Architectures for Extremely Scaled Memories

Paul Franzon

Department of Electrical and Computer Engineering

paulf@ncsu.edu

919.515.7351

High Level Overview

Challenges for Memories

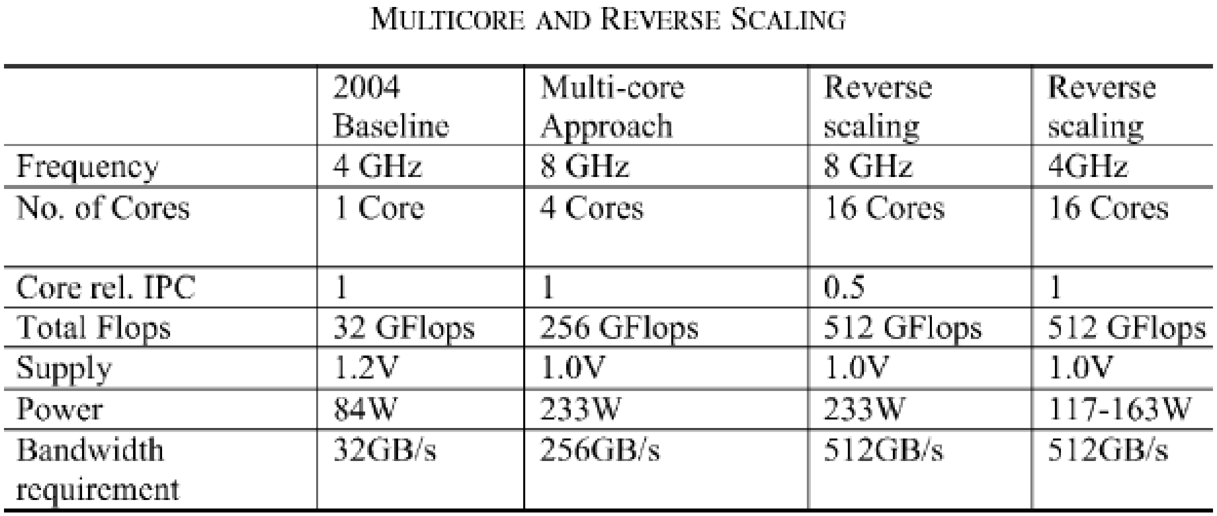
- ▷ Bandwidth
- ▷ Power consumption
- ▷ Resiliency
- ▷ Flexibility
- ▷ Scaling (density, size, level of integration)

Opportunities for Memories

- ▷ 3DIC with TSV
- ▷ Architectural Customization
- ▷ Use of memory in computing

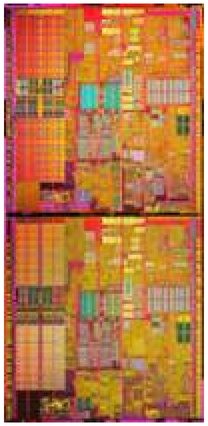
Challenge: Bandwidth

- Soon to exceed 1 TBps

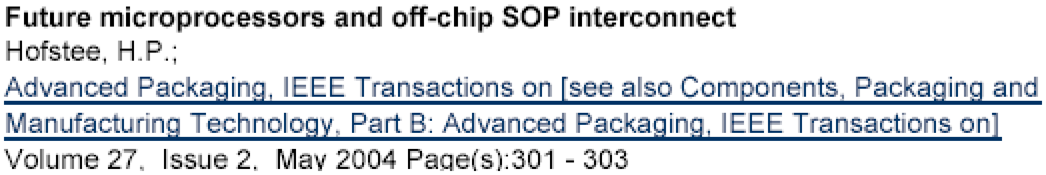


MULTICORE AND REVERSE SCALING

	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



Intel

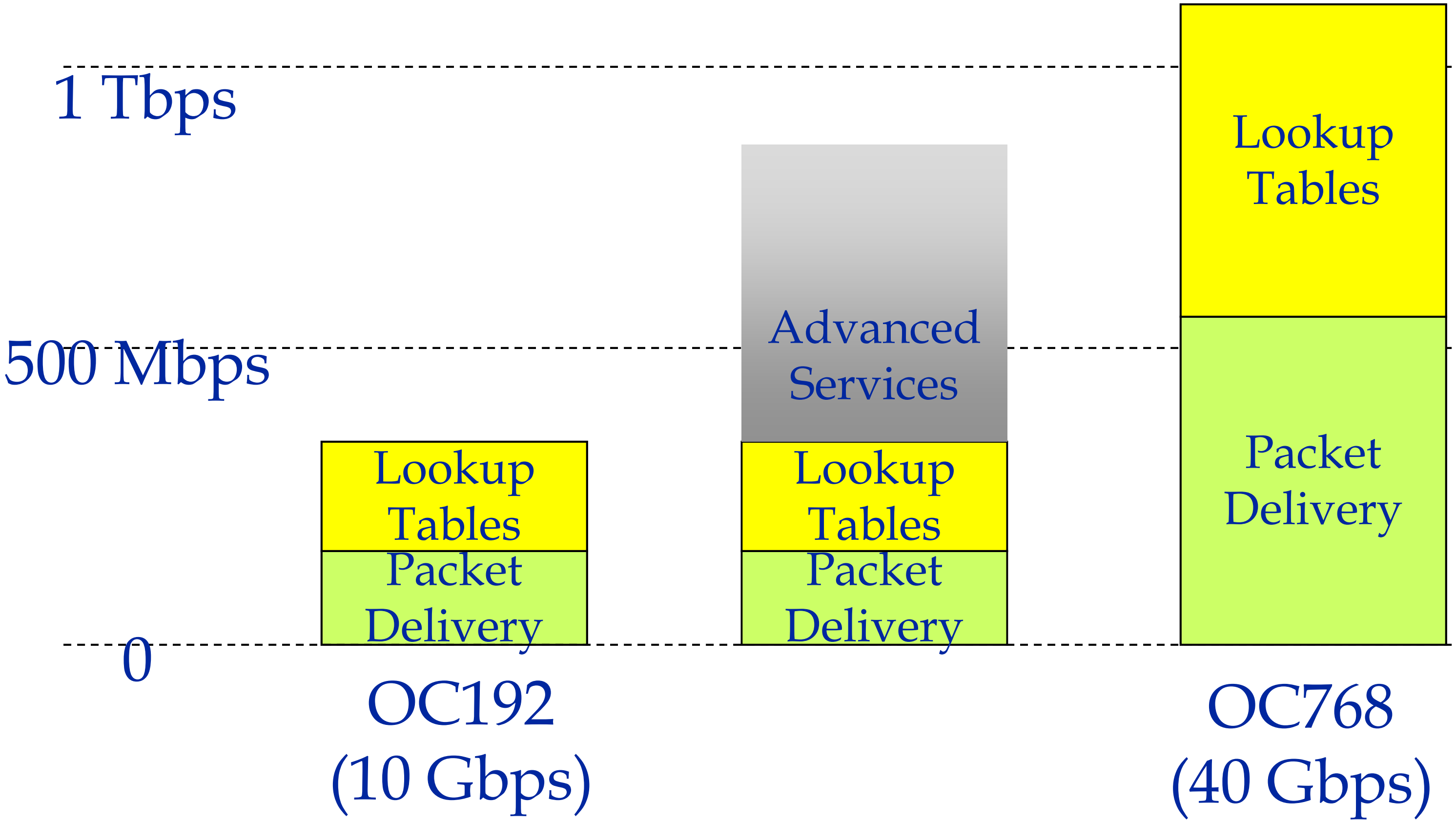


Future microprocessors and off-chip SOP interconnect
Hofstee, H.P.;
Advanced Packaging, IEEE Transactions on [see also Components, Packaging and Manufacturing Technology, Part B: Advanced Packaging, IEEE Transactions on]
Volume 27, Issue 2, May 2004 Page(s):301 - 303

Bandwidth

Graphics, and Networking have similar scaled bandwidth requirements

- 0.2 – 0.5 TBps required soon
- Networking has low latency requirements



Challenge: Power

Specifically providing this bandwidth at reduced power

- **DDR3 : 1 TBps ➔ 600 W of power**

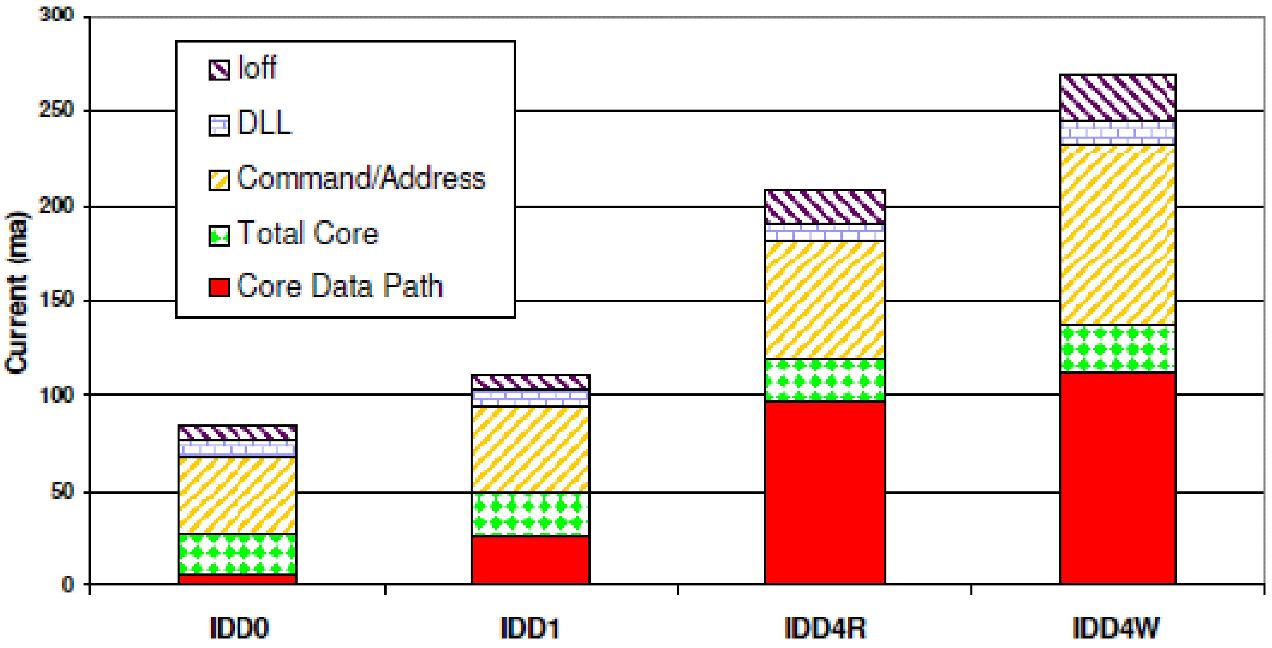
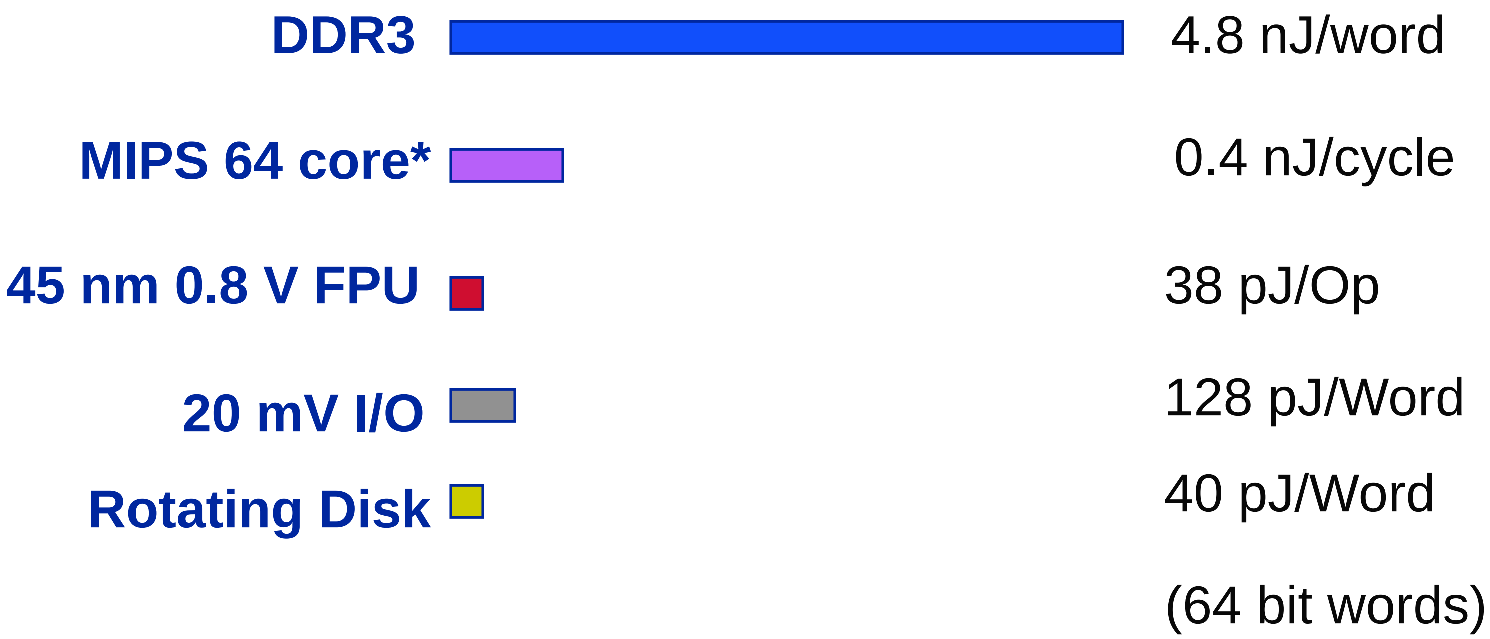


Figure 6.25: DDR3 current breakdown for Idle, Active, Read and Write.

