UCLA COMPUTER SCIENCE DEPARTMENT

# Coping with Vertical Interconnect Bottleneck


**Jason Cong**

***UCLA Computer Science Department***

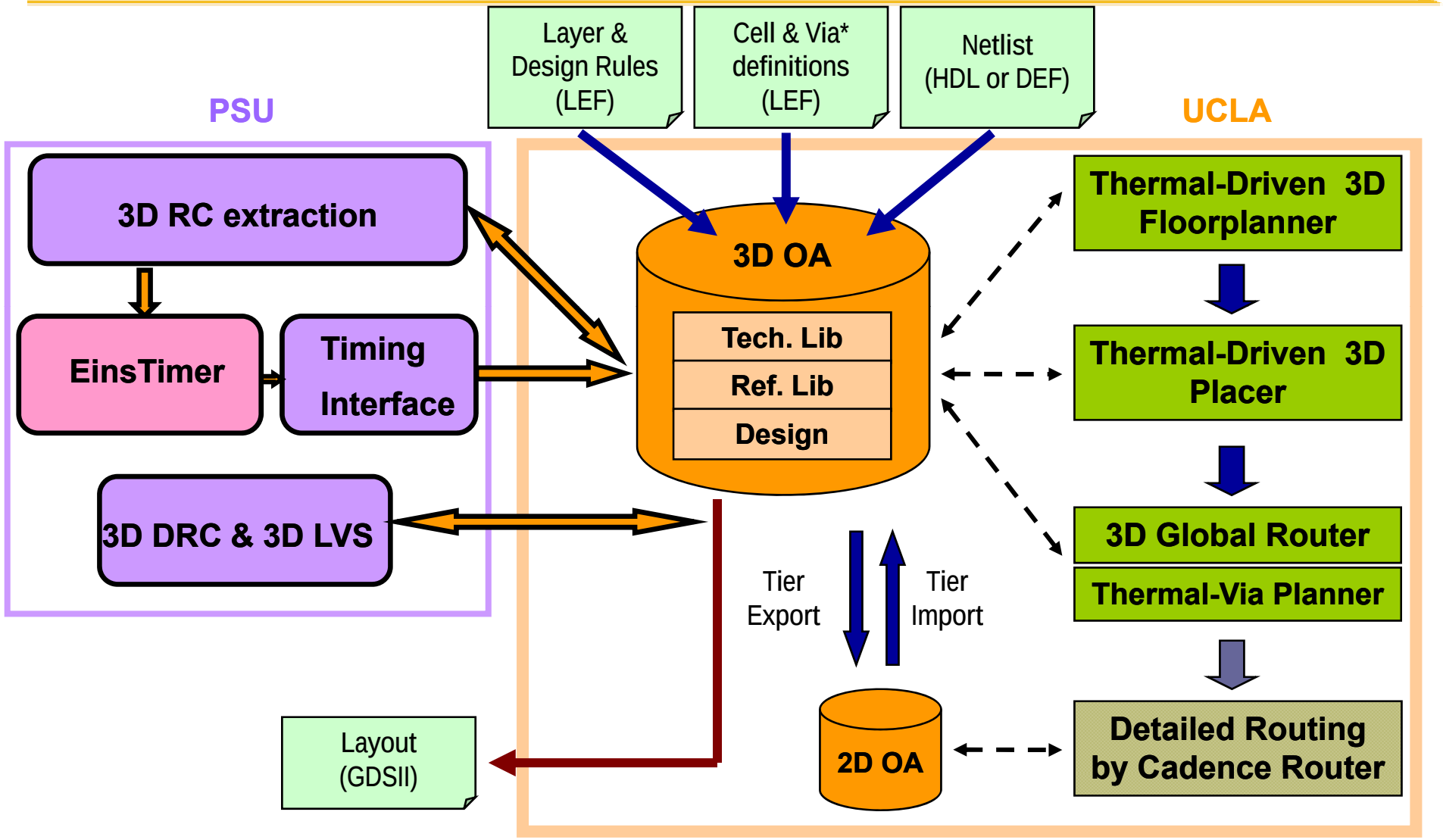
***cong@cs.ucla.edu***

***http://cadlab.cs.ucla.edu/~cong***


# *Outline*

- **Lessons learned**
- **Research challenges and opportunities**

# Recent Work on 3D Physical Design Flow (IBM, UCLA, and PSU) (2006 – 2008)

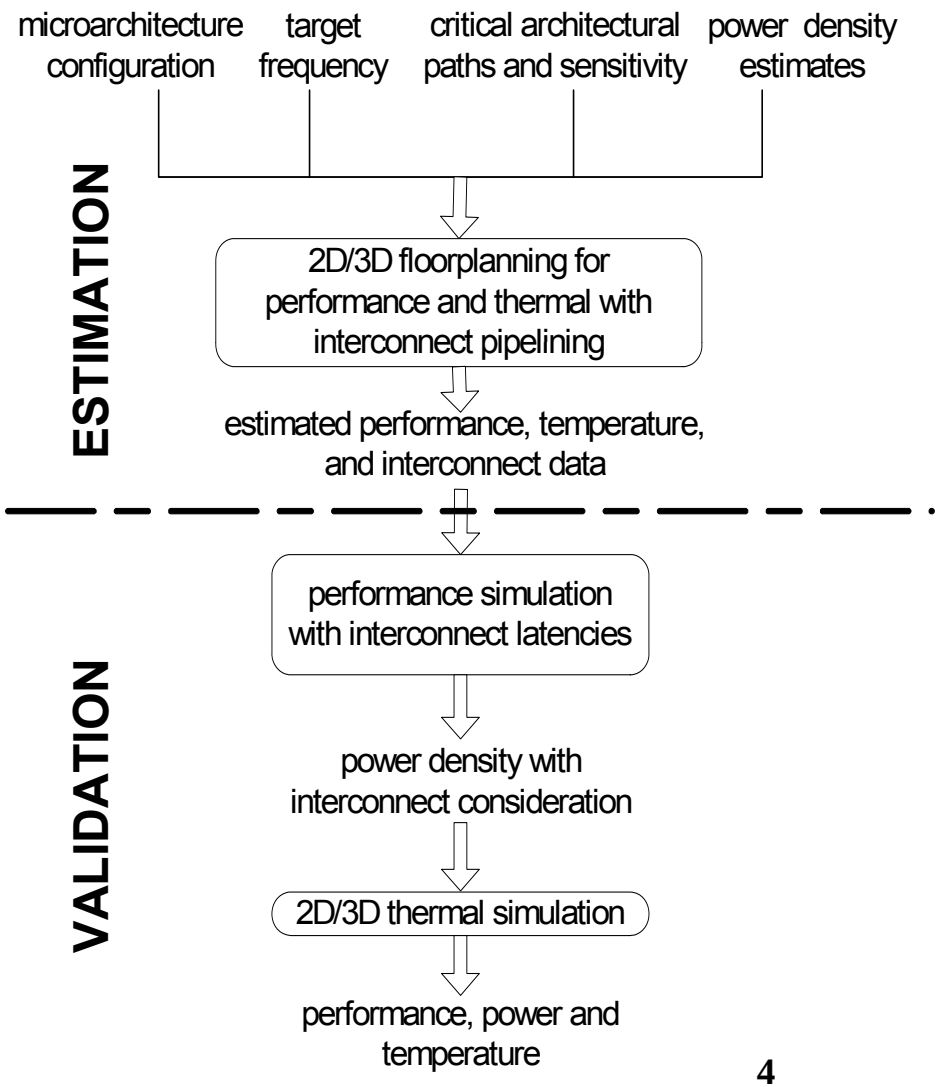





UCLA 3D research started in 2002 under DARPA with CFDRC

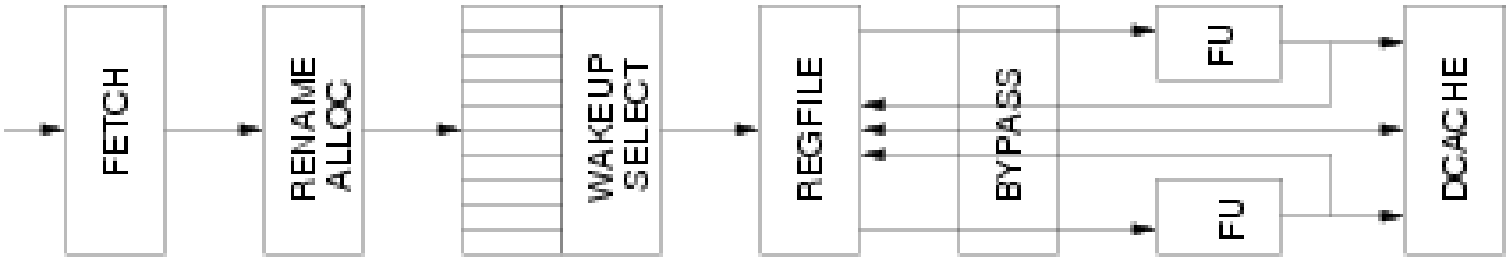
# *3D Architecture Evaluation with Physical Planning -- MEVA-3D [DAC'03 & ASPDAC'06]*

- **Optimize**
