NC STATE UNIVERSITY

Design of 3D-Specific Systems

Paul Franzon

North Carolina State University

Raleigh, NC

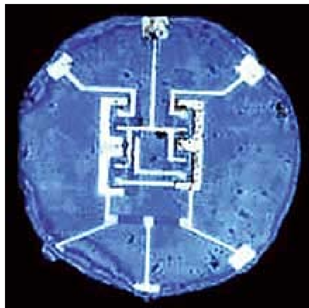
paulf@ncsu.edu

919.515.7351

IEEE CPMT Society

History of Integration

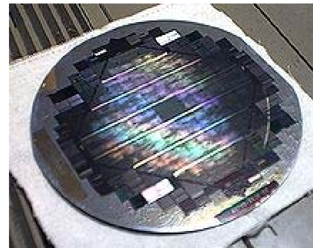
Integrated Circuit



Fairchild

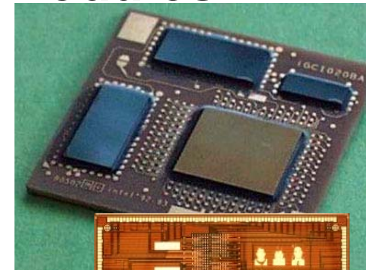
Printed Circuit Board

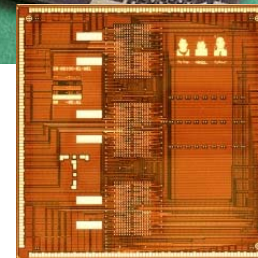
Wafer Scale Integration



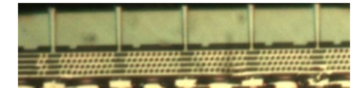
Trilogy

Multichip Modules





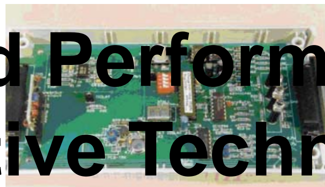
3DIC with TSV

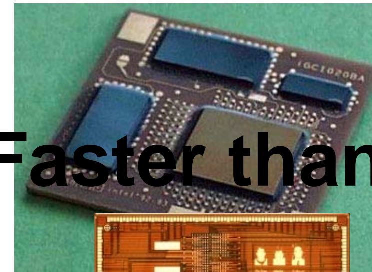
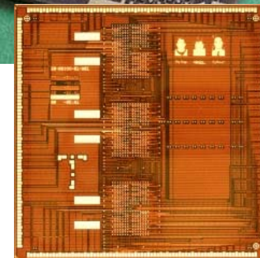


Tezzaron

1960 → **2010**

History of Integration



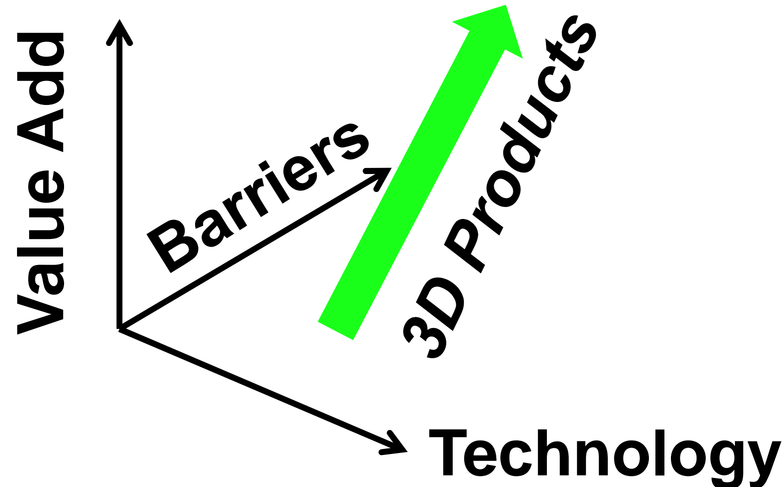



Icreased Performance/Price Faster than Alternative Technologies

How much? >~ 25%

Outline

- Overview of the Vectors in 3D Product Design



- Short term – find the low-hanging fruit
- Medium term – Logic on logic, memory on logic
- Long term – Extreme Scaling; Heterogeneous integration; Miniaturization
- Overcoming Barriers to Employment

Future 3DIC Product Space

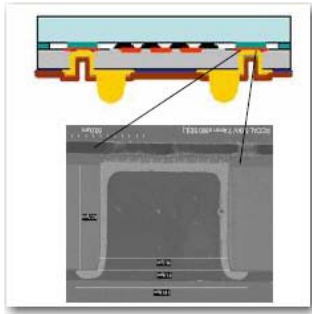
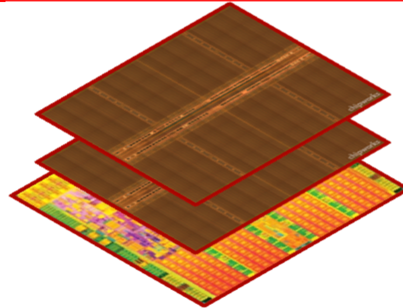
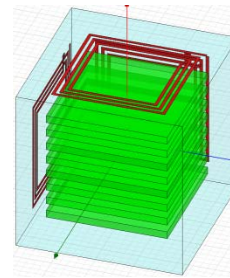


