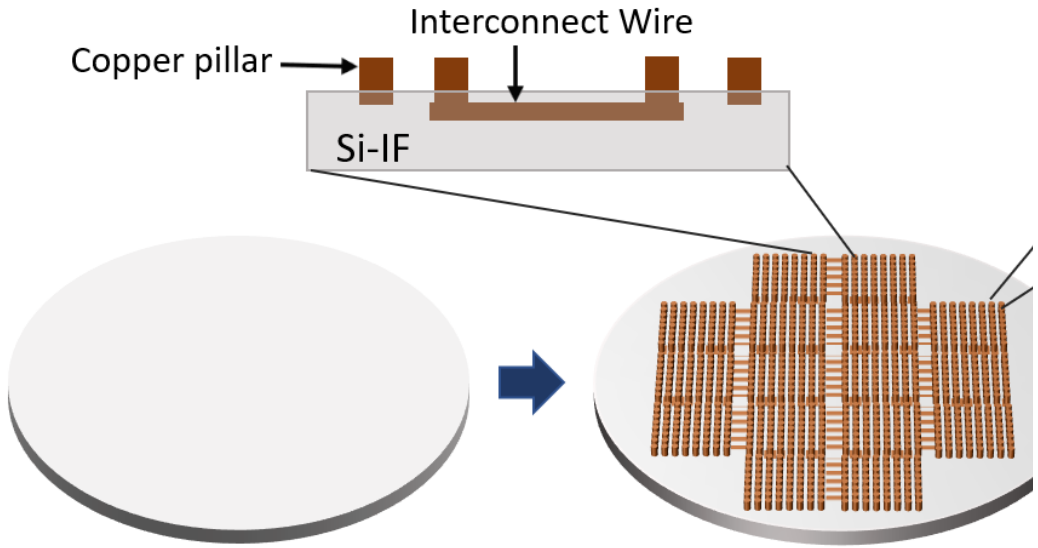
A wafer with interconnect wiring only

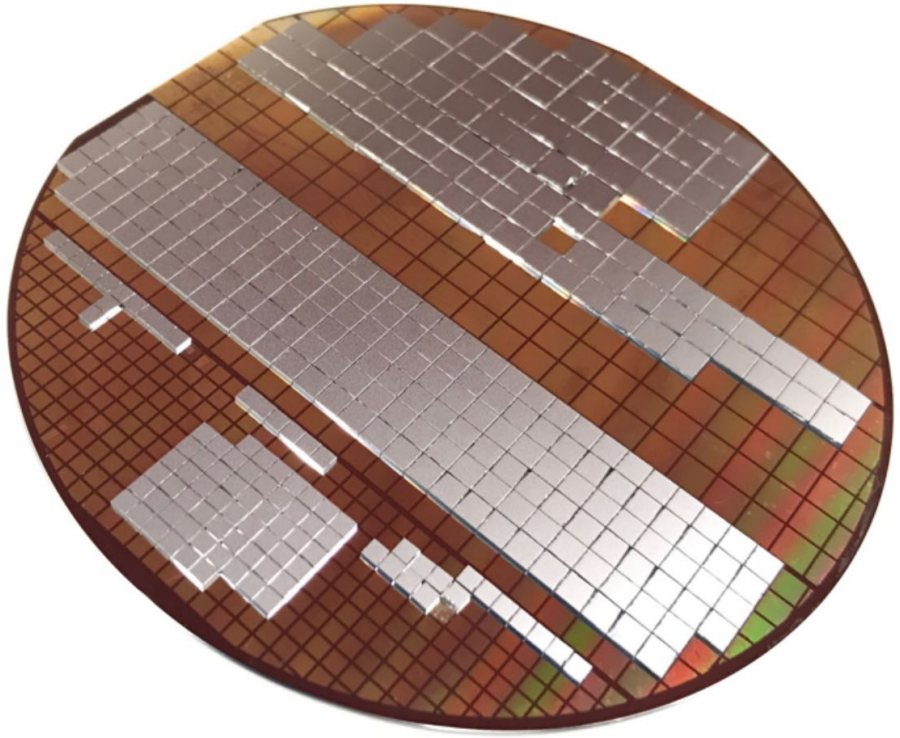
Small known good dies

Bond the dies on to the interconnect wafer

Enabling WSI Technology

Silicon Interconnect Fabric (Si-IF)





