An Interconnect-Centric Design Flow for Nanometer Technologies

Jason Cong
UCLA Computer Science Department
Email: cong@cs.ucla.edu
Tel: 310-206-2775
URL: http://cadlab.cs.ucla.edu/~cong
Exponential Device Scaling

- **Moore’s Law**
  - The min. transistor feature size decreases by 0.7X every three years (Electronics Magazine, Vol. 38, April 1965)
  - True in the past 30 years!

- **National Technology Roadmap for Semiconductors (NTRS’97)**

<table>
<thead>
<tr>
<th>Technology (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>Year</td>
<td>1997</td>
<td>1999</td>
<td>2001</td>
<td>2003</td>
<td>2006</td>
<td>2009</td>
</tr>
<tr>
<td># transistors</td>
<td>11M</td>
<td>21M</td>
<td>40M</td>
<td>76M</td>
<td>200M</td>
<td>520M</td>
</tr>
<tr>
<td>On-Chip Clock (MHz)</td>
<td>750</td>
<td>1200</td>
<td>1400</td>
<td>1600</td>
<td>2000</td>
<td>2500</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>300</td>
<td>340</td>
<td>385</td>
<td>430</td>
<td>520</td>
<td>620</td>
</tr>
<tr>
<td>Wiring Levels</td>
<td>6</td>
<td>6-7</td>
<td>7</td>
<td>7</td>
<td>7-8</td>
<td>8-9</td>
</tr>
</tbody>
</table>
Global/Local Interconnect Delays vs. Gate Delays

Optimization is obtained buffer insertion/sizing and wire sizing.
Coupling noise from two adjacent aggressors to the middle victim wire of 1mm with 2x min. spacing. Rise time is 10% of project clock period.

- Coupling noise depends strongly on both spatial and temporal relations!
Clock cycles required for traveling 2cm line under BIWS (buffer insertion and wire sizing)

Estimated by IPEM
On NTRS’97 technology

Driver size: 100x min gate
Receiver size: 100x min gate
Buffer size: 100x min gate
How Far Can We Go in Each Clock Cycle

- NTRS’97 0.07um Tech
- 5 G Hz across-chip clock
- 620 mm² (24.9mm x 24.9mm)
- IPEM BIWS estimations
  - Buffer size: 100x
  - Driver/receiver size: 100x
- From corner to corner:
  - 7 clock cycles
Two Important Implications

- Interconnects determine the system performance
  
  Interconnect/communication-centric design methodology

- Need multiple clock cycles to cross the global interconnects in giga-hertz designs
  
  Pipelining/retiming on global interconnects
Interconnect-Centric Design Methodology

- **Proposed transition**

  ![Diagram showing interconnect-centric vs. device/function centric design]

- **Analogy**

  ![Diagram showing data/objects vs. programs]

  - **Data/Objects**
    - Programs
  - **Programs**
    - Data/Objects
Interconnect-Centric IC Design Flow
Under Development at UCLA

Architecture/Conceptual-level Design

Design Specification

Interconnect Planning
- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

Synthesis and Placement under Physical Hierarchy

Interconnect Synthesis
- Performance-driven Global Routing
- Pseudo Pin Assignment under Noise Control

Interconnect Layout
- Route Planning
- Point-to-Point Gridless Routing

Interconnect Optimization (TRIO)
- Topology Optimization with Buffer Insertion
- Wire sizing and spacing
- Simultaneous Buffer Insertion and Wire Sizing
- Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

Interconnect Performance Estimation Models (IPEM)
- OWS, SDWS, BISWS

Final Layout

10/16/00
Interconnect-Centric IC Design Flow
Under Development at UCLA

Architecture/Conceptual-level Design

Design Specification

Interconnect Planning
- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

Interconnect Performance Estimation Models (IPEM)
- OWS
- SDWS
- BISWS

Interconnect Optimization (TRIO)
- Topology Optimization with Buffer Insertion
- Wire sizing and spacing
- Simultaneous Buffer Insertion and Wire Sizing
- Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