  - **BIPS (not IPC or Freq)**
    - **Consider interconnect pipelining based on early floorplanning for critical paths**
    - **Use IPC sensitivity model [Jagannathan05]**
  - **Area/wirelength**
  - **Temperature**



# Design Driver 1 (Using Top-Level Floorplan)

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

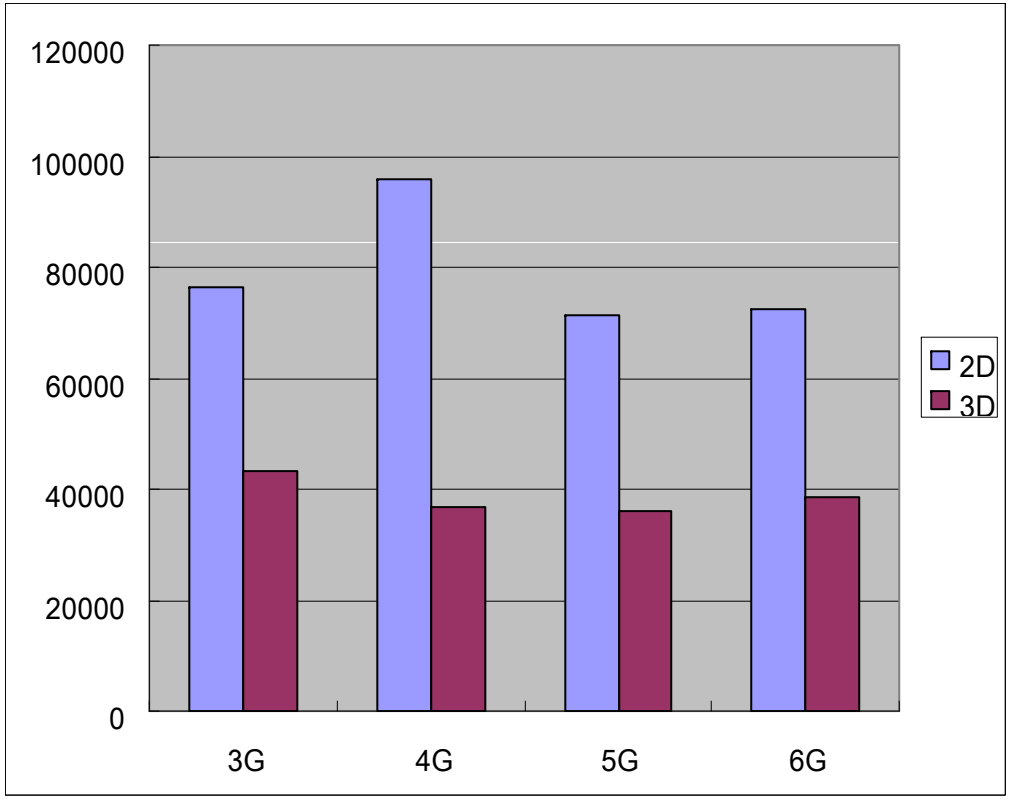



- **Critical paths**

| | |
|---|---|
| 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 |

# *Top-Level Wirelength Improvement from 3D Stacking*

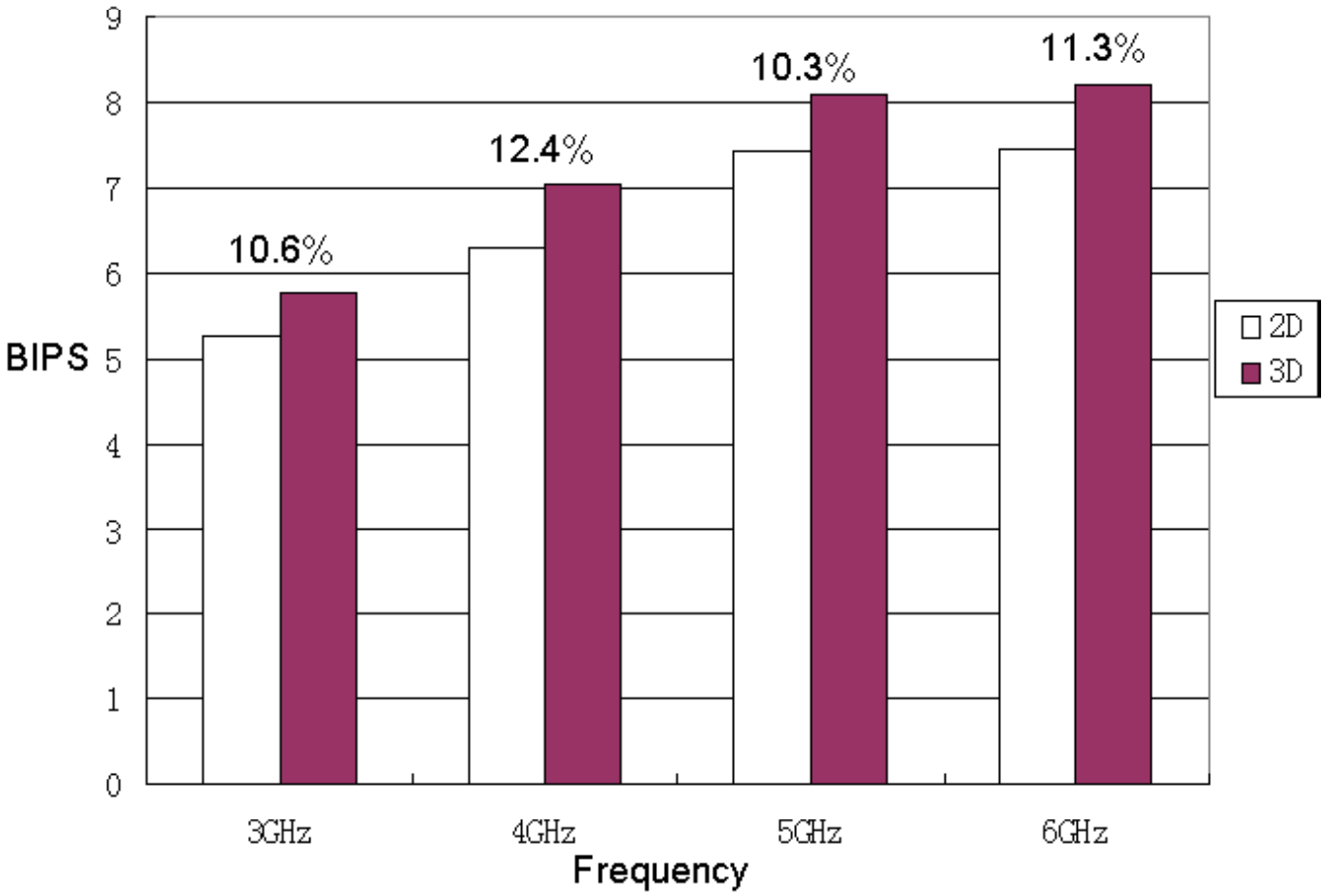
## Close to 2X WL reduction (for top-level interconnects)




