

# EECS251B : Advanced Digital Circuits and Systems

## Lecture 17 – Variability

Borivoje Nikolić

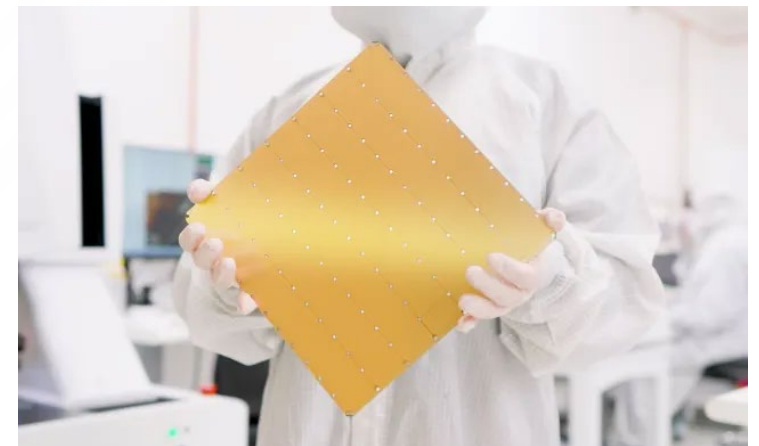


### Cerebras' Third-Gen Wafer-Scale Chip Doubles Performance

**March 13, 2024,** Sally Ward-Foxton, **EETimes.** Cerebras has unveiled a third generation of its wafer-scale chip, offering 125 PFLOPS (at FP16 precision) from a single device. Given a single day, a four-chip installation could fine-tune Llama2-70B, while the biggest installations of 2,048 chips would be able to train it from scratch in the same time.

The wafer-scale engine 3 (WSE3) doubles the large language model (LLM) training speed of the WSE2, in the same 15kW power envelope and at the same cost point, Cerebras CEO Andrew Feldman told EE Times.

...The WSE also features 42 GB of SRAM with 21 PBytes/s memory bandwidth.



Cerebras' third-gen wafer-scale engine.  
(Source: Cerebras)

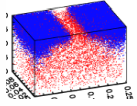
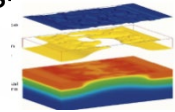
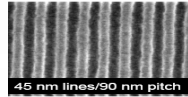
# Announcements

- **Project**
  - Midterm reports due next week
  - Preliminary design review after Spring break
- **Homework 3 due next week**
  - Quiz 3 after Spring break



# Design Variability Sources and Impact on Design

# Systematic and Random Device Variations

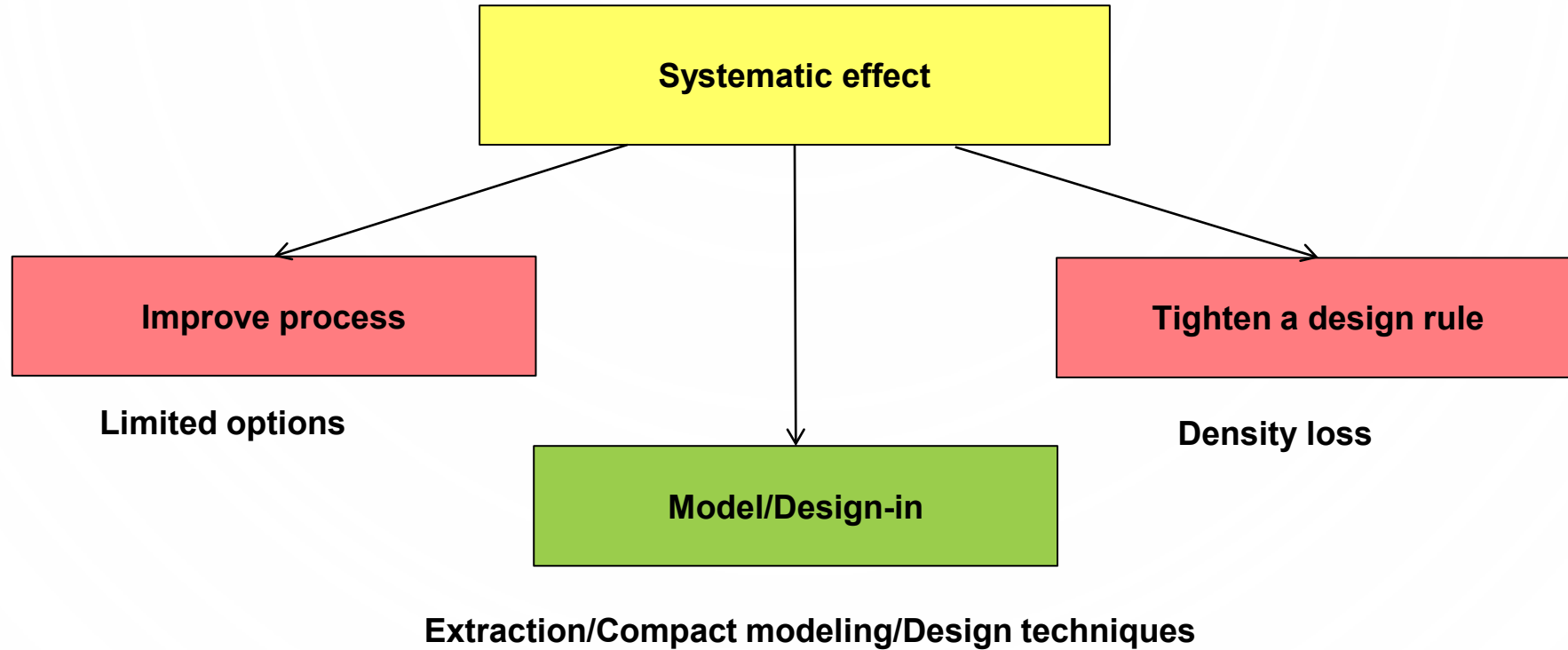
Parameter	Random	Systematic
Channel Dopant Concentration $N_{ch}$	Affects $\sigma_{VT}$ <sup>[1]</sup> 	Non uniformity in the process of dopant implantation, dosage, diffusion
Gate Oxide Thickness $T_{ox}$	Si/SiO <sub>2</sub> & SiO <sub>2</sub> /Poly-Si interface roughness <sup>[2]</sup> 	Non uniformity in the process of oxide growth
Threshold Voltage $V_T$ (non $N_{ch}$ related)	Random anneal temperature and strain effects	Non-uniform annealing temperature <sup>[5]</sup> (metal coverage over gate) Biaxial strain
Mobility $\mu$	Random strain distributions	Systematic variation of strain in the Si due to STI, S/D area, contacts, gate density, etc
Gate Length $L$	Line edge roughness (LER) <sup>[3]</sup> 	Lithography and etching: Proximity effects, orientation <sup>[4]</sup>
Fin geometry/ film thickness variations	Rounding, etc, $\sigma_{VT}$ , mobility.	Systematic fin thickness Systematic Si film/BOX variations

[1] D. Frank et al, *VLSI Symposium*, Jun. 1999 . [2] A. Asenov et al, *IEEE Trans on Electron Devices*, Jan. 2002.

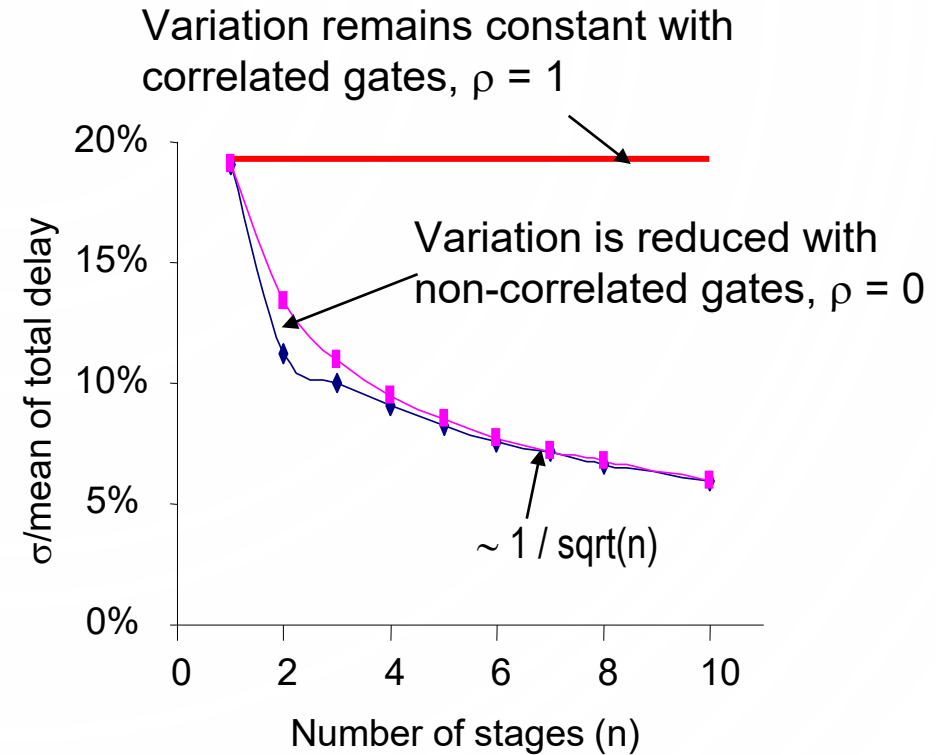
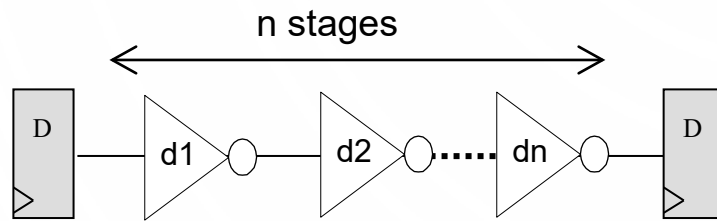
[3] P. Oldiges et al, *SISPAD 2000*, Sept. 2000. [4] M. Orshansky et al, *IEEE Trans on CAD*, May 2002. [5] Tuinhout et al, *IEDM*, Dec 1996