Interconnect Synthesis
- Performance-driven Global Routing
- Pseudo Pin Assignment under Noise Control

Interconnect Layout
- Route Planning
- Point-to-Point Gridless Routing
- Final Layout

HDM
- Structure view
- Functional view
- Physical view
- Timing view
Interconnect-Centric IC Design Flow
Under Development at UCLA

Architecture/Conceptual-level Design

Design Specification

Interconnect Planning
- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

Interconnect Synthesis
- Performance-driven Global Routing
- Pseudo Pin Assignment under Noise Control

Interconnect Layout
- Route Planning
- Point-to-Point Gridless Routing

Interconnect Performance Estimation Models (IPEM)
- OWS, SDWS, BISWS

Interconnect Optimization (TRIO)
- Topology Optimization with Buffer Insertion
- Wire sizing and spacing
- Simultaneous Buffer Insertion and Wire Sizing
- Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

Synthesis and Placement under Physical Hierarchy

abstraction
Interconnect Planning

- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning
Physical Hierarchy Generation

- Designs are hierarchical due to high complexity
- Design specification (in HDL) follows logic hierarchy
- Logic hierarchy may not be suitable to be embedded on a 2D silicon surface, resulting poor interconnect designs
  - RT-level floorplanning is a bad idea!
- Solution: transform logic hierarchy to physical hierarchy
Example of Logic Hierarchy in Final Layout

By courtesy of IBM (Tony Drumm)
Example of Logic Hierarchy in Final Layout

By courtesy of IBM (Tony Drumm)
Transform Logic Hierarchy to Physical Hierarchy

- Simultaneous partitioning, coarse placement, and retiming on the flat netlist to generate a good physical hierarchy
  - Synthesis will follow

- Use multi-level optimization to handle with the complexity
Role of Partitioning

- Importance of Partitioning:
  - Conventional view: enables divide-and-conquer
  - DSM view: defines global and local interconnects

![Diagram showing local and global interconnects with D >> d]
Need of Considering Retiming during Partitioning
- Retiming/pipelining on global interconnects

- Multiple clock cycles are needed to cross the chip
- Proper partitioning allows retiming to hide global interconnect delays.

![Partitioning A](image1)

- $f(A) = 8$

![Partitioning B](image2)

- $f(B) = 8$

- $f(A) = 6$

- $f(B) = 8$
Sequential Arrival Time (SAT)

**Definition** [Pan et al, TCAD98]
- \( l(v) = \max \text{ delay from PIs to } v \text{ after opt. retiming under a given clock period } f \)
- \( l(v) = \max \{ l(u) - f \cdot w(u,v) + d(u,v) + d(v) \} \)

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

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

\[
\begin{align*}
l(u) &= 7 \quad \text{ (u)} \\
l(w) &= 3 \quad \text{ (w)} \\
l(v) &= \max \{ 7-5\cdot1+2+1, 3+2+1 \} = 6
\end{align*}
\]
Simultaneous Partitioning/Placement with Retiming

- Minimize SAT during partitioning/placement
- Apply optimal retiming to the resulting solution (best suitable for retiming)
- Partitioning/placement with retiming can be applied recursively to generate physical hierarchy

- Good news: SAT can be computed efficiently (linear time in practice, quadratic time in the worst case)
- Difficulty: Flattened netlist can be very large!
  - Solution: use multi-level method
Multi-level Partitioning

- Iterative coarsening (clustering) to generate a multi-level hierarchy
- Initial partitioning on the coarsest level
- Iterative de-clustering and refinement
Hierarchical Approach vs Multi-Level Approach

- **Hierarchical approach**: higher-level design constrains lower-level designs
  - Not sufficient information at higher-level
  - Mistake at higher level is impossible or costly to correct

- **Multi-level approach**: finer-level design refines coarse-level design
  - Converge to better solution as more details are considered
Example: Multi-Level Partitioning with Coarse Placement & Retiming