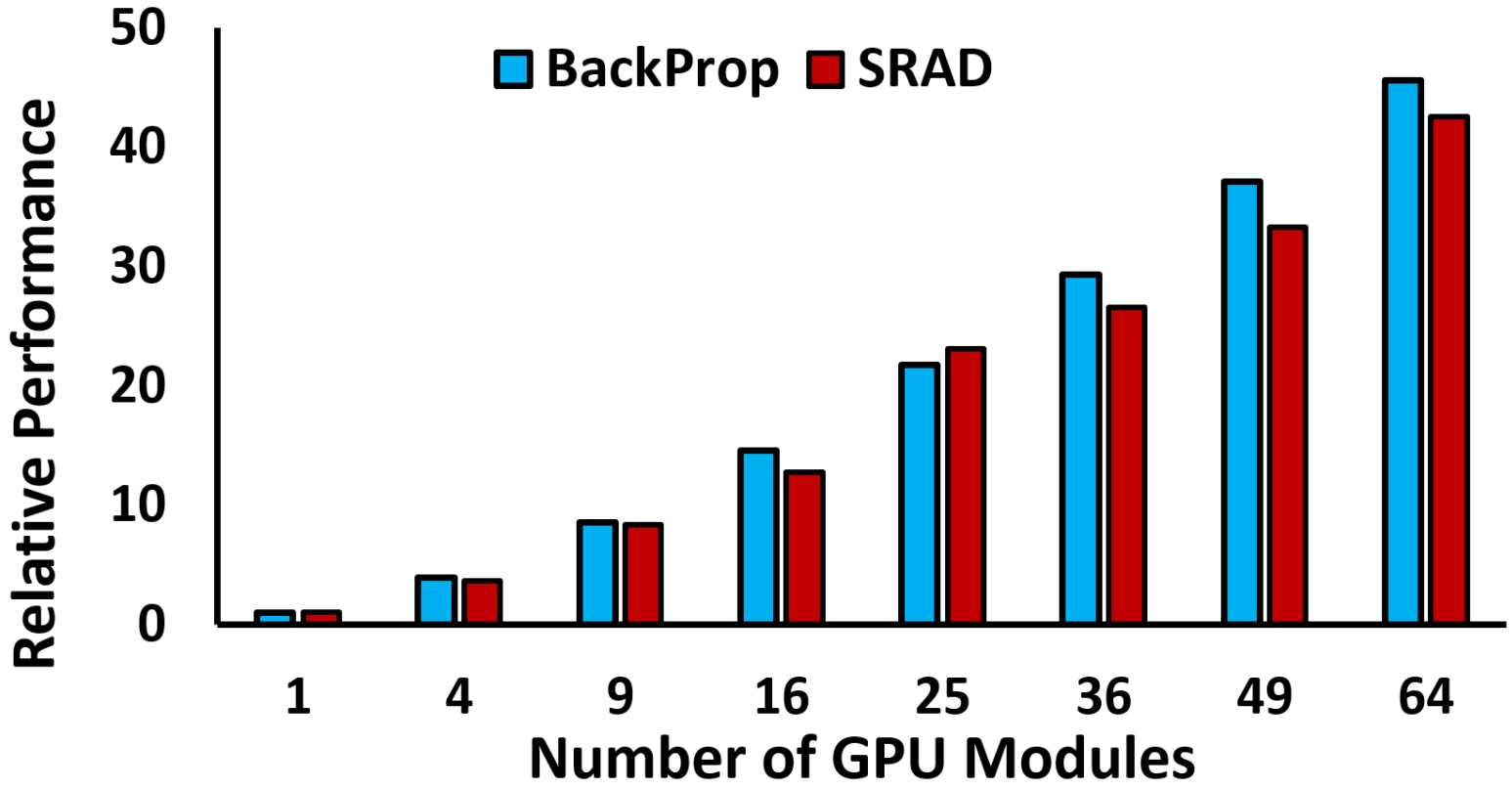
Measured Bond Yield >99%

Allows waferscale integration with high yield

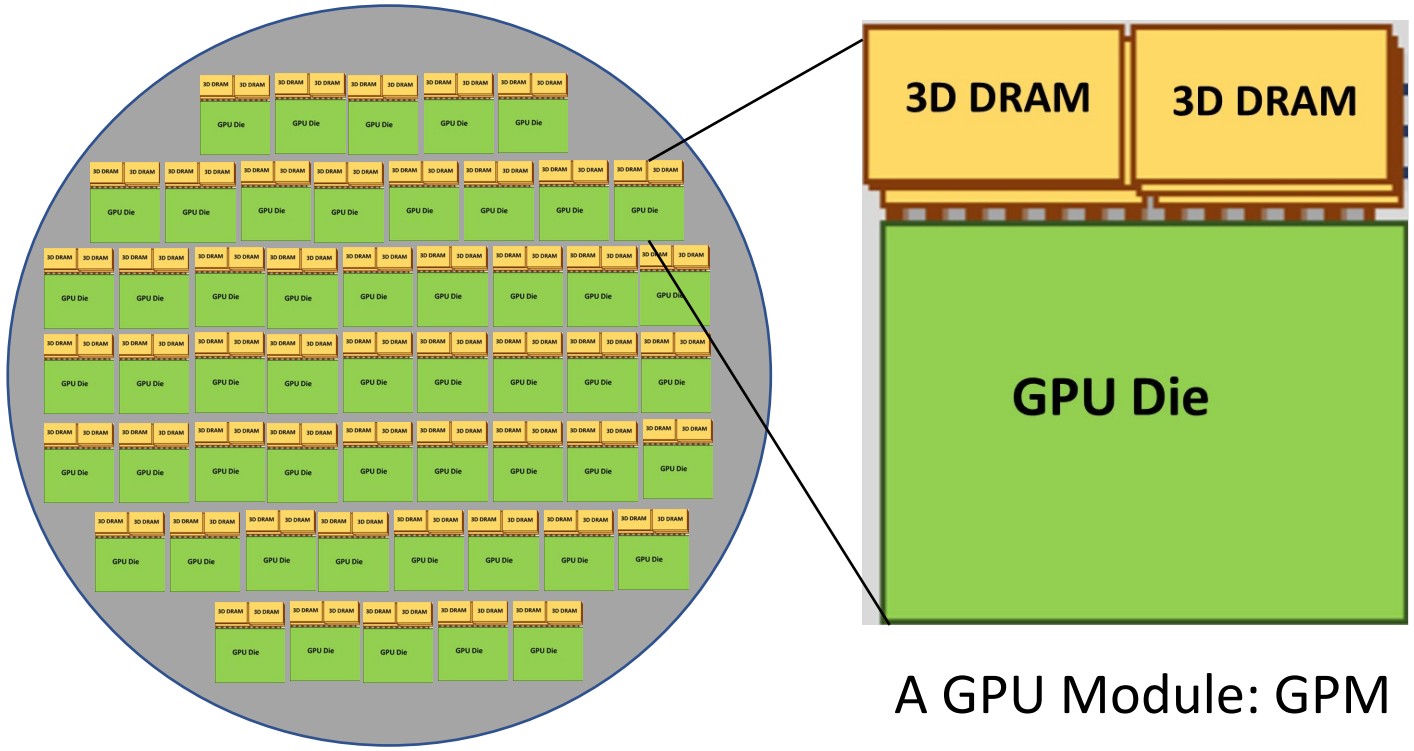
[HPCA2018]

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 ?

