Predictive Variability Modeling and Design Implications

Yu (Kevin) Cao

Nanoscale Integration and Modeling Group (NIMO), ASU

Predictive Variability Modeling

- Process Variations in Light of Scaling
- Intrinsic and Manufacturing Variations
- Future Modeling Needs and Promises

Predictive Variability Modeling

- **Process Variations in Light of Scaling**
- **Intrinsic and Manufacturing Variations**
- **Future Modeling Needs and Promises**

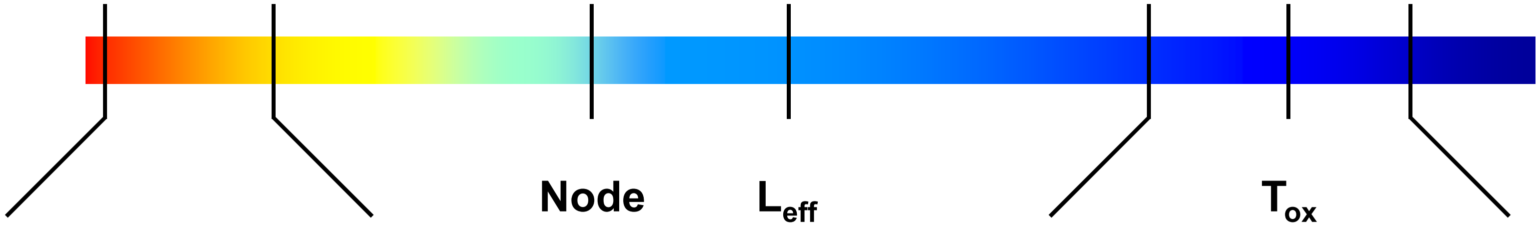
Approaching Physical Limits

Wavelength in lithography

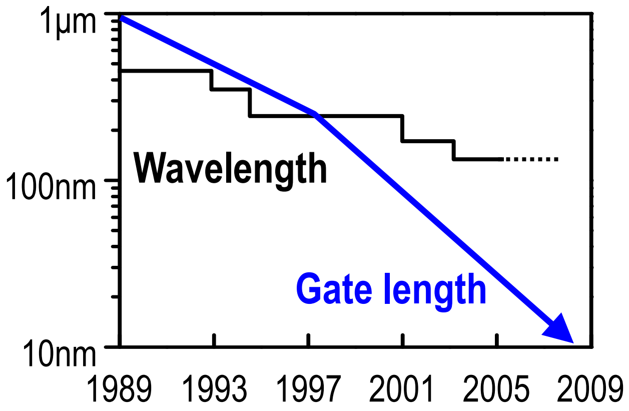
CMOS Technology

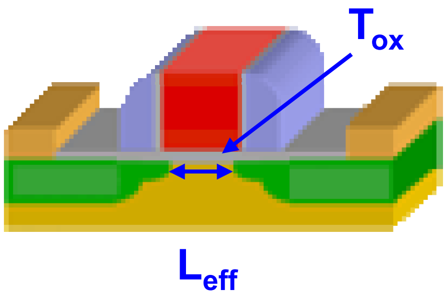
e$^-$ mean free path

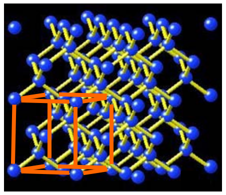
Silicon lattice



		Node	L_{eff}		T_{ox}	
365nm	157nm	32nm	12nm	2nm	1nm	0.54nm





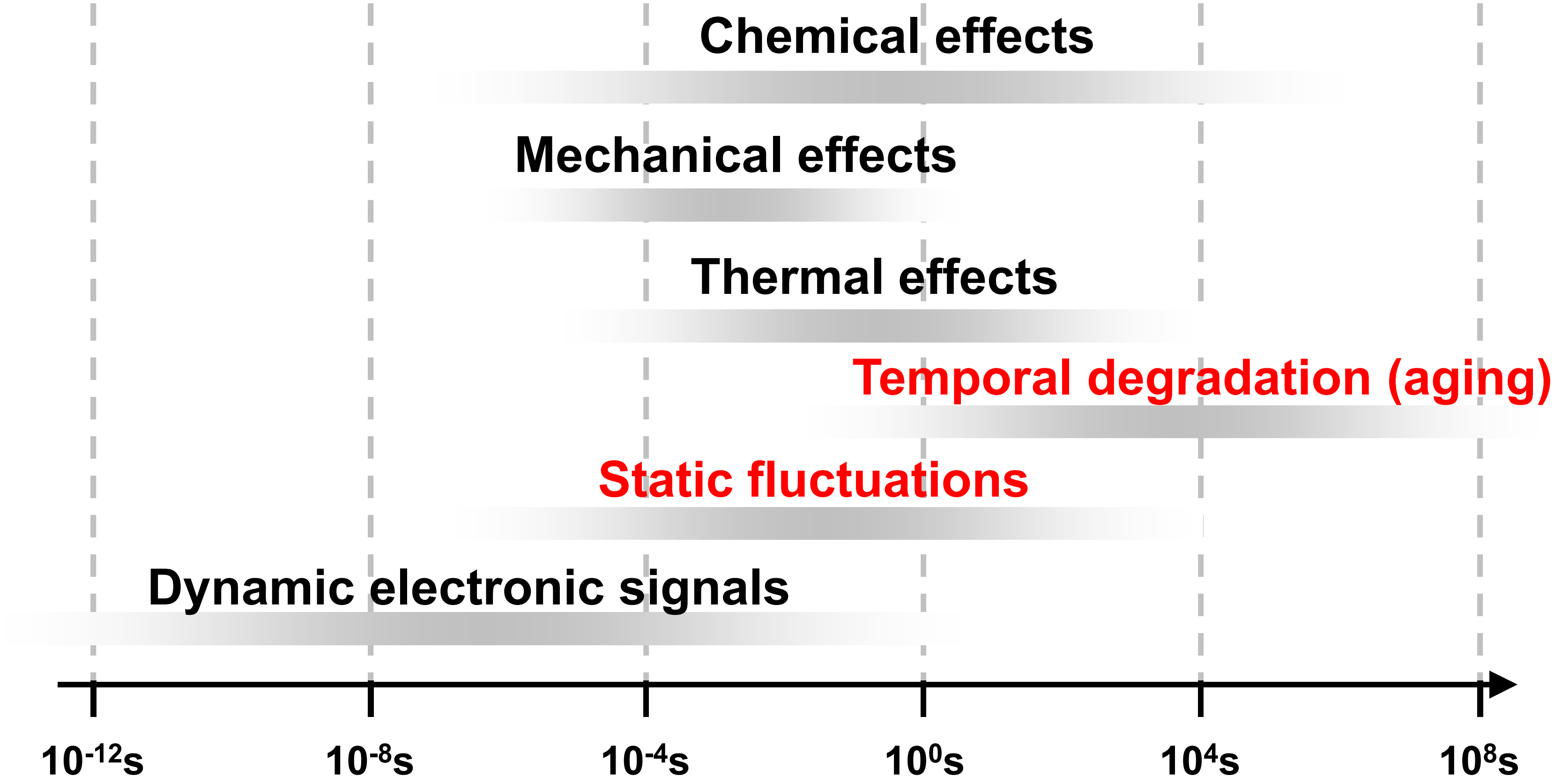


- Many secondary effects are now critical: leakage, variations, reliability, manufacturability, ...

[S. Thompson, U. Florida]

Increased Reliability Concerns

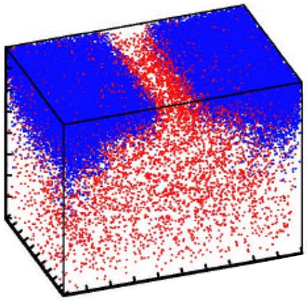
- An inevitable result of aggressive scaling
  - No convenient solution from CMOS technology!



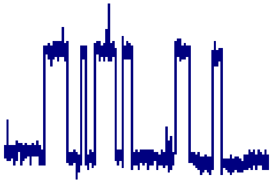
[M. Kole, BMAS 2007]

Intrinsic Variations

- Limited by fundamental physics; **random** in nature
- Representing the lower bound of variations

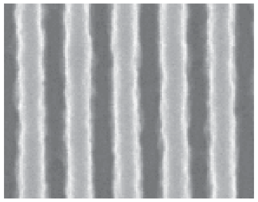


RDF



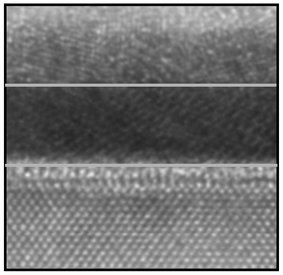
RTN

+



LER

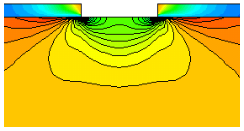
T_{ox}



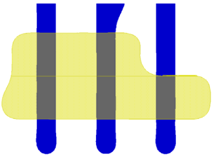
- RDF, RTN, LER, T_{ox} fluctuation, and their **interactions**!
- Approach: joint TCAD and compact modeling

Process Induced Variations

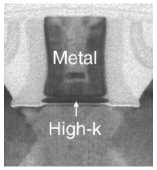
- Induced by the manufacturing process; “**systematic**”



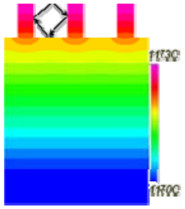
Stress



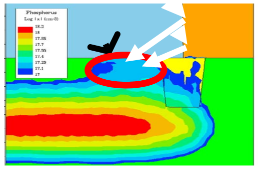
NRG



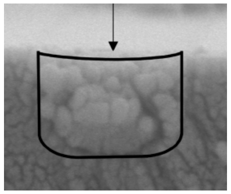
HK/MG



RTA



WPE

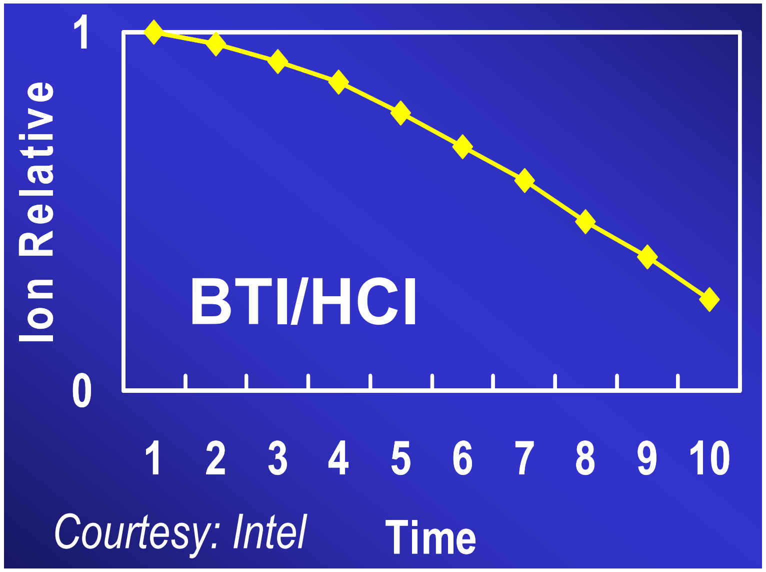


BEOL

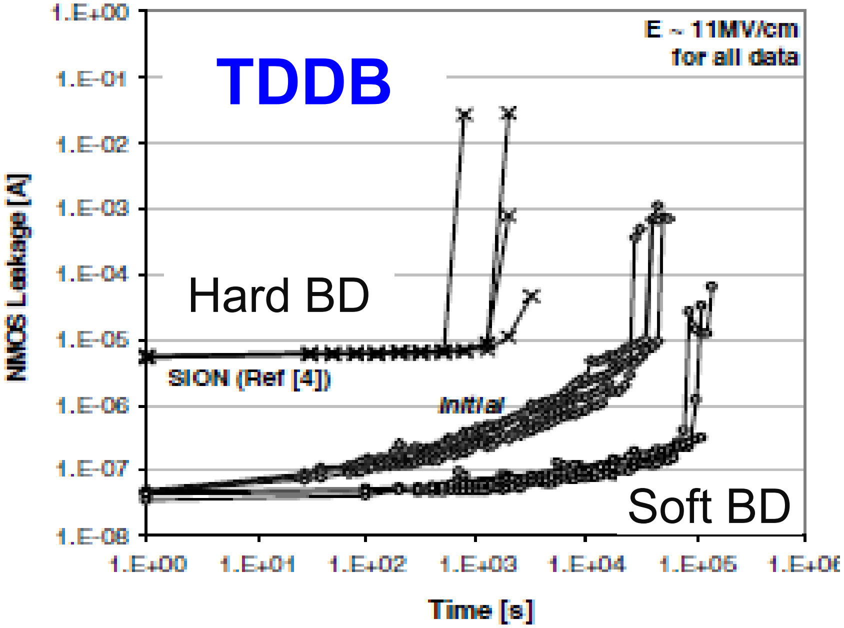
- Usually exhibit layout pattern dependence
- Approach: Compact modeling and in-situ characterization under various process and design conditions

Temporal Degradation (Aging)

- Stressed by circuit operation; “**systematic**”



