UCLA COMPUTER SCIENCE DEPARTMENT

An Automated Design Flow for 3D Microarchitecture Evaluation

Jason Cong[1] Ashok Jagannathan[1] Yuchun Ma[1,2]
Glenn Reinman[1] Jie Wei[1] Yan Zhang[1]

[1]University of California, Los Angeles, California 90095, U. S. A.
[2]Tsinghua University, Beijing, P.R.China

Partially Supported by DARPA/MTO though CFDRC

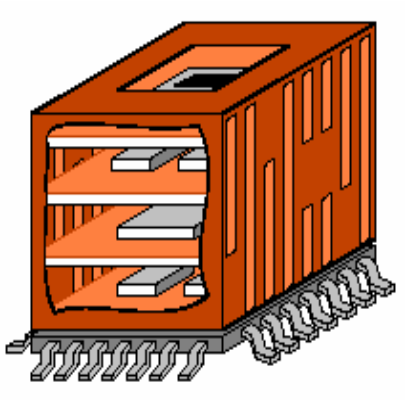
Outline

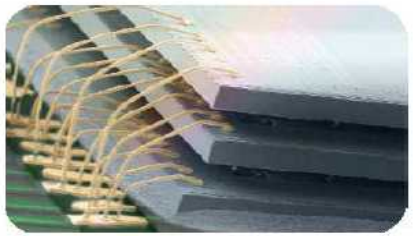
- Introduction
- 3D Microarchitecture Evaluation Flow
  - a. 3D thermal model
  - b. 3D floorplanning
  - c. IPC model and performance simulation
  - d. 3D global routing and thermal via insertion
- Case Study for a Design Driver
- Conclusions and Future Works

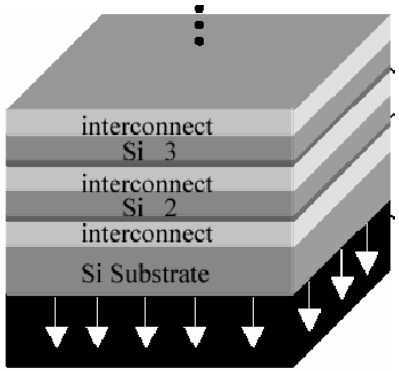
Outline

- Introduction
- 3D Microarchitecture Evaluation Flow
  - a. 3D thermal model
  - b. 3D floorplanning
  - c. IPC model and performance simulation
  - d. 3D global routing and thermal via insertion
- Case Study for a Design Driver
- Conclusions and Future Works

3-D IC Technology Alternatives







•Block level integration

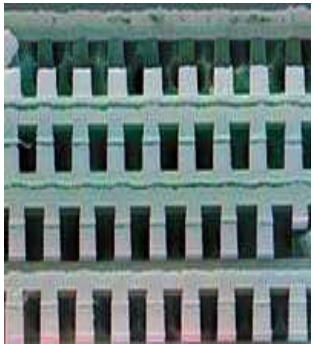
vertical pitch>5μm

vertical density<40k/mm^2

•Chip level integration (3DMCM)

vertical interconnect pitch>50μm

vertical interconnect density< 20/mm (400/mm^2)



•Cell level integration

vertical pitch>200nm

vertical density<25M/mm^2

3D ICs and Microarchitecture

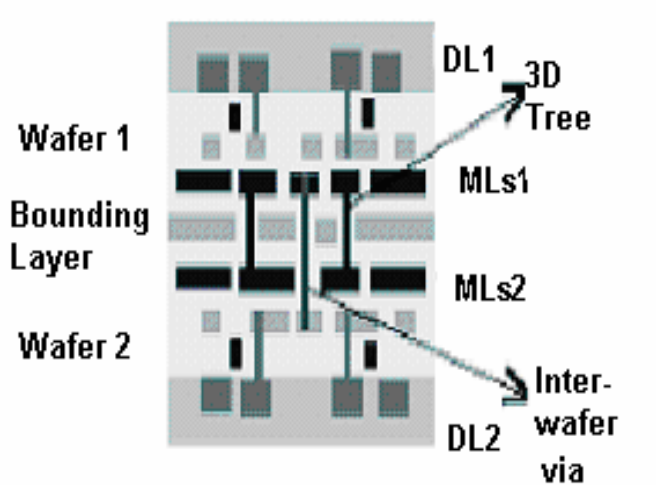
- 3D IC Benefits
  - Reduction of Interconnect delay – increase of performance;
  - Reduction of wirelength – reduction of capacitive power;
  - Increase of integration density;
  - Allow heterogeneous (logic, DRAM, RF, etc) integration
- 3D IC Concerns
  - Thermal constraint
  - Cost
- No existing flow to evaluate 3D implementations of architectures systematically
  - Not clear how much wirelength gain can turn into performance gain (in terms of BIPS)
  - Thermal impact is not known

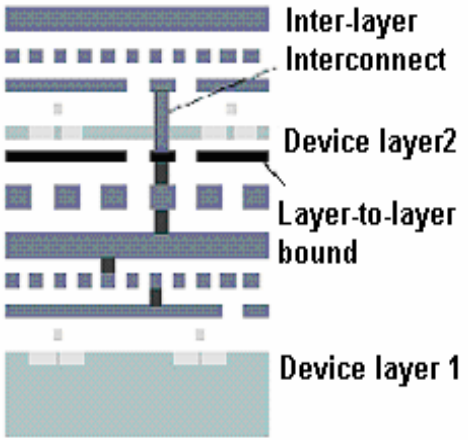
Our Contribution: MEVA-3D

- An Automated Design Flow for 3D Architecture Evaluation (MEVA-3D)
  - Evaluate 3D implementations of micro-architectures systematically in terms of both performance and thermal.
  - Goal: build a physical prototype of the 3D architecture
- MEVA-3D Flow
  - Automated 2D/3D floorplanning;
    - Reduce the latency along critical loops in the mico-architecture by considering interconnect pipelining at a given target frequency.
  - 3D routing and thermal optimization
  - Performance Evaluation
    - Cycle accurate architecture simulation on SPEC2000 benchmarks
  - Thermal Evaluation
    - Resistive network model considering white-space and thermal via insertion.

