Physical-Aware, High-Capacity RTL Synthesis for Advanced Nanometer Designs

Sanjiv Taneja
Vice President, Product Engineering
Cadence Design Systems
ISQED2013, March 4-6
Santa Clara, CA
Agenda

1. Market trend and challenges
2. Physical effects of interconnect and congestion
3. Physical aware RTL synthesis
4. Hierarchical flow
5. Summary
By 2020 there will be over 10 billion mobile internet devices, and the core of each is a specialized semiconductor:

- Tablets
- Smartphones
- MP3 players
- Gaming devices
- Car electronics
- Mobile video
- Home entertainment
- Wireless appliances

Source: Morgan Stanley
## SoC Design Challenges

<table>
<thead>
<tr>
<th></th>
<th>65nm</th>
<th>40nm</th>
<th>28nm</th>
<th>20nm</th>
<th>14nm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5-1GHz</td>
<td>1-2 GHz</td>
<td>1.5-3GHz</td>
<td>2-5 GHz</td>
<td>&gt;5 GHz</td>
</tr>
<tr>
<td><strong>Design size (instances)</strong></td>
<td>10M</td>
<td>20M</td>
<td>50M</td>
<td>100M+</td>
<td>200M+</td>
</tr>
<tr>
<td><strong>Power Density (W/cm²)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic</td>
<td>100</td>
<td>180</td>
<td>250</td>
<td>425</td>
<td>650</td>
</tr>
<tr>
<td>Leakage</td>
<td>50</td>
<td>120</td>
<td>250</td>
<td>425</td>
<td>650</td>
</tr>
<tr>
<td><strong>Mixed Signal Content</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increasing Mixed Signal content in all SoCs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DFM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRC</td>
<td>DRC, Litho</td>
<td>DRC, Litho, LDE</td>
<td>DRC, Litho, LDE, DPT</td>
<td>DRC, Litho, LDE, DPT, FINFET etc.</td>
</tr>
</tbody>
</table>

Source: IBS
## Physical Interconnect Modeling Challenge
### Impact of Physical Effects

<table>
<thead>
<tr>
<th></th>
<th>65nm</th>
<th>40nm</th>
<th>28nm</th>
<th>20nm</th>
<th>14nm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td>0.5-1GHz</td>
<td>1-2 GHz</td>
<td>1.5-3GHz</td>
<td>2-5 GHz</td>
<td>&gt;5 GHz</td>
</tr>
<tr>
<td><strong>Design size (instances)</strong></td>
<td>10M</td>
<td>20M</td>
<td>50M</td>
<td>100M+</td>
<td>200M+</td>
</tr>
<tr>
<td><strong>Physical Effects for Interconnect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Delay Modeling</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Routing Topology / Detoured Nets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Coupling Capacitance / Slew Degradation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Resistance Estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Layer Awareness / Via Estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Layer Assignment:** At 20nm – resistance per unit length from 20 Ohm/Micron (M1-M5) to 8 Ohm/Micron (M6-M7) and finally to 0.05 Ohm/Micron (M8)

**Via Effect:** At 28nm – wire resistance can increase by 2x due to Via resistance from one routing topology to another.

Source: IBS
GigaScale Design Closure Challenges

- **Floorplan Complexity**
  - 1000+ hard IPs, 1000+ pins

- **Block Closure Complexity**
  - 10-30 MMMC views

- **Hierarchical Assembly Complexity**
  - 50-100M instances
  - 10-25 blocks

- **Timing window**
  - (Setup – Hold)

- **Runtime**

- **Complexity**

- **Chip size**

- **Memory consumption**

- **Timing requirements**

- **IP integration**
Congestion Challenge
Impact of Physical Effects

• Congestion due to poor floorplan
  – Adjusting floorplan can be the solution
    – Macro and port placement

• Congestion due to netlist structure
  – Cannot be fixed in physical or can cost PPA(*)
  – Need re-synthesis for best PPA and convergence in physical

High congestion structures:
  – Cross bars
  – Barrel shifters
  – Memory connected Mux chains

PPA = Performance Power Area
Need physically aware high-capacity synthesis to bridge the gap

Increasing design complexity

Physical design challenges
Logical Nets Dominate
Logical physical gap remains

Band-aids include:
- Over-design
- Multiple iterations

Root cause not addressed in deployed solutions

Industry Solution

Issues & Limitations

No modeling method: ignore wires

Known issues: custom WLM best known method

Rectangular area limitation

Extraction method/routing est. method

Nets Dominate
Logical physical gap remains

Need physical awareness in an actionable logic design context
Successive refinement of wire modeling
Multiple wire abstractions

Wire modeling Accuracy vs. Optimization Flexibility

- Standard Synthesis
  - generic loads
- Physical Layout Estimate Synthesis
  - real loads / fanout (pin cap)
  - Accurate wire estimation
  - detailed modeling for placement / congestion / extraction
- Physical Synthesis
  - Incremental optimization, local structuring
  - Datapath mux structuring
  - Gate selection & timing optimization
  - Sequential duplication, physical aware clock gating, critical region resynthesis