## Assume two device layers

# Performance Improvement from 3D Stacking

## Disappointing ….

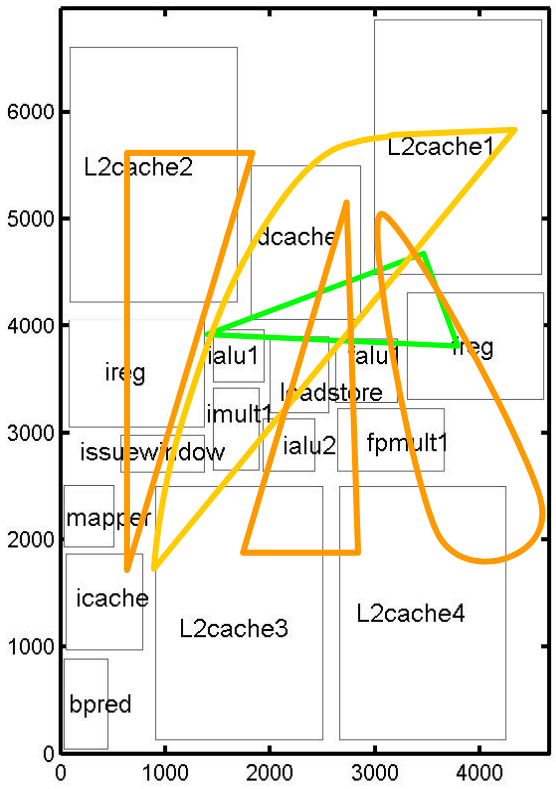



## Assume two device layers

# 2D vs 3D Layout

## Assume two device layers

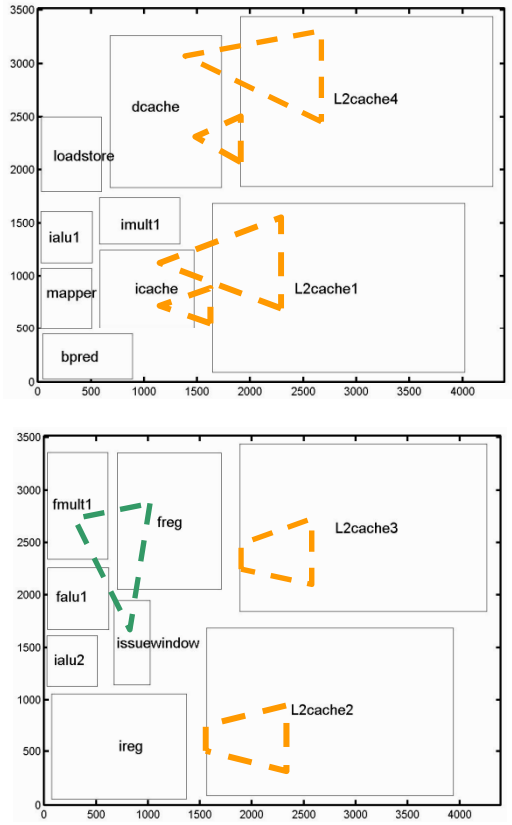
**2D EV6-like core**
**BIPS= 2.75**




**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)**
**BIPS= 2.94**

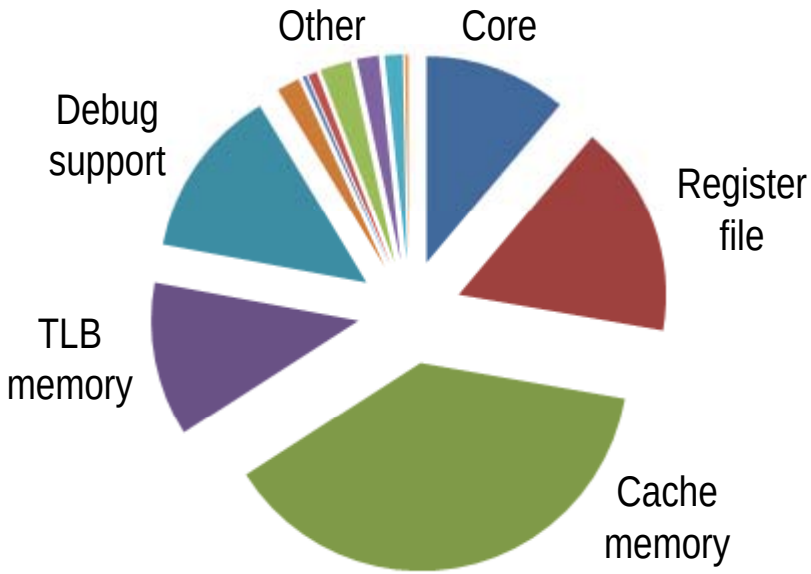


# Design Driver 2 (Using Full RTL)

- An open-source 32-bit processor
  - Compliant with SPARC V8 architecture
- Synthesized by Cadence RTL compiler with UMC 90nm digital cell library and Faraday memory compiler
  - Configuration: Single core with 4KB data cache and 4KB instruction cache as direct-mapped caches
  - statistics:
    - #cell = 34225
    - #macro = 12
    - #net = 36789
    - Total area = 6.67 x $10^5$ $\mu m^2$

# Logical Hierarchy of LEON3

- **LEON3 (77.8% area)**
  - **Processor core (11.1% area)**
    - **Integer unit (6.6% area)**
    - **Multiplier (1.6% area)**
    - **Divider (0.7% area)**
    - **Memory management unit (2.2% area)**
  - **Register file (16.6% area)**
  - **Cache memory (38.1% area)**
  - **TLB memory (12.0% area)**
- **Debug support unit (13.4% area)**
- **Other (8.8% area)**
  - **Memory controller (1.8% area)**
  - **Interrupt controller (0.3% area)**
  - **UART serial interface (0.7% area)**
  - **AMBA AHB bus, AMBA APB bus (4.3% area)**
  - **General purpose timer unit (1.4% area)**
  - **General purpose I/O unit (0.3% area)**