Technology background

- Wafer bonding 3D IC technologies
  - With flipping the top layer;
  - Without flipping the top layer;





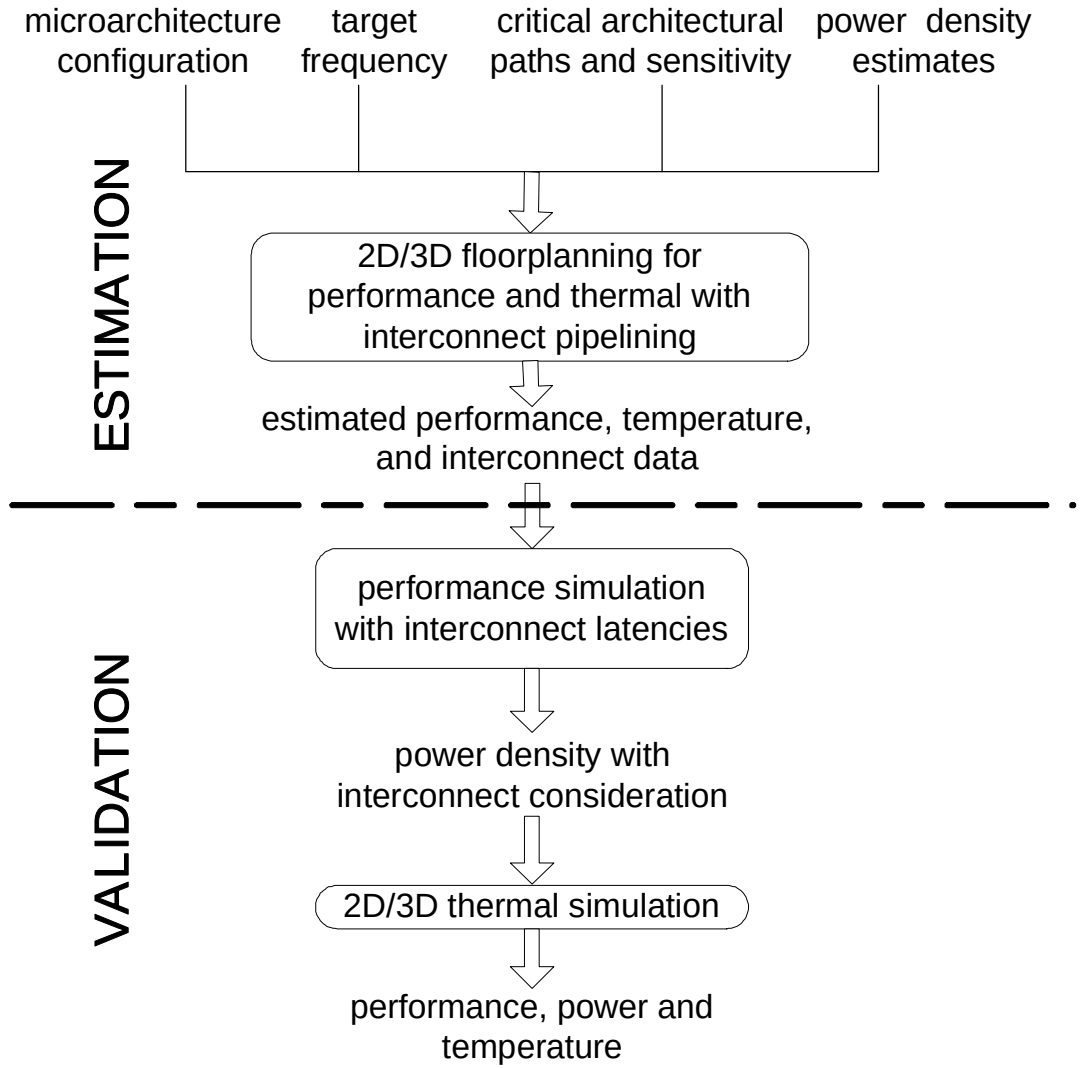
(a) With flipping the top layer (b) Without flipping the top layer