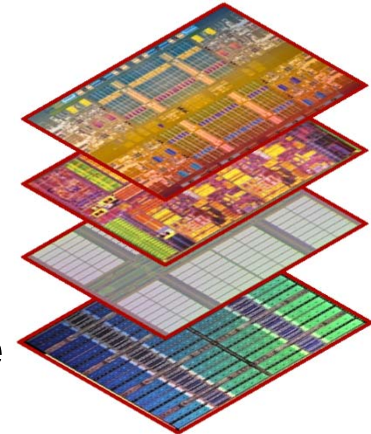
Image sensor



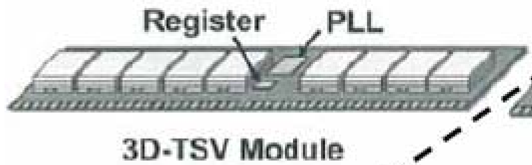
3D Mobile



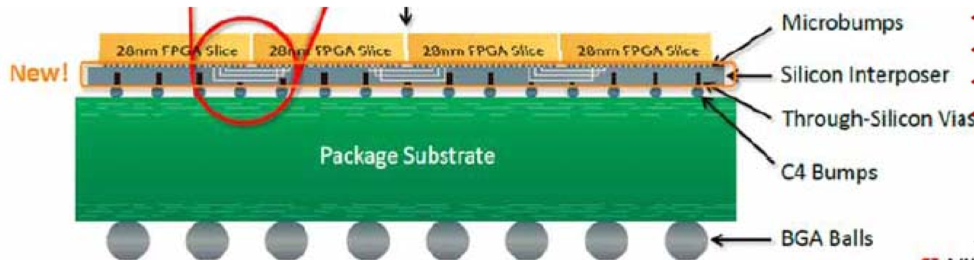
Sensor Node



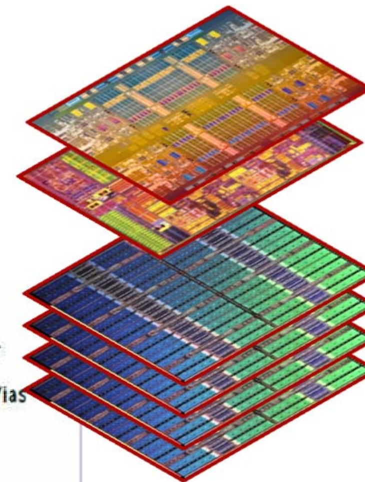
Heterogeneous



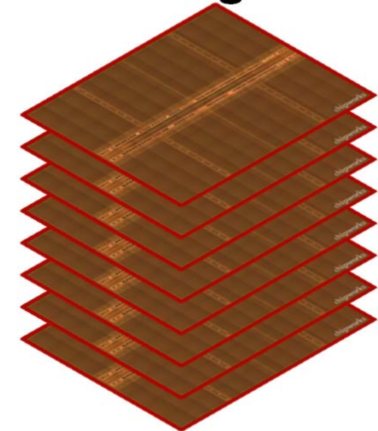
Server Memory



Interposer



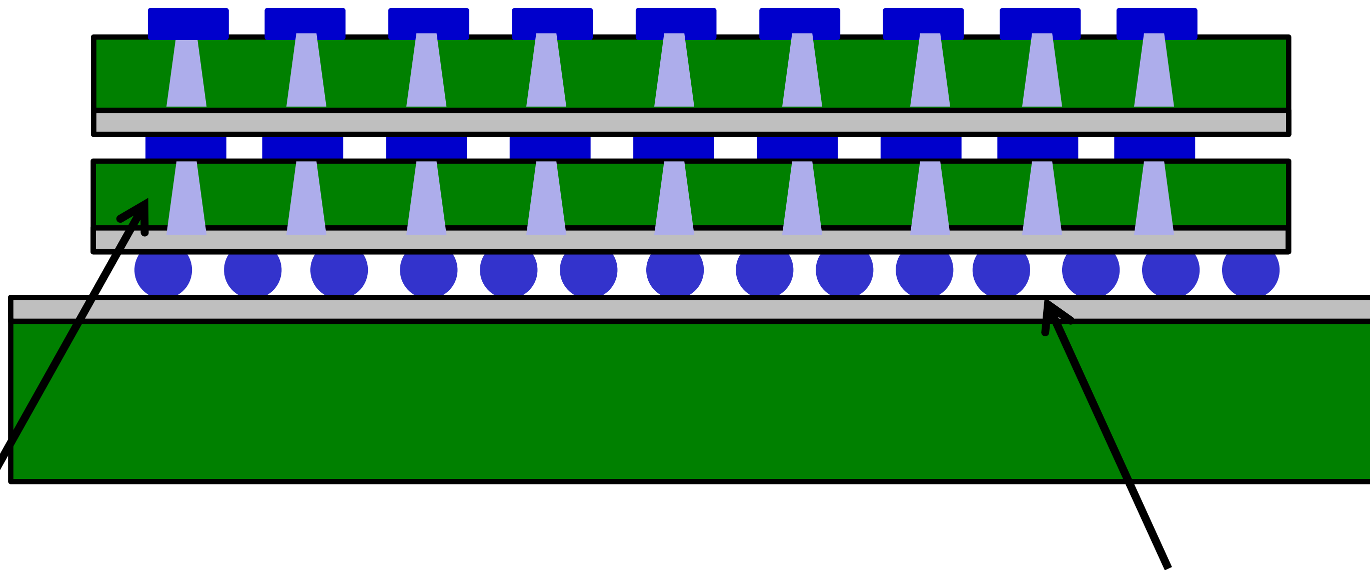
3D Processor



“Extreme” 3D Integration

Time

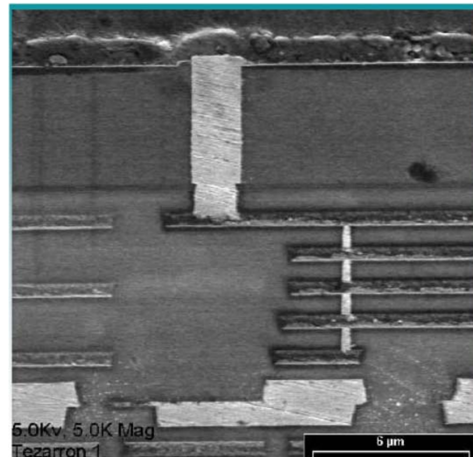
3DIC Technology Set

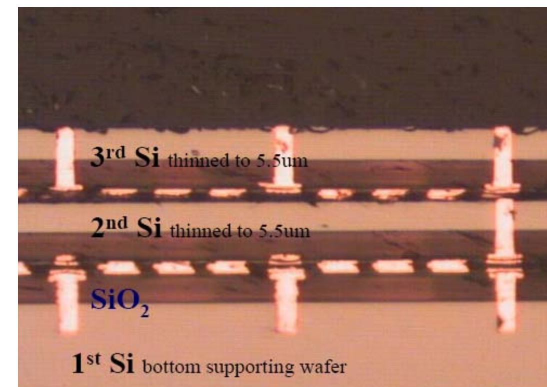


Bulk Silicon TSVs and bumps (25 - 40 μm pitch)

Face to face microbumps (1 - 30 μm pitch)