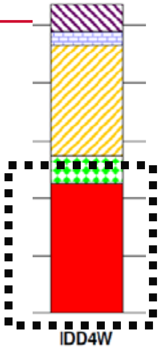
Comparative power consumptions



Without better solutions, memory power will dominate computing

* At 90 nm. Includes 40 kB cache, no FPU

Where does the power go?

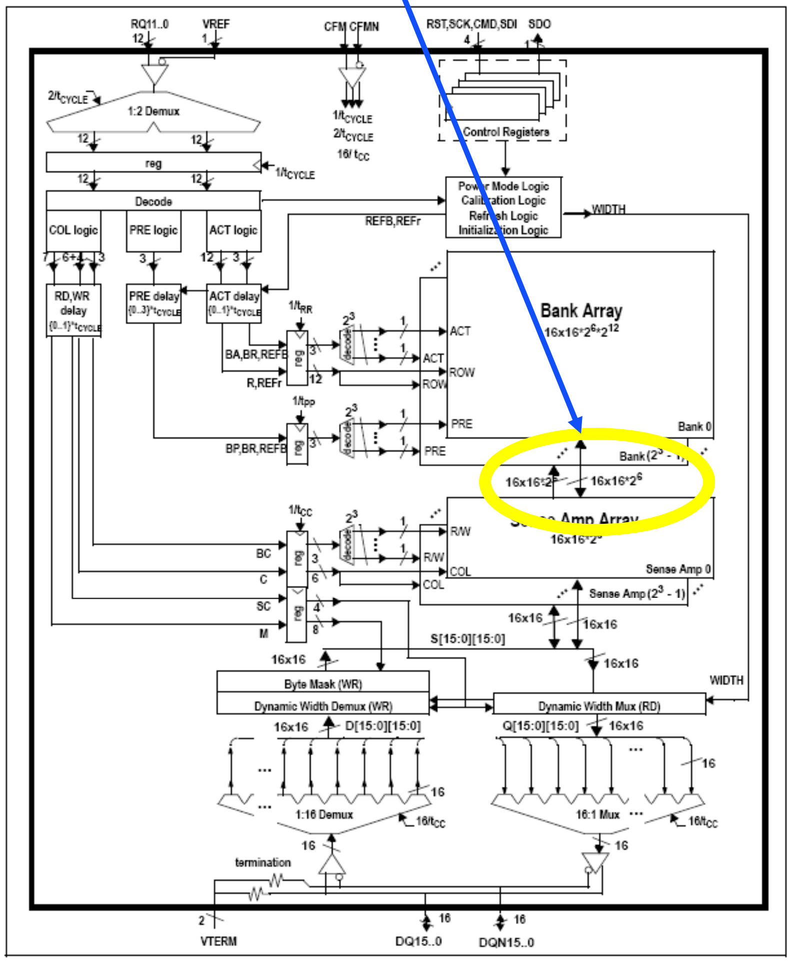


16,000 bit fetch

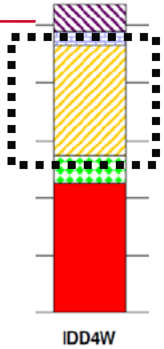
Core

- ▷ Cell: 25 fF @ 1.8 V
  - ▷ 81 fJ per bit
- ▷ Row
  - ▷ 8k to 16kbit wide
  - ▷ Driven at 2.5 V
  - ▷ 2.5 nJ/burst (1-4 bits)
- ▷ Sense amps
- ▷ Charge pumps to supply 1.8 and 2.5 V to core
  - ▷ Inefficient

Figure 2 512Mb (8x4Mx16) XDR DRAM Block Diagram



Where does the power go?



Command, address, data pipeline and “assist” circuits

- ▷ Many flip-flops
- ▷ DRAM process not ideal

Input/Output

- ▷ Difficult timing specs consume considerable power
- ▷ > 40 mW/pin

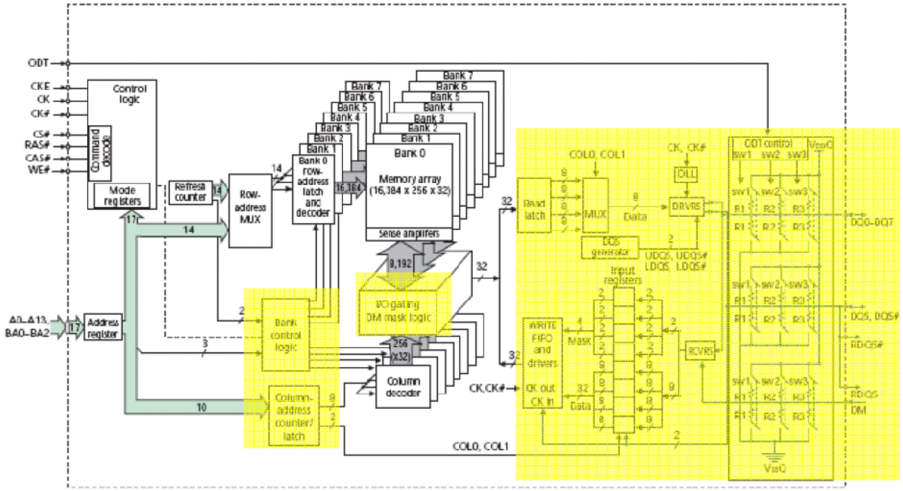


Figure 6.24: Block diagram of 1Gbit, X8 DDR2 device.

Power Scaling

Scaling Core Voltage

- Today 1.8 V
- Tomorrow, possibly 1.0 V, but scaling slowly
- What would be required to scale to 0.6 V?
- Advantages: Core power reduction; Reduced need for charge pumps

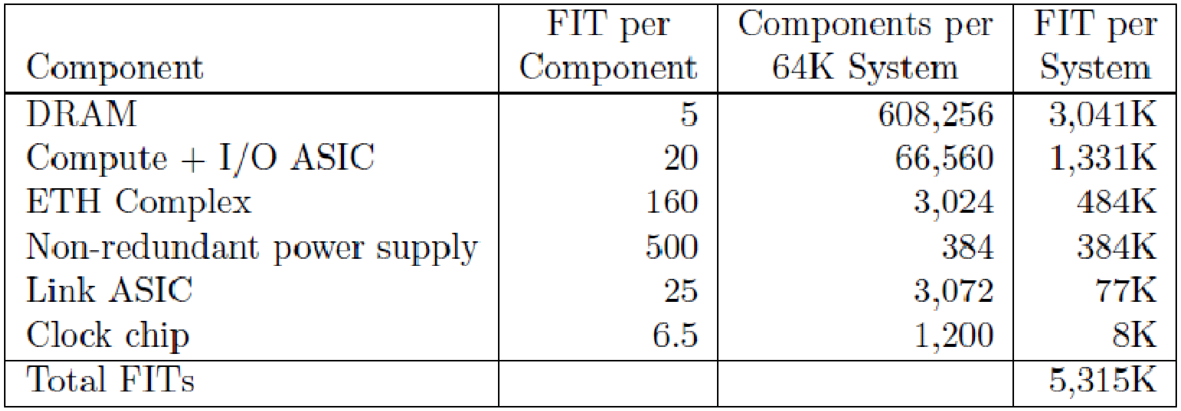
Scaling Command/Address/Data power

- Complex pipeline with many registers
- Increased desire for this pipeline to be configurable, increasing its design challenge and power consumption

Challenge: Resiliency

Issues:

- Soft Error Rate (SEU) of SRAM
- Checkpointing and resiliency of entire processor
- Future scaled server computers could spend 80% of their time checkpointing



Component	FIT per Component	Components per 64K System	FIT per System
DRAM	5	608,256	3,041K
Compute + I/O ASIC	20	66,560	1,331K
ETH Complex	160	3,024	484K
Non-redundant power supply	500	384	384K
Link ASIC	25	3,072	77K
Clock chip	6.5	1,200	8K
Total FITs			5,315K

Table 6.12: BlueGene FIT budget.

Note: DRAM Failures almost all due to packaging

Challenge: Cost per bit

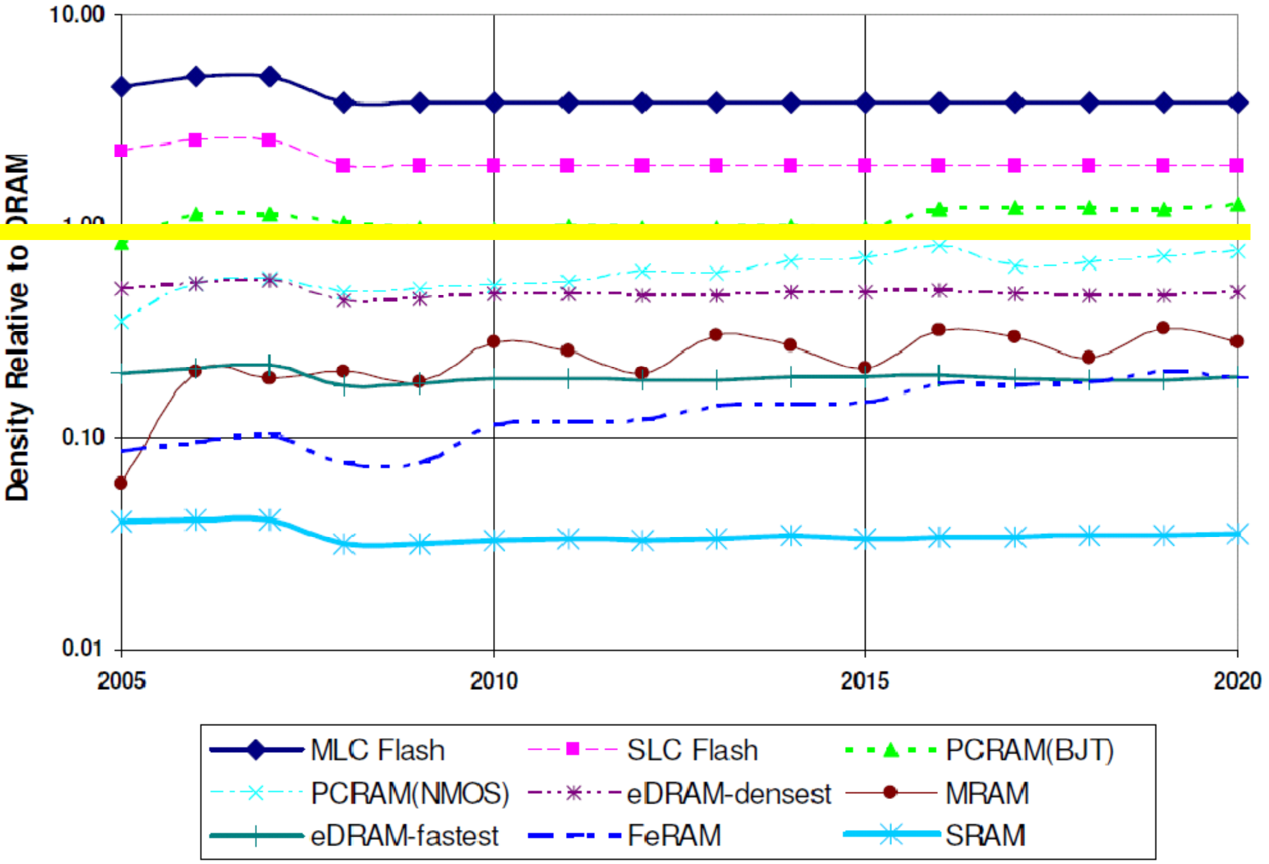
Issues:

- **Cell Size**

Technology	Cell Size	Comments
DRAM	$6F^2$	Capacitance scaling challenge
Flash	$4.5F^2$	Scaling uncertanties
PCRAM	$5.5F^2$	Density Challenges
Resistive RAM	$4F^2 – 6F^2$	Most promising? F can be small.

- **Fill Factor** (% of total silicon area used for memory cells)
  - Sub-array size
  - Area of peripheral and interface circuits
  - Most DRAMS ~ 30% - 40%

Density and Scaling



MLC Flash

PCRAM

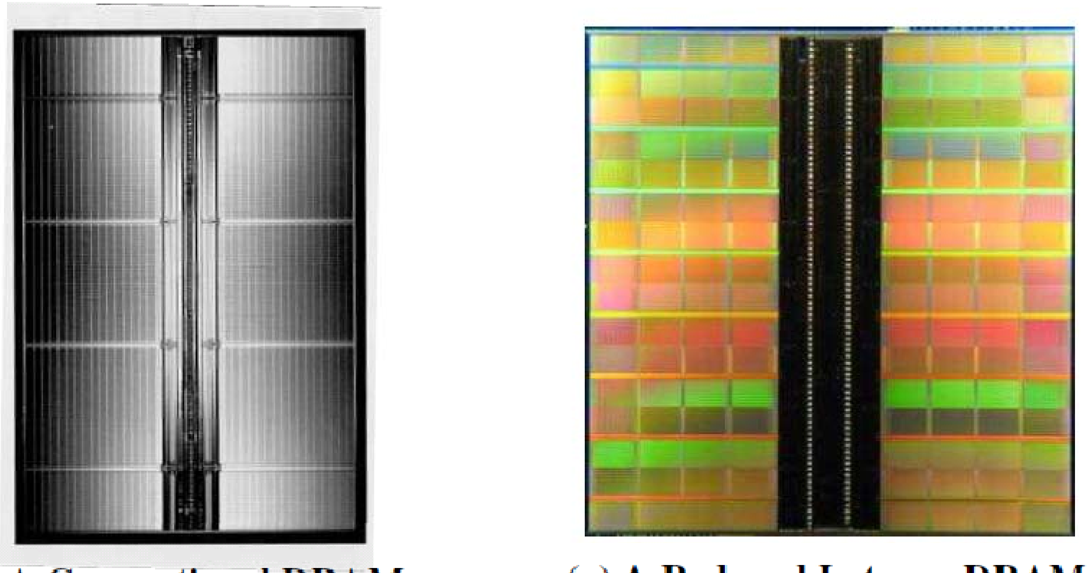
DRAM

MRAM

Figure 6.13: ITRS roadmap memory density projections.

