NC STATE UNIVERSITY

Computing in 3D

Paul Franzon

North Carolina State University

Raleigh, NC

paulf@ncsu.edu

919.515.7351

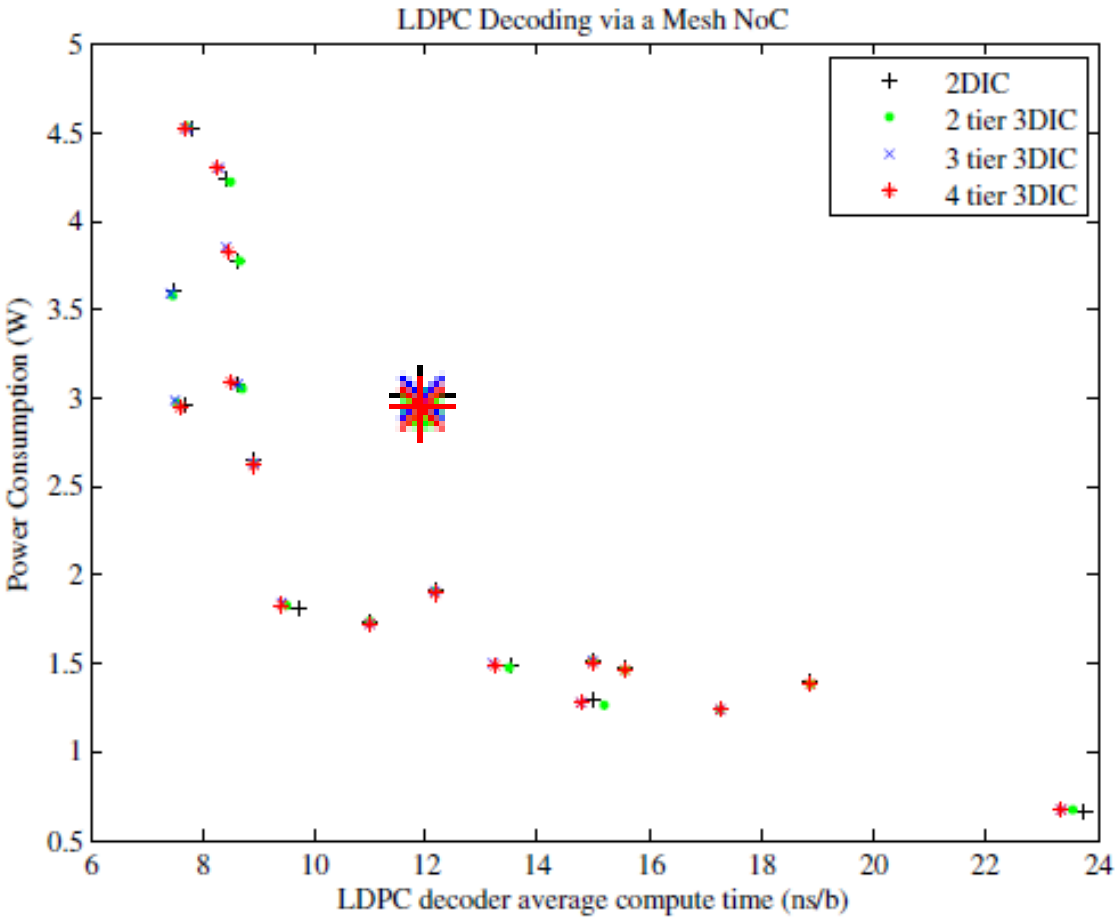
http://www.ece.ncsu.edu/erl/faculty/paulf.html


THIS SUMMER 3D SHOWS ITS TEETH
PIRANHA 3D
AUGUST 20
atmosphere
real 3D
www.PIRANHA-3D.com
DIMENSION

IF
YOU CAN'T
MAKE IT
GOOD
MAKE IT
3D

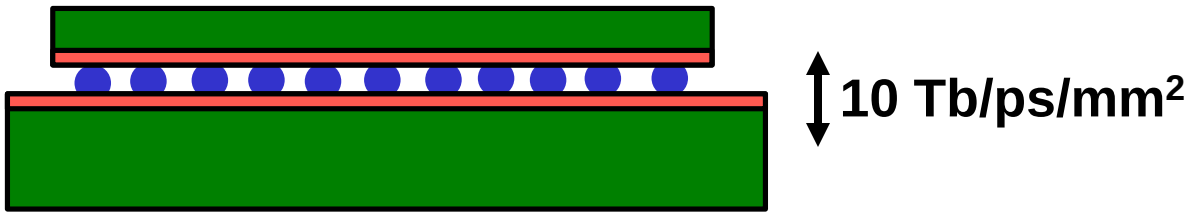
Early days in 3D computing

- Taking a conventional design, and making a 3D version often does not give interesting results



Instead

- Seek a 3D-specific architecture
  - Do something you cant do in 2D!
- Our typical goals:
  - Improve performance/power by > 25%
    - About a node equivalent
  - Improve a key bandwidth, or a highly dominant critical path
  - Often specifically exploiting high density face to face interconnect



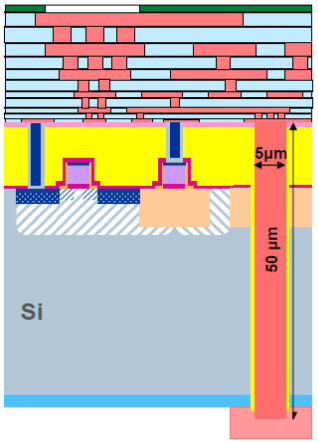
- Now: Memory interfaces
- Next: Logic on Logic & Logic on Memory

Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

3D Technology Set

- 3DIC with TSVs



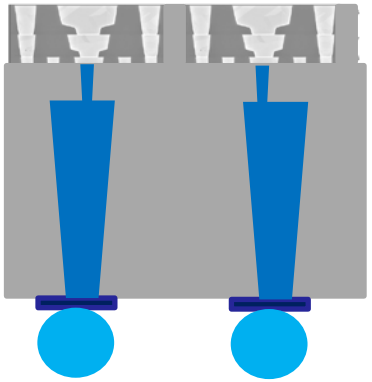
25 to 50 µm

40 µm bump pitch

Tomorrow:

- **DRAM 20 µm longTSV**
- **Logic: 5 µm long TSV**
- **1- 2 µm pitch or below**
- **25 µm bump pitch**

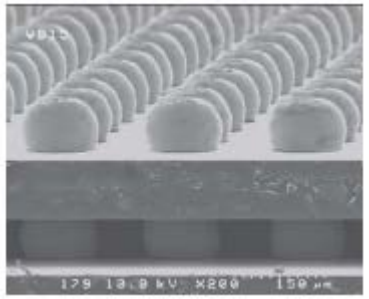
Cvia: 40 fF today → 2 fF tomorrow



- Interposers:
  - 50 µm thick 100 µm TSV pitch
  - Today: 10 µm wire feaures
  - Tomorrow: sub 1 µm wire features

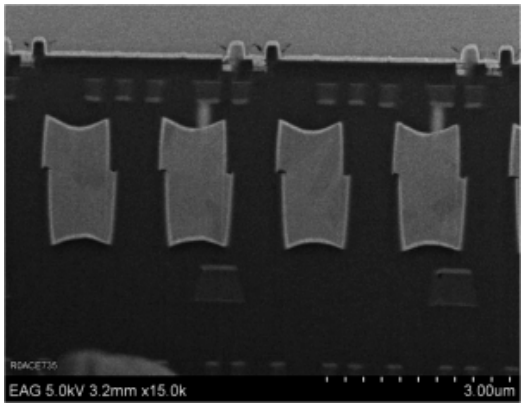
Attachment technologies

- Solder micobumps
  - Today typically 40 μm pitch
  - 25 μm pitch demonstrated
  - Potential for 5 μm pitch