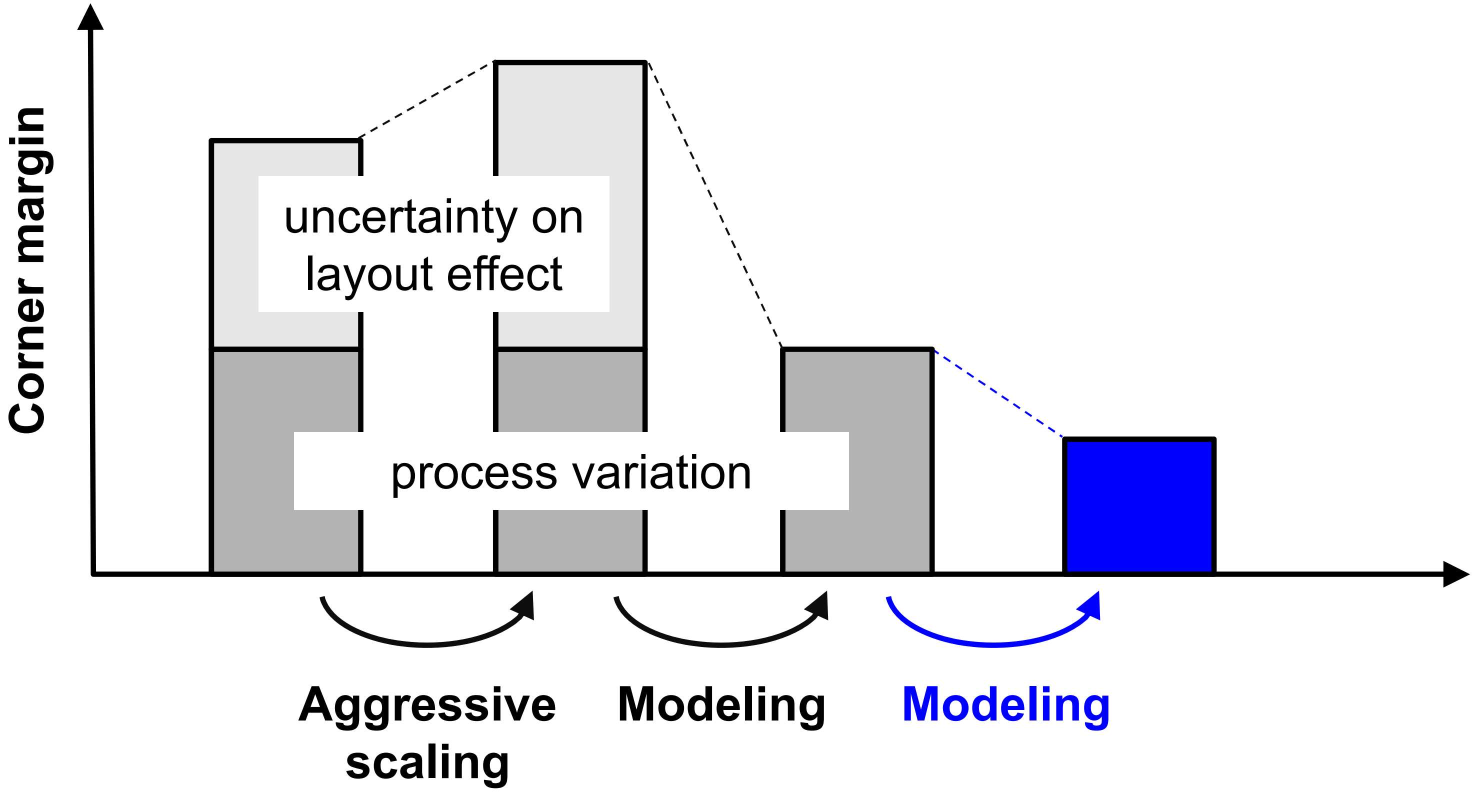
Courtesy: Intel



- Depends on technology and operation conditions
- Approach: Compact modeling and in-situ characterization at device and circuit levels

[J. Hicks, Intel 2008]

Compact Variability Modeling



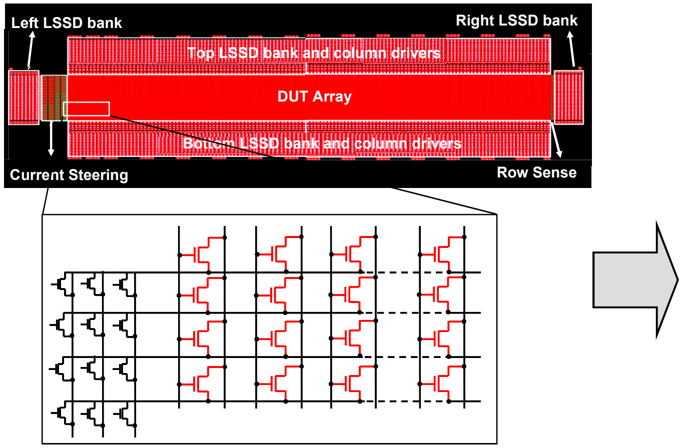
- Turns “random” effects into systematic
- Prepares for design analysis and optimization

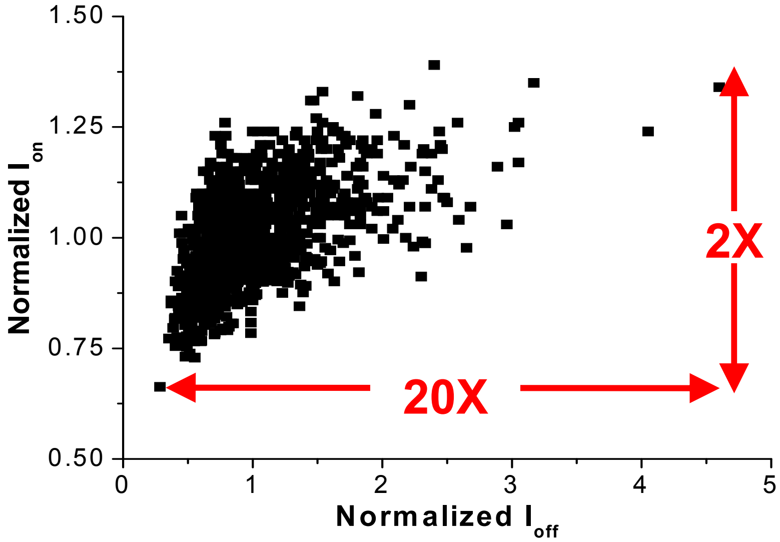
[H. Aikawa, VLSI 2008]

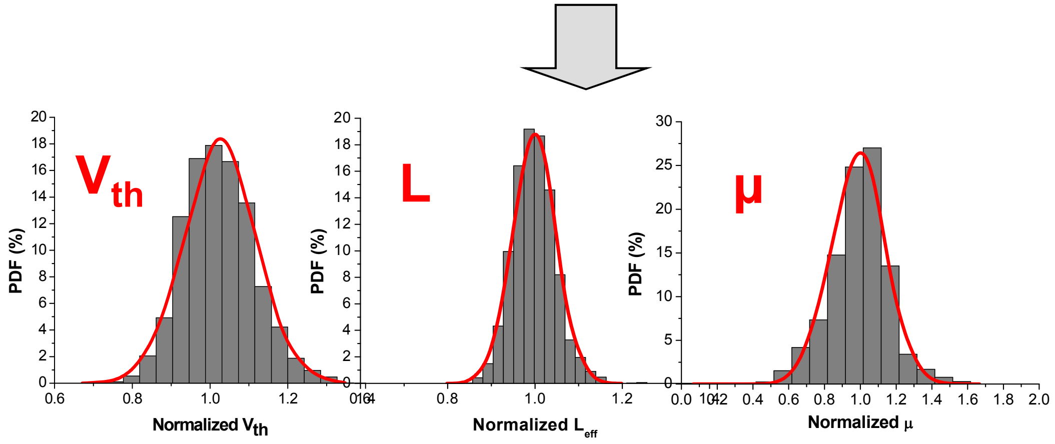
Predictive Variability Modeling

- Process Variations in Light of Scaling
- Intrinsic and Manufacturing Variations
  - Threshold voltage variation
  - Layout dependent effects
  - Temporal degradation
- Future Modeling Needs and Promises

Variation Extraction: Transistor


Left LSSD bank
Right LSSD bank
Top LSSD bank and column drivers
DUT Array
Bottom LSSD bank and column drivers
Current Steering
Row Sense

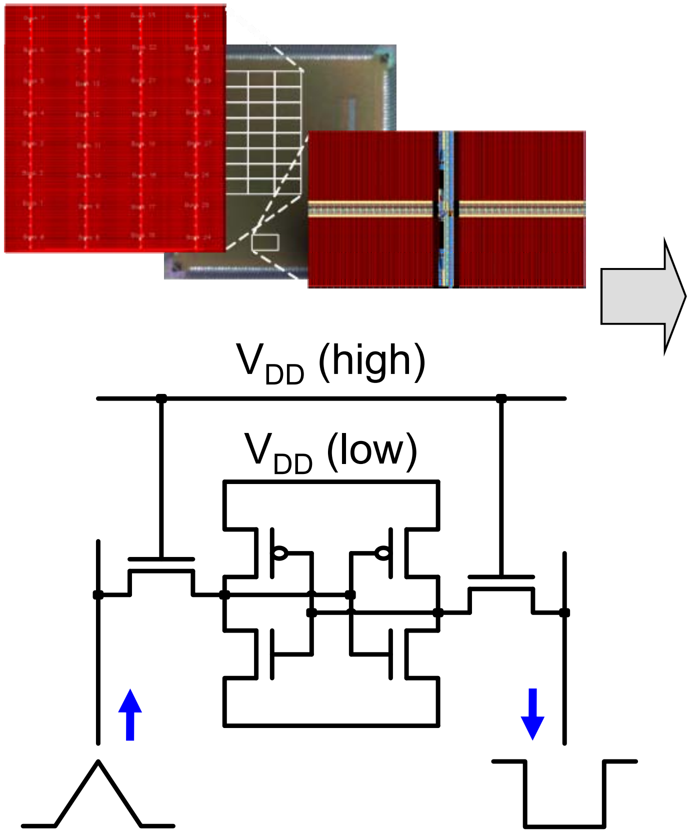


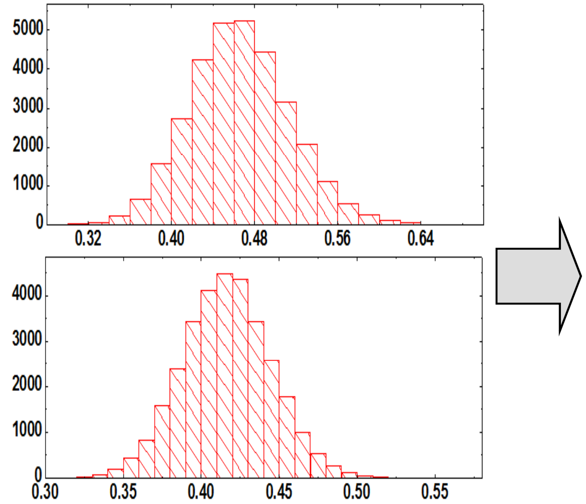


- Physical modeling helps accurate extraction and decomposition of leading variation sources

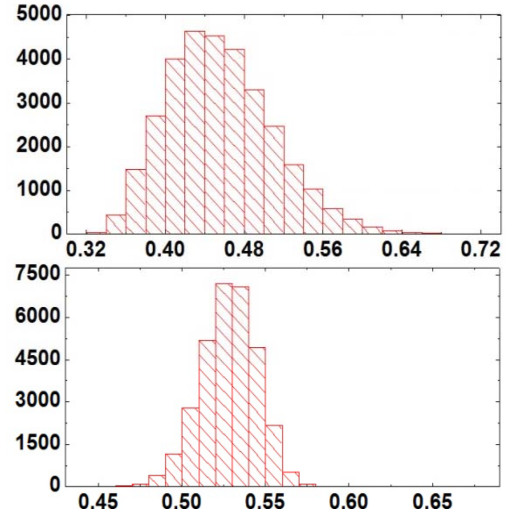
[W. Zhao, TSM 2009]

Variation Extraction: SRAM





Measured distribution of switching points

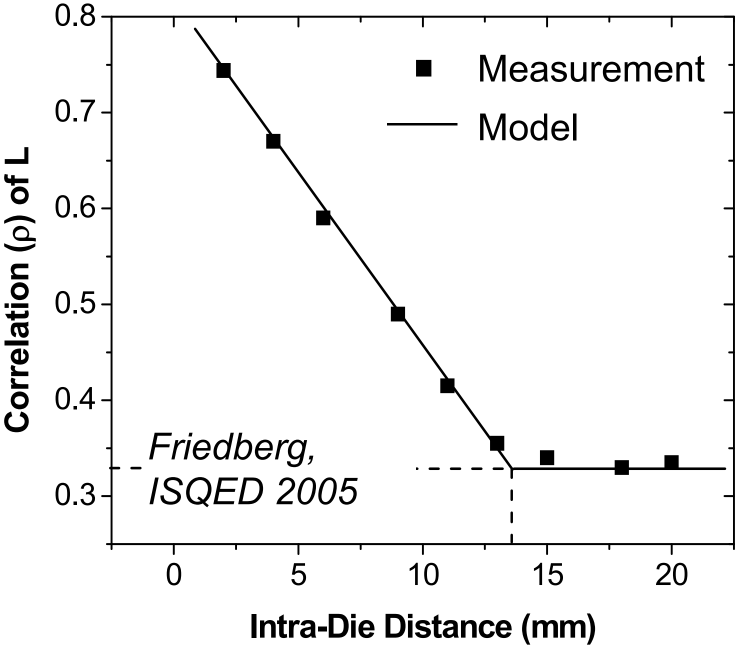


Decomposed V_{th} variability

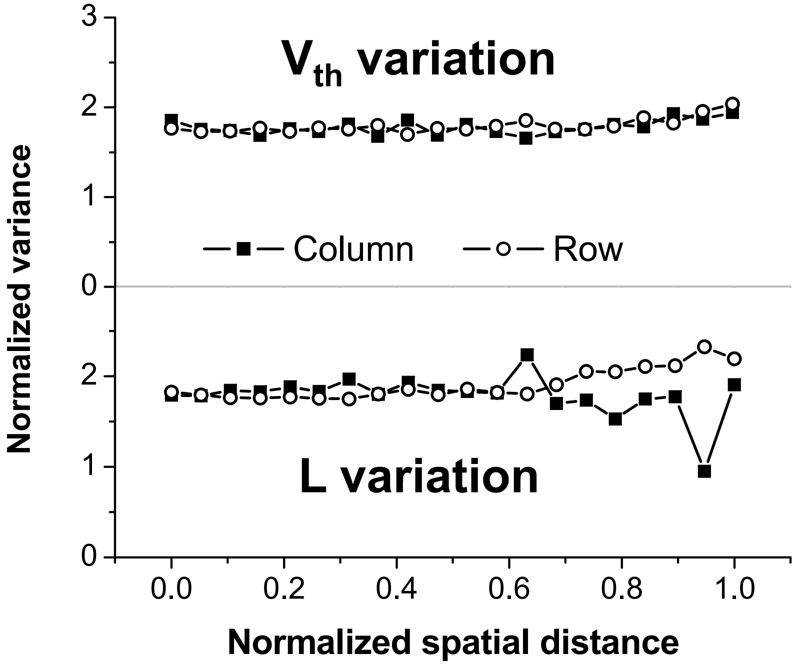
- Appropriate decomposition of as-fabricated SRAM variability helps shed light on joint process-design optimization