Speed/Power ←→ Area tradeoff

Example: DRAM vs. Reduced Latency DRAM (RLDRAM)



(a) A Conventional DRAM

(a) A Reduced Latency DRAM

Figure 6.20: Reduced latency DRAM.

High Level Overview

Challenges for Memories

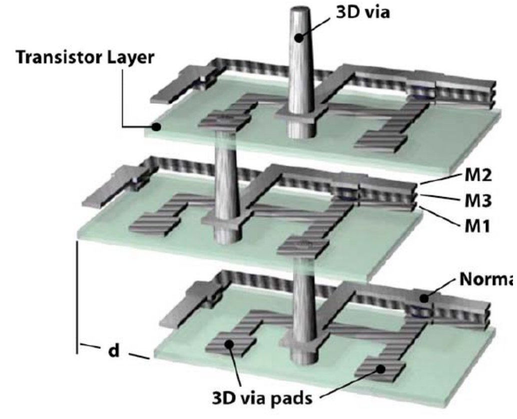
- Bandwidth
- Power consumption
- Resiliency
- Flexibility
- Scaling (density, speed, power)

Opportunities for Memories

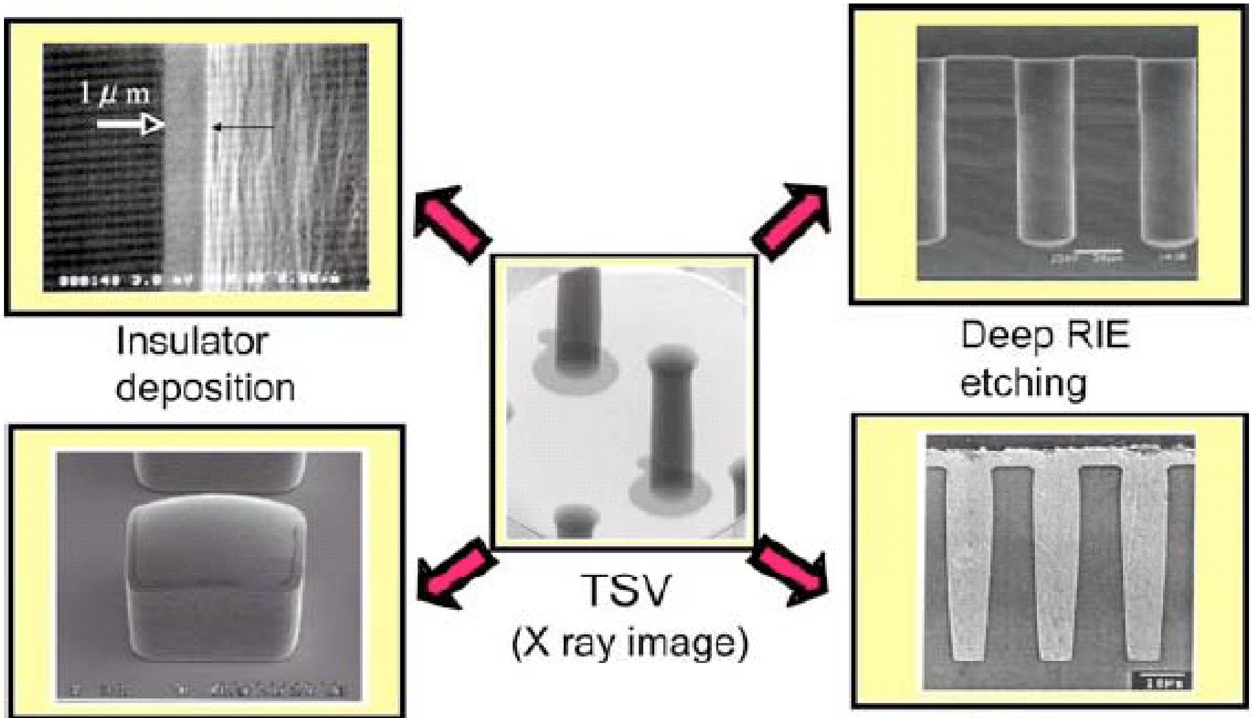
- 3DIC with TSV
- Architectural Customization
- 1R1D cell
- Increased use of memory in logic and routing

3DIC with Through Silicon Vias

Technoloav set:



Jniversity of Alberta.



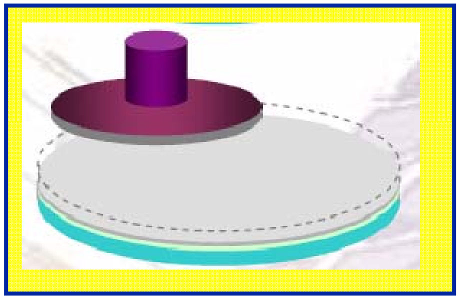
Insulator deposition

Deep RIE etching

TSV (X ray image)

Top and back bump formation

High aspect Cu plating

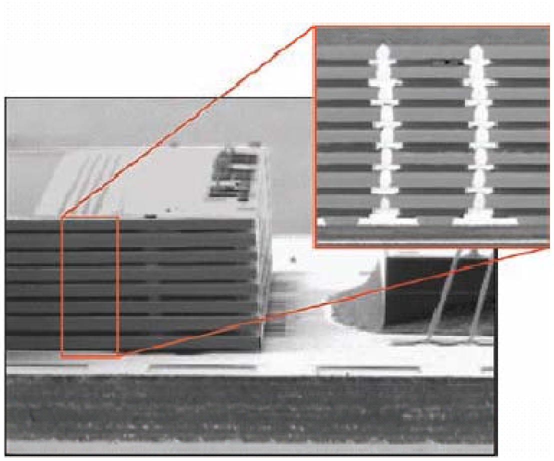


Wafer Thinning

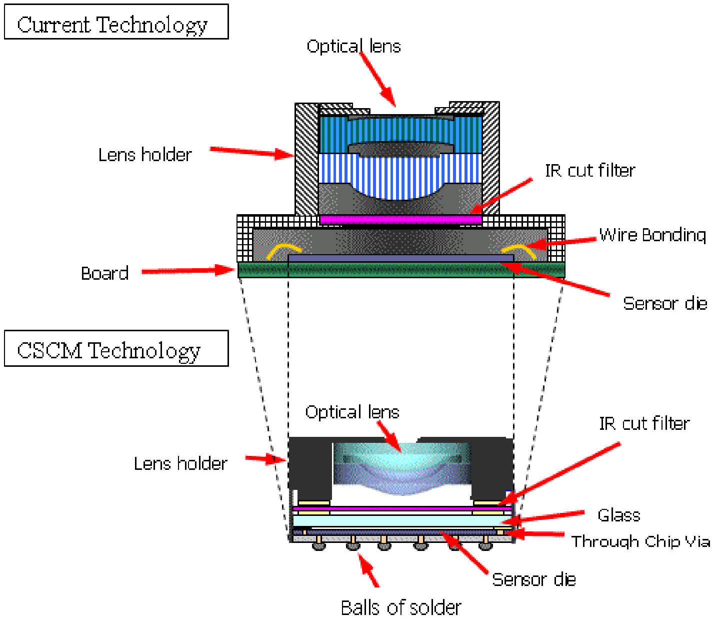
S. Denda, Nagano Prefectural Institute of Technology.

Coarse pitch TSV

- ▷ **Pitch: 40 μm to 250 μm**
- ▷ **Advantages**
  - ▷ Reduces need for wafer thinning
  - ▷ Established production route because of cell phone cameras
- ▷ **Disadvantages**
  - ▷ Limits architectural solutions
  - ▷ Really Advanced Packaging, not advanced integration

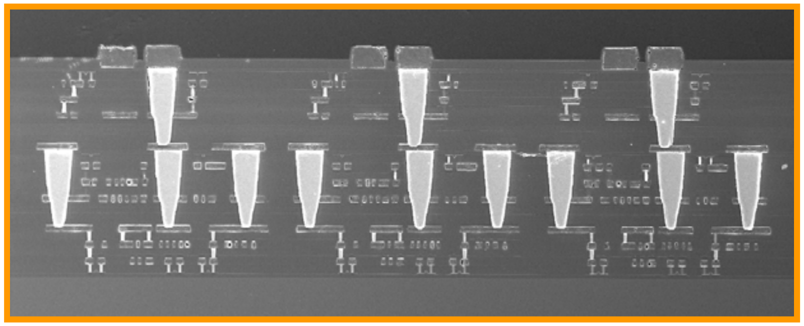


Samsung

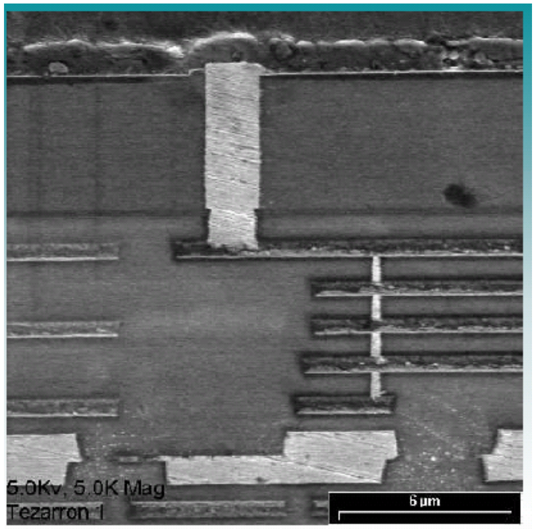


High Density TSV

- **Pitch: 1 µm to 10 µm**
- **Advantages:**
  - Permits architectural optimization
- **Disadvantages**
  - Adds processing cost
  - Adds complexity in design and test
  - Limited supply chain



MIT LL

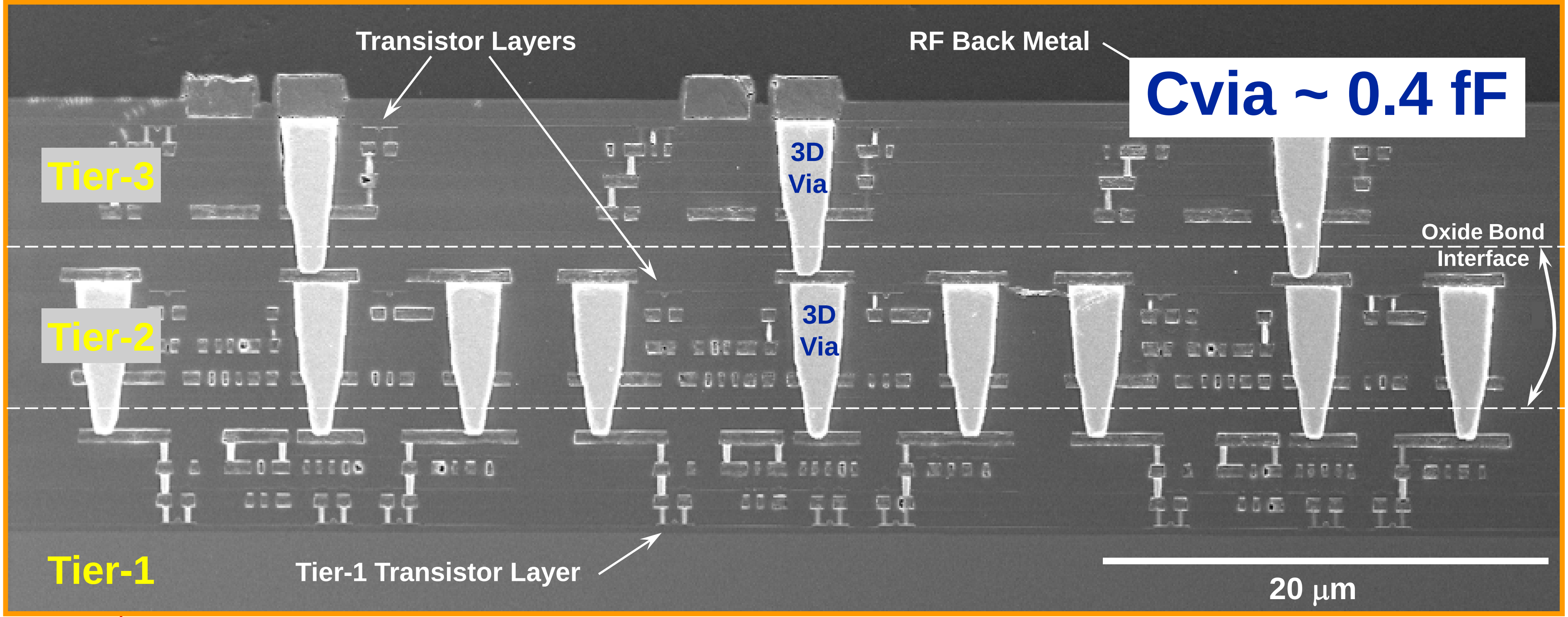


Tezzaron

3-Tier 3DIC Cross-Section

Second DARPA *Multiproject Run* (3DM2)

