Bridging the Gap: Co-Optimization of Physical Design and DFM

Gi-Joon Nam, Shayak Banerjee, Kanak Agarwal, Sani Nassif

June 02, 2013
Outline

- Issues with 22nm design flow
- Designer’s wish list
- Fundamental techniques in timing closure
- Lithography-Aware Physical Synthesis (Top-down)
- Co-Optimization of PD and DFM
  - Wire delay variability impact on PD optimization
Typical Design Flow: How to Make a Chip

Typical Design Flow: How to Make a Chip

Chip (Circuit) Entities

How do we manufacture it efficiently and effectively ????
Transistor Manufacturing Process

[Advanced Micro Devices from Quirck/Serda]

[Check out this too http://www.appliedmaterials.com/HTMAC/animated.html]
Then Interconnect Manufacturing
Design Costs are Increasing per Process Node

Source: International Business Strategies, July 2010
Typical Design Flow: What’s missing…..

Typical Design Flow: What’s missing…..

Concerns with today’s design process (I.e. objective functions)

- Timing
- Power
- Turn-around Time
- Routability
- Signal Integrity
- Clocking
- and more…..
Concerns with today’s design process (I.e. objective functions)

- Timing
- Power
- Turn-around Time
- Routability
- Signal Integrity
- Clocking
- and more.....

Where is DFM Optimization???

Where is Lithography Optimization???
Why? The scale of problem is extraordinary

Even without any manufacturability or lithography awareness
Annotated Design Flow of today

Outline

- Issues with 22nm design flow
- Designer’s wish list
- Fundamental techniques in timing closure
- Lithography-Aware Physical Synthesis (Top-down)
- Co-Optimization of PD and DFM
  - Wire delay variability impact on PD optimization
Designer’s Wish List

- Good (global) optimization engines
  - For each objective function
  - Good balance among multiple objectives
  - Turn-Around Time

Routability example: same netlist, two different placement routing algorithms
Designer’s Wish List

- **Stability**
  - Multiple iterations are inevitable
  - Totally new solution from iteration to iteration is a disaster
  - With delta input change, we expect delta output change
Designer’s Wish List

- ECO friendly
  - No surprise
  - Minimum perturbation to the current solution
  - Incomplete (infeasible) solution is better than no solution at all
Outline

- Issues with 22nm design flow
- Designer’s wish list
- Fundamental techniques in timing closure
  - Courtesy of Paul Villarrubia, Chuck Alpert, Zhou Li, Shyam Ramji, Natarajan Viswanathan, Cliff Sze and many folks from IBM EDA
- Lithography-Aware Physical Synthesis (Top-down)
- Co-Optimization of PD and DFM
  - Wire delay variability impact on PD optimization
PDS (Placement-Driven Synthesis) Timing Closure Flow

**place**

- Pre-processing
  - Congestion-driven placement
    - Early Optimization
      - Timing-driven placement
        - Late Optimization
          - Detailed Placement

**refine**

- Placement-Driven Optimization
Typical Timing Closure Tool

- Eliminate iterations
- Reduce Turn-Around-Time (TAT)
- Tight integration of relevant tools
  - Floorplanning/Placement
  - Logic transforms to correct timing
    - Gate Sizing
    - Buffering
    - Logic restructuring
    - Interconnection restructuring
    - Physical location change
  - Routing
    - Congestion-aware
    - Noise-aware
Placement

- **Input**: a netlist of connected gates and nets
- **Output**: exact locations of circuit elements
- **Optimization**: wire length, congestion, noise etc
Key Themes in Placement

- Placement problem consists of optimizing three orthogonal components:
  - Relative order
  - Spacing
  - Global position

- All in within the context of routability, timing and signal integrity

- Placement within timing closure system is especially sensitive to stability
Analytical Partitioning Placer vs. Force-directed Placer
Buffering: ITRS Projections

This optimistic trend depends on advances in materials that may not be achievable (ITRS red bricks).