Spatial Correlation

130nm



65nm

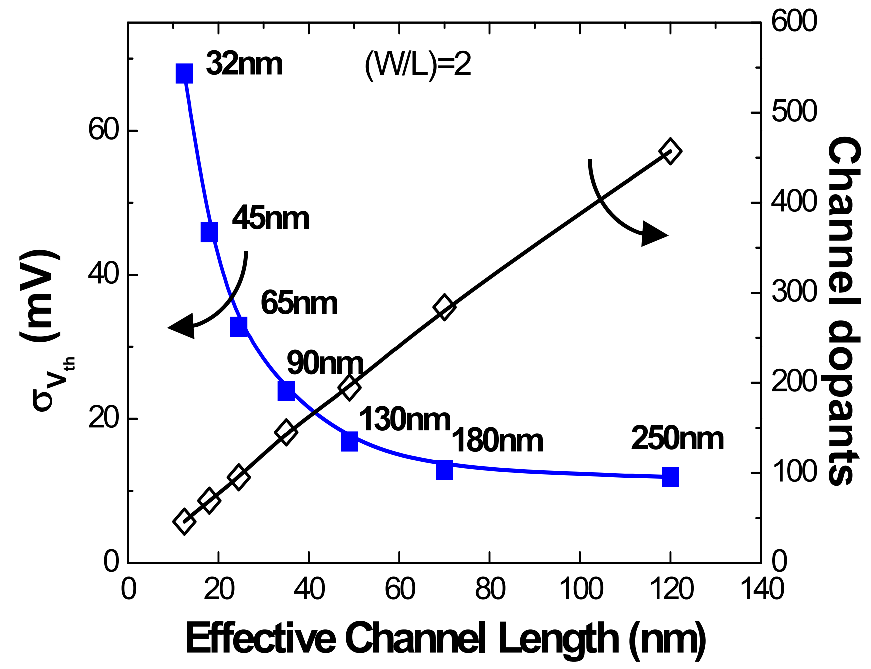


- Spatial correlation is negligible in both directions (1250μm X 110μm), which is different from previous generations
- Possible reason: regular layout; local random variation;

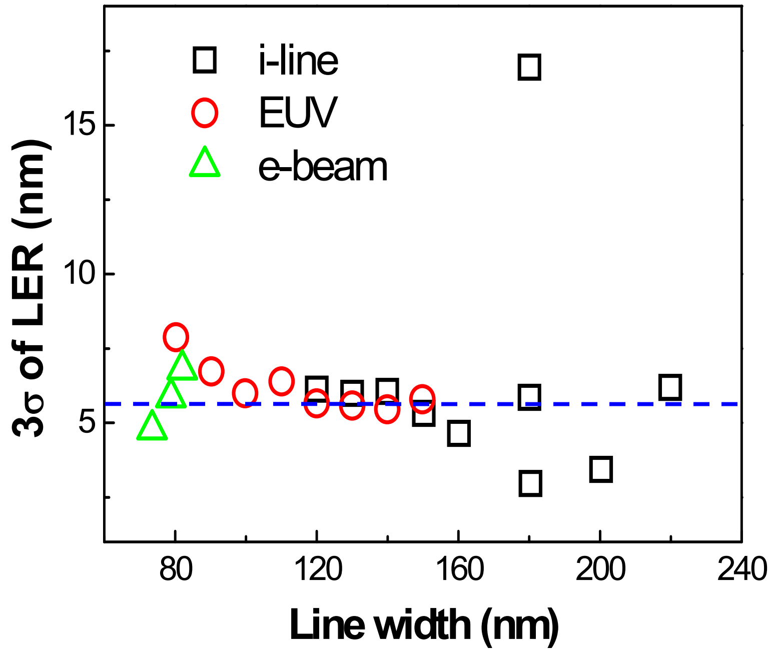
V_{th} Variation: RDF and LER

- **Length scale: nm**; random
- Trend: NOT scaling down with the feature size

RDF



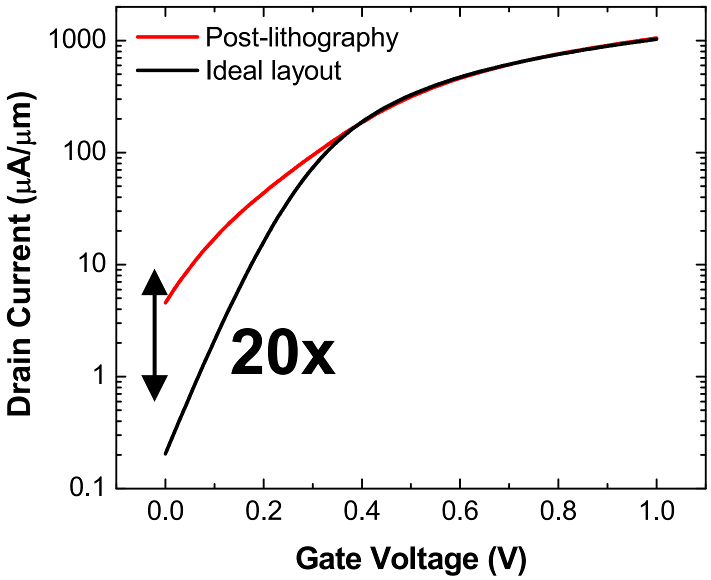
LER

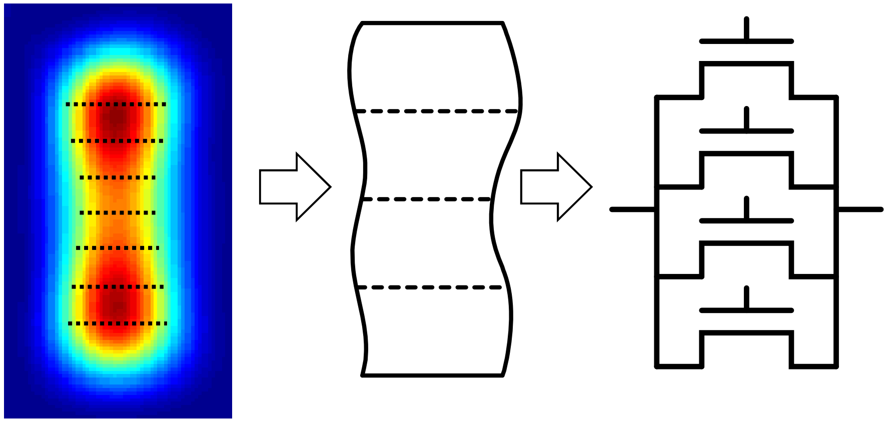


[PTM; J. A. Croon, IEDM 2002]

Gate Slicing Method

- Approach: A SPICE-compatible gate-slicing method
- Gate slice with shorter length dominates the variation and the leakage

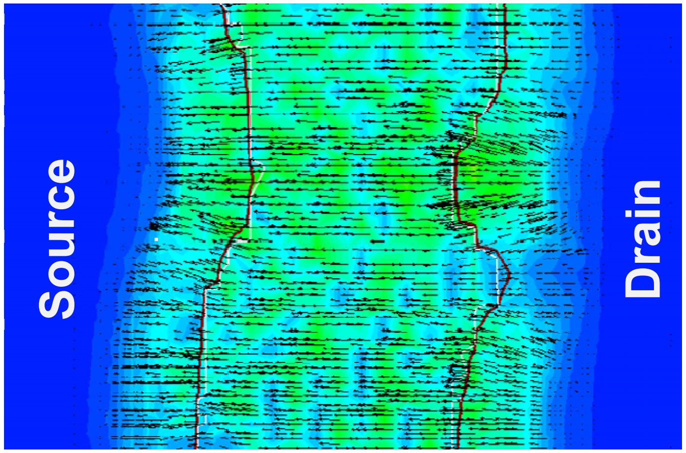


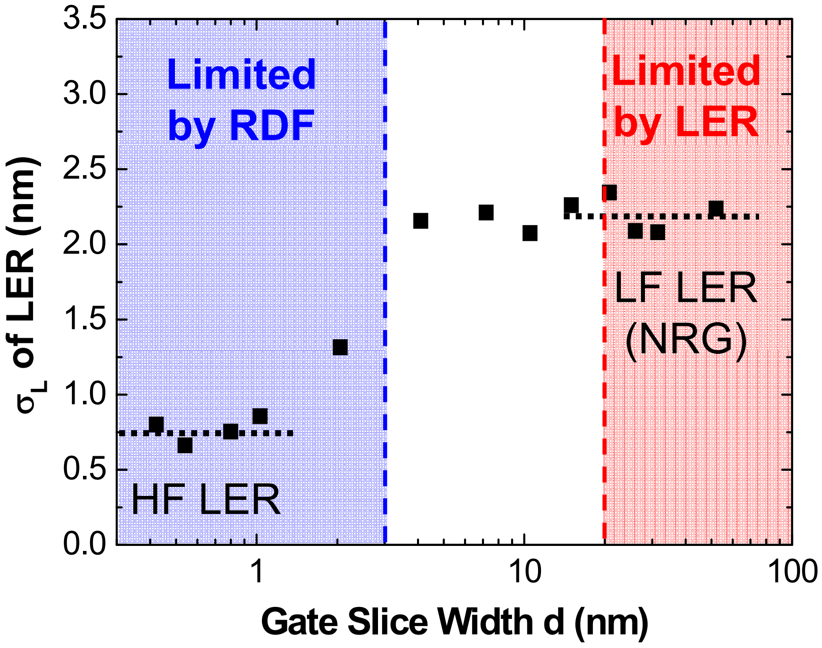


[Y. Ye, DAC 2008]

Limitations on the Slicing Method

- Current distribution
  - Fine for I_{on} if W >> L
- Slice width
  - ~nm
- Linearity
  - I_{ds} should be a linear function of V_{th}
  - Only I_{on} satisfies
  - I_{off} is not suitable





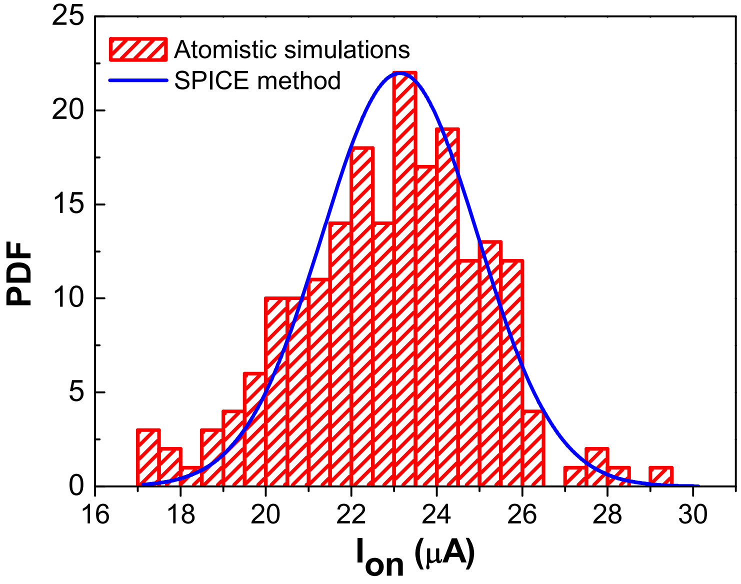
Modeling and Simulation Procedure

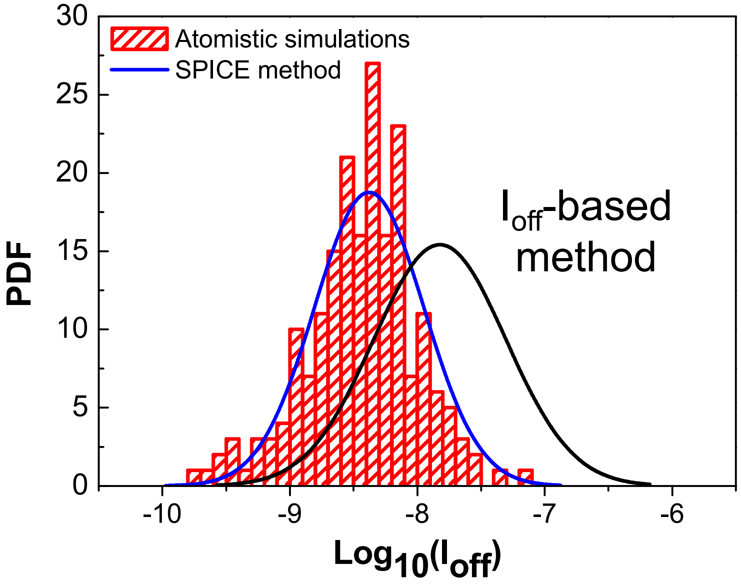
- Starting point: A non-rectangular gate shape with σ_L due to LER and σ_{Vth} due to RDF