A 3D IC example with two device layers

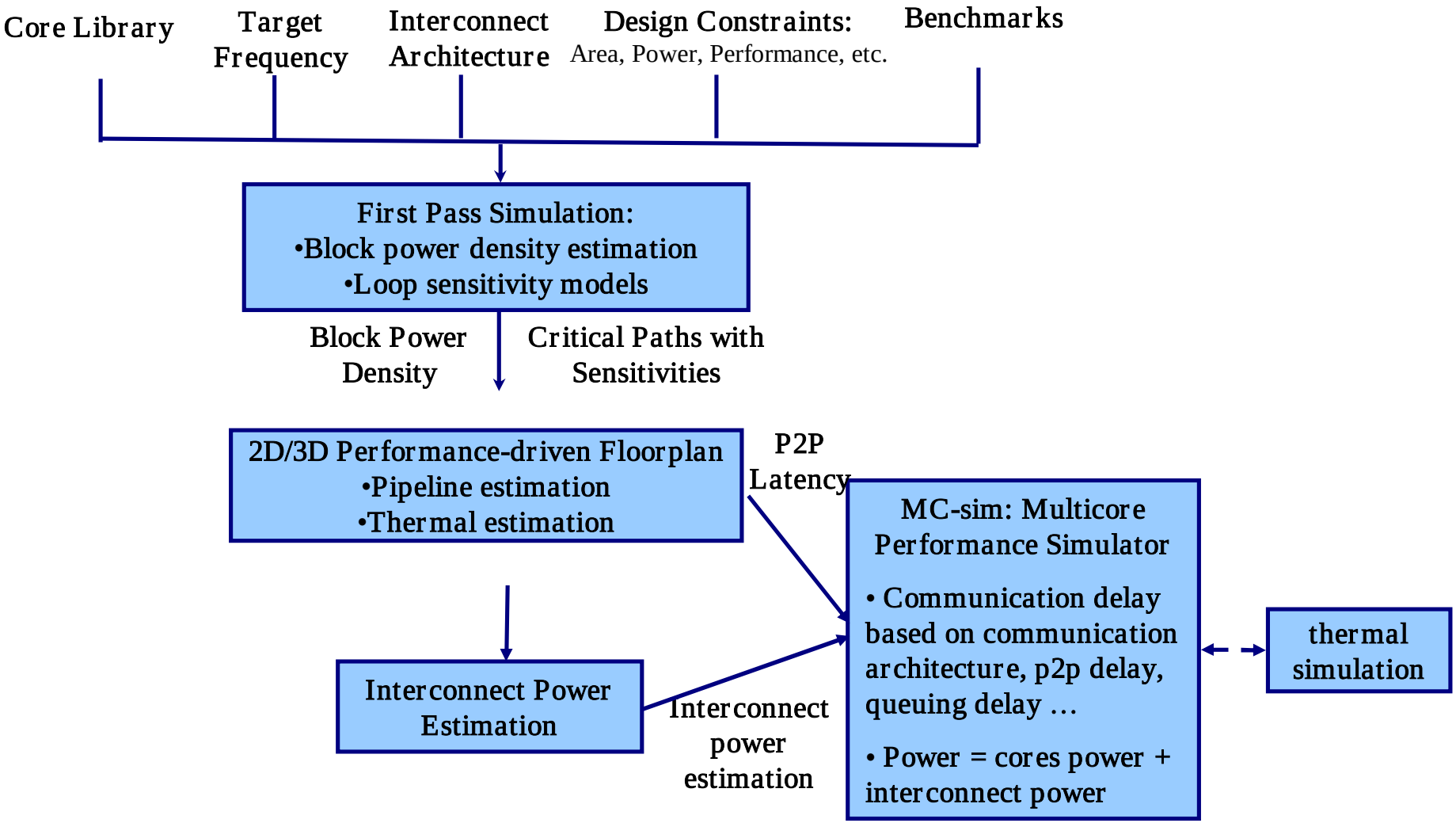
Outline

- Introduction
- 3D Microarchitecture Evaluation Flow
  - a. 3D thermal model
  - b. 3D floorplanning
  - c. IPC model and performance simulation
  - d. 3D global routing and thermal via insertion
- Case Study for a Design Driver
- Conclusions and Future Works

Overview of MEVA-3D

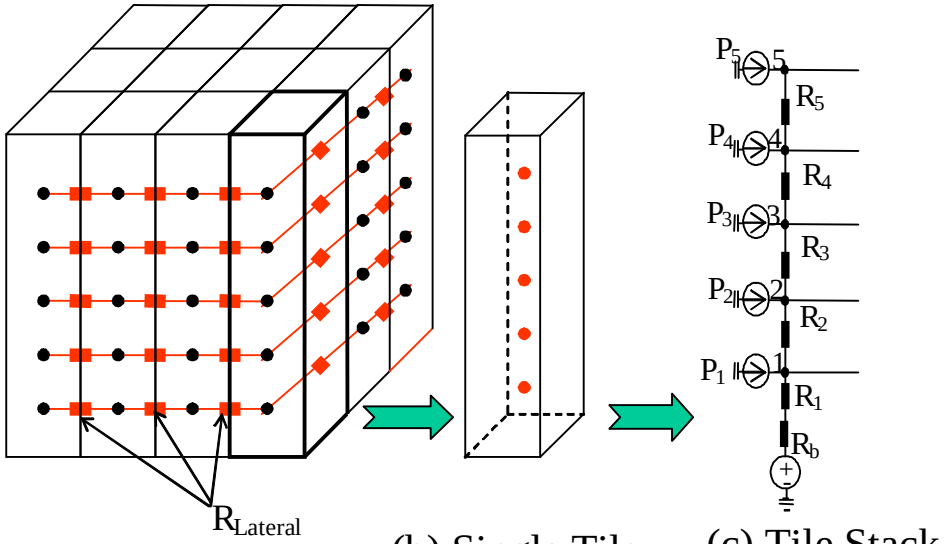


Performance Model and Simulation Flow



3D Thermal Model

- Thermal Resistive Network [Wilkerson 2004]
  - Device layer partitioned into tiles
  - Tiles connected through thermal resistances
  - Heat sources modeled as current sources
  - Heat sinks of fixed temperature T_0
  - TS vias at the center of the tile



(a) Tiles Stack Array

(b) Single Tile Stack

(c) Tile Stack Analysis

Microarchitecture Floorplanning with Wire Pipelining

- Microarchitecture Floorplanning
  - Given:
    - target cycle time T_{cycle}
    - clocking overhead $T_{overhead}$
    - list of blocks in the microarchitecture with their area, dimensions and total logic delay
    - set of critical microarchitectural paths with performance sensitivity models for the paths
    - average power density estimates for the blocks.
  - Objective: Generate a floorplan which optimizes for the die area, performance, and maximum on-chip temperature.

