Variability and Reliability Aware Design for Emerging System-on-Chips

2015. 09. 01

Yongchan Ban, Ph.D.
Principal Engineer, yc.ban@lge.com
System IC R&D, LG Electronics
Outline

- SoC and Embedded Systems
- Technology Scaling: Variability, Reliability, Energy
  - Variability-aware Design
  - Reliability-aware Design
  - Power-aware Design
System-on Chips

- **Today’s IT (Focus on Person)**

- **Tomorrow’s IT (Focus on Objects/World)**

**Research Agenda**
- Cyber-physical System
  - The Internet of Things

**Future Technology**
- Real-time Model of the World
  - Global Intelligence

**Applications**
- Public/Social Issues
  - The Internet of Things

**Research Devices (System on Chips)**

[ slide source: Prof. Jaeyong Jung ]
Technology Scaling in Every 2 Years

<table>
<thead>
<tr>
<th>Node</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 nm</td>
<td>2003</td>
</tr>
<tr>
<td>65 nm</td>
<td>2005</td>
</tr>
<tr>
<td>45 nm</td>
<td>2007</td>
</tr>
<tr>
<td>32 nm</td>
<td>2009</td>
</tr>
<tr>
<td>22 nm</td>
<td>2011</td>
</tr>
<tr>
<td>14 nm</td>
<td>2013</td>
</tr>
</tbody>
</table>

[ picture source: Intel ]
To Go Out ~ 10 Years

[picture source: Intel]
30 Years of Scaling

Contact 1978

Ten 32nm SRAM Cells 2008

Intel Atom™
- dual-core

Rice – single grain

[ source: Intel, Synopsys Forum’11 ]
Goals of Technology Scaling

- Design new devices to be:
  - Faster?
  - Smaller?
  - Lower power?
  - Add new features?

- Growth of semiconductor industry has been fueled by the “ever cheaper” transistor

- Two major issues confront further scaling
  - Energy consumption!!
  - Variability and Reliability!!
Component of Variations

- **Total Variation**
  - Systematic Variation
  - Random Variation

- **Variation Sources**
  - Due to lithography, mask alignment, etching, CMP, deposition, implant, doping profile, temperature, etc.

- **Systematic Variation**
  - Deterministic
  - Depending on layout density, orientation, and/or location

- **Random variation**
  - Due to random process variations
  - Due to fundamental randomness
Lithography Systematic Variation

[ Source: TI ]

Line-end shortening  △CD Proximity  Poly Corner  Line-edge Roughness  Active Corner
Systematic Variation-aware Design: PO/OD

- Active Corner Rounding
- Poly Corner Rounding
- Process Variation
- Line-end Shortening
Impact of Gate Length Variation

- $\Delta L_{\text{gate}}$ is up to 10% @45nm node.

- The small improvement of $\Delta L_{\text{gate}}$ reduction can lead to significant decrease of delay and leakage variations.

[Ban et al., ISPD 2010]
Total Sensitivity based Optimization

Non-Rectangular Gate-aware Cell Charac.

**Total Sensitivity = Circuit Criticality + Manufacturability**  
[Ban et al., ISPD 2010]

\[
\sigma_{\text{NMOS}} = \sigma_{\text{PMOS}} \quad \sigma_{\text{NMOS}} < \sigma_{\text{PMOS}}
\]
The S/D contact resistance is as much as 26% of the gate channel resistance at 45nm node devices.
Impact of Contact Variation

- **S/D Contact Size (CD)**
  - $\Delta S/D$ contact area $\rightarrow \Delta$contact resistance ($R_{co}$) $\rightarrow \Delta$current ($I_{ds}$).
  - 32nm standard cell (nominal CD = 40nm).
  - 10nm Contact $\Delta$CD $\rightarrow$ up to 5% $\Delta I_{ds}$ & over 100% $\Delta R_{co}$.

