Thermal-Aware 3D IC Physical Design and Architecture Exploration

Jason Cong & Guojie Luo

UCLA Computer Science Department

cong@cs.ucla.edu

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

Supported by DARPA
Outline

- Thermal-Aware 3D IC Physical Design Flow (Joint work with IBM and PennState)
  - Thermal Models and Assumptions
  - 3D Routing with Thermal Via Planning
  - 3D Placement
  - 3D Floorplanning

- 3D applications (Joint work with Glenn Reinman)
  - 3D Architecture Exploration

- Summary
3D Physical Design Flow (IBM, UCLA, and PSU)

- Layer & Design Rules (LEF)
- Cell & Via* definitions (LEF)
- Netlist (HDL or DEF)

3D RC extraction
EinsTimer
Timing Interface
3D DRC & 3D LVS

3D OA
Tech. Lib
Ref. Lib
Design

PSU

UCLA

Thermal-Driven 3D Floorplanner
Thermal-Driven 3D Placer
3D Global Router
Thermal-Via Planner
Detailed Routing by Cadence Router

Layout (GDSII)
2D OA
Tier Export
Tier Import
Thermal Resistive Network [Wilkerson04]

- Circuit stack partitioned into tiles
- Tiles connected through thermal resistances
  - Lateral resistances: fixed
  - Vertical resistances $\propto 1/\#\text{via}$
- Heat sources modeled as current sources
  - Current value = power
- Heat sinks modeled as ground nodes

Accurate and slow
Thermal Resistive Chain Model

One-Dimension Heat Flow Analysis

- Elmore delay-like formula [Chiang01]

\[ T_4 = \sum_{i=1}^{4} \left( R_i \sum_{j=i}^{4} P_j \right) \]

- Reduce R: thermal via insertion (routing)
- Permute P: floorplanning

Fast and rough
Multilevel TS-Via Planning and 3D Routing (TMARS)

(1). Power Density Calculation
(2). Heat Flow Estimation
(3). Routing Resource Estimation

(1). Power Density Coarsening
(2). Heat Flow Estimation
(3). Routing Resource Coarsening

Thermal Resistive Network Model

(1). Init Routing Tree Generation
(2). TTS Via Planning
(3). TTS Via Number Adjustment

(1). Routing Refinement
(2). TTS Via Planning
(3). TTS Via Number Adjustment

Downward Pass

Upward Pass

ASPDAC’05
Experimental Results — Temperature Reduction

- With thermal via insertion, temperature can be reduced to the required temperature (77°C)
- Thermal via insertion can reduce the maximum on-chip temperature by over 40%
Temperature Maps of ami33 Top Layer

Before Thermal Via Insertion  After Thermal Via Insertion
3D Placement Problem

Problem Formulation

- Minimize
  - \( WL(x,y,z) + \text{viaCost}(x,y,z) \)
- Subject to
  - Overlap-free condition

WireLength (WL)

- Bounding box model

Via Cost (viaCost)

- Area consumption
- Density congestion

Possible Layout

Placement Model
3D Placement Algorithms

- 2D to 3D transformation by folding/stacking
- 3D placement by nonlinear optimization
3D Placement [ASPDAC’07]

2D to 3D Transformation by Local Stacking
– leveraging the best 2D placers (e.g. mPL6)

1. 2D placement on area $K*A$
   - For 3D chip with $K$ device layers and each with area $A$

2. Shrink: $(x_i, y_i) \rightarrow \left( \frac{x_i}{\sqrt{K}}, \frac{y_i}{\sqrt{K}} \right)$

3. Tetris-style 3D legalization
   - Cost $R = \alpha d + \beta v + \gamma t$
   - Minimize displacement, #via and thermal cost
2D to 3D Transformation by Folding

- Layer assignment and location mapping according to the folded order

- Folding-2

- Folding-4
Window-based Stacking / Folding

1. Divide 2D placement into NxN windows
2. Apply stacking or folding in a window
   - Effect of stacking or folding would be spreaded out, and trade-offs are achieved by varying N
3D Placement Results (1/2)

- Wirelength (stacking) compared to 2D mPL5
- Wirelength v.s. # TS via trade-offs

<table>
<thead>
<tr>
<th>circuit</th>
<th>2D mPL5</th>
<th>T3Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibm01</td>
<td>5.19E+06</td>
<td>2.51E+06</td>
</tr>
<tr>
<td>ibm02</td>
<td>1.44E+07</td>
<td>6.95E+06</td>
</tr>
<tr>
<td>ibm03</td>
<td>1.37E+07</td>
<td>6.67E+06</td>
</tr>
<tr>
<td>ibm04</td>
<td>1.67E+07</td>
<td>8.21E+06</td>
</tr>
<tr>
<td>ibm05</td>
<td>4.23E+07</td>
<td>1.94E+07</td>
</tr>
<tr>
<td>ibm06</td>
<td>2.20E+07</td>
<td>1.09E+07</td>
</tr>
<tr>
<td>ibm07</td>
<td>3.73E+07</td>
<td>1.90E+07</td>
</tr>
<tr>
<td>ibm08</td>
<td>3.94E+07</td>
<td>1.98E+07</td>
</tr>
<tr>
<td>ibm09</td>
<td>3.46E+07</td>
<td>1.78E+07</td>
</tr>
<tr>
<td>ibm10</td>
<td>6.82E+07</td>
<td>3.61E+07</td>
</tr>
<tr>
<td>ibm11</td>
<td>5.02E+07</td>
<td>2.51E+07</td>
</tr>
<tr>
<td>ibm12</td>
<td>7.58E+07</td>
<td>3.78E+07</td>
</tr>
<tr>
<td>ibm13</td>
<td>6.58E+07</td>
<td>3.30E+07</td>
</tr>
<tr>
<td>ibm14</td>
<td>1.42E+08</td>
<td>7.40E+07</td>
</tr>
<tr>
<td>ibm15</td>
<td>1.65E+08</td>
<td>8.42E+07</td>
</tr>
<tr>
<td>ibm16</td>
<td>2.04E+08</td>
<td>1.06E+08</td>
</tr>
<tr>
<td>ibm17</td>
<td>3.05E+08</td>
<td>1.60E+08</td>
</tr>
<tr>
<td>ibm18</td>
<td>2.43E+08</td>
<td>1.28E+08</td>
</tr>
<tr>
<td>avg.</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>
### Effect of temperature optimization