# 3D Placement Restricted By Logical Hierarchies

- **Comparisons**

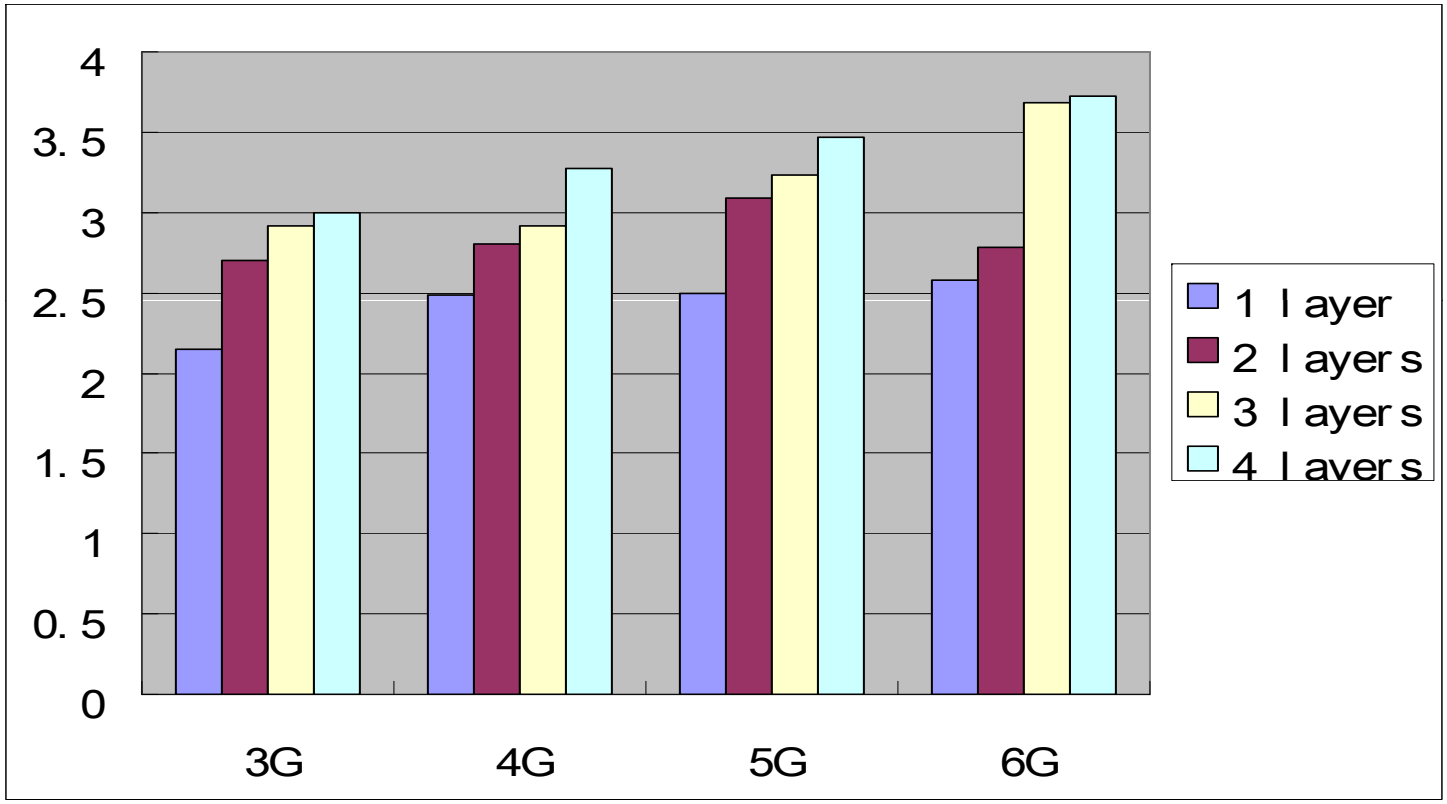
| | Flat | Processor Core restricted | Register file restricted |
|---|---|---|---|
| HWPL | 0.99 (m) | 1.09 (m) | 1.20 (m) |
| #TSV | 3835 | 1715 | 845 |

  - **Flat 3D placement**
  - **Processor core restricted**
    - **Processor core is restricted in only one device layer**
      - Including Integer unit, multiplier, divider and MMU
  - **Register file restricted**
    - **Register file is restricted in only one device layer**

# Outline

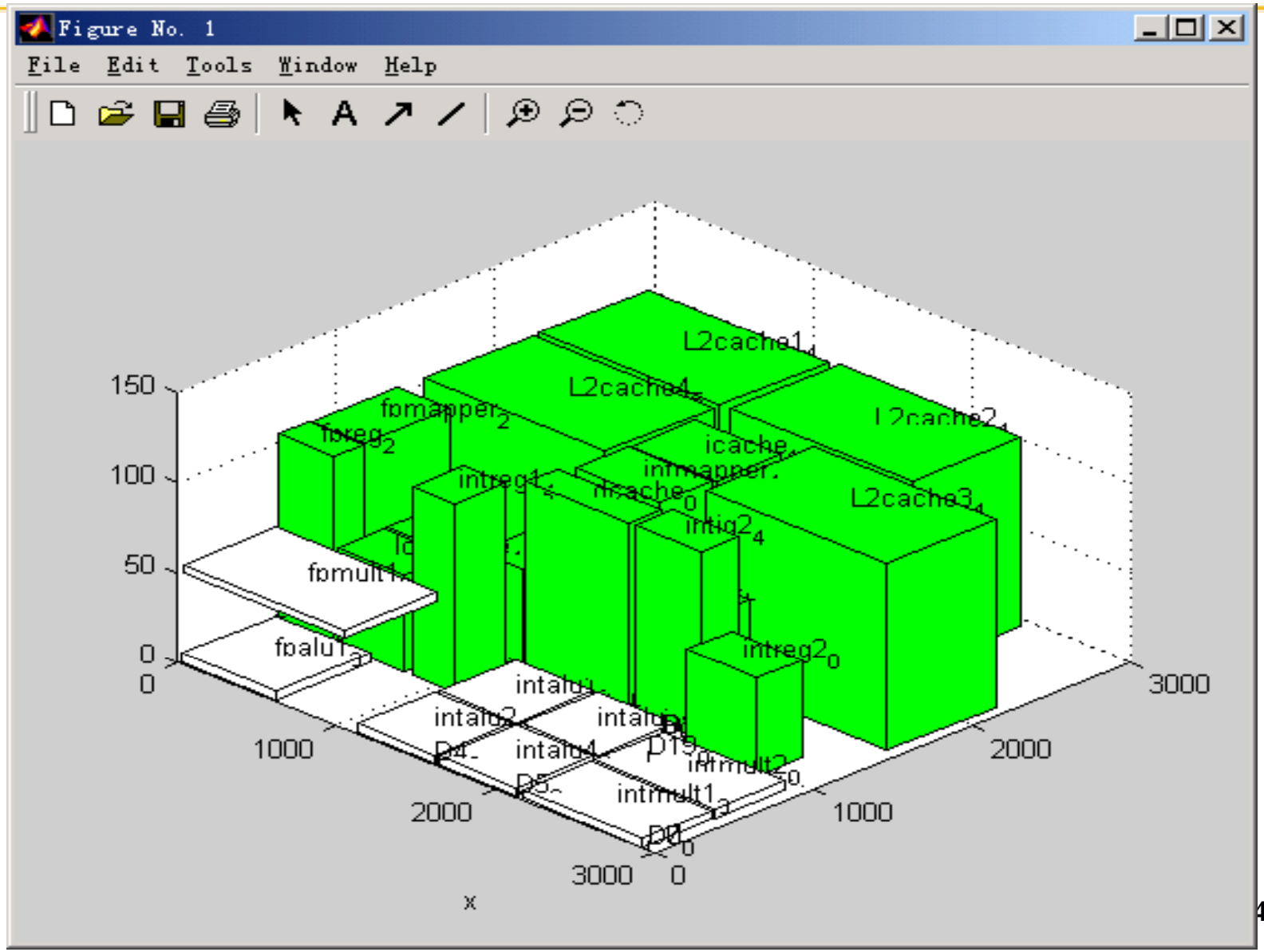
- **Lessons learned**
  - Block stacking gives limited performance and WL reduction
  - Full potential is realized with extensive vertical connections
- **Research challenges and opportunities**
  - Novel 3D architecture component designs that can cope with the vertical interconnect bottleneck,
  - Physical synthesis tools that can fully comply with global and local TSV density constraints,
  - 3D microarchitecture exploration, include generating optimized 3D physical hierarchies under the TSV density constraints
  - New interconnect technologies that can alleviate or eliminate the vertical interconnect bottleneck.