[Ban et al., DAC 2010]
Impact of Contact Distance on $I_{ds}$

- **S/D Contact Position**
  - The $I_{ds}$ degrades as contacts are placed closer to the gate.
  - Neighboring contact holes locally relax the strain in channel.
  - Generate Look-up table for Distance weighting factor ($w_D$)

![Diagram showing the impact of contact distance on $I_{ds}$]

[Ban et al., DAC 2010]

[Eiji Morifuji, IEEE TED'09]
Impact of Contact Shape on \( I_{ds} \)

- **S/D Contact Shape**
  - As the contact length is larger, the \( I_{ds} \) is increased.
  - Less current crowding from the S/D electric field.
  - Generate Look-up table for Shape weighting factor (\( w_S \))

[Ban et al., DAC 2010]
There are two look-up tables for a new contact model

- Distance weighting factor \((w_D)\) due to stress
- Shape weighting factor \((w_S)\) due to the current crowding
- Given \(i^{th}\) slice of a contact,
the resistance is

\[
R_i = \frac{\rho}{w_{D,i} \cdot w_{S,i} \cdot A_i}
\]

where \(\rho\) is resistivity and \(A\) is the area of a slice

- The driving current is

\[
I_{ds} = \frac{V_{ds}}{R_{tot}} = \frac{V_{ds}}{\left(R_{ch} + R_{ds}\right)} \propto \frac{1}{R_{co}} = \sum_i \frac{1}{R_i} = \frac{1}{\rho} \cdot \sum_i \left(w_{D,i} \cdot w_{S,i} \cdot A_i\right)
\]

- The lower \(\omega\), the higher \(R\) and the smaller \(I_{ds}\).

[Ban et al., DAC 2010]
Compact S/D Resistance Model

[Image of printed images, 2D simulation, and 3D simulation]

- **Contact Position**
  - [Graph showing current variation vs. gate to contact space (nm)]
  - [Graph showing current variation vs. contact length along gate (nm)]
  - [Graph showing current variation vs. contact CD (nm)]

- **Contact Shape**
  - [Graph showing current variation vs. gate to contact space (nm)]
  - [Graph showing current variation vs. contact length along gate (nm)]
  - [Graph showing current variation vs. contact CD (nm)]

- **Contact Area**
  - [Graph showing current variation vs. gate to contact space (nm)]
  - [Graph showing current variation vs. contact length along gate (nm)]
  - [Graph showing current variation vs. contact CD (nm)]

[Ban et al., DAC 2010]
S/D Contact Layout Optimization

- **Variability-driven Design**
  - The main goal is to minimize the contact CD variation between the fastest and slowest process corners.

- **Performance-driven Design**
  - The contact CD increases as the pitch decreases.
  - We can make vertically long contacts by reducing the space between contacts.

[Conventional] [Variability-driven] [Performance-driven]

[Ban et al., DAC 2010]
Line-Edge Roughness (LER)

- **What is LER (Line Edge Roughness)?**
  - Random variation of MOS gate length along the gate width.
  - LER is formed in the acid generation, the acid diffusion and development process.

- **Why LER becomes so critical**
  - LER does not scale accordingly and becomes an increasingly larger fraction.
  - Below 30 nm the LER takes over and becomes the dominant fluctuation source.

---

**Exposure**
[Namatsu, JVST’98]

**Development**

[Chandhok, SPIE’07]

**Gate Length**

[Asenov, IEEE TED’03]
Random Variation-aware Design on Std. Cell

- Line-edge Roughness and Contact-edge Roughness

![Line-edge Roughness (LER)](image)

$$S(f) = \frac{2\sigma_{LER}^2 L_c}{\left(1 + (2\pi)^2 f^2 L_c^2 \right)^{0.5+\alpha}}$$

![Power Spectral Density (PSD) LER Modeling](image)