1. Gate slicing at appropriate slice width
2. Assignment of random V_{th} to each slice depending on its W, L, and σ_{Vth}
3. Sum the current together from each slice, then extract V_{th} variation from $\boldsymbol{I_{on}}$
4. Compute equivalent gate length for nominal I-V under non-rectangular gate (NRG)

→ Finish a statistical transistor model under RDF, LER and NRG

Validation with Atomistic Simulations





- A roughly 65nm technology
- I_{on}-based simulation method accurately predicts the variability of both I_{on} and I_{off} under RDF
  - I_{off}-based extraction mis-predicts the distribution

[A. Asenov, TED 2003]

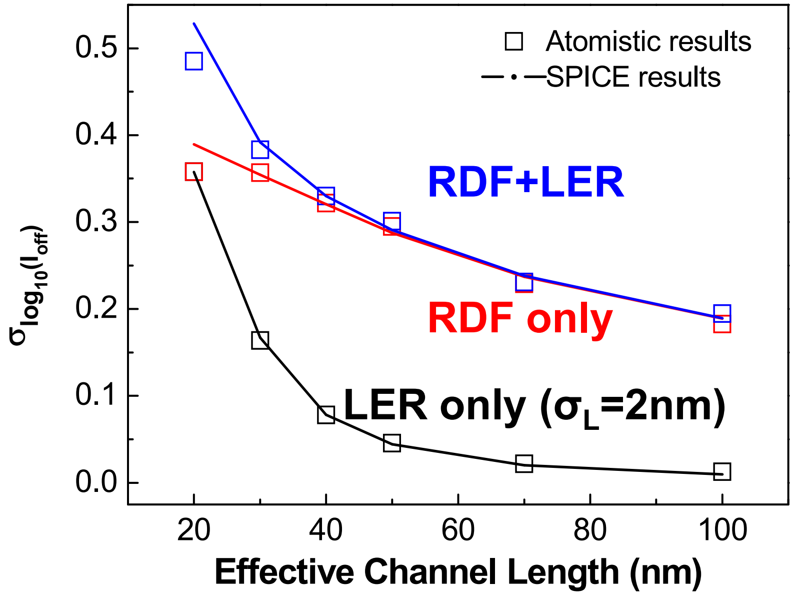
Predictive Modeling

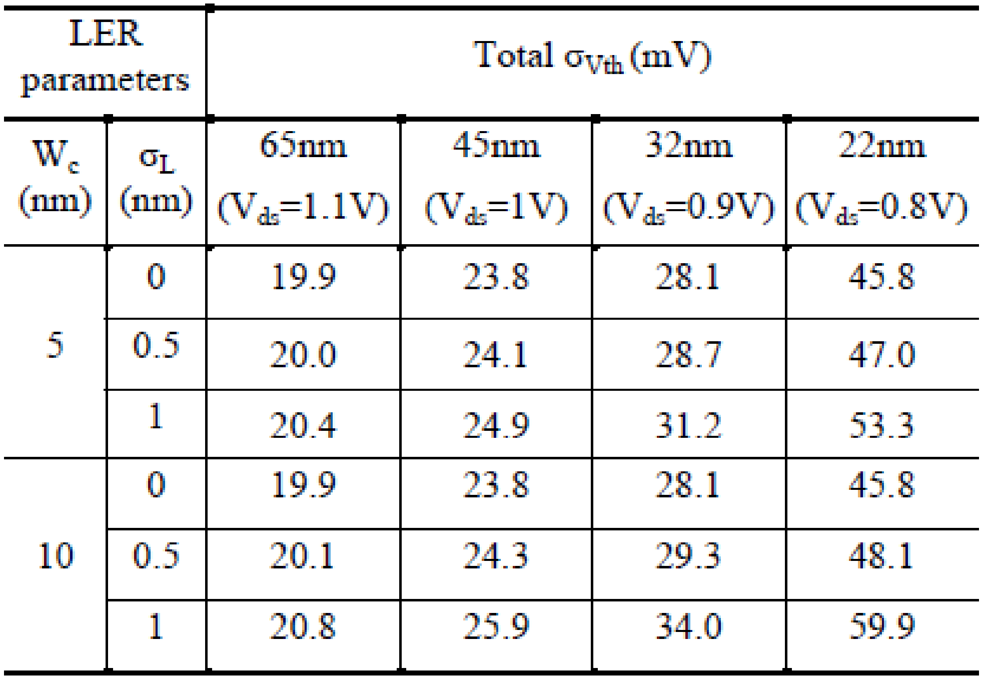
$$\Delta V_{th} = \Delta V_{th0} + V_{ds}\exp\left(-\frac{L}{l'}\right)\cdot\frac{\Delta L}{l'}$$

$$\sigma_{total}{}^{2} = \sigma_{RDF}{}^{2} + \sigma_{NRG}{}^{2}$$

⇒

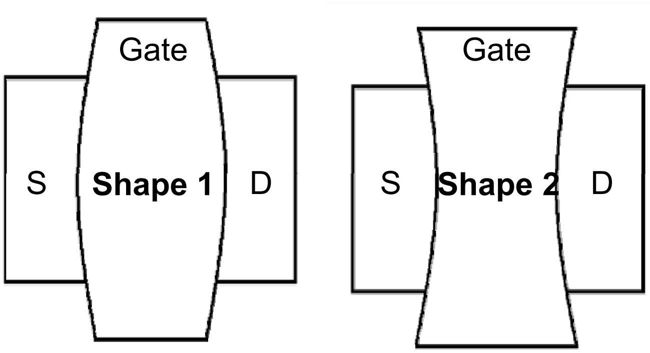
$$\boxed{\sigma_{total}{}^{2} = \frac{C_1}{WL} + \frac{C_2 V_{dd}{}^{2}}{\exp(2L/l')}\cdot\frac{W_c}{W}\cdot\sigma_L{}^{2}}$$

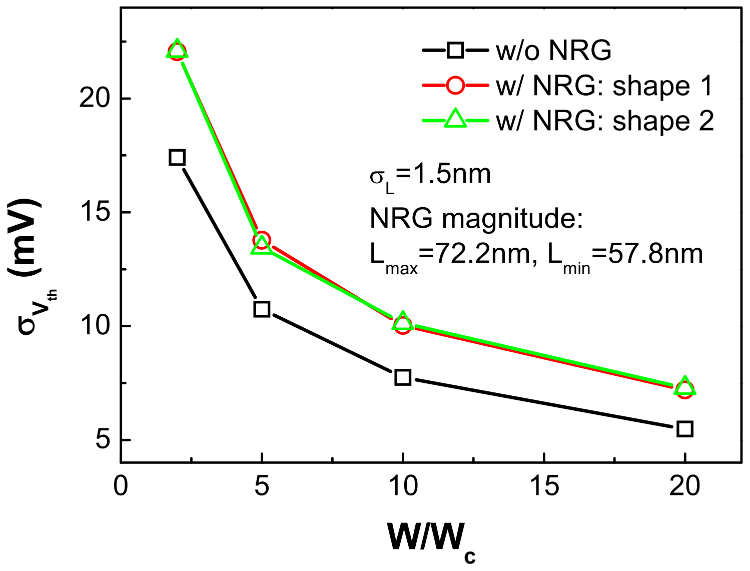




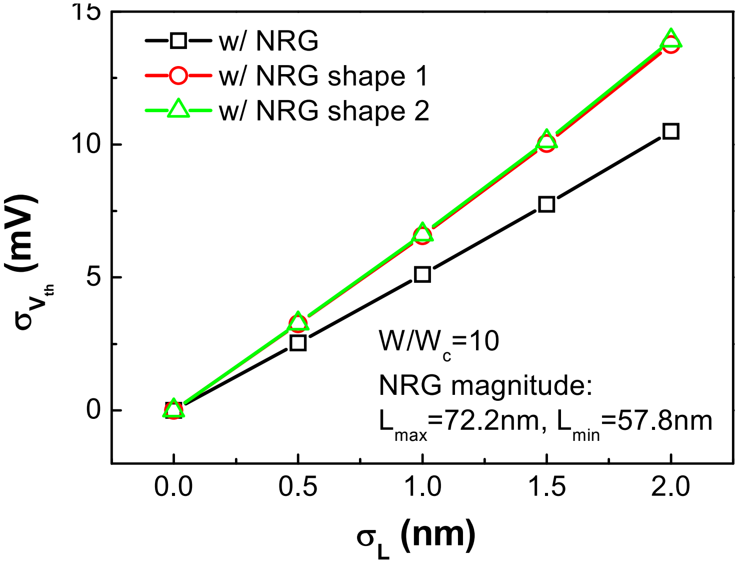
LER parameters		Total σ_{Vth} (mV)			
W_c (nm)	σ_L (nm)	65nm (V_{ds}=1.1V)	45nm (V_{ds}=1V)	32nm (V_{ds}=0.9V)	22nm (V_{ds}=0.8V)
5	0	19.9	23.8	28.1	45.8
	0.5	20.0	24.1	28.7	47.0
	1	20.4	24.9	31.2	53.3
10	0	19.9	23.8	28.1	45.8
	0.5	20.1	24.3	29.3	48.1
	1	20.8	25.9	34.0	59.9

Interaction with NRG

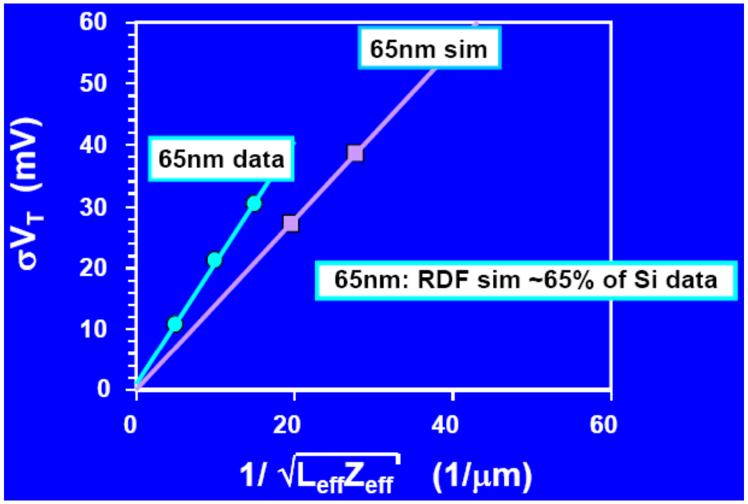


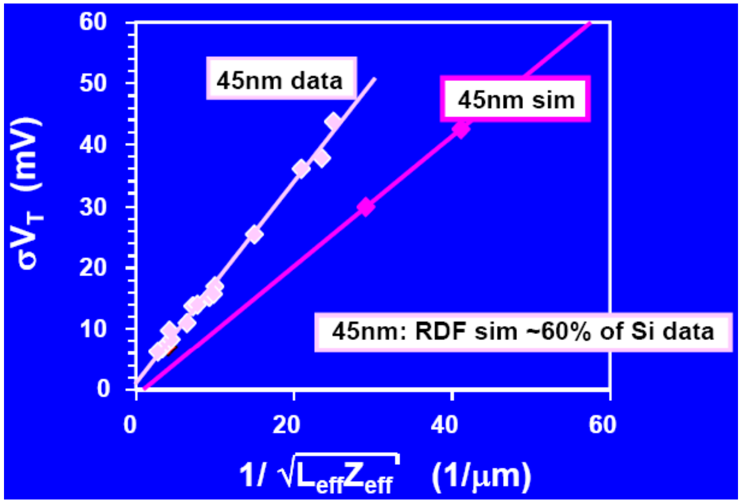


- NRG impacts nominal I-V and V_{th} variance
- But narrow-width effect only affects the nominal behavior, not the variability
- Question: is rectangular gate the optimal shape? How to simulate?