3D floorplanner

- 3D thermal driven floorplan based on TCG representation [Cong 2004]
  - Multi-layer layout
  - Thermal driven
  - Performance evaluation
  - Use simulated annealing with the cost function

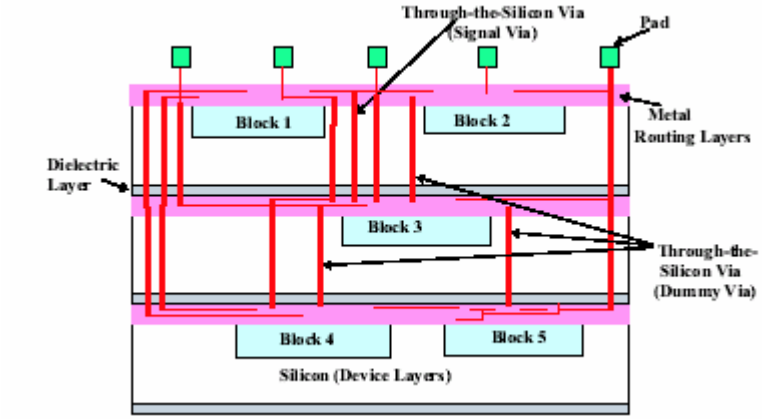
$$cost = w_1 * \frac{1}{BIPS} + w_2 * Area + w_3 * Temp$$

IPC Model and Performance Simulation

- Pipeline evaluation
  - The performance of the micro-architecture in BIPS;
  - IPC performance sensitivity model
    - IPC of a micro-architecture depends on a set of critical processor loops;
    - The information about IPC degradation degree due to extra latency along the path [Sprangle 2002];
    - Calculate the IPC degradation caused by the extra latency introduced by the interconnects in the layout;
- Performance simulation
  - After physical design stage, We adapted the SimpleScalar 3.0 tool set for our simulation framework;
  - Provide feedback to the performance estimators in the floorplanning stage;

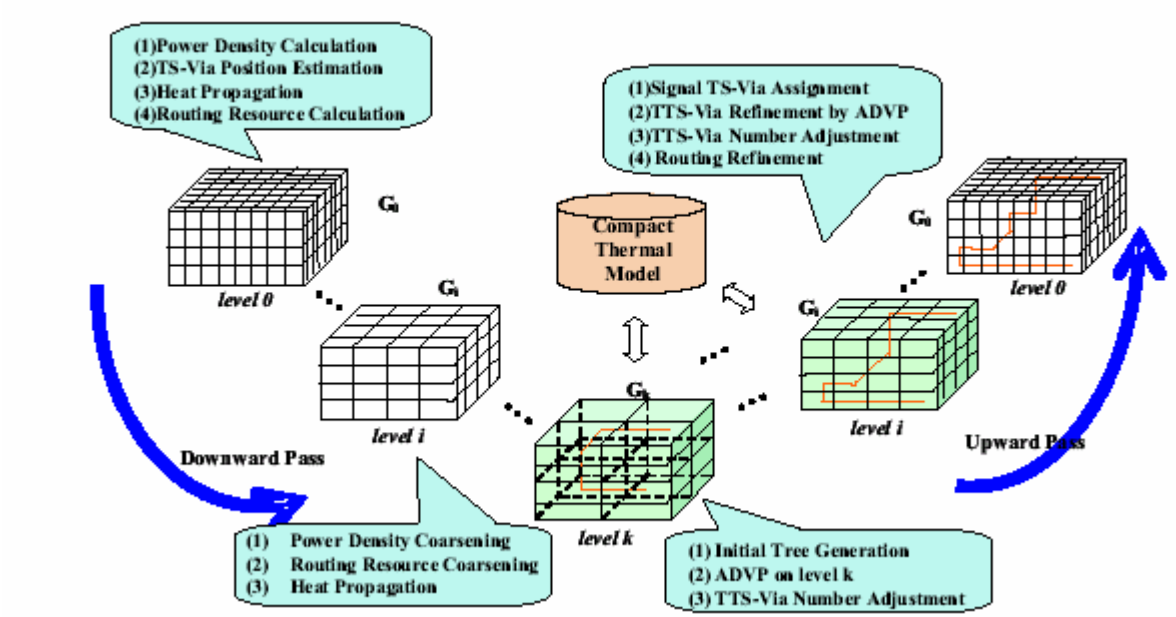
Thermal Optimization Using Though-the-silicon Vias

- Two types of TS vias
  - Signal vias, part of the netlists
  - Thermal vias, with no connections, introduced to reduce temperature
- After floorplanning, we can further reduce the temperature by thermal via insertion.
  - Decrease the maximum temperature by 50%



3D Routing and Thermal Via Planning for 3D ICs

- Simultaneous routing and thermal via planning method. [Cong 2005]
  - Multilevel routing and via planning framework
  - Via planning during and after routing