- Distance weighting factor ($w_D$) due to stress
- Shape weighting factor ($w_S$) due to the current crowding

$$R_i = \frac{\rho}{w_{D,j} \cdot w_{S,j} \cdot A_i}$$

Contact-edge Roughness Modeling
LER-aware Poly Pitch Optimization

\[ S(f) = \frac{2\sigma(x)^2 L_c}{\left(1 + \left(\frac{2\pi f}{2 L_c^2}\right)^2\right)^{0.5+\alpha}} \]

where,

\[ \sigma(x) = \sum_{i=0}^{N} a_i x^i \quad i = 0,\ldots, N \]

or,

\[ \sigma(x) = \begin{cases} 
-\alpha_1 \cdot x + \beta_1 & \text{if } x \leq x_{p1} \\
\alpha_2 \cdot x - \beta_2 & \text{if } x \geq x_{p1} \\
\beta_3 & \text{if } x \geq x_{p2} 
\end{cases} \]

\( x > P \), where \( P \) is process violation

\( x < D \), where \( D \) is design violation
LER-aware Poly Optimization

\[
\min : \sum_{j=1}^{M} \Psi_j \{\sigma(x_L) + \sigma(x_R)\}_j \quad \forall j \in T
\]

where,

\[
\sigma(x_{L,j}) = \sum_{i=0}^{N} a_i (d_{L,j})^i \quad \forall j \in T
\]
\[
\sigma(x_{R,j}) = \sum_{i=0}^{N} a_i (d_{R,j})^i \quad \forall j \in T
\]
\[
d_{L,j} = x_j - x_{j-n}
\]
\[
d_{R,j} = x_{j+n} - x_j
\]

s.t. :

\[
x_{j_{\text{min}}} \leq x_j \leq x_{j_{\text{max}}} \quad \forall j \in T
\]
\[
x_{\text{max}} - x_{\text{min}} = W_{\text{cell}}
\]

• N: a positive integer
• \(d_j\): distances from the edge
• \(\Psi\): device criticality

[Ban et al., DAC 2011]
Cu erosion and dishing change R & C variation $\rightarrow$ timing/power
Topographic variation translates to focus variation for imaging of subsequent layers $\rightarrow$ depth-of-focus $\rightarrow$ manufacturability
CMP impacts both IC parametric and manufacturability

Cupper Interconnect Problem
[Ban et al., KCS 2015]
- CMP variation is highly dependent upon metal density and pitch (line width & line space).

- Dummy metal fill (conventional)

[figure: Prof. Boning, MIT]

[Ban et al., KCS 2015]
The dummy metal fill approach just cares for metal layout density.

- Layout density variation is still huge.
- Layout density is not an only parameter controlling CMP.
- CMP variation mismatch and impact (e.g., $10\% \uparrow \Delta T_{cu} @28\text{nm}$)

28nm M4 Layout Density

28nm M4 Cu Thickness due to CMP

[Ban et al., KCS 2015]
Systematic Variation-aware Design on IP/Chip

- LDE (Layout Dependent Effect)

  - LOD - Length of Diffusion
  - PSE - Poly Spacing Effect
  - WPE - Well Proximity Effect
  - OSE - OD to OD Spacing Effect

  [Ban et al., SPIE 2014]

22% Delay and 31% Leakage Spreading in 16nm Library
LEA variability analysis was used to:
- Quantifies the delay and leakage variability spread
- Verify the characterization context positioning
- Help improve design methodology

Up to 21ps or 22% delay spread
Up to 111pW or 31% leakage spread

[Ban et al., SPIE 2014]
Variability-aware Implementation

- Reducing Variability of Cells/IPs
  - Quantify LDEs related variability in Std. Cells and IPs
  - Pin point root causes of excess variability and optimize cells

- Variability-Aware Placement
  - Context library litho/stress simulation and analysis with LEA
  - A variability score for cell combinations (CCI)
  - Use cells with low variability score

- Timing Margin Tuning

[Cadence, CDNLive’09]
Layout Guidelines to Mitigate LDEs

- **Avoid functional cells on boundaries**
  - Functional cells should be avoided on the block periphery
  - Block periphery should be padded with the non-functional cells (e.g., fillers, decaps, endcaps, and so on).