Remaining Questions



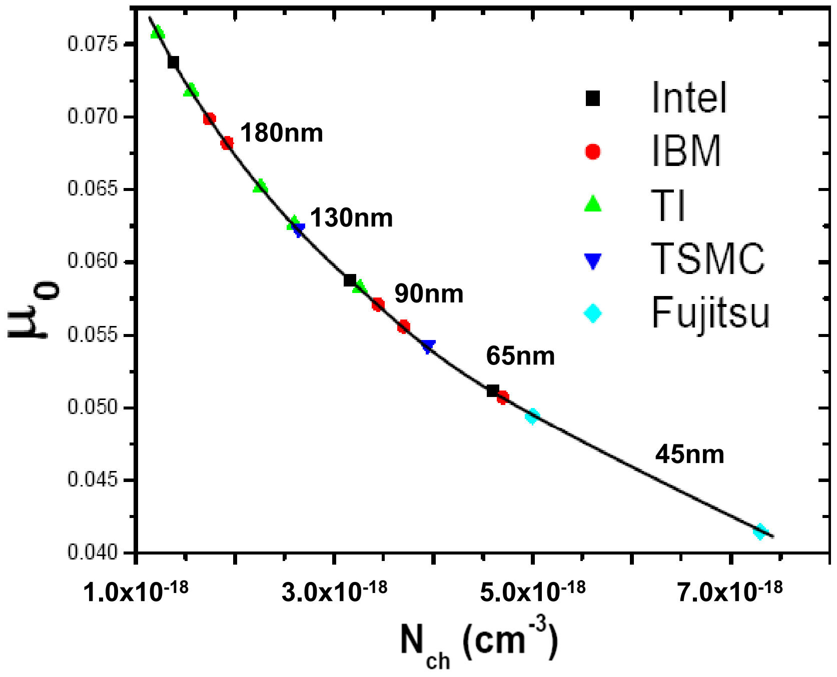


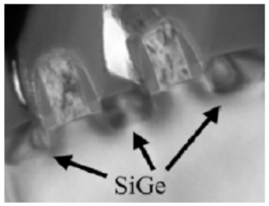
- The dependence on device area maintains during the scaling
- But the slope is larger than RDF only model
- Ongoing: an integral atomistic simulation and modeling for RDF+RTN+LER+T_{ox} for technology-design optimization

[K. J. Kuhn, IEDM 2007]

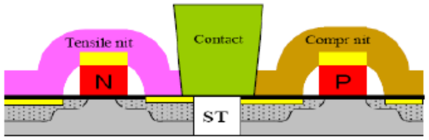
Strain Technology

- Trend: strain is essential for scaled CMOS technology
  - Higher channel doping concentration to define V_{th}
  - But results in carrier mobility degradation

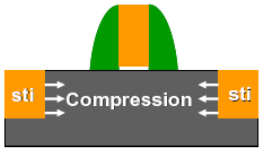




Embedded SiGe (eSiGe)



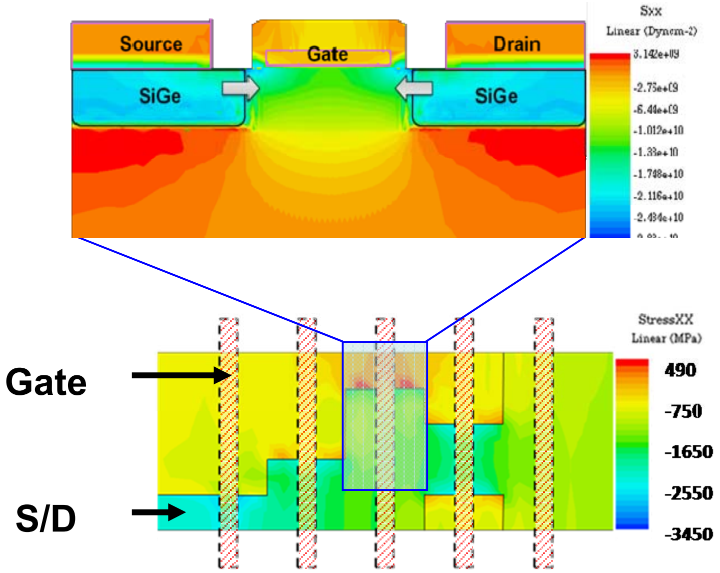
Dual Stress Liner (DSL)

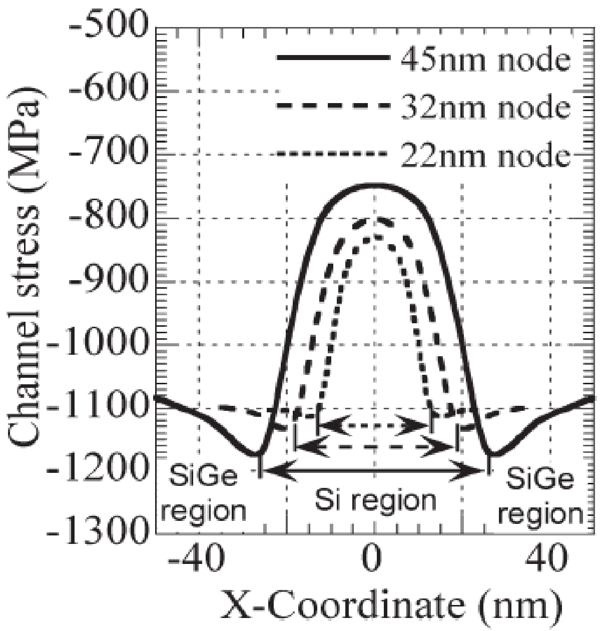


Parasitic Stress STI

Stress Induced Variation

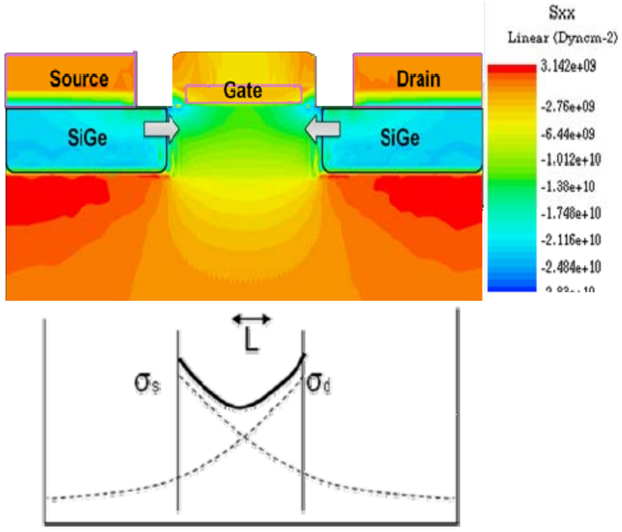
- **Length scale: ~100nm**; layout dependent
- Approach:
  - Physical modeling of layout dependence
  - Systematic layout decomposition for efficient extraction

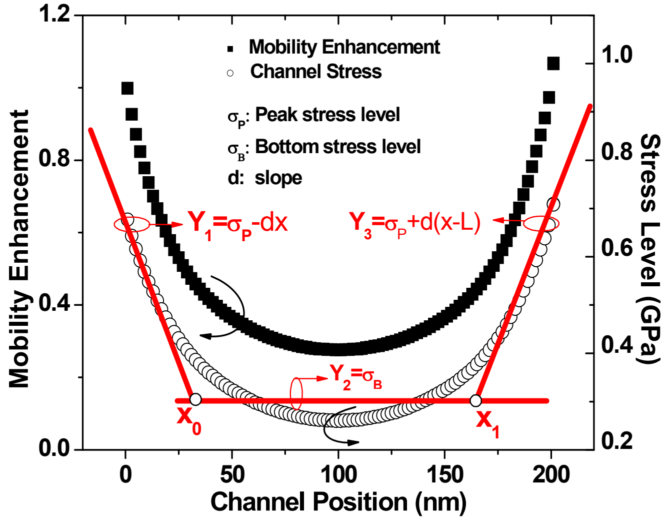




[G. Eneman, TED 2006]

Stress Distribution





$$Y_1 = \sigma_P - dx$$

$$Y_2 = \sigma_B$$

$$Y_3 = \sigma_P + d(x - L)$$

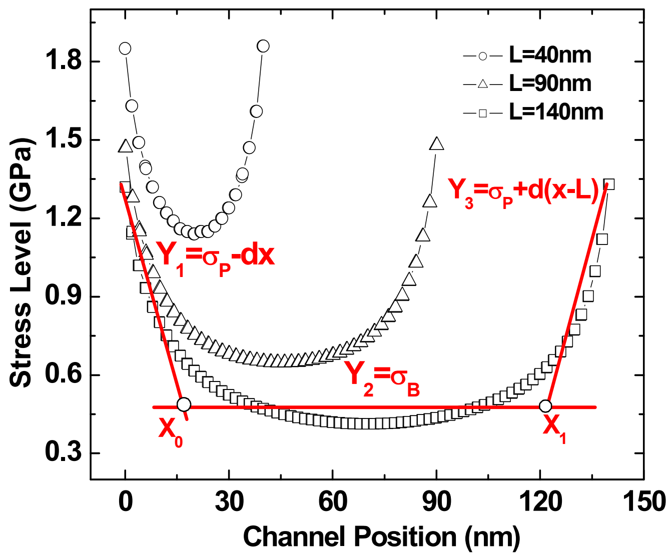
σ_P :Peak stress level

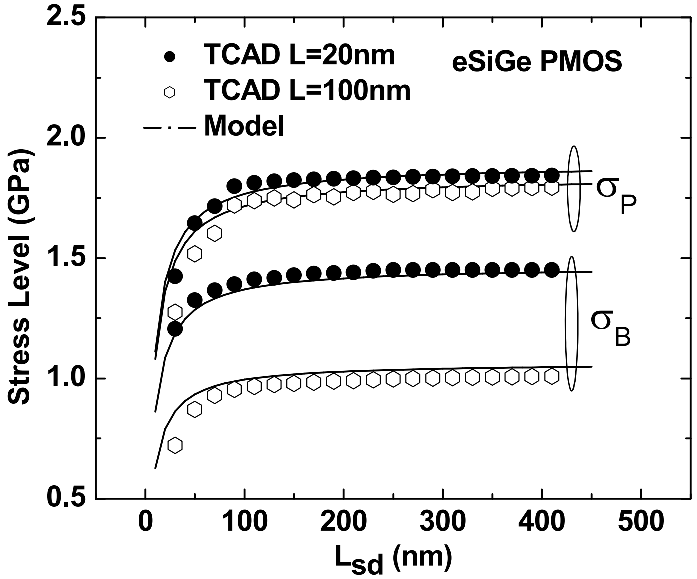
σ_B :Bottom stress level

d: slope

- The stress is induced from both ends of the channel in eSiGe
- **Piecewise linear approximation** for the non-uniform stress distribution in the channel

Layout Dependence





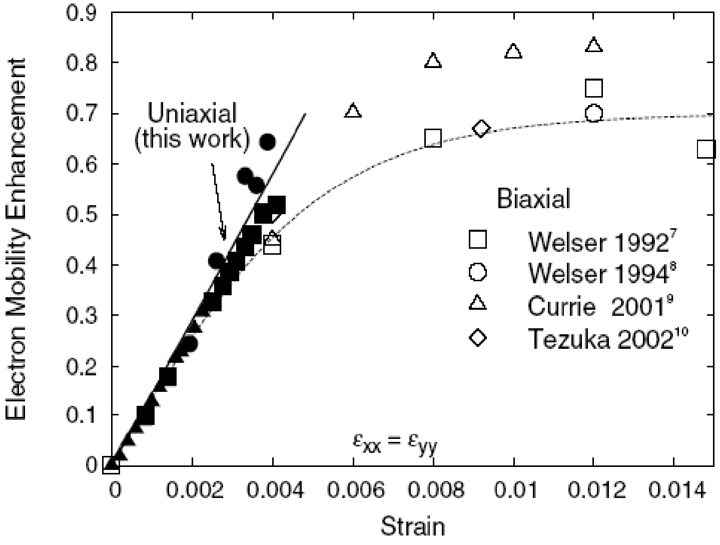
$$\sigma_P = \left(\underbrace{1+\frac{1}{L}}_{\text{Self S/D}} + \underbrace{\frac{1}{L+L_{sd}}}_{\text{1st left neighbor}} + \underbrace{\frac{1}{2L+L_{sd}}}_{\text{1st right neighbor}} \right) \cdot \underbrace{\frac{L_{sd}}{A+L_{sd}} \cdot \sigma_m}_{\text{Stress level in S/D}}$$

$$\sigma_B = \frac{C}{C+L} \cdot \sigma_P$$

- The layout dependence is captured by the **peak and bottom stress levels** in the piecewise-linear stress distribution

[C.-C. Wang, SISPAD 2009]

Mobility Enhancement



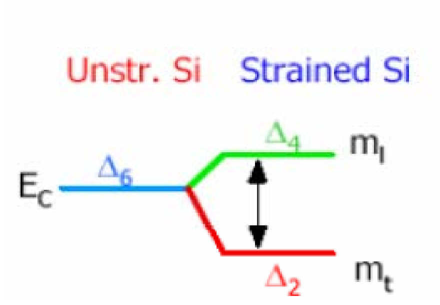
$$\mu_{ii}^{n} = \mu_{n}^{0}\left[1 + \frac{1 - m_{n1}/m_{nt}}{1 + 2(m_{n1}/m_{nt})}\left(\exp\left(\frac{\Delta E_{\mathrm{C}} - \Delta E_{\mathrm{C},i}}{kT}\right) - 1\right)\right]$$

$$\mu^{p} = \mu_{p}^{0}\left[1 + \left(\frac{\mu_{p1}^{0}}{\mu_{p}^{0}} - 1\right)\frac{(m_{p1}/m_{p\mathrm{h}})^{1.5}}{1 + (m_{p1}/m_{p\mathrm{h}})^{1.5}}\left(\exp\left(\frac{\Delta E_{\mathrm{V},1} - \Delta E_{\mathrm{V,h}}}{kT}\right) - 1\right)\right]$$

$$\boxed{\frac{\mu}{\mu_0} = 1 + B \cdot \left[\exp\left(\frac{\Delta E}{kT}\right) - 1\right]}$$

B= -0.3588 for electrons
B= 0.2815 for holes

(Band splitting based model)



$$\boxed{\Delta E = P \cdot \sigma}$$

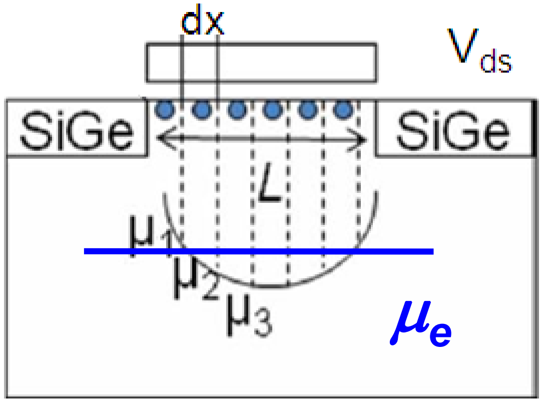
P= 3.9456E-2 (eV/GPa) for
uniaxial stress

P: Energy splitting per GPa

(Deformation Potential Theory
for PMOS device)

[S. E. Thompson, IEDM ’06; J.-S. Lim, EDL ‘04]

Equivalent Channel Mobility



$$I = \frac{Q}{T} = \frac{Q}{\int dt}$$

General expression

$$Q = WLC_{ox}\left(V_{gs} - V_{th}\right)$$

Charge

$$dt = \frac{dx}{v(x)} = \frac{dx}{\mu(x)E(x)}$$

Time to cross dx

$$E(x) = \frac{V_{ds}}{L}$$

Lateral E-Field

$$I = \frac{WC_{ox}\left(V_{gs} - V_{th}\right)V_{ds}}{\int \frac{1}{\mu(x)}dx}$$