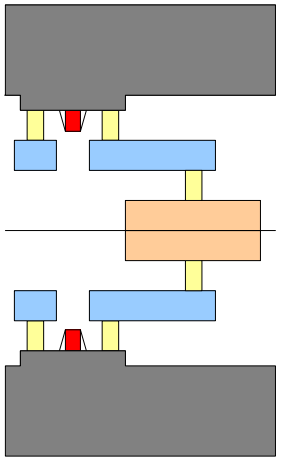
IBM

- Copper-copper
  - @ high temperature (> 400 C)
  - @ low temperature (Ziptronix DBI)
  - Typical 2 – 5 μm pitch
  - Potential for sub-1 μm pitch

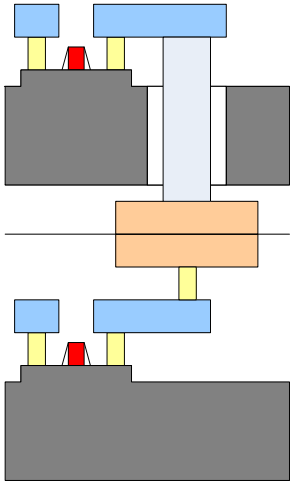


Ziptronix

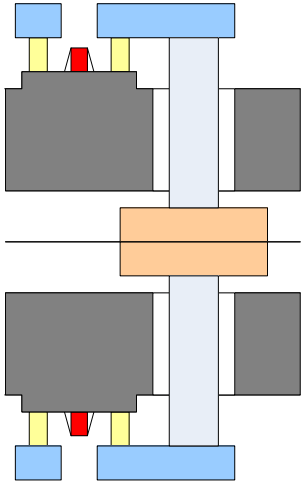
Transistor/TSV Integration Options



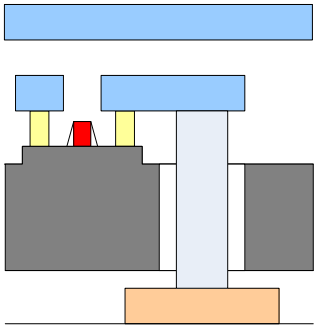
Face-to-Face



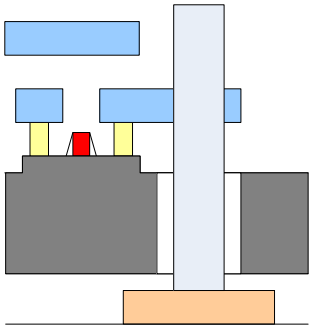
Face-to-Back



Back-to-Back



Via-First/
Via-Middle

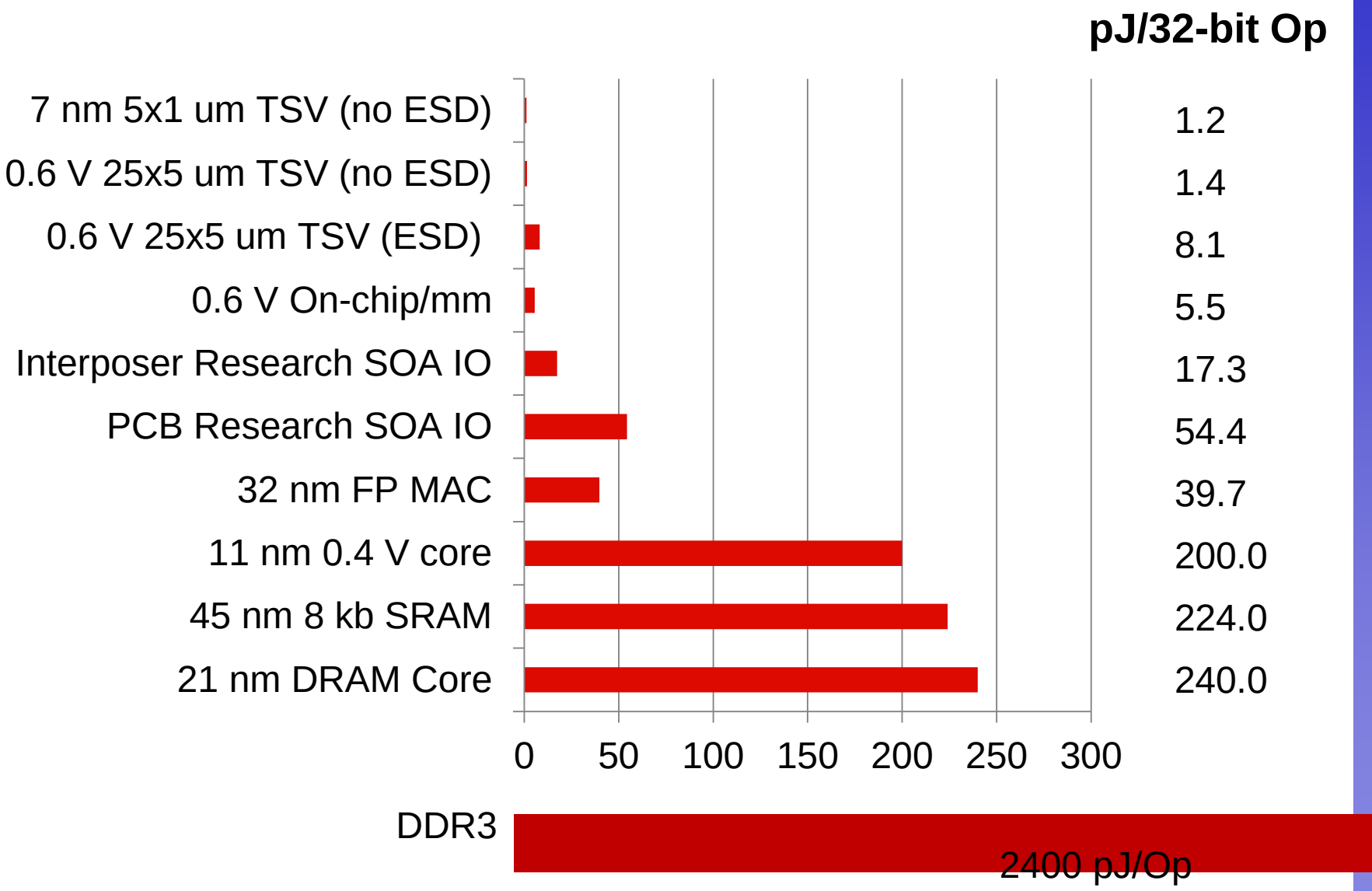


Via-Last

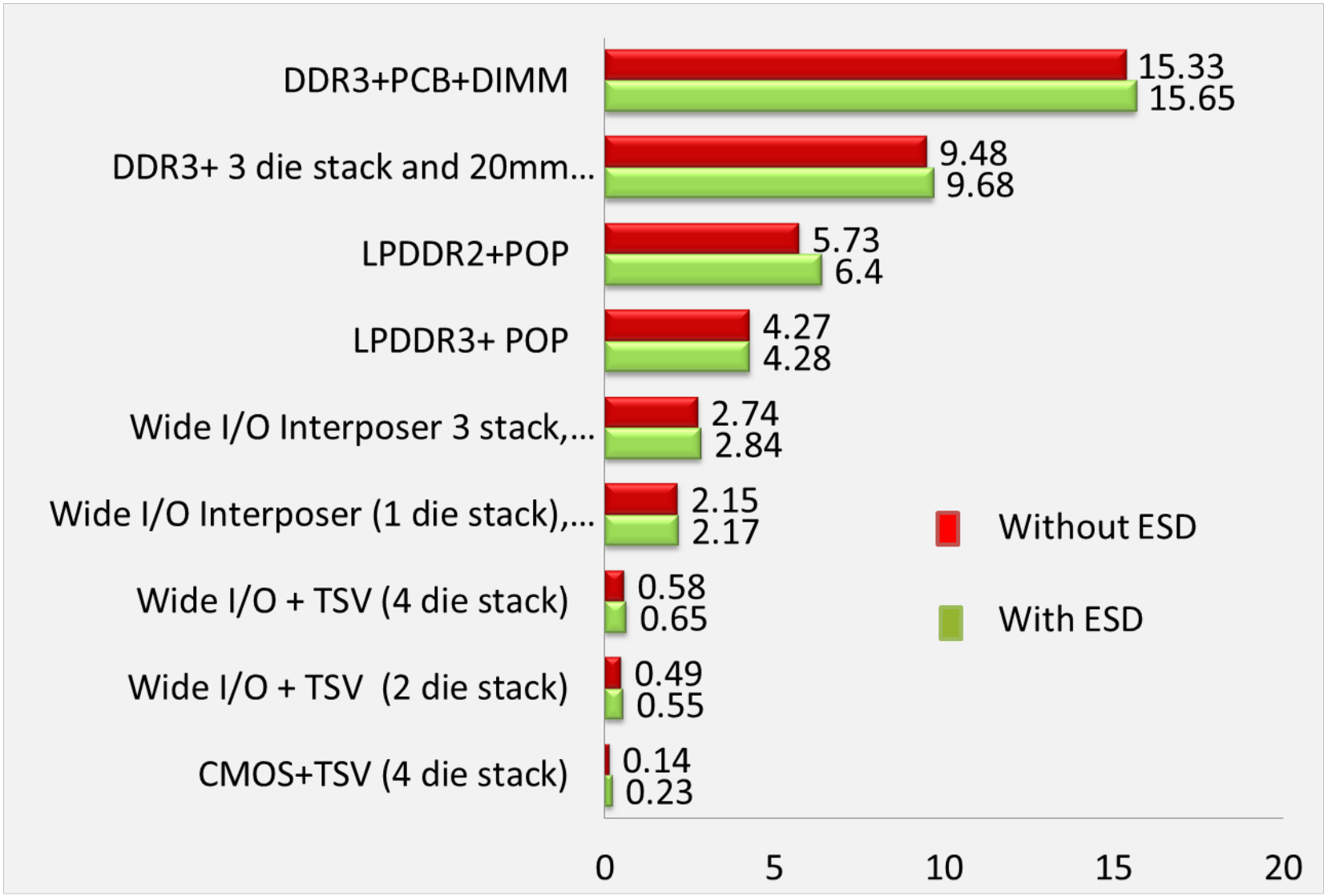
Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

Compute: Energy / 32-bit Operation



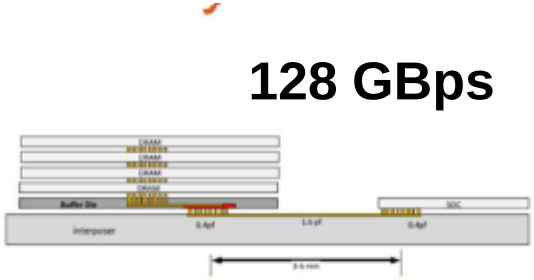
Interface Energy / bit

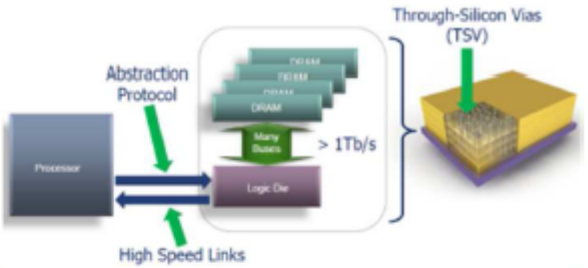


Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

HBM and HMC

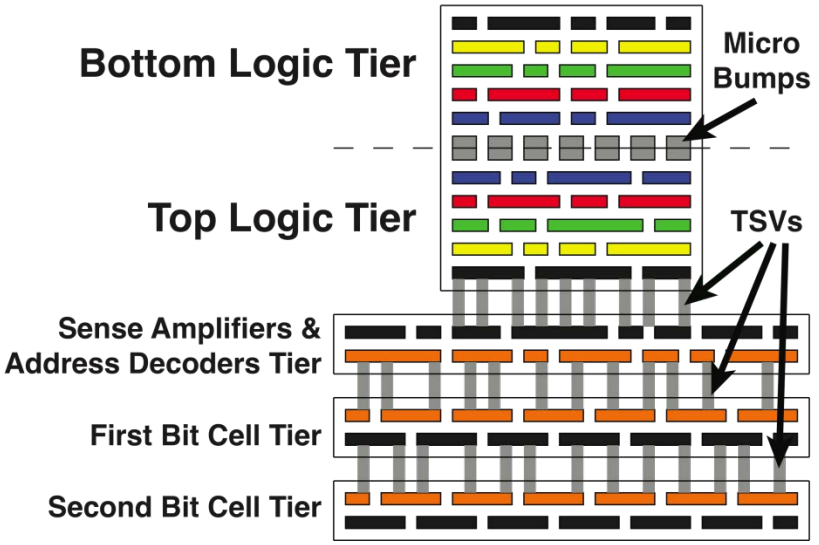




	HBM (High Bandwidth Memory)	HMC (Hybrid Memory Cube)
Standardization	JEDEC	HMC Consortium
Max BW	256GB/s	Scalable, dep. on serial link