- **Fix context of clock cells**
  - Surround all instances of clock cells with an identical context and characterize the clock cells with the context.
  - Minimizes impact of the un-modeled LDEs on the clock cells.

- **Surround double-height cells with non-functional single-height cells**
  - Double-height cells can cause high LDEs around them.
  - Placing non-functional single-height cells around double-height cells reduces LDEs on nearby functional cells.

- **Avoid very small drive strength cells on critical paths and clocks**
  - Small drive strength cells are more susceptible to LDEs.

[Freescale, SPIE 2012]
Routing: a key stage for printability optimization
- Last major design stage
- Wire embedding step

Lithography Aware Routing
- OPC Friendly Detailed Routing [Huang+ DAC’04]
- Multi-level Routing with OPC [Chen+ ASPDAC’05]
- RADAR: Litho-Aware Routing [Mitra+ DAC’05]
Lithography-friendly Detailed Routing

N-th set of Weak Pattern

OPC & Simulation

[Lin, Ban, Pan, and Li, ICCAD 2011]
[Cho, Yuan, Ban, and Pan, TCAD 2009]
[Cho, Yuan, Ban, and Pan, DAC 2008]
In-design DFM w/ Pattern Matching

Pattern Layers

Pattern Marker

Pattern Extent

[courtesy of Synopsys]
(1) Conventional Litho. Check at the signoff stage

- Final GDS → Litho. Check → Hot Spots → Routing Tool Fixing → DFM Clean

(2) In-design Pattern Matching from an early stage of routing

- Routing DB → Pattern Matching → Hot Spots → Routing Tool Fixing → Litho. Check

(3) Correct-by-construction Routing

- Pattern Lib. → Routing w/ Concur. Fix → Litho. Check

<table>
<thead>
<tr>
<th>Flow</th>
<th>Runtime (sec)</th>
<th># Hot spot remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Litho. sim</td>
<td>Pattern Match</td>
</tr>
<tr>
<td>(1) Conv. DFM sign-off</td>
<td>57,120</td>
<td>487 (24 violations)</td>
</tr>
<tr>
<td>(2) In-design DFM</td>
<td>-</td>
<td>499 (49 violations)</td>
</tr>
<tr>
<td>TAT reduction</td>
<td>-</td>
<td>0.97</td>
</tr>
</tbody>
</table>

[Shin, Ban et al., SPIE 2014]
The Evolution of Pattern Matching

Classic DRC Code

Classic Pattern Matching is a specific physical representation

New representation is Topological “Squish” Pattern

[Ban et al., SPIE 2014]
Topological “Squish” Pattern

- Very simple Pattern Abstraction
- More like DRM than DRC Rules
- Easy to maintain
- Natively supports all orientations (mirror, rotations)
- Support multi-level patterns
- High performance search engine

[M2 pattern] + [Via Pattern] + [M3 Pattern] = [Multi-level pattern]

[Ban et al., SPIE 2014]
In-design DFM Optimization

DFM Scoring and Optimization Flow

- Real case optimization of 16FF design

<table>
<thead>
<tr>
<th>16FF Design</th>
<th>Size mm²</th>
<th>Run Time (min)</th>
<th># of Fixes</th>
<th>Min Area Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>atl_cpu*</td>
<td>1.5</td>
<td>15</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Initial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>21%</td>
<td>22%</td>
</tr>
<tr>
<td>M3</td>
<td>12%</td>
<td>23%</td>
</tr>
<tr>
<td>M4</td>
<td>11%</td>
<td>36%</td>
</tr>
</tbody>
</table>

*Courtesy of LG Electronics

(Source: Cadence)
Lithography Challenges in sub 45nm

\[ R = k_1 \frac{\lambda}{NA} \]

- \( R \): lithography difficulty
- \( NA \): numerical aperture
- \( \lambda \): wavelength of source
- \( k_1 \): min. printable half pitch

[S. Borkar, 2014]
Why Double Patterning?