<table>
<thead>
<tr>
<th>circuit</th>
<th>LST, r = 10%,</th>
<th>LST, r = 10%, w/ temp optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Temp. (°C)</td>
<td>WL</td>
</tr>
<tr>
<td>ibm01</td>
<td>276.5</td>
<td>2.81E+06</td>
</tr>
<tr>
<td>ibm03</td>
<td>196.7</td>
<td>7.13E+06</td>
</tr>
<tr>
<td>ibm04</td>
<td>159.6</td>
<td>9.11E+06</td>
</tr>
<tr>
<td>ibm06</td>
<td>160.4</td>
<td>1.23E+07</td>
</tr>
<tr>
<td>ibm07</td>
<td>107.5</td>
<td>2.01E+07</td>
</tr>
<tr>
<td>ibm08</td>
<td>97.7</td>
<td>2.05E+07</td>
</tr>
<tr>
<td>ibm09</td>
<td>96.1</td>
<td>1.94E+07</td>
</tr>
<tr>
<td>ibm13</td>
<td>249.3</td>
<td>3.47E+07</td>
</tr>
<tr>
<td>ibm15</td>
<td>136.5</td>
<td>8.58E+07</td>
</tr>
<tr>
<td>ibm18</td>
<td>89.4</td>
<td>1.31E+08</td>
</tr>
<tr>
<td>Avg.</td>
<td>1.0</td>
<td>1.08</td>
</tr>
</tbody>
</table>
3D Placement by Nonlinear Optimization

◆ Problem Formulation

- **Minimize**
  - WL(x,y,z) + viaCost(x,y,z)

- **Subject to**
  - Overlap-free condition

◆ Relaxed Placement Model

◆ Placement Model
Preliminary Results

- Tradeoff curve compared with [ASPDAC’07] on ibm01
  - Achieve as large as 50% #TSV reduction
  - or 12% WireLength reduction
The exploration of the use of vertical integration on microprocessor design requires consideration for both physical design and architecture.

- **True 3D packing**
- **Architectural Alternative Selection**
  - The number of layers in folded blocks
  - The partition way: block folding or port partitioning
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 layers, 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%)
Outline

- Thermal-Aware 3D IC Physical Design Flow (Joint work with IBM and PennState)
  - Thermal Models and Assumptions
  - 3D Routing with Thermal Via Planning
  - 3D Placement
  - 3D Floorplanning

- 3D applications (joint work with Glenn Reinman)
  - 3D Architecture Exploration

- Summary
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 Example

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

- Critical paths

<table>
<thead>
<tr>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wakeup Latency</td>
<td>Latency to wakeup the dependent instruction</td>
</tr>
<tr>
<td>ALU Bypass</td>
<td>latency of the bypass wires between the ALUs</td>
</tr>
<tr>
<td>DL1 Latency</td>
<td>Load latency though the L1 data cache</td>
</tr>
<tr>
<td>L2 Latency</td>
<td>Latency for access to L2 cache</td>
</tr>
<tr>
<td>MPLAT</td>
<td>latency through the branch resolution path</td>
</tr>
</tbody>
</table>
Performance Impact of 3D Integration

Over 35% performance improvement
5GHz 3 Device Layer Layout
**Temperature Impact of 3D Integration**

**2D Design**

**3D Design**

- **w/o thermal via**
- **w/ thermal via**

1st Layer

2nd Layer
Summary

- **Complete Set of Thermal-Aware 3D IC Physical Design Tool**
  - 3D Routing with Thermal Via Planning
  - 3D Placement
  - 3D Floorplanning
  - Ongoing collaboration with IBM and PennState to include 3D parasitic extraction, timing analysis, etc.

- **3D Architecture Exploration & Design Drivers**
  - Coupled with 3D physical planning
    - Consider both 3D component stacking and folding
3D Multicore Processor with RF-Interconnects

- **Three Silicon Layers**
  - Tier 3: Cache Data Components
  - Tier 2: Interconnect and Cache Tags
  - Tier 1: Cores

- **Non-Uniform Cache Access**
  - Cores see different latencies to different cache banks
  - Data can migrate among distributed caches
    - Can hide latency
    - Adds interconnect traffic