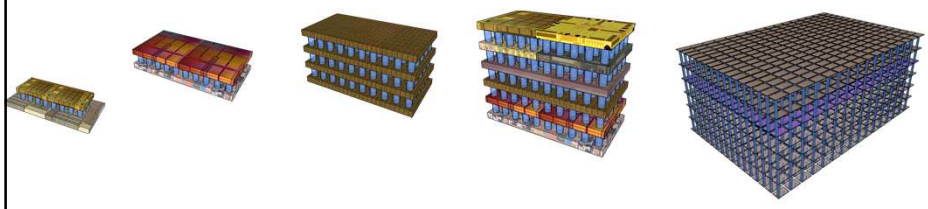
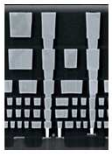


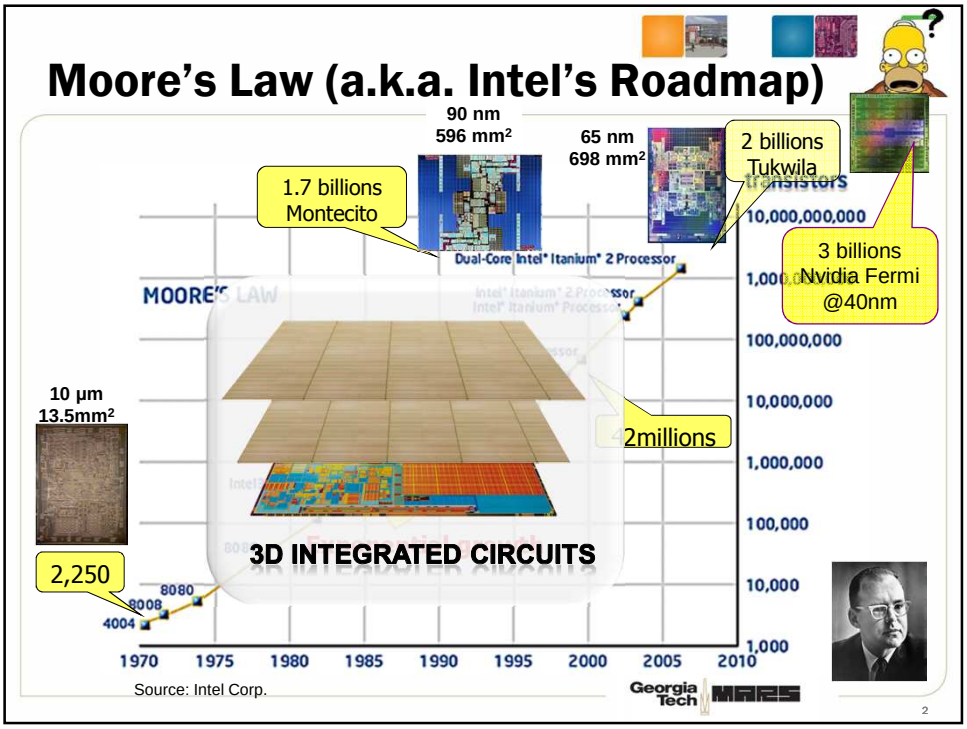
Design, Implementation, and Test for a 3D-IC Many-Core Processor



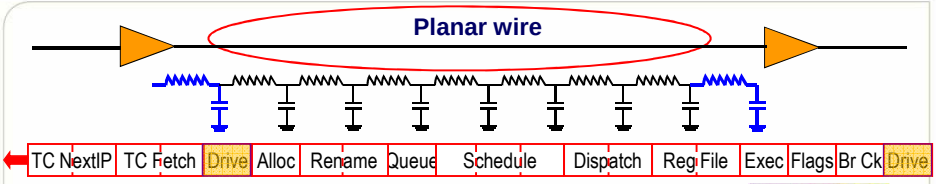
Hsien-Hsin S. Lee
Electrical and Computer Engineering
Georgia Tech

IMPACT 2010, October 21

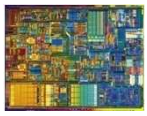
Moore's Law (a.k.a. Intel's Roadmap)

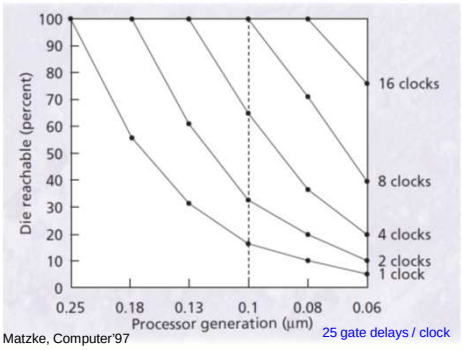


Wire Delay and 3D Integration 101



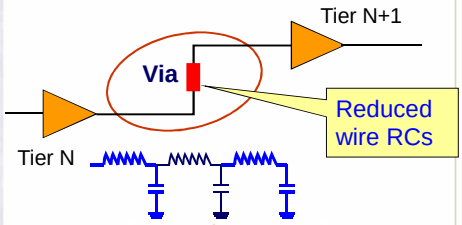
NetBurst (P4) Backend Pipeline (Willamette)





Matzke, Computer'97

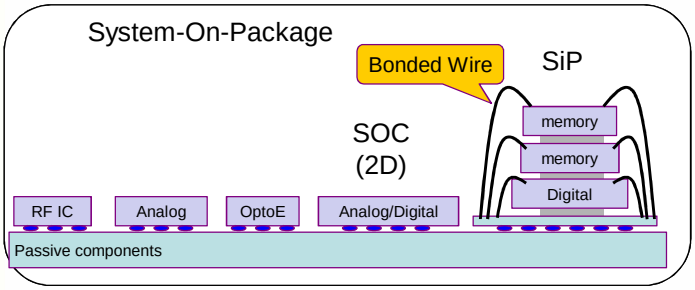
25 gate delays / clock



Georgia Tech MARS

3

State of the Art vs. Emerging 3D ICs



iPad

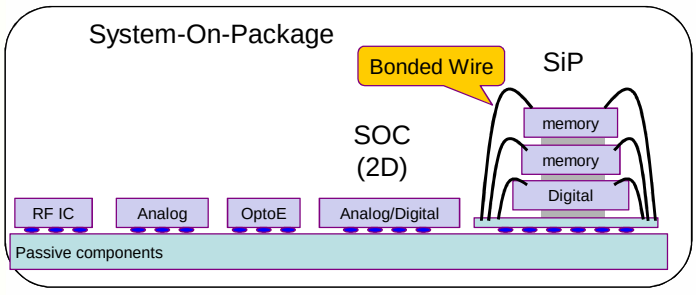
Georgia Tech MARS

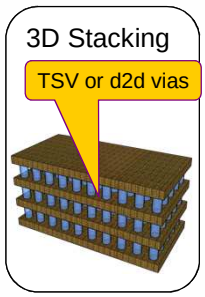
4

State of the Art vs. Emerging 3D ICs

Wire,
wire,
wire

- Length (latency and power)
- Density
- I/O bond (power)





Overhang: 1 to 2 mm	F2F vias: ~3-10µm (< 1 F04) TSV: ~5-50µm
Wirebond Pitch: 40-60 µm	TSV Pitch: 1 to 10 µm
Wires around boundary	TSVs on the entire surface

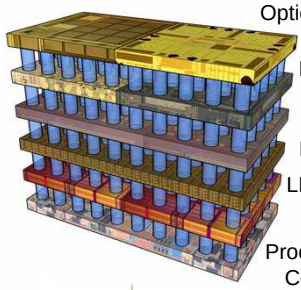
Georgia Tech MARS

5

Heterogeneous Stacking

- Heterogeneous Stacking (Type A)
  - Microprocessor
  - ICN / NoC
  - Analog
  - Power Regulator
  - DSP
  - Memory
  - RF IC
  - (Optical) I/O

- Heterogeneous Stacking (Type B)
  - 45nm
  - 90 nm
- Smaller SoC form factor
- Economy of Scale

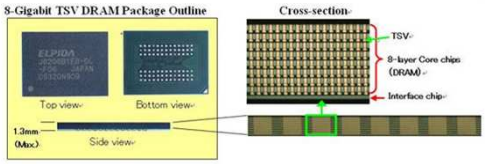


Georgia Tech MARS

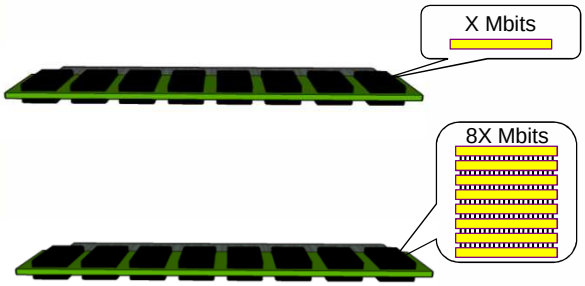
6

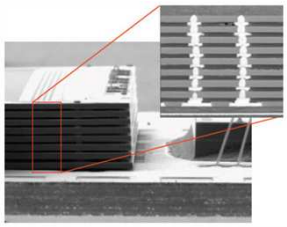
Homogeneous Component-Level Stacking

- Stacked Memory Tiers
  - DRAM, SRAM, NVM
- Capacity-driven
- Cost reduction



Elpida's 8-layer DRAM
(8Gbits, Cu-TSV)





Samsung WSP
8-tier 16Gb NAND flash

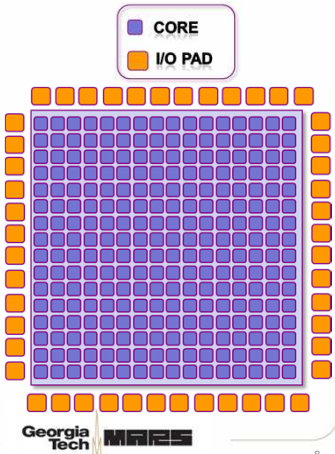
- High Device Density

Georgia Tech MARS

7

Bandwidth Wall and Power Wall

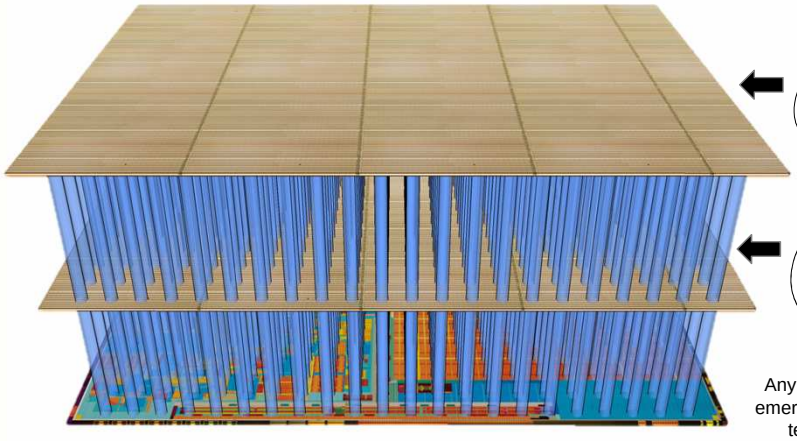
- I/O power — ITRS predicts slow growth in pin count
  - 2/3 for **Power and ground**, 1/3 for **Signal I/O**
  - Limited by physical metal properties and cost
    - DDR3 ~ 40mW per pin
    - **1024** data pins ~ **40W**
    - **4096** data pins ~ **160W**
- Techniques to avoiding I/O pad power
  - Memory Locality optimization
  - Very tight Integration
    - Intel: Cores+MCH+ICH → Cores+PCH
    - 3D integrated circuits