Memory die stack	4 die TSV	4 die TSV
Max Density	scalable	
# of IO	~1600 incl. supply	SR 276 (4link)
Controller integration	Only memory Phy	Integrated controller incl. SerDes
Implementation	**Requires silicon interposer**	Works on PCB or organic MCM
Interconnect power	lowest	low
Package	KGD WLCSP (96x55um bump pitch)	Tested BGA with custom pitch
HVM Maturity	2015	2014
Application	HPC, Networking, Server	

Tezzaron “Dis-integrated RAM”

- Mixed technology concept
  - DRAM arrays in low-leakage DRAM technology
  - Peripheral circuits in high-performance logic process
  - Bit and word lines fed vertically at array edge
  - No repair or test prior to assembly
  - BIST and CAM based remapping in logic layer
- Claimed results
  - Reduced overall cost/bit
    - Two metals only in DRAM tiers
    - Effective ~ 60-70% fill factor (?)
  - Faster timing on interfaces

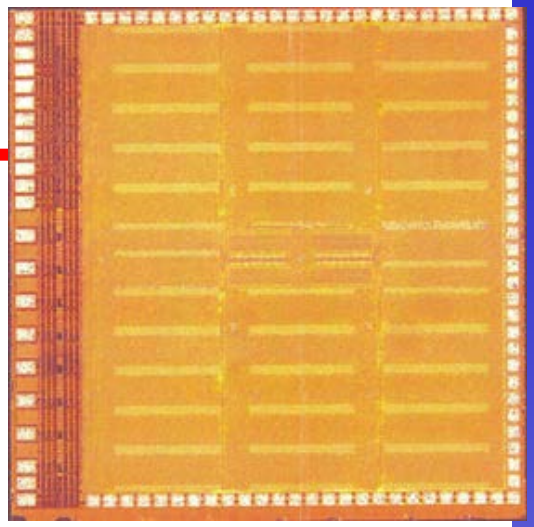


Configuration	8 x 128-bit ports 90 nm DRAM on 130 nm logic
Density	1 Gb/layer of DRAM
Burst access in page/port	1 Gword/s (128 Gbps)

Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

Modular Partitioning

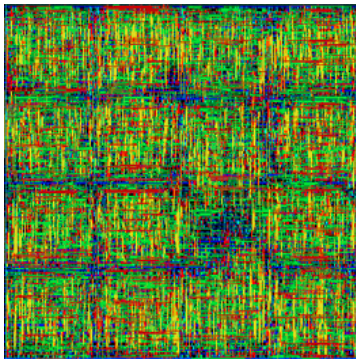


3D FFT Engine in Lincoln Labs SOI

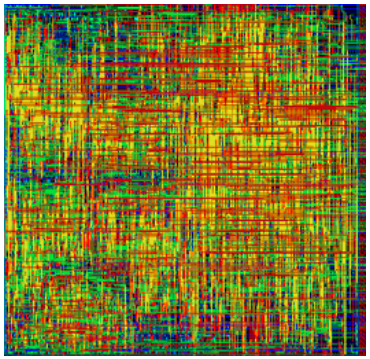
- 60% energy per op savings in memory
- 9% energy per op savings in logic
- 25% less silicon than 2DIC version

Circuit Partitioning

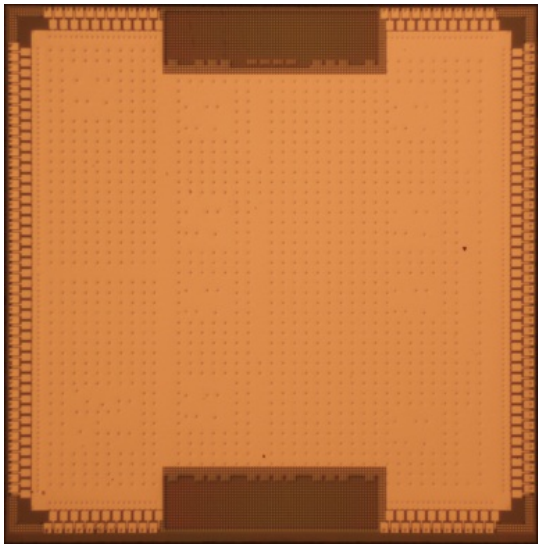
- Complete Synthetic Aperture Radar processor
  - 10.3 mW/GFLOPS
  - 2 layer 3D logic
- All Flip-flops on bottom partition
  - Removes need for 3D clock router
- HMETIS partitioning used to drive 3D placement



Logic only

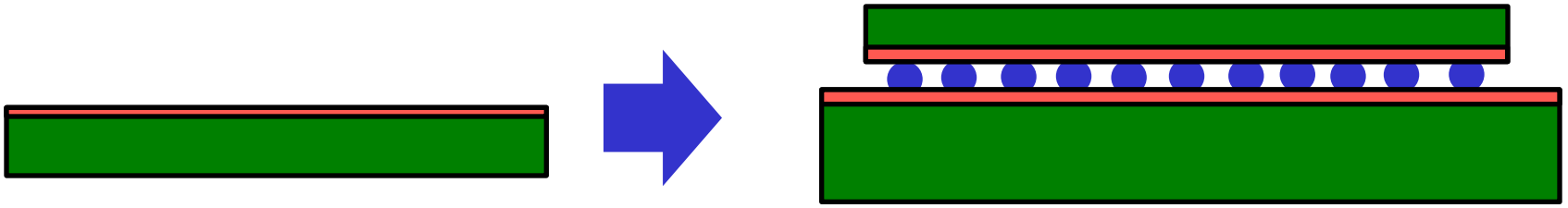


Logic, clocks, flip-flops



Cell level partitioning

- Relying on wire-length reduction



2D Design

0.13 μm Cell Placement split across
6.6 μm face-to-face bump structure

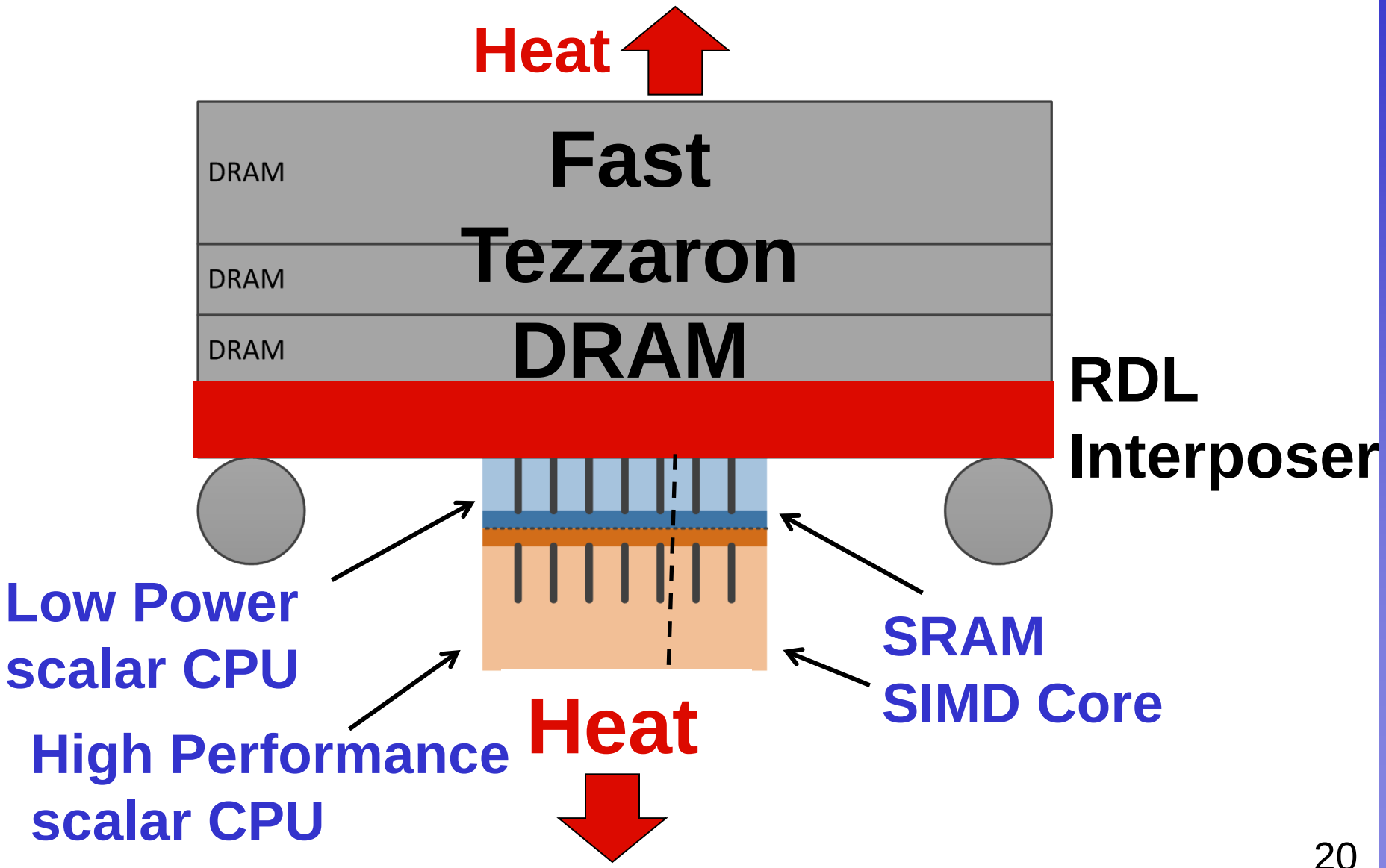
	Total Wire Length (% Change)	Max Frequency (% Change)	Parasitic Power (% Ch	T
PE 3D Seq.	-17.1%	+7.1%		
PE 3D Sim.	-17.7%			-7.7%
PE 3D True	-21.0%			-12.9%
AES 3D Seq.	-8.0		-19.6%	-2.6%
MIMO 3D Seq.	+216.	7.1%	-34.9%	-5.1%