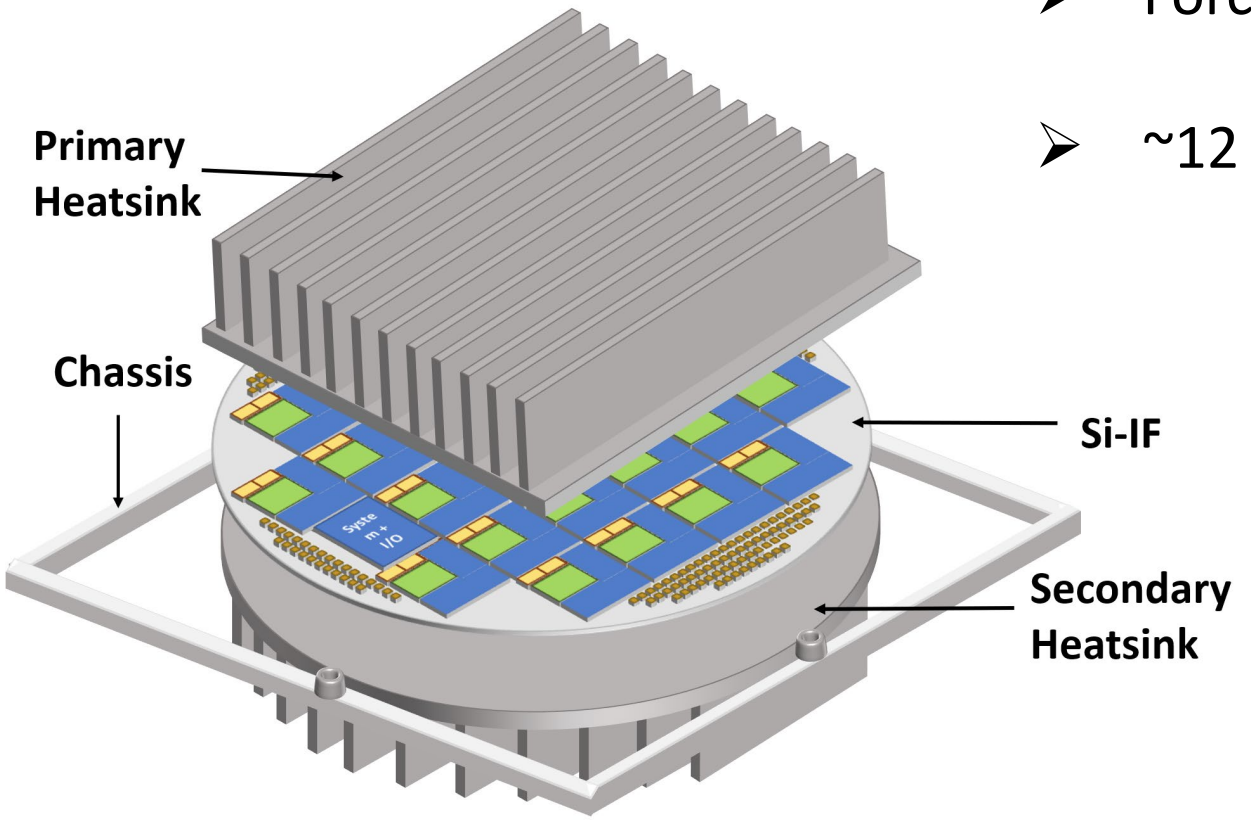
Three major physical constraints:

1. **Thermal**
   - Waferscale GPU would dissipate kWs of power

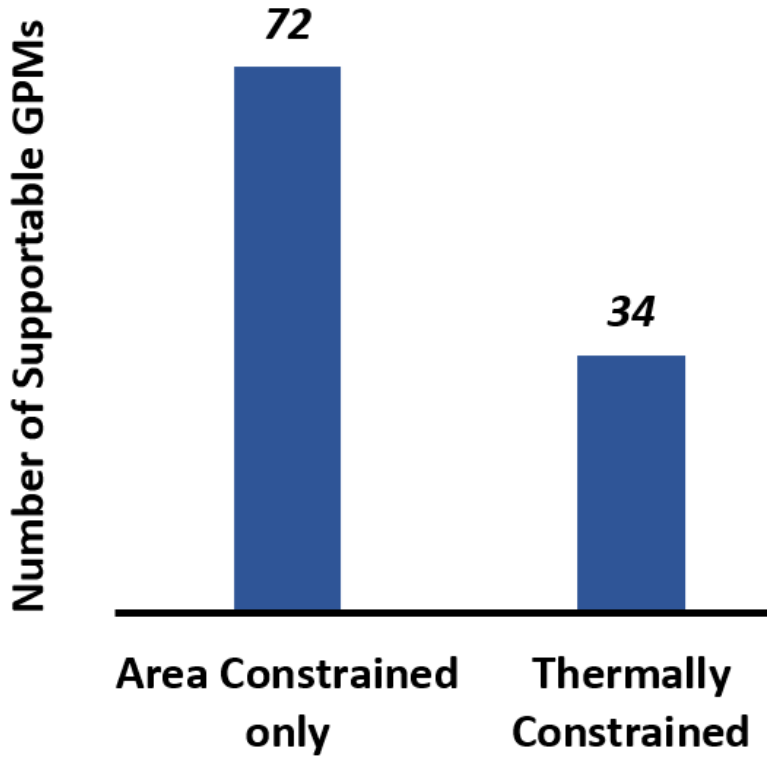
2. **Power Delivery**
   - How to supply kWs of power to the GPU modules?
   - Voltage Regulator Module (VRM) overhead?