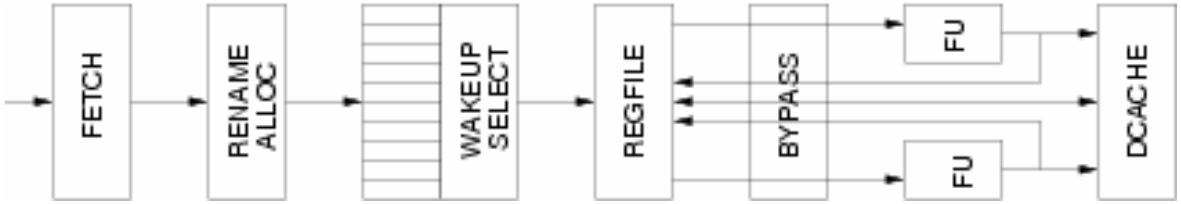


Outline

- Introduction
- 3D Microarchitecture Evaluation Flow
  - a. 3D thermal model
  - b. 3D floorplanning
  - c. IPC model and performance simulation
  - d. 3D global routing and thermal via insertion
- Case Study for a Design Driver
- Conclusions and Future Works

Design Example

- An out-of-order superscalar processor micro-architecture with 4 banks of L2 cache in 70nm technology.



Instruction Cache	32KB, 32B/block, 2-way
Decode Width	4
ROB Size	128 entries
Issue Queue	32 entries
Issue Width	4 ALU ops, 2 MEM ops per cycle
Register File	70 INT and 70 FP
Functional Units	Units 2 IntALU, 1 FPALU, 1 IntMult, 1 FPMult
Load/Store Queue	32 entries
L1Data Cache	16KB, 32B/block, 4-way, 2RW ports
Unified L2 cache	1MB, 64B/block, 8-way

Properties of Design Example

- The critical paths

Wakeup Latency	Latency to wakeup the dependent instruction
ALU Bypass	latency of the bypass wires between the ALUs
DL1 Latency	Load latency though the L1 data cache
L2 Latency	Latency for access to L2 cache
MPLAT	latency through the branch resolution path

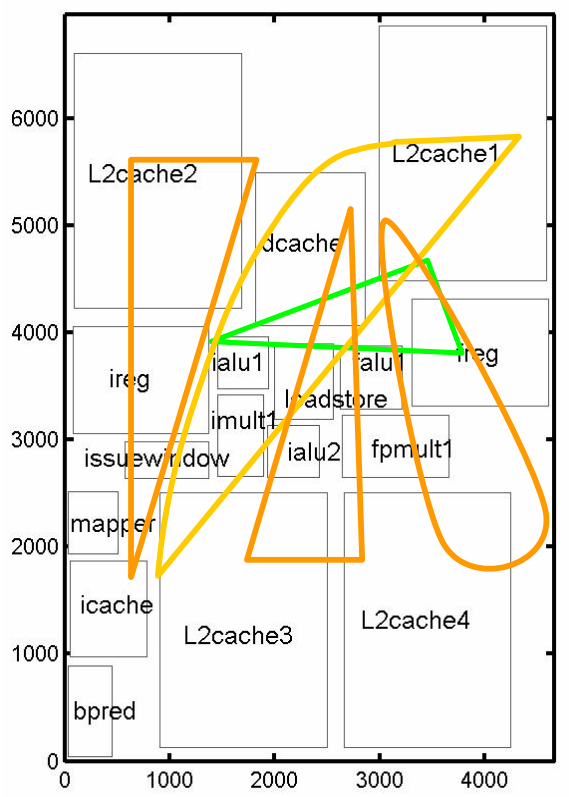
- Baseline processor parameters

Instruction Cache	32KB, 32B/block, 2-way
Decode Width	8
ROB Size	128 entries
Issue Queue	32 entries
Issue Width	8
Register File	70 INT and 70 FP
Functional Units	Units 4 IntALU, 1 FPALU, 2 IntMult, 1 FPMult
Load/Store Queue	32 entries
L1Data Cache	16KB, 32B/block, 4-way, 2RW ports
Unified L2 cache	1MB, 64B/block, 8-way

3D Design Driver

- Alpha EV6-like core – 4GHz clock frequency
  - Design Space Exploration without leveraging 3D for individual architectural blocks

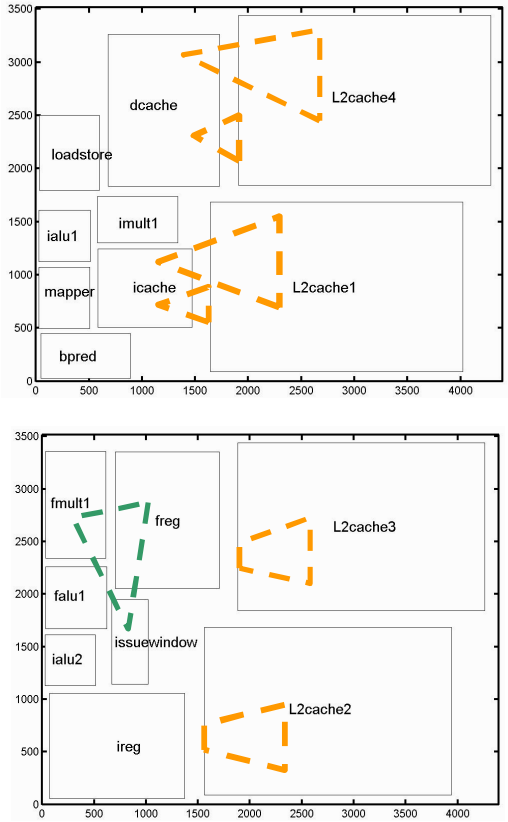
2D EV6-like core



Wakeup loop :
The extra cycle is eliminated.

