UCLA

Modeling Performance Impact of Variability

Puneet Gupta

Dept. of EE, University of California Los Angeles

(puneet@ee.ucla.edu)

Work partly supported by NSF, UC Discovery IMPACT and SRC.

NanoCAD Lab

http://nanocad.ee.ucla.edu/

Outline

- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

Taxonomy of Variations

- Source
  - Process: Litho, CMP, overlay
    - Typically permanent
  - Environment: Vdd, temperature
    - Typically transient
  - Vendor!
- Nature
  - Systematic: metal dishing, stress, RTA, litho proximity effects
  - Random: dopant fluctuations, material variations, LER
- Spatial Scale
  - Intra-die: litho proximity, CMP
  - Inter-die: material variations
    - Includes wafer-to-wafer, lot-to-lot variations

Progress = Random → Systematic

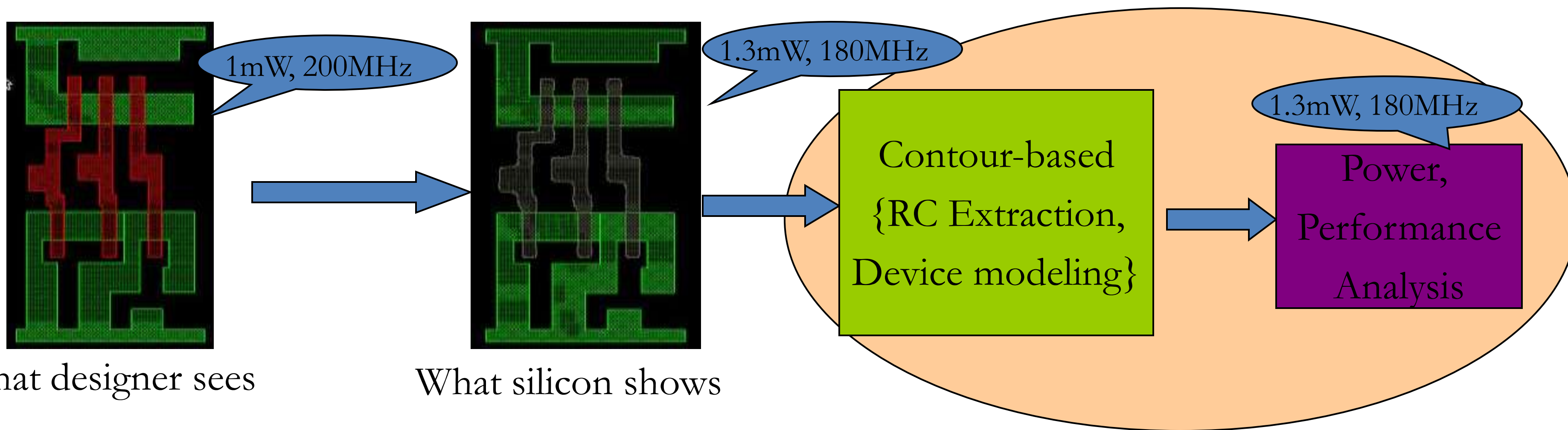
- Random variations
  - Seemingly or truly random behavior
    - E.g., dopant fluctuations
  - Predictable but too complex to model
    - E.g., crosstalk
    - Typically handled by worst-casing or statistics
    - Modeling and computational advancements → more effects can be modeled
- Systematic variations
  - Can be modeled, predicted given layout
    - E.g., CMP-dependent topography variation
  - Some variations are "trend-systematic"
    - E.g., relevant circuit parameter always increases though process parameter may be random
      - E.g., defocus

Variations: random now, systematic tomorrow

Outline

- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

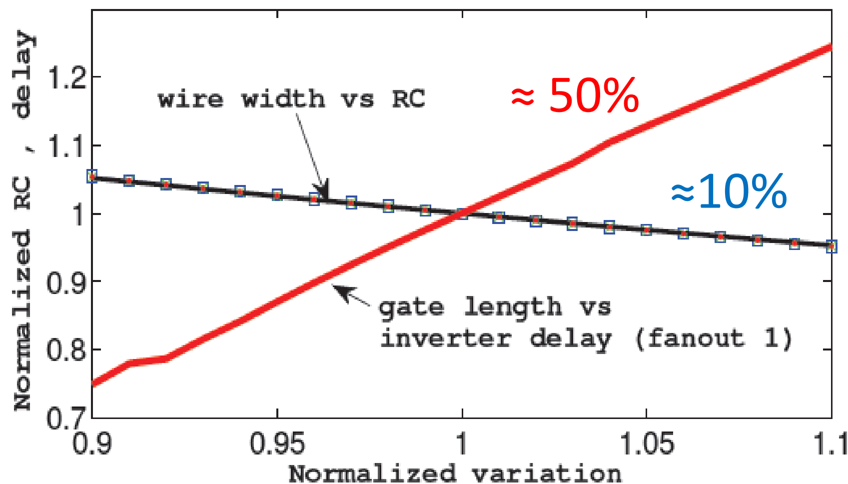
Lithographic WYSIWYG Breakdown



- Existing compact device models (e.g., BSIM) do not handle non-rectangular geometries.
- Where Are Electrical Models of Patterning Imperfections Needed?
  - Cells characterization
  - Electrically-driven OPC
  - Contour-based design analysis
  - Design rule optimization
  - Transistor shape optimization

Why Wires Are Not Important

- Width variation averages over long wires.
- Resistance and capacitance change in opposite directions as line width changes.
- Delay and switching power <3% at chip-level.
  - Impact of wire variation is exaggerated as averaging effect is ignored.



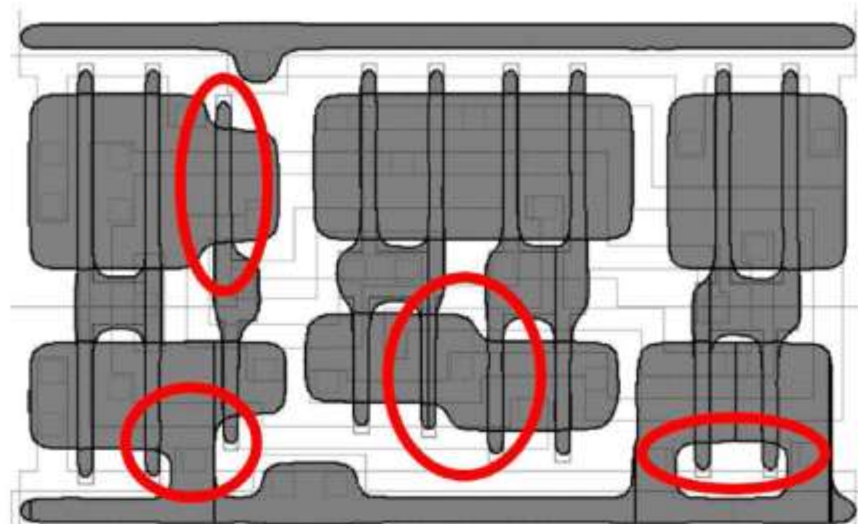
Interconnect layers (variation)	Δ delay (%)	Δ Switching power (%)
M2 (+10%)	0.89	1.46
M2 (-10%)	-0.75	-0.69
M3 (+10%)	1.90	2.83
M3 (-10%)	-1.62	-1.85
M4 (+10%)	0.77	1.64
M4 (-10%)	-0.65	-0.84
M5 (+10%)	0.08	0.50
M5 (-10%)	-0.07	0.13
M6 (+10%)	0.22	0.65
M6 (-10%)	-0.19	0.00

