Designing for Uncertainty: Addressing Process Variations and Aging Issues in VLSI Designs

Puneet Gupta
EE Dept., UCLA
http://nanocad.ee.ucla.edu
puneet@ee.ucla.edu
Outline

- Introduction
- A Quick Intro to Manufacturing Process
- Dealing with Patterning Variations
- Handling Random Variations
  - LER
  - Bowl-shape
  - StatOpt
  - Corners ? (highTemp?
  - Confidence
- Dealing with Hardware Variability in Higher Layers of Abstraction
The Making of an IC

- Expensive hardware and software tools involved at each step
  - $\sim$20M: cost of designing a modern IC
  - $\sim$3B: cost of setting up a modern IC fabrication plant
- Opportunity: Add value at the design-manufacturing interface
Basics of IC Design

- **Design Specification**: Behavior description
- **Logic Synthesis**: Circuit Netlist
- **Physical Design**: Layout
- **Timing Analysis**: Tapeout

**Careabouts:**
- Speed
- Power
- Area
- Correctness
- Schedule
Basics of IC Manufacturing

- Image the design layout onto the silicon wafer (photolithography)
  - Similar to photography but with wave optics
  - Goal: get accurate critical dimension (CD)
- Process the transferred image to create the needed layers (implant, deposit or grow material)
- 600+ processing steps, 30+ layers
  - Every step is an increase in uncertainty and cost

Care-abouts:

1. Shape reproduction:
   - 1. X-Y: linewidth (=CD)
   - 2. Z: thickness/height
2. Electrical model-to-silicon correlation:
3. Cost and Yield
4. Keep the expensive manufacturing equipment busy!
Printing in Subwavelength Regime

Layout

0.25µ

0.18µ

0.13µ

90-nm

65-nm

Figures courtesy Synopsys Inc.
Problem: Uncontrollable Variation

- Chips don’t work as designed
- Implications of loss in predictability
  - Guardbands ⇒ Overdesign
  - Worsened \{time to market, cost, power\}

Across-wafer frequency variation ⇒ What should be the performance spec of this chip?

Figure courtesy IBM
Variability: The Era of Diminishing Returns in Computing

Performance

Technology Generation

130nm  90nm  65nm  45nm  32nm  22nm  post-silicon

nominal scaling

overdesigned scaling

Design-Assisted Scaling

NanoCAD Lab
Problem: Yield and Cost/Risk

- Chips are thrown away
  - Yield = Chips that work at the end vs. that could have worked

- Consider (again) a cellphone chip selling (say) 100M copies
  - Design house pays (say) $5K/300mm wafer in 90nm technology
  - 10mmX10mm chip size at 90nm \( \Rightarrow \) \( \sim \) 700 die/wafer
  - 90% vs. 95% yield
    - 630 vs. 665 good die per wafer
    - 158730 vs. 150370 wafers needed to meet the demand
    - $42M saving!

- High yield key to faster time to volume

- What matters is good die/wafer however you get it
A Quick Intro to Manufacturing Process
Basic Fabrication: Two Steps

• (1) Transfer an image of the design to the wafer
• (2) Using that image as a guide, create the desired layer on silicon
  – diffusion (add impurities to the silicon)
  – oxide (create an insulating layer)
  – metal (create a wire layer)
• Use the same basic mechanism, photolithography, to do (1)
• Use three different methods to do (2)
  – Ion implant - used for diffusion: Shoot impurities at the silicon
  – Deposition - used for oxide/metal: Usually chemical vapor (CVD)
  – Grow - used for some oxides: Place silicon in oxidizing ambient
Introduction to Device Fabrication

Oxidation

Lithography & Etching

Ion Implantation

Annealing & Diffusion

Deposition

Courtesy EE143/Costas Spanos - UCB
The lithographic process

- Complex stuff! Objective lens weights 1000+ pounds

Design => Mask => Wafer

CAD System
- Layout
- Simulation
- Design Rule Checking

Electron Gun
Focus
Deflection

Mask Making
Mask

Light Source
Condenser Lens
Mask
Reduction Lens
Wafer

Wafer Exposure

Courtesy EE143/Costas Spanos - UCB
Lithographic Process

1. Starting wafer with layer to be patterned
2. Coat with photoresist
3. Bake the resist to set its dissolution properties
4. Expose resist by shining light through a photomask
5. Immerse exposed wafer in developer
6. Etch the film

- Dehydration bake
- Adhesion promoter application
- Resist application
- Softbake
- Exposure
- Post exposure bake*
- Develop cycle
- Hardbake
- Resist stabilization*

*Optional steps

Courtesy EE143/Costas Spanos - UCB
Modern Projection Equipment

• Scanner = Illumination equipment
  - Wafer is “stepped” under the mask to print, one field at a time
    - Scan the field by exposing a small “slit” at a time
• Field = area of the mask which is exposed at one time
  - Small field → more exposures → more time to print a wafer BUT less tight lens manufacturing control is needed
  - Maximum field size ~ 26mm x 33mm on wafer scale
    - Dictates biggest chip you can make
  - Reticles are 6inch X 6inch
Projection Printing

What can vary?

- Optical System: illumination and lens
- Resist: exposure, post-exposure bake and dissolution
- Mask: transmission and diffraction
- Wafer Topography: scattering
- Alignment

Courtesy EE143/Costas Spanos - UCB
Resists for Lithography

- Resists
  - Positive
  - Negative

- Exposure Sources
  - Light
  - Electron beams
  - Xrays

Courtesy EE143/Costas Spanos - UCB
Mask (Reticle) Manufacturing

MEBES format and machine, or others
• Place a glass plate covered with chrome covered with resist in a high-vacuum column
• Use an electron beam spot size smaller than the finest resolution of the design
• Mask write
  – Raster Scan: Scan the surface of the mask with the e-beam in a raster-scan order. Modulate the beam to transfer the pattern to the chrome
  – VSB: Shaped Beam: Shot where needed
• Develop the resist, and the chrome, and then remove the resist
• Check and correct the chrome pattern
• Masks shapes are 4X bigger than wafer shape

Courtesy K. Yang, UCLA
Mask Types

- **Bright field masks**
  - opaque features defined by chrome
  - background is transparent
  - used, e.g., for poly and metal

- **Dark field masks**
  - transparent features defined
  - background is opaque (chrome)
  - used, e.g., for contacts
Lithography Primer: Basics

• The famous Raleigh Equation:

\[ \text{Resolution} = k_1 \frac{\lambda}{NA} \]

\( \lambda \): Wavelength of the exposure system

NA: Numerical Aperture (sine of the capture angle of the lens, and is a measure of the size of the lens system)

\( k_1 \): process dependent adjustment factor

• Exposure = the amount of light or other radiant energy received per unit area of sensitized material.

• Depth of Focus (DOF) = a deviation from a defined reference plane wherein the required resolution for photolithography is still achievable. (affects 3D resist)

• Process Window = Exposure Latitude vs. DOF plot for given CD tolerance
Diffraction at the Mask

- Interaction between the illumination and the mask generates a diffraction pattern.
- For periodic structures, the diffraction pattern is discrete plane waves.
- For isolated structures, the diffraction pattern is a continuous spectrum of waves.