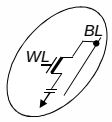


Georgia Tech MARS

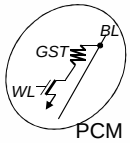
8

High-Bandwidth Memory-Stacked Processor





Or



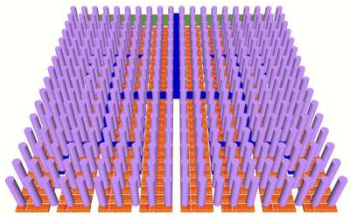
PCM

Or

Any high-density emerging memory technology

Via Bandwidth Potential of 3D-IC

- High 3D via frequency
- High 3D via bandwidth
- Dense 3D via provides many independent channels
- Short 3D via improve
  - Latency
  - Power
  - For both signal and clock
- 1 ~ 10s TB/sec or even higher

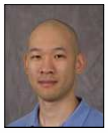


The Georgia Tech 3D-MAPS Team

- Faculty
  - Hsien-Hsin S. Lee, Sung Kyu Lim, Gabriel H. Loh
- Students
  - Architecture, Design and Test team (5): Mohammad Hossain, Dean Lewis, Tzu-Wei Lin, Guanhao Shen, Dong Hyuk Woo
  - CAD team (11): Krit Athikulwongse, Rohan Goel, Michael Healy, Moongon Jung, Dae Hyun Kim, Young-Joon Lee, Chang Liu, Brian Ouellette, Mohit Pathak, Hemant Sane, Xin Zhao
- Collaborators
  - Package/board design: Dr. Daehyun Chung (GT), Prof. Joungho Kim (KAIST), Prof. Madhavan Swaminathan (GT)

LEE, Core/Arch

LIM, CAD Tool



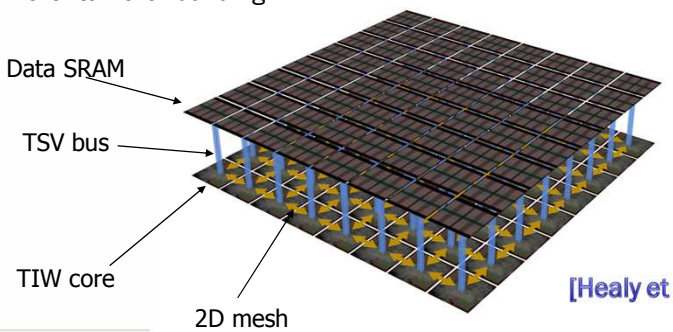
LOH, Memory

Georgia Tech MARS

11/71

3D MAPS V1: 64-Core With Stacked SRAM

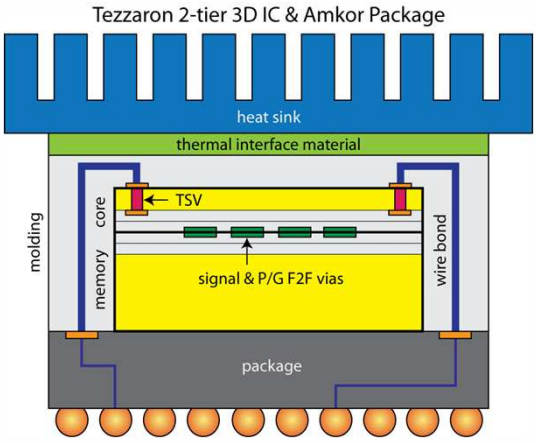
- DARPA MPW Run
  - Chartered + Tezzaron's 3D Process
  - Artisan Library / IP
- Face-to-Face vias
  - Tungsten filling
  - About 1 million vias if fully utilized for our given area
- Wafer-to-wafer bonding



[Healy et al., CICC-2010]

12/71

3D MAPS V1 Package



Core-tier
- thinned to 12um
- TSV height becomes 6um
- closer to heat sink
- talk to package via "wire-bond"
- requires dummy TSVs for density

Memory-tier
- thickness is 765um
- requires dummy TSVs for density

Georgia Tech MAPS

13/71

Architecture Mission for 3D-MAPS Processor

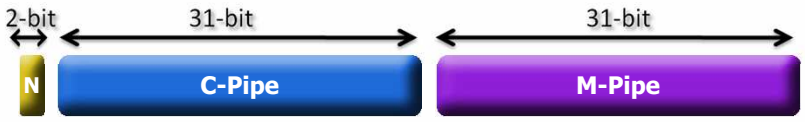
- Keep hardware complexity low
  - Small area (5mm x 5mm die)
  - More processing cores for given die budget
  - Higher power efficiency
  - Higher operating frequency (for .13µm)
  - Very tight design timeframe
  - Small design team, worse, inexperienced graduate students
- Shift complexity to the Software
  - Inexpensive, savvy programmers
  - Intelligent assemblers, compilers

Georgia Tech MAPS

14/71

TIW Custom ISA

- *Two Instruction Word* per core, per cycle
- Total 64 cores in 3D-MAPS v1



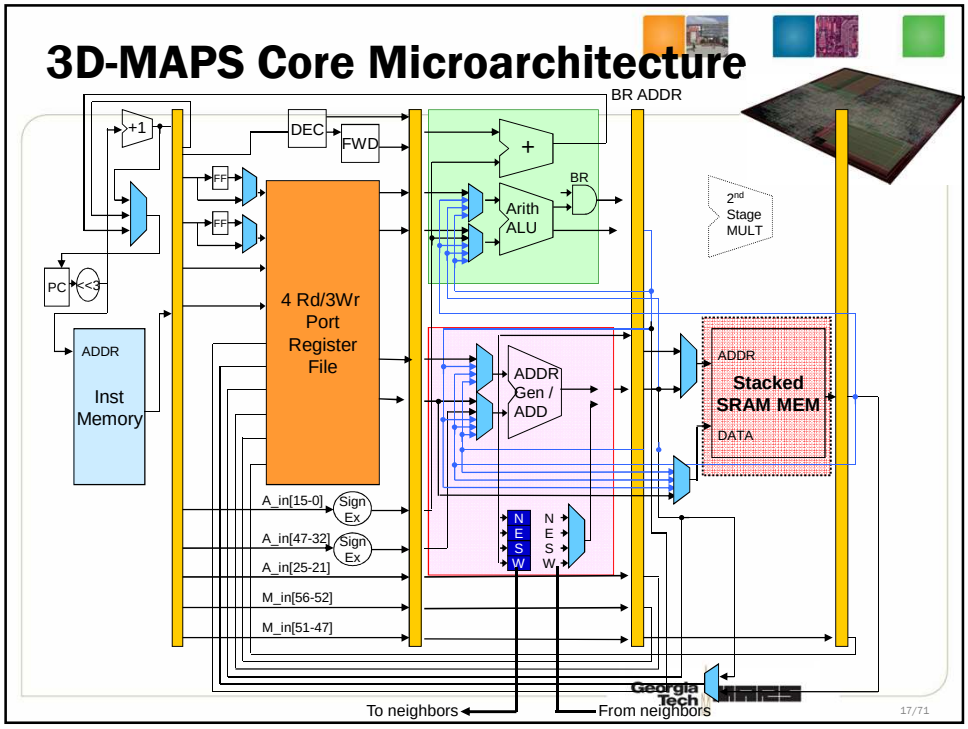
INT Arithmetic instructions
Branch instructions

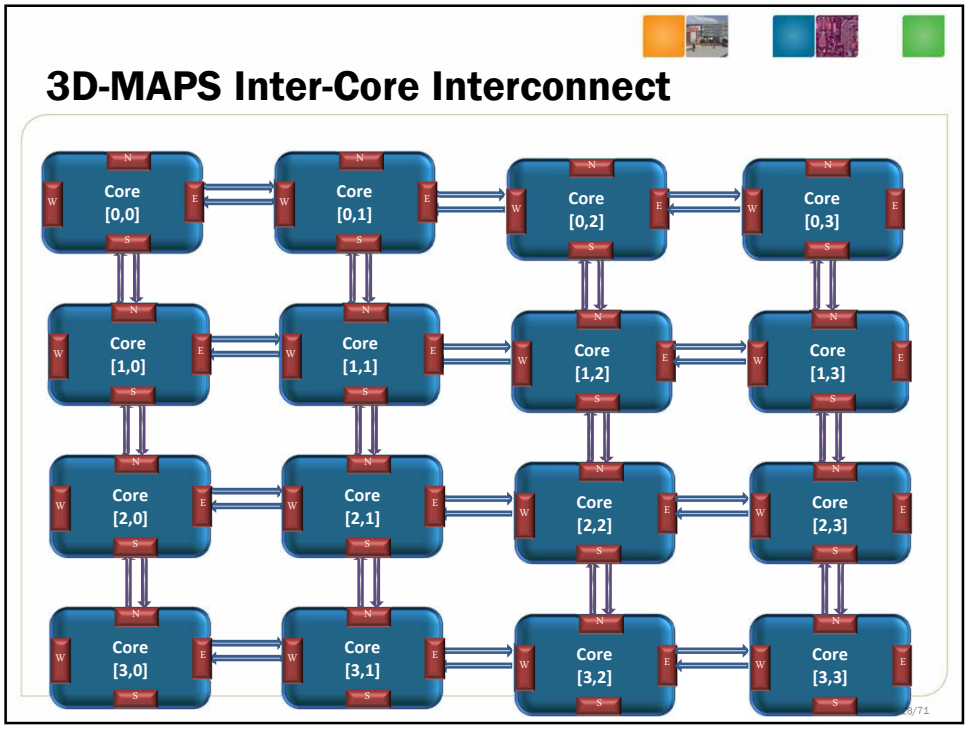
Memory instructions
Communication instructions
ADD / SUB / ADDI