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

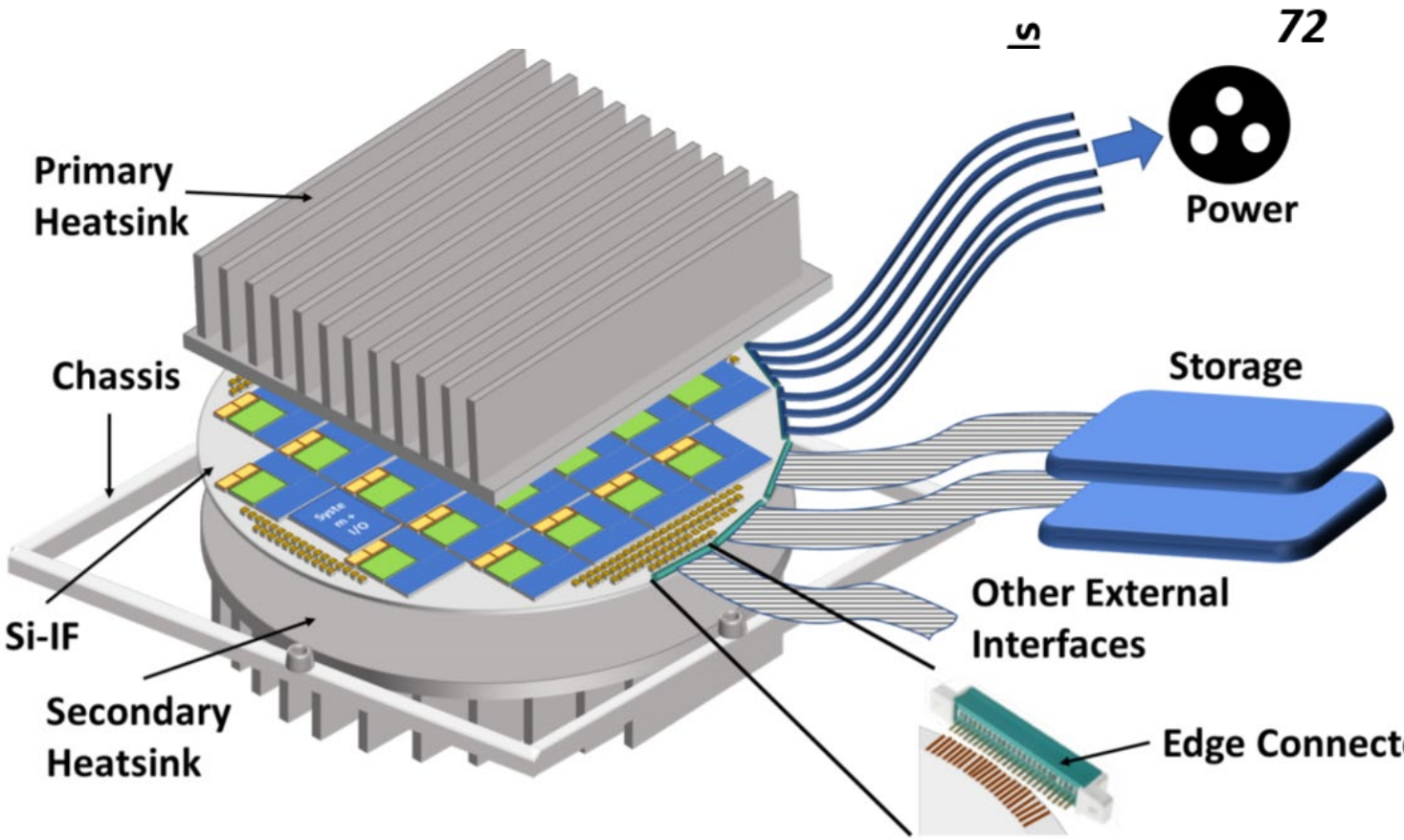
Thermal Design Power

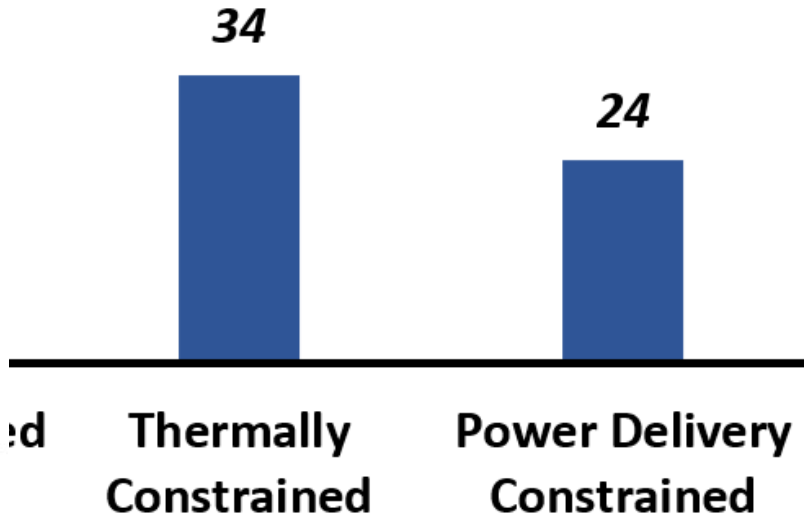


- Forced air-cooling with two heat sinks
- ~12 kW TDP i.e., **34 GPMs** can be supported