18% - 35% improvement in Power.Delay
(21% for SAR)

Logic-on-Logic 3D Integration and Placement

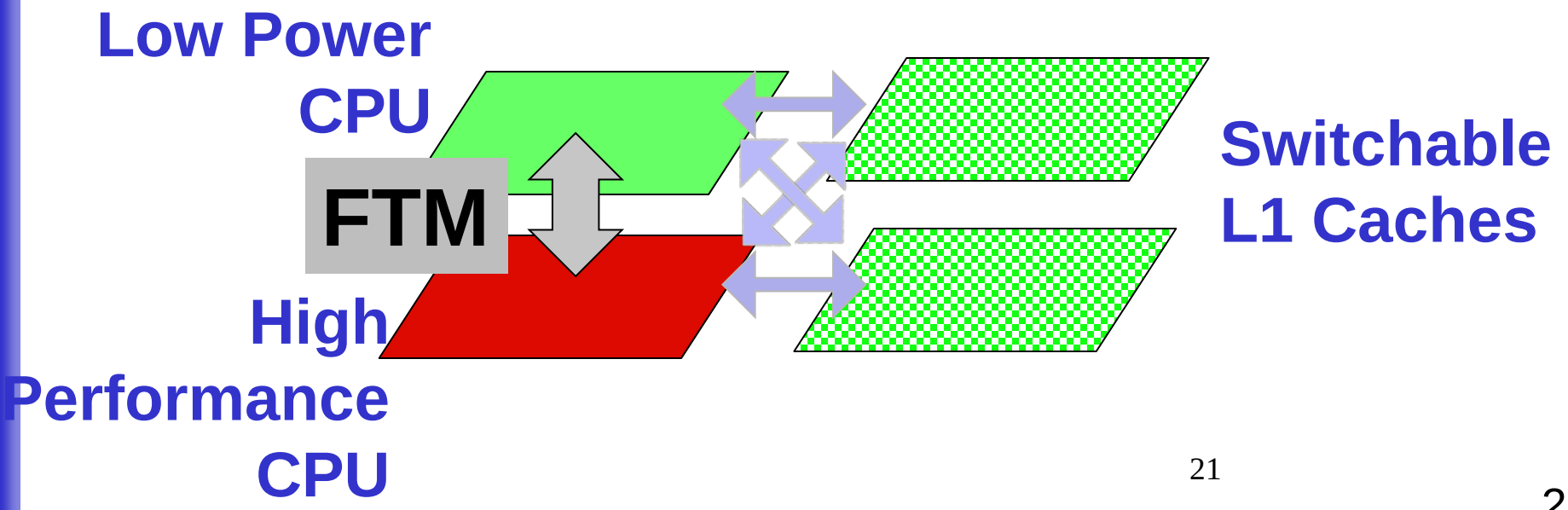
Thorlindur Thorolfsson□ , Guojie Luo† , Jason Cong† and Paul D. Franzon□
□ Department of Electrical & Computer Engineering, North Carolina State University, Raleigh, NC 27695
† Computer Science Department, University of California, Los Angeles, CA 90095
Email: gluo@cs.ucla.edu and cong@cs.ucla.edu

Heterogeneous Computing



Two CPU Stack

- Standardized PnP Fast Thread Migration (FTM) bus permits fast computing thread transfer between HP and LP cores
- CPUs can be designed separately at different times with independent clocks



21

Exploiting Fast Thread Transfer

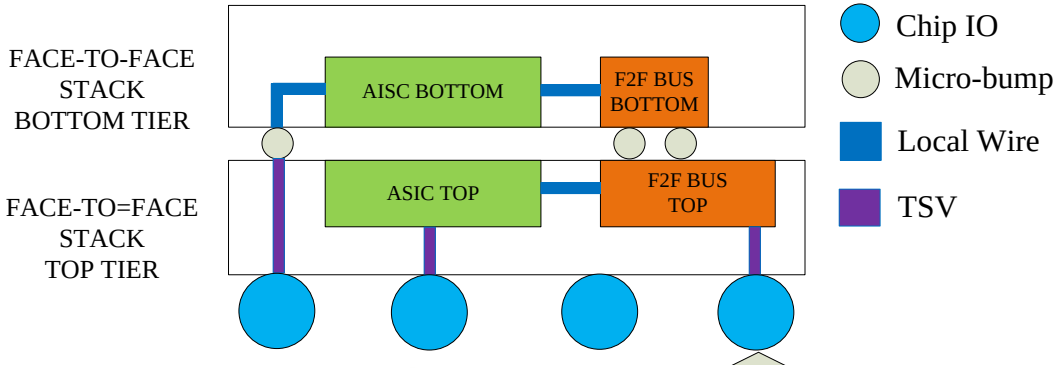
- Sensors determine when to transfer thread
- Threads swapped in around 50 CPU cycles
- Complete SPECint benchmark suite

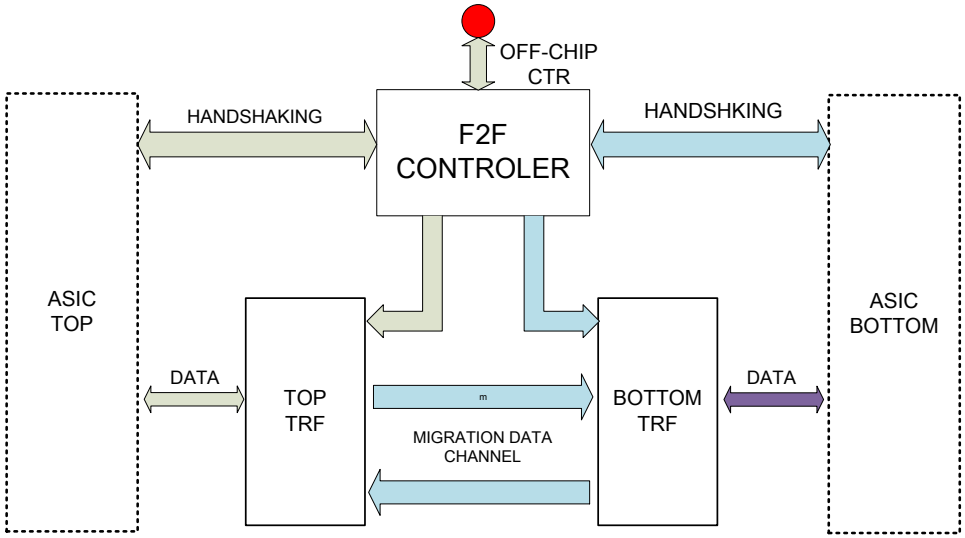
Comparison with Running Data in 2-issue CPU alone:

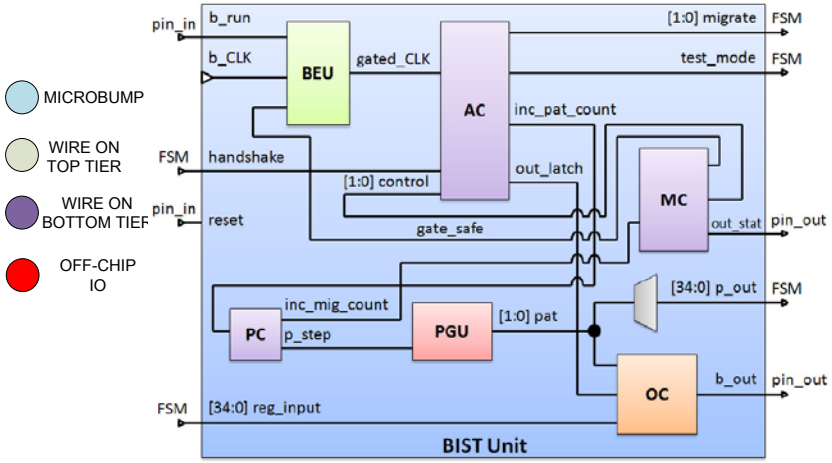
	Energy / op	Performance
1-issue CPU alone	28% savings	39% reduction
Two CPU stack with FTT	27% savings	7% reduction