NOP Compression

C-Pipe	M-Pipe
BEQ $r7, $r0, label	ADD $r3, $r1, $r2
NOP	NOP
NOP	NOP
ADD $r4, $r7, $r2	LW_I $r9, $r3, 0
SHL $r4, $r4, 2	ADDI $r8, $r2, 1

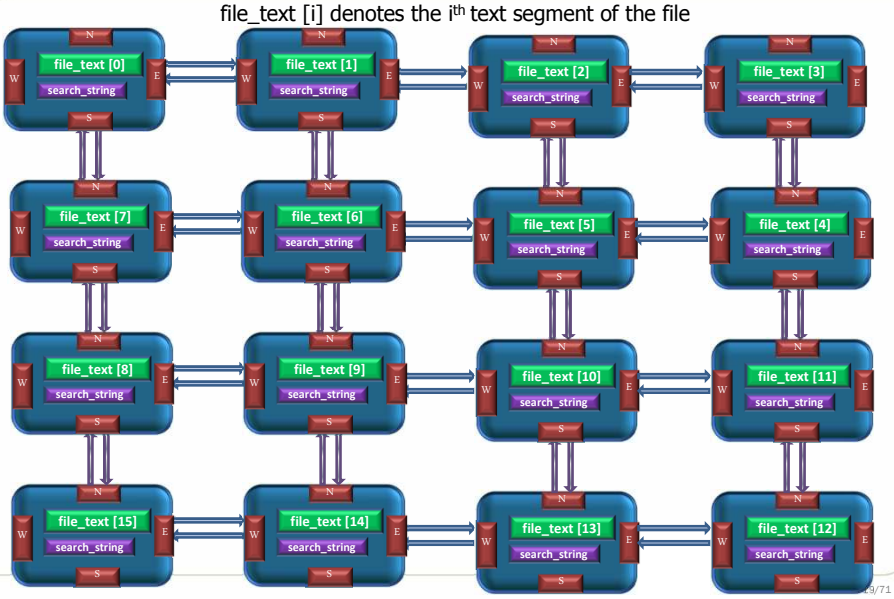
10 BEQ $r7, $r0, label	ADD $r3, $r1, $r2
00 ADD $r4, $r7, $r2	LW_I $r9, $r3, 0
00 SHL $r4, $r4, 2	ADDI $r8, $r2, 1


3D-MAPS Core Microarchitecture
BR ADDR
+1
DEC
FWD
FF
FF
PC
<<3
ADDR
Inst Memory
4 Rd/3Wr Port Register File
+
Arith ALU
BR
2nd Stage MULT
ADDR Gen / ADD
ADDR
Stacked SRAM MEM
DATA
A_in[15-0]
Sign Ex
A_in[47-32]
Sign Ex
A_in[25-21]
M_in[56-52]
M_in[51-47]
N E S W
N E S W
To neighbors
From neighbors
17/71

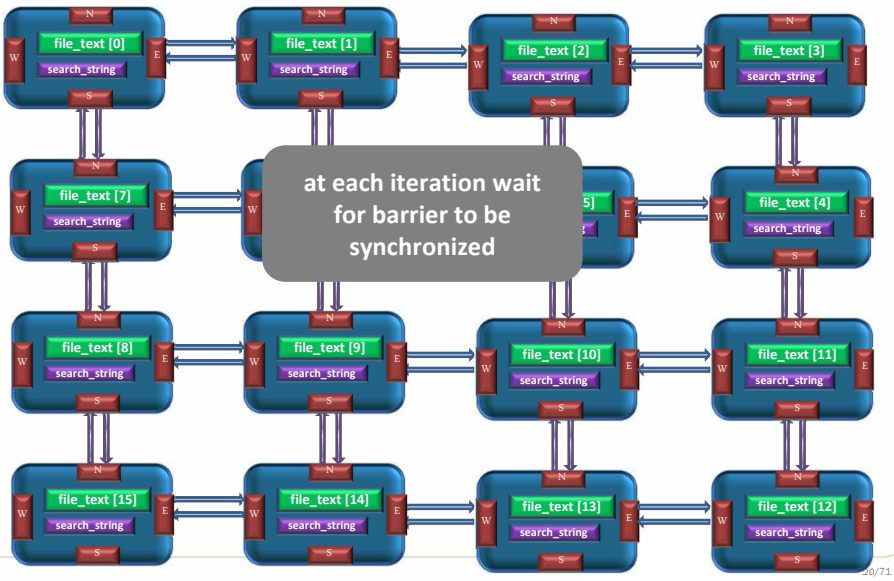

3D-MAPS Inter-Core Interconnect
Core [0,0]
Core [0,1]
Core [0,2]
Core [0,3]
Core [1,0]
Core [1,1]
Core [1,2]
Core [1,3]
Core [2,0]
Core [2,1]
Core [2,2]
Core [2,3]
Core [3,0]
Core [3,1]
Core [3,2]
Core [3,3]

Local Data Partitioning

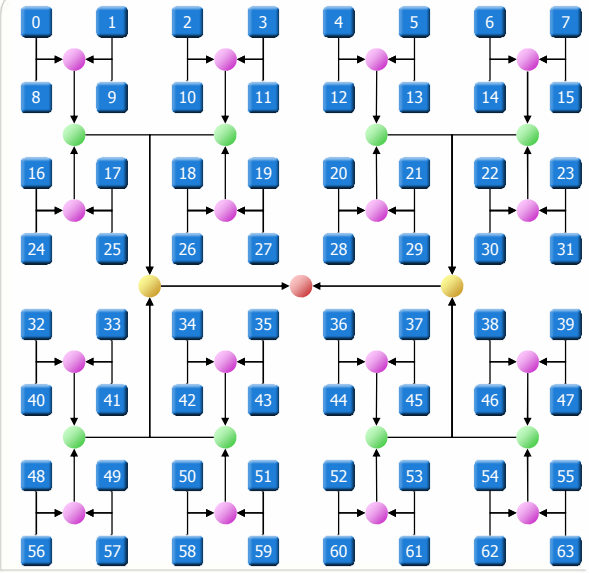
file_text [i] denotes the i^{th} text segment of the file



Barrier Synchronization



Global Barrier

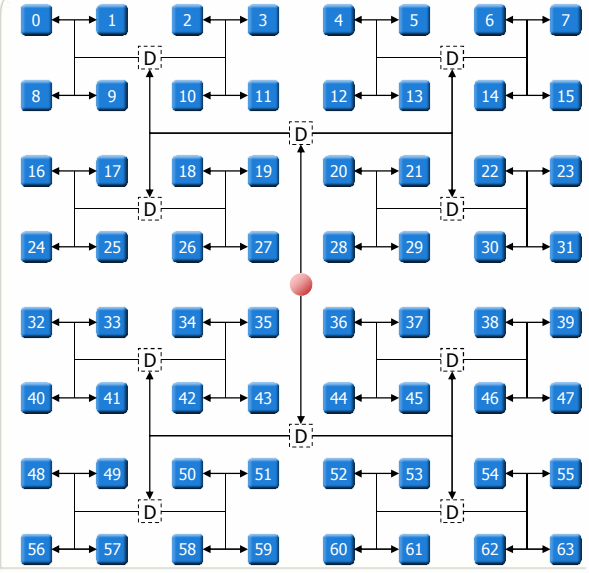


- 4-input AND
- 2-input AND & flipflop
- 4-input AND & flipflop
- 2-input AND & flipflop