Source: Chia Hong Jan, IEDM 2003 Interconnect Short Course
Simple Buffer Insertion Problem

Find: Buffer locations and a routing tree such that slack at the source, $q(s_0)$, is maximized

$$q(s_0) = \min_{1 \leq i \leq 4} \{RAT(s_i) - delay(s_0, s_i)\}$$
Buffering Types

- High fanout buffering

- Long net buffering
Gate Sizing

- Size does matter
- Bigger gate means smaller delay
- Loop through all power levels
- Find best slack without electrical violations
Outline

- Issues with 22nm design flow
- Designer’s wish list
- Fundamental techniques in timing closure
- Lithography-Aware Physical Synthesis (Top-down)
  - Litho/DFM enhanced Design Rules
  - DPT-enhanced routing
  - DPT-conflict removal (ICCAD 2011)
- Co-Optimization of PD and DFM
  - Wire delay variability impact on PD optimization
Lithography-based Manufacturing Process

Light source 193nm

Mask (without optical proximity correction)

Light beam

Mask (with optical proximity correction)

Focusing lens

Wafer

Desired circuit pattern

Exposed circuit pattern

OPC increases the amount of data generated from fracturing

130nm feature

[Schellenberg, IEEE Spectrum’03]
Lithography: Wavelength-based Scaling

Lithography resolution is given by a simple equation:

\[
\text{Resolution (Minimum Pitch/2)} = k_1 \frac{\lambda}{\text{NA}}
\]

NA = numerical aperture = sine of largest diffracted angle captured by the lens.
Not a good parameter for scaling:
• Lens manufacturing becomes hard at NA > .5
• depth-of-focus scales as 1/NA²

\[k_1 = \text{Rayleigh parameter} = \text{measure of lithography complexity, yield is affected at } k_1 < .65\]

For many generations, wavelength (\(\lambda\)) was the primary scaling parameter.
Lithographic Process Variation

Loss of Pattern Fidelity: Shorts & Pinches
Process Variations………ahhhhh

249,403,263 Si atoms
68,743 donors
13,042 acceptors

D. J. Frank et al, Symp. VLSI Tech., 1999
Design Rule Enhancement for DFM

- Design rule example: drawback of via from the metal line-end
  - In 32nm technology, contacts are allowed to be drawn flush on the line-end
  - But printing control on line-ends is poor – so for better yield, the recommended rule asks for a certain amount of drawback

<table>
<thead>
<tr>
<th>505a</th>
<th>-</th>
<th>CA must be within M1.</th>
<th>≥</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>505aR</td>
<td>-</td>
<td>CA must be within M1</td>
<td>≥</td>
<td>x</td>
</tr>
</tbody>
</table>

Excerpt from 32nm design manual

- More constraints at the cost of performance
- Always works (Safeties)
- Coverage and growing size issues
Double Patterning Technology (DPT)

C. Mack, IEEE Spectrum 08
Double Patterning Technology (DPT)

- 436 nm
- 365 nm
- 248 nm
- 193 nm

Feature Size (nm)

0.10
0.01
10
10,000
100
10
100,000

Wavelength (nm)

K1

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

2D Practical Limit

1D Practical Limit

0.5 NA
0.75 NA
0.85 NA
1.2 NA
1.35 NA

ArF (193nm wavelength)

Double patterning
DPT-Enhanced Routing

- The most researched area in DPT-aware CAD tools
- Mostly iterative method
  - Route/Check/Rip-up-Reroute for conflict removal
  - DPT-conflict is not necessarily local
- Constructive phase solution starts to emerge
  - Design Rule-based one
  - Simultaneous routing/color assignment and stitch generation
  - Good recipe for ECO routing
Router needs to solve 2-color mapping internally.

Correct by construction: route on pre-colored tracks

Iterative optimization: route, check, rip-up/re-route:

Hybrid solution: use reasonably restrictive starting rules (limit to even jogs, conservative tip-to-side), route, check, fix limited number of conflicts
DPT Conflict Resolution Framework

- Double Patterning’s biggest challenge
  - Coloring conflicts
- Stitching cannot resolve all conflicts
- Odd cycle in post-stitching conflict graph → native conflict

*Image: C. Mack, IEEE Spectrum 08*
Mask Assignment – Preferred coloring

- Coloring of odd cycle affects efficiency of conflict removal
- Give preference for opposite coloring for certain violations over others → label violations critical vs. less-critical
  - E.g., horizontal spacing violation more critical than vertical or diagonal in case of vertical poly orientation
LP Formulation to Remove Conflicts