# Dealing with Systematic Variations

- Model-to-hardware correlation classifies unknown sources



# Chip Yield Depends on Inter-Gate Correlation

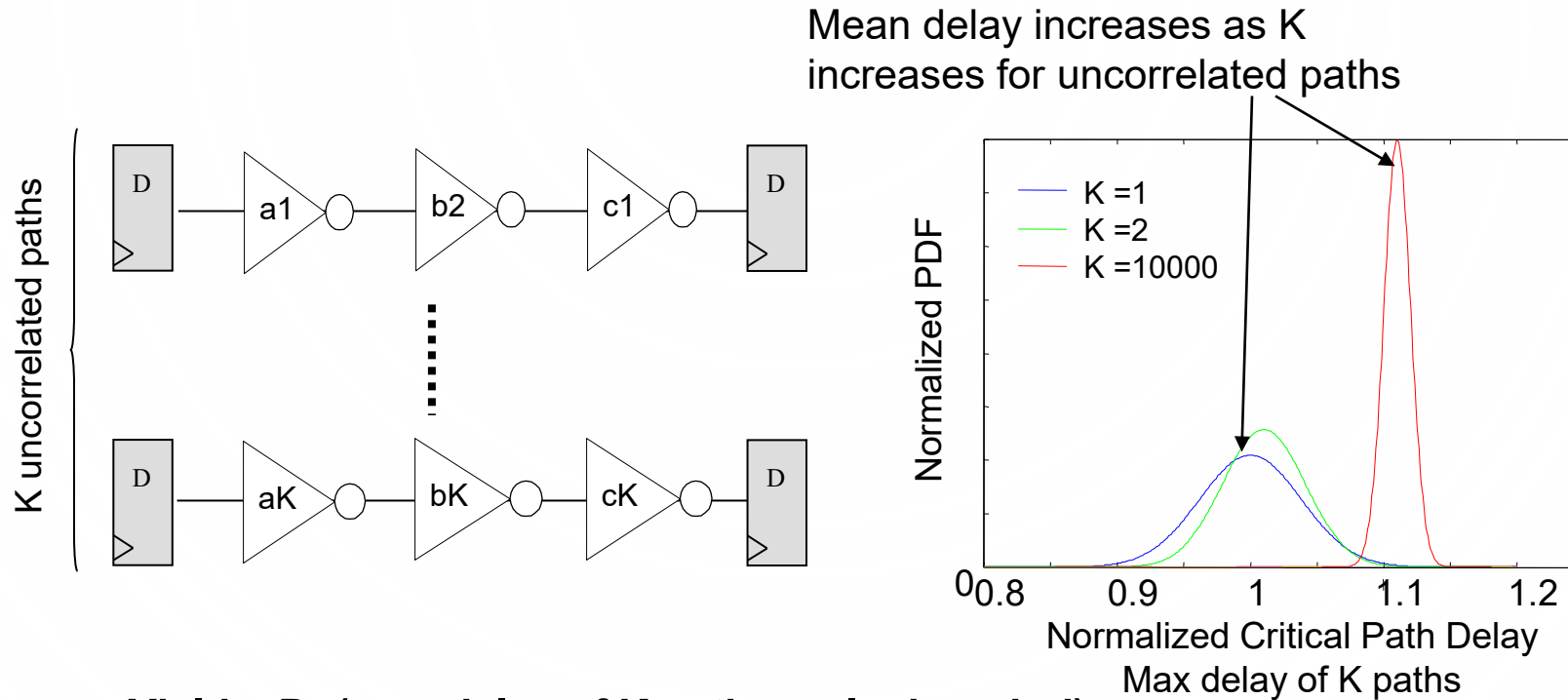


- ▶ **Yield = Pr (sum of n delays < clock period)**
- ▶  **$\rho = 0$  gives highest yield through averaging**

**Non-correlated gates in a path reduce impact of variation**

Bowman et al, *JSSC*, Feb 2002 .

# Chip Yield Depends on Inter-Path Correlation



- ▶ **Yield = Pr (max delay of K paths < clock period)**
- ▶ **K = 1 gives highest yield**

**Correlated paths reduce impact of variation**

Bowman et al, JSSC, Feb 2002 .

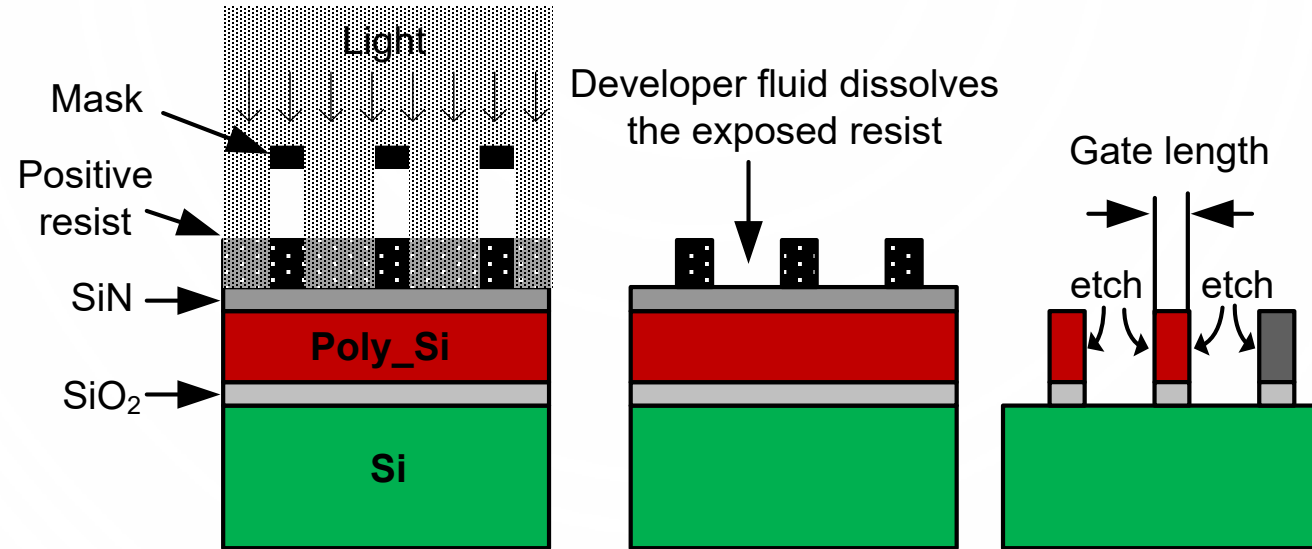


## Design Variability Some Systematic Effects



# Layout: Poly Proximity Effects

- Gate CD is a function of its neighborhood



## Gate length depends on

- Light intensity profile falling on the resist
- Resist: application of developer fluid<sup>[1]</sup>, post exposure bake (PEB) temperature<sup>[2]</sup>
- Dry etching: microscopic loading effects<sup>[3]</sup>

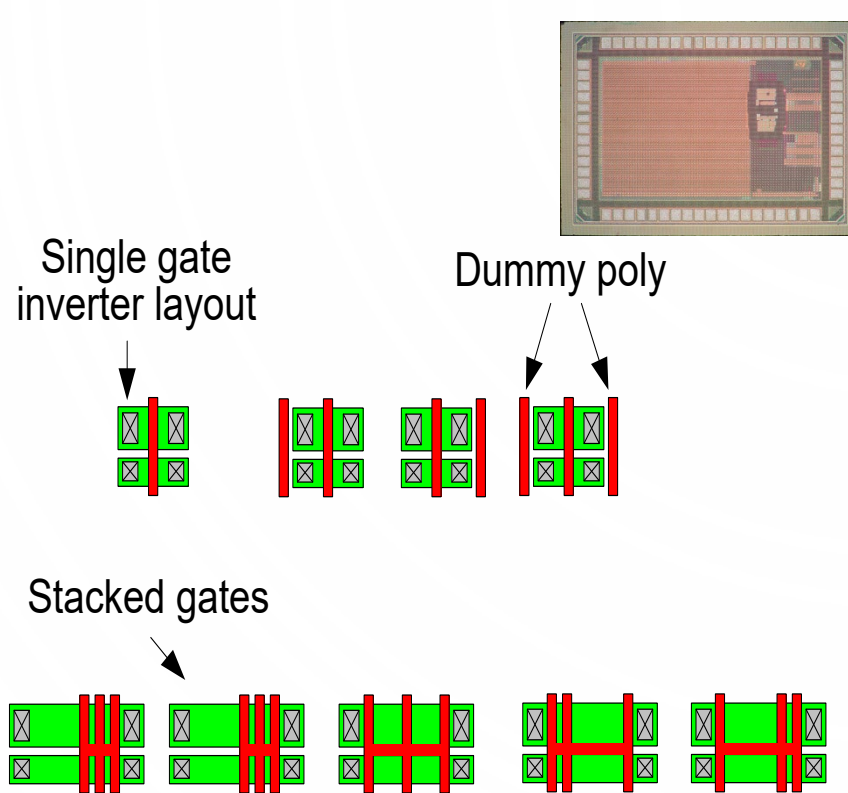
[1] J.Cain, M.S. Thesis, UC Berkeley

[2] D. Steele et al, *SPIE*, vol.4689, July 2002.

[3] J. D. Plummer, M.D. Deal, P.B. Griffin, *Silicon VLSI Technology*, Prentice-Hall, 2000.