Georgia Tech MARS

21/71

Global Barrier

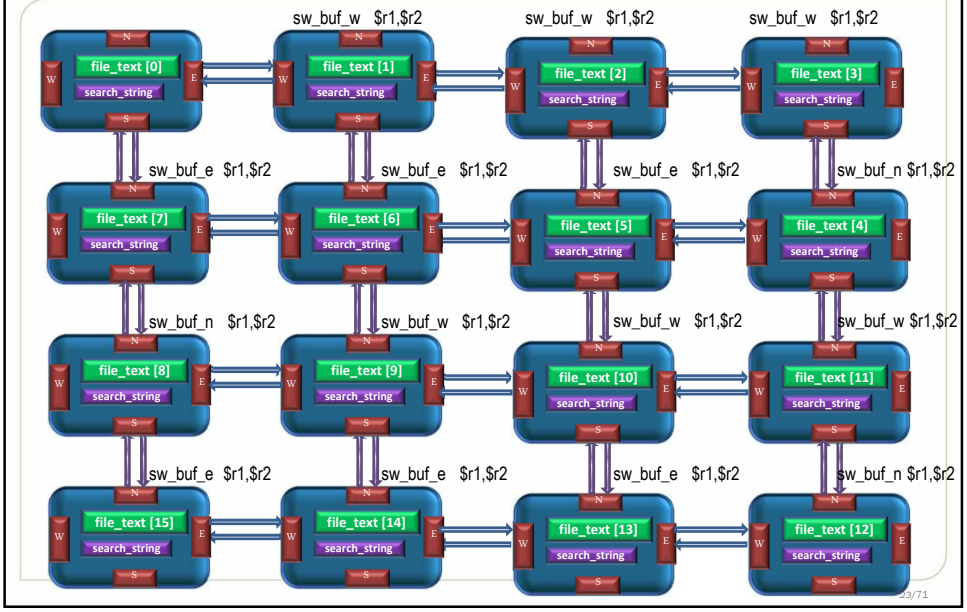


- D Flipflop
- 2-input AND & flipflop

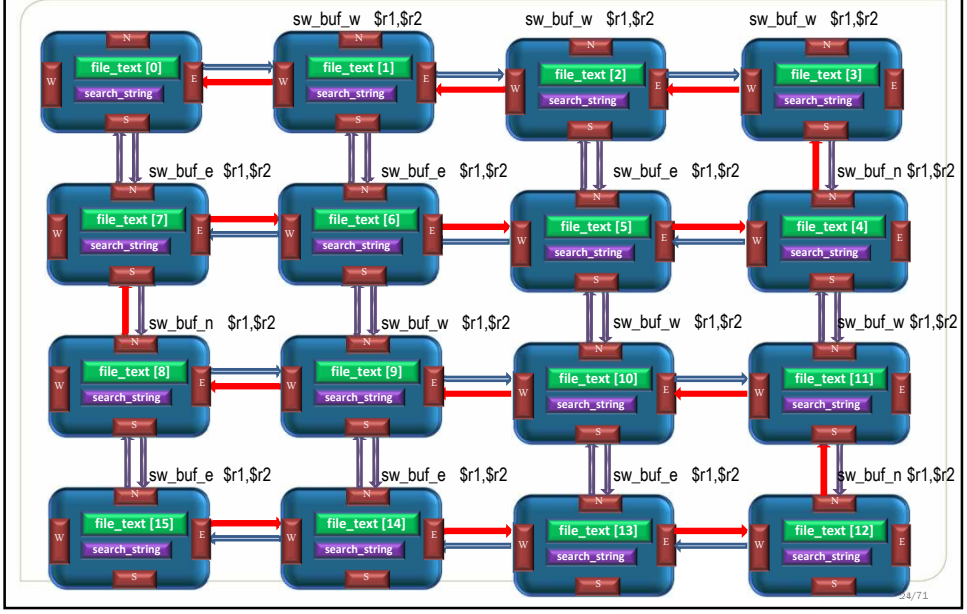
Georgia Tech MARS

22/71

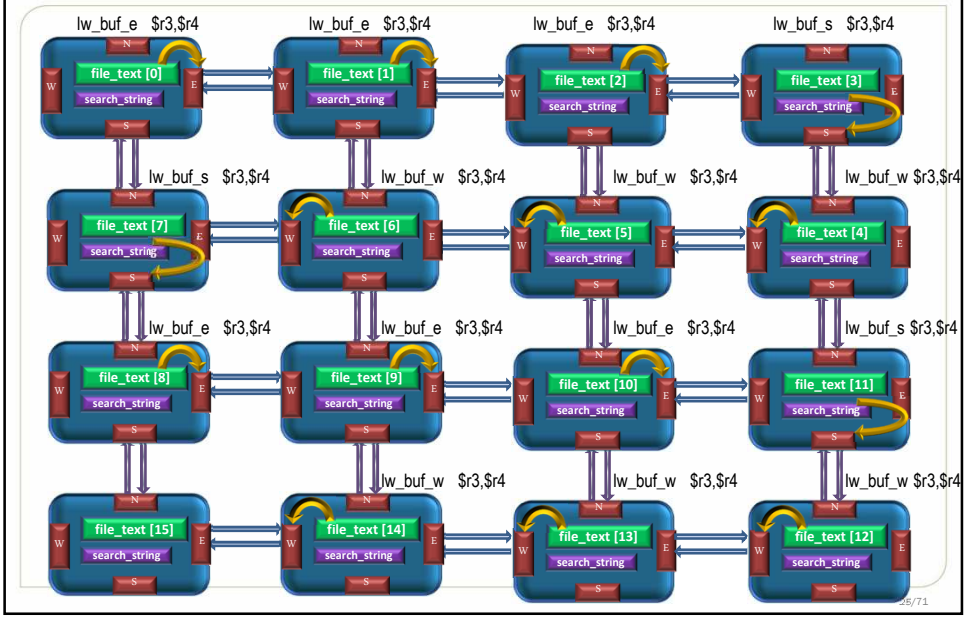
Data Communication: Writing to Neighbor's Buffer



Data Communication: Writing to Neighbor's Buffer



Data Communication: Read from Own Buffer



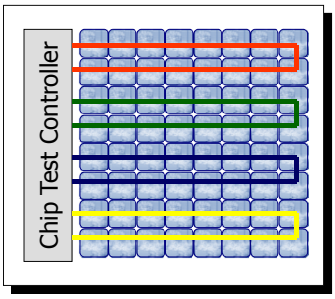
Achievable Data Bandwidth from Stacked SRAM

Benchmark	Theoretical BW	@64, 277MHz Cores (GBps)	Area efficiency Vs. Intel i7
String Search	0.50*Nf*	8.9	4.1x
Matrix Multiplication	0.78*Nf*	13.8	6.5x
AES Encryption	2.79*Nf*	49.5	23.1x
Histogram	1.71*Nf*	30.3	14.2x
Sobel Edge Detector	0.88*Nf*	15.6	7.3x
K-Means	2.29*Nf*	40.6	19.0x
Median Filter	3.60*Nf*	63.8	29.8x
Motion Estimation	1.36*Nf*	24.1	11.3x

- Evaluated based "bandwidth density per mm^2"
- Intel's Glutown i7 @32nm vs. our 3D-MAPS @130nm

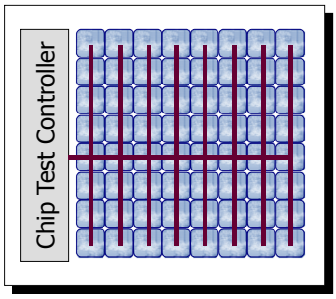
Test Mode and Programming Mode

- Test mode



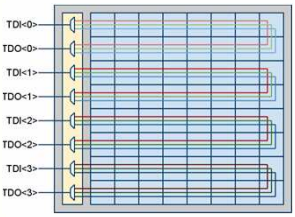
  - Used for scan-based test
  - Scan bus is 1 bit wide
  - Extra control wires will be necessary

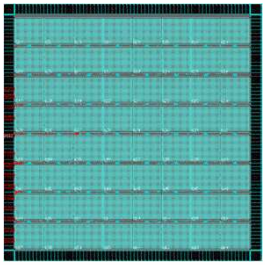
- Programming mode



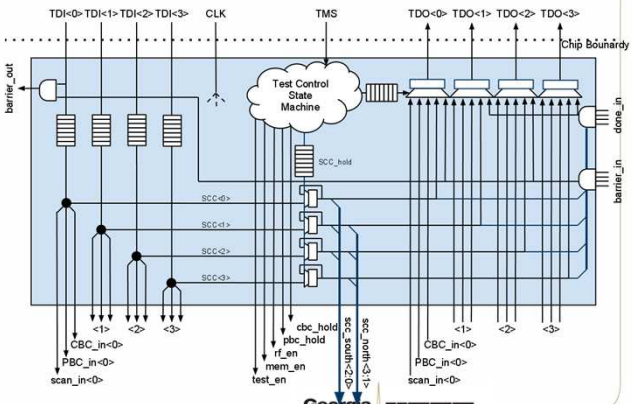
  - Used to load instruction and data memory
  - Scan bus is 4-bit wide

Test State Machine

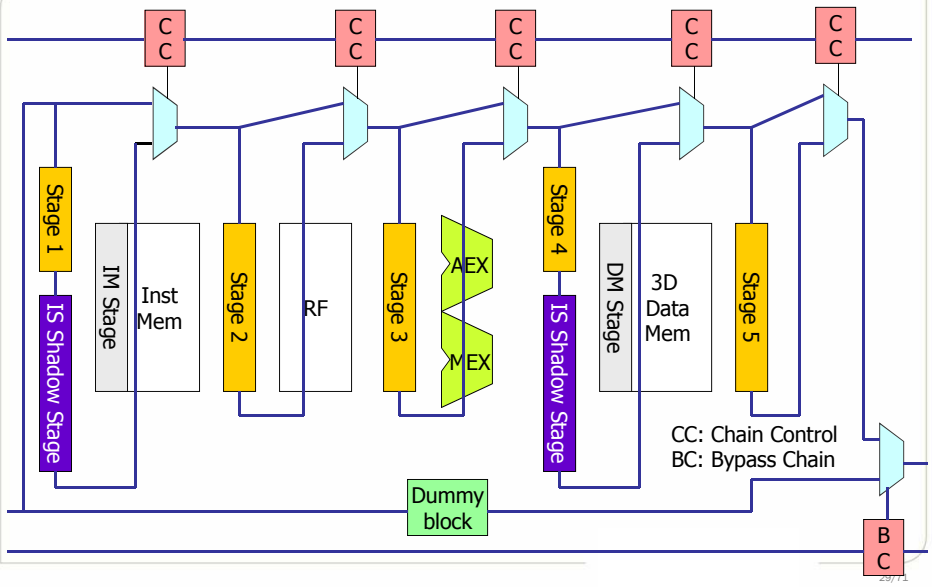




- **64 cores split into 4 sectors, tested independently**
- **Scan I/O pins located on one side**
- **Testing circuitry sitting in between the cores**



Scan Path and Memory Shadow Registers



3D MAPS V1: 64-Core With Stacked SRAM

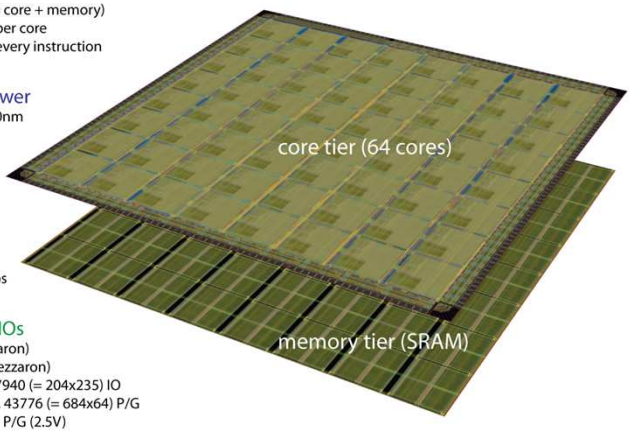
Architecture and Memory Model
- number and type of cores: 64, 5-stage, in-order, 2-way VLIW
- memory capacity: 256KB SRAM
- 3D stacking: 2 tiers face-to-face bonded (= core + memory)
- memory model: dedicated 4KB SRAM tile per core
- memory latency: 1 clock cycle, 1 read per every instruction

Technology, Performance, and Power
- technology: Chartered Semiconductor 130nm
- footprint area: 5mm x 5mm
- clock frequency: 277MHz
- operating voltage: 1.5V
- maximum power consumption: up to 6

Reliability
- maximum IR-drop: up to 78mV
- maximum coupling noise: 574 mV
- clock skew/slew: skew = 82ps, slew = 117ps
- maximum temperature: coming up

TSVs, Face-to-face (F2F) Vias, and IOs
- TSV diameter and pitch: 1.2um, 5um (Tezzaron)
- F2F via diameter and pitch: 3.4um, 5um (Tezzaron)
- total TSV count: 2240 (= 35x64) dummy, 27940 (= 204x235) IO
- total F2F via count: 7424 (= 116x64) signal, 43776 (= 684x64) P/G
- total IO count: 14 signal, 205 P/G (1.5V), 16 P/G (2.5V)

arguably the FIRST many-core 3D processor from academia
- designed to demonstrate memory BW/power benefit of 3D processor

