Leveraging Correlations in Processing and Operation Variations for Design Optimization

Prof. Lei He
Electrical Engineering Department
University of California, Los Angeles 90095
LHE@ee.ucla.edu
Outline

- Clock synthesis considering time variant temperature gradients
- Floorplanning with workload aware temporal temperature variation
- Decap budgeting considering operation and processing variations
Methodologies for Clock Skew Minimization

- The sources of skew
  - Un-balanced clock distribution
  - Process, supply voltage and temperature (PVT) variation
  - Uncertainty from loading

- Methodologies
  - Active de-skew circuit using micro-controller [Rusu'00]
  - Passive balanced embedding by CAD algorithms [Tsay'91] [Edahiro'91] [Chao'92] [Boese'92] [Cong'98]

Variation-induced skew needs to be considered!
Limitation of Existing Passive Method

- The existing work [Cho et al: ICCAD05] ignores the time-variant temperature variations and assumes a fixed temperature map.
- Different work loads lead to different temperature maps (e.g., two SPEC2000 applications: Ammp and Gzip).
- Optimizing skew for one application hurts the skew for another application, this conflict is solved in this work.
Stochastic Temperature Model

- A temperature map is unique for each application or program phase
  - can be obtained by uArch-level simulation
- For each region of the chip, temperature is characterized by its mean and variance over a number of maps
  - Primary component analysis (PCA) to decide # of maps
- Temperature correlation measured as covariance between regions is high over SPEC2000 benchmark set

Considering temperature correlations during optimization can compress searching space!
Problem Formulation

- Given:
  - The source, sinks and an initial tree embedding
  - A set of temperature maps for a benchmark set

- Design freedoms:
  - Re-embedding of clock tree
  - Cross link insertion

- To Minimize the worst case skew among given temperature maps
Bottom-up Greedy-based Re-embedding

Sink

Original merging point

Re-embedding option
Bottom-up Greedy-based Re-embedding

New merging point
Perturbed Modified Nodal Analysis (MNA)

- $x$ is for source, sinks and merging point
- $L$ selects sink responses
- Defining a new state variable with both nominal ($x$) and perturbed ($\delta x$) state variables 

\[ [(G_0 + \delta G_1) + s(C_0 + \delta C_1)] \cdot (x + \delta x) = Bu \]

\[ [(G_0 + \delta G_2) + s(C_0 + \delta C_2)] \cdot (x + \delta x) = Bu \]

\[ \vdots \]

\[ [(G_0 + \delta G_l) + s(C_0 + \delta C_l)] \cdot (x + \delta x_l) = Bu. \]

\[ (G_P + sC_P)x_P = B_P u, \quad y_P = L_P^T x_P, \]

The number of re-embedding configurations $I=5^N$ is huge!

(N is number of merging points)
Compressing Solution Space by Temperature Correlation

**Motivation**
- Highly correlated merging points should be perturbed in the same fashion

**Solution**
- Calculate correlation between two merging points based on temperature correlations
- Cluster merging points based on correlation strength
- Perform the same re-embedding for all points within one cluster
Temperature Correlation Driven Clustering

- Correlation matrix $C$ of merging points is low-rank, and **Singular Value Decomposition** (SVD) reveals the rank $K$

- Partition the merging points into $K$ clusters (K-Means)
  - Maximize the correlation strength within each of $K$ clusters

- Reduced from $5^{70}$ to $5^{4}$
Recap of Skew with Re-embedding

Cluster based reduction
(SVD + K-Means)

K << N

Delay and Skew

Transient time analysis
(Back-Euler)

[Yu et al., DAC'06]
(Best paper award nominee)
Simultaneous Re-embedding and Cross Link Insertion

1. Decide crosslink candidates according to [Rajaram, DAC04]
2. Cluster crosslink candidates again based on the temperature correlation
3. Calculate skew sensitivities w.r.t. crosslink and re-embedding candidates
   - In a fashion similar to the previous triangular block-wise MOR
4. Bottom-up select the best crosslink or re-embedding
Experimental Settings

- Temperature maps obtained by micro-architecture level power-temperature transient simulator [Liao, TCAD’05] with 6 SPEC2000 applications
- 100 temperature maps, one for 10 million clock cycles
- Compare four algorithms (two categories)
  - Traditional optimization under nominal temperature and Elmore delay
    - **DME**: deferred merging-point embedding to minimize wire-length for zero-skew
    - **xlink**: cross-link insertion [Rajaram, ICCAD'04]
  - The proposed algorithms with temperature variation and high-order delay model
    - **re-embed**: re-embedding
    - **xlink+ Re-embed**: simultaneously re-embedding and cross-link insertion
Skew Distribution Over 100 Temperature Maps

- **X+R** = cross link insertion + re-embedding
- **DME** = Deferred Merging points Embedding
For tree structure, \textit{re-embed} reduces the worst-case skew by 3x on average (up to 20x) compared to \textit{DME}.

For non-tree structure, \textit{xlink+re-embed} reduces the worst-case skew by 30\% on average (up to 7x) compared to \textit{xlink}.
For tree structure, *re-embed* has less than 1% wire length overhead compared to *DME*.