# Layout: Proximity Test Structures

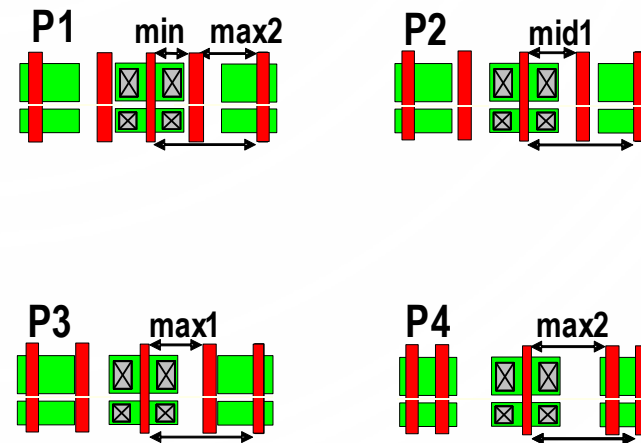
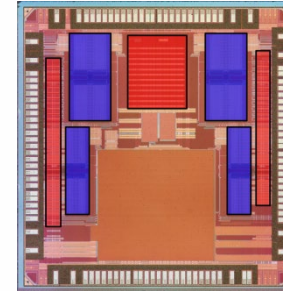
- 90nm experiments



L.T. Pang, VLSI'06

- 45nm experiments

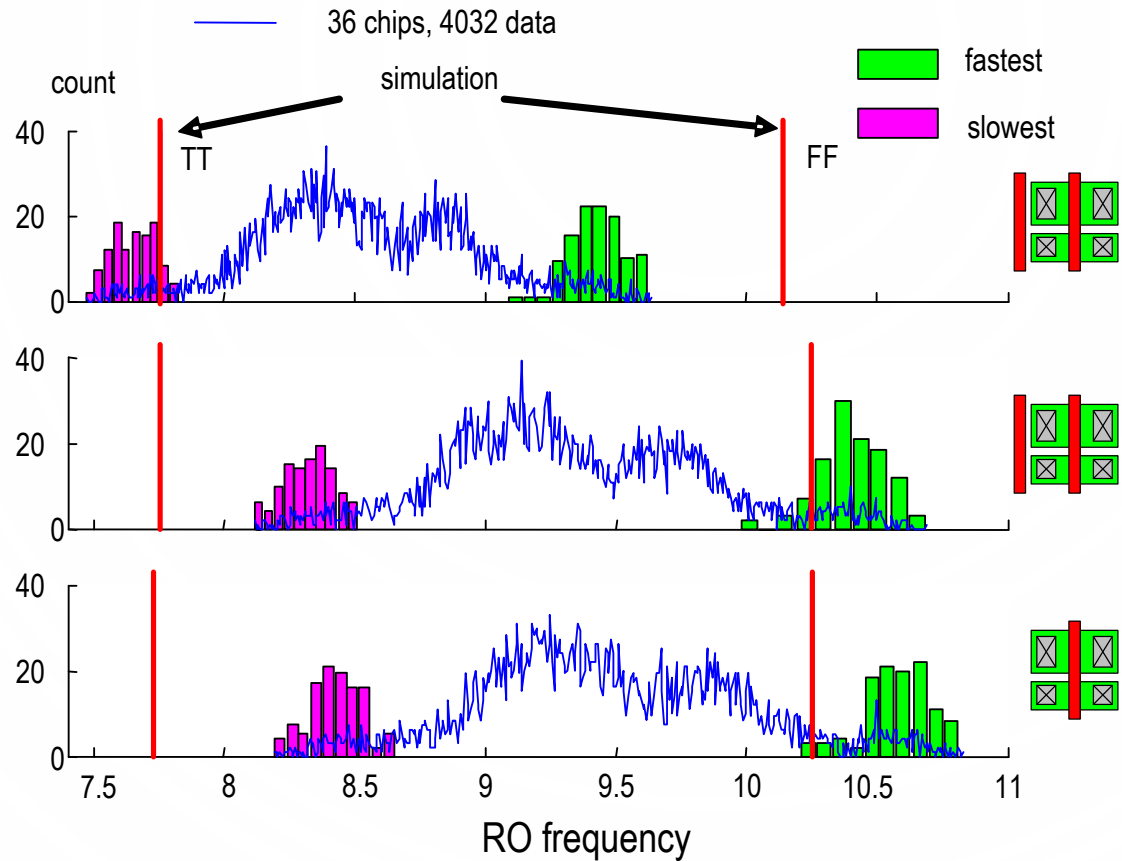
No single gates allowed



L.T. Pang, CICC'08

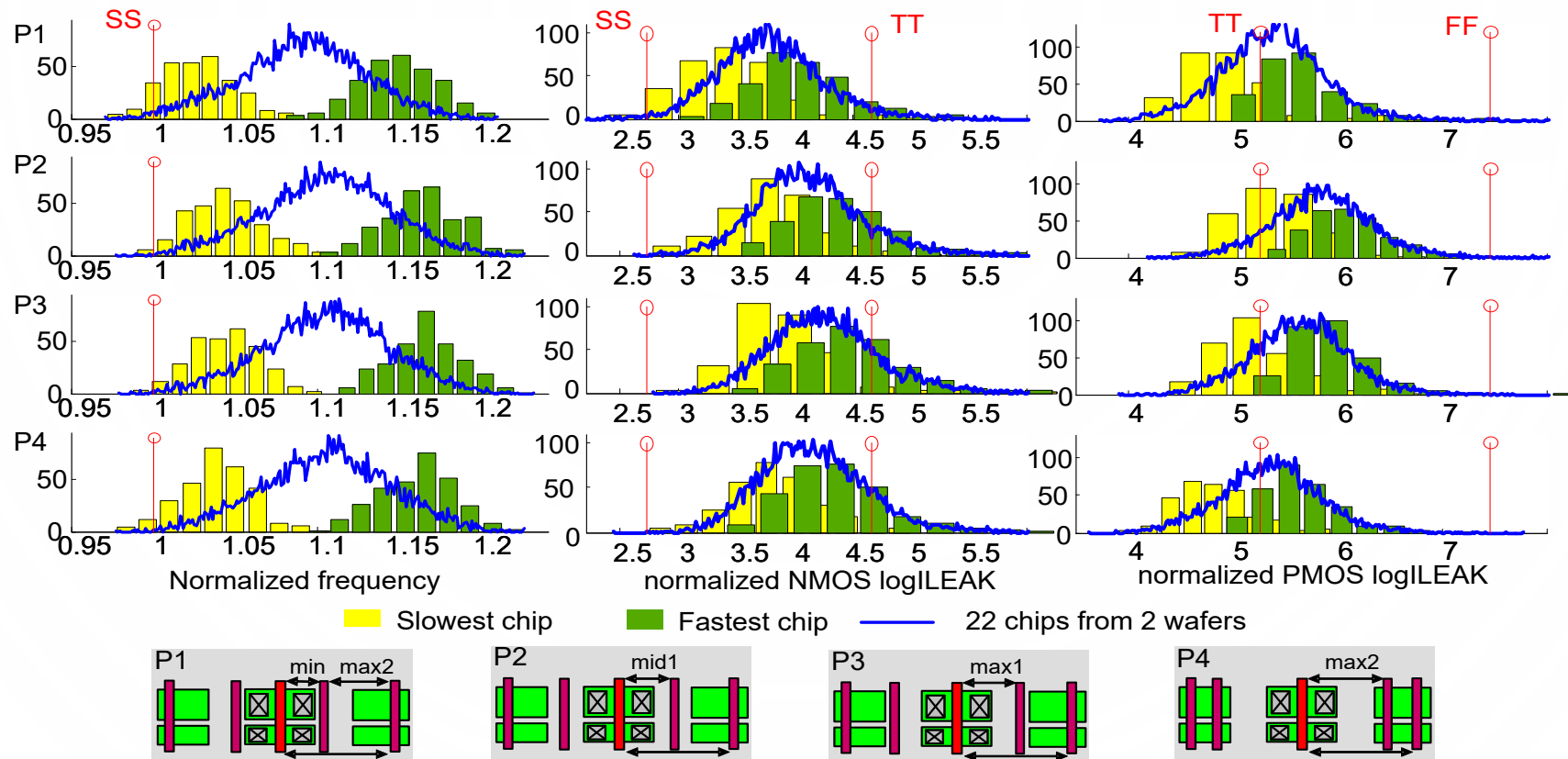
- **Ring oscillators and individual transistor leakage currents**