Total gates=43K Total area=$0.2mm^2$

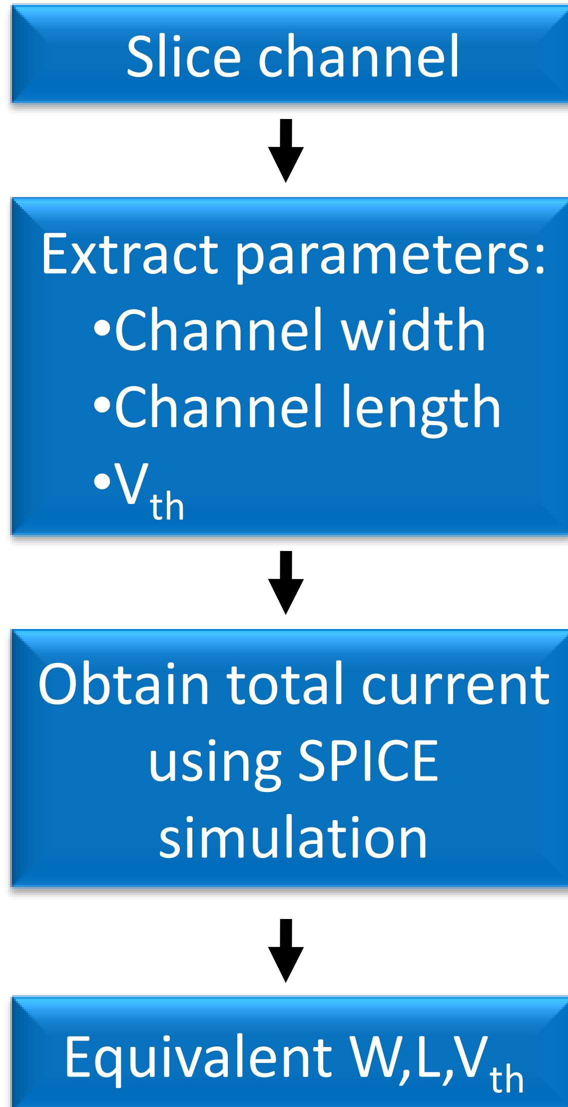
FreePDK 45nm process

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]



Modeling Diffusion+Poly Rounding



Channel width deviation

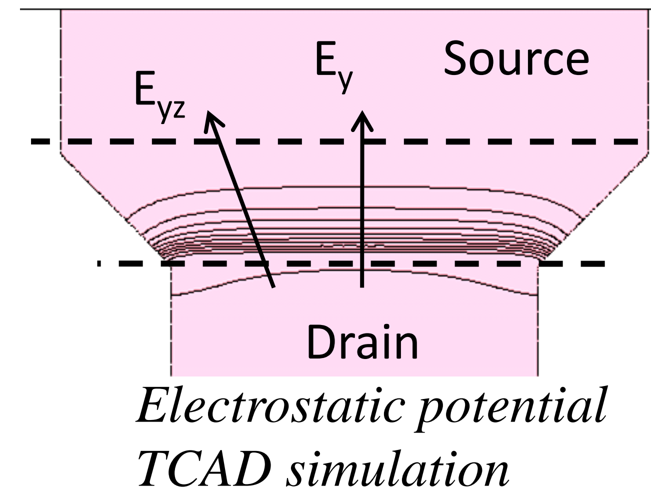
Drain

Source

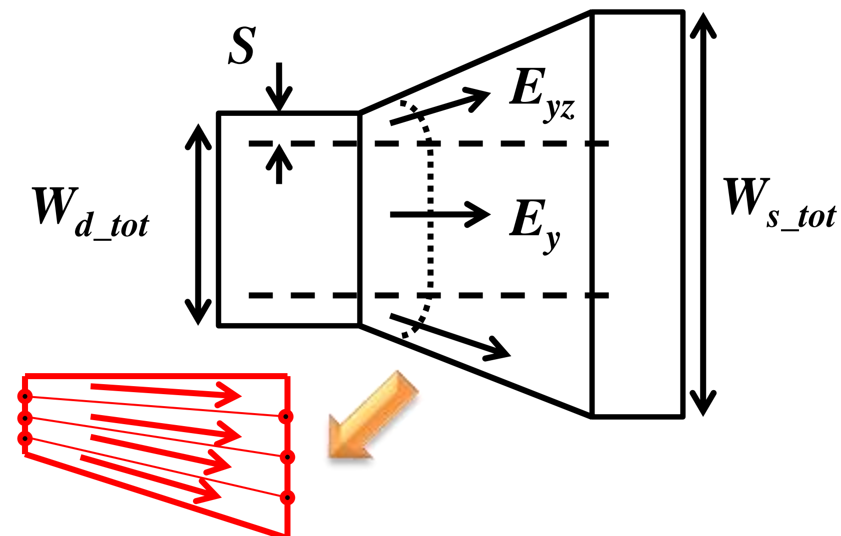
Location dependent channel length

Channel Slicing

- Channel's electrostatic potential is two-dimensional
  - Changes L_{eff} and W_{eff}
- Strategy: divide channel into 3 sections.
- Assume E-field is :
  - ➢ Purely horizontal in middle.
  - ➢ Changing linearly from middle to edges.
  - ➢ Channel length measured along E-field direction



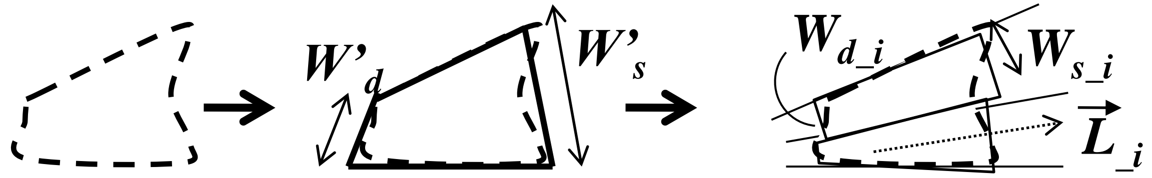
Electrostatic potential TCAD simulation



Effective Channel Width

- Effective width of sliced channel

$$W_{eff} = \frac{(W_s - W_d)}{\ln(W_s / W_d)}$$



W_{d_i} and W_{s_i} are obtained by approximating edges with straight lines orthogonal to the vector of channel length

- W_{eff} is derived based on gradual channel approximation
  → voltage varies gradually from drain to source

$$\int_0^L \frac{I_D.dy}{(W_d + (W_s - W_d)y/L)} = \int_{V_s}^{V_d} \mu Cox[V_G - V_{th} - V]dV$$

Channel width varies along channel

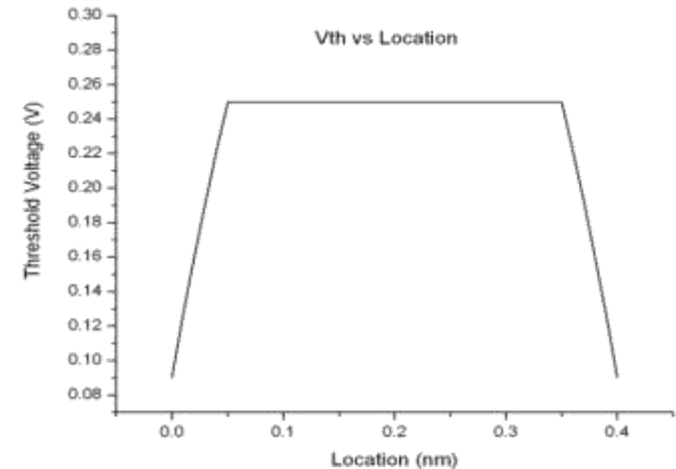
$$I_D = \frac{1}{L}\frac{(W_s - W_d)}{ln(W_s/W_d)}\mu Cox[V_G - V_{th} - \frac{V_{ds}}{2}]V_{ds}.$$

- Second order effects (DIBL, velocity saturation, etc.)
  - Considered by applying effective length, width and V_{th} in SPICE simulation with BSIM model.

ΔVth - Narrow Width Effect (NWE)

$$\Delta V_{th\text{-effective}} = \Delta V_{th-\text{Narrow width}} + \Delta V_{th-CS}$$

- Non-uniform V_{th} along channel width
  - Impact of NWE is modeled by fitting ΔV_{th} as a function of location [SPIE'06]



$$\Delta Vth(x) = \begin{cases} K_1(x-w)^2 + K_2(x-w) & 0 \le x \le w \\ 0 & w \le x \le W-w \\ K_1(W-x-w)^2 + K_2(W-x-w) & W-w \le x \le W \end{cases}$$

w is the maximum width affected by NWE

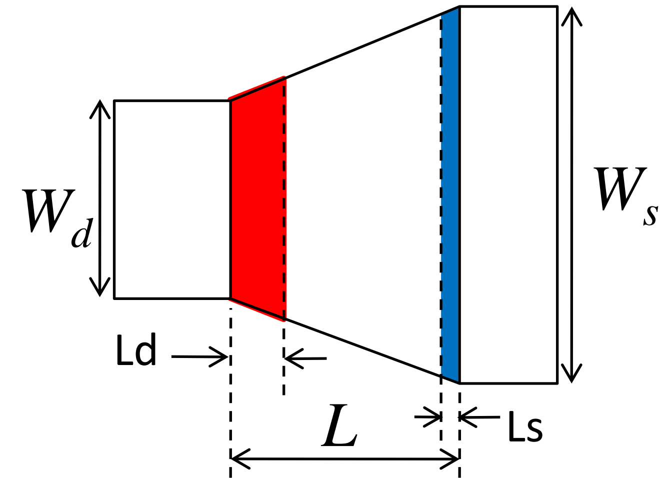
W is device's average width

- The extent of behavior depends on the process

Variation sources	Vth edge/Vth middle
Fringe capacitance	< 1
Well proximity	>= 1
STI Stress	<= 1