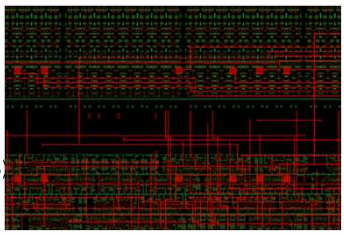


Georgia Tech MARS

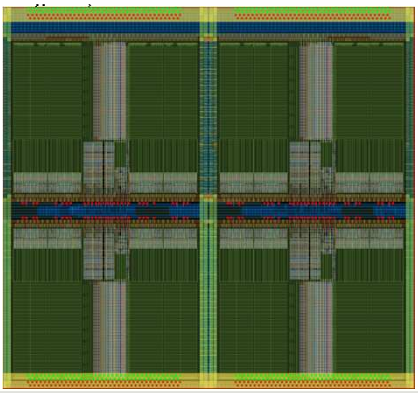
30/71

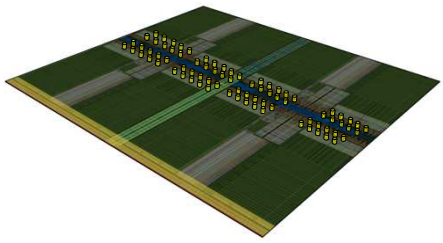
Single D-MEM Tile Layout

- F2F via: 3.4 µm diameter, 5 µ m pitch
- 7424 (116x64) F2F vias: (central located)
  - 32b data in + 32b out
  - 10b address
  - Other controls
- 43776 (684x64) P/G F2F vias (on 2 borders)



Signal face-to-face vias

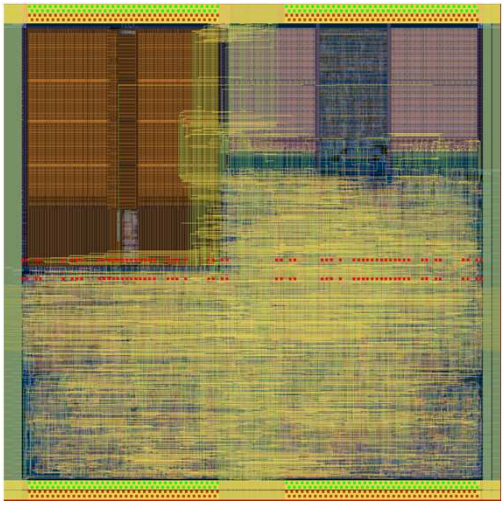




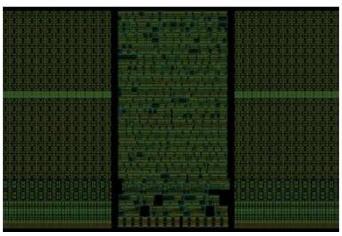
Georgia Tech MAPS

31/71

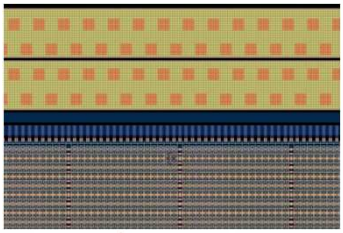
3D-MAPS Single Core Layout



Single core with face-to-face signal and P/G vias



Custom Reg file , 4-read/3-write per cycle



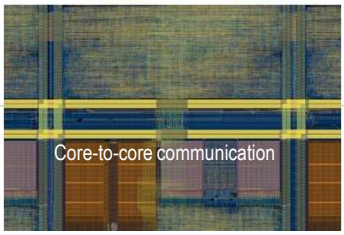
P/G face-to-face vias

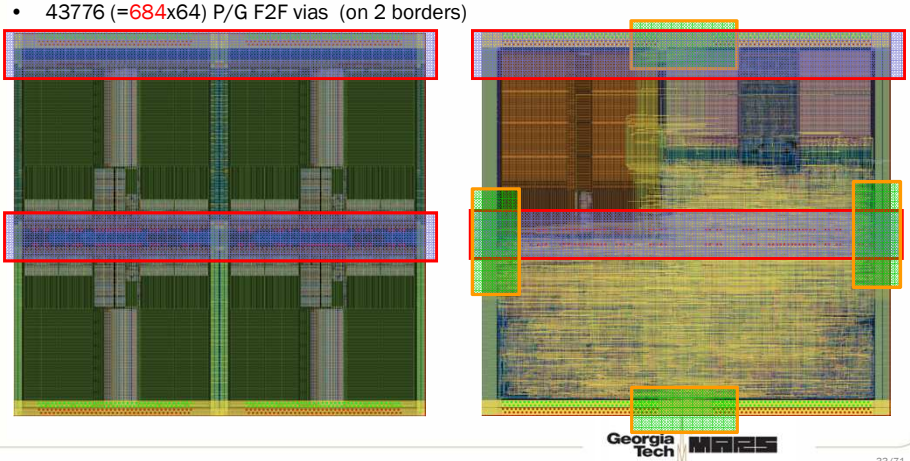
Georgia Tech MAPS

32/71

Single Core/Memory

- F2F via: 3.4 μm diameter, 5 μ m pitch
- 7424 (=116 x64) F2F vias: (central located)
  - 32b data in + 32b out
  - 10b address
  - Other controls
- 43776 (=684x64) P/G F2F vias (on 2 borders)



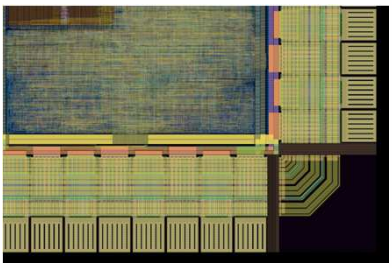


Georgia Tech MARS

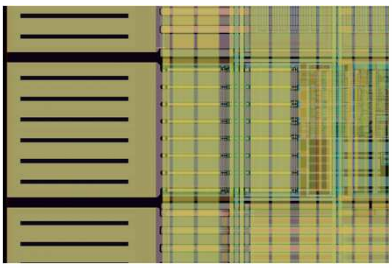
33/71

Through-Silicon Via (TSV)

- Spec: 1.2um diameter, 5um pitch
- Usage: mainly in I/O cells, some as dummy (for density control)
  - 2856 (= 14x204) for signal I/O, 45084 (= 221x204) for P/G IO
  - 3392 (= 53x64) for dummy



IO cells along the periphery

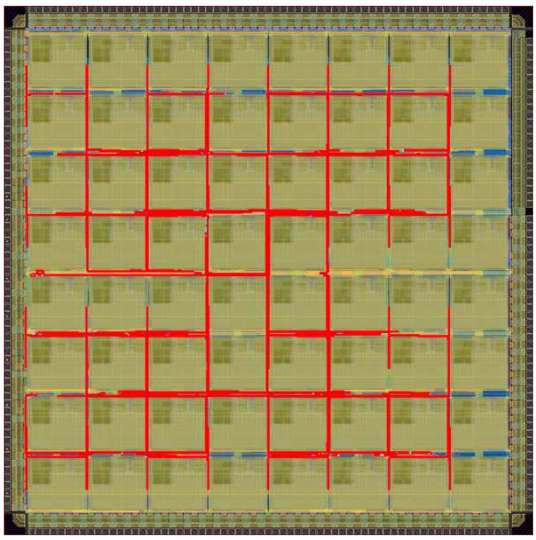


IO cell (zoom-in)

Georgia Tech MARS

34

Clock Distribution on 64-Core Tier



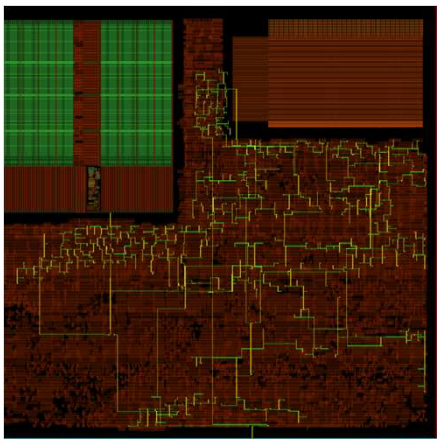
Georgia Tech MARS

35/71

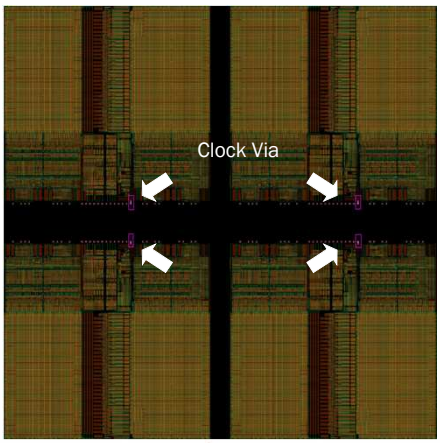
3D Clock Tree for Single Tile

- We provide clock to memory banks using F2F vias

Single Core Tile



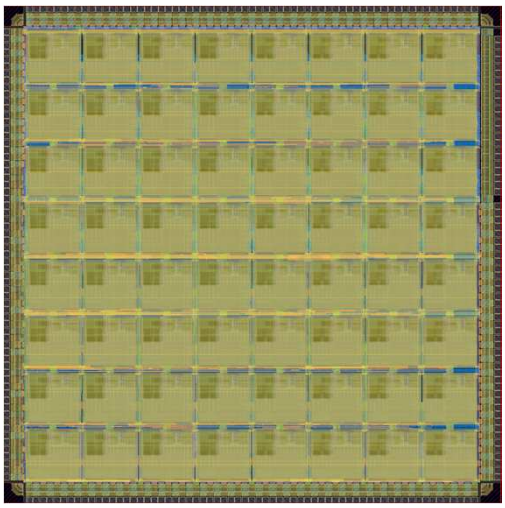
Single Memory Tile



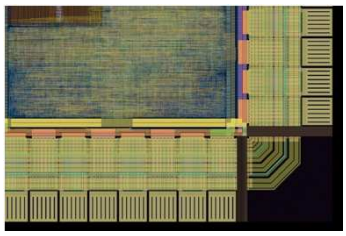
Georgia Tech MARS

36/71

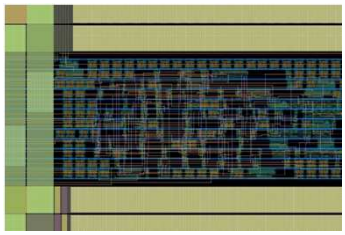
Processor Core Tier Layout



Overall layout of 64 cores + 235 I/O cells (on periphery)