Courtesy Andrew Neureuther - UCB
The propagating light from the diffraction pattern can be described by a set of plane waves.

These waves are commonly drawn as arrows (see left).

For periodic structures with discrete diffraction patterns, the waves are called “orders.”

- They are numbered “-1”, “0”, “+1”
- The wave that goes straight through the mask is called “0”.
- Need at least two orders to pass through the lens to form the image.

Bragg’s Law:

\[ \sin \theta = \frac{m \lambda}{\text{Pitch}} \]

\( \theta \) = angle between \( m^{th} \) order and zero order

Courtesy Andrew Neureuther - UCB
Bigger Lenses: Nothing comes for free!

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

\[ \text{DOF} = k_2 \frac{\lambda}{NA^2} \]

Tradeoff in Projection Litho

Courtesy Lars Liebmann, IBM
Etch Metrics

Degree of Anisotropy

\[ A_f = 1 - \frac{r_{lat}}{r_{ver}} \quad 0 \leq A_f \leq 1 \]

Complete Anisotropic Etching

\[ A_f = 1; \quad B = 0 \]

Etch Selectivity:

\[ S_{AB} = \frac{r_A}{r_B} \]

(a) Film thickness variation across wafer

(b) Film etching rate variation

Etch Uniformity governed by:

Bias \quad B = d_f - d_m

Courtesy EE143/Costas Spanos - UCB
Etch Variability Issues

• Loading effects: non-uniform etching due to difference in supply and demand of reactive radicals
  – Macroloading: global density $\Rightarrow$ determines total area to be etched, i.e., overall etch rate
  – Microloading: local density dependent
  – RIE Lag: aspect ratio dependence $\Rightarrow$ wide vs. narrow trenches $\Rightarrow$ narrow openings have smaller etch depths

• No good scalable models for etch exist yet 😞 (unlike litho)
  – Most OPC flows deal with it in simple (insufficient) rule-based biases
Image Quality Metric: Contrast

Contrast:
\[ C = \frac{I_{\text{MAX}} - I_{\text{MIN}}}{I_{\text{MAX}} + I_{\text{MIN}}} \]

The contrast is always between 0 (no variation) and 1 (perfect minimum).

- Ensures clear distinction between “light” and “dark” → resist can distinguish pattern easily

Courtesy EE143/Costas Spanos - UCB
Image Quality Metric: Image Slope

- Image slope = $dI/dx$
- Image log slope: Normalize w.r.t. dose: $dI/Idx = d(ln I)/dx$
- Normalized image log slope (NILS): normalize w.r.t. linewidth: $w*d(ln I)/dx$

- Small NILS $\Rightarrow$ Small % change in intensity causes large % change in linewidth $\Rightarrow$ higher susceptibility to variations $\Rightarrow$ BAD

* simulated aerial image of an isolated line

Courtesy EE143/Costas Spanos - UCB
Sidewall Angle

- You never get perfect cuboidal features (all depths of resist don’t get same energy)
- CD usually measured at center
Line Edge (width) Roughness

- Caused by variety of random phenomena in litho/resist/etch such as photon shot noise
- Important for small linewidths (<45nm node)
One mask on top of another: What can go wrong: Overlay Errors

“Ideal”

Fox

S/D implant

n+

Electrical short

Solution: Make poly gate longer to overlap the FOX

poly-gate

“With alignment error”

Courtesy EE143/Costas Spanos - UCB
Thin Film Deposition

Applications:
- Metalization (e.g. Al, TiN, W, silicide)
- Poly-Si
- dielectric layers; etc

Physical Methods
- Evaporation
- Sputtering
- Reactive Sputtering

Chemical Methods
- Chemical Vapor Deposition
- Low Pressure CVD
- Plasma Enhanced CVD

Courtesy EE143/Costas Spanos - UCB
Sputtering

Negative Bias (kV)

Al target

Ar plasma

Deposited Al film

Gas Pressure ≈ 1-10 m Torr

Deposition rate = \text{constant} \cdot I \cdot S \quad \text{sputtering yield}

ion current

Courtesy EE143/Costas Spanos - UCB
Step Coverage Problem with PVD

- Both evaporation and sputtering have directional fluxes.

“geometrical shadowing”

Courtesy EE143/Costas Spanos - UCB
Chemical Vapor Deposition (CVD)

More conformal deposition vs. PVD

Shown here is 100% conformal deposition.

Courtesy EE143/Costas Spanos - UCB
Ion Implantation

Reminder: During implantation, temperature is ambient. However, post-implant annealing step (>900°C) is required to anneal out defects.

Preferred Method of Adding Impurities to Wafers

Courtesy EE143/Costas Spanos - UCB
Post-Implantation Annealing

After implantation, we need an annealing step to

(1) Restore Si crystallinity.

(2) Place dopants into Si substitutional sites for electrical activation

Differences in reflectivity/absorption of polysilicon vs. STI vs. active can cause different dopant activation leading to Vth variation → several new poly fill rules originate from here.

- Rapid Heating
- 950-1050° C
- >50° C/sec
Metals: Damascene and Dual-Damascene Process

- Cu is difficult to etch → it is electroplated and then planarized
- CMP: Chemical Mechanical Polishing
  - Wafer is polished using a chemical/-mechanical slurry

Courtesy Andrew Kahng, UCSD
Area Fill & Metal Slot for Copper CMP

- **Dishing** causes thinning → higher-resistance wires or lower-reliability bond pads
- **Erosion** causes a sub-planar dip on the wafer surface, causing short-circuits between adjacent wires on next layer
- **Dishing** and **erosion** arise from different polish rates for copper and oxide → Density Issue

Implications for design: Can’t have really isolated or really wide wires; unforeseen coupling through dummy fill

**Oxide erosion** and **copper dishing** can be controlled by **area filling** and **metal slotting**
Why do we need planarity?

Benefits for Lithography Processes:
• Lower Depth-of-Focus requirement
• Reduced optical reflection effects on resist profiles
• Reduced resist thickness variation over steps

Benefit for Etching Processes:
• Reduced over-etch time required due to steps

Benefit for Deposition Processes:
• Improved step coverage for subsequent layer deposition

Courtesy EE143/Costas Spanos - UCB
Finally, Strain Engineering

THOMPSON et al.: 90–NM LOGIC TECHNOLOGY FEATURING STRAINED SILICON

Type of Stress Needed for Enhanced Mobility

<table>
<thead>
<tr>
<th>Direction</th>
<th>NMOS</th>
<th>PMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal</td>
<td>Tension +++</td>
<td>Compression +++++</td>
</tr>
<tr>
<td>Transverse</td>
<td>Tension ++</td>
<td>Tension +++</td>
</tr>
<tr>
<td>Out-of-plane</td>
<td>Compression ++++</td>
<td>Tension +</td>
</tr>
</tbody>
</table>

Fig. 6. From piezoresistance, the effect of various stress on electron and hole mobility (adapted from [28]).
The strain nitride film effect is proportional to Vol & proximity of the film to the channel