Current expression

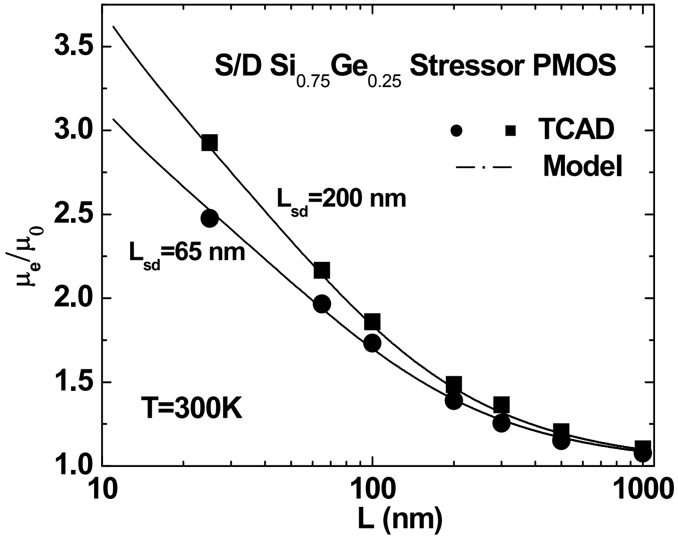
$$\Rightarrow \boxed{\frac{1}{\mu_e} = \frac{1}{L}\int \frac{1}{\mu(x)}dx}$$

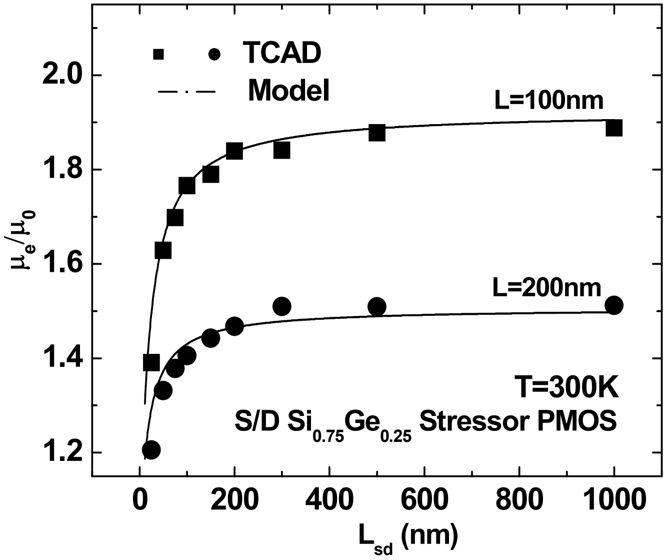
Equivalent mobility

Similar to the form of Mathiessen's rule

[F. Payet, TED '08]

Layout Dependent Mobility

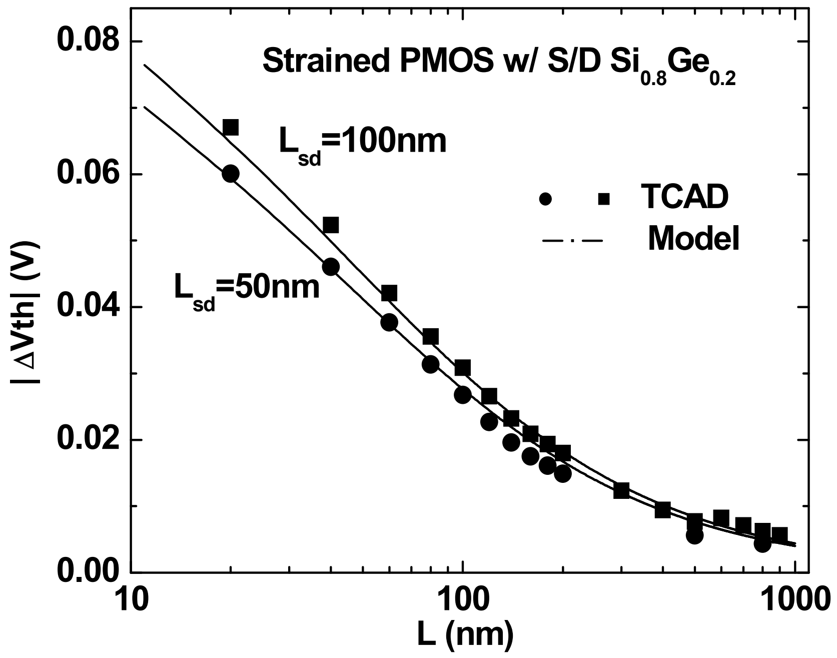


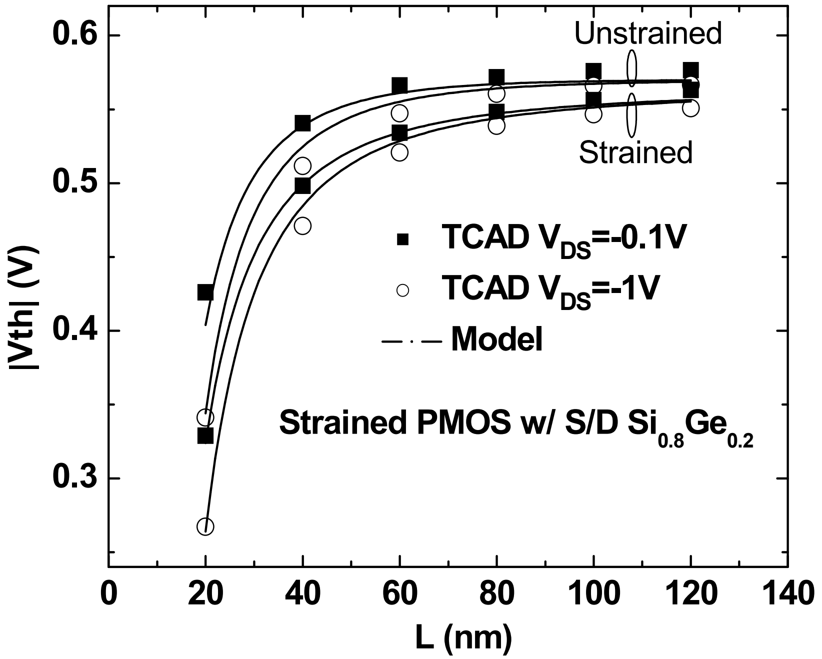


$$\frac{\mu_0}{\mu_e}=\frac{2kT}{dPL(B-1)}\cdot\left\{\frac{-dPx_0}{kT}+\ln\left[\frac{1+B\left(\exp\left(\frac{P\sigma_1}{kT}\right)-1\right)}{1+B\left(\exp\left(\frac{P\sigma_1-dPx_0}{kT}\right)-1\right)}\right]\right\}+\frac{L-2x_0}{L\cdot\left[1-B+B\exp\left(\frac{P\sigma_2}{kT}\right)\right]}$$

- Δμ increases as L goes down because of **higher stress**
- Δμ increases as L_{sd} goes up because of **more S/D stressor**

Threshold Voltage Reduction





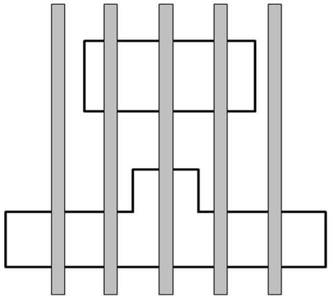
$$\Delta V_{th}(\sigma_B) = VTH\_STR \cdot \sigma_B$$

$$l \propto \left(E_g\right)^{\frac{1}{4}} \propto \left(\sigma\right)^{\frac{1}{4}}$$

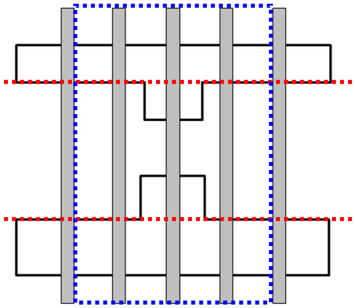
- V_{th} shift becomes larger at shorter channel length
- DIBL and sub-V_{th} swing is relatively insensitive to the stress effect

Impact on Gate Delay

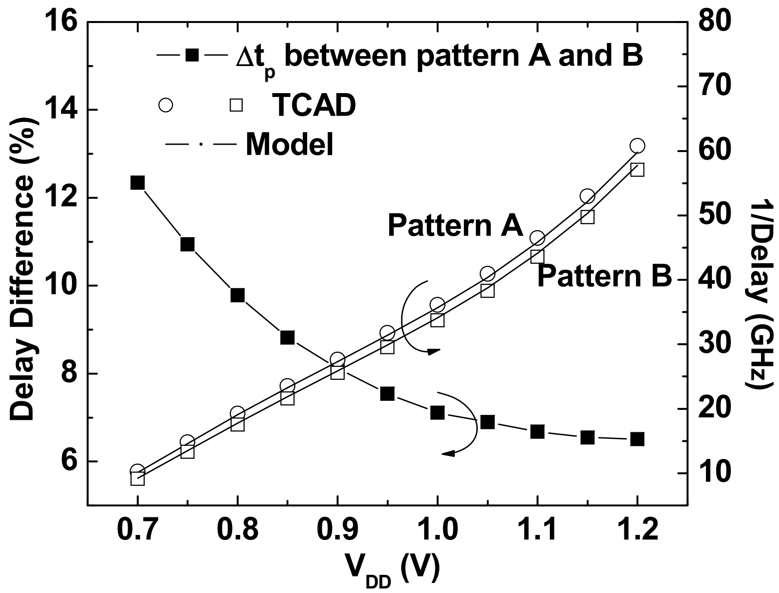
Pattern A:
Weaker Pull-up



Pattern B:
Stronger Pull-up



High to low transition

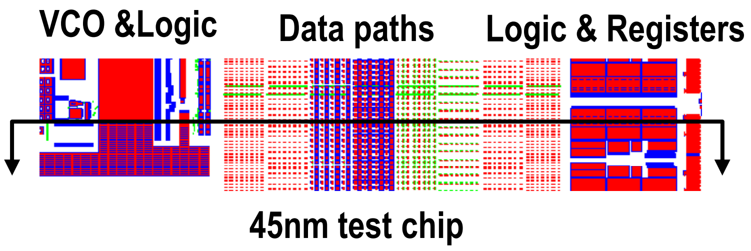


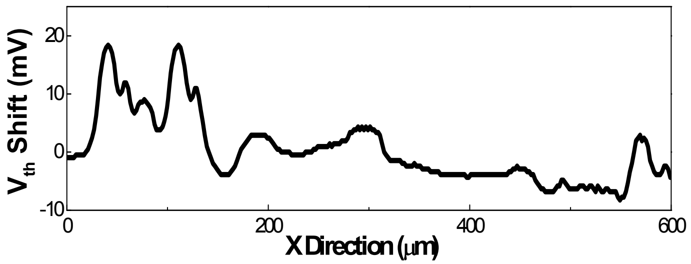
*This example only considers eSiGe, without STI

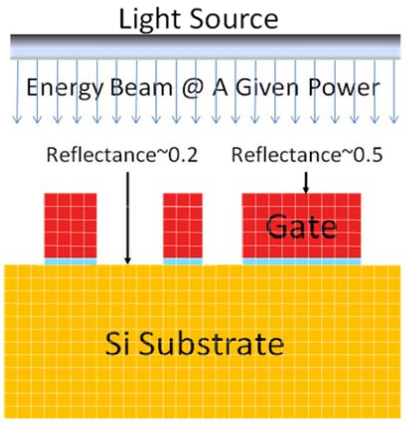
- The full layout is decomposed into basic patterns
- Delay variation due to the stress effect is pronounced
- Ongoing: systematic calibration with Silicon data

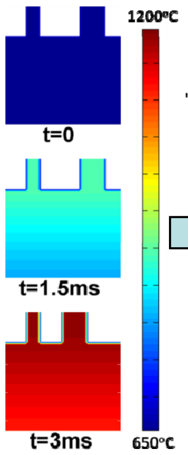
Rapid Thermal Annealing

- **Length scale: ~mm**; layout pattern density dependent
- Approach: Joint TCAD-compact modeling efforts