# Results from 3D Folding and Stacking




**Over 35% performance improvement**

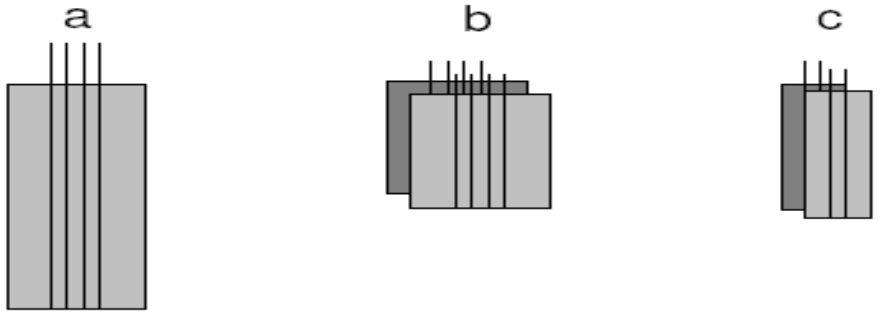
# 5GHz 3 Device Layer Layout



# 3D Architectural Blocks – Issue Queue

- **Block folding**
  - Fold the entries and place them on different layers
  - Effectively shortens the tag lines
- **Port partitioning**
  - Place tag lines and ports on multiple layer, thus reducing both the height and width of the ISQ.
  - The reduction in tag and matchline wires can help reduce both power and delay.

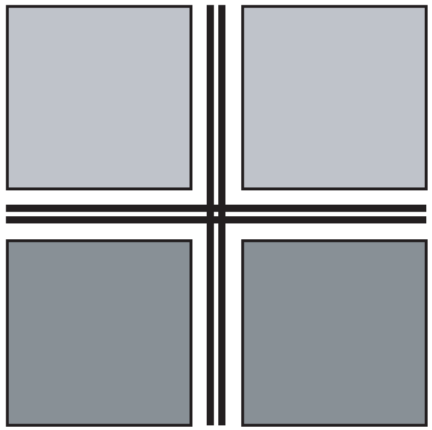
- **Benefits from block folding**
  - Maximum delay reduction of 50%, maximum area reduction of 90% and a maximum reduction in power consumption of 40%



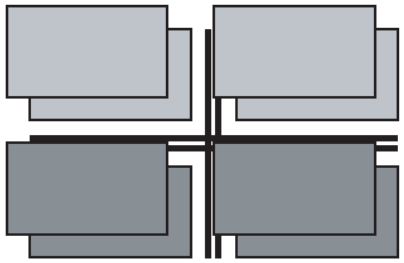

(a) 2D issue queue with 4 taglines;
(b) block folding; (c) port partitioning

# 3D Architectural Blocks – Caches

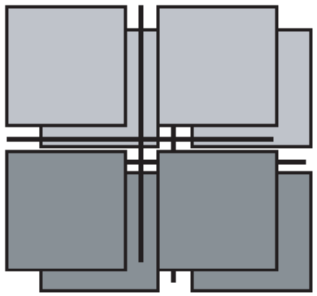
- 3D-CACTI: a tool to model 3D cache for area, delay and power
  - We add port partitioning method
  - The area impaction of vias
- Improvements
  - Port folding performs better than wordline folding for area.(72% vs 51%)
  - Wordline folding is more effective in reducing the block delay (13% vs 5%)
  - Port folding also performs better in reducing power (13% vs 5%)
- Requires dense TSVs



Single Layer Design



Wordline Folding



Port Partitioning

# Outline

- **Lessons learned**
  - Block stacking gives limited performance and WL reduction
  - Full potential is realized with extensive vertical connections
- **Research challenges and opportunities**
  - Novel 3D architecture component designs that can cope with the vertical interconnect bottleneck
  - Physical synthesis tools that can fully comply with global and local TSV density constraints
  - 3D microarchitecture exploration, include generating optimized 3D physical hierarchies under the TSV density constraints
  - New interconnect technologies that can alleviate or eliminate the vertical interconnect bottleneck

# *Current Approaches to Handling TSV Constraints*

- **Approach 1: minimizing**

  **WL + k* #TSVs**

- **Approach 2: minimizing WL (or weighted WL) subject to the total #TSV constraints**

- **None of these can handle local TSV density constraints**

# Outline

- **Lessons learned**
  - Block stacking gives limited performance and WL reduction
  - Full potential is realized with extensive vertical connections
- **Research challenges and opportunities**
  - Novel 3D architecture component designs that can cope with the vertical interconnect bottleneck
  - Physical synthesis tools that can fully comply with global and local TSV density constraints
  - 3D microarchitecture exploration, include generating optimized 3D physical hierarchies under the TSV density constraints
  - New interconnect technologies that can alleviate or eliminate the vertical interconnect bottleneck

# *Example: Impact of Following Logical Hierarchy*

- **Comparisons**

| | Flat | Processor Core restricted | Register file restricted |
|---|---|---|---|
| HWPL | 0.99 (m) | 1.09 (m) | 1.20 (m) |
| #TSV | 3835 | 1715 | 845 |

  - Flat 3D placement
  - **Processor core restricted**
    - **Processor core is restricted in only one device layer**
      - Including Integer unit, multiplier, divider and MMU
  - **Register file restricted**
    - **Register file is restricted in only one device layer**

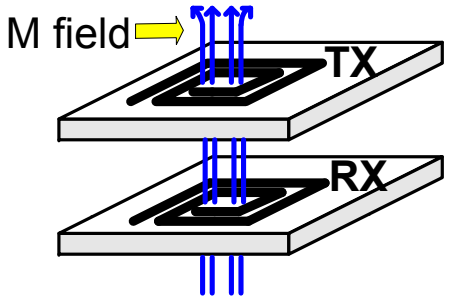
- **Question: how much logic hierarchy to flatten for 3D design/optimization?**

# Outline

- **Lessons learned**
  - Block stacking gives limited performance and WL reduction
  - Full potential is realized with extensive vertical connections
- **Research challenges and opportunities**
  - Novel 3D architecture component designs that can cope with the vertical interconnect bottleneck
  - Physical synthesis tools that can fully comply with global and local TSV density constraints
  - 3D microarchitecture exploration, include generating optimized 3D physical hierarchies under the TSV density constraints
  - New interconnect technologies that can alleviate or eliminate the vertical interconnect bottleneck

# Contactless Interconnects

## Inductor-coupled Interconnect




**Advantages:**

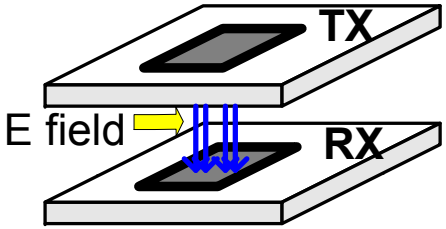
More effective for longer distance communication (hundreds of microns)

**Disadvantages:**

Larger size

Higher cross talks between channels

## Capacitor-coupled Interconnect




**Advantages:**

Smaller size

Lower cross talk

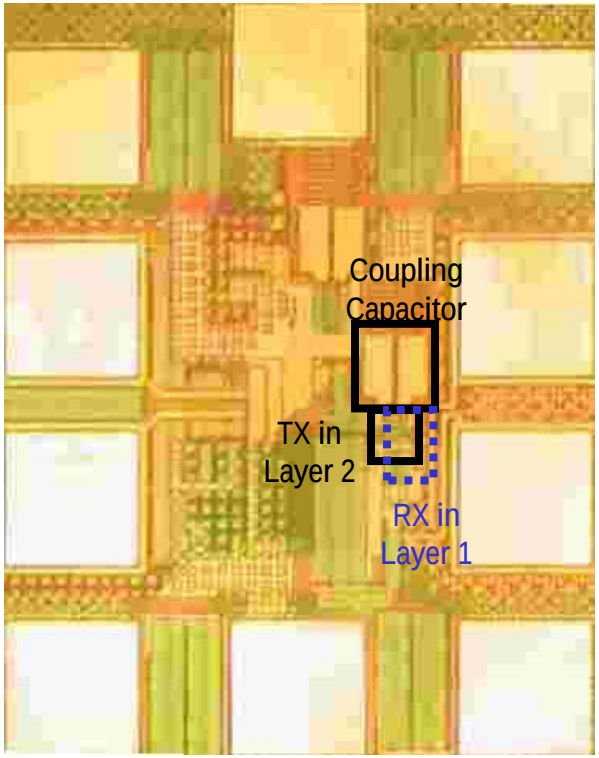
**Disadvantages:**

Effective for short distance communication (several microns)