Strain depends on
- Poly pitch
- Contact Size and Contact-to-Poly space
- Contact pitch
- Length of diffusion
- STI width (diffusion gap)
Remote Versus Nearby Contacts

Low channel stress
Contacts are 60nm from the gate

High channel stress
Contacts are 90nm from the gate

Contacts are 180nm from the gate

Stress Simulation courtesy of V. Moroz (Synopsys)

Courtesy Ban P. Wong, Chartered Semiconductors
The Story does not end Here: New Interactions: WPE

- **WPE = Well Proximity Effect**
- Unintentional dopants $\rightarrow$ $V_t$ depends on proximity to well edge
- Modeled in BSIM4+ by SCA, SCB parameters
  - Connected wells in digital designs a savior for modeling
Effect of Layout Parameters on Mobility

No existing model for drive current variation as a result of different Layout constructs

Contact to gate
Contact pitch
Well to diffusion
Diffusion to gate
Poly spacing
P-channel
N-channel

Source: Synopsys
Courtesy Ban P. Wong, Chartered Semiconductors
Dealing with Patterning Variations
Lithographic WYSIWYG Breakdown

- Existing compact device models (e.g., BSIM) do not handle non-rectangular geometries.
- Where Are Electrical Models of Patterning Imperfections Needed?
  - Cell characterization
  - Electrically-driven OPC
  - Design rule optimization
Is Interconnect Modeling Important?

- Probably not..
  - Litho impacts wire width $\uparrow \downarrow$ ($w$)
    - $w \uparrow \rightarrow R_{\text{wire}} \downarrow, C_g \uparrow, C_c \uparrow$
  - Wire_Delay $\sim 0.5*R_{\text{wire}}*(C_g + C_c + C_L) \sim K_1 + \frac{K_2}{w(P-w)} + \frac{K_3}{w}$
  - Gate_Delay $\sim R_{\text{gate}}*(C_g + C_c + C_L) \sim \frac{K_4}{P-w} + K_5 w + K_6$
  - Wires are long $\Rightarrow$ averaging effects
  - Semi-global and global wiring (M3+) is wide and regular $\Rightarrow$ patterning less of an issue
  - M1/M2 impact on power/performance is small
  - Caveat: contacts (and via) R variation may be non-negligible
Why Wires Are Not Important

- Width variation averages over long wires.
- Resistance and capacitance change in opposite directions as line width changes.

![Graph showing wire width vs RC and gate length vs inverter delay (fanout 1).]

- Approx. 10% change in RC with wire width variation.
- Approx. 50% change in gate length vs inverter delay (fanout 1).

FreePDK 45nm process
Simulation at Chip-Level

- Delay and switching power <3%.
- Impact of wire variation is exaggerated as averaging effect is ignored. → Let us concentrate on devices

<table>
<thead>
<tr>
<th>Interconnect layers (variation)</th>
<th>Δ delay (%)</th>
<th>Δ Switching power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2 (+10%)</td>
<td>0.89</td>
<td>1.46</td>
</tr>
<tr>
<td>M2 (-10%)</td>
<td>-0.75</td>
<td>-0.69</td>
</tr>
<tr>
<td>M3 (+10%)</td>
<td>1.90</td>
<td>2.83</td>
</tr>
<tr>
<td>M3 (-10%)</td>
<td>-1.62</td>
<td>-1.85</td>
</tr>
<tr>
<td>M4 (+10%)</td>
<td>0.77</td>
<td>1.64</td>
</tr>
<tr>
<td>M4 (-10%)</td>
<td>-0.65</td>
<td>-0.84</td>
</tr>
<tr>
<td>M5 (+10%)</td>
<td>0.08</td>
<td>0.50</td>
</tr>
<tr>
<td>M5 (-10%)</td>
<td>-0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>M6 (+10%)</td>
<td>0.22</td>
<td>0.65</td>
</tr>
<tr>
<td>M6 (-10%)</td>
<td>-0.19</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Total gates=43K Total area=0.2mm²

FreePDK 45nm process
Non-Rectangular Transistor Modeling

• Existing compact device models (e.g., BSIM) do not handle non-rectangular geometries

• Device models for shape imperfections:
  – Polysilicon gate shape contours [Gupta SPIE’06]
  – Diffusion rounding [Gupta ASPDAC’08, Chan VLSID’10]
  – Line-end shortening: gate not completely formed [Gupta DAC’07]
  – Line-end rounding: “tapering”, “necking” or “bulging” [Gupta PMJ’08]
Polysilicon Rounding Model

- Line-edge roughness and poly rounding lead to NRG transistor

- Equivalent gate length (EGL) can be used to represent the current behavior of the transistor to communicate to SPICE

\[ I_{\text{eff}} = \sum_{j=1}^{n} I(MOSFET_j) \]

- Transistor’s gate
Narrow Width Effect (NWE)

- Dopant densities, well-proximity effects, line-end capacitive coupling, etc. change with distance from STI edge
  - Non-uniform Vth along channel width
  - Ion/Ioff vs. W plot is not perfectly linear
- The extent and kind of behavior are very process-dependent

<table>
<thead>
<tr>
<th>Variation sources</th>
<th>Vth edge/Vth middle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe capacitance</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>Well proximity</td>
<td>&gt;= 1</td>
</tr>
<tr>
<td>STI Stress</td>
<td>&lt;= 1</td>
</tr>
</tbody>
</table>
Modeling Location Dependent $V_{th}$

- Threshold voltage modeled as a function of location along channel width

$$V_{th}(x) = \begin{cases} 
V_{th}(\text{middle}) - K_1(x - w)^2 + K_2(x - w) & 0 \leq x \leq w \\
V_{th}(\text{middle}) & w \leq x \leq W - w \\
V_{th}(\text{middle}) - K_1(W - x - w)^2 + K_2(W - x - w) & W - w \leq x \leq W 
\end{cases}$$

- $K_1$ and $K_2$ can be fitted purely in SPICE regime
  - NWE effect in BSIM $\Rightarrow$ $I_{\text{off}}$ vs. Width plot
  - $V_{th}$ vs. location can be fitted such that $I_{\text{off}}$ of transistor slices match $I_{\text{off}}$ vs. Width plot

- Parameters of $V_{th}$ model are estimated using $I_{\text{off}}$ data, which is much more sensitive to $V_{th}$
Device Level Modeling Results

TCAD

Uniform $V_{th}$

Location dependent $V_{th}$
Compact Model for Circuit Simulation

- EGLs depend on transistor working states
  - EGLs are extracted at $|V_{gs}| = 0$ and $|V_{gs}| = V_{dd}$ for leakage and timing analysis, respectively

- Alternatives:
  - Model a transistor by multiple smaller transistors connected in parallel [Sreedhar ICCD’08]
    - Accurate but number of transistors increases
  - Fit $L_{eff}$ and $V_{th}$ for $I_{on}$ and $I_{off}$
    - Only a set of parameters for a transistor
Other Circuit Models
• Express gate length as a function of $V_{gs}$ in device’s model (e.g., BSIM)
  – Given $L_{eff}$ at $V_{gs} = 0$ and $V_{gs} = V_{dd}$,
  – Intermediate gate length can be estimated using close form equation [Singhal DAC’07]
• Model the impact of gate length variation using voltage dependent current source [Shi ICCAD’06]
  – $I-V$ curve is calculated based on transistor’s shape.
  – $\Delta I$ due to non-rectangular gate is extracted and modeled as a current source connected in parallel to the transistor
The Flip Side [SPIE’08]