Branch misprediction resolution loop and the L2 cache access latency :
Some of the extra cycles are eliminated

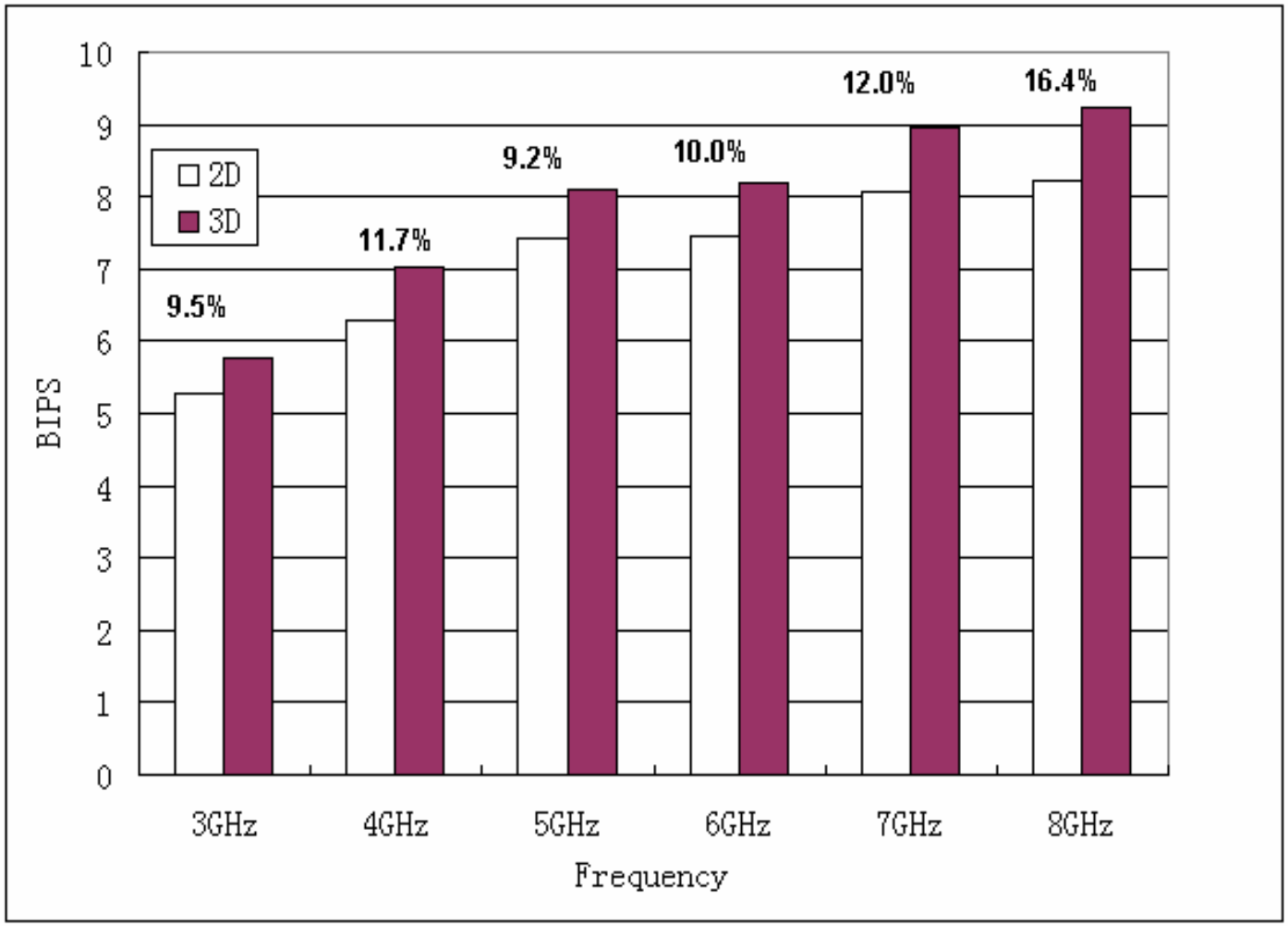
3D EV6-like core (2 layers)