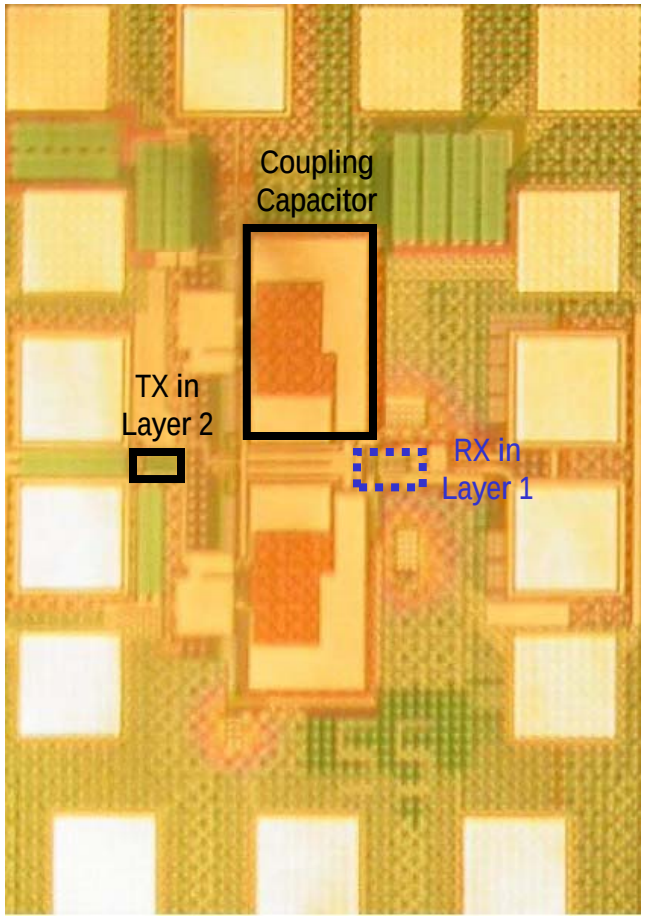
**Suitable for 3DIC integration**

# Die Photos (MIT LL 0.18um)

## BISI die photo

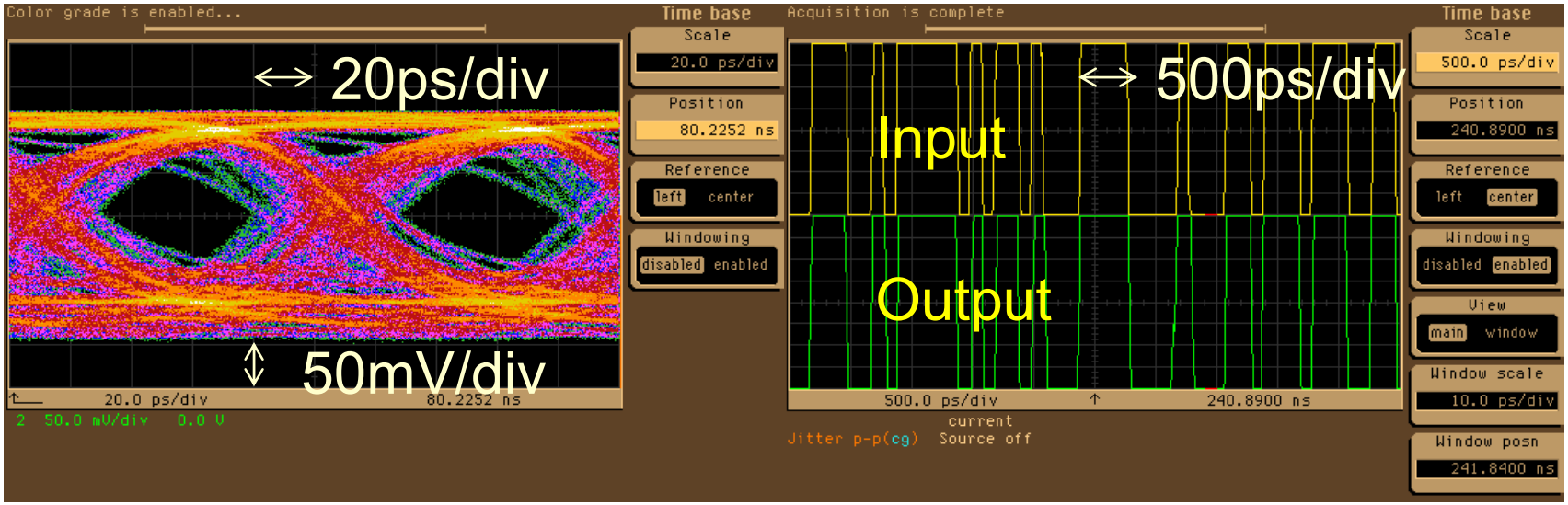



## RFI die photo



# BISI Test Results [ISSCC'07]

**Data rate:10Gbps**




**Output Eye diagram** **Output versus Input**

# *Conclusions*

- **Never enough for vertical interconnects (VIs)**
- **Need to cope with VI constraints**
  - **Novel 3D architecture component designs**
  - **Physical synthesis tools that can fully comply with global and local TSV density constraints,**
  - **3D microarchitecture exploration, include generating optimized 3D physical hierarchies under the TSV density constraints**
- **Need to find ways to break VI bottleneck**
  - **New interconnect technologies**

# *Acknowledgements*

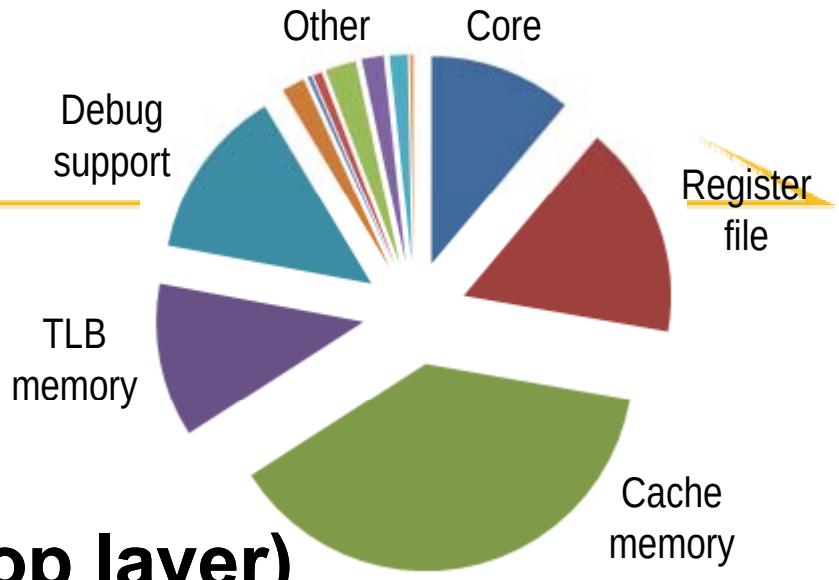

- We would like to thank the supports from DARPA
- Support from the primary contractors -- Collaboration with CFDRC and IBM
- Publications are available from