ΔVth – Asymmetrical Source/Drain

- A portion of depletion region is shared between gate and source/drain
- Asymmetric source/drain sharing regions change effective region supported by gate alone → V_{th} variation
- Charge Sharing Model :
  - $\Delta V_{th} \propto Q_{shared}$,
  - Estimate Q_{shared} based on device's geometry

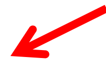


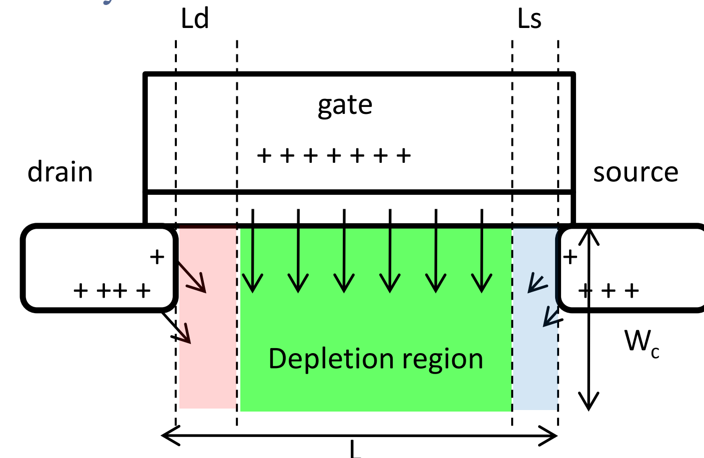
$$V_{th} = Vfb + 2\phi b + \left[\frac{Q_{total} - Q_{shared}}{C_{ox}WL}\right]$$

$$Q_{shared} = qN_a\left(\frac{W_c}{2}\right)(L_d W_d + L_s W_s)$$

$$\Delta V_{th-CS} = \frac{qN_a W_c}{2LC_{ox}}\left[\frac{2(L_d W_d + L_s W_s)}{W_d + W_s} - (L_d + L_s)\right]$$

Asymmetrical source and drain





Total Currents

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

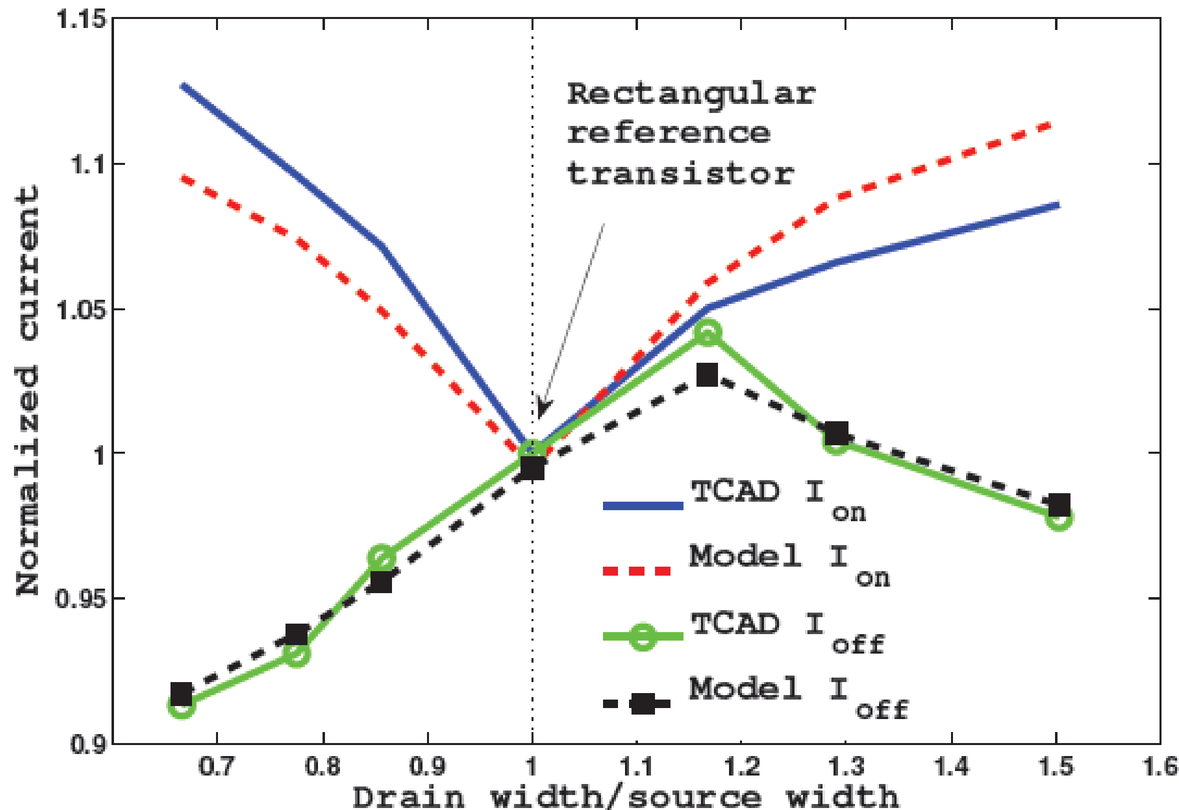
$$I_{total} = \sum_{i=1}^{n} f(L_i, W_i, Vth_i)$$

Can be obtained using conventional compact model e.g., (BSIM).

- 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)
  $V_{gs} = V_{dd}$ $V_{ds} = V_{dd}$ (on)
- With I_{total}, equivalent device for circuit simulation can be obtained using EGL or other methods.

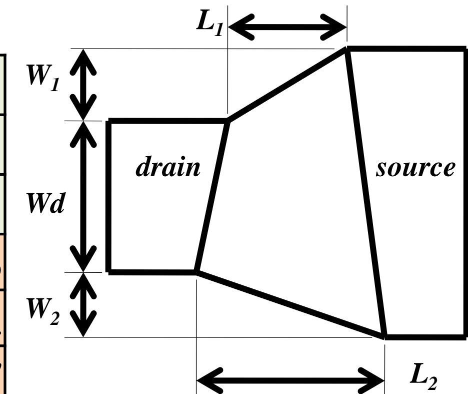
TCAD vs Model (Diffusion Rounding only)

- Asymmetrical I_{on}/I_{off} when rounding happens at Drain/Source terminals
  - ΔVth varies according to drain/source ratio



Poly+Diffusion Rounding

	L1 (nm)	L2 (nm)	W_d (nm)	W_1 (nm)	W_2 (nm)	Error (%)			
						TCAD cal.		SPICE cal.	
						I_{on}	I_{off}	I_{on}	I_{off}
Diffusion rounding only (Source side larger)	45	45	155	26	0	-2.1	-0.8	-2.0	-0.5
	45	45	155	45	0	-2.0	0.7	-1.9	1.1
	45	45	155	78	0	-2.8	0.4	-2.7	0.7
Poly rounding only	55	45	155	0	0	NA	NA	-0.7	2.5
	35	45	155	0	0	NA	NA	-0.2	7.5
Poly+ diffusion rounding	55	45	155	45	0	NA	NA	-1.4	3.1
	55	45	155	0	45	NA	NA	-2.8	-2.7
	35	45	155	45	0	NA	NA	-2.4	0.7
	35	45	155	0	45	NA	NA	-0.7	7.8



Average error :

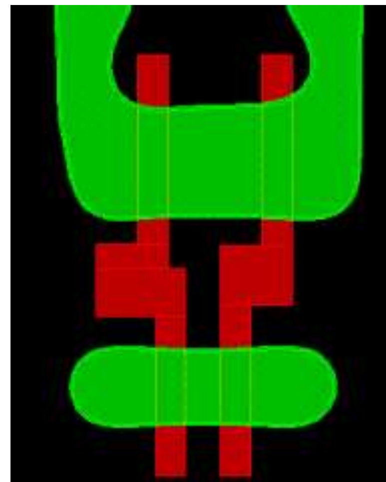
(Diffusion layer rounding only)
TCAD calibrated model = 1.6%
SPICE calibrated model = 1.7%

