Pre and post-silicon techniques to deal with large-scale process variations

Jaeyong Chung, Ph.D.

Department of Electronic Engineering
Incheon National University
Outline

- Introduction to Variability
- Pre-silicon Techniques
  - Basics of traditional static timing
  - OCV
  - AOCV/LOCV
  - SSTA
  - POCV/SOCV
- Post-silicon Techniques
  - Compressed Sensing
  - Compressed Silicon Sensing (CSS)
  - Virtual Probe
  - Our Proposed Framework
  - Application of CSS
Timing Uncertainty

- Add Timing Margins For Delay Uncertainty
  - Process Variation
  - Voltage Variation
  - Temperature Variation
  - Aging Effects

- Associated Costs
  - Area, Power, Design Efforts/Time
Classification of variability

Variability

- Chip-to-chip (including, wafer-to-wafer, lot-to-lot, fab.-to-fab.)
- Within chip
  - Systematic
  - Random
Sources of variation

- FEOL (Front end of line) variation
- BEOL (Back end of line) variation

Random Dopant Fluctuation

Lithography-induced variation /Proximity effects

Erosion and dishing in CMP process
CD (Lgate) Variation

- Critical dimension (a.k.a, Lgate, Leff, …)
  - The effective channel length of transistors
  - Affects delay and leakage substantially
  - Varies across-wafer and within-chip systemically
- A reduction of 1nm of the standard deviation of CD → $7.5/chip for a high end product

CD (Lgate) Variation

- Across-wafer CD variation
  - Post Exposure Bake (PEB) is the greatest variation culprit
  - In areas where the bake plate is relatively cool, CD is larger than average

http://bcam.berkeley.edu/ARCHIVE/theses/Friedberg_PhD.pdf
CD (Lgate) Variation

- Within-chip CD variation
  - Lens aberration induces spatially correlated variation
  - Different layout leads to different spatial patterns due to optical proximity effect

Voltage, temperature, weather, ...
Spatial Correlation

Across-chip variation

- Independently random part
- Spatially correlated part: within-chip distance-related correlation
- Globally correlated part: chip-to-chip, wafer-to-wafer, batch-to-batch variation

Correlation Coefficient
Traditional Static Timing

- Setup check
Traditional Static Timing

- Hold check
Traditional Static Timing

- Use worst/best corners for setup/hold checks

Launch Path $D_L$ (Launch clock path + data path)

Setup Check $D_L < D_c + T_{\text{clock}}$

Capture Path $D_C$

Process Space

CLOCK
Traditional Static Timing

- GBA (Graph-Based Analysis) takes linear time in circuit size
- PBA (Path-Based Analysis) takes exponential time in circuit size

Launch Path $D_L$

Setup Check

Capture Path $D_C$

$D_L < D_c + T_{clock}$
Traditional Static Timing

- GBA finds an upper bound of the worst path delay in linear time through the graph
- GBA is pessimistic than PBA

\[ \text{Slew}_{ac} = F(\text{Slew}_a, \text{Load}_c) \]
\[ D_{ac} = F(\text{Slew}_a, \text{Load}_c) \]

\[ \text{AT}_c = \max(\text{AT}_a + D_{ac}, \text{AT}_b + D_{bc}) \]
\[ \text{Slew}_c = \max(\text{Slew}_{ac}, \text{Slew}_{bc}) \]

\[ \text{Slew}_{bc} = F(\text{Slew}_b, \text{Load}_c) \]
\[ D_{bc} = F(\text{Slew}_b, \text{Load}_c) \]

The upper bound of the worst path delay
On-Chip Variation (OCV)

- Accounts for within-chip variation
- Global derating (e.g., ±5%) → early/late split

\[ D_L < D_c + T_{\text{clock}} \]
On-Chip Variation (OCV)

- Common Path Reconvergence Pessimism Removal (CRPR) (a.k.a, CPPR)

\[ D_L < D_c + T_{\text{clock}} \]

Launch Path: \( D_L \)

Setup Check: \( D_L < D_c + T_{\text{clock}} \)

Capture Path: \( D_C \)
Statistical Cancellation

- **If each gate has a delay of 50 with a standard deviation of 2**
  - Method 1: Set each gate to its $3\sigma$ limit, total delay = $4 \times (50+6) = 224 \ (\text{OCV})$
  - Method 2: Compute the $3\sigma$ value of the sum of 4 random variables
    \[ = 4 \times 50 + 3 \times (2^2 + 2^2 + 2^2 + 2^2)^{1/2} = 212 \]

- **The difference between these two is called RSS credit**
  - RSS = Root of the Sum of the Squares (reflects statistical cancellation)
  - In this case, the credit is 12
Statistical Cancellation