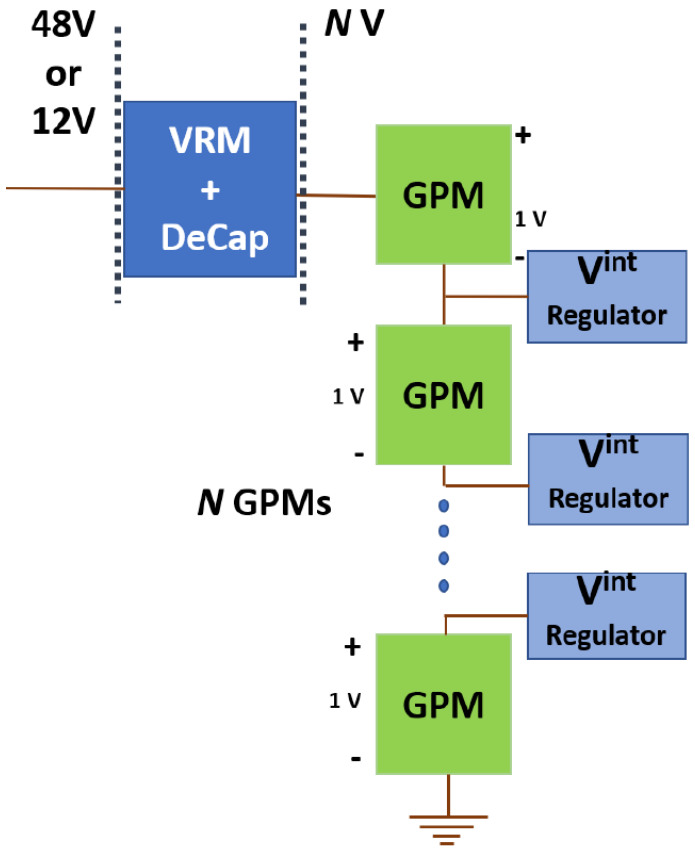


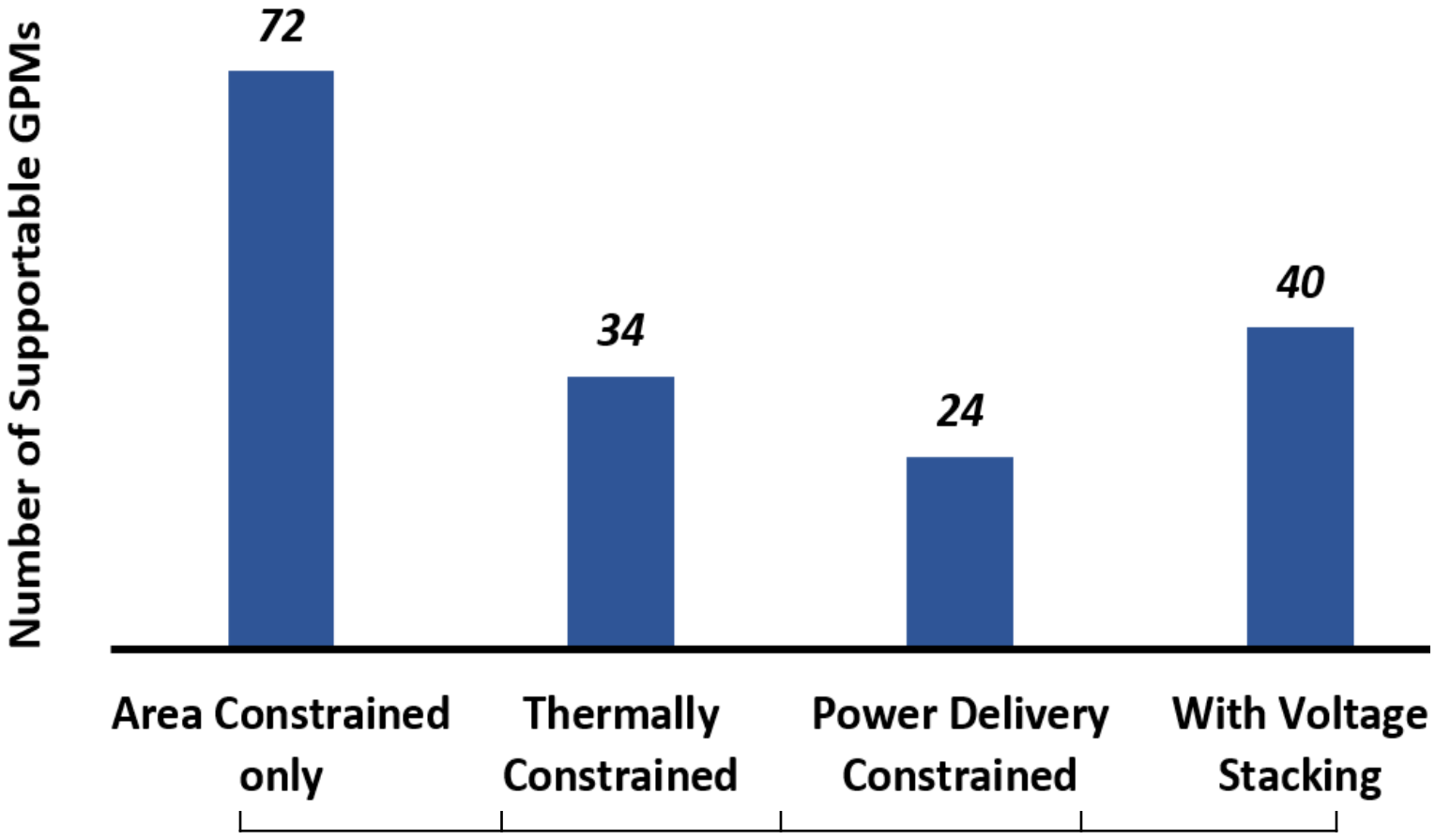
Power Delivery



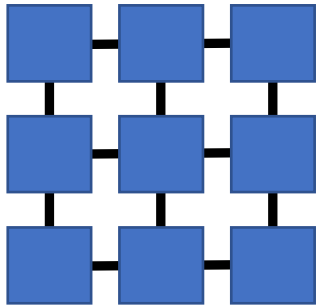


Stacked Power Delivery

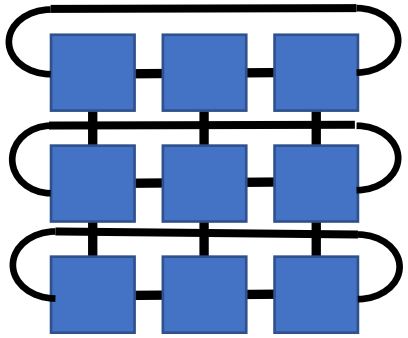




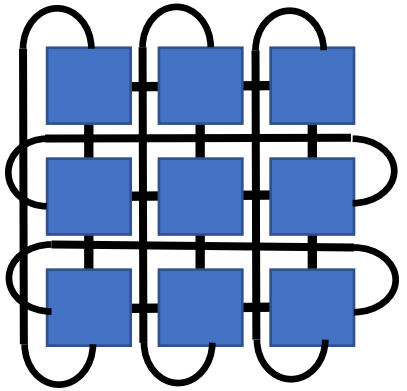
Waferscale Inter-GPM Network