Thread Transfer Bus

- A fast and simultaneous bi-directional bus specially tuned for vertical high-throughout transfer with built in self test

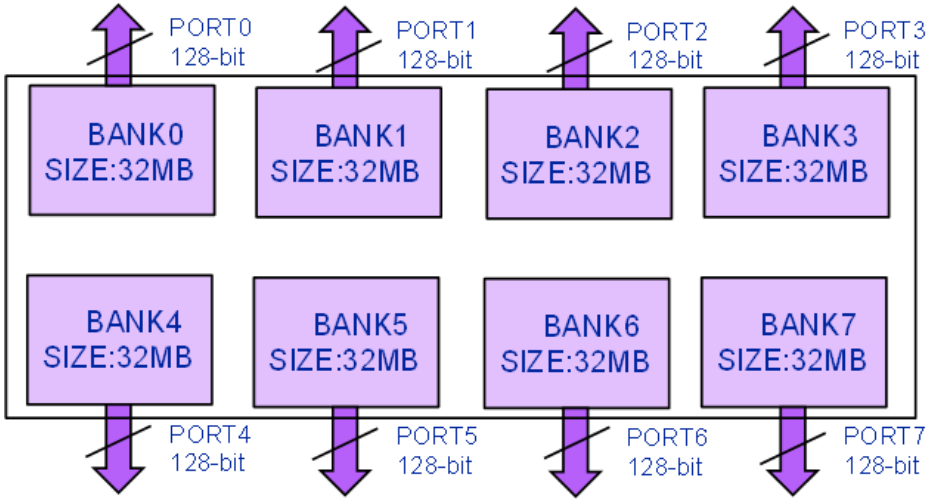






Exploiting Fast DRAM

- Gives speed of L2 cache with better capacity than L3
- Exploit fast RAS-RAS cycle of Tezzaron DRAM

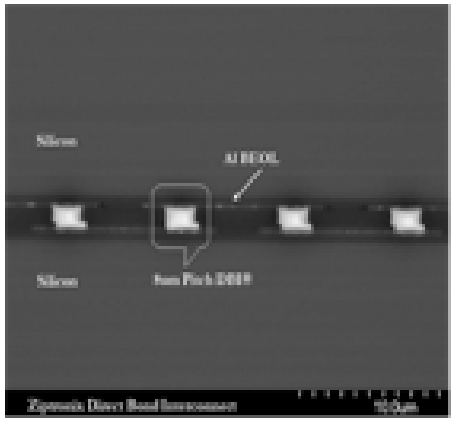


Option	Performance	Power (W)
4MB SRAM cache	1x	2.4 W
240 MB DRAM	1.89 x	0.53 W

Brings 16 core system power down by 15%

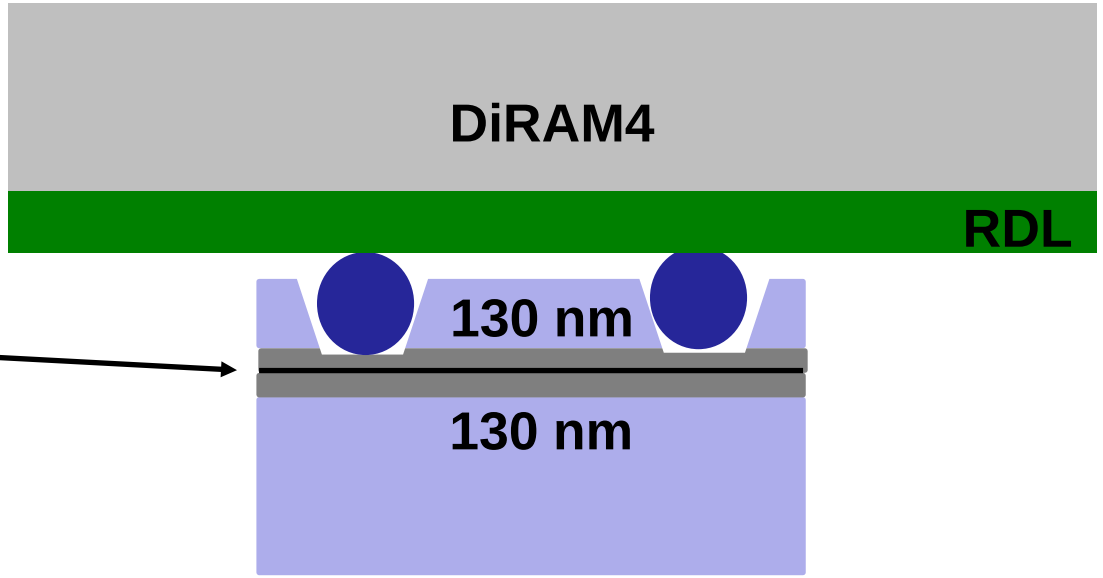
Implementation

- 130 nm 2D Version fabbed and validated
- 130 nm 3D Version to be taped out in November 2015
  - Ziptronix DBI process