(Poly+ Diffusion layers rounding)
SPICE calibrated model =2.7%

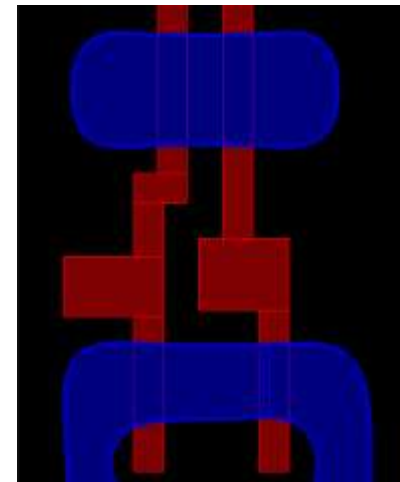
Application on Logic Cells

		NAND_X1		NOR_X1	
		Original	Spacing Reduced	Original	Spacing Reduced
Delay	nominal (no defocus)	1.00	1.00	1.00	0.99
	worst (100nm defocus)	1.05	1.04	1.05	1.05
Leakage	nominal (no defocus)	1.00	1.00	1.00	1.01
	worst (100nm defocus)	0.91	0.91	0.90	0.90
area		1.00	0.95	1.00	0.95

- At 100nm defocus
  - Δ Delay = 5%
  - Δ Leakage = 9%
- Design rule can be optimized.



NAND2_X1

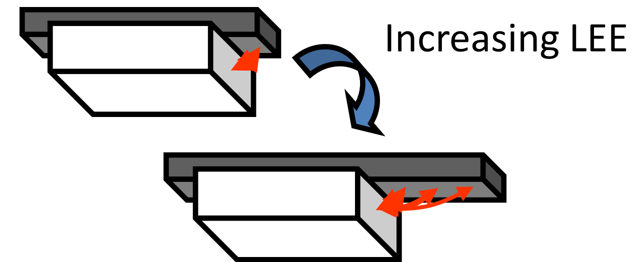


NOR2_X1

Electrical Impact of Line-End Problems

- LEE vs. Capacitance

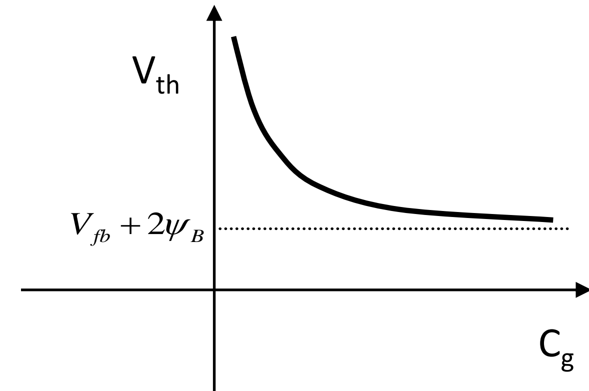
Line-end extension increases C_g because there exists fringe capacitance between line-end extension and channel.



- Capacitance vs. V_{th}

C_g affects V_{th}, narrow width effect

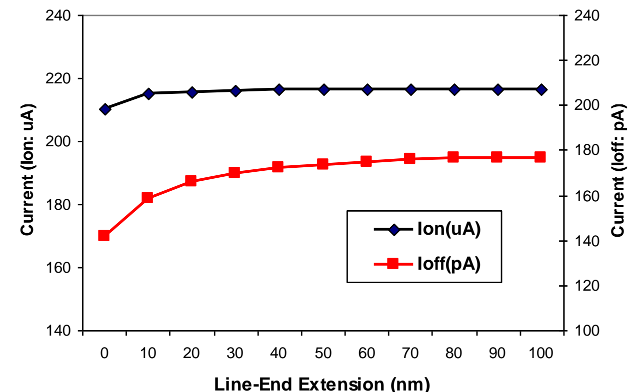
  - C_g increases → V_{th} decreases
  - C_g decreases → V_{th} increases



- V_{th} vs. Current

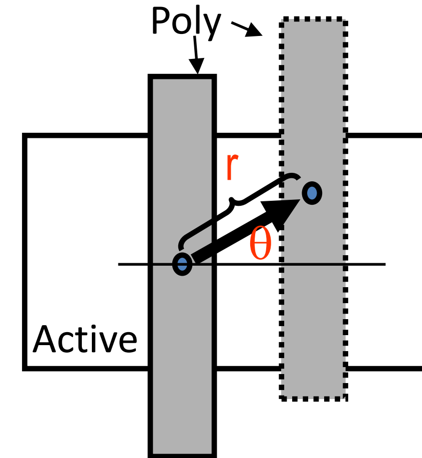
I_{on} and I_{off} are functions of V_{th}

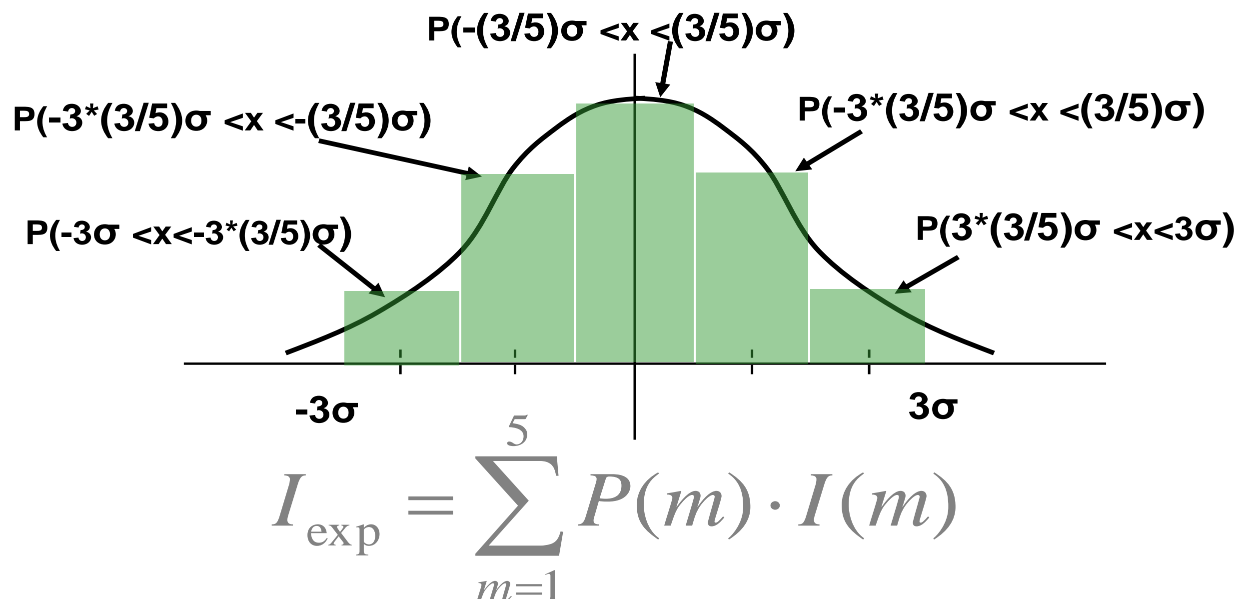
  - V_{th} increases→ I_{on}, I_{off} decrease
  - V_{th} decreases→ I_{on}, I_{off} increase



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)