- a.k.a., RSS Credit (Root of the Sum of the Squares)

\[ E[X + Y] = E[X] + X[Y] \]

\[ Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) \]

\[ E[X + Y] = 2\mu \]

\[ Var(X + Y) = 2\sigma^2 + 2\sigma^2\rho = \begin{cases} 4\sigma^2 & (\rho = 1) \\ 2\sigma^2 & (\rho = 0) \end{cases} \]

\[ \frac{3\sigma}{\mu} \]

\[ \rho = 1 \quad 3 \times \frac{2\sigma}{2\mu} = \frac{3\sigma}{\mu} \]

\[ \rho = 0 \quad 3 \times \frac{\sqrt{2}\sigma}{2\mu} = \frac{3\sigma}{\sqrt{2}\mu} \]

For \( N \) indep. variables, variation is reduced by a factor of \( 1/\sqrt{N} \).
Advanced OCV (AOCV) (aka LOCV)

Stage-based AOCV

[Synopsys Whitepaper]
Advanced OCV (AOCV) (aka LOCV)

[Synopsys Whitepaper]
**Advanced OCV (AOCV) (aka LOCV)**

- Cells get different derate depending on the logic depth and the cell type
- (Design-specific OCV, CLK DA) it also depends on the loads and slews

**Launch Path** \( D_L \)

\[ D_L < D_c + T_{\text{clock}} \]

**Setup Check**

**Capture Path** \( D_C \)
RSS credit in setup/hold check

Setup Check

\[ D_L < D_c + T_{\text{clock}} \]

\[ \Leftrightarrow D_L - D_c < T_{\text{clock}} \]

Method 1 (OCV)

If perfectly correlated, variation will be canceled

\[ \text{Var}(X - Y) = 2\sigma^2 - 2\sigma^2\rho \]

\[
\begin{align*}
4\sigma^2 & \quad (\rho = -1) \quad \text{Method 1 (OCV)} \\
2\sigma^2 & \quad (\rho = 0) \\
0 & \quad (\rho = 1)
\end{align*}
\]
Advanced OCV (AOCV) (aka LOCV)

Distance-based AOCV

\[ D_L < D_c + T_{\text{clock}} \]
Advanced OCV (AOCV) (aka LOCV)

[0.95, 1.05]

5% → 3%

Correlated Delay

Prob. 0.05

+3σ

Perfectly Correlated Delay

Prob. 0.02

+3σ

Random Delay

Prob. 0.03

+3σ

=
Advanced OCV (AOCV) (aka LOCV)

Derating Table for each cell type

[Synopsys Whitepaper]
Advanced OCV (AOCV) (aka LOCV)

- AOCV requires a lot of library characterization efforts
  - Derating values for each cell type, each depth, each location, each slew, each load
  - Worst-case derating is selected across each load and each slew
    - A source of pessimism
- AOCV tables doesn’t have much information
  - Can be predicted by the simple analytic model
    \[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \]
- Path credit is mapped into the credit of segment delays
  - Paths do not consist of a single type of gates
  - Not graph-based analysis (GBA) friendly
Statistical Static Timing Analysis (SSTA)

- Deterministic

- Statistical

- Correlations
  - global
  - spatial
  - none!

\[
E[X + Y] = E[X] + X[Y]
\]
\[
Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)
\]
Statistical Static Timing Analysis (SSTA)

- All timing quantities computed and propagated in a parameterized form
  - ATs, RATs, slacks, slews, delays, etc
- Need to characterize sensitivities for each cell type, each delay, each slew

Canonical Form

\[ a_0 + a_1 \Delta X_1 + a_2 \Delta X_2 + \cdots + a_n \Delta X_n + a_{n+1} \Delta R_a \]

- Constant (nominal) value in the absence of variations
- Sensitivities
- Global random variables; these are probability distributions
- Independently random uncertainty
Statistical Static Timing Analysis (SSTA)

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Statistical Static Timing Analysis (SSTA)

Choice of coefficients

Use 7 variables in the Canonical Form

\[ \rho \]

\[ \begin{array}{c}
1 \\
0 \\
\end{array} \]

\[ \begin{array}{cc}
s & 2s & 3s & 5s \\
\end{array} \]

Distance

\[ \begin{array}{cc}
c = \text{cell}, n = \text{neighbor} \\
\end{array} \]
Statistical Static Timing Analysis (SSTA)

- SSTA benefits
  - Chip-to-chip variation
    - No corners
    - Safe
    - RSS credit
  - Within-chip variation
    - RSS credit down a path
    - RSS credit in setup/hold check

\[
E[X + Y] = E[X] + E[Y]
\]
\[
Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)
\]
Statistical Static Timing Analysis (SSTA)

Derating Factors of ±0% (Chip 2)

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Statistical Static Timing Analysis (SSTA)

Derating Factors of ±5% (Chip 2)

Setup Tests

Hold Tests

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Statistical Static Timing Analysis (SSTA)

Derating Factors of ±13% (Chip 2)

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Statistical Static Timing Analysis (SSTA)

Derating Factors of ±25% (Chip 2)

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Statistical Static Timing Analysis (SSTA)

- Apples-to-apples comparison of statistical flow to:
  - 2 corner foundry-like timing with derating
  - ‘n’ corner industry-standard flow
  - Exhaustive corner timing

- 380ps total
  - 200ps from RSS credit in chip-to-chip variation
  - 80ps from RSS credit in on-chip variation