- Use the models to draw non-rectangular transistors intentionally to reduce power
- Proposed alternative: shape the transistor channel to create a dominant device
  - Lower leakage, faster delay, smaller capacitance
- 90nm simulation results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C5315</td>
<td>1.96</td>
<td>1.95</td>
<td>31.93</td>
<td>30.46</td>
<td>4.6</td>
</tr>
<tr>
<td>C6288</td>
<td>5.62</td>
<td>5.61</td>
<td>39.66</td>
<td>38.38</td>
<td>3.2</td>
</tr>
<tr>
<td>C7552</td>
<td>3.19</td>
<td>3.19</td>
<td>36.78</td>
<td>35.08</td>
<td>4.6</td>
</tr>
<tr>
<td>i2</td>
<td>0.86</td>
<td>0.86</td>
<td>13.55</td>
<td>12.80</td>
<td>5.5</td>
</tr>
<tr>
<td>i3</td>
<td>0.45</td>
<td>0.45</td>
<td>6.07</td>
<td>5.74</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Leakage Improvement vs Width
Why are Rectangular Gate Channels Suboptimal?

- Ion/Ioff densities depend on distance from device (STI) edge
  - Line-end capacitance, dopant scattering, well proximity effects $\Rightarrow$ lowered $V_{th}$ near edges $\Rightarrow$ Better delay-leakage tradeoff at the center than at edges $\Rightarrow$ Uniform channel length suboptimal
  - Longer $L$ at edges and shorter $L$ in center $\Rightarrow$ lower leakage for same delay
Its not only “L”: Diffusion Rounding

- Diffusion rounding occurs due to printing imperfection.
  - Diffusion routing
  - Pwr/Gnd connections
- Modeled as trapezoid gate to investigate electrical performance.

Victor Moroz, Munkang C. & Xi-Wei Lin SPIE 2009
Developing a Physical Diffusion+Poly Rounding Model [ASPDAC’09, VLSID’10]

• To capture two dimensional E field, slice channel according to its distribution
  • For each slice, $L_{\text{eff}-i} = L_i$
• Effective width is derived using gradual channel approximation:
  $$W_{\text{eff}-i} = \frac{(W_{s-i} - W_{d-i})}{\ln(W_{s-i} / W_{d-i})}$$
• $V_{\text{th}}$ varies due to NWE and asymmetry between source and drain
• Using charge sharing model:
  $$\Delta V_{\text{th-effective}} = \Delta V_{\text{th-Narrow width}} + \Delta V_{\text{th-CS}}$$
  $$\Delta V_{\text{th-CS}} = \frac{qN_a W_c}{2L C_{\text{ox}}} \left[ \frac{2(L_d W_d + L_s W_s)}{W_d + W_s} - (L_d + L_s) \right]$$
Total Currents

- Each slice is trapezoidal with equivalent (rectangular) $L, W$ and $V_{th}$:

$$I_{total} = \sum_{i=1}^{n} f(L_i, W_i, V_{th_i})$$

- Second order effects (DIBL, short channel effects, etc) are implicitly considered in BSIM.
- Evaluate $I_{total}$ at $V_{gs} = 0V$, $V_{ds} = V_{dd}$ (off)
  $$V_{gs} = V_{dd}, \quad V_{ds} = V_{dd} \text{ (on)}$$
- With $I_{total}$, equivalent device for circuit simulation can be obtained using EGL or other methods.
TCAD vs Model (Diffusion Rounding only)

- Asymmetrical $I_{on}/I_{off}$ when rounding happens at Drain/Source terminals
  - $\Delta Vth$ varies according to drain/source ratio
  - Behavior is NOT symmetric w.r.t source/drain
# Poly+Diffusion Rounding

## Table

<table>
<thead>
<tr>
<th>L1 (nm)</th>
<th>L2 (nm)</th>
<th>W_d (nm)</th>
<th>W_1 (nm)</th>
<th>W_2 (nm)</th>
<th>Error (%)</th>
<th>TCAD cal.</th>
<th>SPICE cal.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I_{on}</td>
<td>I_{off}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I_{on}</td>
<td>I_{off}</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>155</td>
<td>26</td>
<td>0</td>
<td>-2.1</td>
<td>-0.8</td>
<td>-2.0</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>155</td>
<td>45</td>
<td>0</td>
<td>-2.0</td>
<td>0.7</td>
<td>-1.9</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>155</td>
<td>78</td>
<td>0</td>
<td>-2.8</td>
<td>0.4</td>
<td>-2.7</td>
</tr>
<tr>
<td>55</td>
<td>45</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>-0.7</td>
</tr>
<tr>
<td>35</td>
<td>45</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>-0.2</td>
</tr>
<tr>
<td>55</td>
<td>45</td>
<td>155</td>
<td>45</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>-1.4</td>
</tr>
<tr>
<td>55</td>
<td>45</td>
<td>155</td>
<td>0</td>
<td>45</td>
<td>NA</td>
<td>NA</td>
<td>-2.8</td>
</tr>
<tr>
<td>35</td>
<td>45</td>
<td>155</td>
<td>45</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>-2.4</td>
</tr>
<tr>
<td>35</td>
<td>45</td>
<td>155</td>
<td>0</td>
<td>45</td>
<td>NA</td>
<td>NA</td>
<td>-0.7</td>
</tr>
</tbody>
</table>

## Diagram

- **Diffusion rounding only** (Source side larger)
- **Poly rounding only**
- **Poly+ diffusion rounding**

## Average error:

**(Diffusion layer rounding only)**
- TCAD calibrated model = 1.6%
- SPICE calibrated model = 2.7%

**(Poly+ Diffusion layers rounding)**
- TCAD calibrated model = 1.6%
- SPICE calibrated model = 1.7%
Application on Logic Cells

<table>
<thead>
<tr>
<th></th>
<th>NAND_X1</th>
<th>NOR_X1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Spacing Reduced</td>
</tr>
<tr>
<td>Delay</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nominal (no defocus)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>worst (100nm defocus)</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>Leakage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nominal (no defocus)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>worst (100nm defocus)</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>area</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- At 100nm defocus
  - Δ Delay = 5%
  - Δ Leakage = 9%
- Design rule can be optimized.
Electrical Metrics for Line-end Tapers [DAC’07, SPIE’08, JM3’10]

- Line end imperfections \(\rightarrow\) “taper”
  - Impact
    - Worsened impact under misalignment
    - Major RET concern
    - Line-end extension rule: major cause of area increase
  - Metrics
    - Geometric: pullback, CD at gate edge
    - Electrical: currently non-existent
Electrical Impact of Line-End Problems

• LEE vs. Capacitance
Line-end extension increases $C_g$ because there exists fringe capacitance between line-end extension and channel.