Mesh



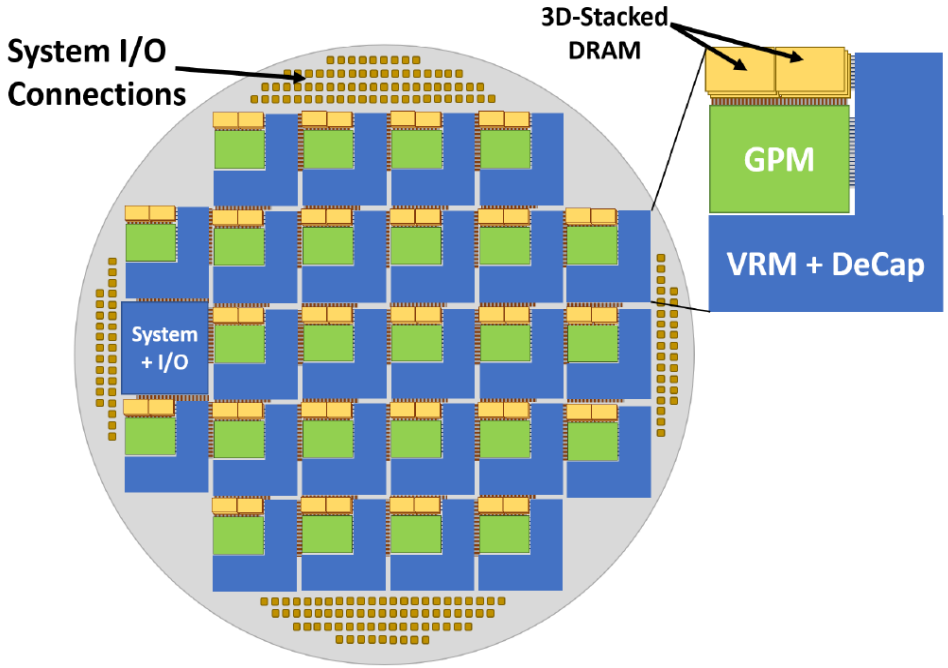
1D-Torus

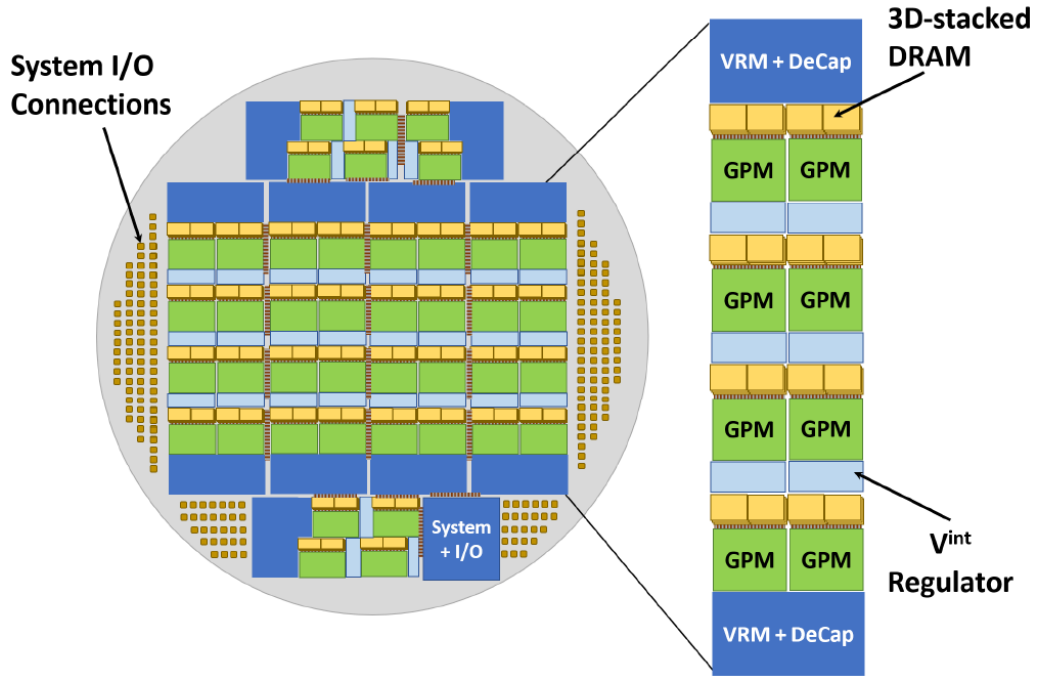


2D-Torus

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**

Thread Block Scheduling and Data Placement

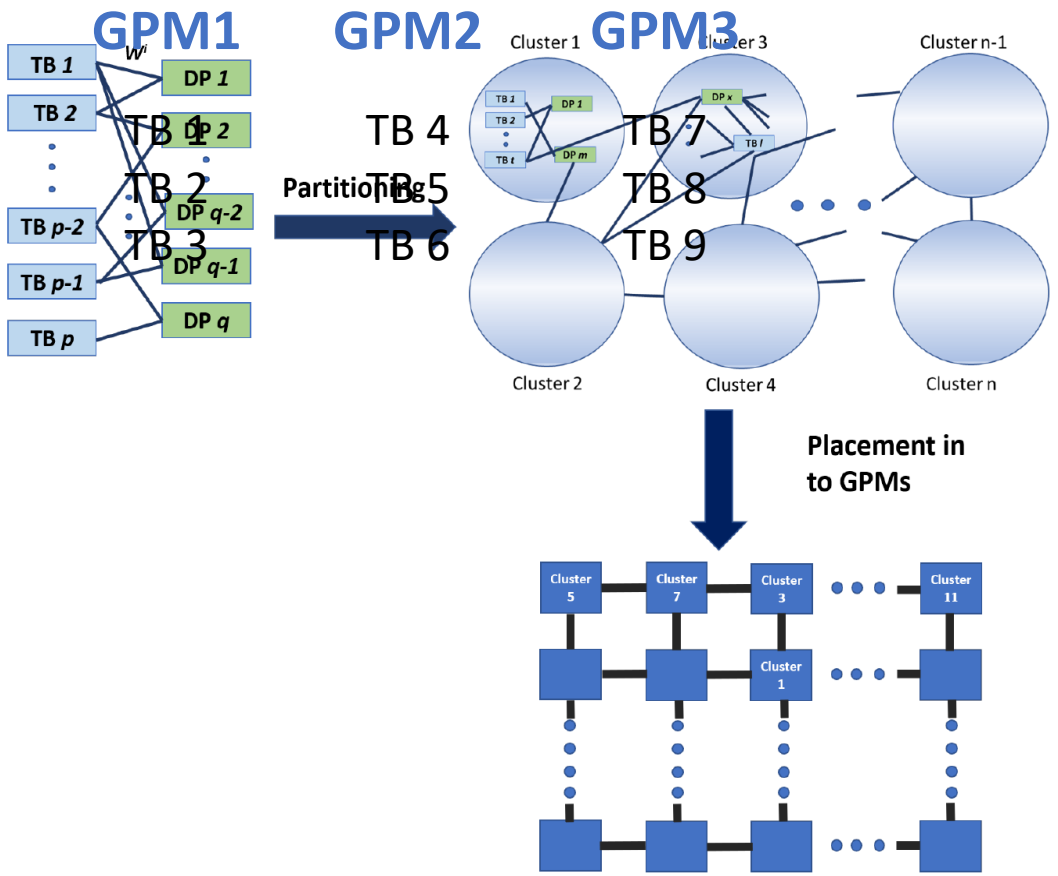
Dynamic Online [1]:

- Contiguous TBs placed in the same GPM
- First-touch data placement

Static Offline:

- Recursive Partitioning based on *Fiduccia-Matthessey* algorithm

- Logical Cluster to Physical GPM mapping → *Simulated Annealing (SA)* based placement



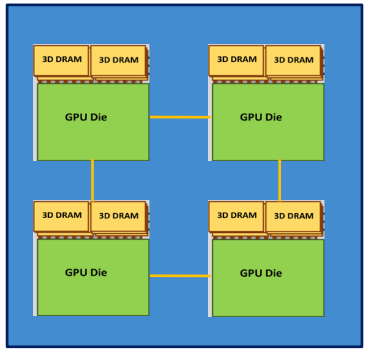
[1] "MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability", A. Arunkumar et. al., ISCA 2017

Experimental Methodology

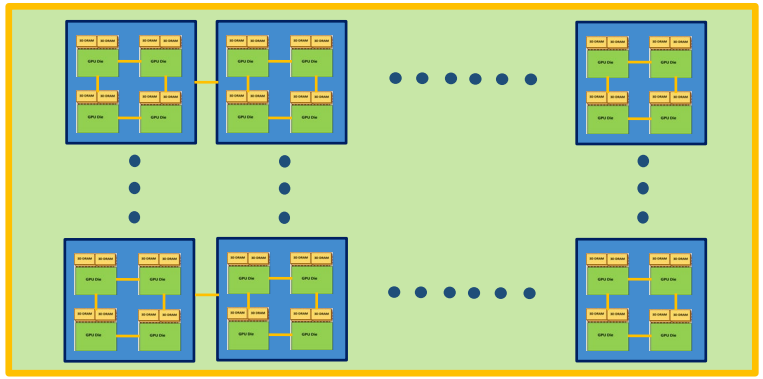
Simulator: In-house Trace-based GPU Simulator (Validated against Gem5-GPU)