I/O cells along the periphery

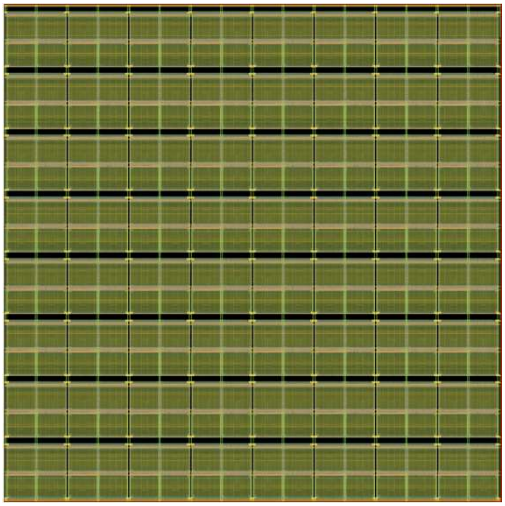


Testing circuitry

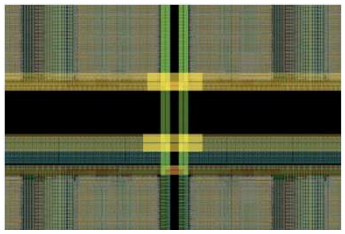
Georgia Tech MARS

37/71

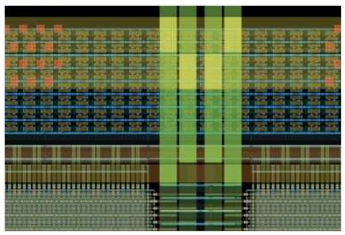
64 Tile SRAM Tier layout



Overall layout of 64 SRAM memory tiles (64 x 4KB)



P/G rings for SRAM tiles



Decaps attached to the P/G rings

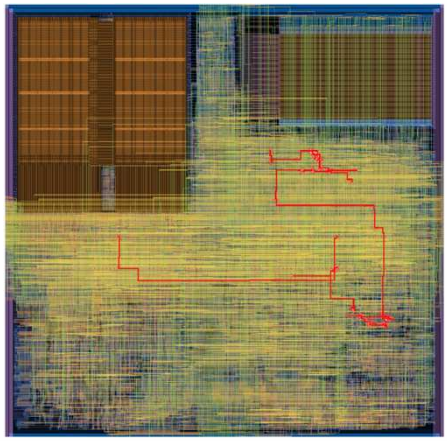
Georgia Tech MARS

38/71

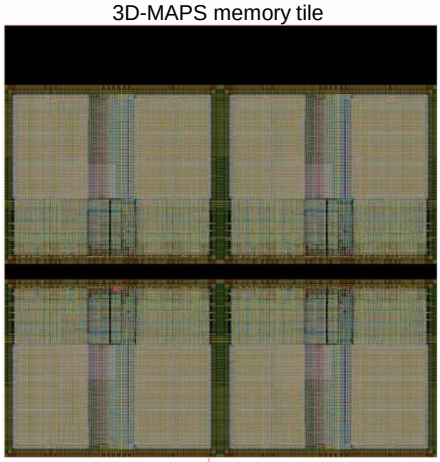
3D Timing Analysis

- Based on Synopsys PrimeTime: this 3D path has 3.6ns delay

3D-MAPS core tile



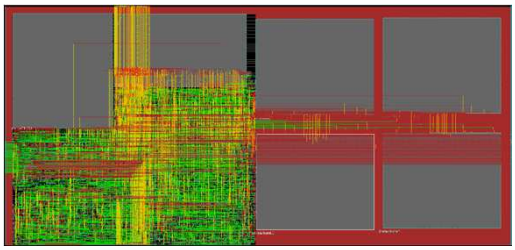
3D-MAPS memory tile



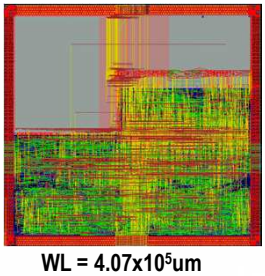
Georgia Tech MARS

39

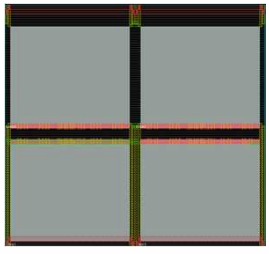
2D-MAPS vs. 3D-MAPS Comparison



WL = 4.4×10^5um



WL = 4.07×10^5um



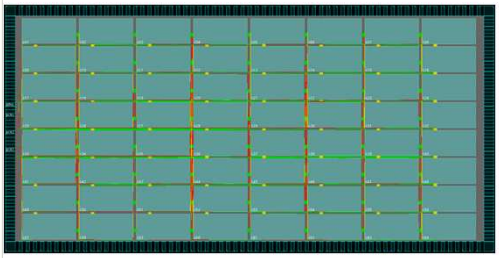
WL = 0.02×10^5um

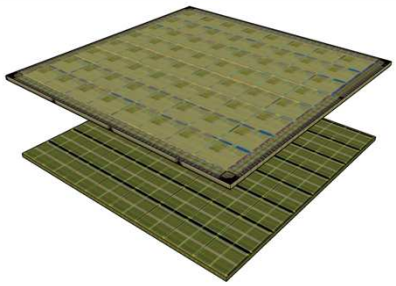
Georgia Tech MARS

40

2D-MAPS vs. 3D-MAPS Comparison

- Inter-core WL = 16.9×10^5um (2D) vs 9.19×10^5um (3D)
- Overall WL = 298.5×10^5um (2D) vs 270.1×10^5um (3D)





Georgia Tech MAPS

41

Fact Sheet of 64-Core 3D MAPS V1

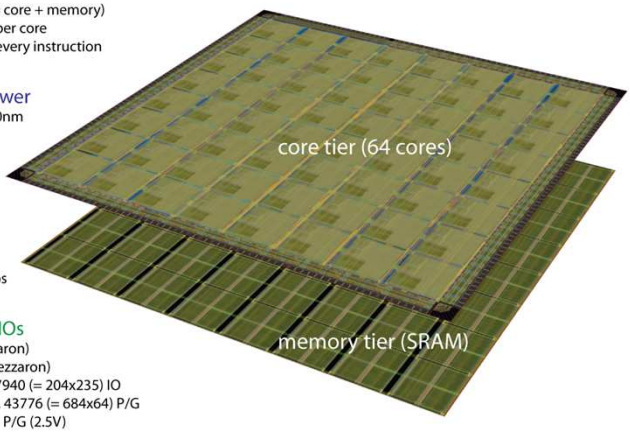
Architecture and Memory Model
- number and type of cores: 64, 5-stage, in-order, 2-way VLIW
- memory capacity: 256KB SRAM
- 3D stacking: 2 tiers face-to-face bonded (= core + memory)
- memory model: dedicated 4KB SRAM tile per core
- memory latency: 1 clock cycle, 1 read per every instruction

Technology, Performance, and Power
- technology: Chartered Semiconductor 130nm
- footprint area: 5mm x 5mm
- clock frequency: 277MHz
- operating voltage: 1.5V
- maximum power consumption: up to 6

Reliability
- maximum IR-drop: up to 78mV
- maximum coupling noise: 574 mV
- clock skew/slew: skew = 82ps, slew = 117ps
- maximum temperature: coming up

TSVs, Face-to-face (F2F) Vias, and IOs
- TSV diameter and pitch: 1.2um, 5um (Tezzaron)
- F2F via diameter and pitch: 3.4um, 5um (Tezzaron)
- total TSV count: 2240 (= 35x64) dummy, 27940 (= 204x235) IO
- total F2F via count: 7424 (= 116x64) signal, 43776 (= 684x64) P/G
- total IO count: 14 signal, 205 P/G (1.5V), 16 P/G (2.5V)

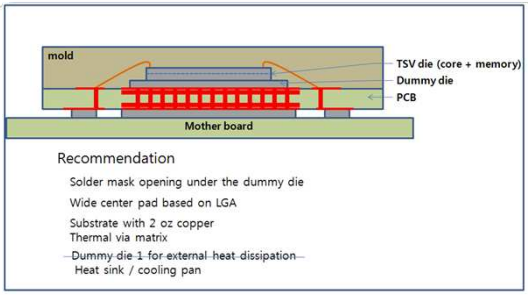
arguably the FIRST many-core 3D processor from academia
- designed to demonstrate memory BW/power benefit of 3D processor



Georgia Tech MAPS

42/71

Package Side View (Amkor)



Recommendation

Solder mask opening under the dummy die

Wide center pad based on LGA

Substrate with 2 oz copper
Thermal via matrix

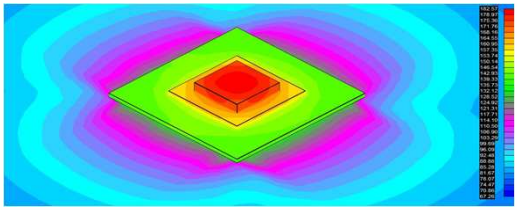
Dummy die 1 for external heat dissipation
Heat sink / cooling pan

Boundary Condition

Power (W)	6.32
Ambient Temperature (˚C)	50

Result

	LGA (JEDEC Board)	
Air Velocity (m/s)	θ_{JA} (˚C/W)	Die Temp. (˚C)
0	21	182.6
1	18.8	168.7
2.5	17.7	162



Isotherms of Theta JA

	Heatsink + JEDEC 1S2P		Heatsink + 8Layer Board		Heatsink + 12Layer Board	
Air velocity (m/s)	Max Die Temp (˚C)	θ_{JA} (˚C)	Max Die Temp (˚C)	θ_{JA} (˚C)	Max Die Temp (˚C)	θ_{JA} (˚C)
0	117.8	10.7	114.5	10.2	114.3	10.2
1	102.6	8.3	100.4	8.0	100.2	7.9
2.5	96.5	7.4	94.8	7.1	94.7	7.1

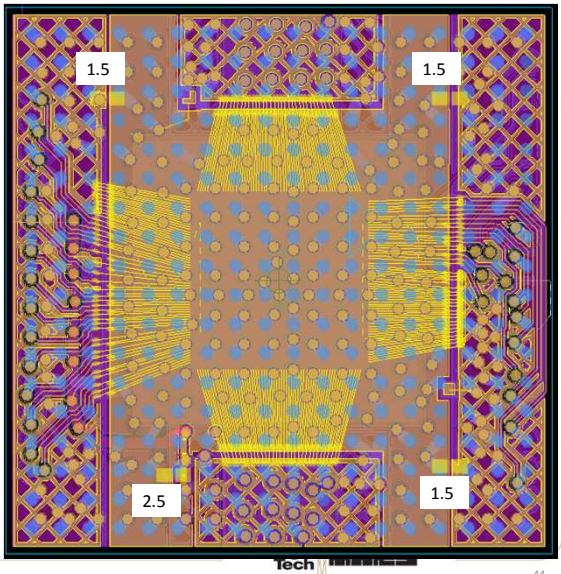
**JEDEC board: 1S2P + 36 thermal vias

Georgia Tech MARS

43

Package Overview (Amkor)

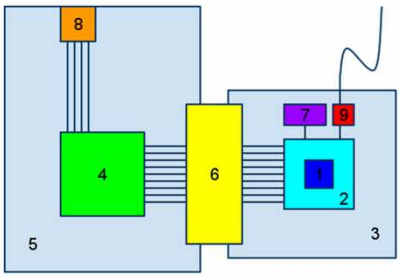
- Package-level decaps
  - 1.5V – 300 nF
  - 2.5V – 100 nF
- Board-level decaps
  - 1.5V – TBD (500 µF)
  - 2.5V – TBD (100 µF)