- Bottom-up multi-level clustering
- Top down cell move based multi-level partitioning
- Sequential timing analysis at each level

[Cong and Lim, ICCAD00]
Success of Multi-Level Approach

- First used to solve partial differential equations (multi-grid method)
- Successfully applied to circuit partitioning (hMetis [Karypis et al, 1997])
  - Best partitioner for cut-size minimization
- Successfully applied to physical hierarchy generation (HPM and GEO [Cong et al, DAC’00 & ICCAD’00])
  - 30-40% delay reduction compared to hMetis
- Successfully applied to circuit placement [Chan et al, ICCAD’00]
  - 10x speed-up over GordianL
Experimental Results

- Comparison with existing algorithms
  - hMetis [DAC97] + retiming + slicing floorplan [Algo89]
  - HPM [DAC00] + slicing floorplan [Algo89]
  - GEO: simultaneous partitioning + coarse placement + retiming

Close to 40% delay reduction!
Interconnect Planning

- Physical Hierarchy Generation
- **Floorplan/Coarse Placement with Interconnect Planning**
  - Example: Buffer Block Planning in Floorplanning
- Interconnect Architecture Planning
Demand of Buffers in Nanometer Designs

- Need to insert buffers in long global interconnects for performance optimization

<table>
<thead>
<tr>
<th>Technology (um)</th>
<th>0.25</th>
<th>0.18</th>
<th>0.13</th>
<th>0.10</th>
<th>0.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>#buffer per chip</td>
<td>5k</td>
<td>25k</td>
<td>54k</td>
<td>230k</td>
<td>797k</td>
</tr>
</tbody>
</table>

Source: [Cong’97, SRC Work Paper]
http://www.src.org/research/frontier.dgw

( Estimated based on NTRS’97 & [Davis-Meindl’97] )
Buffer Block Planning Problem

[Cong-Kong-Pan, ICCAD’99]

- Restriction from hard IP blocks
- Implications on P/G routing
- Impact on floorplan configuration

=> need to plan ahead for buffers.
Optimal Buffer Location Can Be Relaxed

- **Closed-form** formula of feasible region (FR) for inserting one buffer to meet delay constraint

\[
\begin{align*}
\mathbf{x} &\in [x_{\text{min}}, x_{\text{max}}] \\
x_{\text{min}} &= M \left. A \right. X \left( 0, \frac{K_2 - \sqrt{K_2^2 - 4 K_1 K_3}}{2 K_1} \right) \\
x_{\text{max}} &= M \left. I N \right. \left( l, \frac{K_2 + \sqrt{K_2^2 - 4 K_1 K_3}}{2 K_1} \right)
\end{align*}
\]
Feasible Region (FR) Is Very Large

- Even under tight delay constraint, FR for BI can still be very large!

- Delay budget is \((1+\Delta) T_{opt}\) (the best delay by optimal buffer insertion)

<table>
<thead>
<tr>
<th>Delta</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>19%</td>
</tr>
<tr>
<td>5%</td>
<td>43%</td>
</tr>
<tr>
<td>10%</td>
<td>60%</td>
</tr>
<tr>
<td>20%</td>
<td>86%</td>
</tr>
</tbody>
</table>

=> FR provides a lot of flexibility to plan buffer location
Extension: 2D Feasible Region

- FR extended to 2-dimension with obstacles

source

sink

2-D FR

Locus of min-delay BI (Restricted lines)
Experimental Results of Buffer Block Planning

Buffer block planning reduces # buffer blocks, better meets timing constraints, and use smaller area
Concluding Remarks

- Interconnects determine system performance
- Interconnect-centric design is needed
  - Interconnect planning
  - Interconnect synthesis
  - Interconnect layout
- Physical hierarchy generation is crucial for interconnect planning
- A good combination of partitioning/placement and retiming can hide global interconnect delays, and lead to good physical hierarchy
- Multi-level method is an effective way to cope with complexity