Problem formulation: Min layout perturbation

Min \[ \sum_{i} W_i |X_i - X_i^{\text{init}}| \]

St: \[ X_j - X_i \geq d_{ij}, \forall A_{ij} \]

DP rules

- LP formulation allows fast polynomial time solving
- Working with pre-colored layout
  - Allows legalization across multiple layers simultaneously
  - Solving conflict on one layer cannot create another elsewhere

![Diagram showing layout, constraint graph, and constraints](image)
Conflict Removal Example

Original
5 conflicts

Same area,
2 conflicts

No conflicts, 6.2% area increase
Outline

- Issues with 22nm design flow
- Designer’s wish list
- Fundamental techniques in timing closure
- Lithography-Aware Physical Synthesis (Top-down)
- Co-Optimization of PD and DFM
  - Wire delay variability impact on PD optimization
Motivation

- Growing contribution of wire delay to overall circuit delay
- Reducing wire dimensions ($W, l$) with scaling
  - Resistance ($R$) x Capacitance ($C$) represents wire delay
  - RC product scales poorly with feature shrinking
- But what about the nature of wire delay variability?
  - Explore the characteristics of variability in wire delay
  - Highlight important trends
  - Evaluate impact on physical design

Source: Shekhar Borkar, Intel
Wire Delay Models - 1

- Resistance and capacitance can be modeled as functions of wire dimensions
- Resistance per unit length
  \[ R_w = \frac{\rho}{W \cdot T} \]
- Capacitance to ground plane per unit length (Sakurai et. al. IEEE Trans. Elec. Devices)
  \[ C_{wg} = \varepsilon \left( k_1 \frac{W}{H} + k_2 \left( \frac{T}{H} \right)^\gamma \right) \]
- Separating out primary dimension of interest - wire width (W)
  \[ R_w C_{wg} = \alpha + \frac{\beta}{W} \]
Wire Delay Models - 2

- Capacitance to neighboring wire can be similarly written as a function of width (W) and space (S)

\[ C_{ww} = \varepsilon \left( k_3 \frac{W}{H} + k_4 \left( \frac{T}{H} \right)^\gamma - k_5 \left( \frac{T}{H} \right)^\gamma \right) \left( \frac{S}{H} \right)^{-\eta} \]

- Total capacitance is sum, so the total RC product can be similarly written as a function of width

\[ C_w = C_{wg} + C_{ww} \]

\[ R_w C_w = \alpha + \frac{\beta}{W} + \alpha' S^{-\eta} + \beta' \frac{S^{-\eta}}{W} \]
Sample Values of $\alpha$, $\beta$

- $\alpha$, $\beta$, $\alpha'$, $\beta'$ are shown for five layers of 1x metallization
  - M1 through M5

- $\alpha$, $\beta$ are at least an order of magnitude greater than $\alpha'$, $\beta'$
  - Especially at M1 and M2
  - Implies that $(1/W)$ dependence is greater than space $(S)$ dependence

<table>
<thead>
<tr>
<th>Level</th>
<th>$\lambda$</th>
<th>$\beta$</th>
<th>$\alpha'$</th>
<th>$\beta'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>2.030</td>
<td>1.198</td>
<td>0.011</td>
<td>0.028</td>
</tr>
<tr>
<td>M2</td>
<td>1.243</td>
<td>1.075</td>
<td>0.012</td>
<td>0.029</td>
</tr>
<tr>
<td>M3</td>
<td>0.741</td>
<td>0.876</td>
<td>0.012</td>
<td>0.031</td>
</tr>
<tr>
<td>M4</td>
<td>0.514</td>
<td>0.808</td>
<td>0.014</td>
<td>0.030</td>
</tr>
<tr>
<td>M5</td>
<td>0.222</td>
<td>0.518</td>
<td>0.010</td>
<td>0.036</td>
</tr>
</tbody>
</table>
Wire Delay Sensitivity

- $\alpha$ is wire RC delay per unit length when $W$ is large
  \[ RC = \alpha + \beta \frac{W_{nom}}{W} \]

- $\alpha$, $\beta$ are constants for a particular metal layer

- Observation 1: Asymmetry of sensitivity of RC Delay
  - To width variation around $W_{nom}$
  - Small negative change in $W$ has larger impact on RC than small positive change
Wire Delay Trends

- $\alpha, \beta$ depend on
  - Metal Level: Tall, thin 1x wires different from thicker wires at higher levels of routing
  - Technology: $W, T, H$ scaling at different rates

- Plotting $\gamma = \beta/\alpha$
  - Higher indicates more impact of wire width variability

- Observation 2: Impact of wire width variability on RC delay growing with scaling

\[
RC = \alpha + \beta \frac{W_{\text{nom}}}{W}
\]
What does this mean for Wire Delay Variability?

