Physical Planning with Retiming

Jason Cong and Sung Kyu Lim
UCLA Computer Science Department

Support: MARCO/DARPA GSRC and a grant from Intel Corporation
Outline

• Motivation
• Problem Formulation
• GEO Algorithm
• Experimental Results
• Conclusions & Ongoing Works
Motivation

- DSM VLSI technology
  - Interconnect delay far exceeds gate delay
  - Post-layout optimization is still not enough [NTRS97, Cong97]

<table>
<thead>
<tr>
<th>Tech (um)</th>
<th>0.25</th>
<th>0.18</th>
<th>0.15</th>
<th>0.13</th>
<th>0.10</th>
<th>0.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate delay (ns)</td>
<td>0.071</td>
<td>0.051</td>
<td>0.049</td>
<td>0.046</td>
<td>0.039</td>
<td>0.022</td>
</tr>
<tr>
<td>1mm (ns)</td>
<td>0.059</td>
<td>0.049</td>
<td>0.051</td>
<td>0.044</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td>2cm un-opt (ns)</td>
<td>2.08</td>
<td>1.97</td>
<td>2.06</td>
<td>2.07</td>
<td>2.89</td>
<td>3.52</td>
</tr>
<tr>
<td>2cm opt (ns)</td>
<td>0.89</td>
<td>0.79</td>
<td>0.77</td>
<td>0.70</td>
<td>0.77</td>
<td>0.67</td>
</tr>
<tr>
<td>Exp. clk (GHz)</td>
<td>0.8</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
<td>2.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Exp. dly (ns)</td>
<td>1.3</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- Multiple clock cycle is necessary for long wires
- Interconnect-centric design methodology is essential
Role of Partitioning in Interconnect-Centric Design Flow

- **New perspective**
  - Conventional view: enables divide-and-conquer
  - DSM view: defines global and local interconnects
  - Small blocks (50K-100K gates) can be synthesized reliably
  - Key is simultaneous optimization of delay and cutsize

\[ D \gg d \]
Impact of Retiming during Partitioning

- Wider solution space [Cong et al, DAC99]
  - allows retiming to hide some global interconnect delays

\[ \varphi(A) = 8 \]
\[ \varphi(B) = 8 \]

\[ D = 2, \ d = 1 \]

same cutsizes

Partitioning A

\[ \varphi(A) = 8 \]

\[ \varphi(A) = 8 \]

Partitioning B

\[ \varphi(B) = 8 \]

\[ \varphi(B) = 6 \]
Impact of Floorplanning in Physical Planning

- Limitation of non-geometric delay model
  - May mislead partitioning with retiming

Gate delay model [ICCAD91]
\[ d(p_1) = d(p_2) \]

Geometric delay model
\[ d(p_1) < d(p_2) \]
Problem Definitions

• **Problem instance**
  – netlist $NL = (C, N)$, area $a(c)$ and delay $d(c)$ for each cell $c$, and area constraints $A = (a_i, b_i)$ for each partition

• **Performance driven partitioning**
  – is there a partition $P$ of $C$ into non-empty disjoint sets $B_1, B_2, \ldots, B_K$ such that $a_i \leq |B_i| \leq b_i$, and such that $\alpha \cdot \text{cutsize} + \beta \cdot \text{delay}$ is minimized?

• **Physical planning**
  – is there a partition $P$ of $C$ into non-empty disjoint sets $B_1, B_2, \ldots, B_K$ with their locations such that $a_i \leq |B_i| \leq b_i$, and such that $\alpha \cdot \text{cutsize} + \beta \cdot \text{delay} + \gamma \cdot \text{wirelength}$ is minimized?
Existing Algorithms

- **PKL Algorithm [Pan et al, TCAD98]**
  - First proposed sequential arrival time
  - Quasi-optimal clustering with retiming
  - Limitation: space $O(n^2)$ and time $O(n(n+m)\log^2 n)$ complexity

- **PRIME Algorithm [Cong et al, DAC99]**
  - Overcome PKL’s complexity problem
  - Significant reduction of candidate set for label update
  - Limitation: large cutsize

- **HPM Algorithm [Cong et al, DAC00]**
  - Overcome PRIME’s cutsize problem
  - Cutsize of hMetis + delay of PRIME
  - Limitation: non-geometric delay model
Our Contribution: GEO Algorithm

• Limitations of existing approaches
  – Separation of partitioning, retiming, and floorplanning
  – Partitioning: unrealistic delay model
  – Retiming: FF locations are limited
  – Floorplanning: global interconnects are already fixed

• GEO algorithm: unified approach
  – Tight-couple partitioning, retiming, and floorplanning
  – Partitioning: exploits geometric delay model
  – Retiming: retime on global interconnects
  – Floorplanning: global perturbation possible
Multi-level Method

- Recursive coarsening and refinement
  - From coarse-grain into finer-grain optimization
  - Successfully used in circuit partitioning [Karypis et al, DAC97] and placement [Cong et al, ICCAD00]
GEO Algorithm: Overview

- Multi-level block placement
  - Bottom-up multi-level clustering [Cong and Lim, ASPDAC00]
  - Top down cell move based multi-level and multiway partitioning [Cong and Lim, ICCAD98]
Sequential Arrival Time (SAT)

- **Definition [Pan et al, TCAD98]**
  - \( l(v) = \text{delay from PIs to } v \text{ after retiming under given } \phi \)
  - \( l(v) = \max\{l(u) - \phi \cdot w(u,v) + d(u,v) + d(v)\} \)
  
  ![Diagram](image)
  - \( l(u) \) \( w(u,v) \) \( d(v) \)

- **Relation to retiming**: \( r(v) = \left\lceil \frac{l(v)}{\phi} \right\rceil - 1 \)