# Results: Single Gates in 90nm



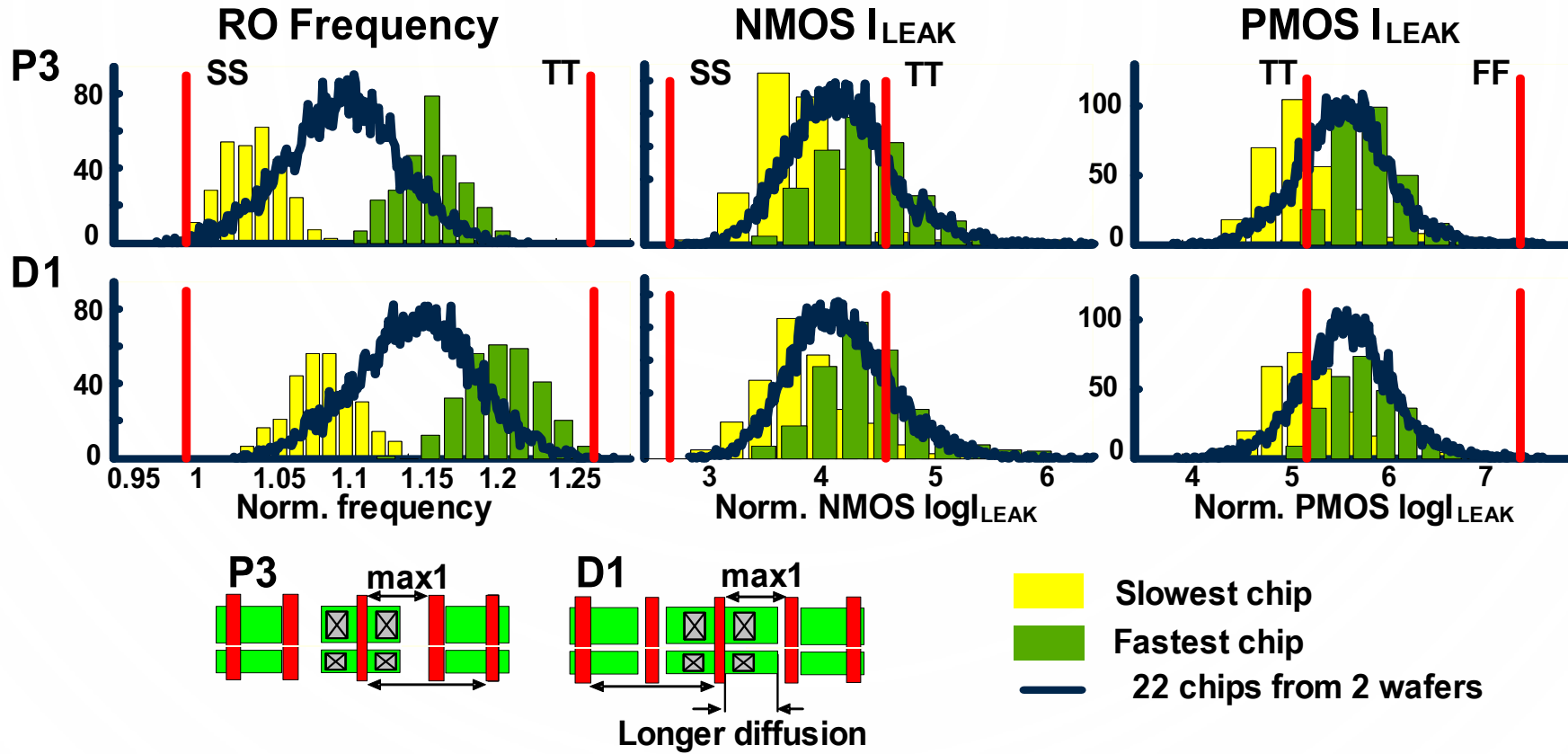
- **Max  $\Delta F$  between layouts  $> 10\%$**
- **Within-die  $3\sigma/\mu \sim 3.5\%$ , weak dependency on density**

# Results: Single Gates in 45nm



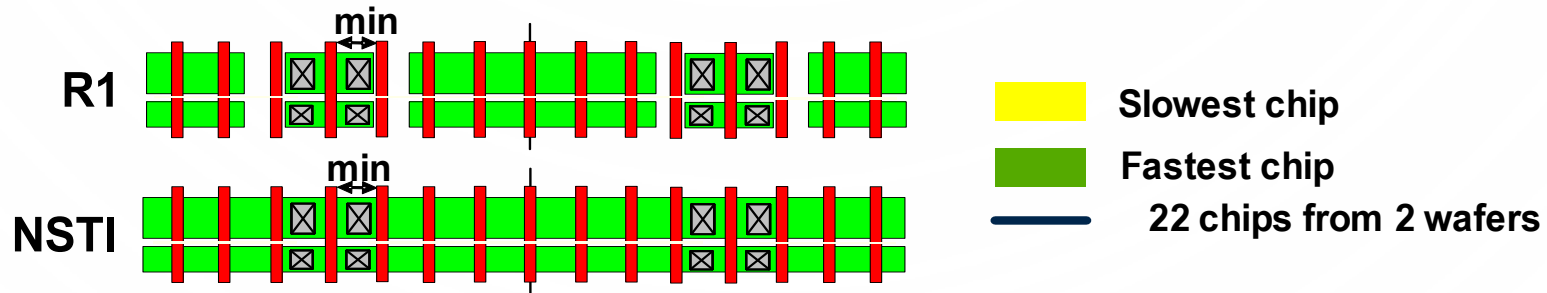
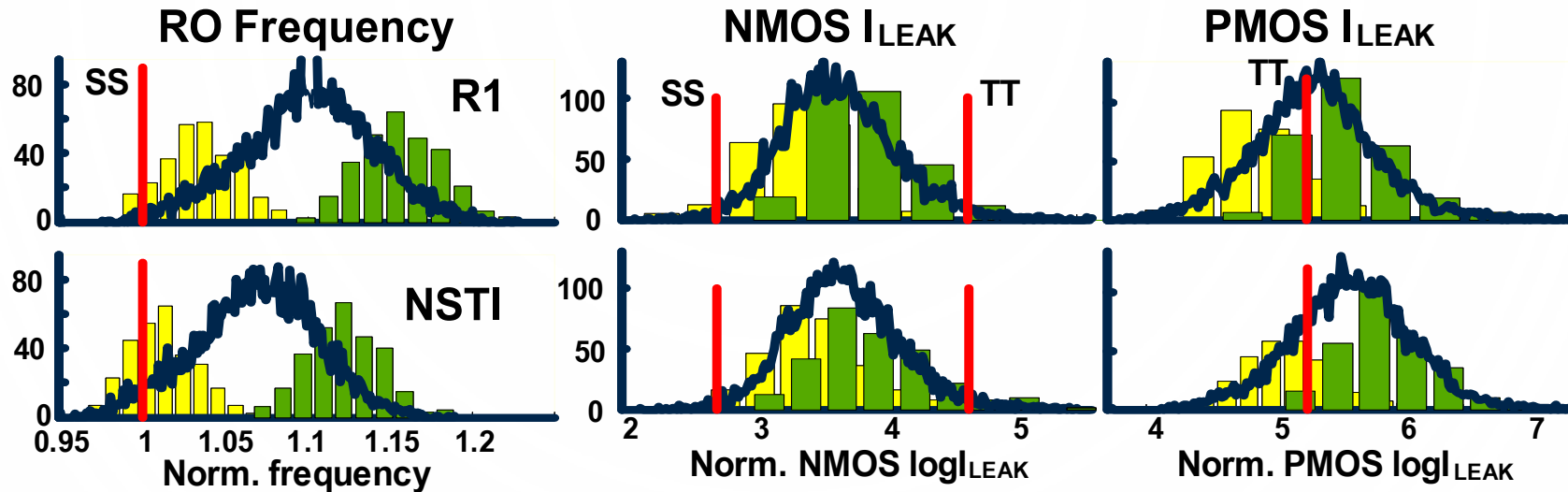
- **Weak effect on performance.  $\Delta F \sim 2\%$**
- **Small shifts in NMOS leakage and bigger shifts in PMOS leakage**

# Impact of Longer Diffusion in 45nm



- Strongest effect measured in 45nm,  $\Delta F \sim 5\%$
- No significant shift in  $I_{LEAK}$