• Capacitance vs. $V_{th}$
$C_g$ affects $V_{th}$, narrow width effect
  - $C_g$ increases $\rightarrow V_{th}$ decreases
  - $C_g$ decreases $\rightarrow V_{th}$ increases

• $V_{th}$ vs. Current
$I_{on}$ and $I_{off}$ are functions of $V_{th}$
  - $V_{th}$ increases $\rightarrow I_{on}, I_{off}$ decrease
  - $V_{th}$ decreases $\rightarrow I_{on}, I_{off}$ increase
Misalignment Model

- There exists misalignment error between gate and diffusion processes
- Overlapping region (=actual channel) can vary according to misalignment error
  - Increase linewidth variation
- Misalignment has a probability, $P(m)$

\[
I_{\text{exp}} = \sum_{m=1}^{5} P(m) \cdot I(m)
\]
Optimizing Line-End of SRAM

SRAM Bitcell Layout vs. Line-End Design Rule

(Line-End Length, Sharpness) vs. (Leakage, Area)

Large $n$ is better for leakage variation but it increases OPC and Mask costs.

According to the taper shape, LEE design rule can be optimized to reduce bitcell size.
Line-End Shortening (LES)

- Polysilicon does not cover active region completely
  - Sources: Misalignment and line-end pullback

- Transistor suffering LES:
  - Functionally correct
  - High Leakage power
  - May have hold time violation
Design Flow Integration

- Full-custom/Analog designs
  - SPICE or SPICE-like analyses flows
  - $W_{eq}$, $L_{eq}$ per transistor is sufficient
- Cell-based digital designs
  - Static analysis flows based on standard cell abstraction
    - One cell is 2-100 transistors
    - Timing/power views stored in pre-characterized “.lib” files
  - Analysis done at PVT “corners”
  - State of art 45nm logic designs have 10M+ cells and 50M+ transistors \(\Rightarrow\) Hierarchy preservation essential
Hierarchy recovery using Type-Mapping

- Hierarchy flattened due to *layout dependent* parameters
  - Stress, annealing, etch, lithographic variability
- Layout contexts for each device may be limited
  - Due to repetitive nature of the designs
- Create *type-variants*
  - Variants on the original hierarchical blocks
Hierarchical Type-Mapping

Type-Map “B”
Hierarchical Type-Mapping

This induces clusters in “A”
Hierarchical Type-Mapping

and in “C”
Hierarchical Type-Mapping

Type-Map “A”
Hierarchical Type-Mapping

This induces clusters in “C”
Hierarchical Type-Mapping

Type-Map “C”: Induced clusters can be merged
Application: Timing

- ISCAS `89 designs + Open Cores ALU in 45nm lib
  - Litho. Simulation at defocus 160nm + NRG models
- Create type variants of standard cells
- Slacks can be used to reduce the number of type-variants that are needed for accurate timing
  - Less than 1% with just 1 hour of library characterization time
Lithographic Defocus

- **Defocus = deviation from best focus**
  - Causes blurring of the image
- **Photolithography + defocus**
  - Causes bad printing, linewidth (e.g., gate length) variation
  - Is caused by wafer not being flat enough
  - Few 100nm of defocus $\rightarrow$ 10%-20% change in CD
  - One of the major causes of gate length variation
Layout Composability for SRAFs

- **SRAF insertion**
  - Leads to more allowed pitches
  - Needs discrete spacings between primary features
  - More is better

- **Assist-aware layout**
  - Assist-correct cell libraries
    - Cell layout to avoid forbidden pitches
  - Assist-correct placement
    - Intelligent whitespace adjustment in the placer to remove forbidden pitches

---

![Graph showing pitch vs. CD for different OPC (Optical Proximity Correction) scenarios: W/O OPC (Best DOF), W/O OPC (Defocus), Bias OPC (Defocus), SRAF OPC (Defocus). The graph indicates a reduction in CD with increased pitch for SRAF OPC compared to other scenarios.](image)
Dealing with Defocus in Placement

- Dummy non-printing shapes or sub-resolution assist features (SRAFs) inserted to make isolated lines look like dense \(\Rightarrow\) remove iso-dense bias
  - Forbidden pitch: spacing that does not accommodate assist features

- **Idea:** Perturb the **standard-cell placement to eliminate forbidden pitches**
  - Preserve timing by minimal perturbation
  - Preserve legality of placement by restricting movements to placement sites
AFCorr: SRAF-Correct Placement
[ASPDAC’05, TCAD’07]

• By adjusting whitespace, additional SRAFs can be inserted between cells
  – Resist image improves and avoids open fault at worst-case defocus

• **Problem:** Perturb given placement minimally to achieve as much SRAF insertion as possible
Horizontal AFCorr (H-AFCorr)

- Horizontal-forbidden pitch is caused by interactions of poly geometries in the same row
Vertical AFCorr (V-AFCorr)

- Vertical-forbidden pitch is caused by interactions of poly geometries between adjacent cell row.

Before V-AFCorr

After V-AFCorr
Perturbation (H- + V- AFCorr)

- AFCorr: H-AFCorr + V-AFCorr
  - Adjusting whitespace $\rightarrow$ additional SRAFs $\rightarrow$ reduce # of forbidden pitch
Solving AFDcorr

- Dynamic programming based solution
  - Similar in nature to placement legalization
- Similar ideas can be used for improving etch variability, double patterning composability, recommended rule coverage, etc

![Graph showing printing error reduction vs placement utilization.](graph.png)

**130nm example**

**90nm example**
Now something to NOT worry about: Double patterning impact on BEOL

- Two interleaved exposures → overlay error translates to linewidth/spacing error → possible R, C variations on interconnect
Design Considerations Reducing The Overlay Impact

[SPIE’09, TSM’10]

- Congestion reduces the impact
  - 3% WC $\Delta RC$ for positive tone in 3-line structure vs. 17% in 2-line structure
  - Not the case for negative tone

- Layout decomposition in +ve tone, causes increased S in some cases and reduced S in others (Same happens for W in –ve tone)
  - Use wire sizzling to reduce the impact
Overlay Impact in CPs

- Expected WC impact in CPs (overlay budget = 20% of the DR)

- Impact on Crosstalk Noise
What about “Correct By Construction”? 