C_{TSV} ~ 30 fF

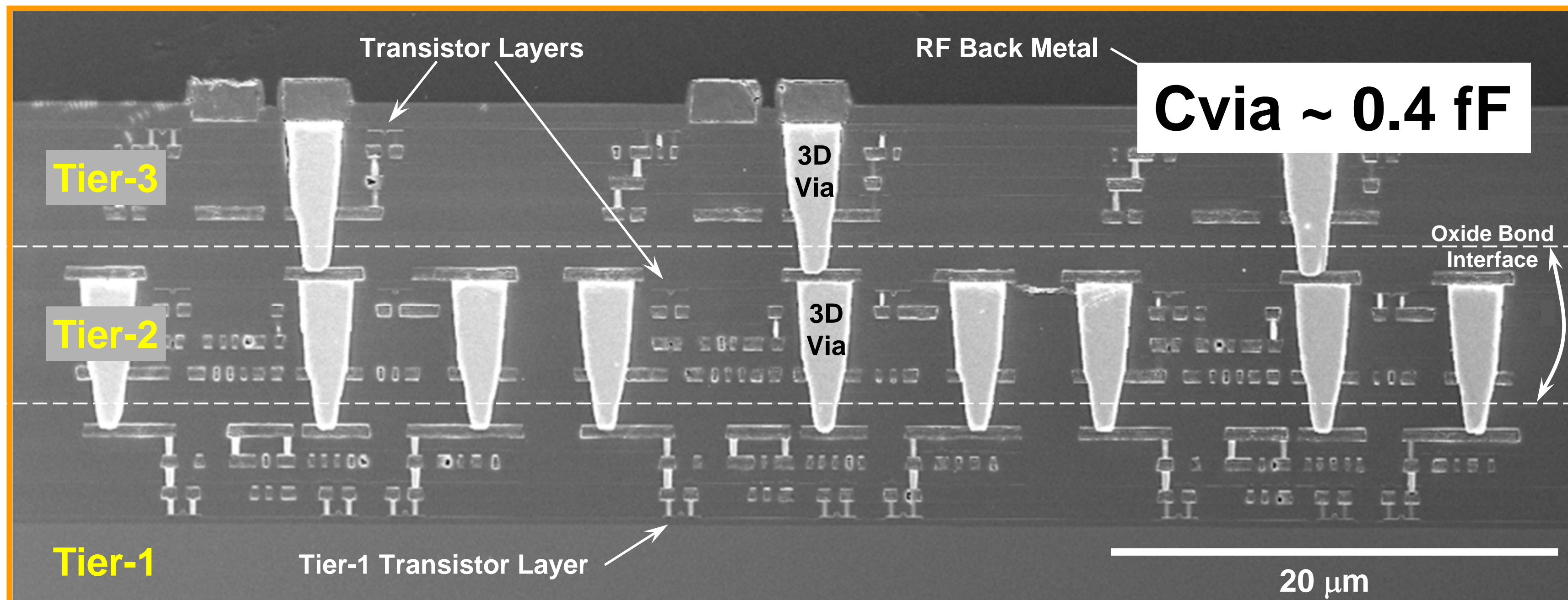




Tezzaron

… 3DIC Technology Set

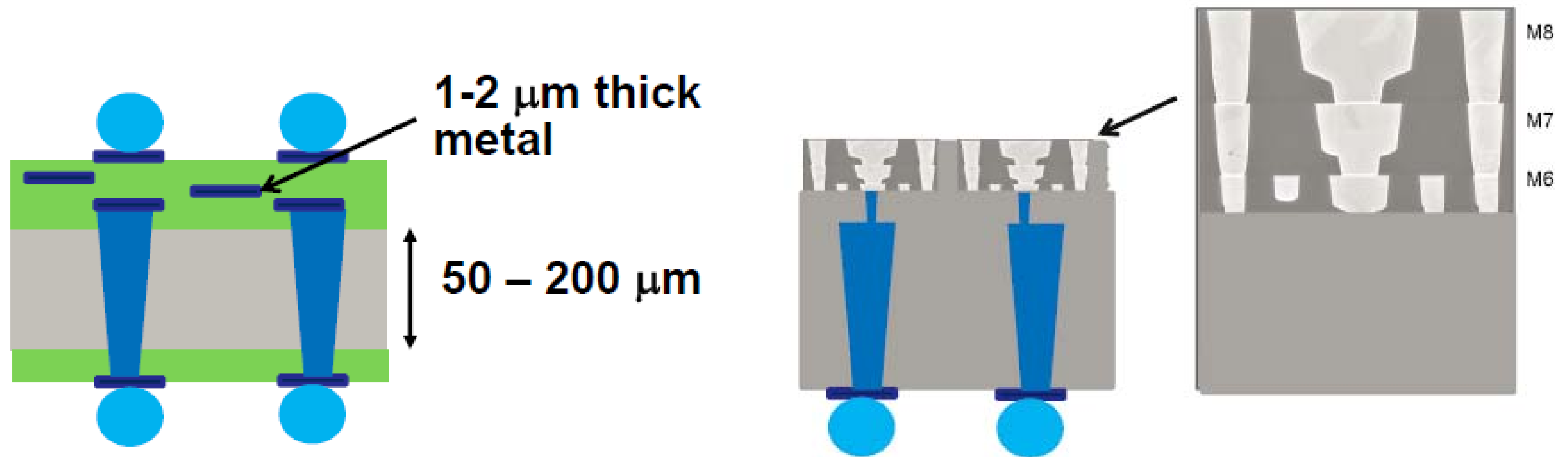
TSVs in an SOI process



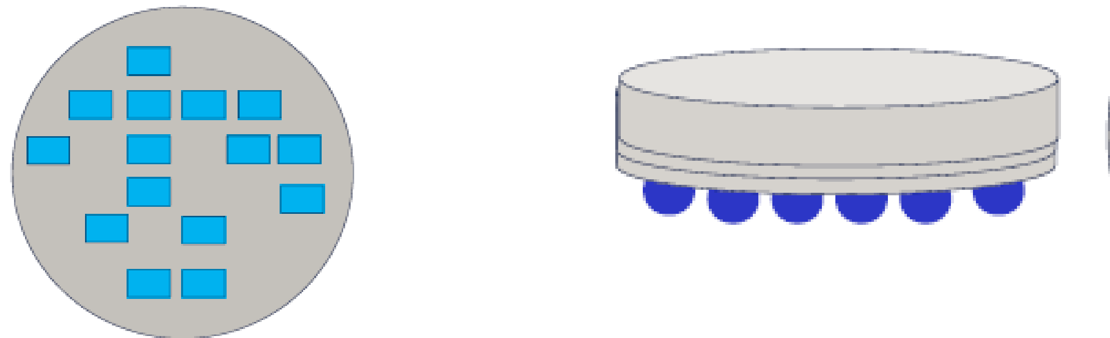
MIT Lincoln Labs

3DIC Technology Set

- Interposers: Thin film or 65/90 nm BEOL



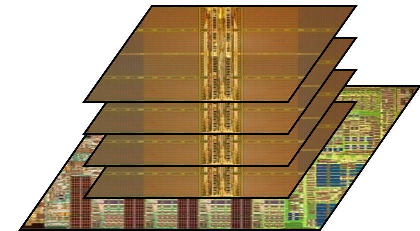
- Assembly : Chip to Wafer or Wafer to wafer



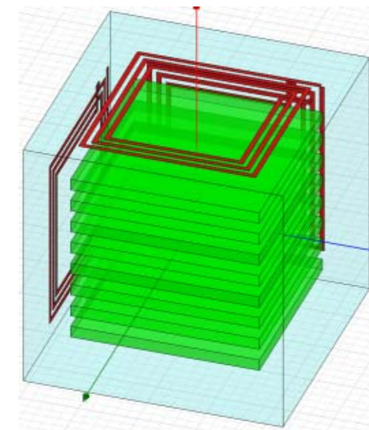
Value Propositions

Fundamentally, 3DIC permits:

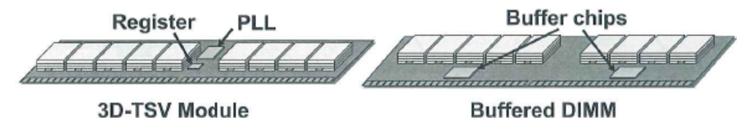
- Shorter wires
  - consuming less power, and costing less
  - The memory interface is the biggest source of large wire bundles



- Heterogeneous integration
  - Each layer is different!
  - Giving fundamental performance and cost advantages, particularly if high interconnectivity is advantageous



- Consolidated “super chips”
  - Reducing packaging overhead
  - Enabling integrated microsystems



The Demand for Memory Bandwidth

Computing

Future microprocessors and off-chip SOP interconnect
Hofstee, H.P.;
Advanced Packaging, IEEE Transactions on [see also Components, Packaging and Manufacturing Technology, Part B: Advanced Packaging, IEEE Transactions on]
Volume 27, Issue 2, May 2004 Page(s):301 - 303

MULTICORE AND REVERSE SCALING

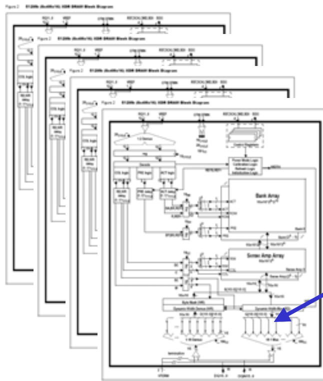
	2004 Baseline	Multi-core Approach	Reverse scaling	Reverse scaling
Frequency	4 GHz	8 GHz	8 GHz	4GHz
No. of Cores	1 Core	4 Cores	16 Cores	16 Cores
Core rel. IPC	1	1	0.5	1
Total Flops	32 GFlops	256 GFlops	512 GFlops	512 GFlops
Supply	1.2V	1.0V	1.0V	1.0V
Power	84W	233W	233W	117-163W
Bandwidth requirement	32GB/s	256GB/s	512GB/s	512GB/s

Similar demands in Networking and Graphics

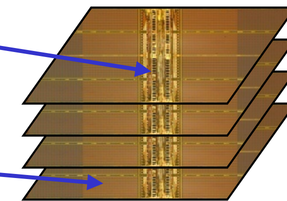
Ideal: 1 TB / 1 TBps memory stack

Memory on Logic

Conventional | **TSV Enabled**



Less Overhead



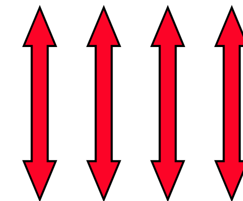
Flexible bank access

x32 to x128

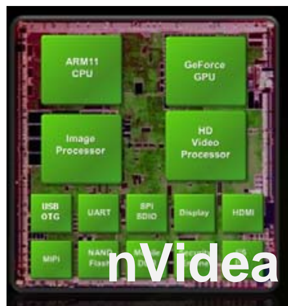
Less interface power

3.2 GHz @ >10 pJ/bit

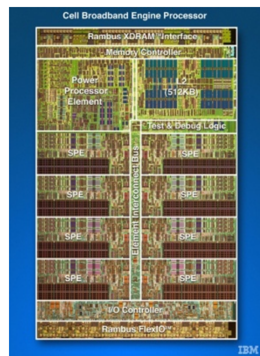
→ 1 GHz @ 0.3 pJ/bit



N x 128 "wide I/O"

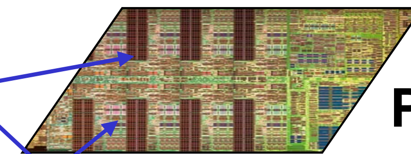


or



Flexible architecture

Short on-chip wires



Processor

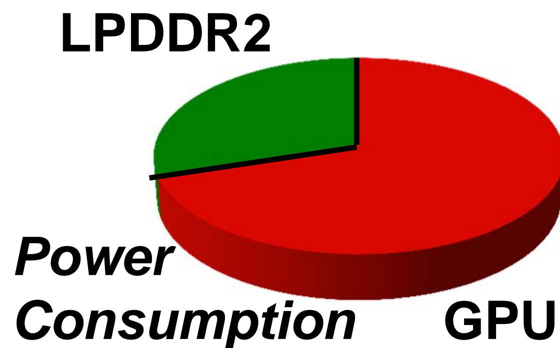
or

Mobile

Mobile Graphics

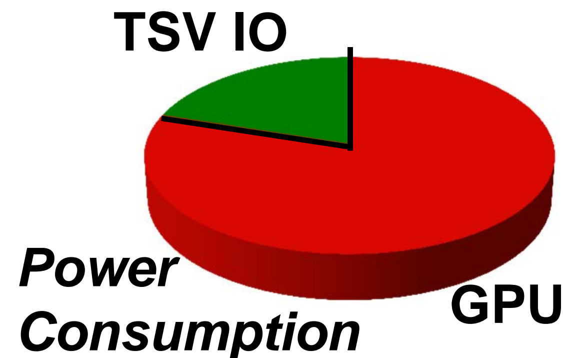
- **Problem**: Want more graphics capacity but total power is constrained
- **Solution:** Trade power in memory interface with power to spend on computation