Two Digital & One RF 180-nm 1.5V FDSOI CMOS Tiers



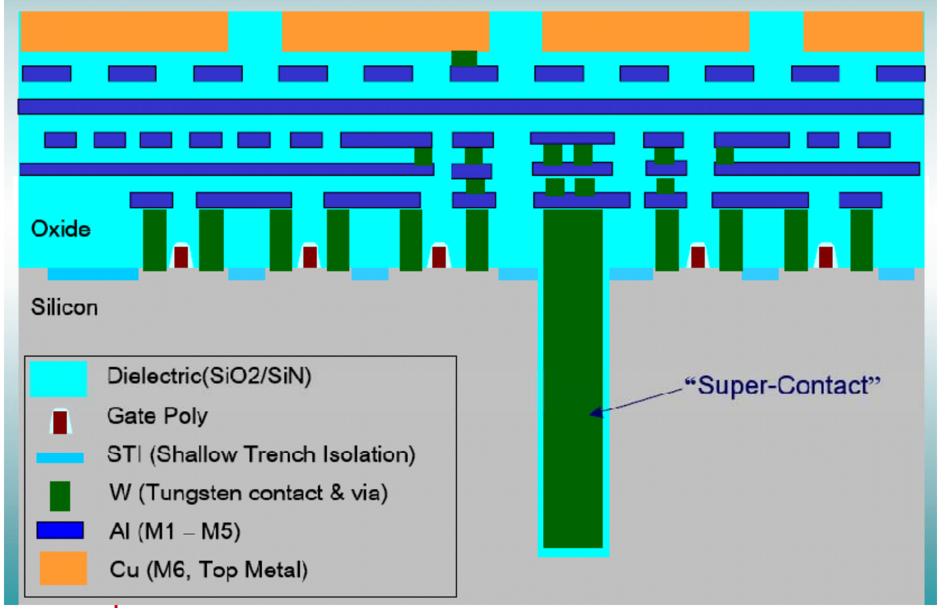
3DM2 Process Highlights

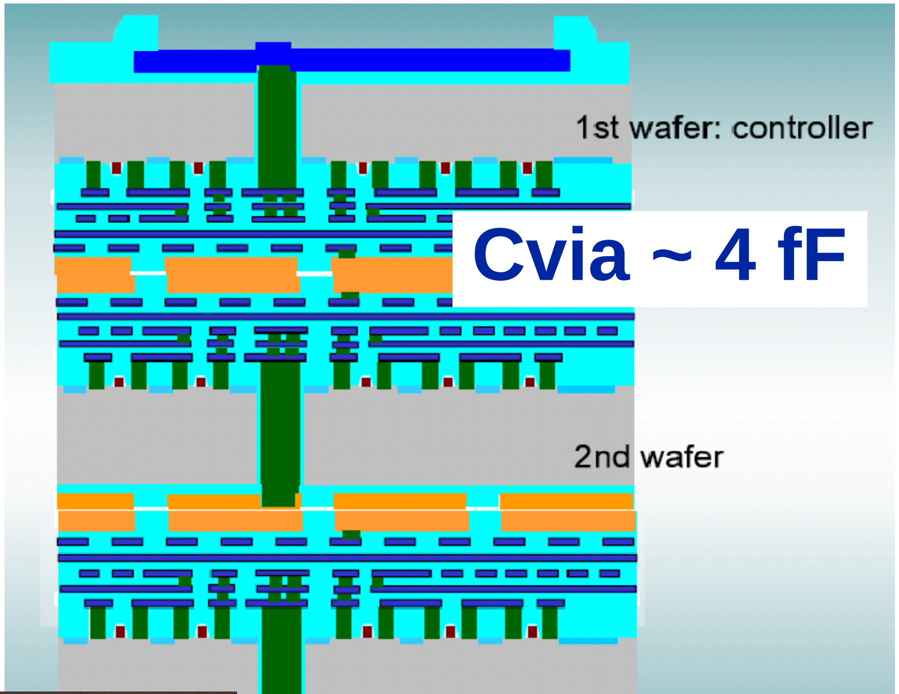
11 metal interconnect levels
1.75-µm 3D via tier interconnect
Stacked 3D vias allowed
Tier-2 back-metal/back-via process

2-µm-thick RF back metal
Tier-3 W gate shunt
Tier-3 silicide block

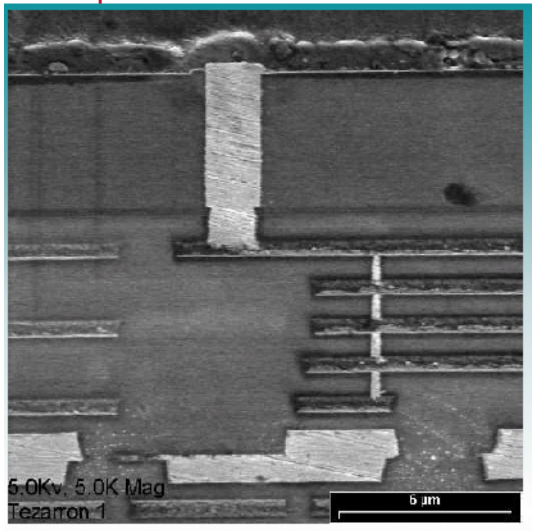
MIT Lincoln Labs

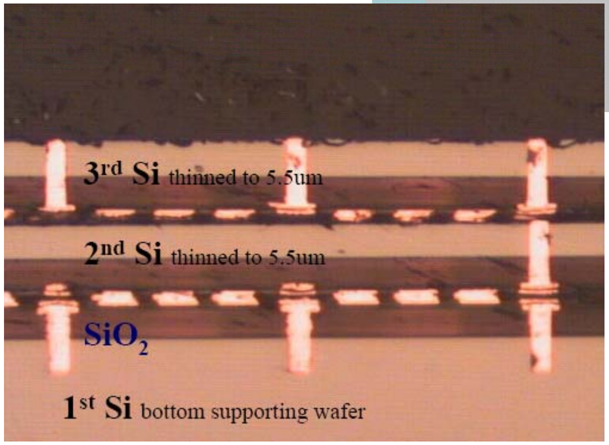
Tezzaron 3D Technology: 0.13 um Bulk CMOS





Cvia ~ 4 fF





Tezzaron

3DIC and Memory

Immediate application space:

- 3D memory stacking with coarse pitch TSVs
- Challenges:
  - Justifying initial cost
  - Cost scaling

More exciting application space:

- 3D-specific architectures
  - Memory-on-logic
  - High-density TSVs
- Challenges
  - Cost; test; design complexity

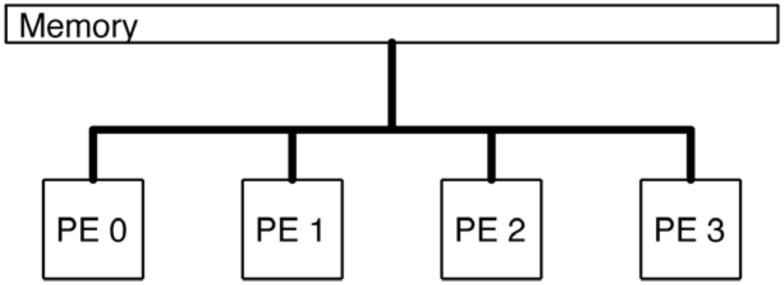
Example

- 3D Synthetic Aperture Radar Processor
  - Specifically FFT engine
- Opportunities Exploited
  - Co-architected memory and logic
- 3D specific design achieved the following
  - **65% power reduction**
  - **800% increase in memory bandwidth**
  - At cost of 22% increase in total silicon area (for the re-partitioned memory)
- 1024 point FFT:

  16 GFLOPS, 50 GBps in 2.6 x 3 mm

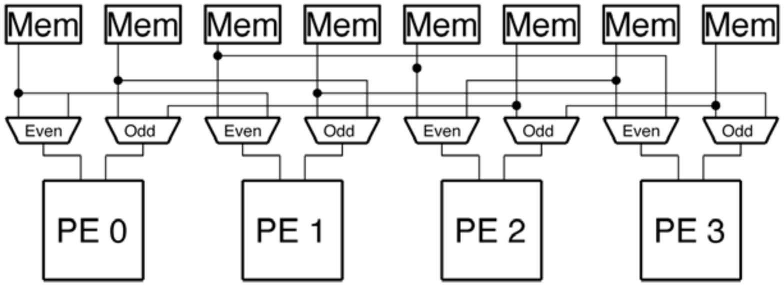
3D FFT for Radar Processor

2DIC “optimal” design (+/-)



One Big Slow Memory on Shared Bus

3DIC Optimal design



Multiple Individual Fast Memories

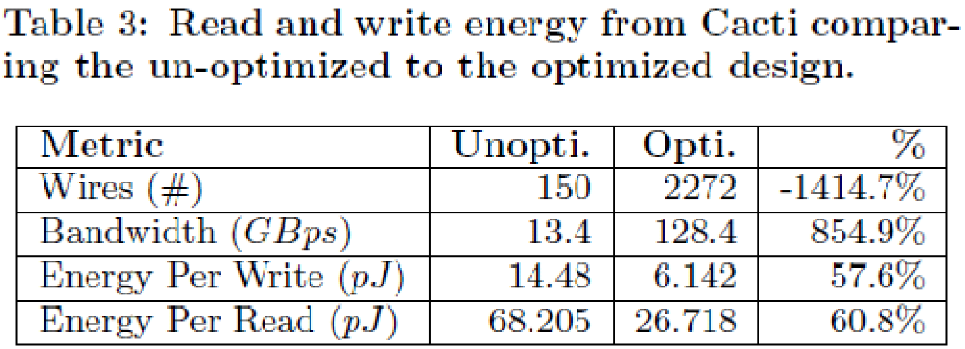


Table 3: Read and write energy from Cacti comparing the un-optimized to the optimized design.

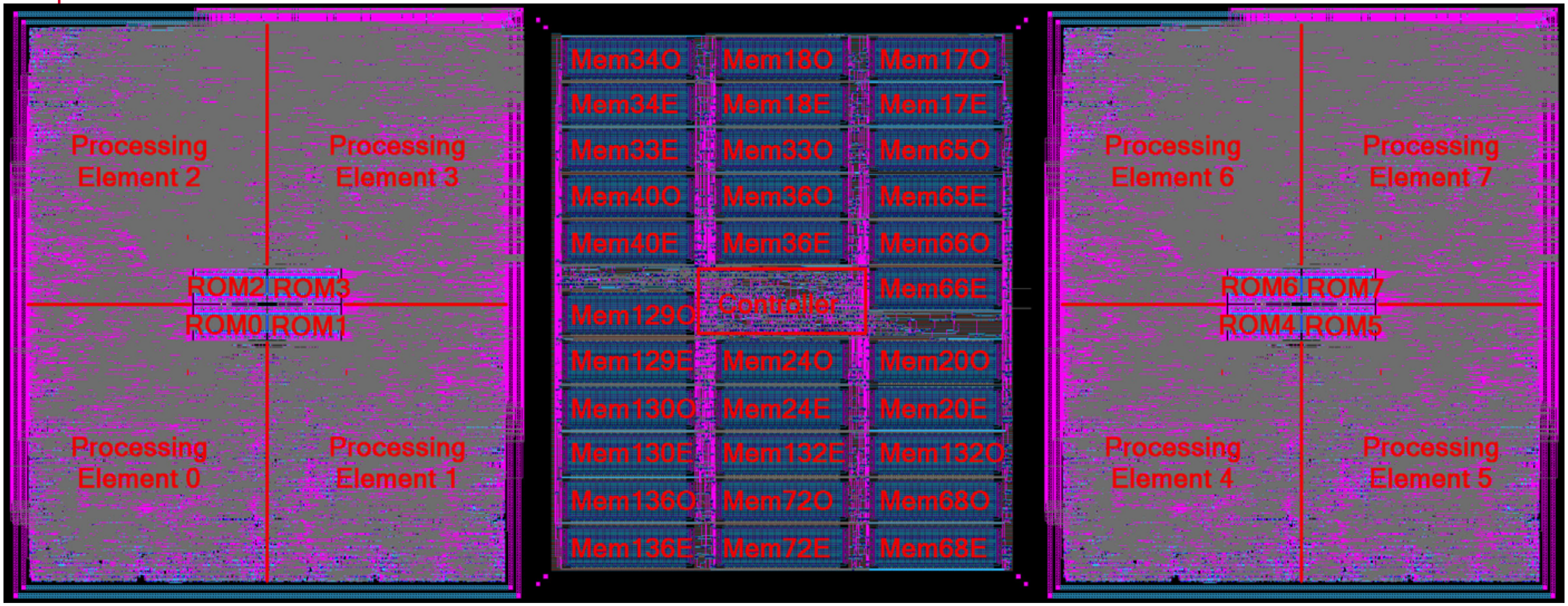
Metric	Unopti.	Opti.	%
Wires (#)	150	2272	-1414.7%
Bandwidth ($GBps$)	13.4	128.4	854.9%
Energy Per Write (pJ)	14.48	6.142	57.6%
Energy Per Read (pJ)	68.205	26.718	60.8%