- **2D** 
- **Limited** 
- **1D** 
- **Fixed pitch** 

### Area [μm²]

<table>
<thead>
<tr>
<th></th>
<th>2D</th>
<th>Ltd</th>
<th>1D</th>
<th>Ltd-fixed pitch</th>
<th>1D-fixed pitch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>51000</td>
<td>55000</td>
<td>63000</td>
<td>67000</td>
<td>71000</td>
</tr>
</tbody>
</table>

- **15%** increase in area for **2D** compared to **1D**

### Functional Yield

- **2D**: 0.955
- **Ltd**: 0.960
- **1D**: 0.965
- **Ltd-fixed pitch**: 0.970
- **1D-fixed pitch**: 0.975

- **3x** improvement in functional yield for **1D-fixed pitch** compared to **2D**

### Variability

- **2D**: 3.0%
- **Ltd**: 2.5%
- **1D**: 2.0%
- **Ltd-fixed pitch**: 1.5%
- **1D-fixed pitch**: 1.0%

- **37%** reduction in variability for **1D-fixed pitch** compared to **2D**

- Fast layout estimation ← Fast topology generation + congestion estimation
  - < 1 hour runtime for a 45nm 100 cell library
  - Publicly available at http://nanocad.ee.ucla.edu/Main/DownloadForm

![Diagram of Design Rule Evaluator (DRE)]

- Layout policies
- Set of DR
- DR to evaluate
- Range of DR values

- Area: In progress
- Performance
- Power
- Designability
- Variability
- Manufacturability

![Graph showing estimated vs. actual area]

Estimate = actual
Handling Random Variations
First an Example of what you may not need to worry about (too much): FinFET LER

- LER is random device to device $\rightarrow$ on a digital design it averages out nicely especially for high-performance circuits.
Typical Random Variation Models

• Process variation is decomposed to inter-die, within-die spatial, and within-die random variation

\[ X = X_g + X_s + X_r \]

– Within-die spatial variation \( X_s \) assumed spatially correlated

• Several complex models of correlation exist

– Let's take a step back: what causes spatial variability?
The Reason: Across-Wafer Variation

• Across-wafer frequency variation e.g., [Qian, SPIE’09]
  – Usually parabolic
  – From the die point of view, the parabolic across-wafer systematic variation appears to be spatially correlated variation
  – After subtracting across wafer variation, pure random within-die variation is almost uncorrelated e.g., [Friedberg, SPIE’06]

• Across-wafer variation is not purely random \(\Rightarrow\) cannot be modeled as random correlated variation
Delving Deeper: Physical Origins

• Overlay error
  – Position and rotation of the wafer
  – Wafer stage vibration
  – Distortion of the wafer

• Nonuniformity
  – Higher temperature near the center of the wafer (PEB)
  – Center peak shape of the electric field distribution and chamber wall conditions in plasma etch

• Nonuniformity and distortion varies radially
  – Wafer are rotated to improve uniformity in the tangential direction

• All these are largely systematic phenomena ⇒ need to model them
Slope Augmented Across-Wafer Variation Model (SAAW)

\[
V_p (x, y) = a(x_c + x')^2 + b(y_c + y')^2 + c(x_c + x') + d(y_c + y') + s_x x' + s_y y' + m_w + r
\]

- The location of the die in the wafer is not known to designer
  - Model \( x_c \) and \( y_c \) as random variables evenly distributed in the circular wafer

- Advantage
  - Exactly models the across wafer variation
  - Only 6 random variables: \( X_c, Y_c, s_x, s_y, m_w, \) and \( r \)
  - Number of random variables does not depend on chip size
  - Number of random variables of grid based spatial variation model depends on number of grids
    - Larger chips have more grids
  - Process does not see die boundaries, only wafer (and field) boundaries!
**Few Observations**

- **Different locations on die have different means and variances**
  - Difference depends on ratio between die size and wafer size

![Graphs showing μ vs r_d and σ² vs r_d](image)

- Correlation coefficient $\rho$ is within a narrow range but covariance is not
  - This explains why people find that correlation coefficient only depends on distance but incomplete picture!
Accuracy/Runtime Tradeoffs

<table>
<thead>
<tr>
<th></th>
<th>SAAW</th>
<th>LDAW</th>
<th>SPC</th>
<th>IW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>μ</td>
<td>σ</td>
<td>95%</td>
<td>T</td>
</tr>
<tr>
<td>C1908</td>
<td>1.4</td>
<td>1.8</td>
<td>2.0</td>
<td>26</td>
</tr>
<tr>
<td>C3540</td>
<td>0.6</td>
<td>1.1</td>
<td>1.9</td>
<td>35</td>
</tr>
<tr>
<td>c7552</td>
<td>1.5</td>
<td>1.4</td>
<td>1.6</td>
<td>101</td>
</tr>
</tbody>
</table>

Absolute error percentage for 2cmX2cm Chip

- **SAAW**: Slope-augmented Across-Wafer Model
- **LDAW**: Location Dependent Across Wafer Model (change mean and sigma depending on (x,y) location of gate)
- **SPC**: Grid-based spatial correlation model
- **IW**: Simple within-die, die-to-die model
- **SAAW** is \(~5X\) faster and 50% more accurate than spatial correlation
## Comparison of Different Models

<table>
<thead>
<tr>
<th>Across-wafer models</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Model</th>
<th># of RVs</th>
<th>Case to be applied</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Efficient and accurate</td>
<td>Need die tracking to extract</td>
<td>SAAW</td>
<td>6</td>
<td>Large chip, non-parabolic across-wafer variation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QAW</td>
<td>4</td>
<td>Large chip, parabolic across-wafer variation</td>
</tr>
<tr>
<td>Old Models</td>
<td>Easy to extract</td>
<td>Not accurate</td>
<td>LDAW</td>
<td>2</td>
<td>Small chip</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPC</td>
<td>Depends on Chip size</td>
<td>Large chip</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IW</td>
<td>2</td>
<td>Small chip</td>
</tr>
</tbody>
</table>

Suggested approach:

- **SAAW model for moderate to large chips (≥6mm x 6mm)**
- **LDAW model for everything else**
The Too Many Models Conundrum

• Different types of variation models
  – Differing accuracy/runtime tradeoffs
    • Corners, 2-level global/local, spatial correlation, etc
  – Different design tools require different models
  – Too much calibration maintenance effort at foundry end
• Idea: just fit one (e.g., SAAW) model and derive (closed-form) others from it → a levelized modeling structure
**Example Levelized Variation Model**

- General variation model

\[ v(x, y) = v_0 + m_l + m_w + m_d + v_w(x, y) + v_f(x, y) + v_d(x, y) + r \]

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Complexity</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-lot</td>
<td>Inter-Wafer</td>
<td>Inter-die random</td>
</tr>
<tr>
<td>General</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sim 1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sim 2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Inter-/within die</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Spatial</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

1 across-field variation is lumped into inter-die and across-die variation
2 across-die variation is lumped into within-die random variation
3 across-wafer variation is lumped into inter-die random and within-die random variation
4 across-wafer, across-field, and across-die variations are modeled implicitly as spatial variation
5 spatial variation model is more accurate than inter-within-die model but less efficient and less accurate than all other models
Comparison for Different Models

- Run time and accuracy comparison

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>Sim2</th>
<th>Inter-/within-die</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error %</td>
<td>T (ms)</td>
<td>Error %</td>
<td>T (ms)</td>
<td>Error %</td>
</tr>
<tr>
<td>C1908</td>
<td>1.0</td>
<td>146</td>
<td>2.3</td>
<td>54</td>
</tr>
<tr>
<td>c3540</td>
<td>0.7</td>
<td>212</td>
<td>1.2</td>
<td>76</td>
</tr>
<tr>
<td>c7552</td>
<td>0.2</td>
<td>435</td>
<td>1.4</td>
<td>115</td>
</tr>
</tbody>
</table>

Accuracy of model simplification

<table>
<thead>
<tr>
<th></th>
<th>Sim2</th>
<th>Inter-/within-die</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract from Measurement</td>
<td>2.3</td>
<td>6.9</td>
<td>3.8</td>
</tr>
<tr>
<td>Obtain from Level1 model</td>
<td>2.9</td>
<td>7.4</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Error % for C1908
Are Variation Models Reliable?

