Cross-Boundary Inductive Timing optimization for 2.5D Chiplet-Package Co-Design

MD Arafat Kabir\textsuperscript{1}, Dusan Petranovic\textsuperscript{2}, Yarui Peng\textsuperscript{1}
CSCE Dept., University of Arkansas, Fayetteville, AR\textsuperscript{1}
Mentor, a Siemens Business, Fremont, CA\textsuperscript{2}
Introduction

- Package becomes increasingly critical in post-Moore’s Law era
  - Transistor scaling is saturated, and chips are reaching reticle limit.
  - 2.5D and 3D packages provide high bandwidth and compact size.
  - Novel design techniques like plug-and-play, Drop-in, Hardware security
  - Heterogeneous integration capabilities (AMD Milan-X, Intel Lakefield)
  - Supports large systems with tens of Known-Good-Dies

- Need for a cross-boundary package-aware design strategy
  - Interactions between the package and chiplets are significant
  - Package inductance is expected to play significant role on performance and signal integrity.

- Objectives
  - Study of RDL inductance impact on 2.5D system performance
  - A cross-boundary inductance aware timing optimization flow
Need for Inductance-Aware Flow

Need for cross-boundary Inductance-aware timing optimization
- Package nets are the bottlenecks in a 2.5D system [1]
- Large I/O drivers are used for inter-chiplet communication [2]
- Bigger driver mean large power and area
- Custom drivers can save power, area, and improve performance
- Requires consideration of all circuit elements (RLC) and careful analysis and optimization

Limitation of existing flows
- Support for only RC elements in STA tools
- Driver optimization based on capacitive load only
- Timing impact of interconnect inductance is completely ignored

Interconnect Delay Study

Interconnect model

- Inductance is modeled using the following partial inductance equation
- \( k = l/r \), \( l \) is the length and \( r \) is the thickness of the wire
- At 2 GHz, skin depth of copper is 1.45 \( \mu m \)

<table>
<thead>
<tr>
<th>RDL Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>10 ( \mu m )</td>
</tr>
<tr>
<td>Spacing</td>
<td>10 ( \mu m )</td>
</tr>
<tr>
<td>Thickness</td>
<td>1 ( \mu m )</td>
</tr>
<tr>
<td>Resistance</td>
<td>0.05 ( \Omega/\mu m )</td>
</tr>
<tr>
<td>Capacitance</td>
<td>0.068 fF/( \mu m )</td>
</tr>
<tr>
<td>Partial Inductance</td>
<td>Equation (1)</td>
</tr>
</tbody>
</table>

\[
L_{l,k} = \frac{\mu_0 l}{2\pi} \left[ \ln(\sqrt{k^2 + 1 + k}) - \sqrt{k^{-2} + 1} + \frac{0.9054}{k} + 0.25 \right] \quad (1) \ [3]
\]

Simulation setup

- Nangate45 cell library based on FreePDK45
- 2 GHz pulse source with 10 ps rise/fall time
- Ignoring the IO pads in this simulation
- Multiple simulation runs with different interconnect length
- Both RC and RLC models are simulated
Using only RC model can underestimate the propagation delay by approximately 30%
- Following result is for INV_X16 as the driver
- Consistent with previous studies
We are using a transmission line model to estimate the RLC delay

- The model equation is developed based on some previous studies [3]
- RLC delay is approximated from the RC delay using a scaling factor
- The scaling factor is later used in the parasitic scaling flow

### Line Parameter Definition

<table>
<thead>
<tr>
<th>Line Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_t$</td>
<td>Total line resistance</td>
</tr>
<tr>
<td>$C_t$</td>
<td>Total line capacitance</td>
</tr>
<tr>
<td>$L_t$</td>
<td>Total line inductance</td>
</tr>
<tr>
<td>$C_L$</td>
<td>Total input capacitance of the receiver</td>
</tr>
<tr>
<td>$\zeta_{line}$</td>
<td>Damping ratio of the line</td>
</tr>
</tbody>
</table>

scalingFactor = $k + a\zeta_{line}^3 + b\zeta_{line}^2 + c\zeta_{line} + d\zeta_{line}^2 C_T$

$RC\ Delay = scalingFactor \times RC\ Delay$

$C_T = \frac{C_L}{C_t}, \quad \zeta_{line} = \frac{R_t}{2} \sqrt{\frac{C_t}{L_t}}$

All driver cells are simulated and fitted to the delay model

- Fitted parameters of some of the Nangate45 library cells
- RC model is equivalent of $k=1.0$ and $a=b=c=d=0$
- Larger the driver, larger is the deviation from the RC model due to reduced driver resistance.