- **Too hard** with the current infrastructure!
  - Increase $NA$ (3rd generation fluid >1.4, larger lens)
  - Use shorter wavelengths (13.5nm-EUV)

- Have no choice but to bear with the current state, but increase $HP$ to print 32/22nm designs
  - **Double patterning**!
  - Reuse the current infrastructure

\[
R = k_1 \frac{\lambda}{NA}
\]

Ex)
\[
\begin{align*}
\lambda &= 193\text{nm} \\
NA &= 1.4 \\
K1 &= 0.26 \\
R &= 35.8 \\
\Rightarrow \text{Around pitch 70nm ↓ is a lithography limit.}
\end{align*}
\]
Double Patterning Tech (DPT)

Conventional LELE (litho-etch-litho-etch) DPT

Spacer Type SA (self-aligned) DPT

[Ban et al., DAC 2011]
- Layout decomposition for DPT is more complex than phase assignment or 2-coloring

[Cho, Ban, ICCAD'08]
LELE DPT Aware Routing

a) # of stitches
b) wirelength, via, and so on

WL=34, Stitch=1
WL=28, Via=2

[Cho, Ban, ICCAD’08]
SADP Aware Routing

- SADP Layout Decomposition

- SADP-awra Routing

[C. Kodama+, ASPDAC’13]

[Ban et al., DAC 2011]
Reliability Problems in 3D ICs

- Mainly due to CTE (coefficients of thermal expansion) mismatch between Si and Cu TSV

Better Performance

- Massive Bandwidth
- Reduced Interconnect Delays
- Power Reduction (Less IO driver)
- Higher Functionality/Space
- Heterogeneous Integration

[ Prof. S. Lim, Gatech ]
TSV-TSV & TSV-Transistor Interactions

Stresses add up

Stresses cancel out

[Source: IMEC]
Keep-Out Zone (KOZ)

- KOZ tradeoff: area vs delay vs reliability

[ Prof. S. Lim, Gatech ]
TSV Stress Impact on Mobility

Thermal Mechanical Stress

Hole Mobility Variation

Electron Mobility Variation

[ D.Pan, ASPDAC’12 ]
TSV Stress-aware Placement

Hole Mobility Variation

Electron Mobility Variation

Regular TSV

Irregular TSV

Hole Critical Cell

Electron Critical Cell

[ K.Athikulwongse, ICCAD'10 ]
Reliability-aware Design (EM)

- Signal Electro-migration (EM)
  - Mean-to-time-failure (DC Current)
    \[ \text{MTTF} = \frac{A}{J^n} e^{\left(\frac{E_a}{kT}\right)} \]
  - Mean-to-time-failure (AC)
    \[ \text{MTTF}_{\text{AC}} = \frac{A}{(J_+ - \gamma J_-)^n} e^{\left(\frac{E_a}{kT}\right)} \]
  - Mean-to-time-failure (AC+\Delta\text{Process})
    \[ \text{MTTF}_{\text{AC(\Delta\text{process})}} = \frac{(A + \Delta A)}{\left[(J_+ + \Delta J_+) - \gamma (J_- + \Delta J_-)\right]^n} e^{\left(\frac{E_a}{kT}\right)} \]
  - Joule Heating, Signal interconnects
  - Prone to process variation (lithography, etch and CMP)
  - Higher driving strength in FinFET and higher circuit frequency

[Ban et al., SPIE 2014]
[Ban et al., KTC 2014]
Signal-EM in 16nm Design

\[
MTTF_{AC(\Delta \text{process})} = \frac{(A + \Delta A)}{[J_+ + \Delta J_+ - \gamma(J_- + \Delta J_-)]^n} e^{(E_a/kT)}
\]

(a) When the impact of the geometry is higher than one of the current density, MTTF increases.

(b) While when the impact of the current density is higher than one of the geometry change, MTTF decreases.