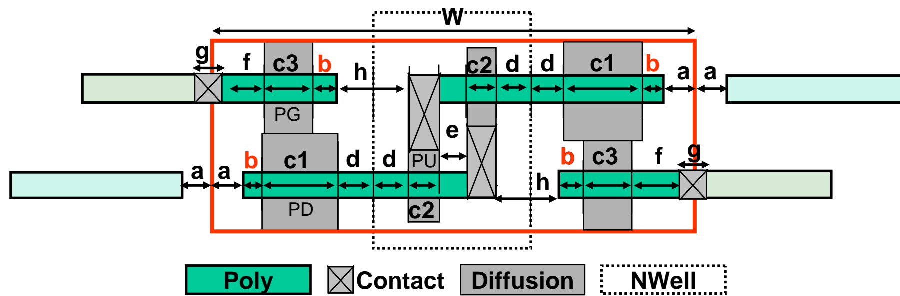


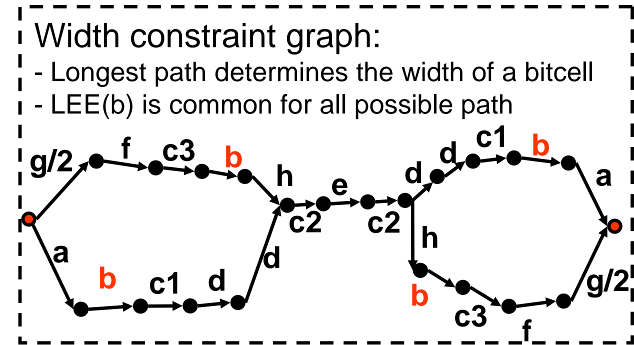


$$I_{\exp} = \sum_{m=1}^{5} P(m) \cdot I(m)$$

Optimizing Line-End of SRAM

SRAM Bitcell Layout vs. Line-End Design Rule

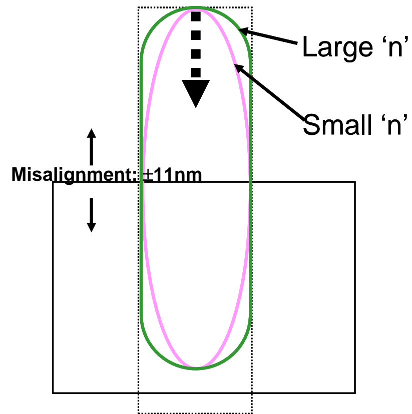


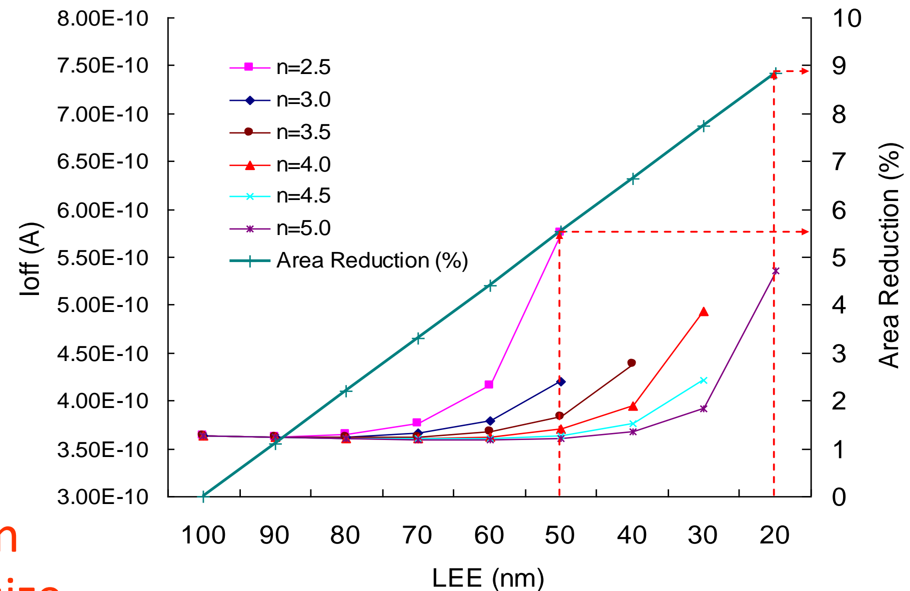


(Line-End Length, Sharpness) vs. (Leakage, Area)

Large *n* is better for leakage variation

but it increases OPC and Mask costs.

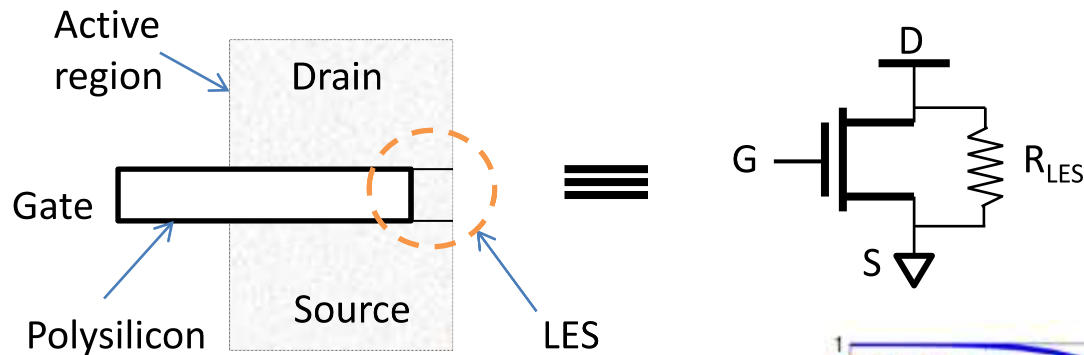




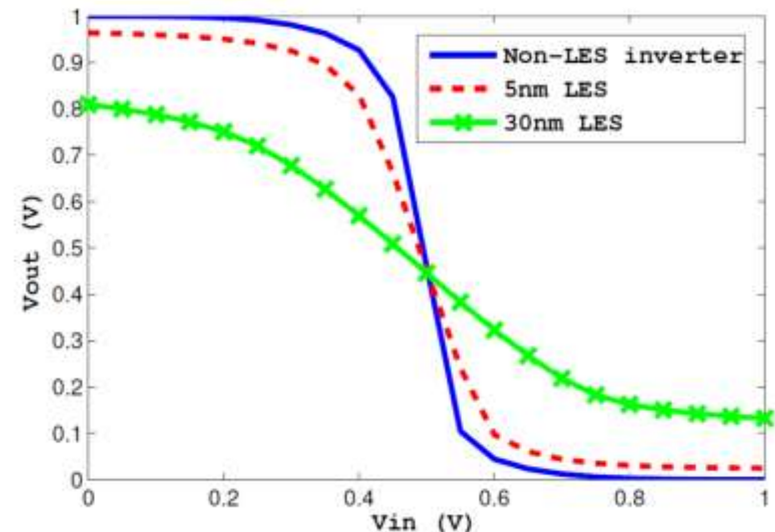
According to the taper shape, LEE design rule can be optimized to reduce bitcell size.

Line-End Shortening (LES)

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

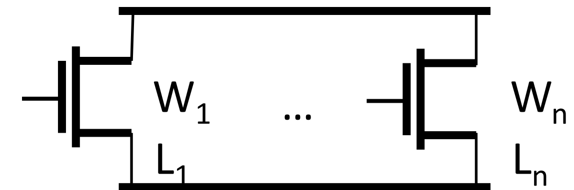


- Transistor suffering LES :
  - Functionally correct
  - High Leakage power
  - May have hold time violation

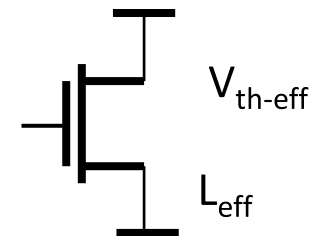


Compact Model for Circuit Simulation

- 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

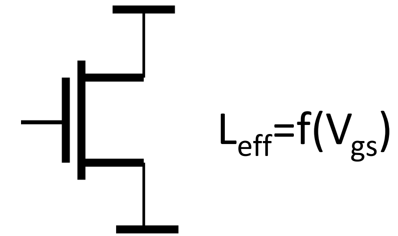


  - Fit L_{eff} and V_{th} for I_{on} and I_{off}
    - Only a set of parameters for a transistor

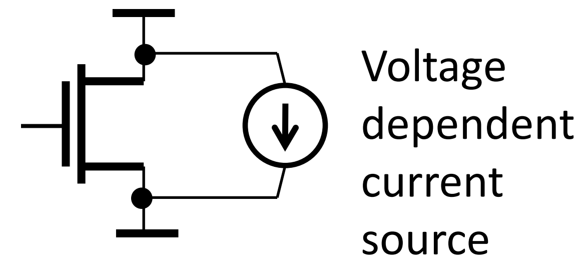


Other Circuit Models

- 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.
  - ΔI due to non-rectangular gate is extracted and modeled as a current source connected in parallel to the transistor



Other Layout Dependent Sources of Variability

- Layout-dependent stress variation (e.g., 15% ΔI_{on})
- Well proximity effect on V_{th} (e.g., up to 10% delay increase)
- Etch introduces CD variability with strong dependence on pattern-density within a few microns range
- RTA used in the fabrication of ultra-shallow junctions
  - Long-range effect (few millimeters)
  - Affects I_{on} / I_{off} ratio and V_{th}.
- CMP imperfections of dishing and erosion
  - Causes interconnect *RC* variability
  - Depends on line-width/spacing and pattern-density within a long-range (up to 100micron)