- 60% reduction in memory power
- 67% increase in memory area
- 8x increase in bandwidth

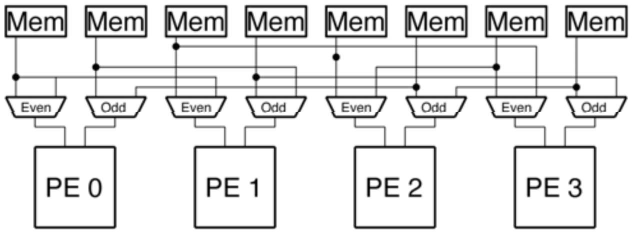
3D FFT Floorplan

Support multiple small memories WITHOUT an interconnect penalty

- Gives 60% memory power savings
- Memories communicate vertically only

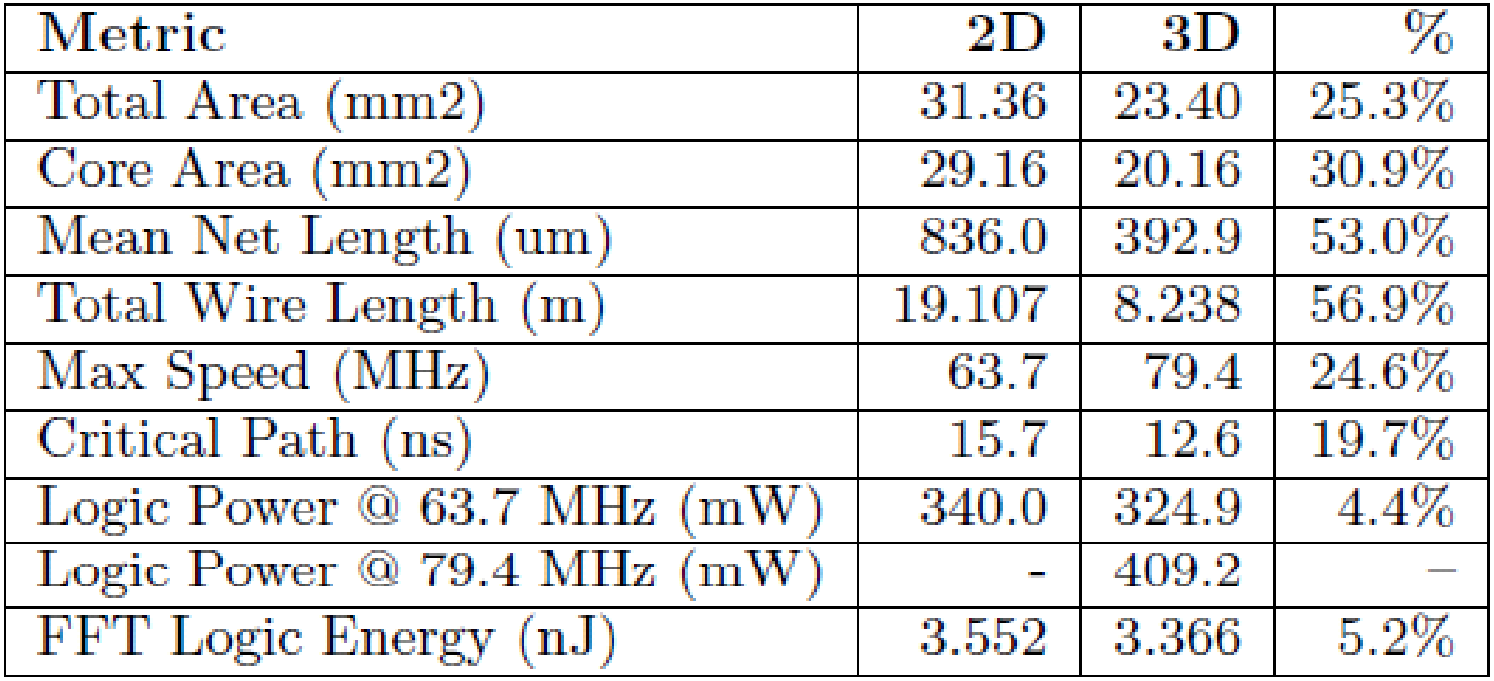


Implications of 3D



Multiple Individual Fast Memories

What are differences between 2D and 3D implementations of THIS architecture?

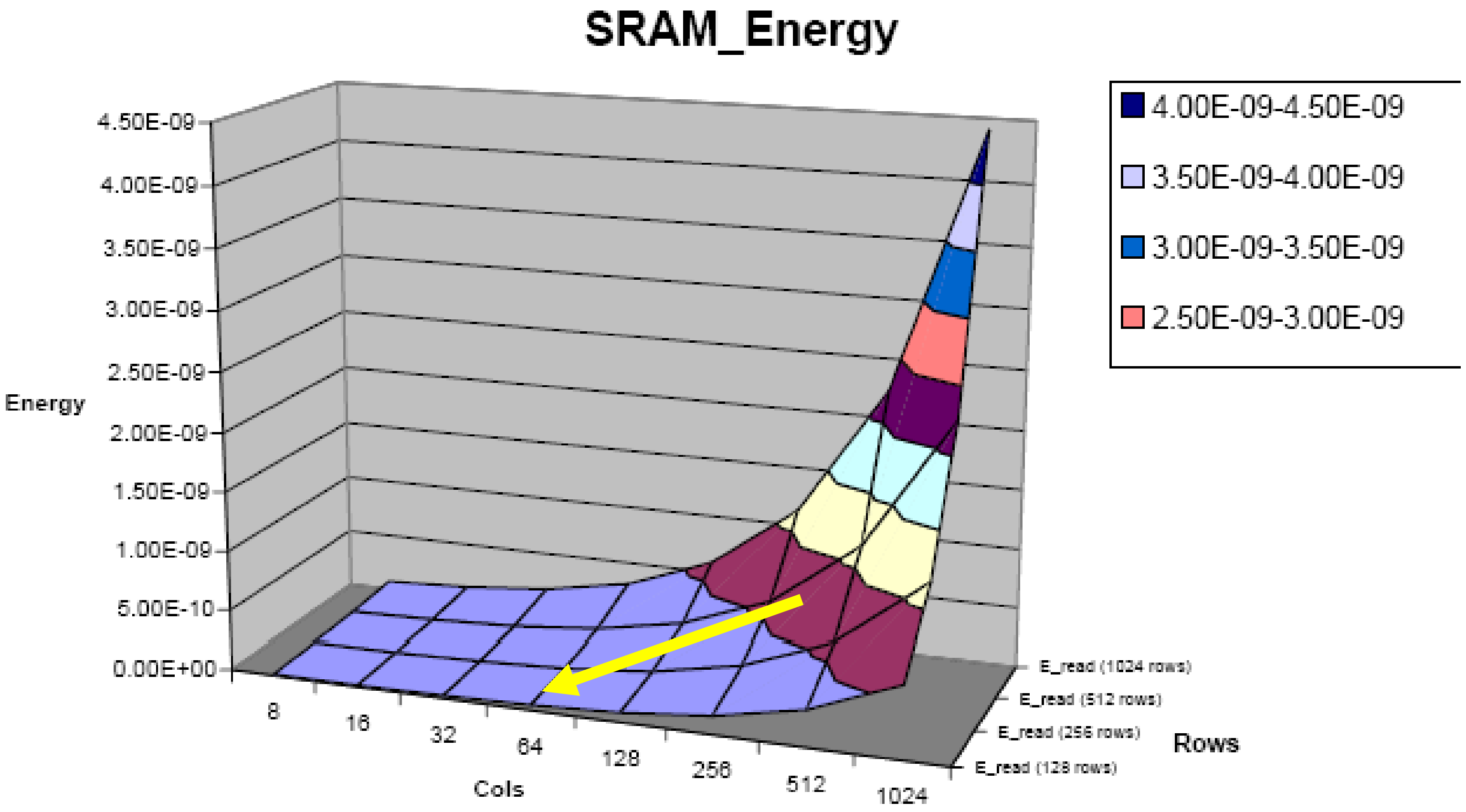


Metric	2D	3D	%
Total Area (mm2)	31.36	23.40	25.3%
Core Area (mm2)	29.16	20.16	30.9%
Mean Net Length (um)	836.0	392.9	53.0%
Total Wire Length (m)	19.107	8.238	56.9%
Max Speed (MHz)	63.7	79.4	24.6%
Critical Path (ns)	15.7	12.6	19.7%
Logic Power @ 63.7 MHz (mW)	340.0	324.9	4.4%
Logic Power @ 79.4 MHz (mW)	-	409.2	–
FFT Logic Energy (nJ)	3.552	3.366	5.2%

Memory bank size tradeoffs

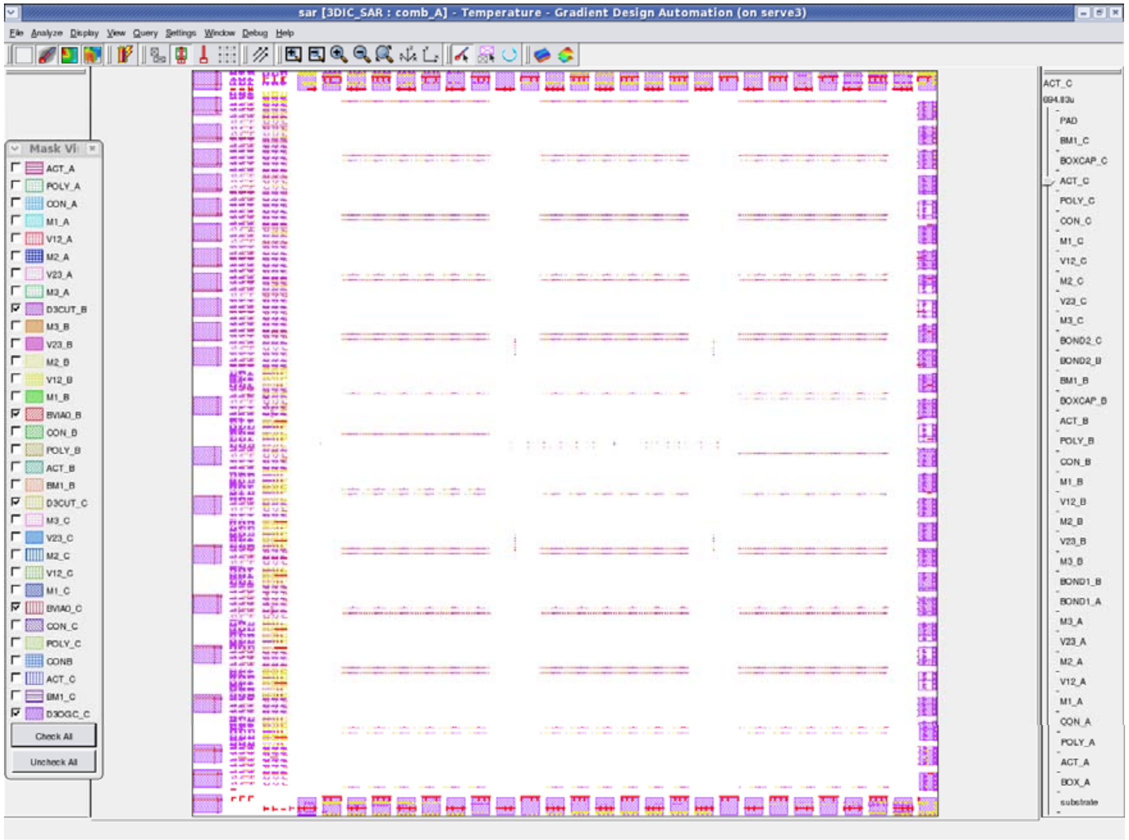
E.g. 32 x 2 kbit SRAM 10x less energy/bit than 1 x 64 kbit SRAM

- With 17% increase in area (partially recoverable by in 3D)



TSV Placement

Floorplanning, TSV placement and partitioning are easier in a memory-on-logic device than a logic-on-logic design



sar [3DIC_SAR : comb_A] - Temperature - Gradient Design Automation (on serve3)

Mask Vi

□ ACT_A
□ POLY_A
□ CON_A
□ M1_A
□ V12_A
□ M2_A
□ V23_A
□ M3_A
☑ D3CUT_B
□ M3_B
□ V23_B
□ M2_B
□ V12_B
□ M1_B
☑ BVIA0_B
□ CON_B
□ POLY_B
□ ACT_B
□ BM1_B
☑ D3CUT_C
□ M3_C
□ V23_C
□ M2_C
□ V12_C
□ M1_C
☑ BVIA0_C
□ CON_C
□ POLY_C
□ CONB
□ ACT_C
□ BM1_C
☑ D3OGC_C

17,634 TSVs

Power/Ground:

- 4554 A ←→ B
- 4800 B ←→ C

Signal:

- 4140 A ←→B
- 4140 B ←→ C

0.14 mm^2 of TSV
(1.7% area)

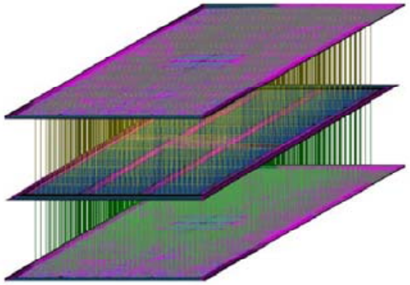
TSV Tradeoffs in FFT Processor

Process	Area loss
Lincoln Labs SOI	0.14 mm^2 1.7%
Tezzaron bulk CMOS	0.02 mm^2 0.3%
Package style TSV	2 mm^2 or more* (18%)

* Assumed “aggressive” effective 15 μm pitch (i.e. TSV + keepout)

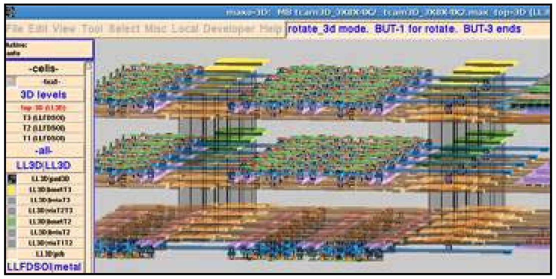
Circuit Level Partitioning

Above is block level partitioning



What about circuit level partitioning?