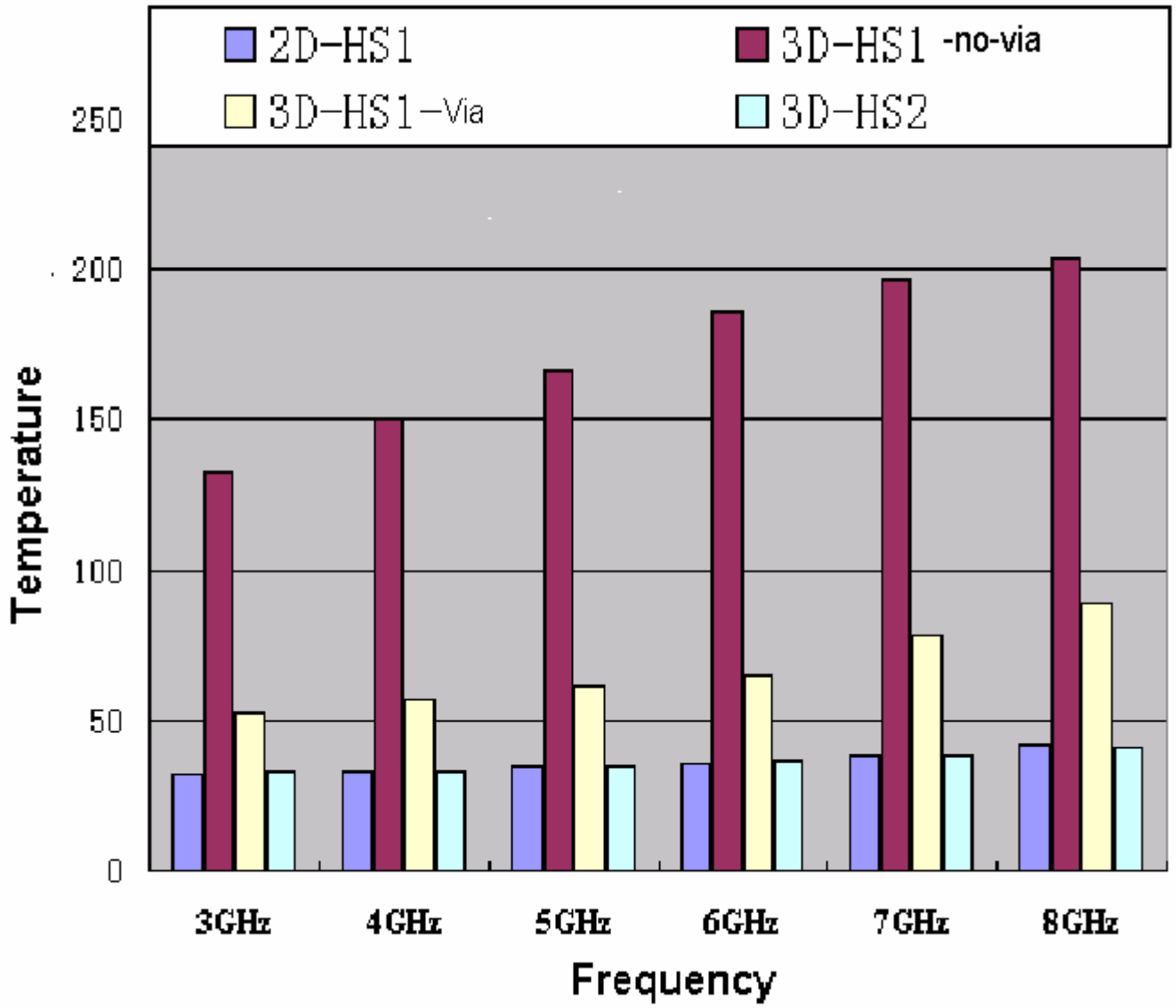
The number of extra cycles for critical paths group

Freq(Hz)	3G		4G		5G		6G		7G		8G	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
wakeup	1	0	1	1	1	1	1	0	2	1	2	1
ALU	0	0	1	0	0	0	0	0	1	0	1	0
DL1	0	0	0	0	1	0	2	1	1	1	2	1
L2	2	0	2	1	3	1	5	1	5	2	5	1
MPLAT	1	0	1	0	2	1	2	1	2	1	5	2

Performance for the micro-architecture with 2D and 3D layout at different target frequencies