For non-tree structure, *xlink+re-embed* has 5% LESS wire length compared to *xlink*. 
Temperature-aware optimizations (re-embed and xlink+re-embed) is about 10x slower compared to DME and xlink, respectively, but
- Our work uses high-order delay model
- DMR and xlink use Elmore delay

<table>
<thead>
<tr>
<th>Input(node)</th>
<th>DME</th>
<th>xlink</th>
<th>re-embed</th>
<th>xlink+re-embed</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1(267)</td>
<td>0.5</td>
<td>1.1</td>
<td>1.1</td>
<td>1.4</td>
</tr>
<tr>
<td>r2(598)</td>
<td>1.0</td>
<td>3.2</td>
<td>4.5</td>
<td>5.3</td>
</tr>
<tr>
<td>r3(862)</td>
<td>1.4</td>
<td>4.7</td>
<td>6.1</td>
<td>13.2</td>
</tr>
<tr>
<td>r4(1903)</td>
<td>2.1</td>
<td>5.5</td>
<td>33.6</td>
<td>59.0</td>
</tr>
<tr>
<td>r5(3101)</td>
<td>6.2</td>
<td>11.4</td>
<td>86.6</td>
<td>191.4</td>
</tr>
<tr>
<td>ave</td>
<td>2.24</td>
<td>5.18</td>
<td>26.38</td>
<td>54.06</td>
</tr>
<tr>
<td>ratio</td>
<td>1</td>
<td>2x</td>
<td>12x</td>
<td>24x</td>
</tr>
</tbody>
</table>
Outline

- Clock synthesis considering time variant temperature gradients
- Floorplanning with workload aware temporal temperature variation
- Decap budgeting considering operation and processing variations
Decap Budgeting with Correlated Currents and Process Variation

- **Problem Formulation**
  - Given total white space for decap, find the distribution and location of the white space so the noise-over-time integral on power network is minimized

- **Correlation between Injected Currents due to**
  - **Operation variation**
    - Currents at different ports have *logic-induced* correlation
      - Large number of ports with limited control bits
    - Currents at the same port have *temporal* correlation
      - System takes several clock cycles to execute one instruction
  - **Process variation**
    - Currents have intra-die variation due to process variation
      - Mainly from Leff variation
      - Spatially Correlated
Algorithm Overview

- Apply Independent Component Analysis (ICA) to remove the correlation between current models.

- With the parameterized MNA formulation, iteratively solve the following two problems by sequential linear or quadratic programming:
  
  1. Find the optimal decap budgeting for the given max droop/bounce.
  2. Find the input corresponding to the max. droop/bounce for the given decap budgeting.

  Update the max droop/bounce.
Impact of Current Correlations

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Maximum current at all ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2</td>
<td>Stochastic model with logic-induced correlation</td>
</tr>
<tr>
<td>Model 3</td>
<td>Model 2 + temporal correlation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node #</th>
<th>Noise (V*s)</th>
<th>Runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>1284</td>
<td>6.33e-7</td>
<td>1.28e-7</td>
</tr>
<tr>
<td>10490</td>
<td>5.21e-5</td>
<td>1.09e-5</td>
</tr>
<tr>
<td>42280</td>
<td>7.92e-4</td>
<td>5.38e-4</td>
</tr>
<tr>
<td>166380</td>
<td>1.34e-2</td>
<td>5.37e-3</td>
</tr>
<tr>
<td>avg</td>
<td>1</td>
<td>37.3%</td>
</tr>
</tbody>
</table>

- Compared with the model assuming maximum currents at all ports, under the same decap area,
  - Considering spatial correlation reduces noise by 3X
  - Considering both spatial and temporal correlations reduces noise by 9X
## Impact of Leff Variation

<table>
<thead>
<tr>
<th>Node number</th>
<th>Variation</th>
<th>sLP mean (V*s)</th>
<th>sLP std (V*s)</th>
<th>runtime (s)</th>
<th>sLP + Leff mean (V*s)</th>
<th>sLP + Leff std (V*s)</th>
<th>runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1284</td>
<td>10%</td>
<td>9.28e-7</td>
<td>3.97e-7</td>
<td>184.2</td>
<td>6.14e-7</td>
<td>1.38e-7</td>
<td>332.8</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>9.43e-7</td>
<td>4.55e-7</td>
<td></td>
<td>6.38e-7</td>
<td>1.86e-7</td>
<td></td>
</tr>
<tr>
<td>10490</td>
<td>10%</td>
<td>1.03e-4</td>
<td>4.79e-5</td>
<td>1121</td>
<td>7.22e-5</td>
<td>1.23e-5</td>
<td>3429</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>1.22e-4</td>
<td>4.38e-5</td>
<td></td>
<td>7.94e-5</td>
<td>2.06e-5</td>
<td></td>
</tr>
<tr>
<td>42280</td>
<td>10%</td>
<td>2.29e-3</td>
<td>9.72e-4</td>
<td>2236</td>
<td>8.23e-4</td>
<td>1.01e-4</td>
<td>6924</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>4.43e-3</td>
<td>1.01e-3</td>
<td></td>
<td>8.28e-4</td>
<td>1.92e-4</td>
<td></td>
</tr>
<tr>
<td>166380</td>
<td>10%</td>
<td>2.06e-2</td>
<td>9.91e-3</td>
<td>3824</td>
<td>5.31e-3</td>
<td>8.92e-4</td>
<td>11224</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>2.31e-2</td>
<td>1.03e-2</td>
<td></td>
<td>5.92e-3</td>
<td>9.33e-4</td>
<td></td>
</tr>
<tr>
<td>avg</td>
<td>10%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>49.5%</td>
<td>19.8%</td>
<td>2.73X</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>51.2%</td>
<td>24.7%</td>
<td></td>
</tr>
</tbody>
</table>

- Considering Leff variation reduces mean noise by 74% and 3-sigma noise by 92%
- Spatial correlation function is $\rho(v) = e^{-3|v|}$
Conclusions

- Correlation exists for PVT variations

- Correlation can be used to reduce over-design, obtain more robust designs, and reduce design time

- Our ongoing work studies runtime tuning for operation variations
  - As operation variation is the largest uncertainty