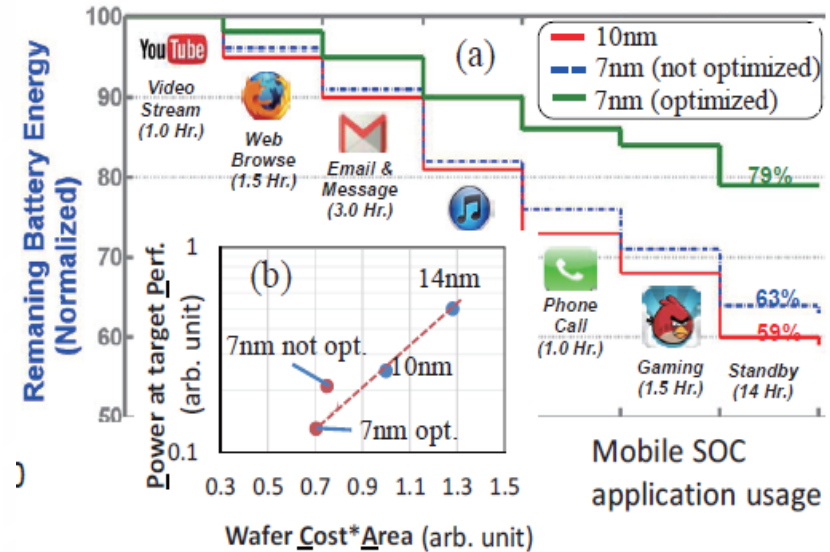
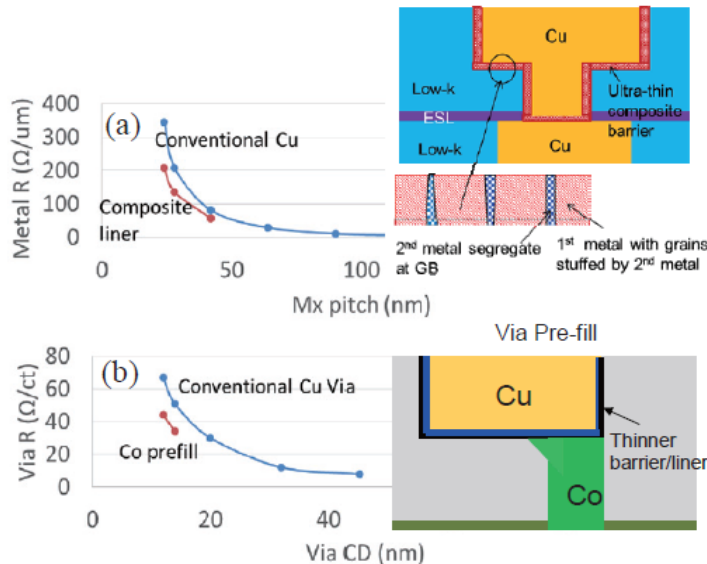
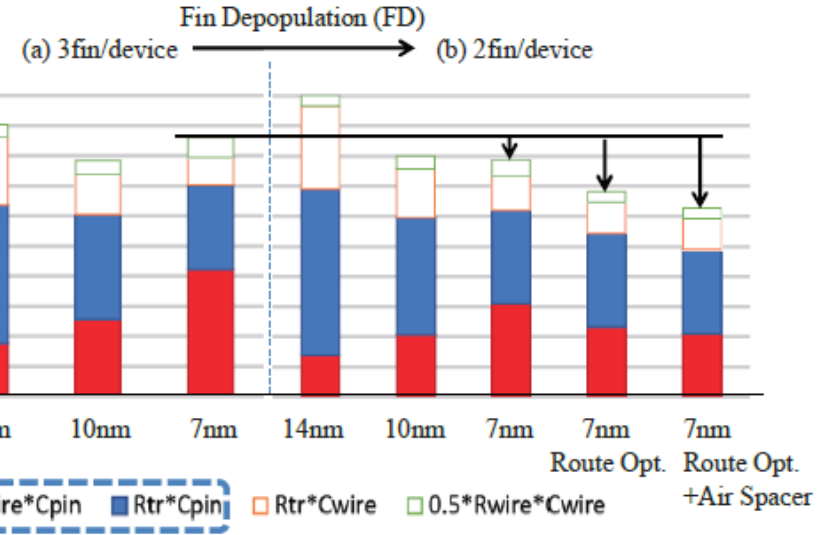
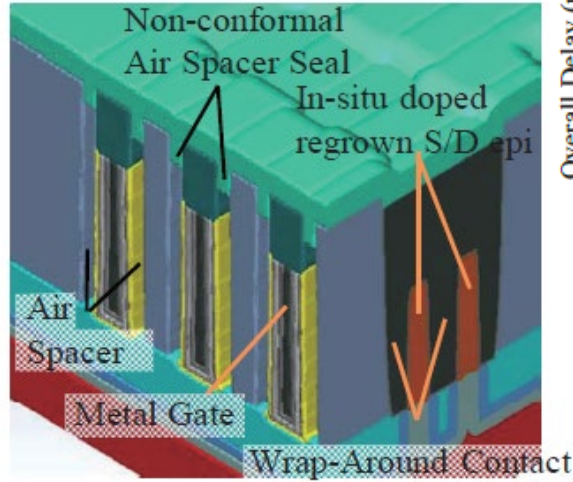
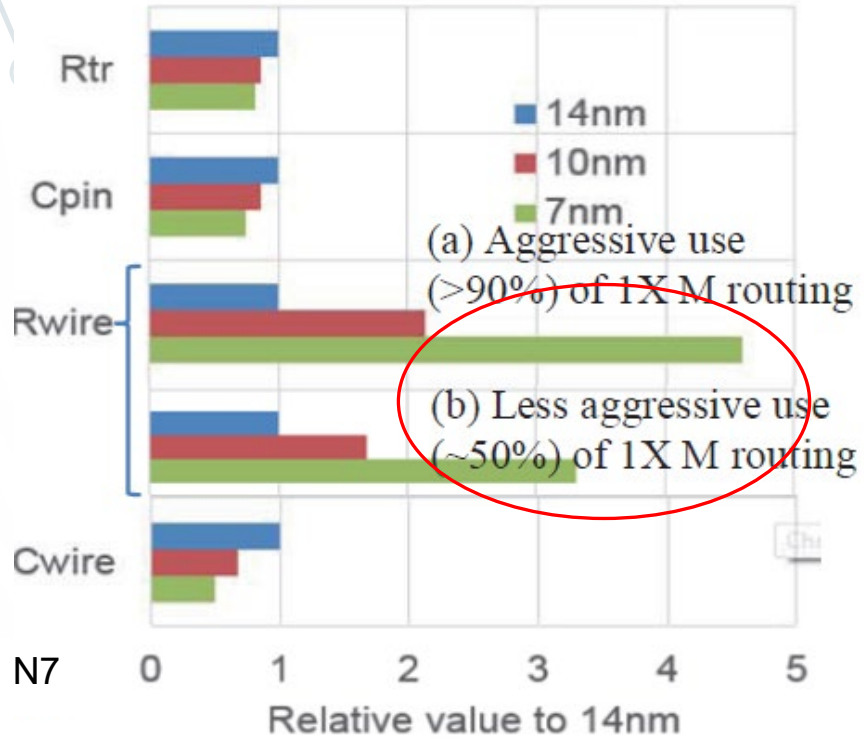
# Impact of Shallow Trench Isolation (STI)



- $\Delta F \sim 3\%$ , small changes in  $I_{LEAK}$
- Due to STI-induced stress

# Patterning and process impact on FinFETs

$$\text{Delay} = R_{\text{wire}} \times C_{\text{pin}} + R_{\text{tr}} \times C_{\text{pin}} + R_{\text{tr}} \times C_{\text{wire}} + 1/2 \times R_{\text{wire}} \times C_{\text{wire}}$$



Song VLSI'15

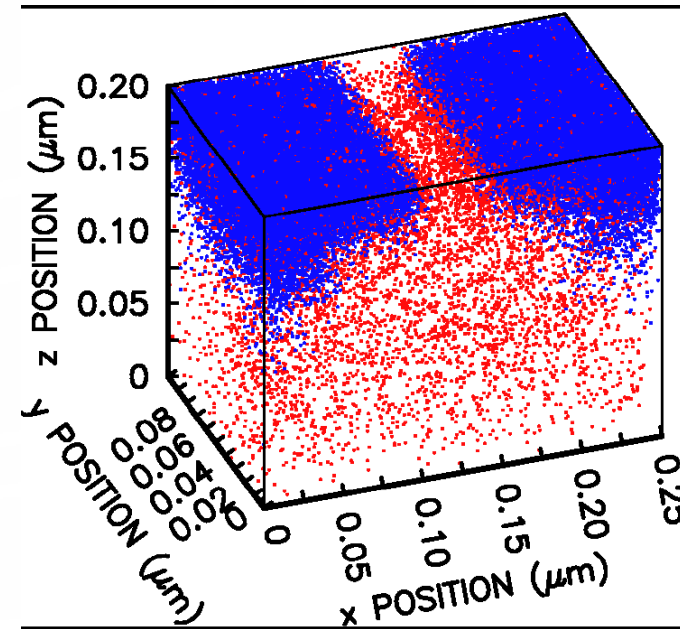
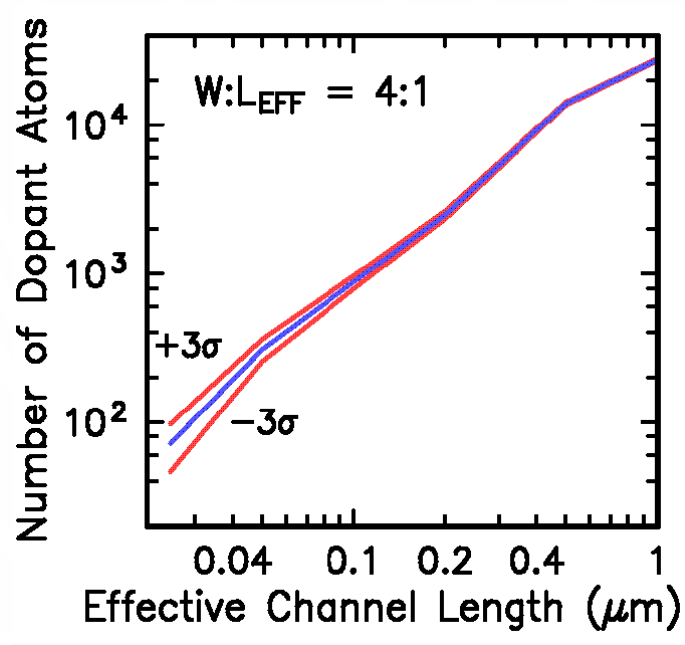