POP with LPDDR2



532 M triangles/s

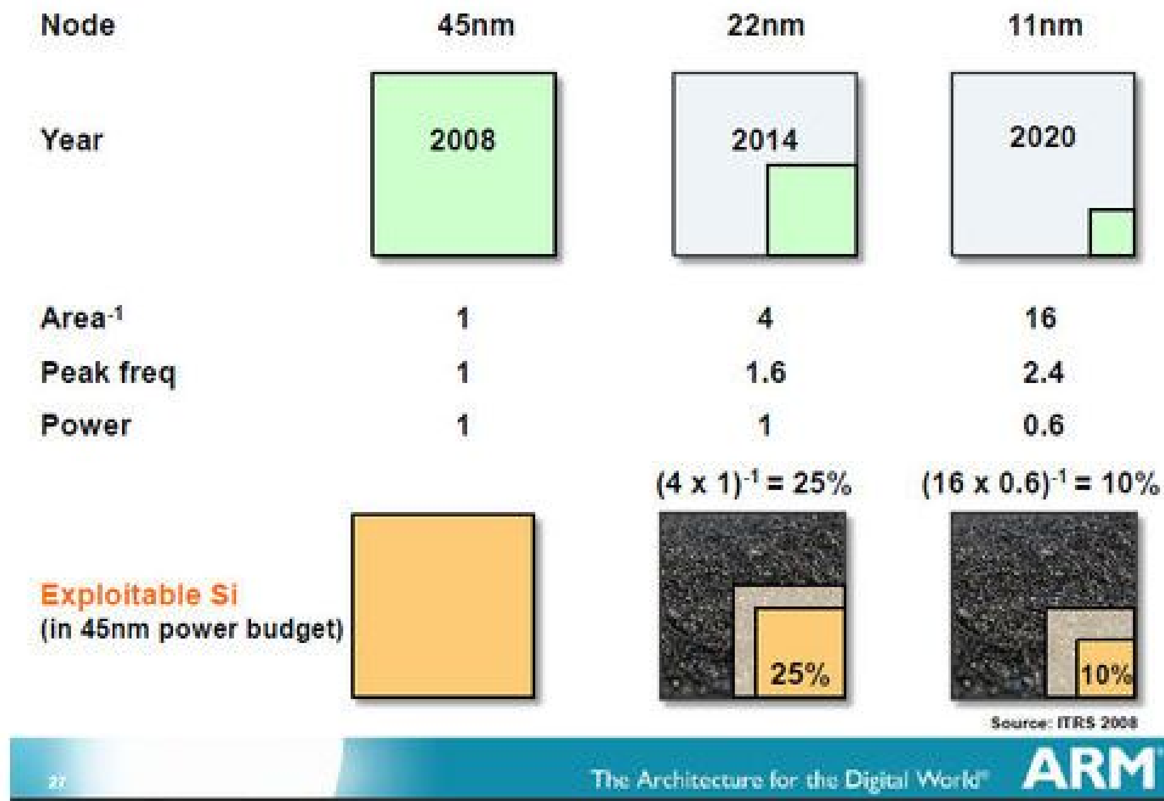
TSV Enabled



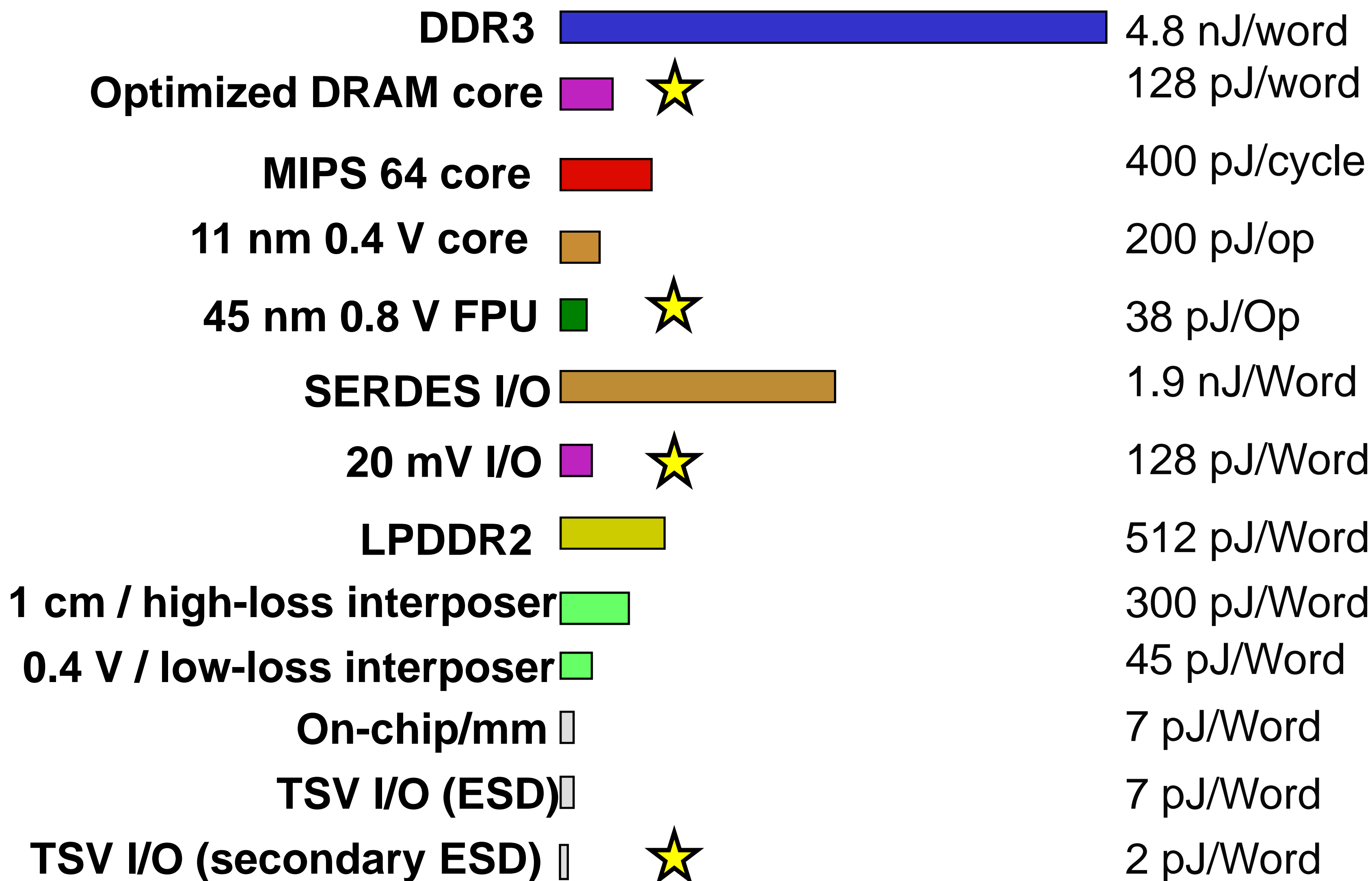
695 M triangles/s

Dark Silicon

- Performance per unit power
  - Systems increasingly limited by power consumption, not number of transistors
  - ➔ "**Dark Silicon**" : Most of the chip will be OFF to meet thermal limits



Energy per Operation

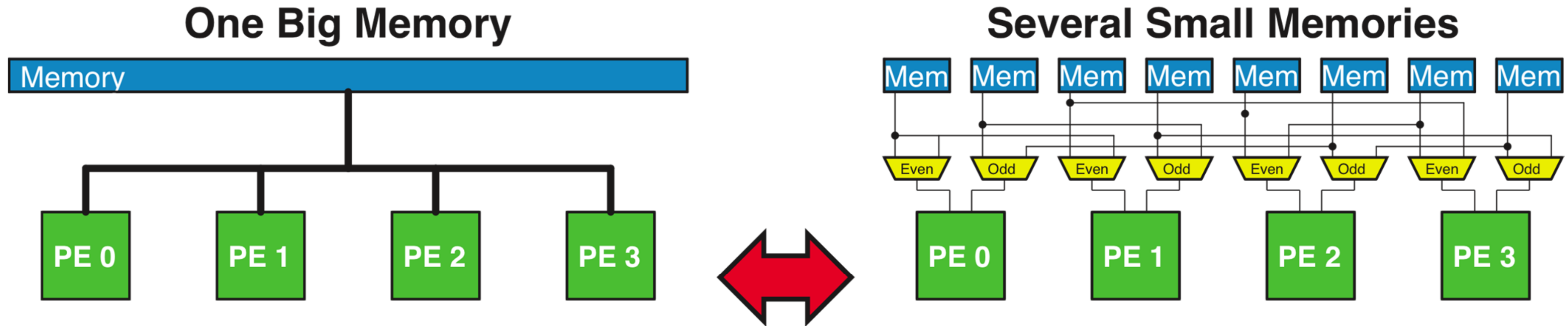


(64 bit words)