- Let’s assume (for now) $W$ is normally distributed
  - Monte Carlo sampling in $W$
  - Typical 1x metal values used ($\gamma \approx 4$)
  - Plot distribution of RC delay through simulation

- Distribution is asymmetric
  - Tail to the right longer than tail to the left
  - Skewness is a common statistical measure (third moment)
  - Skewness = +0.833

\[
\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\text{E}[(X - \mu)^3]}{(\text{E}[(X - \mu)^2])^{3/2}}
\]
Another common statistical measure is inter-quantile ratio

- Estimate the 1% and 99% quantiles ($q_1$ and $q_{99}$)
- Compute the ratio of the difference between the quantiles and mean

$$\zeta = \frac{q_{99} - \mu}{\mu - q_1}$$

- Depends largely on percentage of $W$ variation
- Plot shows that inter-quantile ratio can be 2 or more as width variation increases

What does this mean for Wire Delay Variability?
Variation of Wire Dimensions

- Variation in T, H due to chemical mechanical polish (CMP)
  - Typically ~5% [Gupta et al. ICCAD 2003]

- Variation in W due to lithography
  - Typically ~10% or larger
  - Getting worse with feature scaling

- Low-k1 lithography
  - Feature sizes small fraction of wavelength of light used
  - More susceptible to lithographic variations
Variation of Wire Width Due to Lithography

- Primary sources of lithographic variation
  - Dose errors
  - Focus errors
  - Mask / overlay errors

- Impact of dose on width is approximately linear [Smith et. al. SPIE 2002]
  \[ w = w_0 + s_0 \Delta d \]

- Impact of focus on width is parabolic in nature
  - Manufacturing data fitted to parabolic function
  \[ w = w_0 + (s_1 - s_2 (\Delta f)^2) \Delta d - s_3 (\Delta f)^2 \]
Impact of Lithography on RC Variation

- Dose and focus can be assumed to be approximately normally distributed
  - Central limit theorem
  - Zero mean Gaussian processes

\[ w = w_0 + (s_1 - s_2 (\Delta f)^2) \Delta d - s_3 (\Delta f)^2 \]

- Wire width distribution
  - Monte Carlo sampling in dose/focus space

- Observation 3: Lithographic variations push RC delay distribution towards further asymmetry
Impact of Lithography on RC Variation - 2

- RC histogram for typical range of dose/focus variations
- Assumed different ranges of defocus in the process
  - Inter-quantile ratio plotted
  - Shows that amount of skew in RC distribution very quickly increases as focus gets out of control

Inter-quantile Ratio as function of Range of Defocus

\[ RC = \alpha + \beta \frac{w_0}{w_0 + (s_1 - s_2(\Delta f)^2)\Delta d - s_3(\Delta f)^2} \]
What does this mean for Physical Design?

- **Hold (H) and setup (S) time constraints**
  - Defined by late arrival time ($\Delta$) and early arrival ($\delta$) time

\[
\Delta_{\text{clk-src}} + \Delta_{\text{data}} \leq T - S + \delta_{\text{clk-dest}}
\]

\[
\delta_{\text{clk-src}} + \delta_{\text{data}} \geq H + \Delta_{\text{clk-dest}}
\]

- **Most current methodologies consider only nominal wire delay during timing**
  - But M3 1x wire can be 2% slow or 21% fast equally likely!
  - Introduced new terms for interconnect delay $\Delta_{\text{ILate}}$ and $\delta_{\text{IFast}}$
Timing Analysis with Wire Delay Asymmetry

- Late mode timing analysis on sample 32nm ASIC macros with wire delay asymmetry accounted for
  - Compared against no wire delay variability (only nominal wire delay)
  - Total negative slack (TNS) increases by ~15%
  - Total number of paths with negative slack increases by ~12%
  - Paths would have been undetected as causing timing failures if only nominal wire delay was considered