## Design Variability Some Random Effects



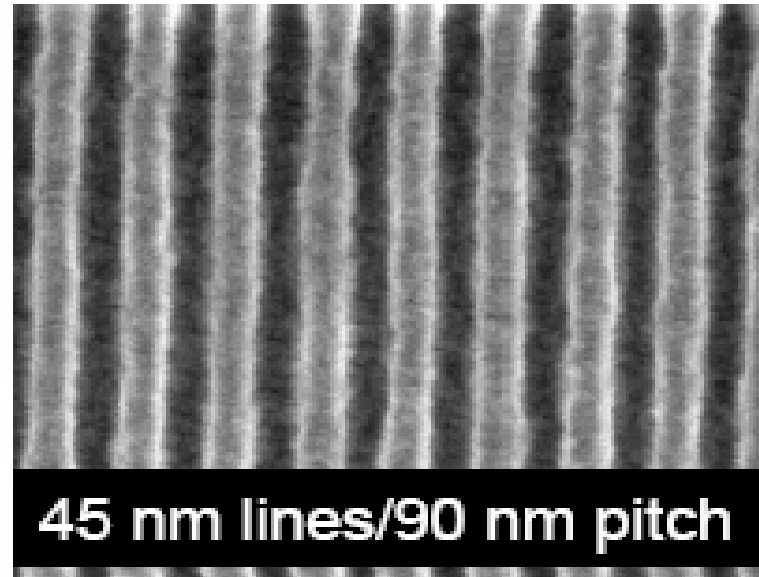
# Random Dopant Fluctuations

- Number of dopants is finite



Frank, IBM J R&D 2002

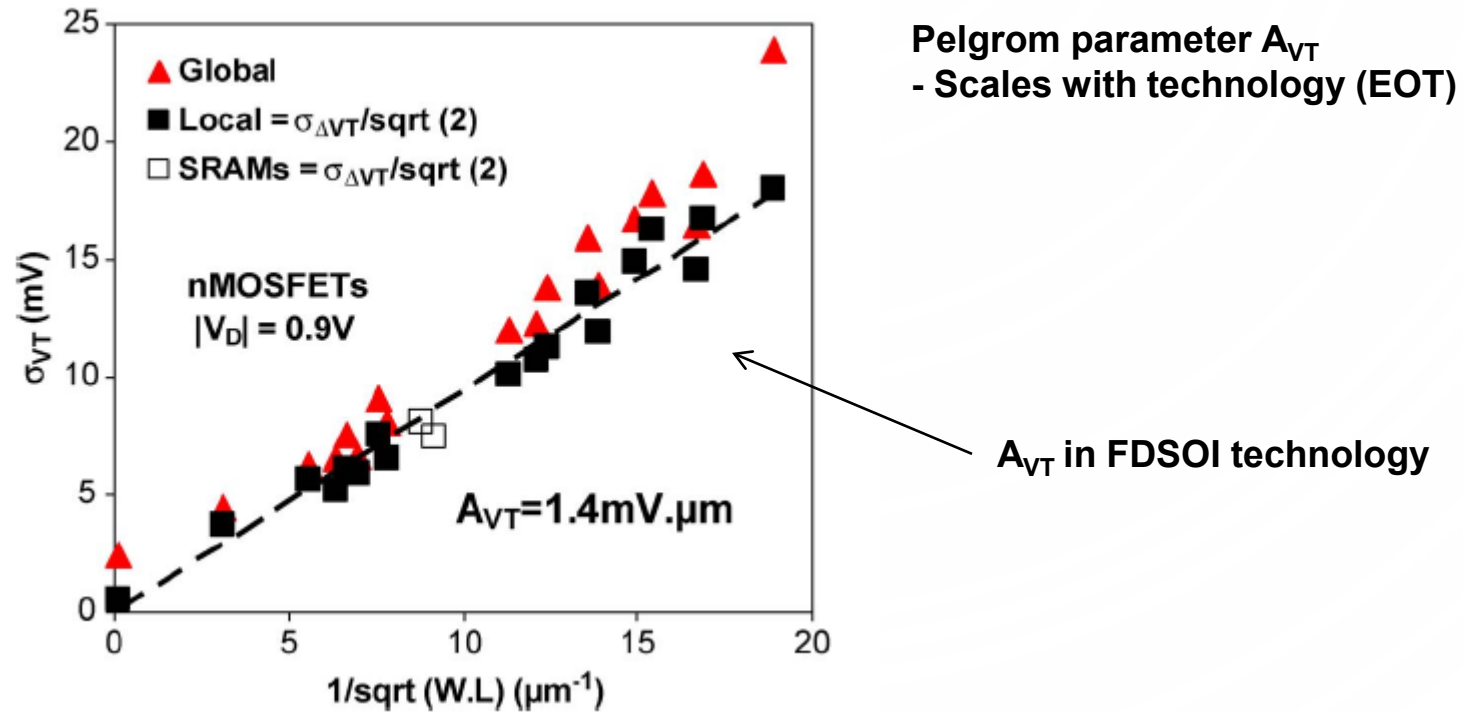
# Processing: Line-Edge Roughness



- Sources of line-edge roughness:
  - Fluctuations in the total dose due to quantization
  - Resist composition
  - Absorption positions
- Effect:
  - Variation (random) in leakage and power

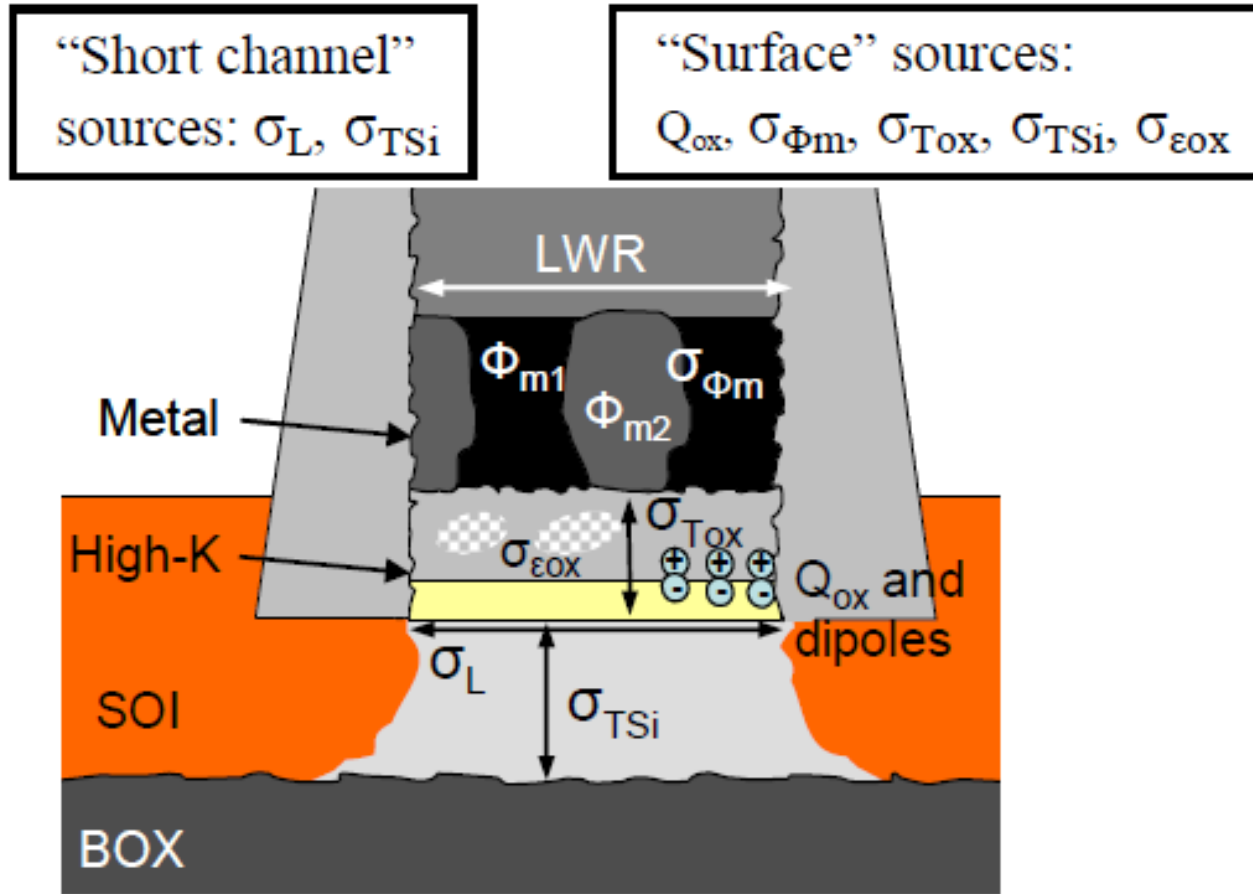
# Transistor Matching

- $V_{Th}$  matching of geometrically identical transistors varies with size  $\sim \sqrt{WL}$  and distance



J. Mazurier, Trans E.D., 2011.

# FDSOI example



- All the effects follow  $1/\sqrt{WL}$  dependence

Short channel effect:  $\sigma_L, \sigma_{TSi}$

$$\sigma_{Vt,SCE} \propto \sqrt{\left(\frac{\partial V_t}{\partial L} \sigma_L\right)^2 + \left(\frac{\partial V_t}{\partial T_{Si}} \sigma_{TSi}\right)^2} \quad (1)$$

$V_t$  long channel [18] :

$$V_t = \Delta\phi_{mi} + \frac{k.T}{q} \ln\left(\frac{2.C_{ox}.k.T}{q^2.n_i.T_{Si}}\right) + \frac{\hbar^2.\pi^2}{2.q.m^*.T_{Si}^2} \quad (2)$$

with  $\Delta\phi_{mi}$  the gate WF with respect to intrinsic Si.

Oxide charges:

$$\sigma_{Vt,Qox} \propto \frac{q.T_{ox}}{\epsilon_{ox}} \frac{\sqrt{N_{it} + N_{ox}}}{\sqrt{W.L}} \quad (3)$$

Oxide thickness and permittivity [19]:  $\sigma_{Tox}, \sigma_{\epsilon_{ox}}$

$$\sigma_{Vt,Tox} \propto \frac{k.T}{q} \frac{\alpha}{\sqrt{W.L}} \sqrt{\left(\frac{\sigma_{\epsilon_{ox}}}{\epsilon_{ox}}\right)^2 + \left(\frac{\sigma_{Tox}}{T_{ox}}\right)^2} \quad (4)$$

$T_{Si}$  thickness:  $\sigma_{TSi}$

$$\sigma_{Vt,TSi} \propto \frac{k.T}{q} \frac{\beta}{\sqrt{W.L}} \frac{\sigma_{TSi}}{T_{Si}} \quad (5)$$

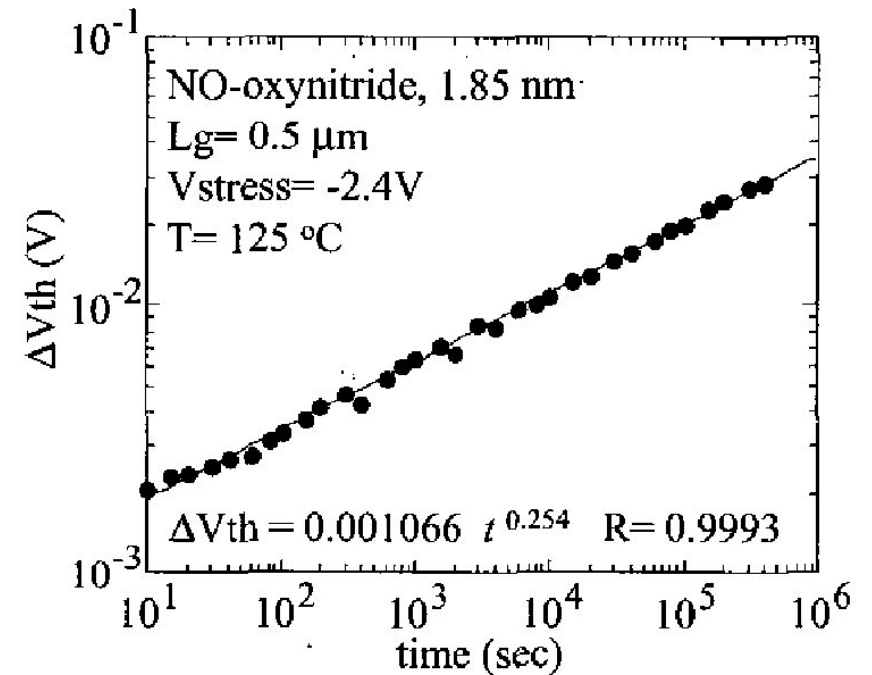
Metal gate workfunction :  $\sigma_{\Phi_m}$