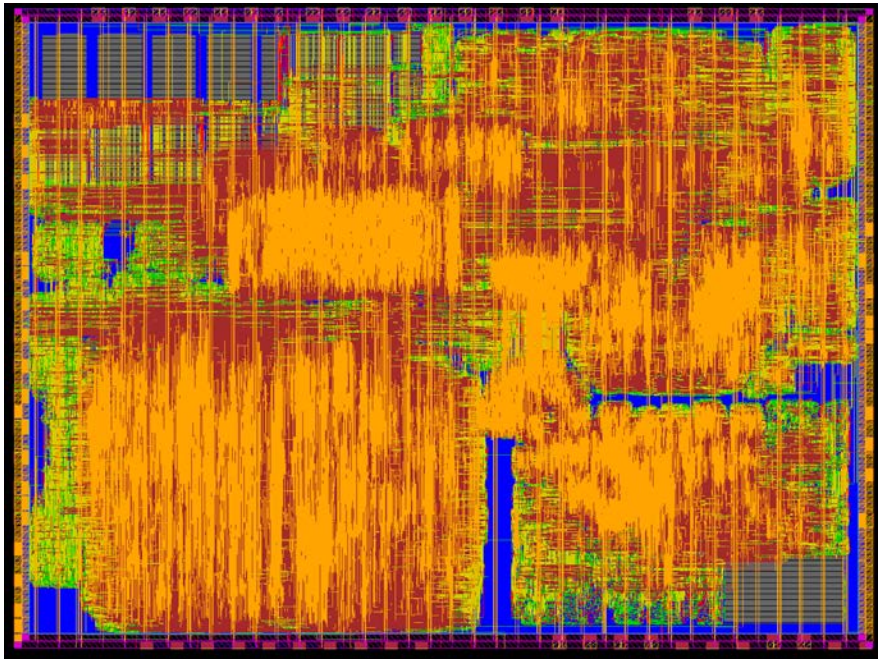
8 um

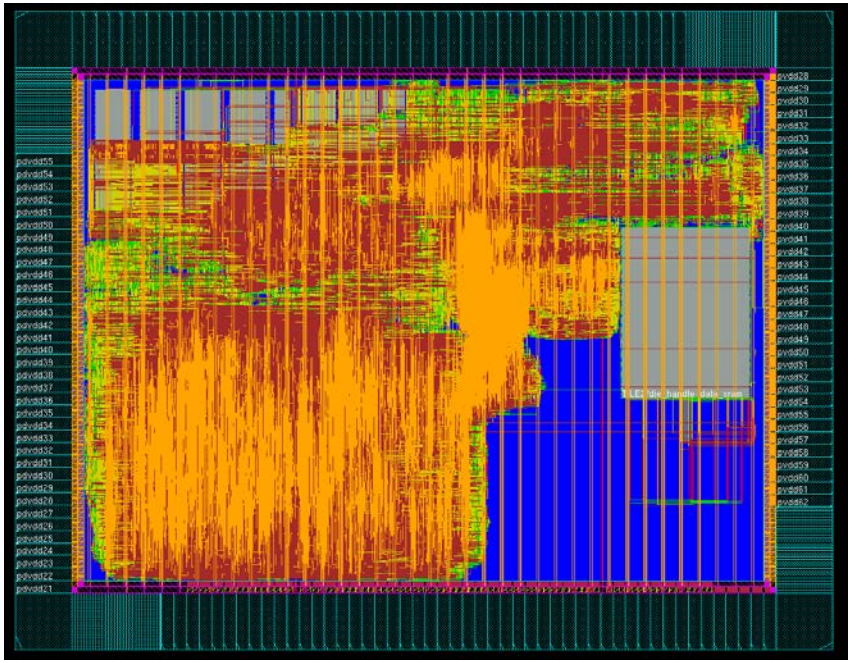
DBI



November Tapeout – 2 chip stack

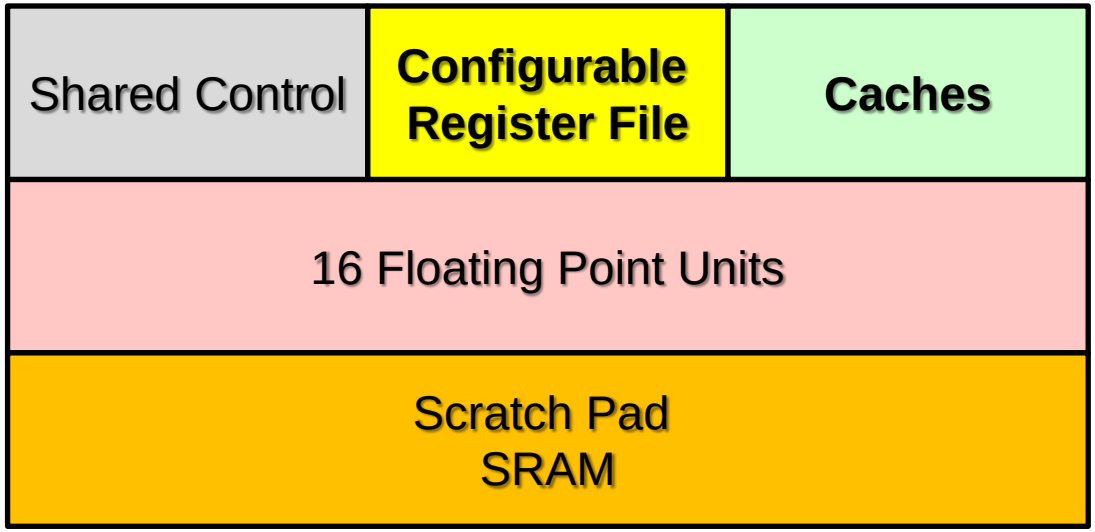
- Heterogeneous Processor stack and SIMD on SRAM stack



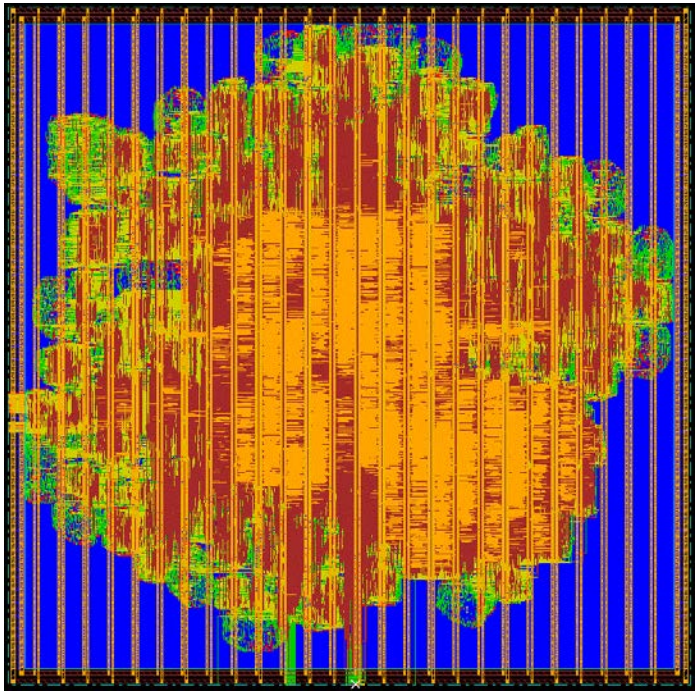


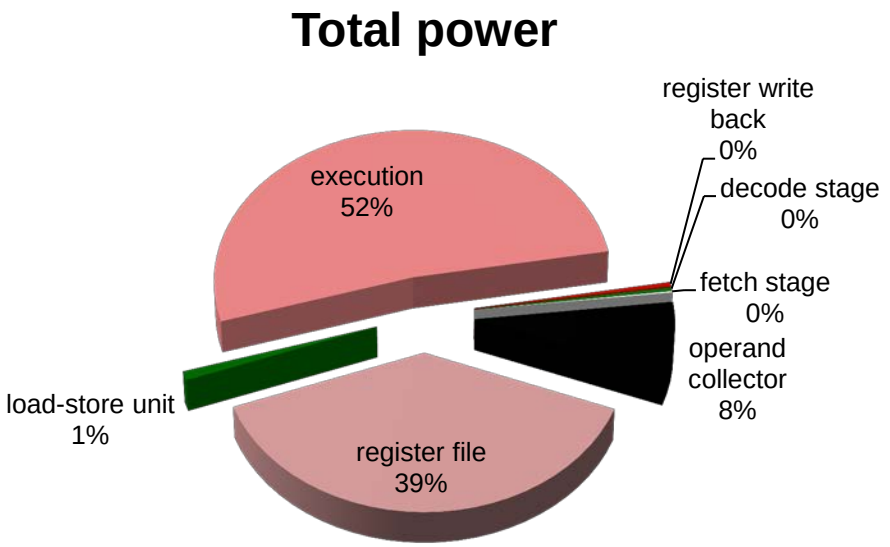
SIMD Compute Tile

- Problem:
  - Power consumed in instruction management, control, and data management = 10x power of computation
- Solution:
  - Low overhead SIMD tile
  - SIMD = Simultaneous Instruction Multiple Data
  - 16 Floating point lanes with shallow pipelines
  - Logic on SRAM on DRAM to manage memory BW needs