[Ban et al., SPIE 2014]
- Front-end: Typical process corner
- Back-end: 110 °C and RCbest parasitic corner
- The higher deriving current from FinFET causes more EM violations.

[Ban et al., KTC 2014]
Signal-EM with Ambient Temperature

- Front-end: FF (Fast-Fast) process corner
- Back-end: 110 °C and RCbest parasitic corner
- The more FinFET current is induced as temperature increases due to temperature inversion.

[Ban et al., KTC 2014]
Signal-EM with Metal Temperature

- Front-end: Typical process corner
- Back-end: 110 °C and RCbest parasitic corner
- The signal-EM is so sensitive to the metal temperature.
- 110 °C is recommended for sub-20nm node design.

[Ban et al., KTC 2014]
Signal-EM with Metal RC Corners

- Front-end: FF (Fast-Fast) process corner
- RCBest corner gives the worst EM violation.
- RC reduction on signal nets might induce more current.
- The DPT corners minor changes EM violations.

[Ban et al., KTC 2014]
Reliability-aware Design

16nm Cortex-A57 CPU

EM on Via

EM on Clock Net

Redundant Via

Metal Dummy Fill

Non-Default Rule

EM Fix Approaches

[Ban et al., SPIE 2014]
[Ban et al., KTC 2014]
The total power dissipated in a device consists of two components.

\[ P_{total} = P_{static} + P_{dynamic} \]

Static or leakage power is state, temperature, process and voltage dependent.

\[ P_{static} = Vdd \times I_{leak} \]

Dynamic Power is comprised of internal power and switching power.

\[ P_{dynamic} = P_{internal} + P_{switching} \]

\[ P_{dynamic} = C_{eff} \times V^2 \times f_{clk} \]
Dynamic Voltage Drop

- Overall Voltage drop is

\[ V_{\text{drop}} = IR + L \frac{di}{dt} \]

- In addition to IR drop, inductance for power ground network also affects the voltage drop.

[Voltage drop @ transistor level] [Hot-spot Area]

[Ban et al., ISOCC 2014]
Dynamic IR Drop w/ Process Corners

![Image of thermal maps]

<table>
<thead>
<tr>
<th>RC corners</th>
<th># DVD Max.</th>
<th># DVD Min.</th>
<th>Worst Instance</th>
<th>Worst Wire</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCBest_DPT</td>
<td>531</td>
<td>426</td>
<td>27.8%</td>
<td>27.0%</td>
</tr>
<tr>
<td>RCBest_dPV</td>
<td>712</td>
<td>433</td>
<td>29.1%</td>
<td>27.6%</td>
</tr>
<tr>
<td>RCBest</td>
<td>561</td>
<td>433</td>
<td>28.4%</td>
<td>27.6%</td>
</tr>
<tr>
<td>Typical</td>
<td>1537</td>
<td>513</td>
<td>32.9%</td>
<td>31.6%</td>
</tr>
<tr>
<td>RCWorst</td>
<td>4722</td>
<td>1199</td>
<td>40.5%</td>
<td>38.7%</td>
</tr>
<tr>
<td>RCWorst_dPV</td>
<td>2365</td>
<td>400</td>
<td>33.5%</td>
<td>30.5%</td>
</tr>
<tr>
<td>RCWorst_DPT</td>
<td>4543</td>
<td>1086</td>
<td>40.3%</td>
<td>38.5%</td>
</tr>
</tbody>
</table>

*Dvd with Metal RC-corner*

[Ban et al., ISOCC 2014]
Future Research Directions

- **Reliable SoC/Embedded System**
  - Manufacturability
  - Reliability
  - Testability
  - Energy Efficiency

- **Circuit/System-level Techniques for New Processes/Devices**
  - Pi-gate, Wire, CNTs
  - EUV, DSA, Nanowires
  - Memristor-based Systems

- **3D SoC**
  - TSV (Through Silicon Via)
  - Heterogeneous Integration

- **HW/SW Co-design**
  - HW Acceleration IPs
  - System Software
Thank You!