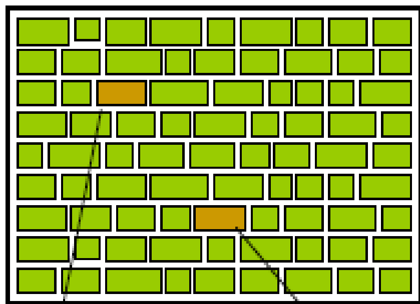
Outline

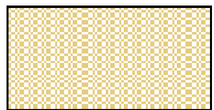
- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

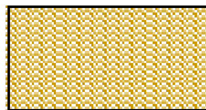
Design Flow Integration

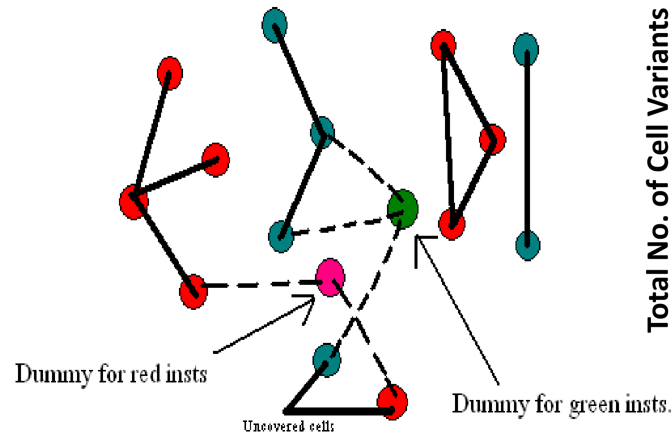
- Full-custom/Analog designs
  - SPICE or SPICE-like analyses flows
  - Weq, Leq per transistor is sufficient
- Cell-based digital designs
  - Static analysis flows based on standard cell abstraction
    - One cell is 2-100 transistors
    - Timing/power views stored in pre-characterized ".lib" files
  - Analysis done at PVT "corners"
  - State of art 45nm logic designs have 10M+ cells and 50M+ transistors →Hierarchy preservation essential
- *Problems are the same for other layout-dependent systematic variations*
  - Stress, etch, RTA

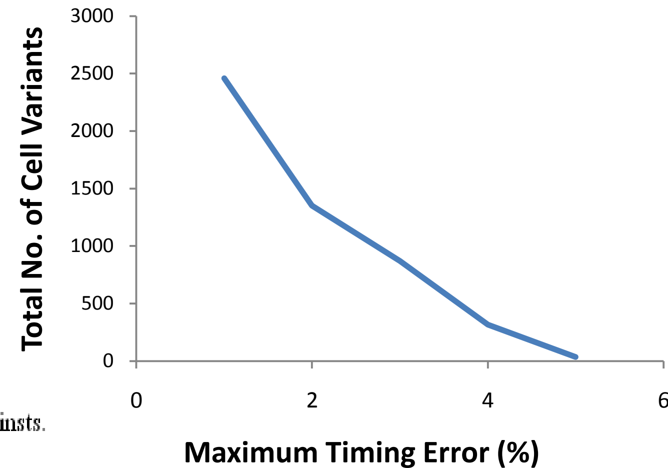
Recovering Hierarchy Parametrically











- Cluster "flattened" instances of a cell if they are parametrically (delay/power) close enough
  - Introduce "dummy" cell masters; or
  - Snap to pre-characterized masters
- Recover hierarchy, reduce characterization load

Adoption Challenge: SPICE vs. Litho Corners

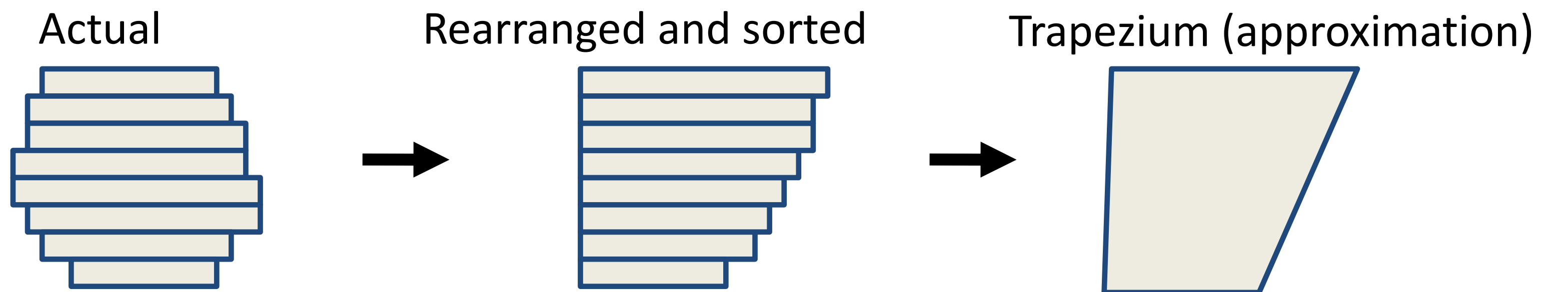
- Typical BSIM corner methodology
  - Based on a reference pattern context
    - FF, SS & TT correspond to the device placed in the reference context
    - Within this context, parameters (tox, Vt0, etc.) are fitted from silicon over multiple L and W bins
  - Litho-dependency in the pattern contexts outside the reference pattern is not accounted for
    - Prohibitive to cover all contexts
    - Some limited context-dependent "re-centering" of the model
- Typical litho process window
  - Across focus, exposure with multiple patterns
- No explicit connection between L/W variation in litho vs. SS-FF L/W variation in SPICE → No way to connect litho simulation across PW to circuit power/performance analysis

A Unified Corner Methodology

- Need to establish SPICE corner models that both lithography and SPICE communities can agree on
  - Filter out systematic, litho-dependent variation
  - Compatible with current SPICE corner model
- Possible solution #1
  - Reference context based correlation of litho corners and SPICE corners
    - Use SPICE calibration test patterns to calibrate F/E skew
    - BSIM corner model to contain only random and unmodeled systematic variation
- Possible Solution #2: Generate context-dependent BSIM corner models
  - Too many contexts → complex model extraction
  - No need for litho simulation
    - Ignores complicated 2D, long range effects

Decoupling Extraction and Modeling

- A clean flow (mimics the current BSIM + RCX flow)
  - Contour generation and shape extraction is better done by RCX tools
  - Modeling is done by foundry, contained in SPICE models
- Starting point: a compact model of the shape
  - NRG transistor are modeled as transistor slices connected in parallel
  - Detailed description of transistor slices is costly
    - (transistor #) x (slices #) x (geometrical info)
- Example Compact Shape Model :
  - Ignore narrow width effect → slices are independent → can be rearranged



  - L and W replaced by Lmin, Lmax, W → 1 extra layout-dependent parameter extracted by device extraction

Thanks: discussions with Dr. Sani Nassif, IBM

Outline

- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

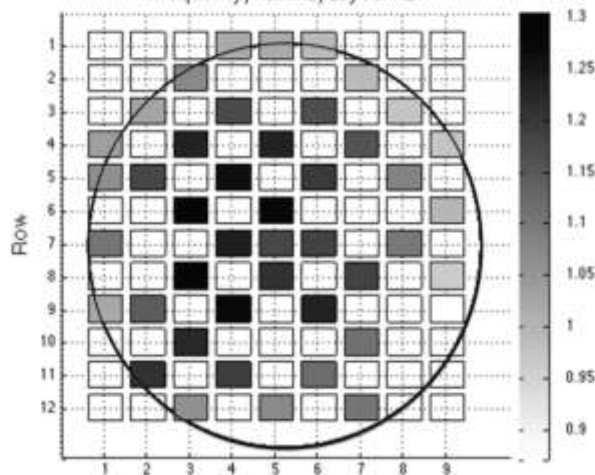
Typical Random Variation Models

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

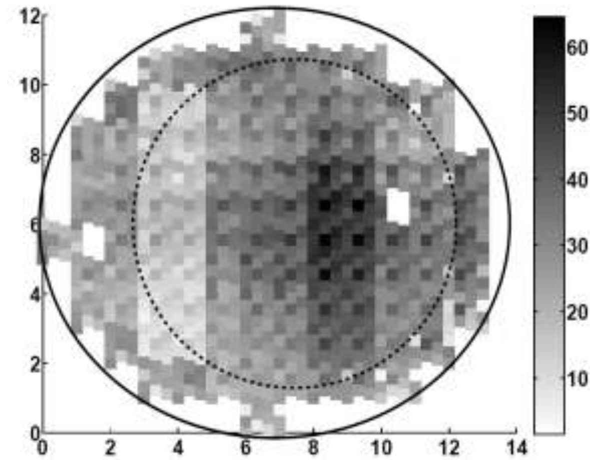
$$X = X_g + X_s + X_r$$

  – Within-die spatial variation X_s assumed spatially correlated
    - Several complex models of correlation exist
  – *Lets take a step back: what causes spatial variability ?*