  http://cadlab.cs.ucla.edu/~cong

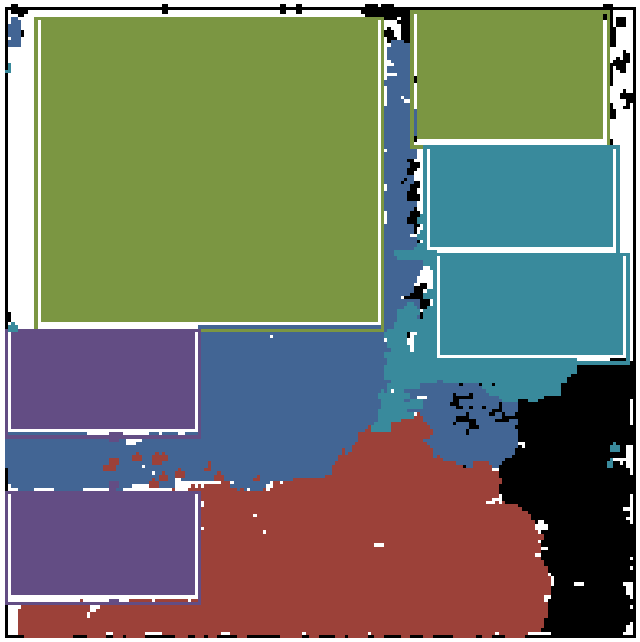
# Example 1 Processor Parameters

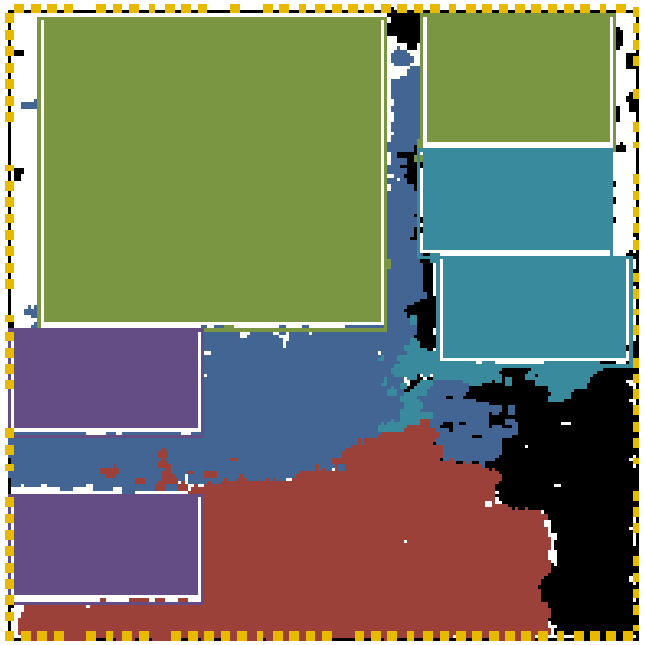
| | |
|---|---|
| Instruction Cache | 32KB, 32B/block, 2-way |
| Decode Width | 8 |
| ROB Size | 128 entries |
| Issue Queue | 32 entries |
| Issue Width | 8 |
| Register File | 70 INT and 70 FP |
| Functional Units | Units 4 IntALU, 1 FPALU, 2 IntMult, 1 FPMult |
| Load/Store Queue | 32 entries |
|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

# Example 2 Flat 3D Placement

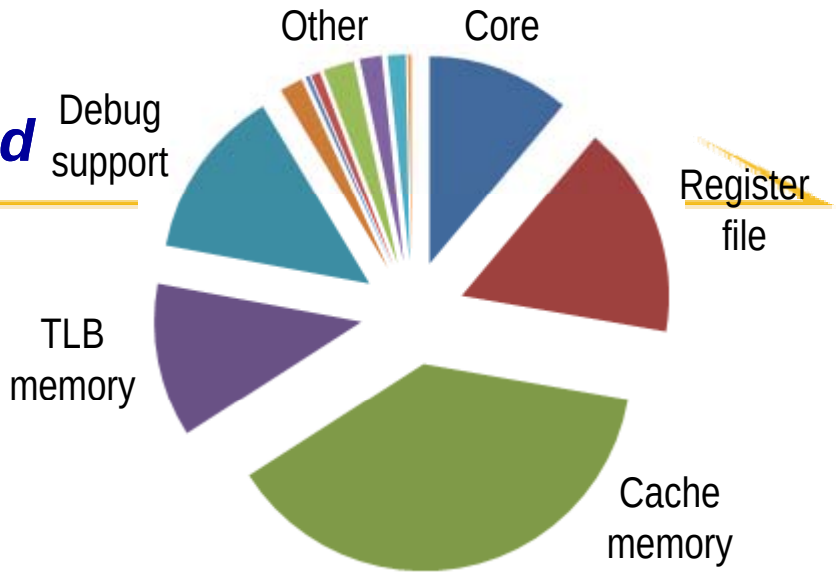



- HPWL = 0.99 (m)
- #TSV = 3835
- Placement (bottom layer, top layer)