$$\sigma_{Vt,\Phi_m} \propto \frac{\gamma}{\sqrt{W.L}} \sigma_{\Phi_m} \quad (6)$$

with  $\alpha, \beta, \gamma$  spatial correlation lengths of the fluctuations

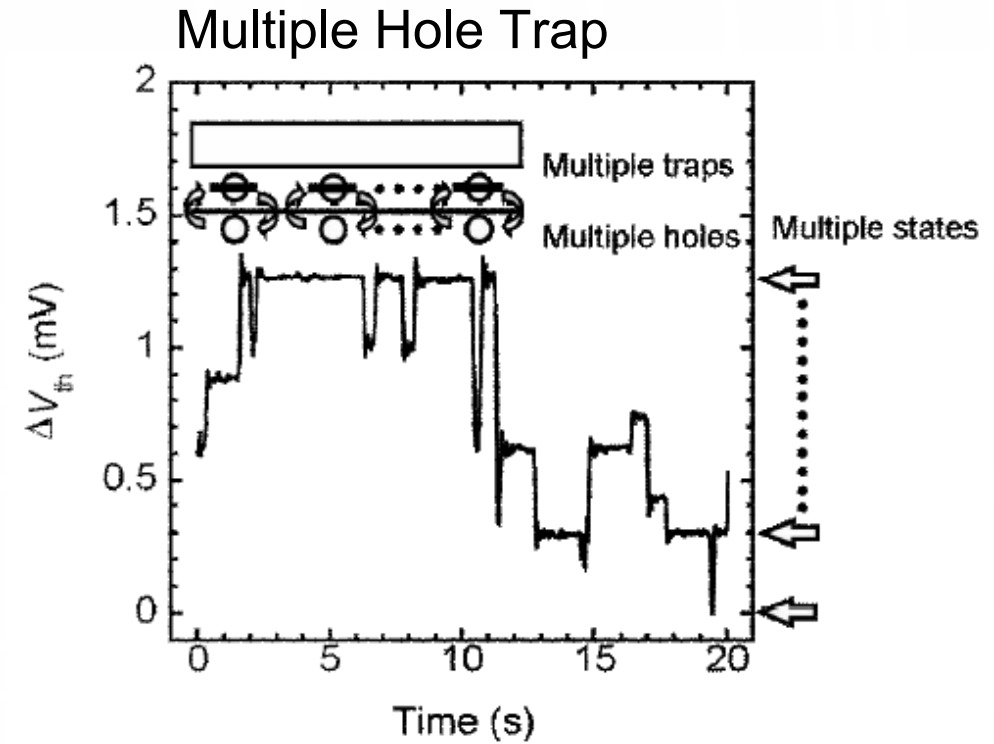
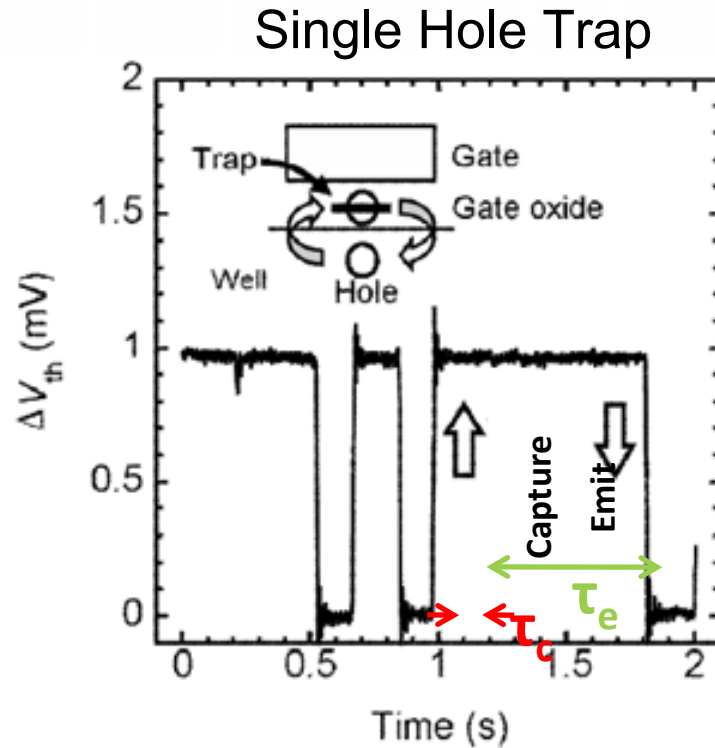
# Negative Bias Temperature Instability

- PFET  $V_{Th}$ 's shift in time, at high negative bias and elevated temperatures
- The mechanism is thought to be the breaking of hydrogen-silicon bonds at the Si/SiO<sub>2</sub> interface, creating surface traps and injecting positive hydrogen-related species into the oxide.
- Also other charge trapping and hot-carrier defect generation
- Systematic + random shifts



Tsujikawa, IRPS'2003

# Random Telegraph Signal (RTS)



- Trapping of a carrier in oxide traps modulates  $V_{th}$  or  $I_{ds}$
- $\tau_e$  and  $\tau_c$  are random and follow exponential distributions

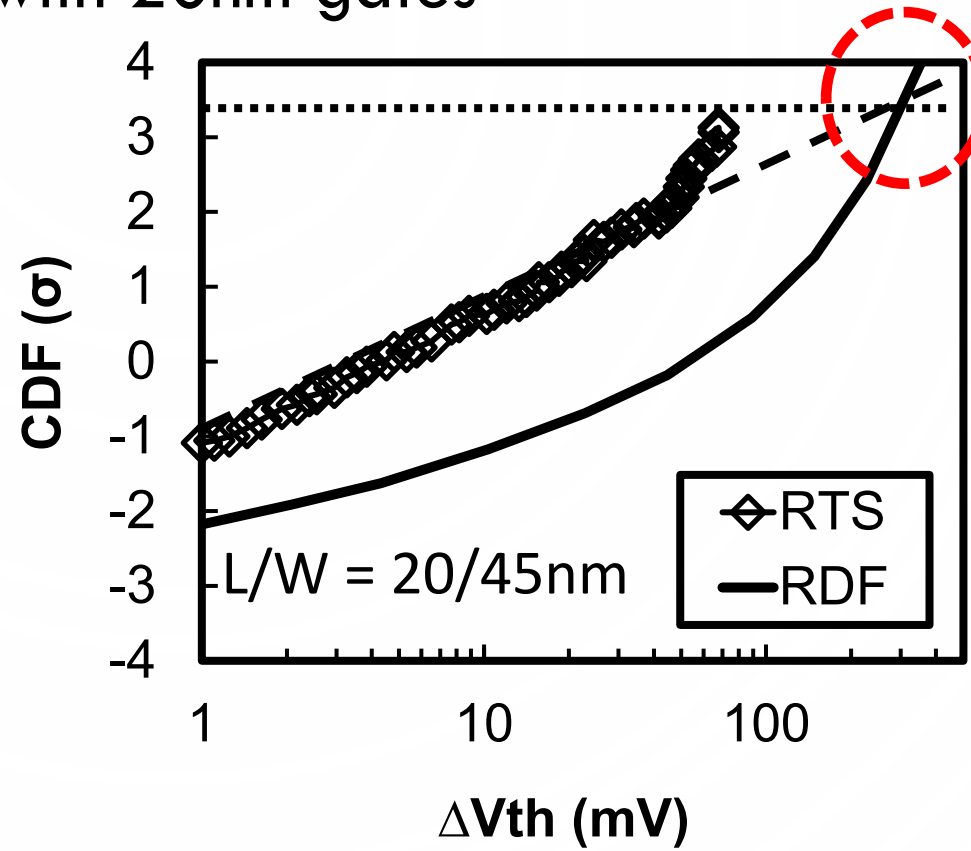
N. Tega et al, IRPS 2008.

# RTS and Technology Scaling

- RTS exceeds RDF at 3 sigma with 20nm gates

$$\Delta V_{th, RTS} \sim \frac{1}{WL}$$

$$\Delta V_{th, RDF} \sim \frac{1}{\sqrt{WL}}$$



Tega *et. al*, VLSI Tech. 09

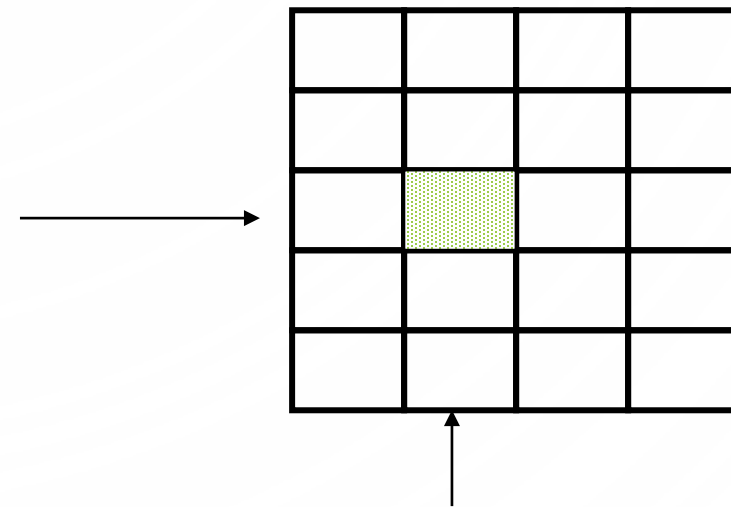
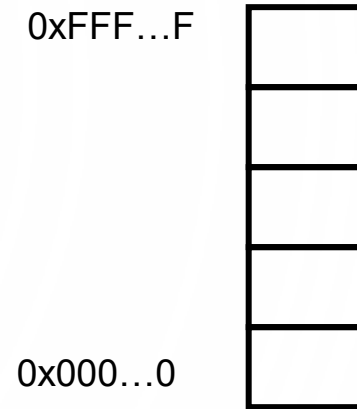