$$\text{scalingFactor} = k + a\zeta_{\text{line}}^3 + b\zeta_{\text{line}}^2 + c\zeta_{\text{line}} + d\zeta_{\text{line}}^2 C_T$$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>INV X1</th>
<th>INV X4</th>
<th>INV X16</th>
<th>BUF X1</th>
<th>BUF X4</th>
<th>BUF X16</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.023</td>
<td>0.132</td>
<td>3.312</td>
<td>0.004</td>
<td>-0.036</td>
<td>1.931</td>
</tr>
<tr>
<td>b</td>
<td>0.047</td>
<td>-0.242</td>
<td>-5.783</td>
<td>0.007</td>
<td>0.000</td>
<td>-3.788</td>
</tr>
<tr>
<td>c</td>
<td>-0.013</td>
<td>0.156</td>
<td>2.804</td>
<td>-0.006</td>
<td>0.048</td>
<td>2.085</td>
</tr>
<tr>
<td>d</td>
<td>-0.008</td>
<td>-0.136</td>
<td>-1.009</td>
<td>-0.007</td>
<td>-0.076</td>
<td>-0.561</td>
</tr>
<tr>
<td>k</td>
<td>1.003</td>
<td>1.001</td>
<td>0.961</td>
<td>1.004</td>
<td>1.008</td>
<td>0.938</td>
</tr>
</tbody>
</table>
Our RLC Delay Model

- Our RLC delay model has only 1% error compared to SPICE simulation of the RLC Interconnect model
  - The simulation covers up to 5 mm RDL wirelength
Our Holistic Flow

- **Exchange of cross-boundary design information in planning, design, analysis, and optimization steps**

(a) Holistic Co-Optimization Flow

(b) Inductance Impact Modeling

- **Gate Level Design**
- **Partitioning**
- **Chip-Level Design**
- **Timing Budget Extraction and Hierarchical Sub-Design Formation**
- **Package Plan**
- **Chiplet Plan**
- **RDL Routing**
- **Placement & Routing**
- **Design Assemble**
- **Holistic Extraction**
- **Scaling for Inductance**
- **Analysis and Context Creation**

- **Chip Design Tool**
- **Our In-House Tool**
- **Design Information**

- **Cell Timing Library**
- **Design Data**
- **RC Parasitics**

- **Scaling for Inductance**
  - Parse Design Data and RC Parasitics
  - Calculate RLC equivalent Delay Factors
  - Calculate Ctot, eq, and Parasitics Scaling Factors
  - Scale RDL Net Parasitics to Simulate RLC Delay

- **RLC Equivalent Parasitics**
Parasitics Scaling for Inductance

- **Industry standard tools do not directly support Inductance in the STA and timing optimization steps**
  - Standard parasitics formats (like SPEF) support inductance.
  - STA tools (like PrimeTime) ignore the inductances in timing analysis.
  - PDK and timing optimization do not consider inductance either.
  - Fundamental changes are required in existing tool flows for inductance.

- **We come up with parasitic (RC) scaling as a direct solution**
  - No need to modify the existing timing analysis tool flow.
  - Inherently compatible with the timing optimization flow.
RC parasitics are scaled to emulate RLC equivalent delay

- Our in-house tool performs the parasitic scaling
- Design data and RC parasitics are obtained from P&R tool and holistic extraction
- RLC delay for RDL wires is estimated using our RLC equivalent model
- The scaling factor to emulate RLC delay is determined
- Capacitance values of the RDL wires are scaled to generate RLC equivalent parasitic
Scaling for Inductance

- RLC equivalent parasitics is computed using equation (2)
  - Cell delay depends on input transition and total output capacitance.
  - Net delay is calculated using Elmore delay model.
  - Total Elmore-delay is scaled if all capacitances are scaled keeping all resistances constant

\[
RLC = cell\ delay + net\ delay
= LUT (C_{tot,eq}, tr) + scalePar \times (RC\ net\ delay) \tag{2}
\]

Where,
- \(C_{tot}\) : Total Capacitance in the RC network,
- \(t_r\) : Input transition time of the driver cell,
- \(C_{tot,eq}\) : Equivalent total capacitance required to simulate RLC delay,
- \(LUT\) : Cell timing library look-up table
- \(scalePar\) : \(C_{tot,eq}/C_{tot}\)
ARM Cortex-M0 based micro-controller system

- Consists of an ARM Cortex-M0 core, 16 KB memory, and some common peripheral devices
- Two-chiplet system: Core and Memory
- The 16 KB memory is divided into two parts, 8 KB each.
We use Nangate45nm as our PDK
- M1-M7 used for chiplet routing

We modify the top three layers to include 2.5D package RDLs
- Dimensions are similar to the TSMC 2.5D InFO technology

<table>
<thead>
<tr>
<th></th>
<th>M6</th>
<th>via6</th>
<th>M7</th>
<th>via7</th>
<th>RDL1</th>
<th>viar1</th>
<th>RDL2</th>
<th>viar2</th>
<th>RDL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>2.28</td>
<td>3.08</td>
<td>3.9</td>
<td>7.5</td>
<td>12.5</td>
<td>17.5</td>
<td>22.5</td>
<td>27.5</td>
<td>32.5</td>
</tr>
<tr>
<td>Thickness</td>
<td>0.8</td>
<td>0.82</td>
<td>3.6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Width</td>
<td>0.4</td>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Spacing</td>
<td>0.4</td>
<td>0.44</td>
<td>2</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>