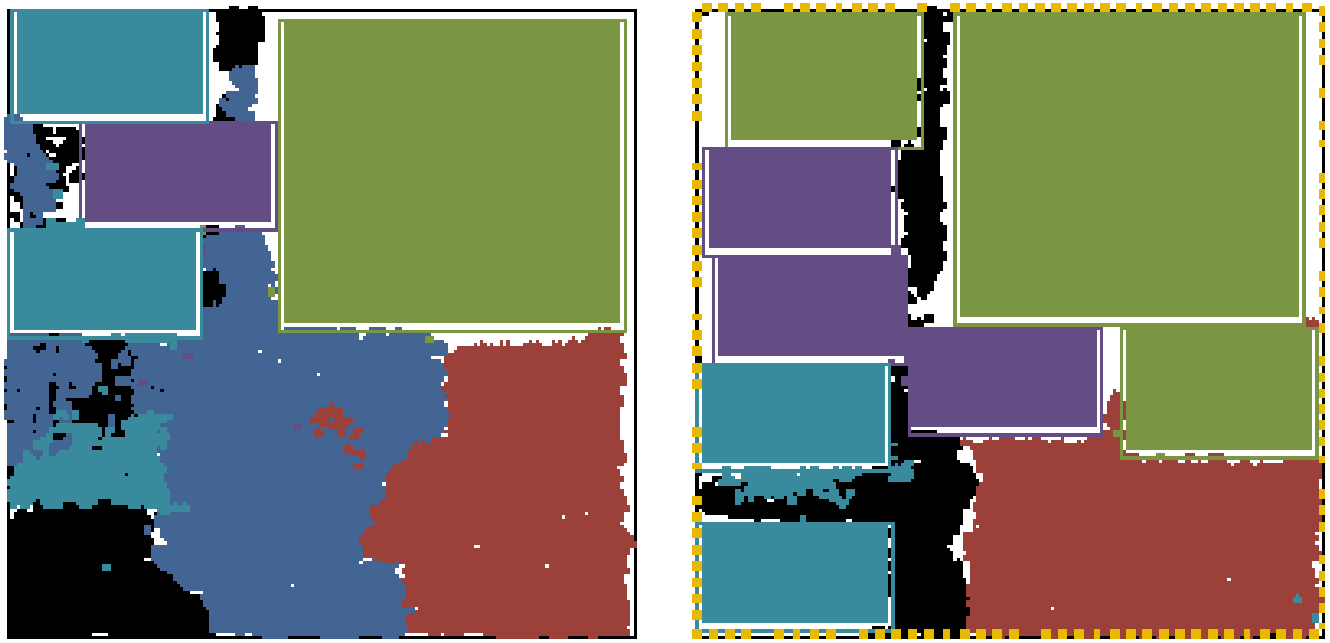




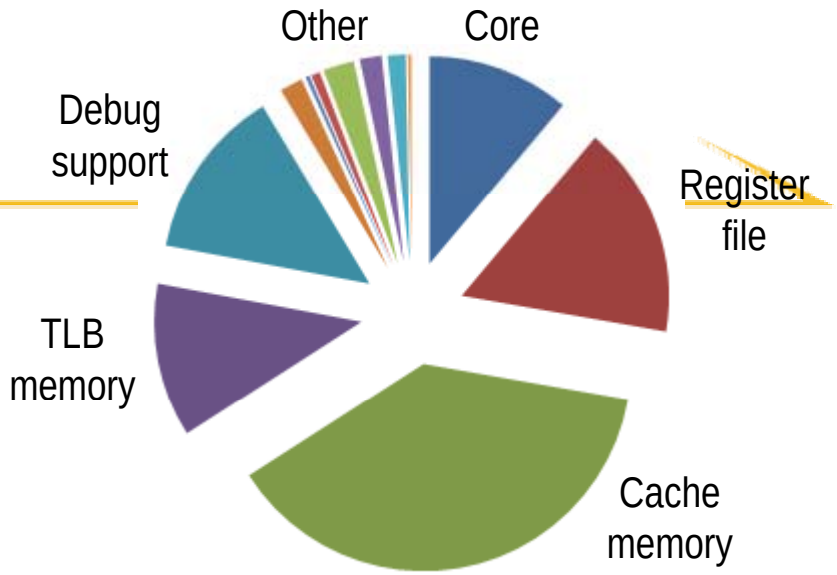
## *Example 2 Processor Core Restricted*



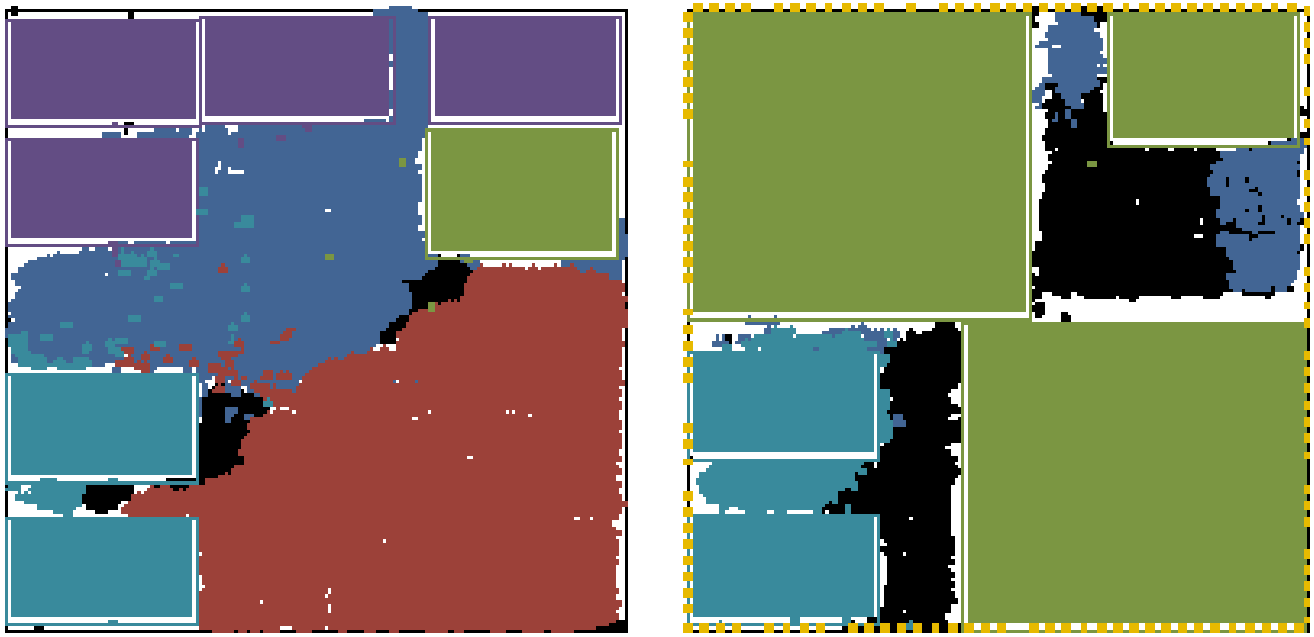

- **HPWL = 1.09 (m)**
- **#TSV = 1715**
- **Placement (bottom, top)**



## *Example 2 Register File Restricted*




- **HPWL = 1.20 (m)**
- **#TSV = 845**
- **Placement (bottom, top)**



# Further Discussions of Example 2

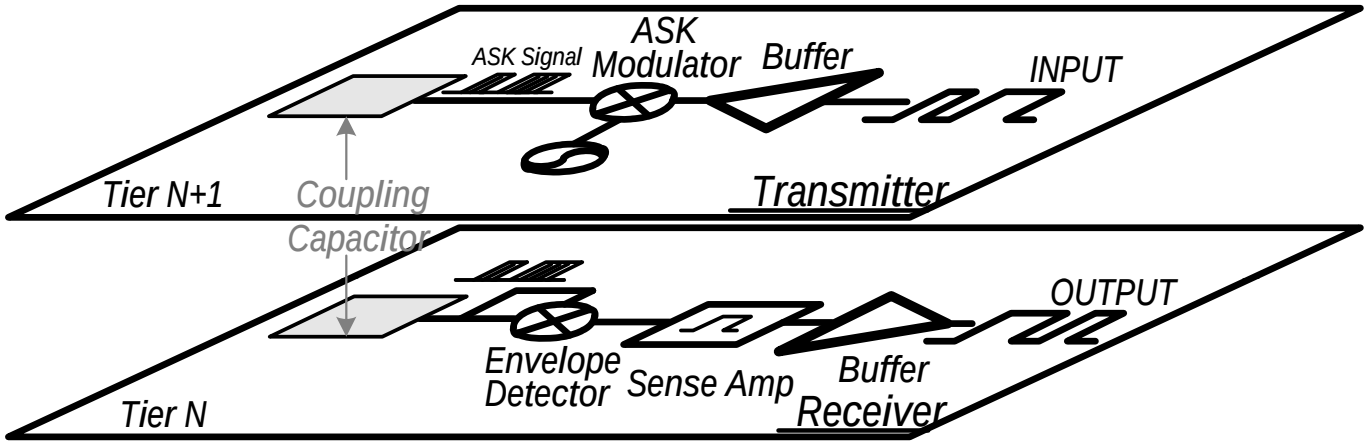
- **Comparisons**

| | Flat | Processor Core restricted | Register file restricted |
|---|---|---|---|
| HWPL | 0.99 (m) | 1.09 (m) | 1.20 (m) |
| #TSV | 3835 | 1715 | 845 |

  - **Example: TSV in MIT Lincoln 180nm SOI 3D technology**
    - Resistance: one TSV is equivalent to a 8-20 μm metal 2 wire
    - Capacitance: one TSV is equivalent to a 0.2 μm metal 2 wire

- **Conclusion**
  - **TSV impact on the RC is not significant**
  - **Some logical units are preferred to be distributed on different device layers**
    - **E.g., the register file in the LEON3 circuit**
  - **Flat 3D placement is preferred to optimize total RC**

# 3D Capacitive RF-Interconnect




***NRZ baseband signal is up-converted by an RF carrier at the transmitter (tier N+1) using ASK (Amplitude-Shift-Key) modulation; and an RF envelope detector at the receiver (tier N) recovers NRZ data in the receiver.***