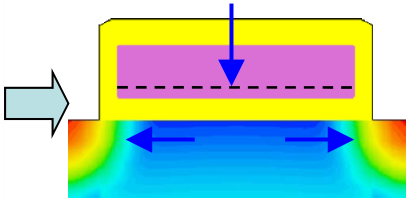




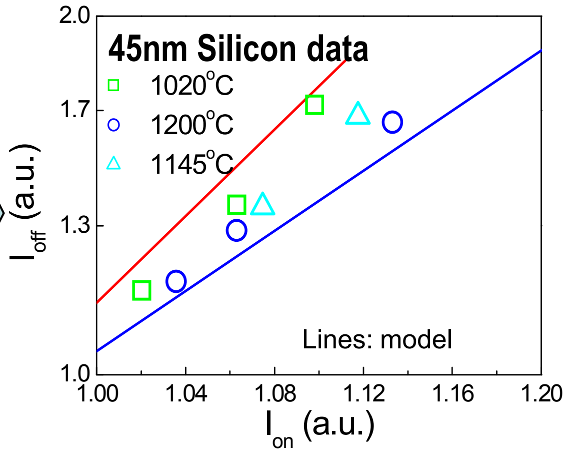




Dopant Activation and Poly-Depletion



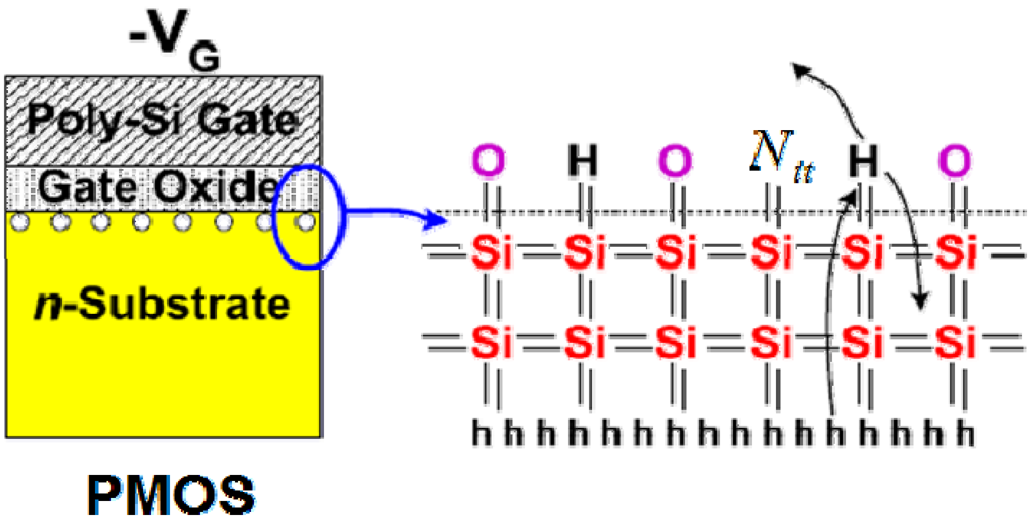
Lateral S/D Diffusion



[Y. Ye, DAC 2009]

Temporal Shift: NBTI

- **Time scale: hours to years**, depending on PVT & activity
- Two steps: Reaction–Diffusion
  - Other possible mechanisms involve fast interface traps
- Two phases: stress and **recovery**
- Approach: TCAD, modeling, and silicon characterization



$$\Delta V_{th} = \frac{qN_{it}}{C_{ox}}$$

[D. K. Schroder, MR 2007]

Time and Technology Dependence

- Power-law dependence on time (t) – **diffusion**

$$N_{it} = Kt^{n} + \delta \qquad \Delta V_{th} = \frac{qN_{it}}{C_{ox}}$$

n is 0.1~0.3 (e.g., 0.16 for H_2)

- Dependences on voltages, temperature, and oxide thickness – **reaction**

$$K = A \cdot \underbrace{\sqrt{C_{ox}\left(V_{gs} - V_{th}\right)}}_{\text{Hole density}} \cdot \exp\underbrace{\left(\frac{E_{ox}}{E_0}\right)}_{\text{Electric field dependence}} \cdot \exp\underbrace{\left(-\frac{E_a}{kT}\right)}_{\text{Temperature dependence}}$$

Modeling of Dynamic NBTI

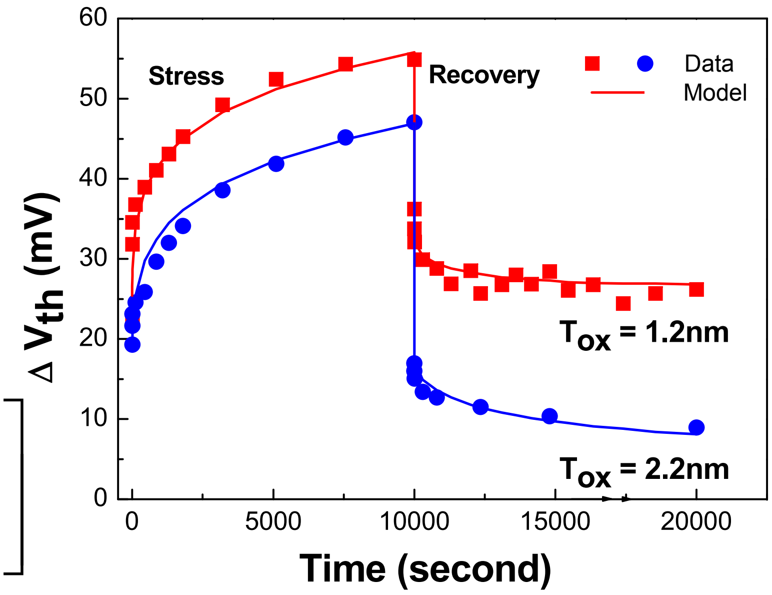
- A transistor with thicker t_{ox} recovers more than that of thinner t_{ox}
- Model is continuous at the boundary

Stress:

$$N_{it} = \left[K\left(t - t_0\right)^{0.5} + \sqrt[2n]{N_{it0}}\right]^{2n}$$

Recovery:

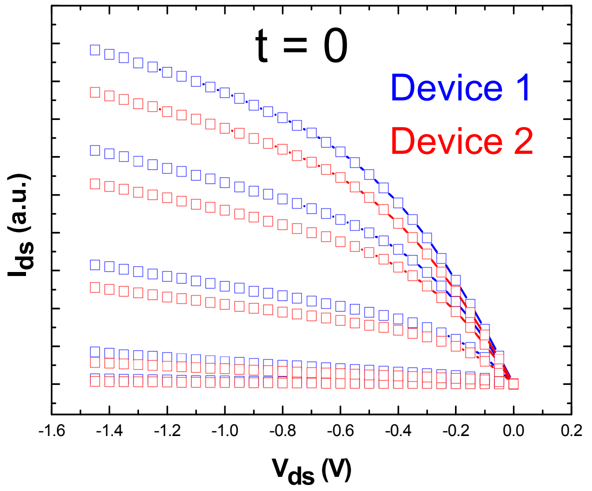
$$N_{it} = N_{it0} \cdot \left[1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C\left(t - t_0\right)}}{2t_{ox} + \sqrt{Ct}}\right]$$



90nm, V_{gs} = -1.2V, IEDM 2005

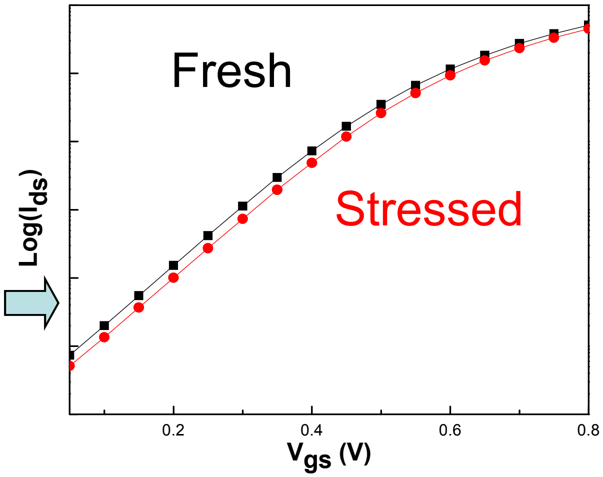
[W. Wang, CICC 2007]

Parameter Extraction



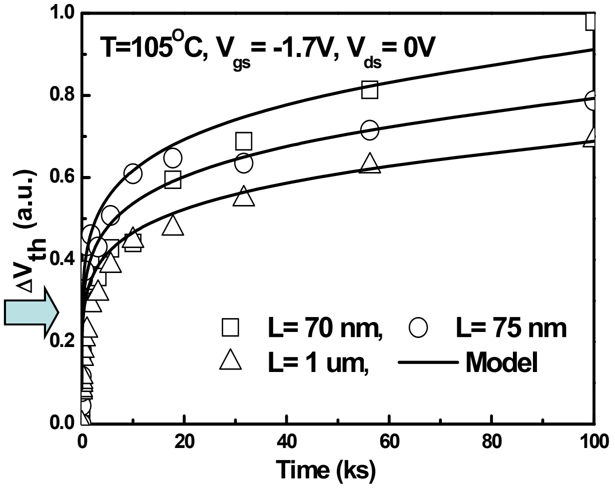
Decouple variations

(V_{th}, L, μ, etc.)



Extract the degradation

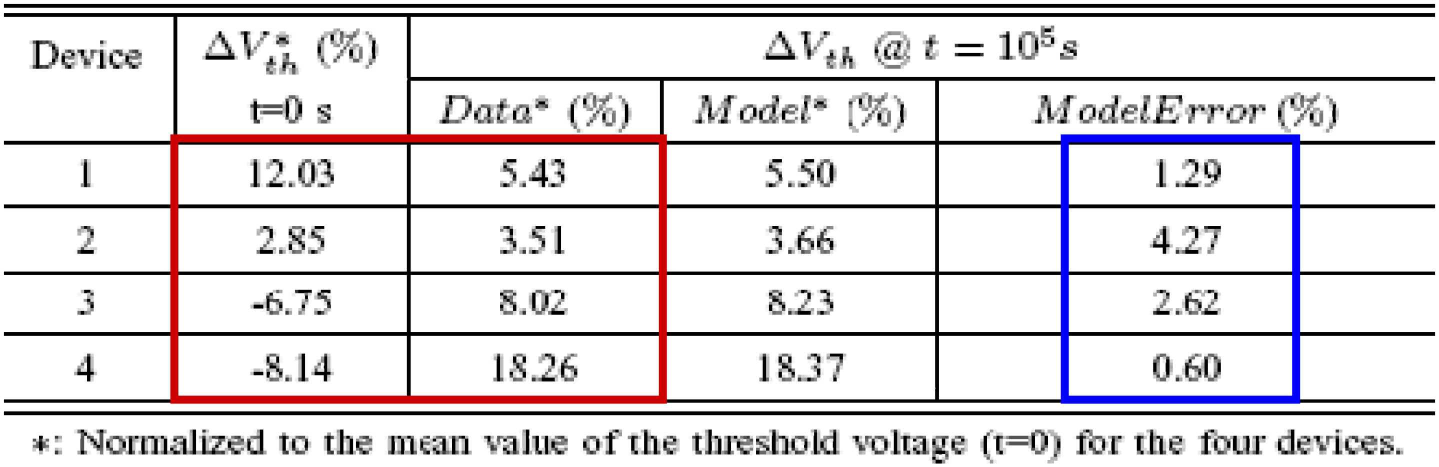
(V_{th} from $I_{leakage}$, μ from I_{linear})



Reliability model

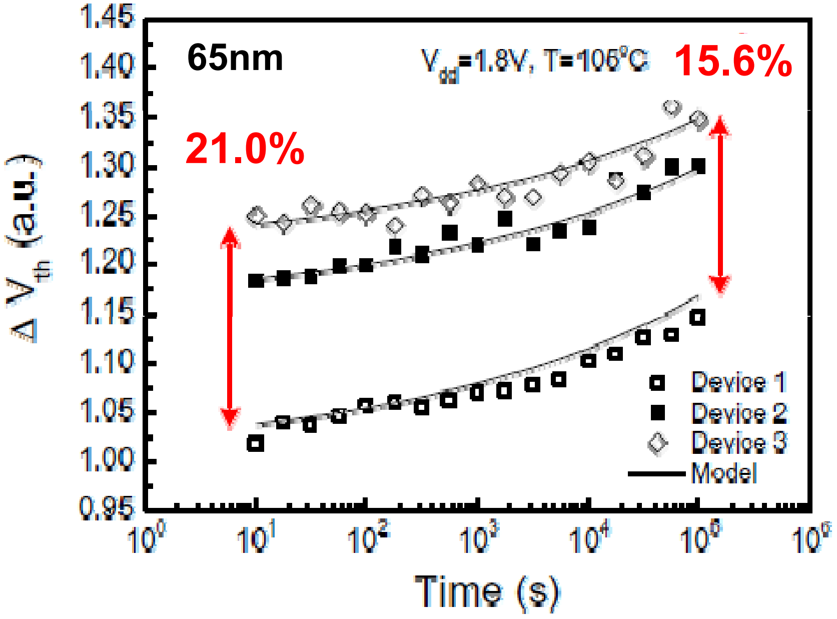
- Only 5-6 model parameters need to be extracted
- Reliability model is scalable with primary process and design parameters

Decoupling from Variations



Device	ΔV_{th}^{*} (%) t=0 s	ΔV_{th} @ $t = 10^5 s$		
		*Data** (%)	*Model** (%)	*ModelError* (%)
1	12.03	5.43	5.50	1.29
2	2.85	3.51	3.66	4.27
3	-6.75	8.02	8.23	2.62
4	-8.14	18.26	18.37	0.60

∗: Normalized to the mean value of the threshold voltage (t=0) for the four devices.