Various Sources

Synthetic Aperture Radar Processor

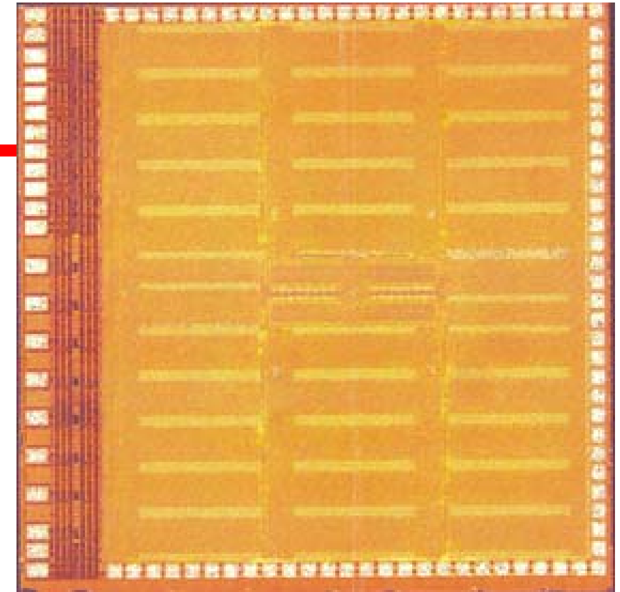
- Built FFT in Lincoln Labs 3D Process



Metric	Undivided	Divided	%
Bandwidth (GBps)	13.4	128.4	+854.9
Energy Per Write(pJ)	14.48	6.142	-57.6
Energy Per Read (pJ)	68.205	26.718	-60.8
Memory Pins (#)	150	2272	+1414.7
Total Area (mm^2)	23.4	26.7	+16.8%

Thor Thorolfsson

3D FFT Floorplan



- All communications is vertical
- Support multiple small memories WITHOUT an interconnect penalty
  - AND Gives 60% memory power savings

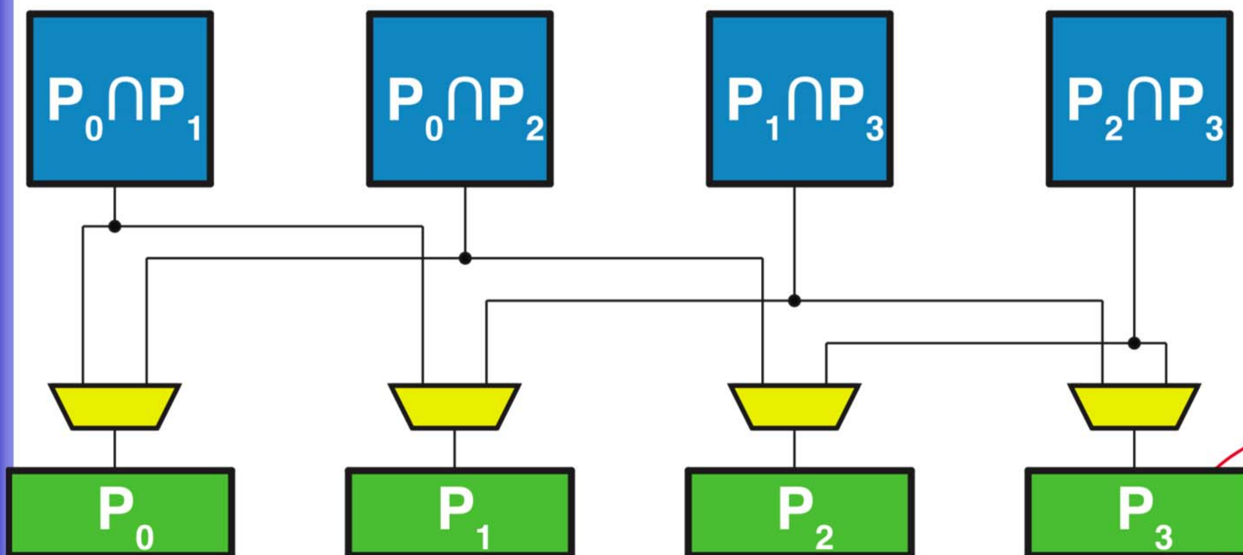
Processing Element 2
Processing Element 3
ROM2 ROM3
ROM0 ROM1
Processing Element 0
Processing Element 1

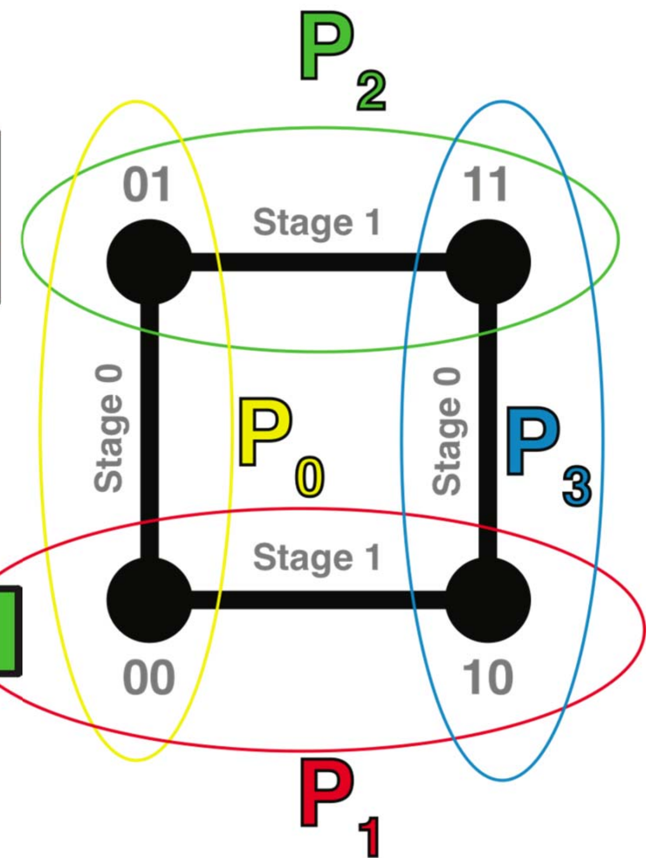
Mem34O Mem18O Mem17O
Mem34E Mem18E Mem17E
Mem33E Mem33O Mem65O
Mem40O Mem36O Mem65E
Mem40E Mem36E Mem66O
Mem66E
Mem129O Controller
Mem129E Mem24O Mem20O
Mem130O Mem24E Mem20E
Mem130E Mem132E Mem132O
Mem136O Mem72O Mem68O
Mem136E Mem72E Mem68E

Processing Element 6
Processing Element 7
ROM6 ROM7
ROM4 ROM5
Processing Element 4
Processing Element 5