- Distributing banks amongst tiers?
- Distributing peripheral circuits



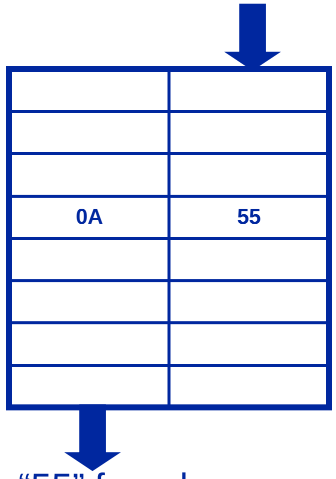
rotate_3d mode. BUT-1 for rotate. BUT-3 ends

- Issues:
  - Size of TSV vs. memory cell
  - Capacitance of TSV

Distributing banks amongst tiers

- **SRAM, DRAM:**
  - Potential advantages in a homogeneous technology memory stack are small
  - Little potential to decrease power or area
- **Content Addressable Memory**
  - Searches memory for content
  - Significant potential advantage
    - Due to high capacitance of match line
    - Match line == “found”

Search for “55”

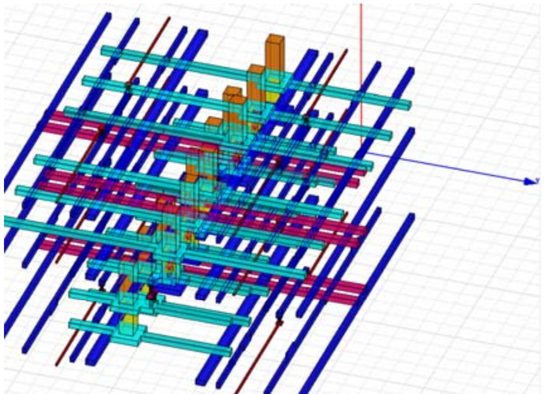


“55” found.
At address 0A

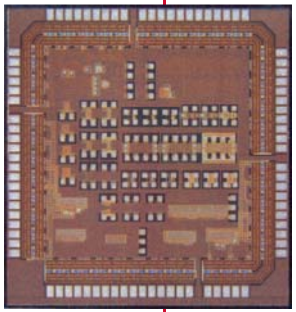
3D CAM: Advantages over 2D

In CAM Memory Core,

- 40% C_ML (matchline capacitance) reduction
- 27% P_ML (matchline power) reduction
- 23% overall power reduction
- Area (footprint) reduction of CAM core cells: ~50%



<Q3D model of interconnects for capacitance analysis>

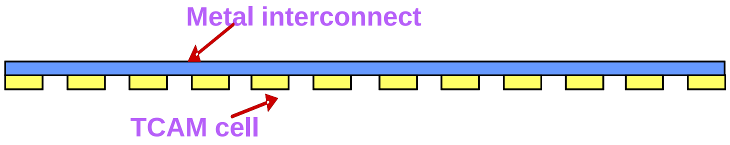


	2D Structure	3D Structure with 3 Tiers	Power reduction in %
P_ML	2.9p	2.1p	27%
P_total	8.0p	6.2p	23%

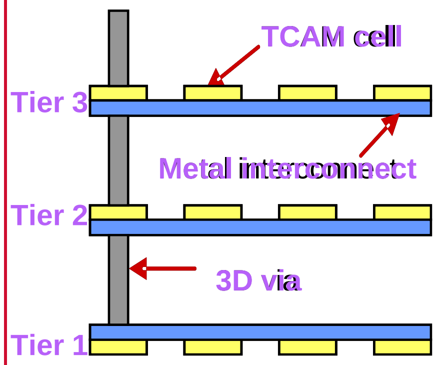
Oh

Only makes sense in low-capacitance SOI process

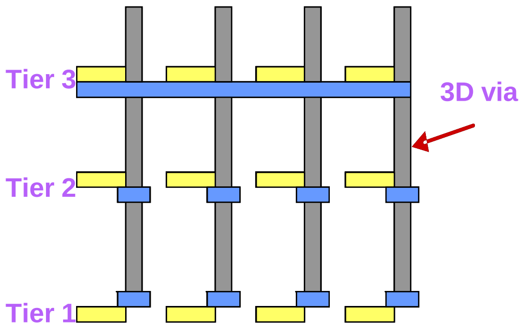
Partitioning Choices



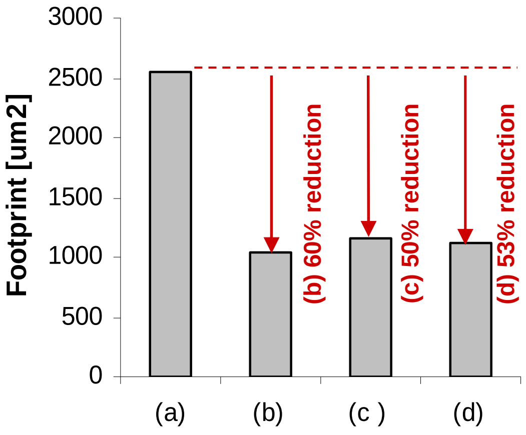
(a) 2D cross-sectional view



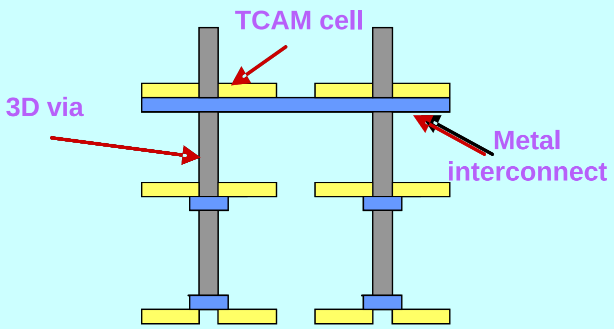
(b) Block partitioning



(c) Stub cell partitioning

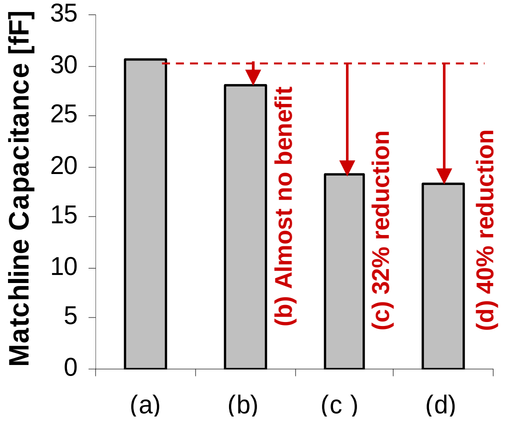


Footprint comparison



21% P_tot reduction vs. 2D TCAM core

(d) Stub shared cell partitioning



Matchline capacitance comparison
(Q3D field simulation)

Tezzaron "Dis-integrated RAM"

- **Mixed technology concept**
  - DRAM arrays in low-leakage DRAM technology (at node N)
  - Peripheral circuits in high-performance logic process (at node N-1)
  - Bit and word lines fed vertically at array edge

- **Expected results**
  - Reduced overall cost/bit
  - Faster interfaces
  - Lower latency
  - Reduced power/bit
  - Greater architectural flexibility

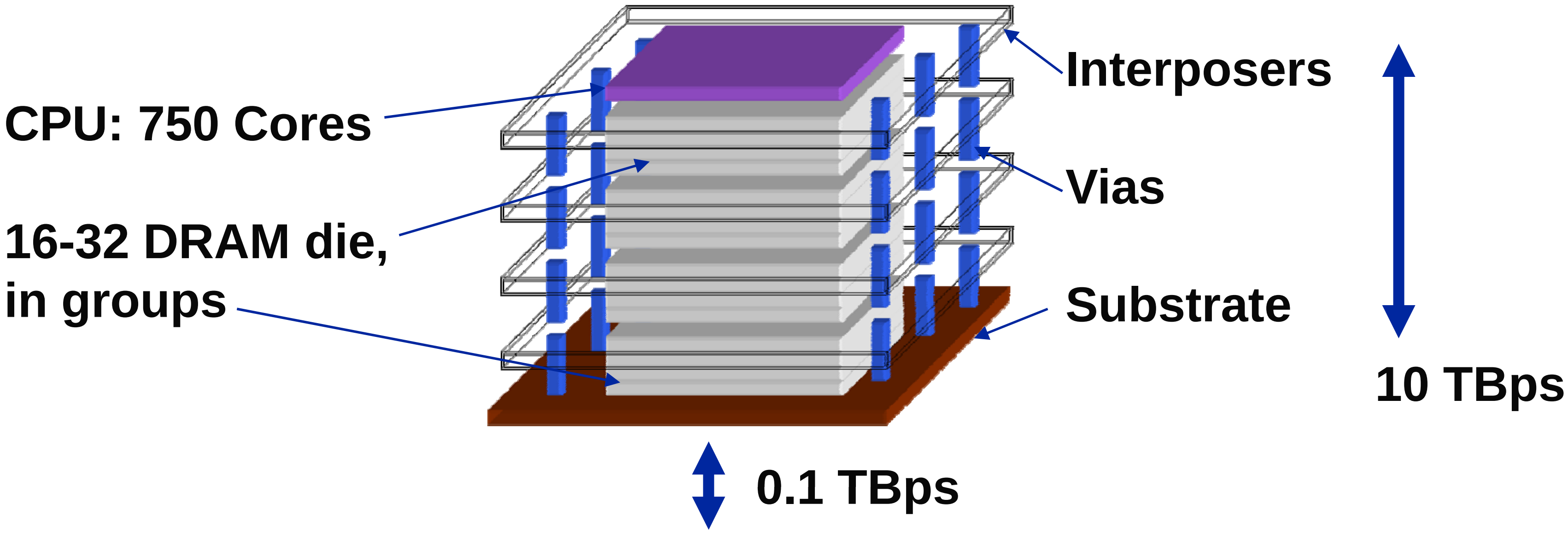
3DIC “Issues”

1. Cost
   - Cost in low volumes with 12” equipment will be high
   - Currently at bottom of volume and cost reduction learning curve
   - Try to recover through unique product advantage and reduced silicon area
2. Test
   - Known Good Die (or wafer) issues
   - Changes RAM test and burn-in strategies
3. Thermal
   - Power delivery / thermal dissipation codesign issue
   - Must keep DRAM below 90 C

Exascale Computing Node

Snapshot of the future?

- “Extreme” stacking needed to manage bandwidth and energy
- One computing node:



CPU: 750 Cores

16-32 DRAM die, in groups

Interposers

Vias

Substrate

10 TBps

0.1 TBps

Architectural Solutions

DDR optimized towards cache row refill

- ▹ And well suited for little else

Architectural Opportunities created by 3DIC RAM:

- ▹ Can separate memory array structure from architectural specification
- ▹ E.g. Tezzaron supplies “raw” multi-bank memory with SDRAM style interface
- ▹ Permits co-optimization of floorplan, logic, and memory
- ▹ With CPU cores, fast 3D RAM removes need for L2 cache

Nanoscale Emerging Memory Solutions

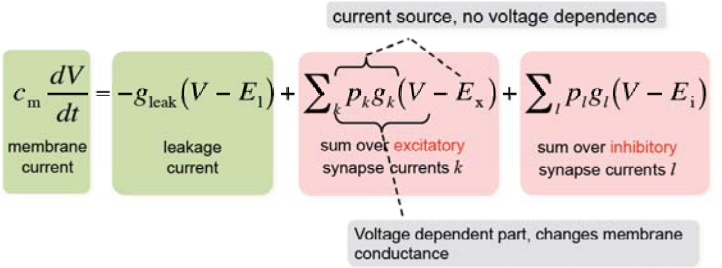
- **3DIC**
  - “Dis-integrate” with non-MOSFET based memories
- **Non-volatile memory**
  - Integrated functionality to improve resiliency of computers and logic
    - E.g. Embedded check-pointing
- **Neuromorphic computing**
  - Need: Analog memory or high density digital memory with DAC
- **Non-memory applications of emerging memory**
  - Routing; Analog functions

NC STATE UNIVERSITY

Neuromorphic Computing

Need for scaling:

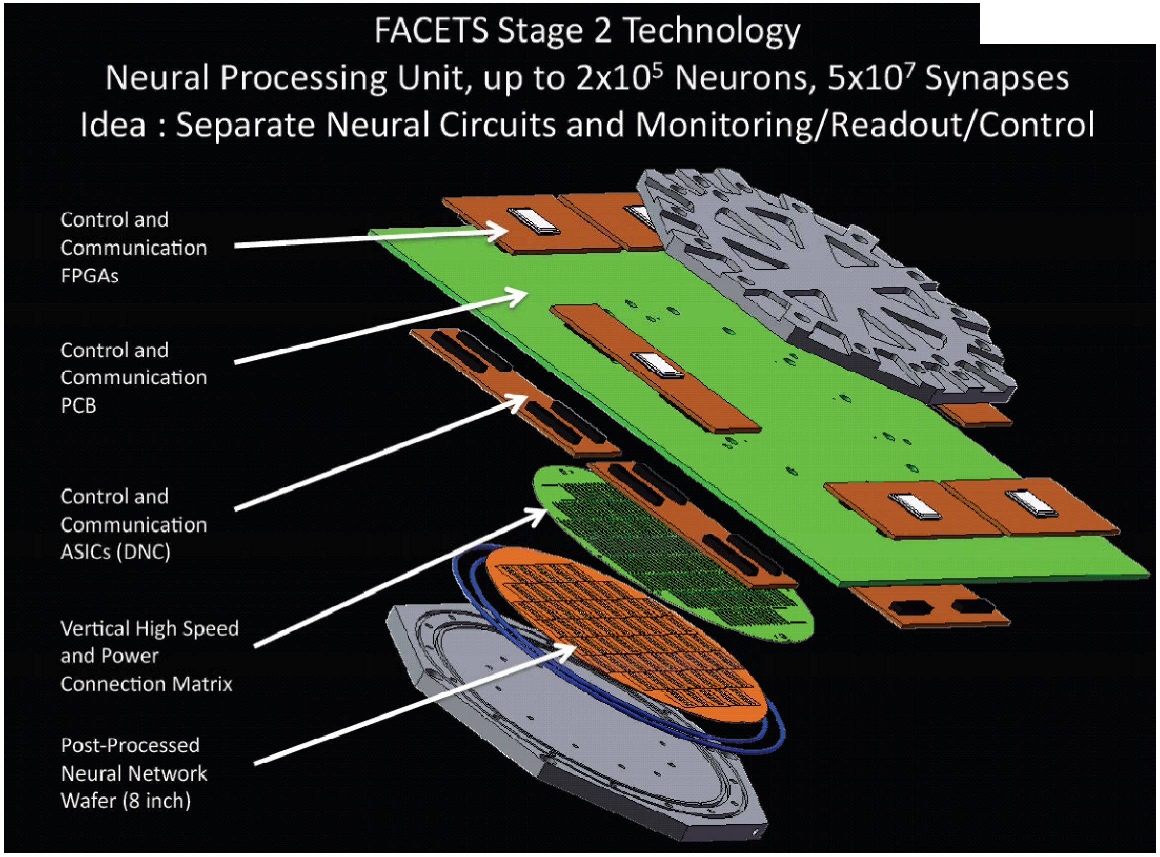
- ▷ Fast compact analog memory
- ▷ 3DIC



$$c_m \frac{dV}{dt} = -g_{\text{leak}}(V - E_l) + \sum_k \overbrace{p_k g_k}(V - E_x) + \sum_l p_l g_l (V - E_i)$$

current source, no voltage dependence
membrane current
leakage current
sum over excitatory synapse currents k
sum over inhibitory synapse currents l
Voltage dependent part, changes membrane conductance

Synaptic Computation Model

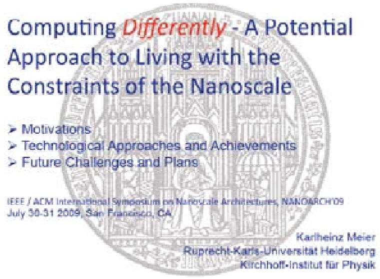


FACETS Stage 2 Technology

Neural Processing Unit, up to 2×10^5 Neurons, 5×10^7 Synapses

Idea : Separate Neural Circuits and Monitoring/Readout/Control

Control and Communication FPGAs
Control and Communication PCB
Control and Communication ASICs (DNC)
Vertical High Speed and Power Connection Matrix
Post-Processed Neural Network Wafer (8 inch)



Computing *Differently* - A Potential Approach to Living with the Constraints of the Nanoscale

- ➢ Motivations
- ➢ Technological Approaches and Achievements
- ➢ Future Challenges and Plans

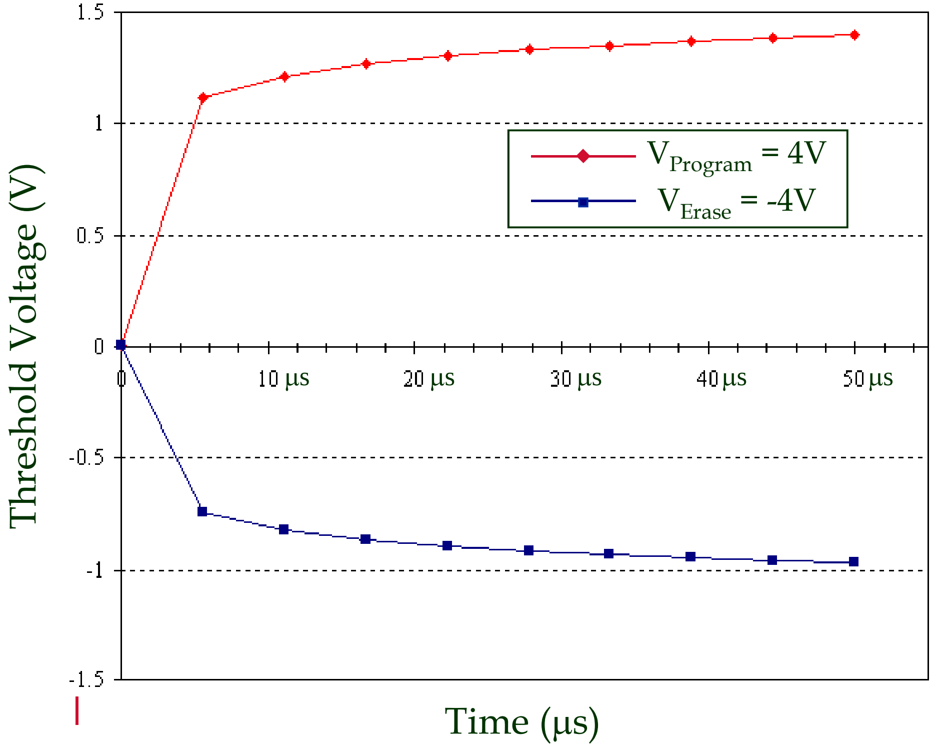
IEEE / ACM International Symposium on Nanoscale Architectures, NANOARCH'09
July 30-31 2009, San Francisco, CA

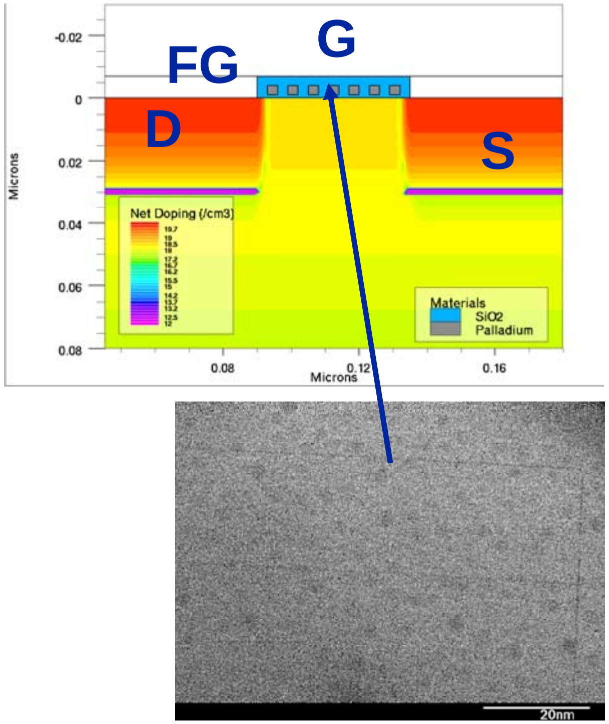
Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg
Kirchhoff-Institut für Physik

Nano In Logic

Key: Highly programmable, analog FET based on nanocrystal metal floating gate

Threshold Voltage vs. Program/Erase Time





Metal Nanocrystal Floating Gate

- High density of states
- Reliable
- Good retention

Example: NC FG-based FPGA

1. Shows benefit of a memory device in a static reconfigurable interconnect application
2. Palladium Metal nanocrystal flash reduces programming voltage to 3-4 V

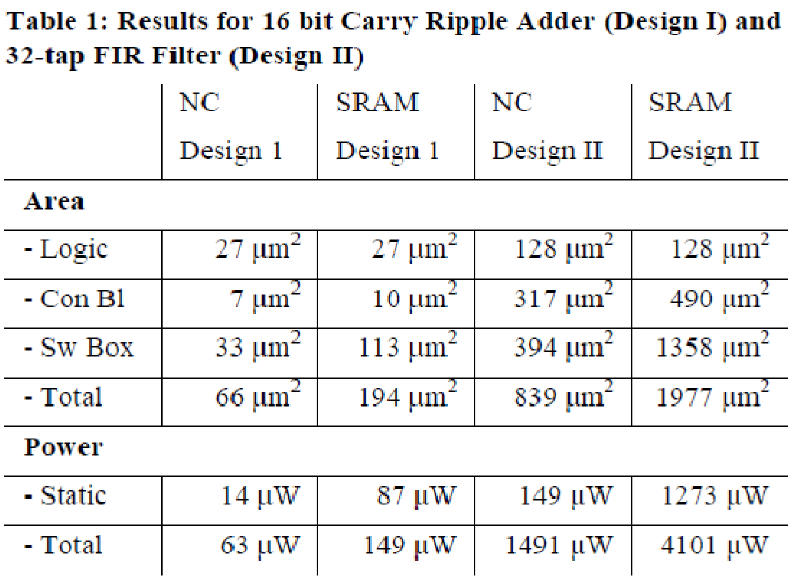
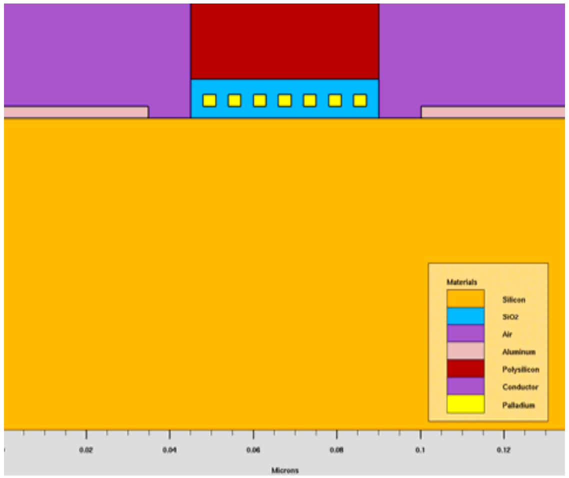
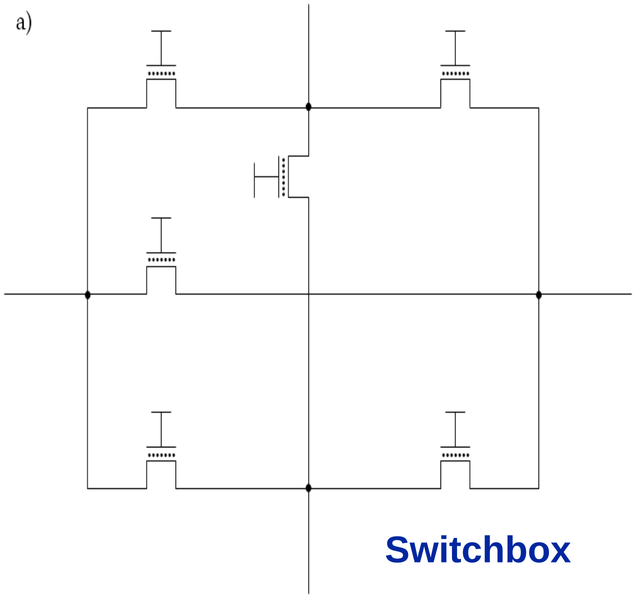


Table 1: Results for 16 bit Carry Ripple Adder (Design I) and 32-tap FIR Filter (Design II)

	NC Design 1	SRAM Design 1	NC Design II	SRAM Design II
Area				
- Logic	27 µm^2	27 µm^2	128 µm^2	128 µm^2
- Con Bl	7 µm^2	10 µm^2	317 µm^2	490 µm^2
- Sw Box	33 µm^2	113 µm^2	394 µm^2	1358 µm^2
- Total	66 µm^2	194 µm^2	839 µm^2	1977 µm^2
Power				
- Static	14 µW	87 µW	149 µW	1273 µW
- Total	63 µW	149 µW	1491 µW	4101 µW