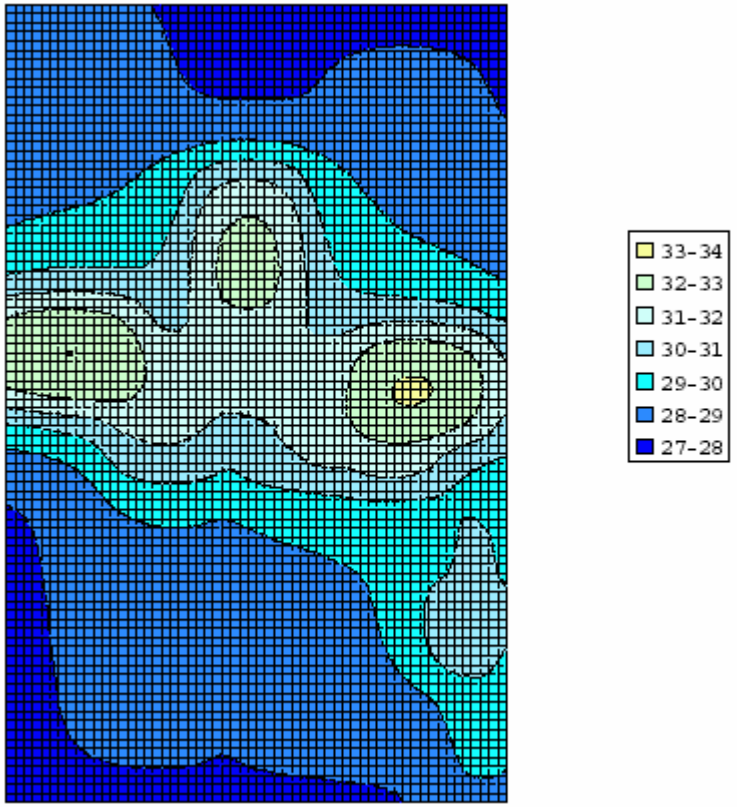


Maximum On-Chip Temperature



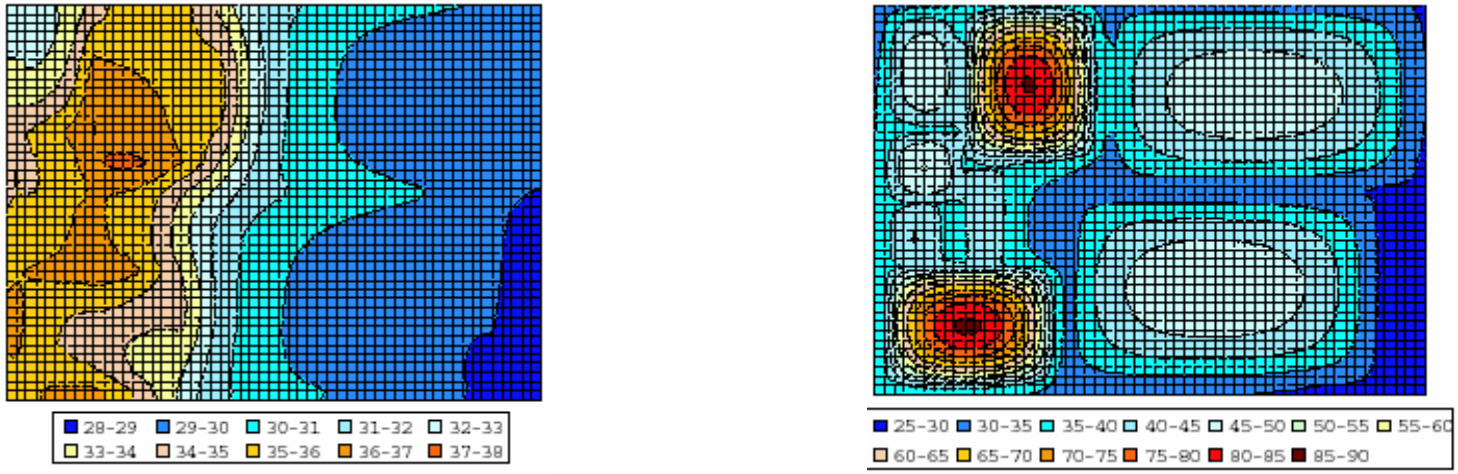
HS denotes a heat sink, and the 3D integration allows to insert thermal vias to reduce the temperature.

Thermal Profiles for 2D chip(4Ghz)

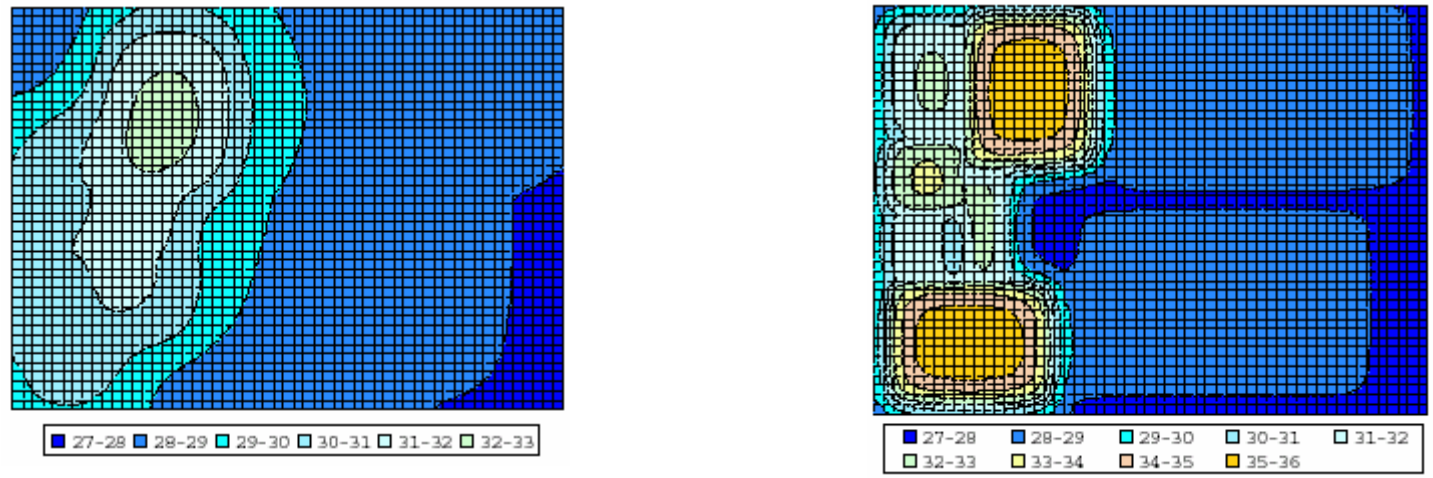


Temperature distribution in 2D integration.

Thermal Profiles for 3D chip(4Ghz)



Temperature distribution in 3D integration with one heat sink.



Temperature distribution in 3D integration with two heat sinks and flipped upper layer.

Conclusions

- 3D integration can eliminate most of the extra cycles incurred by interconnects in 2D micro-processor designs.
- MEVA-3D can systematically evaluate the 3D architecture both from the performance side and from the thermal side.
- By evaluating one Alpha-like out-of-order, superscalar microprocessor, we show that 3D integration can improve the performance by 10% with comparable maximum on- chip temperature.

On-going Research

- 3D Architecture Modeling and Exploration
  - 3D layout can only reduce the global interconnect delay. Additionally, the 3D implement of components can reduce internal delay and power consumption;
  - 3D area, performance and power models for key architectural components including register files, caches and issue queues;
  - Automated tool for microarchitectural and physical floorplanning co-design to explore the use of 2D and/or 3D architecture blocks;
  - Preliminary study shows 20% performance improvement over a 2D architecture.

THANKS