<table>
<thead>
<tr>
<th>CKT</th>
<th>#Nets</th>
<th>Var</th>
<th>#Negs</th>
<th>TNS</th>
<th>Δ Negs</th>
<th>Δ TNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKT1</td>
<td>117K</td>
<td>NO</td>
<td>196</td>
<td>-3123</td>
<td>N.A</td>
<td>N.A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>244</td>
<td>-3784</td>
<td>24.49%</td>
<td>21.16%</td>
</tr>
<tr>
<td>CKT2</td>
<td>180K</td>
<td>NO</td>
<td>650</td>
<td>-10908</td>
<td>N.A</td>
<td>N.A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>763</td>
<td>-13065</td>
<td>17.38%</td>
<td>19.77%</td>
</tr>
<tr>
<td>CKT3</td>
<td>469K</td>
<td>NO</td>
<td>2631</td>
<td>-66711</td>
<td>N.A</td>
<td>N.A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>2882</td>
<td>-75066</td>
<td>09.54%</td>
<td>12.52%</td>
</tr>
<tr>
<td>CKT4</td>
<td>530K</td>
<td>NO</td>
<td>2114</td>
<td>-63140</td>
<td>N.A</td>
<td>N.A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>2302</td>
<td>-69780</td>
<td>08.89%</td>
<td>10.52%</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.69%</td>
<td>14.68%</td>
</tr>
</tbody>
</table>
Solutions for Reducing Asymmetry - 1

- Route critical signals using higher layers
  - Combination of W, L, H not as bad at higher levels of routing
  - ~2x reduction in skew by routing on 2x wires

- Route using wider wires, or at larger pitch
  - Increased W
  - Reduce sensitivity of RC to W
  - Reduce sensitivity of W to defocus variations
  - Increased S – reduces coupling capacitance to neighbors

<table>
<thead>
<tr>
<th>Width Variations</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1X skew</td>
<td>0.44</td>
<td>1.39</td>
<td>2.72</td>
</tr>
<tr>
<td>1X (\zeta)</td>
<td>1.29</td>
<td>2.07</td>
<td>3.25</td>
</tr>
<tr>
<td>2X skew</td>
<td>0.19</td>
<td>0.68</td>
<td>1.33</td>
</tr>
<tr>
<td>2X (\zeta)</td>
<td>1.15</td>
<td>1.50</td>
<td>2.07</td>
</tr>
</tbody>
</table>
Solutions for Reducing Asymmetry - 2

- Perform OPC at defocus condition
  - Parabolic dependence on defocus is major contributor to asymmetry
  - If OPC is performed at zero defocus
    - Probability of printing larger < Probability of printing smaller
    - Solution: Choose to perform OPC at defocus condition ($\Delta f_{OPC}$)
  - Let $\varphi(x)$ represent a standard normal distribution ($0, \sigma_f$)
  - Select $\Delta f_{OPC}$ such that:
    $$\int_{0}^{\Delta f_{OPC}} \Phi(x) dx = \int_{\Delta f_{OPC}}^{\infty} \Phi(x) dx$$
  - Solution is $\Delta f_{OPC} = 0.674 \sigma_f$
Conclusion

- Need to worry about wire delay variability in addition to magnitude
- Demonstrated the asymmetric nature of wire delay variability
  - Skewness increasing with scaling
- Highlighted three major contributing factors
  - Sensitivity of delay to smaller wire widths in current technologies
  - Increase of this sensitivity with scaling
  - Skewed variation of wire width itself due to lithographic variations
- Demonstrated impact on timing analysis
  - Not considering wire delay variability can lead to ~12% more failing paths
- Evaluated techniques to reduce delay asymmetry
  - Design methods – route at higher levels, route using wider wires
  - Manufacturing method – perform OPC at defocus condition
Summary

- Lithography issue will affect design process in 20nm and beyond
  - Current practice is rather ad-hoc with bunch of heuristics mostly at post-design stage

- Co-optimization is desirable
  - Litho-aware design optimization
  - Design-aware litho optimization

- Minimum perturbation-based optimization is important
  - No further source of late surprise