The Reason: Across-Wafer Variation



Process 1



Process 2

- 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 → 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)

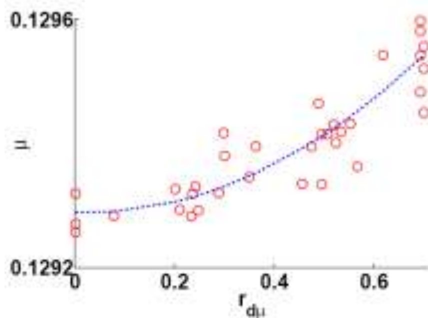
$$V_p(x,y) = a(x_c + x')^2 + b(y_c + y')^2 + c(x_c + x') + d(y_c + y') + s_x x' + s_y y' + m_w + r$$

Quadratic across-wafer variation Model | Linear fitting of residual | inter-die random variation | within-die random variation

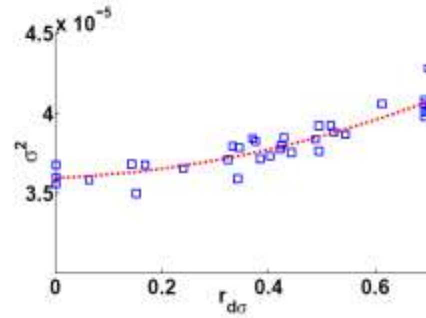
- The location of the die in the wafer is not known to designer
  - Model x_c and y_c as random variables evenly distributed in the circular wafer
- Advantage
  - Exactly models the across wafer variation
  - Only 6 random variables: X_c, Y_c, s_x, s_y, m_w, and r
  - Number of random variables does 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

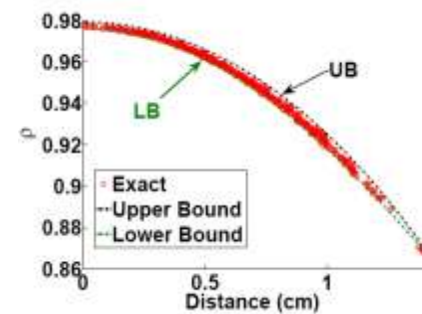
- ***Different locations on die have different means and variances***
  - Difference depends on ratio between die size and wafer size

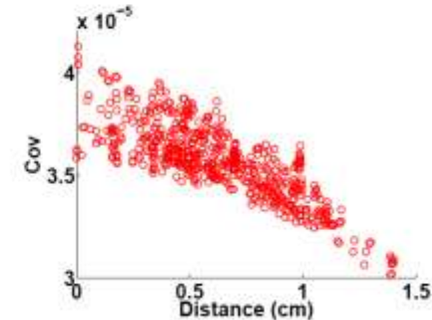


(a) μ Vs r_d



(b) σ^2 Vs r_d





- Correlation coefficient ρ 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

- Assume ISCAS benchmarks are stretched over a 2cmX2cm chip

	SAAW				SPC			
	μ	σ	95%	T	μ	σ	95%	T
C1908	**1.4**	**1.8**	**2.0**	***26***	**2.1**	**4.4**	**4.0**	**135**
C3540	**0.6**	**1.1**	**1.9**	***35***	**2.0**	**6.5**	**5.7**	**202**
C7552	**1.5**	**1.4**	**1.6**	***101***	**3.3**	**3.5**	**4.3**	**433**

Absolute error percentage for 2cmX2cm Chip

- SAAW is more accurate than QAW with a small increase of run time
- SAAW is ~**5X faster** and **50% more accurate** than spatial correlation
  - Far fewer random variables to deal with

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

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

Lump as inter-die

Accuracy Complexity ↑

Efficiency ↓

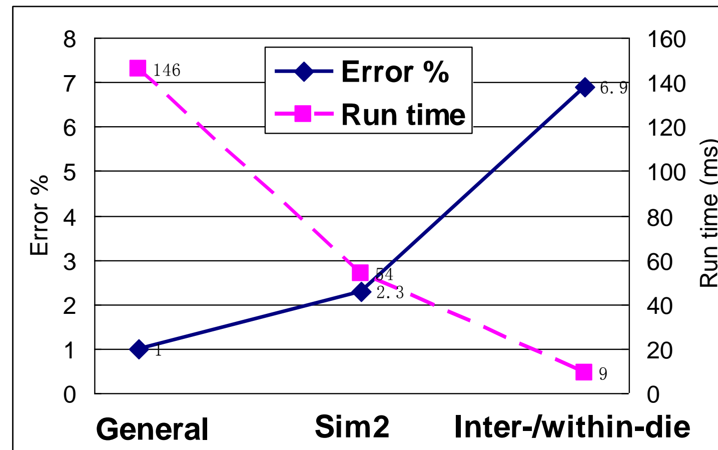
	Inter-lot	Inter-Wafer	Inter-die random	Across-wafer	Across-field	Across-die	Within-die random
General	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sim 1	Yes	Yes	Yes	Yes	No [1]	Yes	Yes
Sim 2	Yes	Yes	Yes	Yes	No	No[2]	Yes
Inter-/within die	Yes	Yes	Yes	No [3]	No	No	Yes
Spatial[5]	Yes	Yes	Yes	Yes[4]	Yes[4]	Yes[4]	Yes

[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

	General		Sim2		Inter-/within		Spatial	
	Error %	T (ms)	Error %	T (ms)	Error %	T (ms)	Error %	T (ms)
C1908	1.0	146	2.3	54	6.9	9	3.8	1450
c3540	0.7	212	1.2	76	4.6	13	4.0	4210
c7552	0.2	435	1.4	115	4.0	20	2.9	8182



Comparison for c1908

- Accuracy of model simplification

	Sim2	Inter-/within-die	Spatial
Extract from Measurement	2.3	6.9	3.8
Obtain from Level1 model	2.9	7.4	4.2

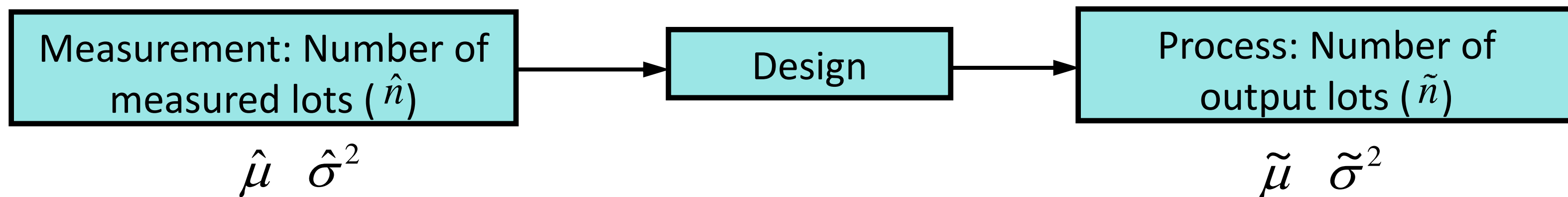
Error % for c1908

Outline

- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

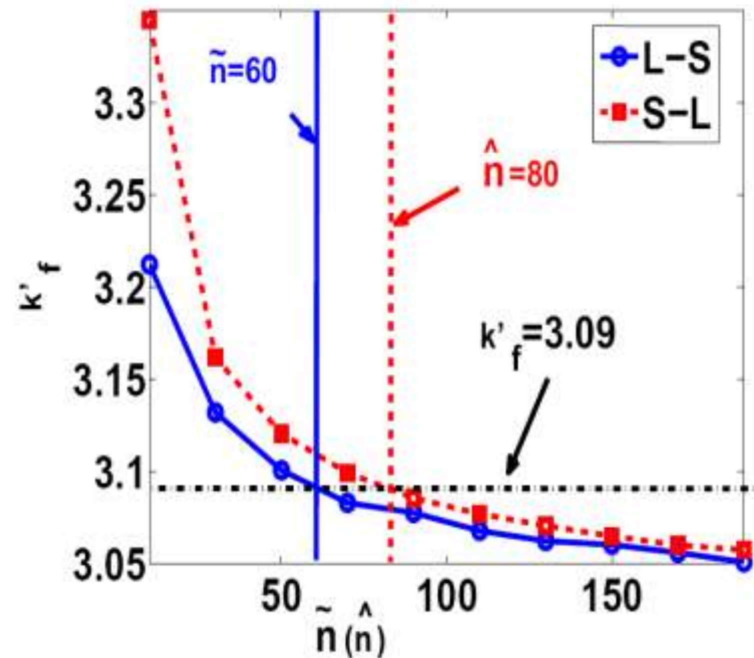
Are Variation Models Reliable ?