- Implementation

Generic Optimization
  - Technology Mapping
  - Physical Transformations
  - Detailed physical Optimization
Physical Layout Estimation

Wireload Model Replacement Improves Netlist Creation

What is PLE?
- A physical modeling technique to capture timing closure P&R tool behavior for RTL synthesis optimization
  - Result: better timing-power-area balance
- Uses actual design and physical library info
- Dynamically adapts to changing logic structures in the design
- Same runtime as WLM

Does a good job modeling the short wires in a design (80-90%)

Improves QoS and predictability over WLM
Congestion Optimization Techniques
Produce a cleaner design to begin physical implementation

Morphing:
• Incrementally estimates and optimizes congestion
• Uses native, real-time congestion estimation

Global whitespace distribution:
• Re-distribute “whitespace” around placed instances to reduce pin access problem
• Amount of whitespace is calculated based on:
  - Instance pin density, Local congestion severity, Global interconnect

Congestion & placement aware IOpt:
• Dynamically estimate congestion for every move
• Incrementally place new gates to reduce congestion
• Structuring and cell selection is congestion-aware and placement-aware
Congestion driven DFT logic placement

• Challenges:
  – Due to dense connectivity, standard placement algorithms can “clump” compression logic causing local Routing congestion

• Solution:
  – Specialized DFT aware congestion driven placement algorithms can mitigate congestion without disturbing signal path placement
Physical-aware Scan Chain Optimization
Superior results

• Challenges:
  – Lack of physical information
  – Scan chain congestion
  – Impact on timing/SI

• Solution:
  – Scan chain built using DEF physical information
  – Proven to reduce scan wire congestion by 40%
  – Improved balancing
  – Shorter physical scan chains, reducing area and wire congestion
Synthesis capabilities for advanced node

- Leakage optimization (MVt opto) for advanced nodes
  - Enhanced multi-Vt cell selection during Global Mapping
  - Leakage power more significant at 28nm and below

- Improved slew degradation estimation
  - Enhanced slew degradation estimation in RC timing analysis
  - Multi-threaded RC timing calculation to deliver fast runtime with accuracy

- Layer assignment estimation & modeling
  - Estimate layer assignments for critical nets
  - Pass forward layer assignment assumption to physical implementation to ensure convergence

- Advanced OCV (on-chip-variation) support
  - Logic depth based cell delay variation (derating)
  - More accurate than plain OCV with common cell derating value for all cells
Evolution of Design Flow for a Better Timing Closure

- Physical aware logic synthesis
  - Incremental congestion prevention
  - Structural datapath support
  - Physical aware clock gating
  - Physical aware logic structuring
  - Physical aware mapping
Physical Aware Structuring
Minimizes Congestion & Improves Timing

- Target high congestion structures
  - Cross bars
  - Barrel shifters
  - Memory connected Mux chains

- Physical aware
  - Tradeoff Mux structuring

- Physical aware
  - Tree rebalancing (huge inputs)
Improved Timing & Cleaner Placement
Physical Aware Re-Structuring

• Targets OR trees with some shared sub functions
  – Example shown: Four inputs on four sides of module, eight 256-bit OR trees

• RCP results without considering OR tree’s input position
• Results in related logic being pulled to the center

• RCP results when considering OR tree’s input position
• Results in a cleaner placement and improved timing
  – 10% faster & 25% smaller
Physical Aware Mapping
Optimized Timing with Increased Correlation

• Initial map is purely logical
  - “Logic schematic” with wire estimation
  - Long wires not predictable
  - Enables initial placement

• Map w/ placement has register location
  - Automatic path requirement adjustment
  - Mapping w.r.t. long wire-aware timing
  - Logic “squeezed” to meet timing
Hierarchical Flow

• Partitioning/Prototyping
  – Generate partitions from top level design
    – Logical and physical partitions
  – Create timing budget for individual partitions
  – Optimize each partition

• Assembling top level design
  – Assemble partitions together
  – Various models can be used to represent partitions
    – Lib/LEF
    – Full netlist/def/spec
    – ILMs

• Performing Top level timing closure
  – Interface logic optimization
Benefits of using ILMs

• Highly accurate representations of the original design.
  – ILMs do not abstract, they simply discard what is not required for modeling boundary timing.

• Small memory footprint and runtime
  – Up to 90% of the logic can be discarded

• Can be adapted easily to any stage of the design process

• Easy to replace one ILM with another
Summary

- Advanced technology nodes pose new challenges for RTL Synthesis
- Physical effects of interconnect and congestion need to be modeled at RTL for a convergent, predictable flow
- Logic structuring and global mapping need to be physically aware in order to generate layout friendly netlist
- Hierarchical methodology needs to evolve to manage complexity while providing accurate modeling in a bottom-up flow