- Process variation is decomposed into 4 components:
  - within-die (21%), Tens of measured device per die
  - die-to-die (39%), Hundreds of dies per wafer
  - wafer-to-wafer (21%), Tens of wafers per lot
  - lot-to-lot (19%)

- Number of measured lots or output lots is usually not large
  - Uncertainty of mean and variance mainly comes from lot-to-lot variation

<table>
<thead>
<tr>
<th>Cases</th>
<th>$\hat{n}$</th>
<th>$\tilde{n}$</th>
<th>Confidence interval</th>
<th>Reliability of Statistics analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-L</td>
<td>Large</td>
<td>Large</td>
<td>Small</td>
<td>High</td>
</tr>
<tr>
<td>S-L</td>
<td>Small</td>
<td>Large</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>L-S</td>
<td>Large</td>
<td>Small</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>S-S</td>
<td>Small</td>
<td>Small</td>
<td>Large</td>
<td>Low</td>
</tr>
</tbody>
</table>
Comparison for S-S, L-S and S-L

90% confidence worst case fast corner for different $n$

\[
\tilde{c} n_{f/s} = \tilde{\mu}_t \pm k_{f/s} \tilde{\sigma}_t
\]

- Example computation of “fast” corner of a parameter
  - Need 3.3% margin even with 80 characterization lots!
  - Need 3.3% margin even with 60 manufactured lots (~1.5M chips) → Low volume designs should be really worried
SPICE Fast/Slow Corner Model

• SPICE corners obtained from measuring inverter chain delay
• Up to 3.4% guard band value needs to be added to achieve high confidence
  – Remember SS-TT corners are usually separated by 10% - 20%
• Similar numbers for SSTA, etc

<table>
<thead>
<tr>
<th>conf_l</th>
<th>conf_h</th>
<th>L_f</th>
<th>V_{tnf}</th>
<th>V_{tpf}</th>
<th>L_s</th>
<th>V_{tns}</th>
<th>V_{tps}</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>60</td>
<td>0.07</td>
<td>0.21</td>
<td>0.20</td>
<td>0.13</td>
<td>0.39</td>
<td>0.36</td>
</tr>
<tr>
<td>70</td>
<td>80</td>
<td>0.29</td>
<td>0.86</td>
<td>0.81</td>
<td>0.32</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>90</td>
<td>95</td>
<td>0.79</td>
<td>2.36</td>
<td>2.22</td>
<td>0.71</td>
<td>2.13</td>
<td>2.00</td>
</tr>
<tr>
<td>95</td>
<td>99</td>
<td>1.15</td>
<td>3.44</td>
<td>3.23</td>
<td>0.90</td>
<td>2.71</td>
<td>2.54</td>
</tr>
</tbody>
</table>

Guard band percentage of different variation sources $\hat{n} = 10$, $\tilde{n} = 15$
Another Case for Simplicity: Statistical Power Optimization

- Costs
  - Slower Tools
  - Modeling: Extract statistics
    - Library Characterization, test chips, etc
- Do we really need statistical power optimization? → can we do with modeling power in traditional deterministic fashion (Note: this is just about power, independent of SSTA or not)
  - Practical Solvers: Solution Rankings
    - If a deterministic solution is in the top 10% of all deterministic sizings, will it in the top 10% of all statistical solutions?
  - Sub-optimality Bounds
    - “What is the maximum improvement that can be gained by optimizing statistically?”
  - Results for benchmark designs in 45 nm library with gate width sizing (similar results for L biasing, Vt assignment)

![Graph showing leakage, total power, and performance variability from 2007 to 2015.](From ITRS Roadmap 2007 (Design))
Solution Rankings

**Goal:** Relate the deterministic and statistical powers

- **Input:** 10,000 different assignments
  - Gate widths, lengths, vth,
- **Output:** Statistical Power vs. Deterministic Power Plot

**Benchmark c7552**

Mean + 3 Sigma

Power

Relation is nearly linear!
(Small amount of noise)

~10%

~10%
Solution Rankings

Quantitative Measurements using Kendall’s $\tau$:

- $\tau = 1 \rightarrow$ rankings are identical ($\tau = -1 \rightarrow$ reversed)
- $\tau = 0 \rightarrow$ rankings are uncorrelated
Suboptimality bounds

**Given:** Optimal deterministic sizing solution

**Find:** Maximum improvement that can be gained by optimizing statistically?

**Method:**

- Power optimization finds the minimum power solution in the *timing feasible region*
  - Timing feasible region is difficult to characterize quickly
- Approximate the timing feasible region
  - Find the *minimum power point in the relaxed region*
  - This is a **lower bound**
Calculating lower bounds: Example

- Visual example: two gates

Minimum Statistical Value in half-space (lower bound)

Corresponding Statistical Value

Deterministic Power Measure

Statistical Power Measure

Optimize over the relaxed region

Timing feasible region

Iso-power plane

Deterministic Optimum Bounded by half-space!!

Sizes with greater deterministic power

Optimize over the relaxed region

Timing feasible region

Deterministic Optimum Bounded by half-space!!

Visual example: two gates

- Gate size 1
- Gate size 2

Power

NanoCAD Lab
Suboptimality bounds: Results

\[ \sigma_L = 1 \text{ nm} \ (dtd) / 0.5 \text{ nm} \ (wid) \]

\[ \sigma_{Vth} = 4.7\% \ (dtd \text{ and } wid) \]

Worst-case Suboptimality

Benchmark

\[ \sigma_L = 2 \text{ nm} \ (dtd) / 0 \text{ nm} \ (wid) \]

\[ \sigma_{Vth} = 4.7\% \ (dtd \text{ and } wid) \]

Mean Power

Mean + 3 Sigma Power
Proxy for full statistical power optimization

- For **Mean + 3 Sigma** Power Optimization
  - Bounds are inconclusive (>10%) for large (2nm) $l$ variations
  - Measure is not linear – requires the computation of correlations between gate leakages
- Simple (**linear**) proxy measure can be used in deterministic power optimization to get within 3.5% of optimal power:

  $$p_{\text{approx}} = p_{\text{nominal}} \cdot \exp\left(\sigma_{\text{dtd}} + \sigma_{\text{wid}} / 2\right)$$

  $\sigma$: fitted lognormal parameter

  - Similar to a 1-$\sigma$ measure
  - Additive – does not require complex correlation estimates
Suboptimality of the Proxy Measure

- For $w$, $l$ and $v_{th}$ sizing

Worst-case Suboptimality of Proxy Measure

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>c432</th>
<th>c499</th>
<th>c880</th>
<th>c1355</th>
<th>c1908</th>
<th>c2670</th>
<th>c3540</th>
<th>c5315</th>
<th>c6288</th>
<th>c7552</th>
<th>ALU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suboptimality</td>
<td>0.0%</td>
<td>2.0%</td>
<td>4.0%</td>
<td>6.0%</td>
<td>8.0%</td>
<td>10.0%</td>
<td>12.0%</td>
<td>3.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dealing with Hardware Variability in Higher Layers of Abstraction
Sources of Variability

Frequency variation in an 80-core processor within a single die in Intel’s 65nm technology

Semiconductor Manufacturing

Vendor Differences

Power variation across five 512 MB DDR2-533 DRAM parts [Hanson07]

Normalized frequency degradation in 65 nm due to NBTI [Zheng09]

Ambient Conditions

Aging

Variation in $P_{\text{sleep}}$ with temperature across five instances of an ARM Cortex M3 processor
The Hardware-Software Interface...