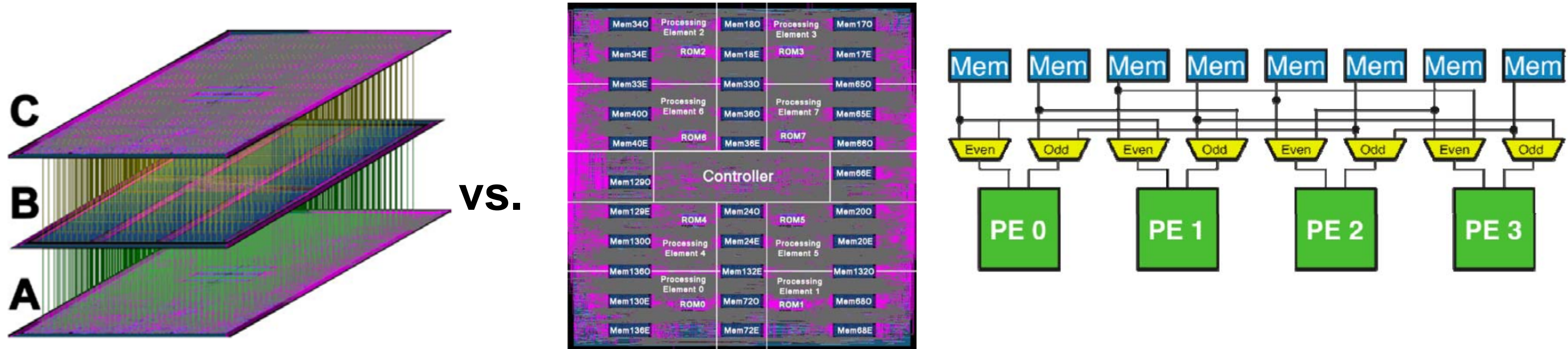
RePartition FFT to Exploit Locality

- Every partition is a PE
- Every unique intersection is a memory





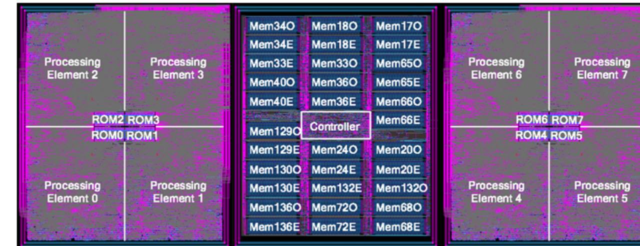
2DIC vs. 3DIC Implementation



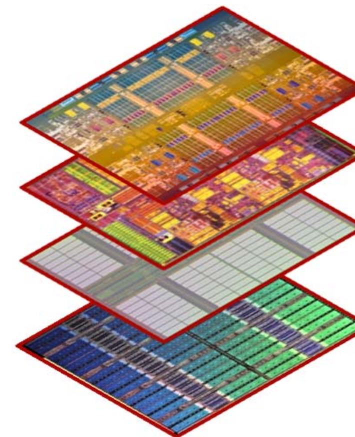
Metric	2D	3D	Change
Total Area (mm^2)	31.36	23.4	-25.3%
Total Wire Length (m)	19.107	8.238	-56.9%
Max Speed (Mhz)	63.7	79.4	+24.6%
Power @ 63.7MHz (mW)	340.0	324.9	-4.4%
FFT Logic Energy (μJ)	3.552	3.366	-5.2%

Increasing the Return

1. 3D specific architectures

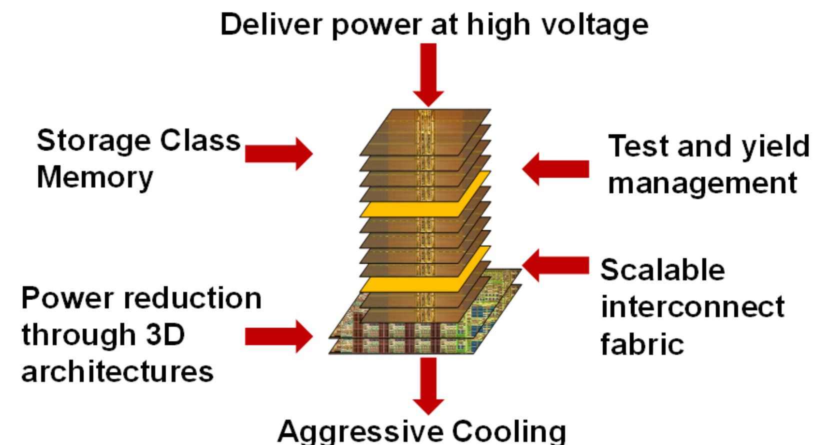


2. Exploiting Heterogeneity



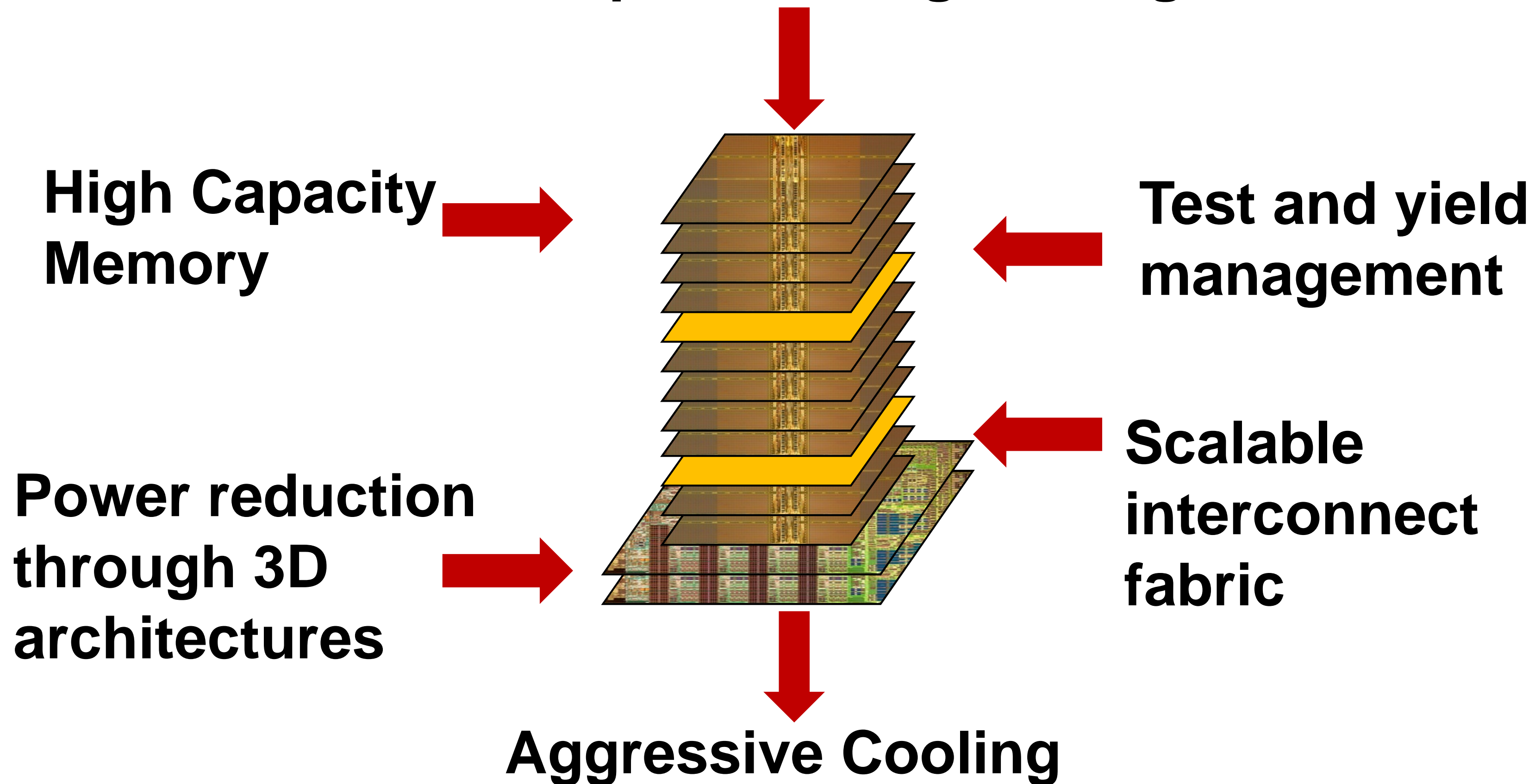
High Performance
Low Power/Accelerat
Specialized RAM
General RAM

3. Ultra 3D Scaling



Extreme Integration

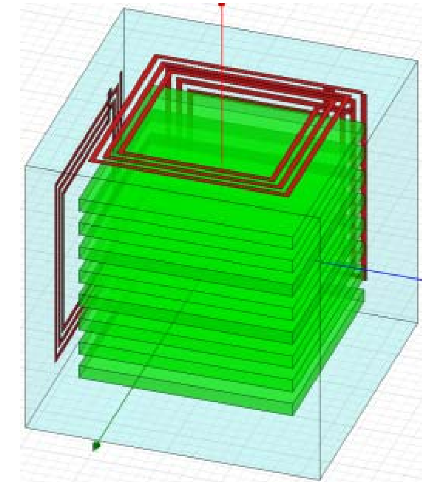
Motivation: Database Servers; High End DSP



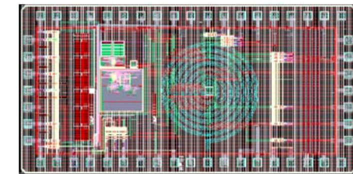
3D Miniaturization

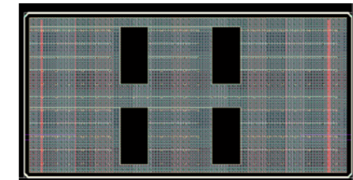
Miniature Sensors

- mm^3 scale - Human Implantable (with Jan Rabaey, UC(B))
- cm^3 scale - Food Safety & Agriculture (with KP Sandeep, NCSU)



- Problems:
  - Power harvesting @ any angle (mm-scale)
  - Local power management (cm scale)