Power-per-component for 16 EX Lanes





Control overhead <10% of total power

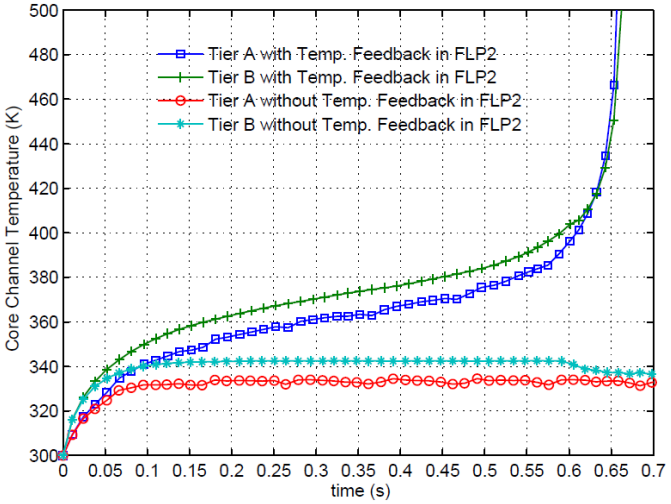
32 GFLOPS/W in 65 nm on FFT Benchmark

Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

Thermal Issues

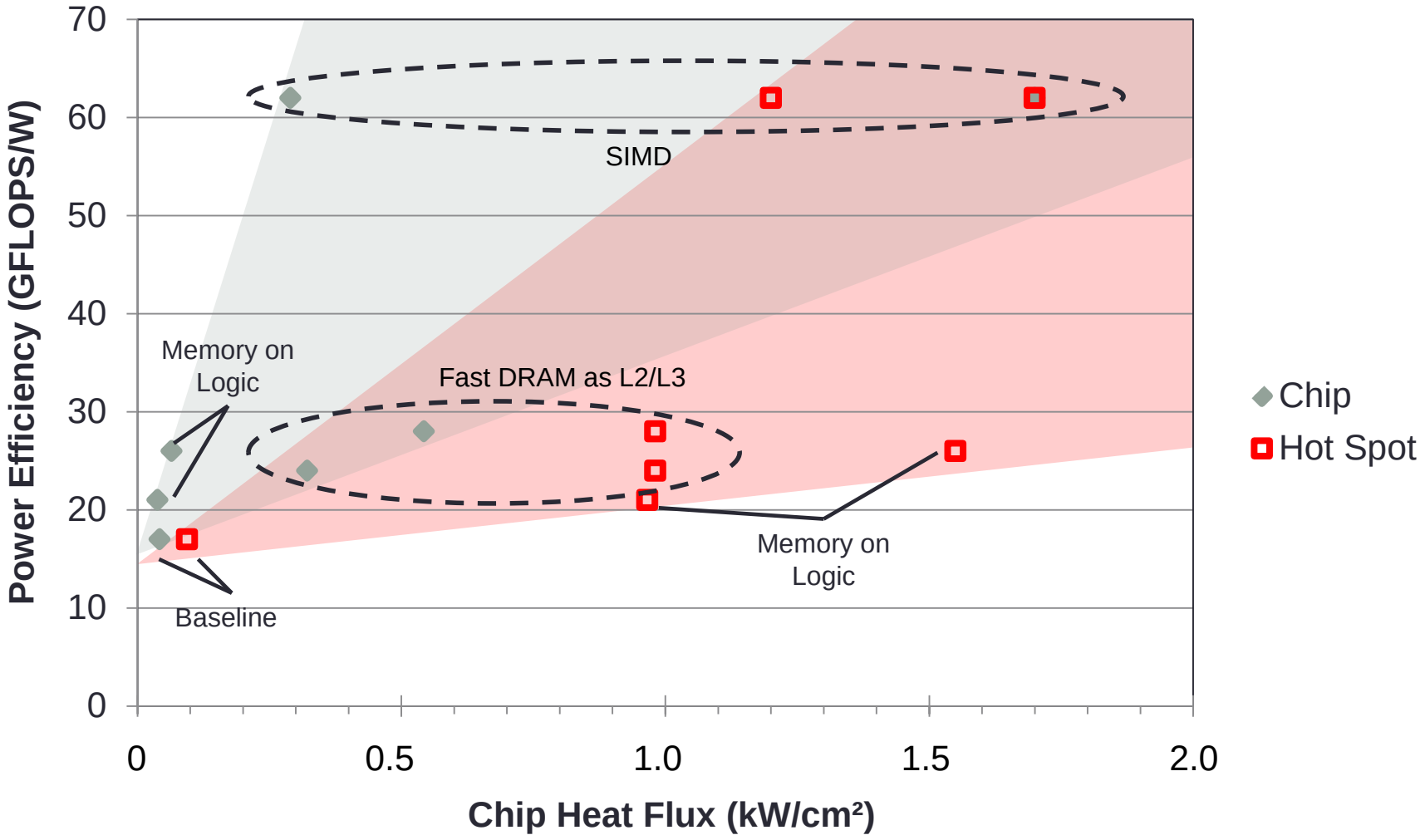
- Air cooling of 3DICs can be done
  - But requires active power management, including task migration and performance throttling
  - Significant threat: Co-heating of DRAM reduces refresh time
  - Also: Need transient simulation to capture self heating of SRAM



Thermal Runaway in SRAM on Logic

- Servers increasingly going to liquid cooling
  - 40% reduction in cooling power
  - Permits higher performance due to increased system density

Flux vs. Efficiency Tradeoff

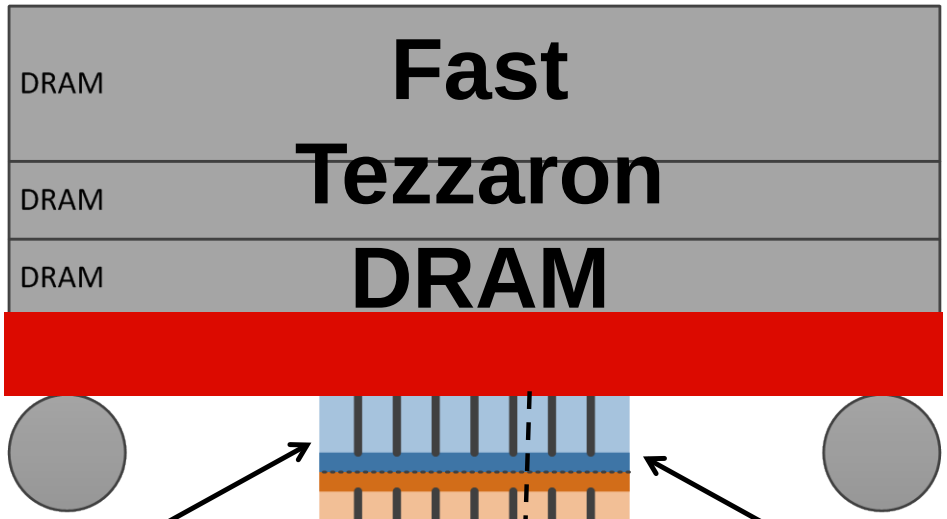


Outline

- 3D Technology Set
- Motivations
- 3D Memories
- Computing beyond 3D Memory
  - Logic stacking
  - Heterogeneous Computing
  - 3DECC – Parallel numerical at low power
- Thermal Challenges
- Conclusions

Conclusions

- Performance/Power improvement for each concept



L2/L3 cache in DRAM (8.5x over SRAM L2/L3)

RDL Interposer

Low Power scalar CPU

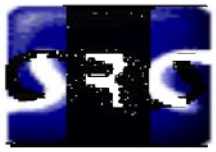
High Performance scalar CPU

SRAM SIMD Core

10^5 over GPU for Machine Learning

Fast Thread Transfer (23% over HP scalar)

Acknowledgements



Faculty: Rhett Davis, Michael B. Steer, Eric Rotenberg, James Tuck, Huiyang Zhou
Professionals: Steven Lipa, Eric Wyers
Current Students: Joonmu Hu, Brandon Dwiel, Zhou Wang, Marcus Tishibanqu, Ellliott Forbes, Randy Wilkiansano, Joshua Ledford, Jong Beom Park,
Past Students: Hua Hao, Samson Melamed, Peter Gadfort, Akalu Lentiro, Shivam Priyadarshi, Christopher Mineo, Julie Oh, Won Ha Choi, Ambirish Sule, Gary Charles, Thor Thorolfsson,
Department of Electrical and Computer Engineering
NC State University