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM
Parametric OCV (POCV) (aka SOCV)

- Use SSTA for within-chip variation only
- Eliminate a lot of characterization burden from SSTA, giving up the benefits in chip-to-chip variation
- Use a few variables only in the canonical form
  \[ delay = d_0 + \Delta d + \frac{\partial d}{\partial r} \Delta r + \frac{\partial d}{\partial c} \Delta c + \frac{\partial d}{\partial c_L} \Delta c_L \]
- Statistical OCV (SOCV) is a similar technique
- In theory, POCV/SOCV is clearly a better engineering than AOCV
  - Better accuracy and less characterization effort

“A parametric approach for handling local variation effects in timing analysis”, DAC 2009, Mutlu. A (Extreme DA)
Remaining Pessimism in SSTA/POCV

- Refactoring - CRPR for Combinational Networks

\[
D_L = D_{cd} + \max(a + b, a + c) + d + e \\
= D_{cd} + a + \max(b, c) + d + e
\]

Launch Path

Capture Path

Using Distributivity Of + over \( \max \)

\[
D_L < D_c + T_{\text{clock}}
\]

[Chung and Abraham, ICCAD 2009] (Best Paper Award Nomination)
[Chung and Abraham, TCAD 2012]
Compressed Sensing

- Well-known that natural signals are *compressible*
- Traditional DSP Systems

Acquisition (Sampling) → RGB2YCbCr Color Converter → 8X8 Block 2D Discrete Cosine Transform → Quantization → Huffman Encoding

Sparse

\[ f_{\text{max}} \]
Compressed Sensing

- Tremendous impact on signal processing, machine learning, statistics,..
- The original groundbreaking paper [Donoho 2004] has been cited 8769 times (200+ papers in the last 3 years.)

**Linear measurements**

\[
\begin{align*}
    y_1 &= \langle \text{Image 1}, \text{Image 2} \rangle \\
    y_2 &= \langle \text{Image 3}, \text{Image 4} \rangle \\
    y_3 &= \langle \text{Image 5}, \text{Image 6} \rangle \\
    \vdots \\
    y_k &= \langle \text{Image k}, \text{Image k+1} \rangle 
\end{align*}
\]

**Non-uniform sampling**

**Decoding or Recovery**

\[
\begin{bmatrix}
    Y_1 \\
    \vdots \\
    Y_k
\end{bmatrix} = \left[ \begin{array}{ccc}
    T_{1,1} & \cdots & T_{1,n} \\
    T_{2,1} & \cdots & T_{2,n} \\
    \vdots & \cdots & \vdots \\
    T_{n,1} & \cdots & T_{n,n}
\end{array} \right]^{-1} \begin{bmatrix}
    \theta_1 \\
    0 \\
    \vdots \\
    0 \\
    \theta_m \\
    0
\end{bmatrix}
\]

- **Classical answer:**
  - Underdetermined \( \rightarrow \) cannot solve
  - We have \( k \) equations and \( 2m \) unknowns,
    - If \( k>2m \), we may have a unique solution
- **New answer:** Information on \( 2m \) unknowns are encoded into \( k \) measurements, and we can recover it *perfectly and efficiently*
  (In practice, around \( 4m \) are needed)
Compressed Sensing

- Images and sounds have continuation
- Samples adjacent in time or space are highly correlated (high energy at low frequencies)
  - Conventional measurements are not efficient
  - CS recovers/predicts unobserved quantities from a few observations

After acquisition at t0

Lower entropy, less information
Compressed Silicon Sensing

- In IC manufacturing, measurements are expensive
  - IC cost = die cost + test cost + package cost
- Could be applicable to pre-silicon as well (where some simulations are expensive or interpolation is used)
Virtual Probe

- Framework for wafer characterization
- Many wafer test results are spatially correlated across wafer

Spatially correlated data (282 measurements)  Random 50 measurements  Predicted from 50 samples

1.8% Error
Our CSS Framework

- Test-items are also correlated strongly
  - VP recover results of each test-item independently
  - Our approach does it simultaneously

<table>
<thead>
<tr>
<th>Synthetic wafer</th>
<th>Normalized Flushed delay</th>
<th>Normalized Log(IDDQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP’s Prediction</td>
<td>(12% Error)</td>
<td>Our Prediction</td>
</tr>
<tr>
<td>50 samples/item</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our CSS Framework

- Can decompose it into correlated variation and random variation

**Synthetic wafer**

![Synthetic wafer images]

50 samples/item
Applications of CSS

- What can we do if we have a very good predictor?
- At wafer characterization step, sampled measurements are common
Conclusions

- Robustness is the key to success in nanometer technologies
  - Margins are the easiest way to obtain robustness
  - Margins eat up competitiveness
  - Needs sophisticated engineering for margining (OCV, AOCV, POCV,...)
- Post-silicon engineering (silicon debug, characterization, etc) is very important under large-scale process variations
  - Compressed Silicon Sensing
    - CS is a revolutionary theory
    - Let’s take advantage of it at IC design and manufacturing!
Thank you!