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



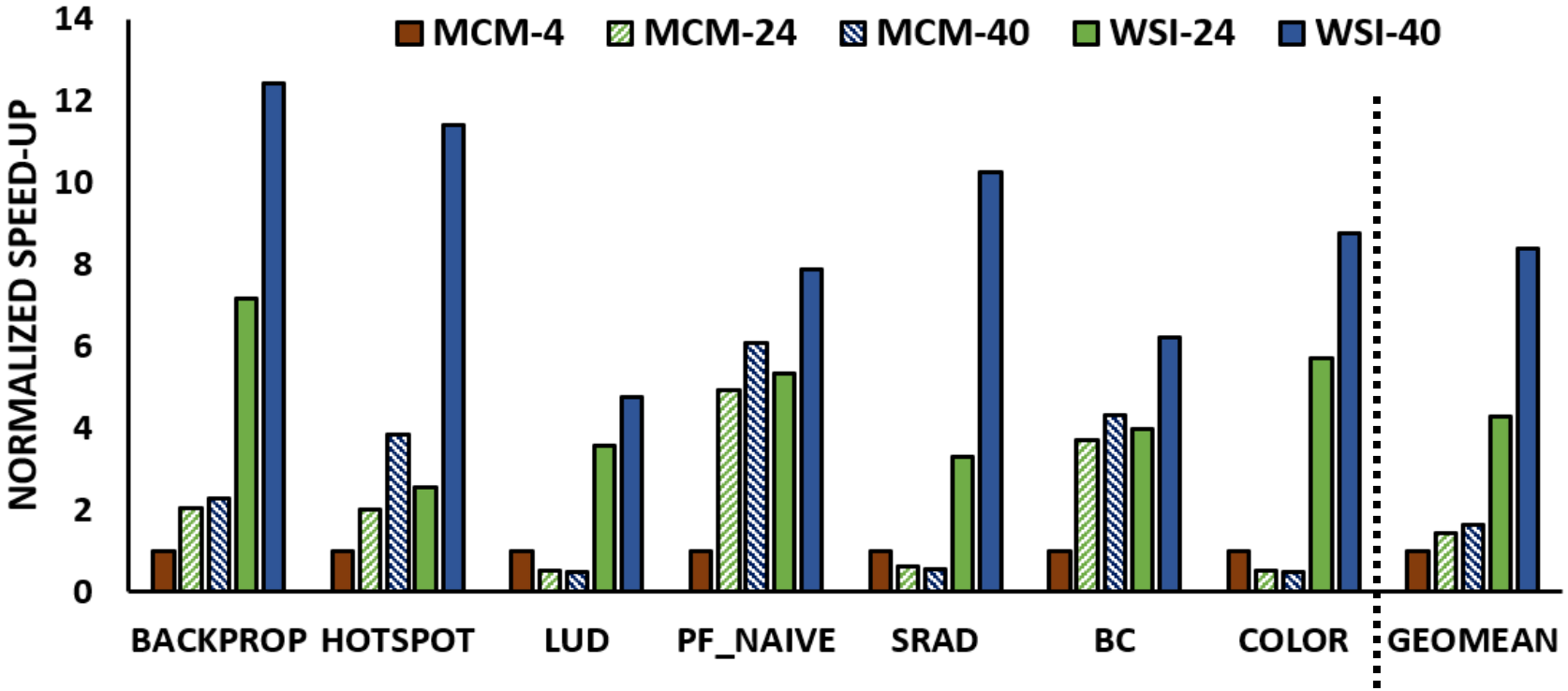
MCM Package



PCB

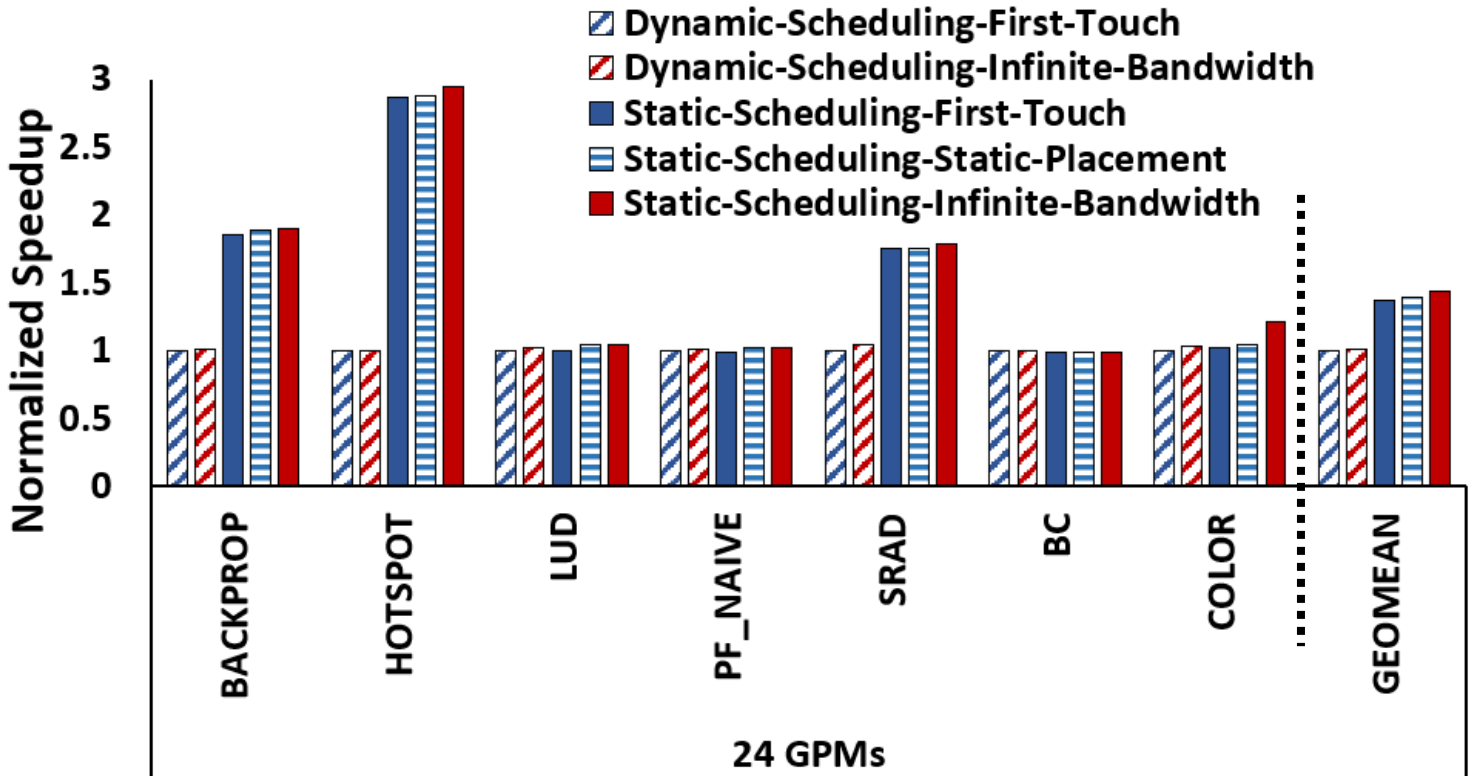
	PCB	Package	WSI
Bandwidth	256GBps	1.5TBps	1.5TBps
Energy	10pJ/b	0.54pJ/b	1pJ/b

Results – WS-GPU Performance Improvement



- 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**)
- With dynamic online scheme, WSI's speedup improves by another ~**2x**

Results – Speedup using the Static Scheme

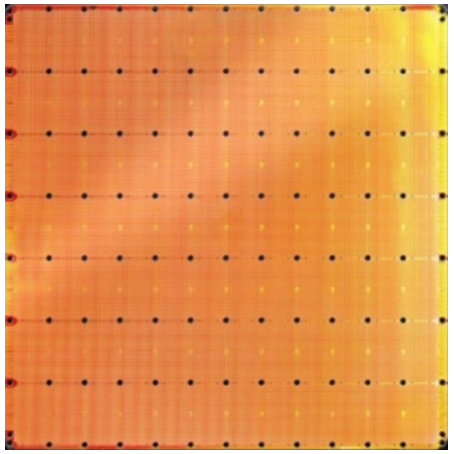


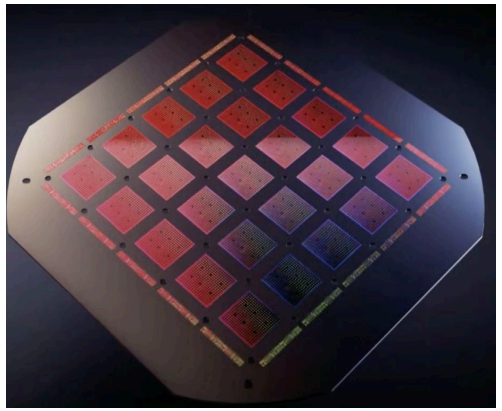
- Improvement of up to **2.88x (average 1.4x)**
- Optimization in scheduling impacts speedup more than data placement

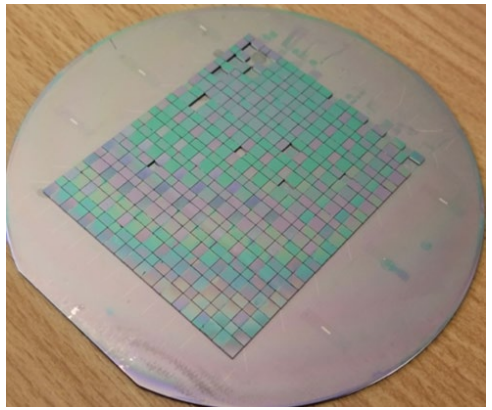
Summary and Conclusion

- Communication between packaged processors is a major bottleneck
- Si-IF technology enables waferscale integration
- Waferscale GPU versus multi-MCM system:
  - **5.2x** performance improvement
  - **22.5x** EDP improvement
- Intelligent scheduling can provide up to **2.88x** (average 1.4x)speedup
- Advanced power and thermal architecture has the potential to improve performance further

Re-emergence of Waferscale Technologies







	Cerebras	Tesla Dojo	Ours
Heterogeneous Integration	**No**	**Yes**	**Yes**
Core Count	**High**	**High**	**High**
Memory Capacity	**Low**	**Low**	**High**
Network Bandwidth	**High**	**High**	**High**
Inter-Die Hop Latency	**Low**	**High**	**Low**

23