Cont. Pads (via7)
- RDL1 (M8)
- RDL2 (M9)
- RDL3 (M10)
Physical Design

- The system is implemented keeping the chiplets 1 mm apart
  - A small system is easy to control for experimental study
  - RDL wirelength varies between 1-2.50 mm
  - Final system frequency of 300 MHz
  - Two different designs are prepared,
    - Using holistic RC analysis and optimization flow: **RC-Design**
    - Using our proposed flow with parasitic scaling: **RLC-Design**

(a) Assembled 2.5D system with chiplets and the package together
(b) Core-Chiplet
(c) Memory Chiplet
Holistic RC Extraction

- Package and chiplet designs are assembled for holistic extraction
  - The extraction environment has everything together
  - The extraction tool can capture the cross-boundary RC parasitics between the package and chiplets

<table>
<thead>
<tr>
<th>Coupling Capacitance (CCAP)</th>
<th>Metal Layer</th>
<th>M1-M5</th>
<th>M6</th>
<th>M7</th>
<th>RDL1</th>
<th>RDL2</th>
<th>RDL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1-M5</td>
<td>6116</td>
<td>413.1</td>
<td>38.45</td>
<td>57.60</td>
<td>10.13</td>
<td>7.316</td>
<td></td>
</tr>
<tr>
<td>M6</td>
<td>413.1</td>
<td>494.4</td>
<td>92.67</td>
<td>109.2</td>
<td>12.09</td>
<td>10.17</td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td>38.45</td>
<td>92.67</td>
<td>41.11</td>
<td>18.32</td>
<td>2.097</td>
<td>2.354</td>
<td></td>
</tr>
<tr>
<td>RDL1</td>
<td>57.60</td>
<td>109.2</td>
<td>18.32</td>
<td>721.7</td>
<td>2646</td>
<td>45.63</td>
<td></td>
</tr>
<tr>
<td>RDL2</td>
<td>10.13</td>
<td>12.09</td>
<td>2.097</td>
<td>2646</td>
<td>750.2</td>
<td>2623</td>
<td></td>
</tr>
<tr>
<td>RDL3</td>
<td>7.315</td>
<td>10.17</td>
<td>2.353</td>
<td>45.62</td>
<td>2623</td>
<td>1135</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground Capacitance (GCAP)</th>
<th>Metal Layer</th>
<th>M1-M5</th>
<th>M6</th>
<th>M7</th>
<th>RDL1</th>
<th>RDL2</th>
<th>RDL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacitance</td>
<td>21640</td>
<td>2142</td>
<td>288.9</td>
<td>2118</td>
<td>365.3</td>
<td>681.7</td>
<td></td>
</tr>
</tbody>
</table>
RC only analysis and optimization keeps 35% of the paths in timing violation

- These violations remain undetected in RC analysis.
- The worst violation is by 0.15 ns
- Finished system will fail to run at nominal speed
Our optimization flow automatically adjusts drivers for inductance delay overhead

- Smaller drivers are used in the RC-Design.
- Drivers are upsized in the to compensate for the inductance impact.
- This shift in driver size distribution is ONLY to compensate for the inductance overhead.

![Bar chart showing cell count without and with inductance impact for different driver sizes.](chart.png)
Receiver Optimization

- **Receiver cells are adjusted to reduce total load capacitance**
  - Shift in receiver distribution to reduce TOTAL input capacitance
    - Larger receiver cells are downsized
    - Many small receivers (X1) replaced with single a bit larger receiver (X2)
  - **Reduction in total path delay**
    - The logic cell itself if downsized instead of inserting a smaller buffer.

![Cell Count Graph with Receiver Size](attachment:
Cell Count Graph)

<table>
<thead>
<tr>
<th>Design</th>
<th>Path-1</th>
<th>Path-2</th>
<th>Path-3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BUF_X4</td>
<td>AOI21_X1</td>
<td>AOI22_X4</td>
</tr>
<tr>
<td>RC</td>
<td>BUF_X1</td>
<td>NAND4_X1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BUF_X8</td>
<td>XOR2_X1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BUF_X2</td>
<td>BUF_X1</td>
<td></td>
</tr>
<tr>
<td>RLC</td>
<td>BUF_X2</td>
<td>BUF_X1</td>
<td>AOI22_X2</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

Conclusions

- Chiplet-package interactions need to be considered during analysis and optimization of 2.5D systems.
- RDL wire delays are significantly underestimated with RC-only model.
- Our RLC delay model can accurately capture the inductance impact on timing delay through RDL wires.
- Parasitic scaling for inductance is compatible with the existing tools.
- Our parasitic scaling flow performs cross-boundary optimization to reduce RDL overhead.

Future Work

- Model 2.5D interconnects with native RLC models and CAD tools
- Extend our model to multi-point connections
- Signal and Power Integrity Study with RCLM elements
Thank You