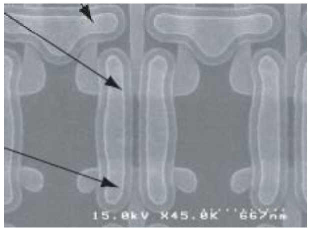


$$\Delta V_{th} \propto \exp\left(\frac{E_{ox}}{E_0}\right) = \exp\left(\frac{V_{gs}}{T_{ox} \cdot E_0}\right)$$

- The amount of temporal degradation is comparable to static variations

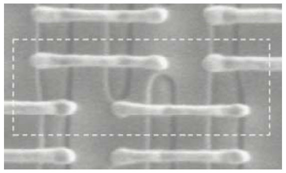
Brief Summary of Variations

130nm
2.34µm²

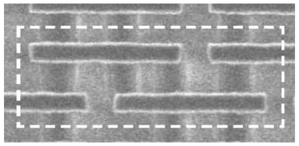


90nm
1.0µm²

65nm
0.57µm²



45nm
0.346µm²



Sources: IBM, Intel; picture size not to scale

	Statistical Property	Spatial Correlation	Design Solution
Intrinsic Variations	Random (σ predictable)	Weak (nm)	Joint tech-design optimization
Manufacturing Variations	“Systematic”	Strong (100μm to mm)	Modeling and design optimization

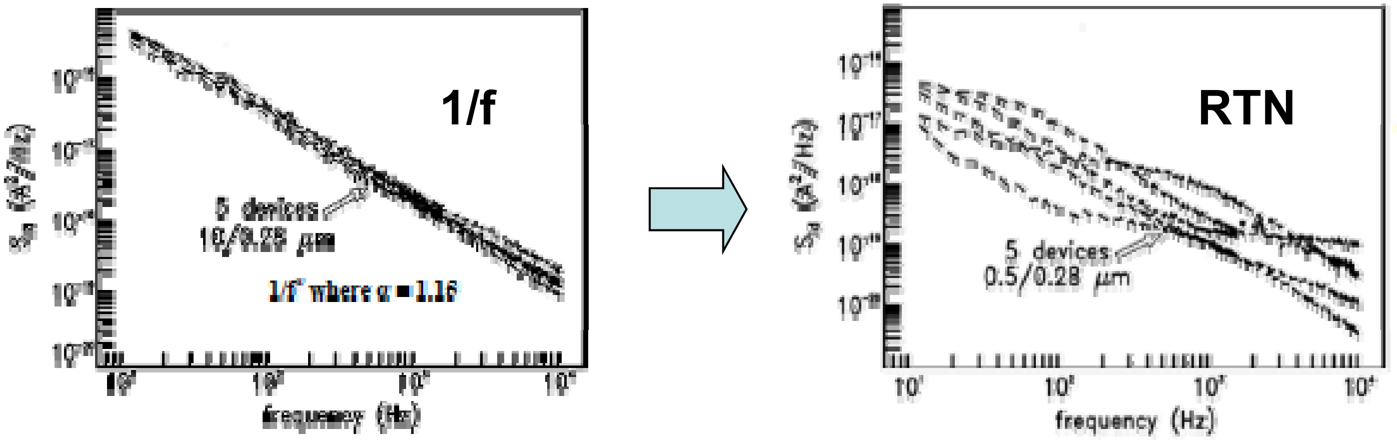
***Additional variations from dynamic operations, e.g., V_{dd} noise, NBTI, etc.**

Predictive Variability Modeling

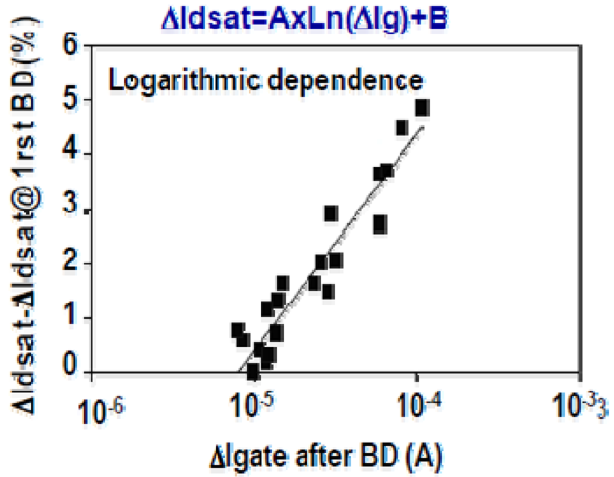
- Process Variations in Light of Scaling
- Intrinsic and Manufacturing Variations
- **Future Modeling Needs and Promises**

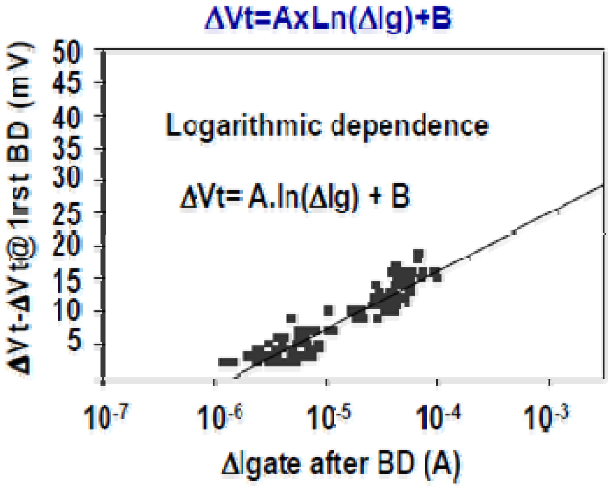
Emerging Variation Effects

- Toward random, discrete, atomistic variations



- Correlation among various mechanisms: e.g., TDDB + NBTI

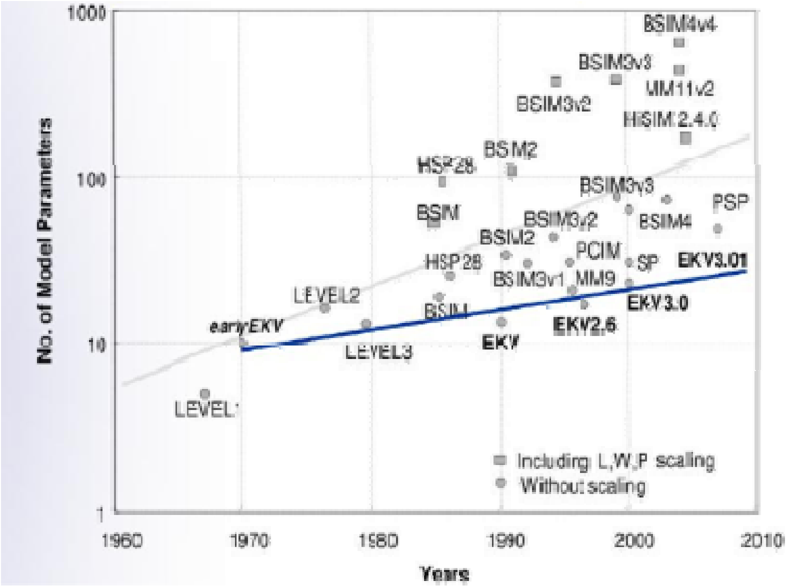




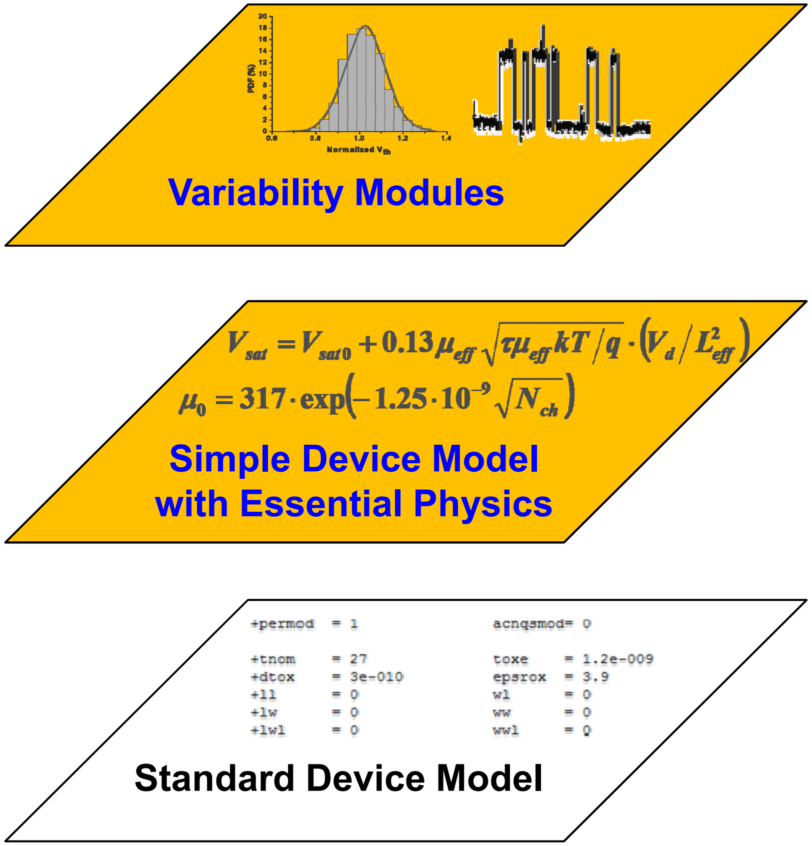
[A. J. Scholten, TED 2003; G. Ribes, IRPS 2008]

Model Efficiency and Flexibility

- Device modeling
  - From corner based to statistical, and maybe hybrid
  - But much heavier with too many instances



- Solution: hierarchical, module based modeling structure



Variability Modules

$$V_{sat} = V_{sat0} + 0.13\mu_{eff}\sqrt{\tau\mu_{eff}kT/q}\cdot\left(V_d/L_{eff}^2\right)$$

$$\mu_0 = 317\cdot\exp\left(-1.25\cdot10^{-9}\sqrt{N_{ch}}\right)$$

Simple Device Model
with Essential Physics

```
+permod  = 1            acnqsmod= 0

+tnom    = 27           toxe     = 1.2e-009
+dtox    = 3e-010       epsrox   = 3.9
+ll      = 0            wl       = 0
+lw      = 0            ww       = 0
+lwl     = 0            wwl      = 0
```

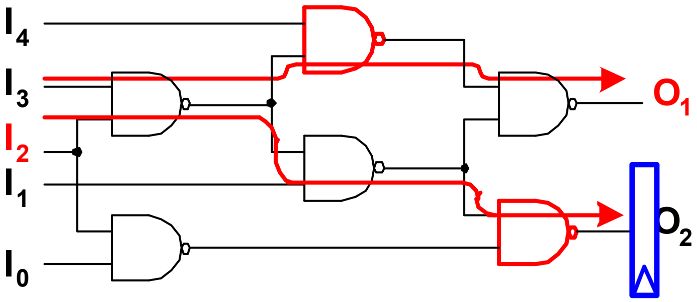
Standard Device Model

[W. Grabinski, 2008]

Integration with Design Practice

- Modeling and simulation tools enabling quantitative assessment during the design stage
- Support adaptive detection and protection scheme
  - Predict statistical performance change
  - Detect critical units, evaluate the overhead, and optimize the solution

A 90nm Ethernet Controller



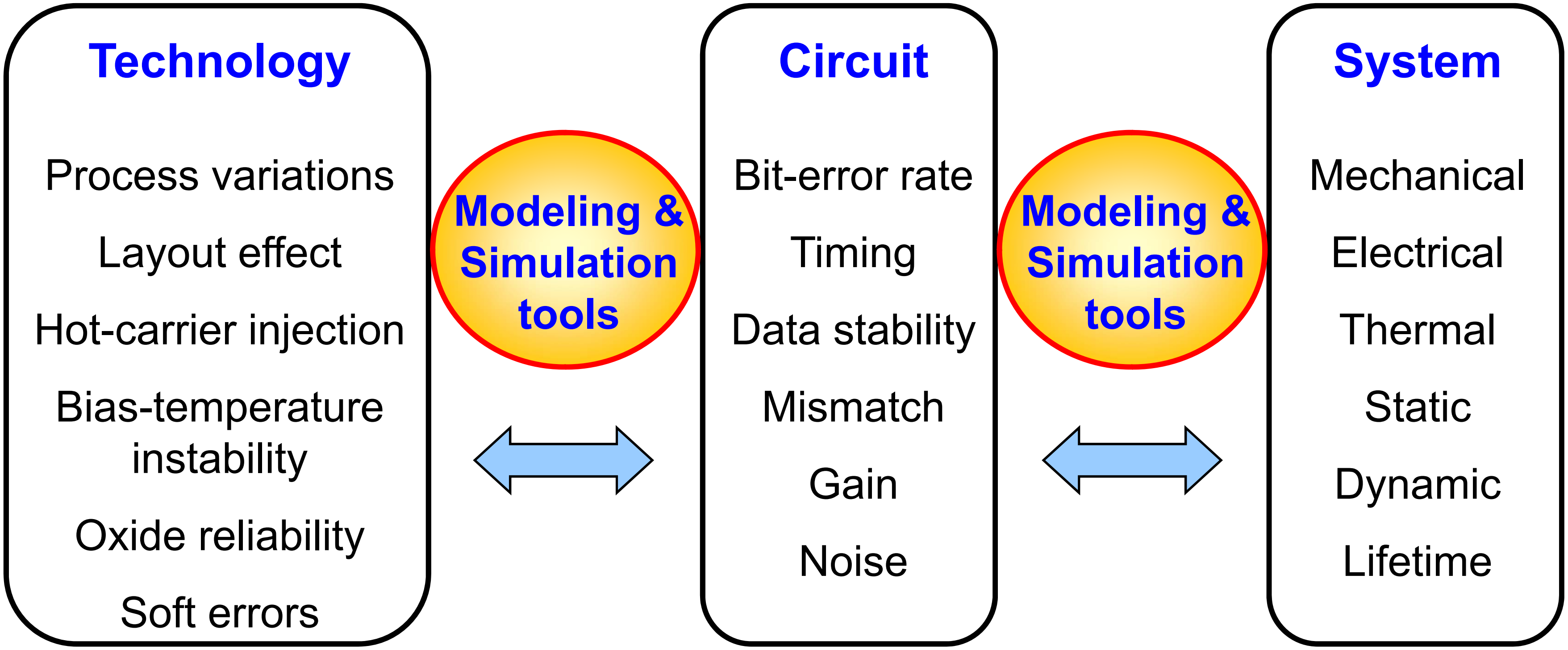
Protection (20% slack)	Flip-Flops	Percentage	Area overhead
w/o aging analysis	1667	87 %	10%
w/ aging analysis	479	25 %	2%

[ITC 2008]

Design for Variability & Reliability

- "Germany began using it during WWII and it is cited as a reason for Japan's current electronic dominance."

  – B. E. Hegler, 1988



Technology

Process variations

Layout effect

Hot-carrier injection

Bias-temperature instability

Oxide reliability

Soft errors

Modeling & Simulation tools

Circuit

Bit-error rate

Timing

Data stability

Mismatch

Gain

Noise

Modeling & Simulation tools

System

Mechanical

Electrical

Thermal

Static

Dynamic

Lifetime