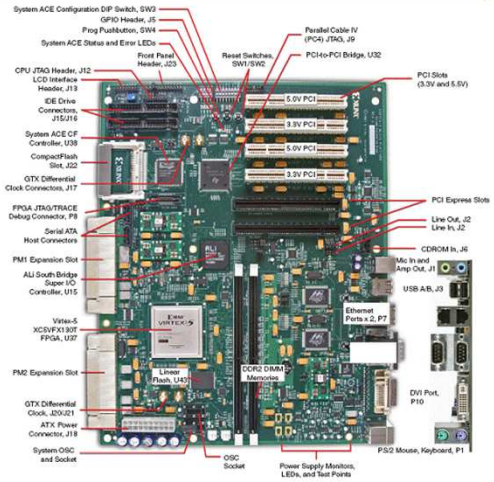


Tech MARS

44

Board Design





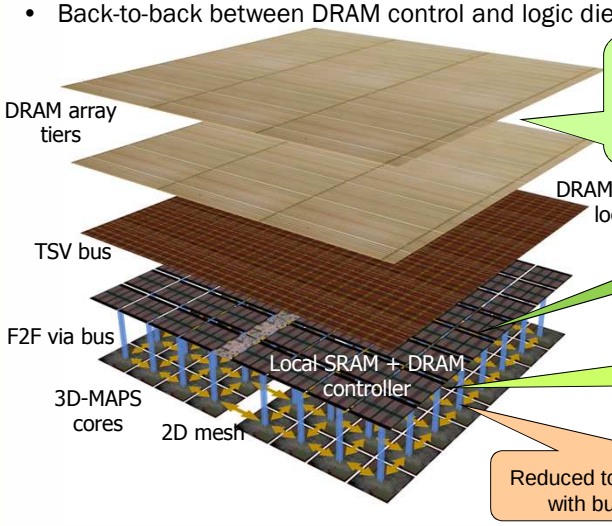
1. 3D-MAPS chip
2. BGA package
3. Daughter card
4. FPGA
5. FPGA development board (Xilinx ML510)
6. High-speed connector
7. PLL (200MHz - 1GHz continuous)
8. High capacity connector (e.g. USB, SATA, Ethernet, etc.)
9. Wall power supply for 1.5 DC rail.

Georgia Tech MARS

45

3D MAPS V2: Many-Core With Stacked SRAM + DRAM

- Chip-to-wafer bonding (no need to be the same size, not drawn to scale)
- Back-to-back between DRAM control and logic die



Optimized 3D DRAM design by moving peripheral logic to one separate die tier. (Tezzaron)

Four cores share one DRAM controller

Break up SRAM banks across 2 logic layers (2 back each layer)

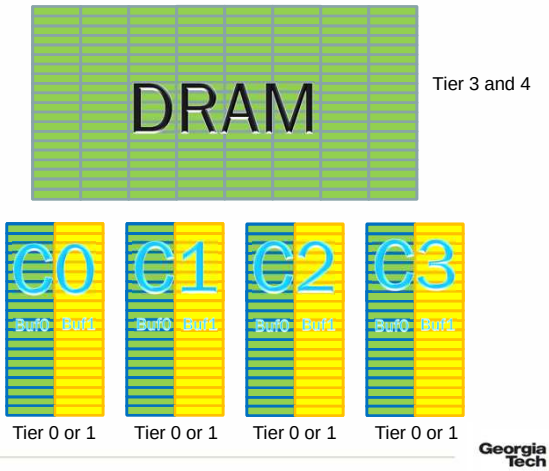
Reduced to **128** cores for 10x10 mm^2 with bump-down SRAM banks

Georgia Tech MARS

46/71

Double Buffering Memory Architecture

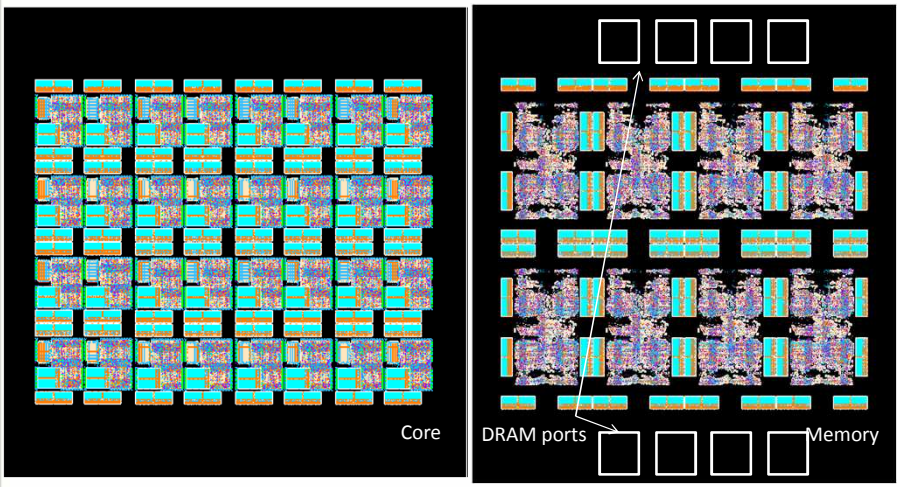
- Concurrent execution of computation and DMA streaming



Georgia Tech MARS

47

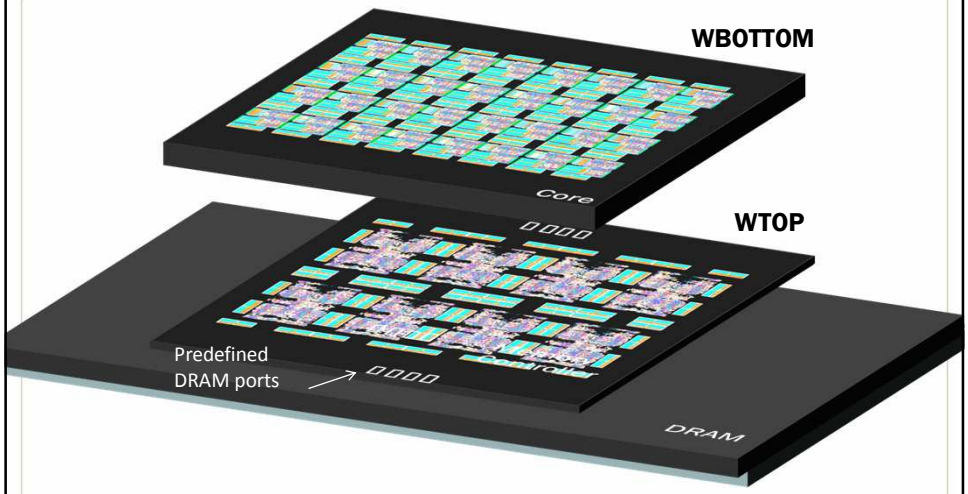
Layout of Core+Memory Layers (32-Core quad)



Georgia Tech MARS

48

3D MAPS V2 Stacking



Georgia Tech MARS

49

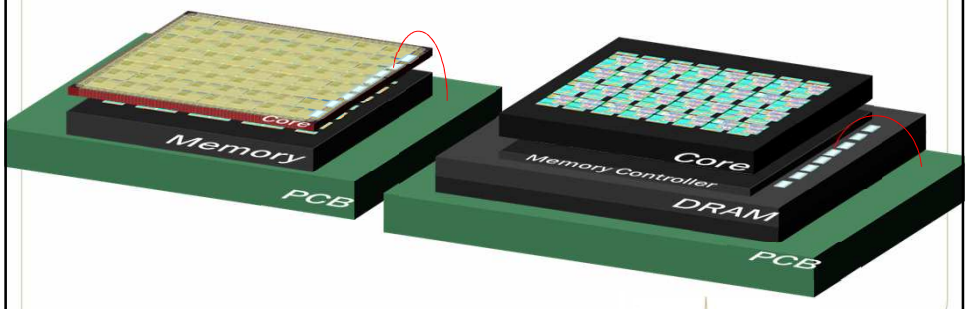
3D-MAPS: V1 versus V2 Stacking

V1

- Core die- thinner die
- Memory die- thicker die

V2

- Core die- thicker die
- Memory die- thinner die



Georgia Tech MARS

50

Rethinking 3D Memory Architecture

D. H. Woo, N. K. Seong, D. L. Lewis, H.-H. S. Lee,
"An Optimized 3D-Stacked Memory Architecture by Exploiting Excessive, High-Density TSV Bandwidth"
In **HPCA-16**, 2010

Georgia Tech MARS

51

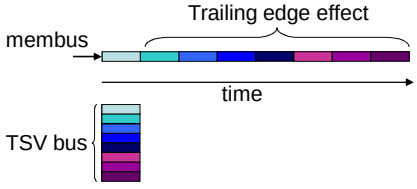
Revisiting Prior Architectural "False Truth"

- Common wisdom

  Bandwidth problems can be cured with money, latency problems are harder because the speed of light is fixed —you can't bribe God.

- We challenge this to ameliorate "latency" using bandwidth

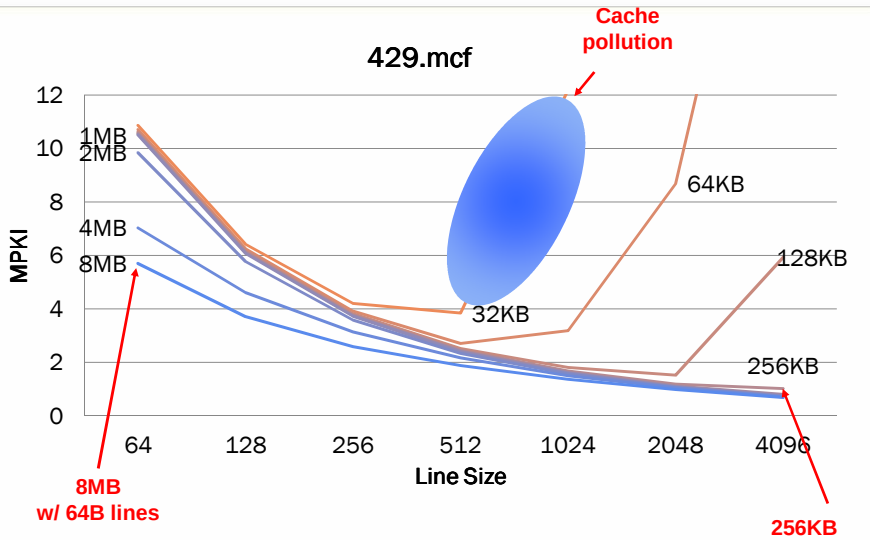
- TSV or F2F vias
  - Are fast (< 1FO4)
  - Are high density
  - Eliminate trailing-edge



Georgia Tech MARS

52

Revisiting Cache Line Size



SMART-3D: Optimized 3D Memory Architecture

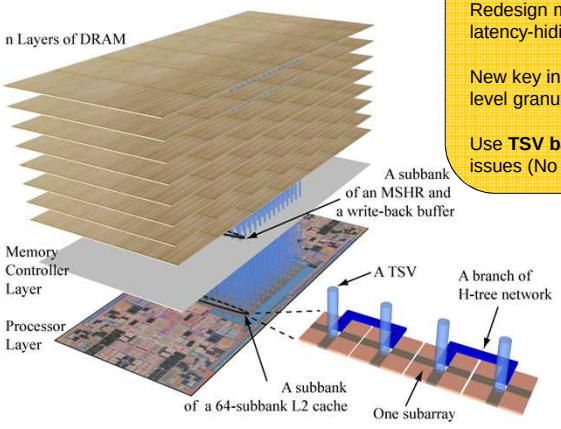
TSV could deliver 100s GB/sec or more bandwidth.