8x power savings
4x area savings

Switchbox

Conclusions

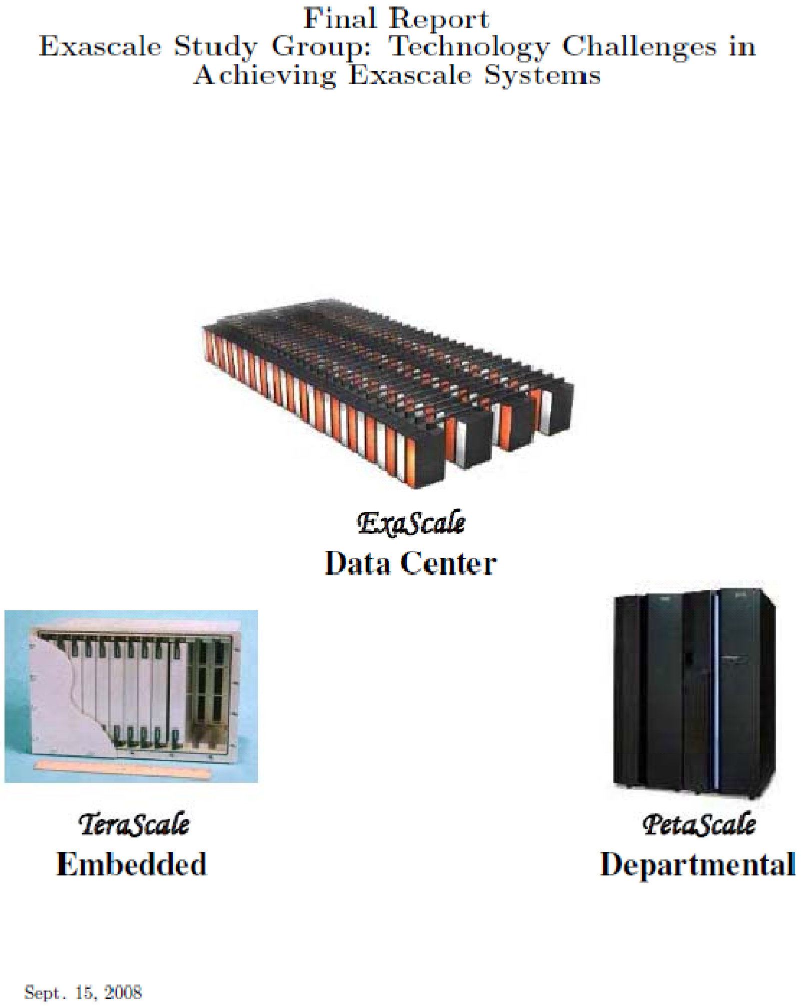
- **Memory Business readying for disruptive change**
  - Mix of rising challenges and emerging opportunities
  - Key: Delivering new technological responses cost-effectively
- **Challenges**
  - Bandwidth
  - Power at this bandwidth
  - Cost
- **Opportunities**
  - 3DIC
  - 1D1R memory
  - Non-traditional architectural mixes

Acknowledgements

William Rhett Davis, Michael B. Steer, Mehmet Ozturk,
Hua Hao, Steven Lipa,
Sonali Luniya, Christopher Mineo, Julie Oh,
Ambirish Sule, Thor Thorolfsson, Chirs Amsinck,
Neil DiSpigna, Shep Pitts, Daniel Schinke,
Department of Electrical and Computer Engineering
NC State University

Acknowledgments

My colleagues on



Final Report
Exascale Study Group: Technology Challenges in
Achieving Exascale Systems

ExaScale
Data Center

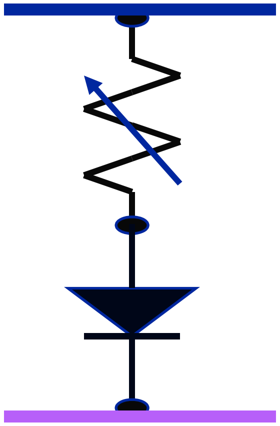
TeraScale
Embedded

PetaScale
Departmental

Sept. 15, 2008

Benefits of 1R1D cell

- ▷ Permits highest core density
- ▷ With high on:off ratio, large arrays are possible



On:off Ratio	Max. Array
7:1	64x64
13:1	128X128
100:1	1225X1225
1000:1	12kX12k
8000:1	1MX1M

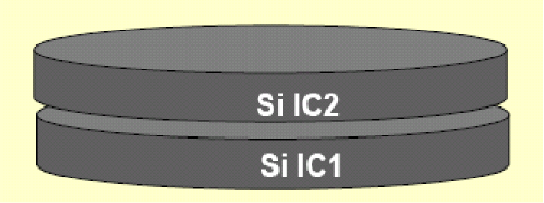
C. Amsinck, N. DiSpigna, D. Nackashi, P. Franzon, “Scaling constraints in nanoelectronic random-access memories,” Nanotechnology 16(10), Oct. 2005, pp. 2251 – 2260.

3DIC Test

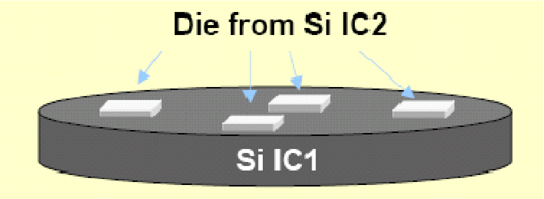
- Problem: Yield impact of accumulated (untested) silicon area

One tier	Two tiers	Three tiers	Four tiers
95%	90%	85%	81%

- Wafer on wafer stacking
  - Test before assembly has uncertain utility

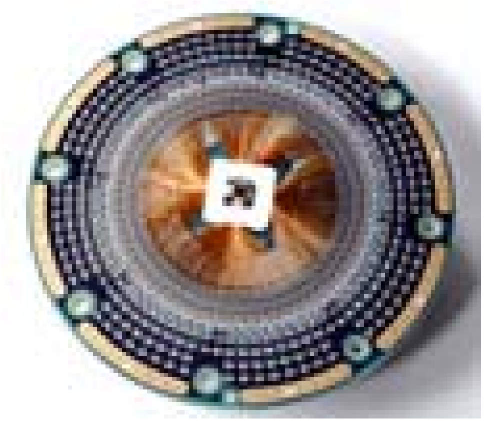


- Chip on wafer stacking
  - Known Good Die potentially highly useful



3DIC Test

Wafer probing a multi-thousand pin TSV field is unscalable



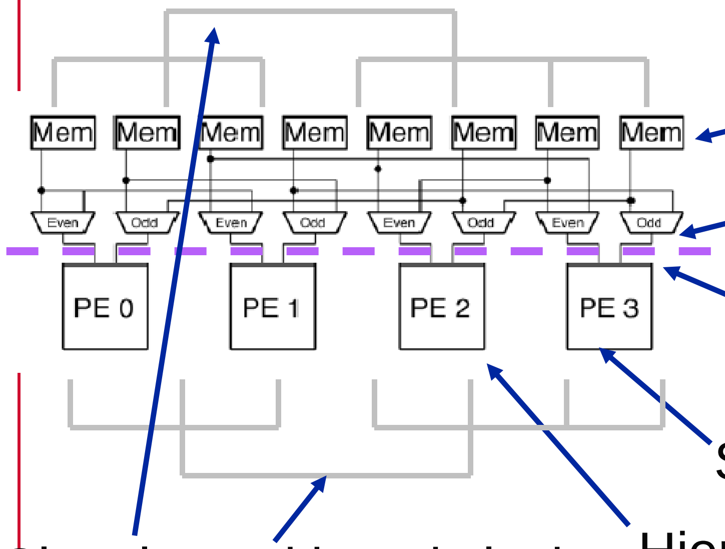
5 µm pad alignment

100 kg contact force

- **Logic die:**
  - Need Known Good Die solution with compact test set
- **Memory stack:**
  - Need yield management and Known Good Die solution

Design For Test Sub-flow

Partitioning choices and DFT planning dramatically impact ability and cost to achieve this



Directed pattern speed scan testing

Shadow DFT scan registers

TSV integrity self-test

Registered IO to permit scan testing

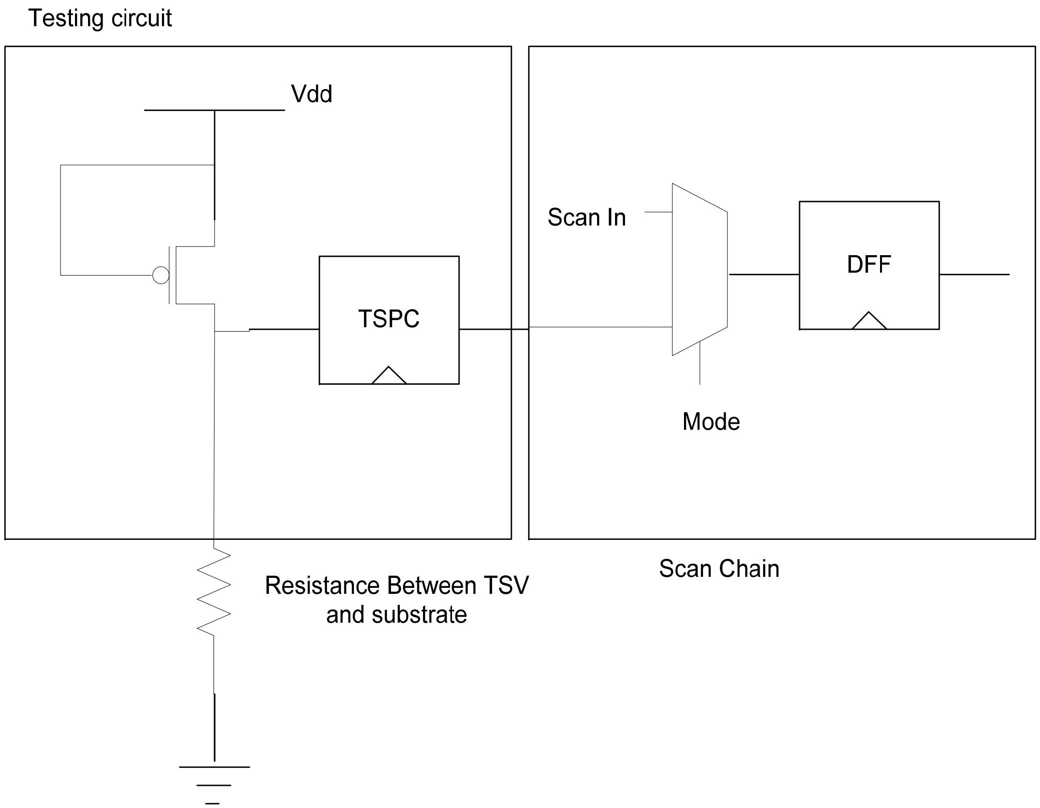
Scan-based speed sorting

Hierarchical test

Cleanly partitioned clocks

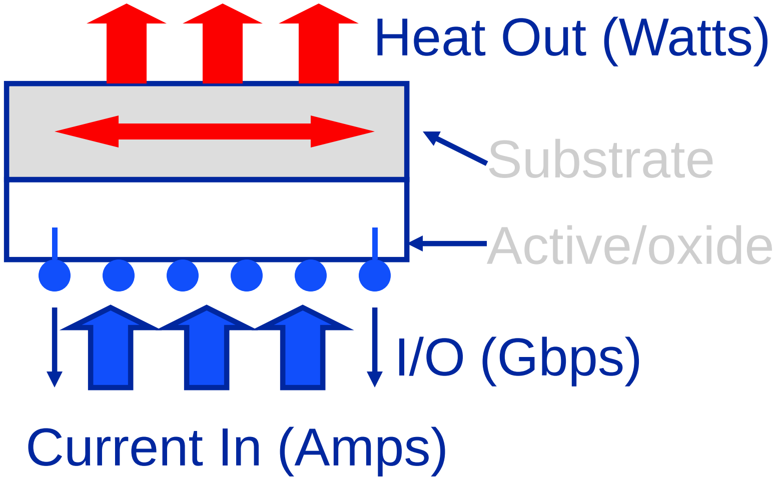
TSV Self-Test

1. Self-test for leakage easy to implement
2. Gives 1/0 answer for read-out via scan chain

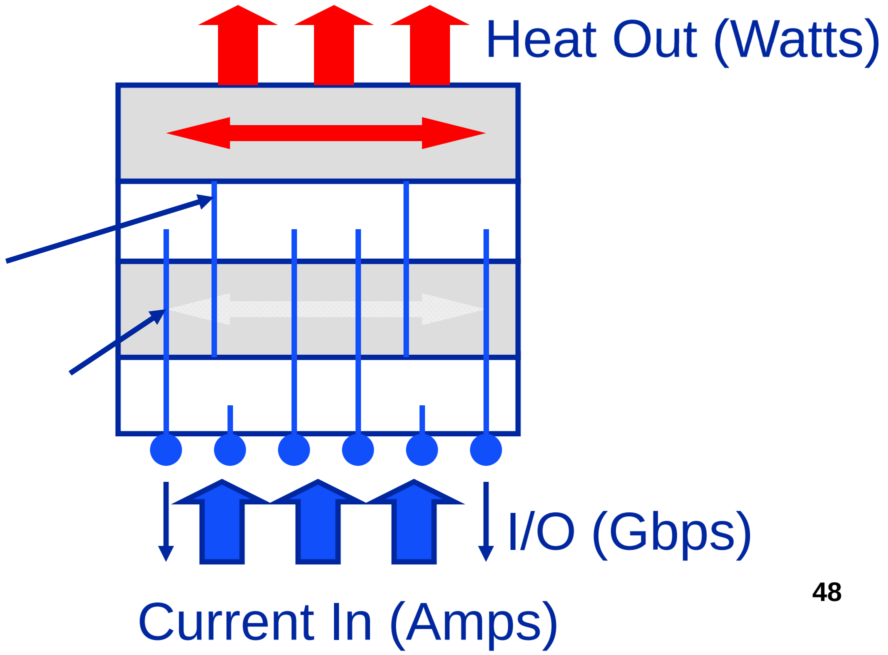


Power delivery, I/O and thermal

1. **2D chip:**
   - Heat spreader next to heat source
   - Short Idd Iss wires
   - Short I/O wires over oxide



2. **3D chip:**
   - Bottom side power and signal delivery
   - Top-side heat dissipation
   - Through TSVs needed for thermal dissipation
   - Through TSVs increase LCR of Vdd, Gnd and IO


Heat Out (Watts)
I/O (Gbps)
Current In (Amps)