- **Theorem**: \( P \) can be retimed to \( \phi + \max\{d(e)\} \iff l(\text{POs}) \leq \phi \)

  ![Diagram](image)
  - \( l(u) = 7 \) \( u \) \( d(v) = 1, \; d(e) = 2, \; \phi = 5 \)
  - \( l(w) = 3 \) \( w \) \( l(v) = \max\{7-5\cdot1+2+1, \; 3+2+1\} = 6 \)
Computation of SAT

• **Single source longest path algorithm**
  – Positive loops and negative edge length due to FFs
  – Bellman-Ford variant
  – Requires multiple iterations before convergence
  – Complexity: $O(n^2)$, practically $O(n)$
Sequential Required Time (SRT)

- **Definition**
  
  - $q(v) = \text{timing constraint for } v \text{ after retiming under given } \phi$
  
  - $q(v) = \min\{q(u) + \phi \cdot w(v,u) - d(v,u) - d(v)\}$

  ![Diagram](image)
  
  - $d(v) = 1$, $d(e) = 2$, $\phi = 5$
  
  - $q(v) = \min\{3+5\cdot1-2-1, 7-2-1\} = 4$

  - Can be computed together with $l(v)$ and $r(v)$
  
  - Slack $s(v) = q(v) - l(v)$

  ![Diagram](image)
  
  - $q(u) = 3$
  
  - $q(w) = 7$
Sequential Timing Analysis with Consideration of Retiming

- **Complete path analysis**
  - Compute SAT, SRT, and slack
  - Perform at the original circuit: expensive $O(n)$
    - Once per cell move vs once per pass

- **Optimization engine: cell move based partitioning**
  - Increase edge weights in e-network
  - Gain represents reduction in total weighted edge lengths
  - Runtime overhead negligible
    - Gain update can be done by examining neighbors
  - Works well in multi-level partitioning framework
**FF Placement**

- **Fine-grained FF placement**
  - Limitation of existing scheme: place FF at front of cell
  - Geometric embedding: place FF on edges
  - **Old Theorem:** $P$ can be retimed to $\phi + \max\{d(e)\}$ iff $l(\text{POs}) \leq \phi$
  - **New Theorem:** $P$ can be retimed to $\phi + \max\{d(v)\}$ iff $l(\text{POs}) \leq \phi$
Summary of GEO Algorithm

- Geometric embedding based multi-level partitioning with retiming = coarse placement with retiming
- Complexity is $O(n \log n)$

Perform multi-level clustering
From top to bottom
- perform STA
- find e-network from current level
- perform cell move
- generate more cutlines

Perform FF placement
**Experimental Settings**

- **Benchmarking**
  - 7 MCNC and 4 industrial ckt (upto 100K cells)
  - geometric embedding onto 8x8 grid
  - measure delay, cutsize, and wirelength

\[
delay = 6d_i + 5D_j
\]
\[
cutsize = 5
\]
\[
wirelength = 2 + 5 + 4 + 4 + 5
\]
Experimental Results

• Comparison with existing algorithms
  – hMetis [DAC97] + retiming + slicing floorplan [Algo89]
  – HPM [DAC00] + slicing floorplan [Algo89]
Conclusions & Ongoing Work

• Paradigm shift
  – Interconnect delay dominates performance
  – Partitioning defines global and local interconnects
  – Proper physical planning allows retiming to hide (some) global interconnect delays
  – Coarse placement in physical planning is key for accurate delay estimation

• Algorithms for physical planning developed at UCLA
  – PRIME [DAC99], HPM [DAC00], and GEO [ICCAD00]

• Ongoing work
  – Floorplanning with interconnect planning
  – Performance-driven cell placement with retiming
## Benchmark Characteristics

- **MCNC and Industrial ckts**

<table>
<thead>
<tr>
<th>Name</th>
<th>GA</th>
<th>PI</th>
<th>PO</th>
<th>FF</th>
<th>Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>s9234</td>
<td>1290</td>
<td>28</td>
<td>39</td>
<td>135</td>
<td>1492</td>
</tr>
<tr>
<td>s5378</td>
<td>1443</td>
<td>35</td>
<td>49</td>
<td>163</td>
<td>1690</td>
</tr>
<tr>
<td>s13207</td>
<td>3146</td>
<td>59</td>
<td>152</td>
<td>486</td>
<td>3843</td>
</tr>
<tr>
<td>s15850</td>
<td>3784</td>
<td>76</td>
<td>150</td>
<td>515</td>
<td>4525</td>
</tr>
<tr>
<td>bigkey</td>
<td>8599</td>
<td>228</td>
<td>197</td>
<td>224</td>
<td>9248</td>
</tr>
<tr>
<td>s38584</td>
<td>13209</td>
<td>38</td>
<td>304</td>
<td>1423</td>
<td>14974</td>
</tr>
<tr>
<td>clma</td>
<td>30552</td>
<td>61</td>
<td>82</td>
<td>33</td>
<td>30728</td>
</tr>
<tr>
<td>ind1</td>
<td>29780</td>
<td>2630</td>
<td>3242</td>
<td>603</td>
<td>36255</td>
</tr>
<tr>
<td>ind2</td>
<td>26060</td>
<td>2772</td>
<td>6242</td>
<td>1755</td>
<td>36829</td>
</tr>
<tr>
<td>ind3</td>
<td>52197</td>
<td>2801</td>
<td>3070</td>
<td>2001</td>
<td>60069</td>
</tr>
<tr>
<td>ind4</td>
<td>101531</td>
<td>4155</td>
<td>4575</td>
<td>8333</td>
<td>118566</td>
</tr>
<tr>
<td>IBM</td>
<td>73238</td>
<td>129</td>
<td>112</td>
<td>74</td>
<td>73883</td>
</tr>
</tbody>
</table>