- 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



Cases	$\hat{n}$	$\tilde{n}$	Confidence interval	Reliability of Statistics analysis
L-L	Large	Large	Small	High
S-L	Small	Large	Large	Low
L-S	Large	Small	Large	Low
S-S	Small	Small	Large	Low

Comparison for S-S, L-S and S-L



90% confidence worst case fast corner for different $\hat{n}$ $\tilde{n}$

$$\widetilde{cn}_{f/s} = \tilde{\mu}_t \pm k_{f/s}\tilde{\sigma}_t$$

- 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 are usually 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

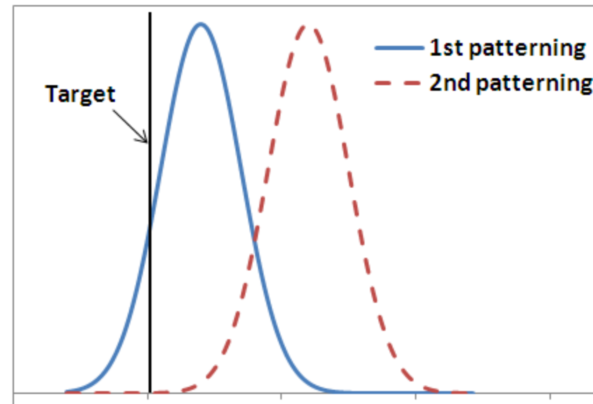
$conf_t$	$conf_f$	L_f	V_{tnf}	V_{tpf}	L_s	V_{tns}	V_{tps}
50	60	0.07	0.21	0.20	0.13	0.39	0.36
70	80	0.29	0.86	0.81	0.32	0.97	0.91
90	95	0.79	2.36	2.22	0.71	2.13	2.00
95	99	1.15	3.44	3.23	0.90	2.71	2.54

Guard band percentage of different variation sources $\hat{n} = 10, \tilde{n} = 15$

Outline

- Introduction
- Modeling Systematics: Litho Example
- Design-Flow Adoption Challenges
- Revisiting Random Spatial Variation Modeling
- Confidence in Variation Models
- Conclusions

What Lies Ahead: DPL



- Two Different exposure/etch steps → two CD populations
- Large CD/delay variability (e.g., 34% 3σ increase - by ASML study)

$$3\sigma^2_{pooled} = \frac{3\sigma^2_{p1}}{2} + \frac{3\sigma^2_{p2}}{2} + \left(\frac{3}{2}\left|\mu_{p1} - \mu_{p2}\right|\right)^2$$

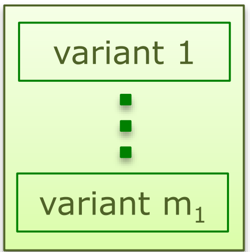
- Loss of spatial correlation between neighbors
- If used on poly, may require radically different modeling/characterization methods

Predicting Variability Trends

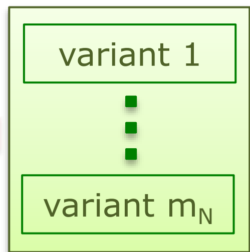
- Motivation:
  - Rapidly changing process and device technologies are a norm → Need to predict their variability impact at all layers (device, design, system).
- Key observation:
  - Silicon scaling is *evolutionary* →basic set of process steps (Litho, CVD, RIE, etc) do not change a lot → can leverage pre-characterized variation models of process "unit steps" to extrapolate variability of unknown devices.
- The key underlying model could be the "Level 1" or general variability model coupled with some description of systematic variations.

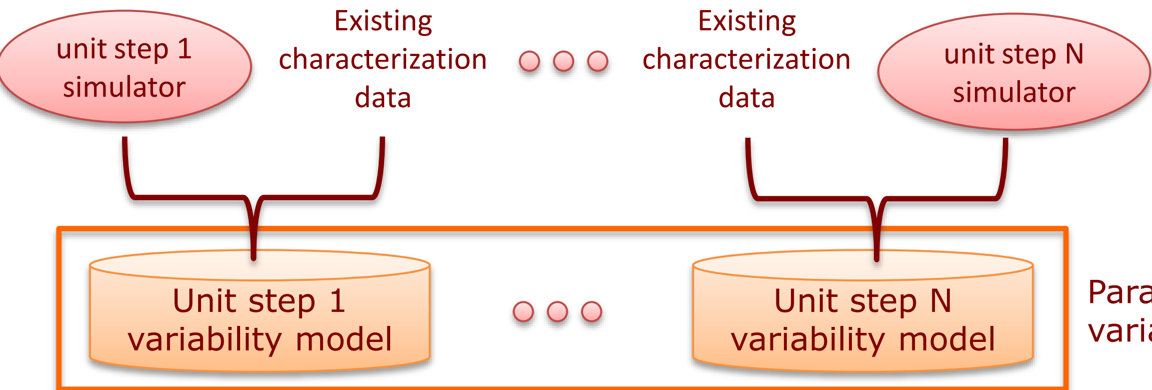
VirtualFab Variability Projection Framework

spanning variants for step *1* (XML)



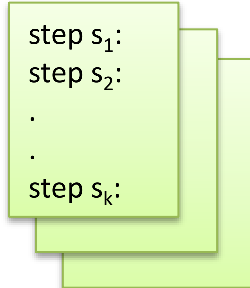
spanning variants for step *N* (XML)



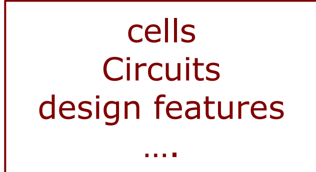


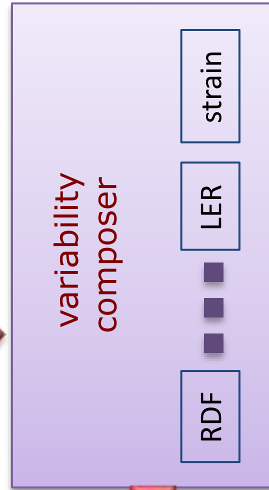
Parameterized unit step variability models (XML)

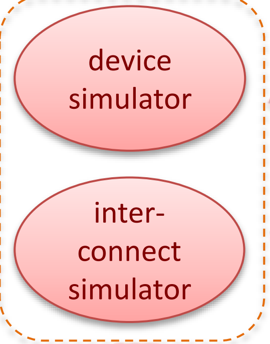
manufacturing steps (XML)

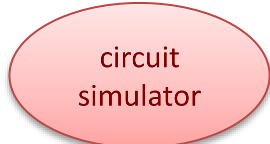


Component-level variability data









variability trends

Conclusions

- Leveraging systematic variation models requires tight integration SPICE modeling/extraction frameworks
- Variation models need to be *physically justifiable* AND *statistically reliable* AND *computationally tractable*
  - These need not be conflicting objectives. E.g., SAAW model is faster AND more accurate AND more physically justifiable than conventional spatial correlation models
  - True "DFM" models should have understanding of "M" beyond just the data
- Variability characterization should be done carefully
  - Enough samples for all sources
  - Low-volume parts should *expect* models to not be accurate

Acknowledgements

- Graduate students: Tuck-Boon Chan, Rani Ghaida, Lerong Cheng
- Collaborators: Costas Spanos (UCB), Andrew B. Kahng (UCSD), Sherief Reda (Brown)
- Industry help: Sani Nassif (IBM), Victor Moroz (Synopsys), Andres Torres (Mentor Graphics)
- Support: UC Discovery IMPACT (http://impact.berkeley.edu), SRC, NSF