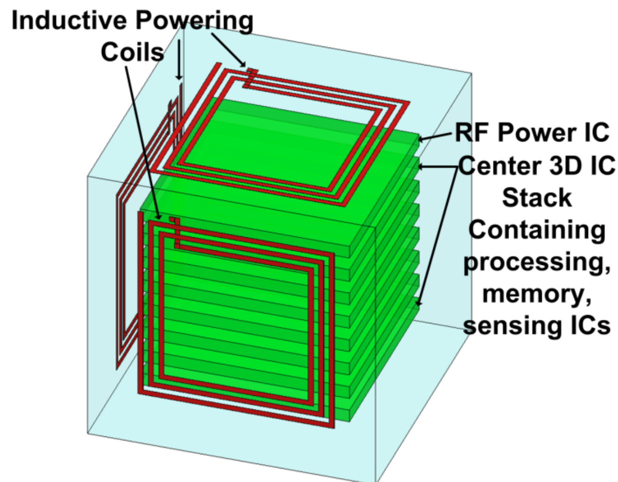


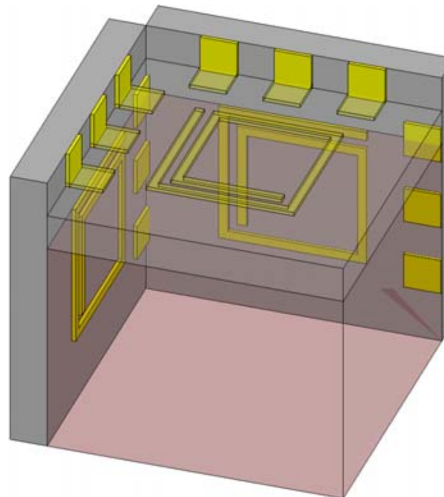


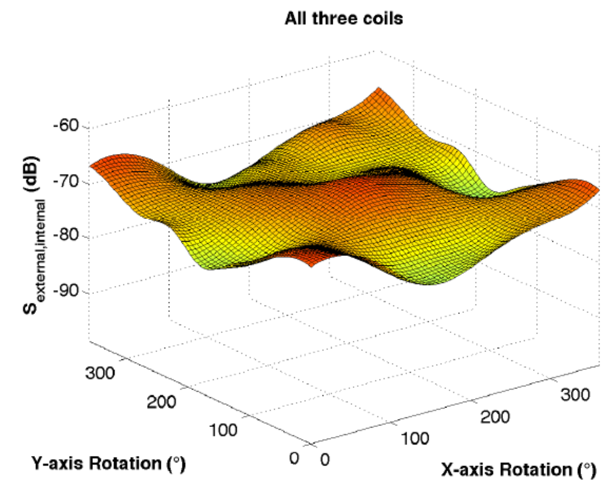
Peter Gadfort, Akalu Lentiro, Steve Lipa

"True" 3D Integration

- Orientation of mm-scale sensor will be random
  - Building antenna "through" 3DIC chip stack on edge will be very lossy
  - Need power harvesting on all 3 sides
  - Developed packaging integration flow to achieve this







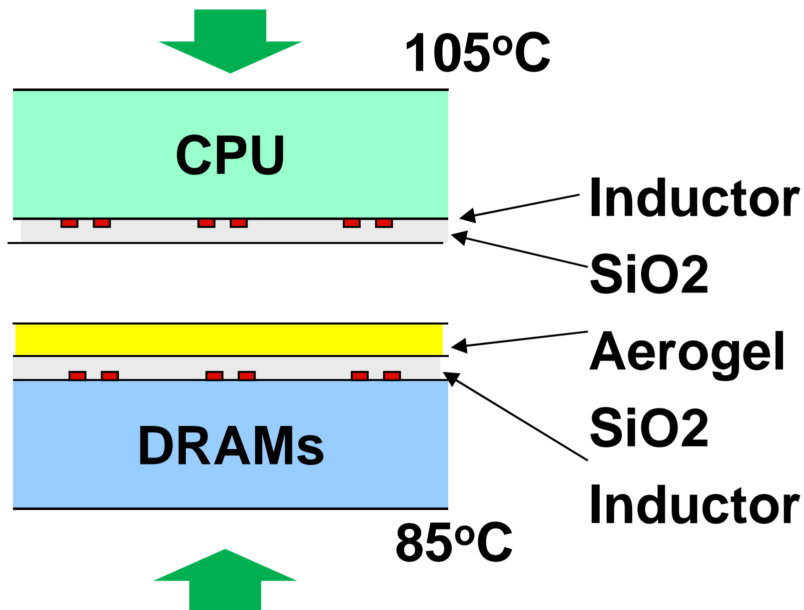
Peter Gadfort

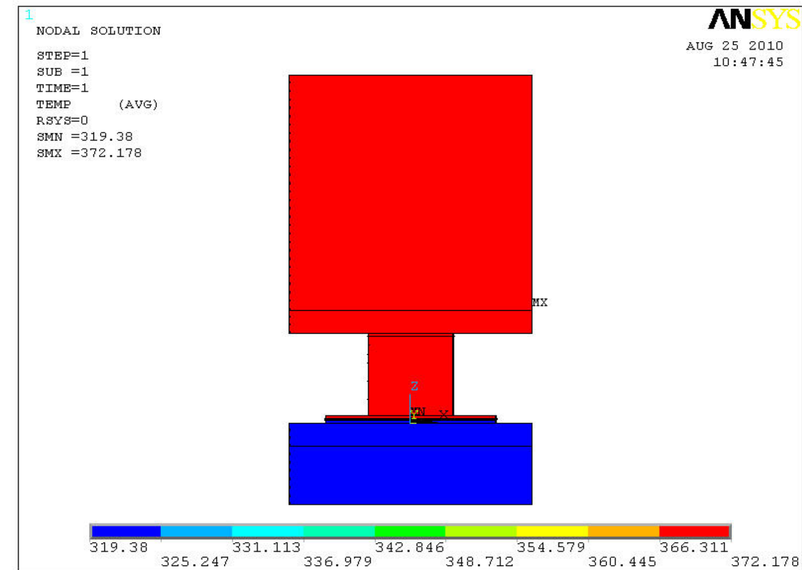
Mid-term Barriers to Deployment

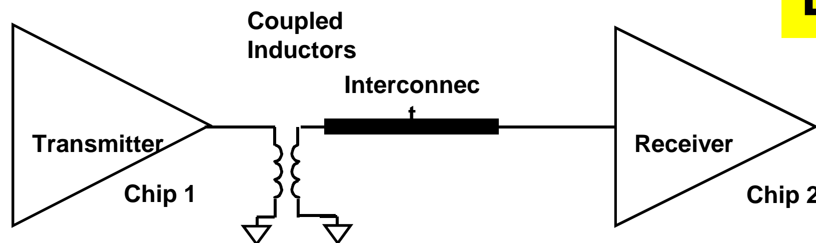
Barrier	Solutions
Thermal	Early System Codesign of floorplan and thermal evaluation
	DRAM thermal isolation
Test	Specialized test port & test flow
Codesign	"Pathfinding" in SystemC
	CAD Interchange Standards
Cost & Yield	Supporting low manufacturing cost through design

Technologies for Thermal Isolation

- Introduce thermal isolation material between CPU and DRAM
- Use inductive coupling for communications