TSV are not fully utilized for single-thread apps.

Redesign memory hierarchy to use TSV to enable latency-hiding

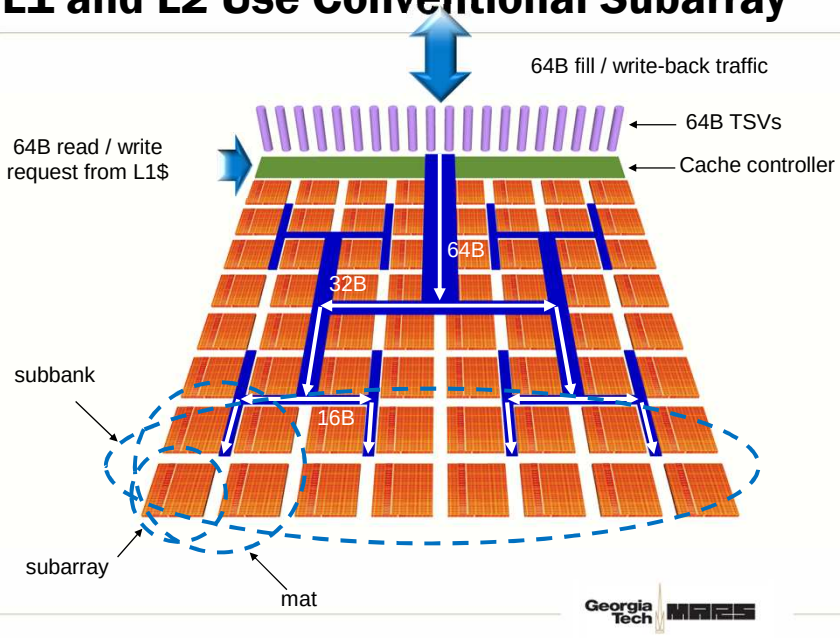
New key insight to fetch data into cache at page-level granularity (e.g., 4KB)

Use **TSV bandwidth** to ameliorate memory latency issues (No trailing-edge effect)

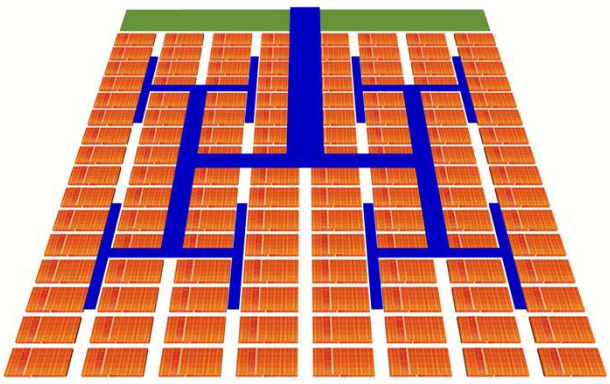


SMART-3D: Stacked Memory-Aware Rich TSV-enabled 3D (HPCA-16, 2010)

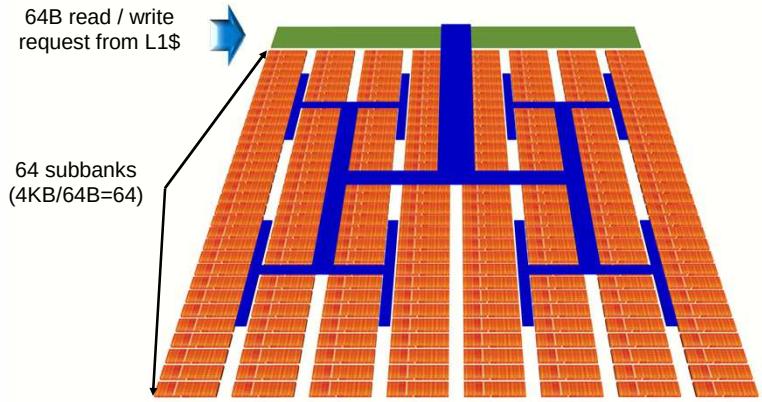
L1 and L2 Use Conventional Subarray



SMART-3D: L2 and 3D DRAM

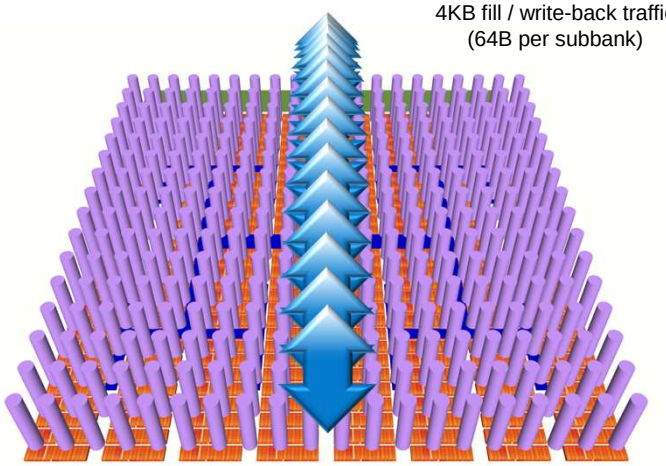


SMART-3D: L2 and 3D DRAM



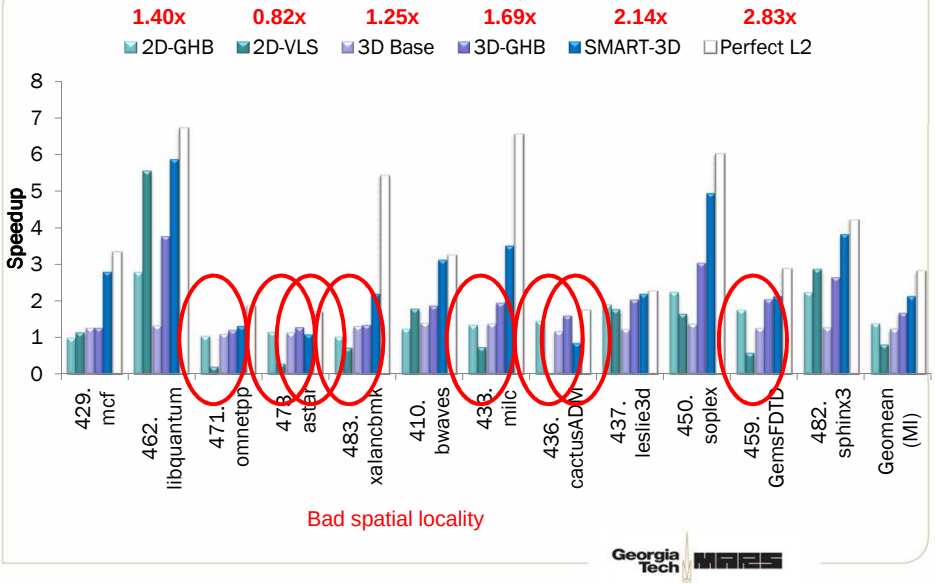
Georgia Tech MARS

SMART-3D: L2 and 3D DRAM

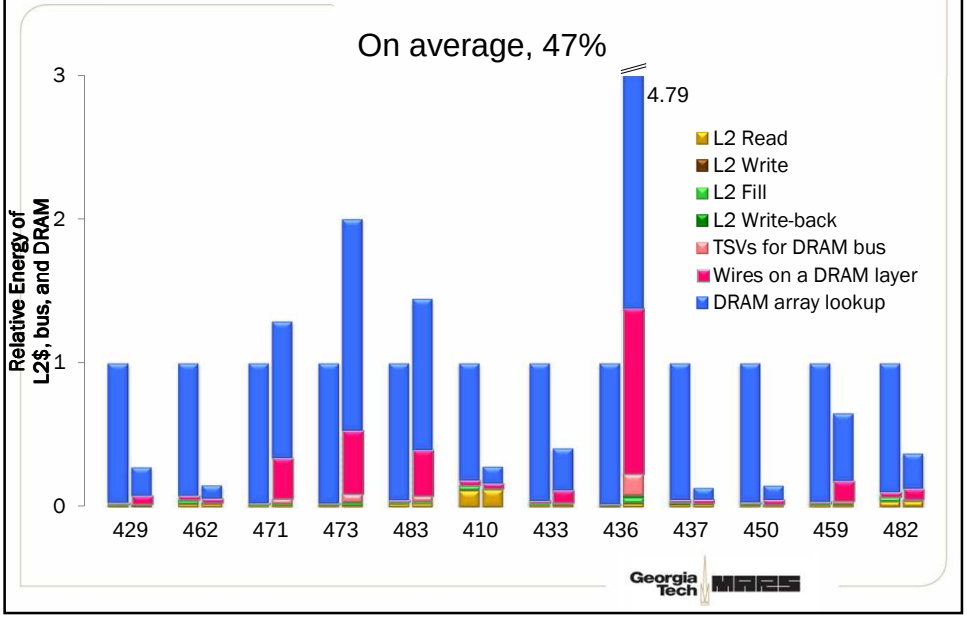


Georgia Tech MARS

Single Core Results



Energy Effect





Summary

- Research Scope @ Georgia Tech
  - 3D Stacked Architecture
  - Physical Design and Tools
  - Design for Testability
  - 3D-MAPS Prototyping and Fabrication
- 3D Integration provides
  - Opportunities for future^2 scaling and system integration
  - Wirelength reduction, i.e., latency and energy consumption
  - New insights for processor/system architecture design

http://arch.ece.gatech.edu
http://www.3d.gatech.edu

Georgia Tech MARS

61