*Practice: over-design & guard-banding for illusion of rigidity*

Variation:
- 20x in sleep power
- 50% in performance

Time or part
Imagine a New Hardware-Software Interface..

- Minimal variability handling in hardware

- Hardware Abstraction Layer (HAL)

- Operating System

- Application

- Traditional Fault-tolerance

- Opportunistic Software

- Underdesigned Hardware
UnO Computing Machines: From *Crash-and-Recover* to *Sense-and-Adapt*

- **Do Nothing**
  - (Elastic User, Robust App)
- **Change Hardware Operating Point**
  - (Disabling parts of the cache, Changing V-f)
- **Change Algorithm Parameters**
  - (Codec Setting, Duty Cycle Ratio)
- **Change Algorithm Implementation**
  - (Alternate code path, Dynamic recompilation)

**Variability signatures:**
- cache bit map
- cpu speed-power map
- memory access time
- ALU error rates

**Variability manifestations:**
- faulty cache bits
- delay variation
- power variation

**Hardware can avoid overdesign as well as self-healing → lowered cost**

**Software leverages hardware maximally**

**Underdesign Mechanisms**
- fluid hw constraints
- application intent

NanoCAD Lab
UnO Adaptation

- Hardware can avoid overdesign as well as self-healing \(\rightarrow\) lowered cost
- Software leverages hardware maximally
- Example hardware signatures
  - Achievable frequency of different components
  - Power consumption vs. temperature
Example 1: Dutycycling in Embedded Sensing
[Hotpower’10, DATE’11]

$\uparrow$active time = $\uparrow$quality of sensing

$$\text{DC} = f \left( P_A, P_S, L, E \right)$$

DC: Duty Cycle
$P_A$: Active Power (W)
$P_S$: Sleep Power (W)
$L$: Lifetime (s)
$E$: Energy (J)

Joint work with Mani Srivastava, UCLA
Sleep Power Variability

ARM Cortex M3 Sleep Current (Room Temperature)

Atmel SAM3U4E Cortex M3 Sleep Mode, 32KHz Slow Oscillator Room Temperature
Sleep Power Variability Across Temperature

14x Variation

Current (μA) vs Temperature
Variability-Aware Duty Cycle

- Determine $P_A$, $P_S$ for each instance across a temperature profile → Duty cycle is determined for each instance, time unit

- When the difference ($P_A - P_S$) is constant across temperature, a uniform duty cycle throughout the lifetime is optimal
Software Architecture (TinyOS)

Duty Cycle Scheduler: \( DC = f(P_A, P_S, ...) \)

Hardware Signature

Adaptable Task

Adaptable Task

Traditional Task

allowable DC

\( P_A, P_S, ... \)
Improvement over Worst-Case Duty Cycle

average: 7x improvement
Lifetime reduction with Datasheet Spec DC

average: 50 days short for one year’s lifetime
Example 2: H.264 Video Encoding

- Inherent application adaptability can be leveraged to compensate performance variation and avoid software/hardware overdesign
  - Motion estimation, transform and entropy coding blocks with correlated or uncorrelated variation

All results on Mobile video sequence
Results

- 8% decrease in overdesign to ensure 80% yield
  - yield defined by a jitter (no frame loss) constraint
- 2.5dB PSNR improvement at 80% yield
  - 0.5-1dB is noticeable. See videos at http://nanocad.ee.ucla.edu/Main/Codesign
Now, Something that May not Work: Dynamic Reliability Management [DATE’11]

- **NBTI**: $|V_{th}|$ increases for negatively biased PMOS
  - $|V_{th}|$ increase causes **delay increase**
  - Delay increase can cause **timing failures**

- Degradation depends on **stress time** and **Vdd**
  \[ \Delta V_{th} \propto f(V_{dd}) \cdot t_{stress}^{time_{-}exponent} \]

- NBTI degradation is **front-loaded**
  - 50% of lifetime degradation occurs within 1.6 months
**NBTI Reaction-Diffusion Model**

- Holes interact with H-passivated Si atoms
- Holes break Si—H bonds at Si/SiO$_2$ interface, generating traps and freeing H atoms
- H atoms anneal a trap or diffuse through SiO$_2$
- $|V_{th}|$ increase proportional to number of traps

\[
\Delta V_{th} = \frac{q N_{it}}{C_{ox}}
\]

- Diffusion can also drive H atoms back toward interface when stress is relaxed $\rightarrow$ **Recovery**
Device-level Analytical Model

- Architecture-level techniques have been based on device-level analytical models

\[
\Delta V_{th} = A_{NBTI} \cdot \tau_{ox} \cdot \sqrt{C_{ox} (V_{dd} - V_{th})} \cdot e^{\frac{V_{dd} - V_{th}}{\tau_{ox} E_0} - \frac{E_a}{kT}} \cdot t_{stress}^{0.25}
\]

- Model fine for device-level analysis, but NOT valid for “dynamism” (Vdd change, power gating, etc) → used too much out of context! → possibly invalid conclusions

  - Some other issues (e.g., ignoring “CMOS” nature of designs) as well

- Solution: Flexible, Numerical but Fast Aging Model for NBTI
  - Arbitrary activity patterns, Vdd schedules, etc can be applied correctly
  - Freely downloadable from
    http://nanocad.ee.ucla.edu/Main/DownloadForm
NBTI Mitigation Techniques

• **Guardbanding** is traditional way to deal with NBTI
  – Increase voltage / Reduce Frequency / Increase Area

• Many works propose techniques to reduce the cost of provisioning for NBTI
  – **Dynamic Voltage Scaling**
    • Always use lowest possible supply voltage
  – **Activity Management**
    • Attempt to put PMOS in idle state
  – **Power Gating**
    • Relax all nodes by turning power off
Limitations of Device-level Models

- Analytical equation does not model physical degradation phenomenon
- Changing voltage in analytical equation is like instantaneously changing internal device state
Limitations of Device-level Models

- Alternating values in idle state models averaging effect of

![Graph showing normalized delay over time with different states and error bars.]

- <1% estimation error
- ±5% delay estimation error
Guardbanding vs. DRM

Dynamic Voltage scaling < 7% energy benefit not accounting for overheads

Power gating: useful but have to shut down processor for 6 out of 10 year lifetime to get 5% frequency benefit
Conclusions

• Variability can be and should be managed at all levels of hardware-software stack but carefully:
  – Managing at “too high” an abstraction level → lose accuracy, physical justification of models
  – Managing at “too low” an abstraction level → lose {design, application} intent

• In many cases, simpler and conventional methods don’t do too bad!