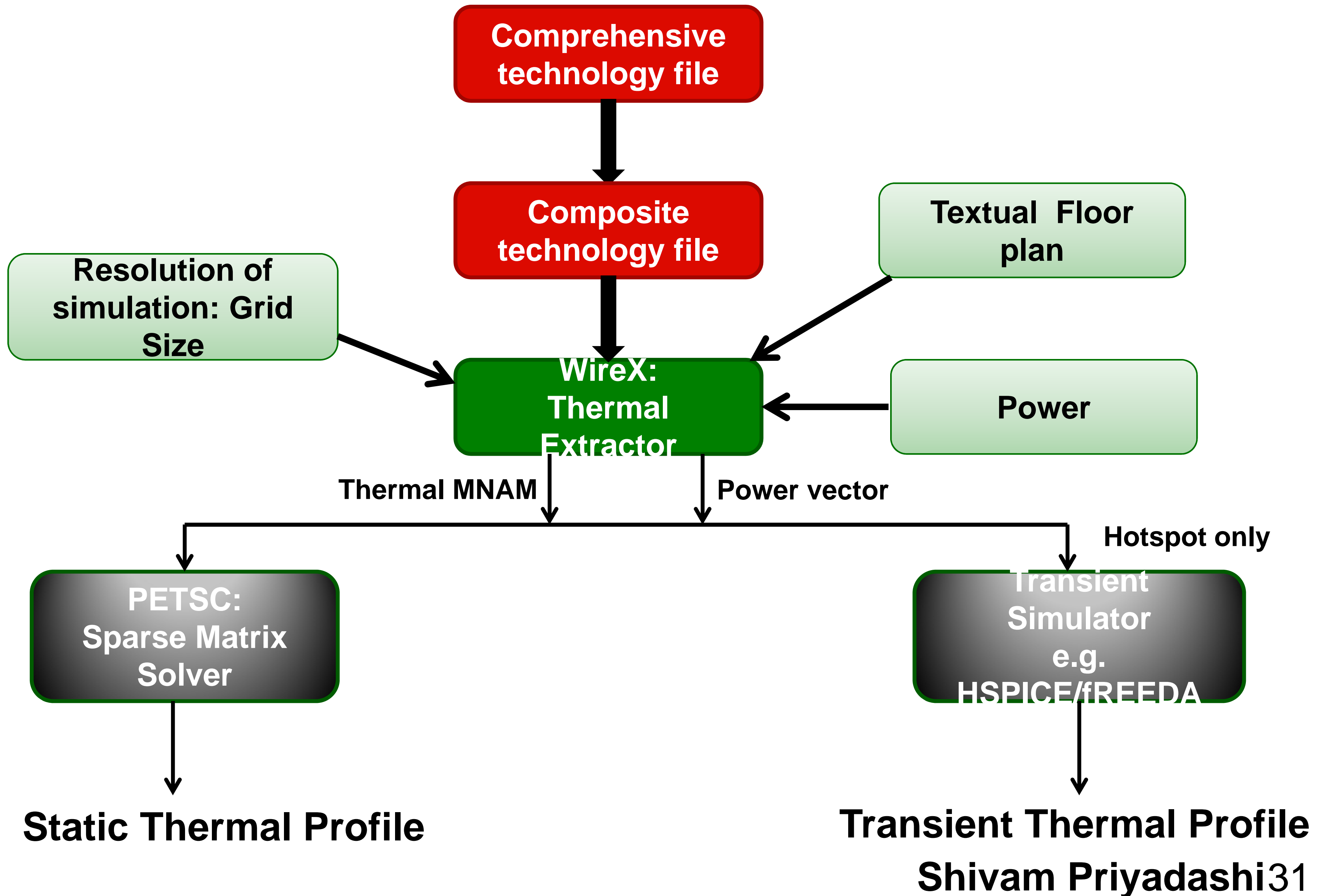


DRAM surface 50°C cooler than SOC

vs. Only 9°C cooler with direct attachment

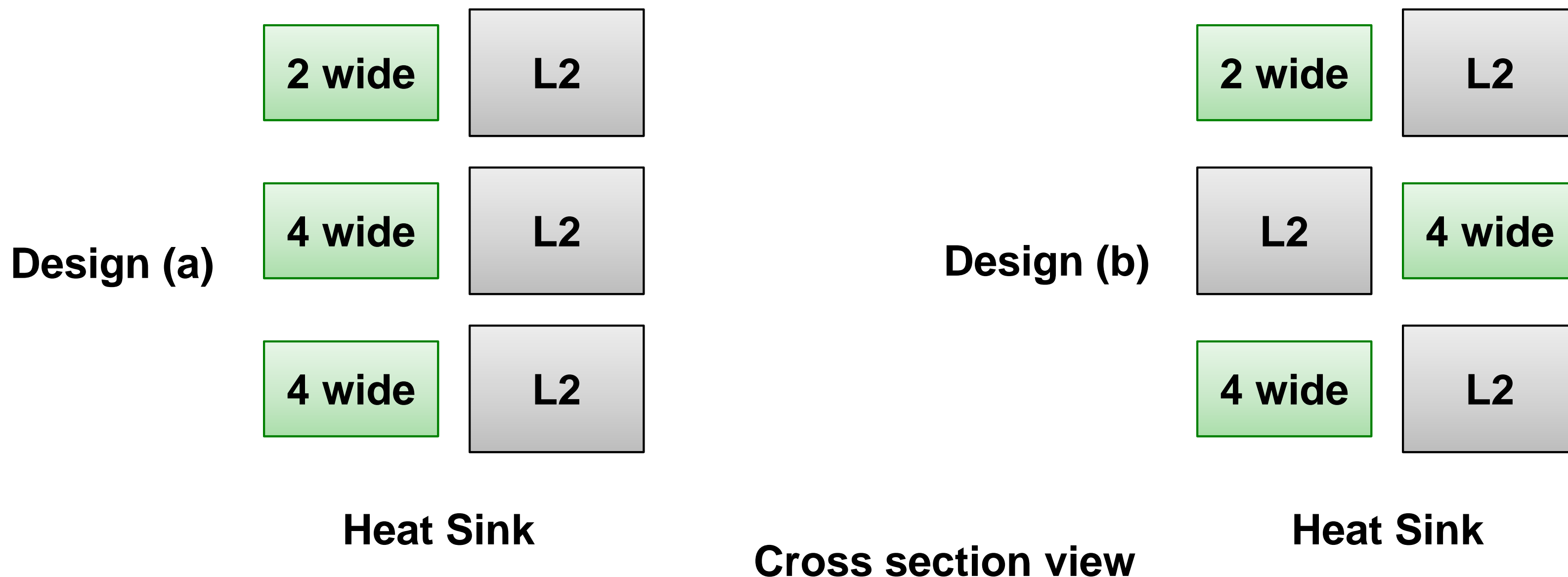
Ming Li, John Wilson, Neil DiSpigna

Thermal and Physical Flow:



Pathfinder 3D:

- Goals:
  - Electronic System Level (ESL) codesign for fast investigation of performance, logic, power delivery, and thermal tradeoffs
  - Focus to date: Thermal/speed tradeoffs – static and transient
- Test case : Stacking of Heterogeneous Cores



Long-term Barriers to Deployment

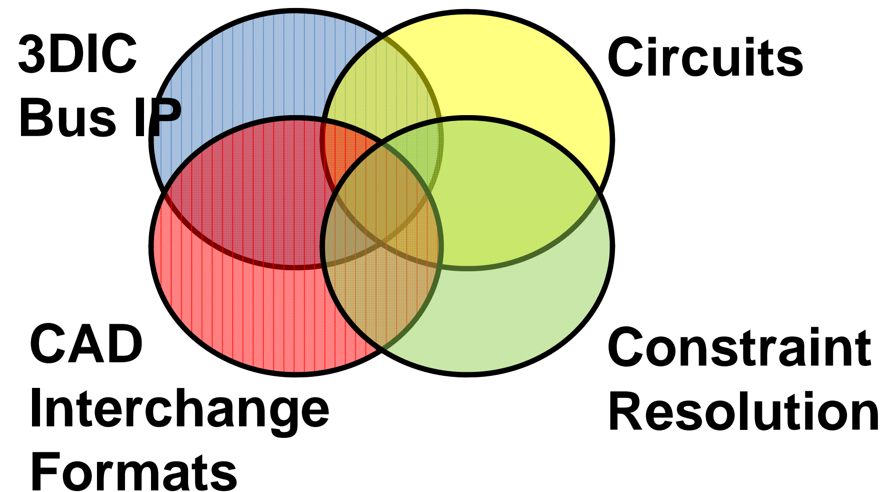
Barrier	Solutions
Thermal & Power Deliver	3D specific temperature management
	New structures and architectures for power delivery
Test & Yield management	Modular, scalable test and repair
Co-implementation	Support for Modularity and Scalability
Cost & Yield	Supporting low manufacturing cost through design

3D Specific Interface IP

Proposal:

Open Source IP for 3D and 2.5D interfaces

An interface specification that supports signaling, timing, power delivery, and thermal control within a 3D chip-stack, 2.5D (interposer) structure and SIP solutions



(Proc. 3DIC 2011)

Conclusions

- Three dimensional integration offers potential to
  - Deliver memory bandwidth power-effectively;
  - Improve system power efficiency through 3D optimized codesign
  - Enable new products through aggressive Heterogeneous Integration
- Main challenges in 3D integration (from design perspective)
  - Effective early codesign to realize these advantages in workable solutions
  - Managing cost and yield, including test and test escape
  - Managing thermal, power and signal integrity while achieving performance goals
  - Scaling and interface scaling

Acknowledgements

Faculty: William Rhett Davis, Michael B. Steer, Eric Rotenberg,
Professionals: Steven Lipa, Neil DiSpigna,
Students: Hua Hao, Samson Melamed, Peter Gadfort, Akalu Lentiro,
Shivam Priyadarshi, Christopher Mineo, Julie Oh, Won Ha Choi,
Zhou Yang, Ambirish Sule, Gary Charles, Thor Thorolfsson,
Department of Electrical and Computer Engineering
NC State University