# Memory



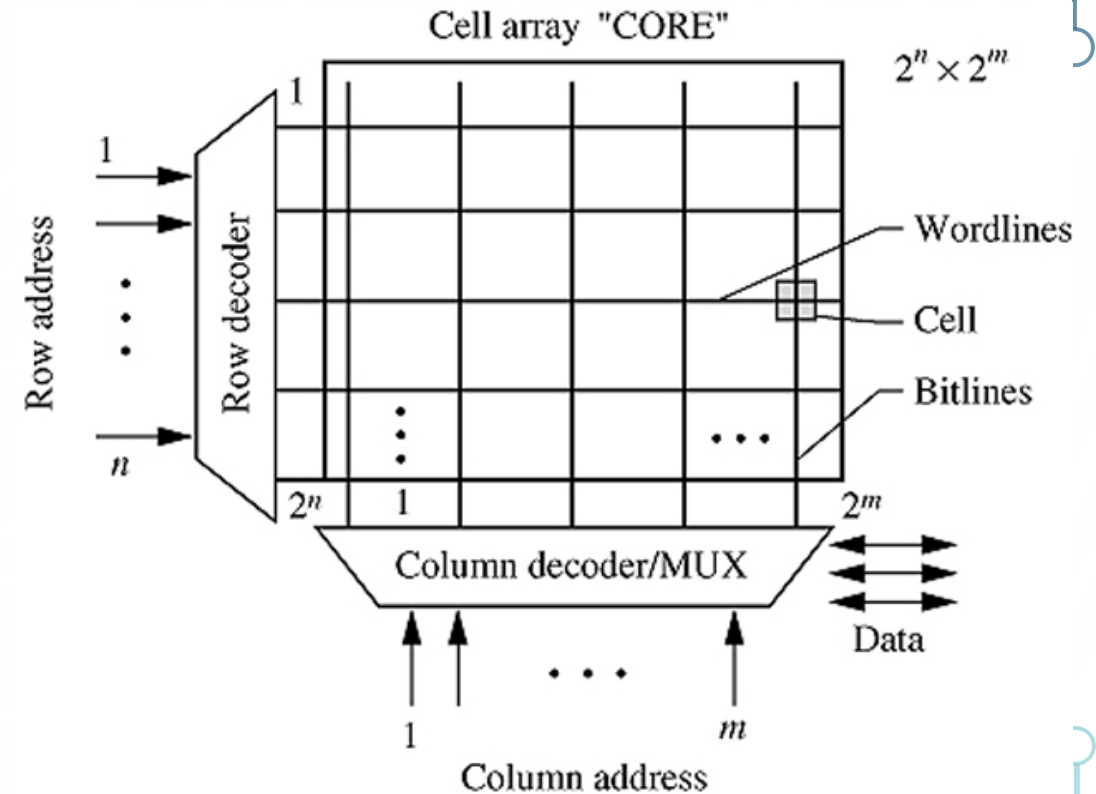
# Random Access Memory Architecture

- **Conceptual: Linear array of addresses**
  - Each box holds some data
  - Not practical to physically realize
    - millions of 32b/64b words
- **Create a 2-D array**
  - Decode Row and Column address to get data



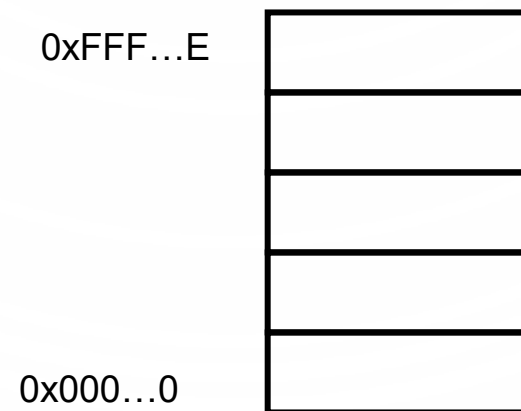
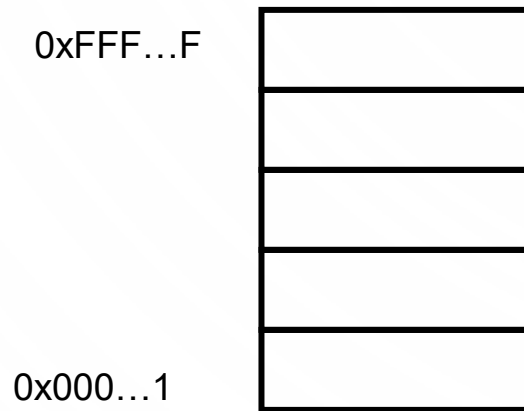
# Basic Memory Array (From 151/251A)

- Core
  - Wordlines to access rows
  - Bitlines to access columns
  - Data multiplexed onto columns
- Decoders
  - Addresses are binary
  - Row/column MUXes are 'one-hot' - only one is active at a time
- Important to optimize the aspect ratio to balance the delays

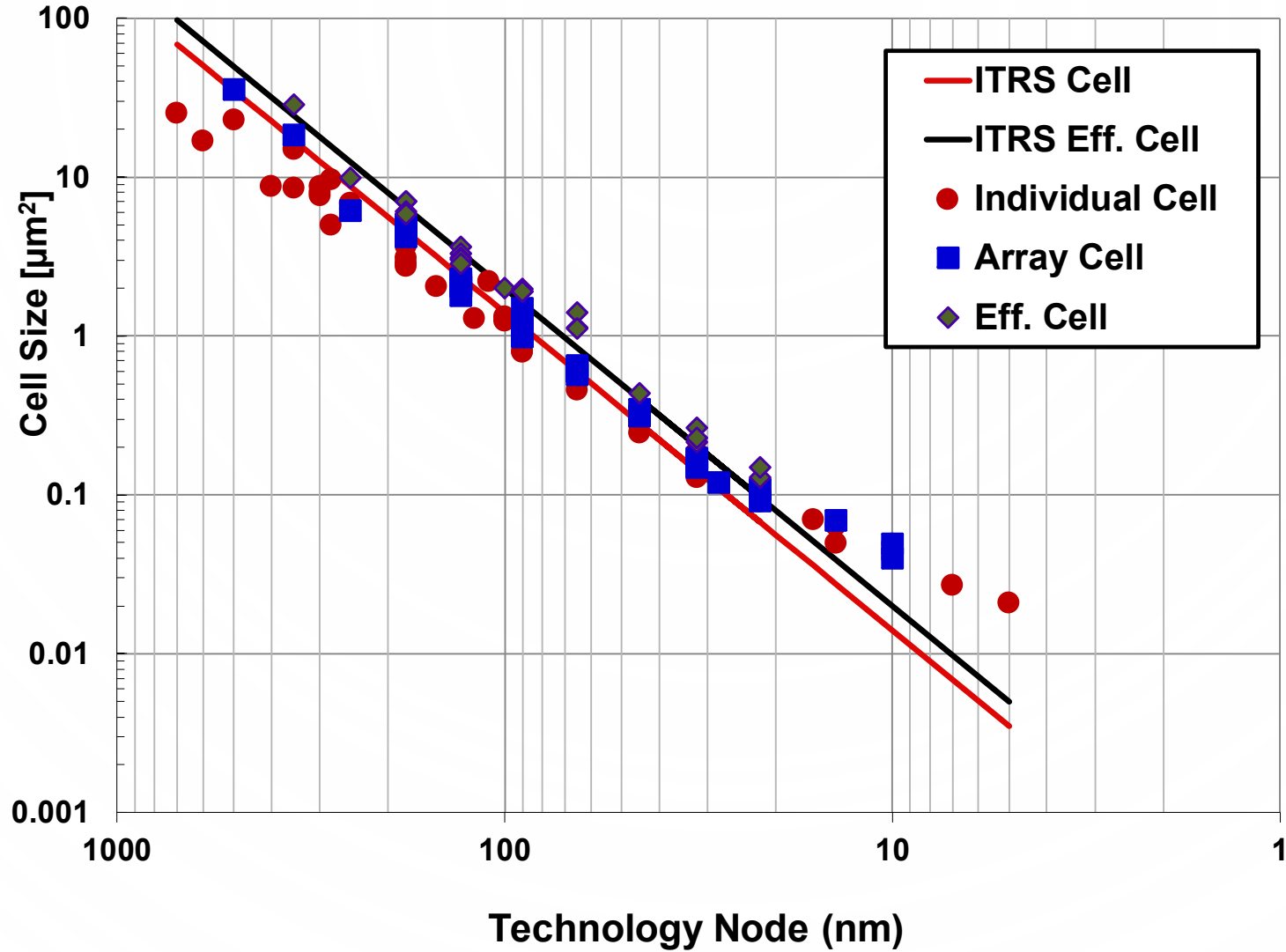


# Memory Banks

- Traditionally addressed by the LSB
  - Example two-bank memory
  - Odd and even banks

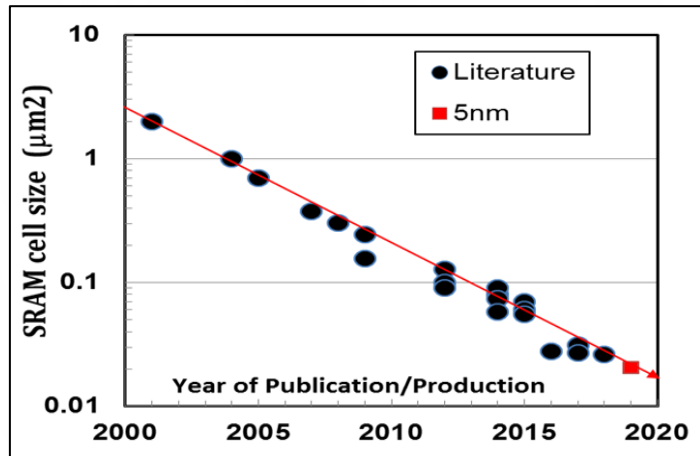


# SRAM Cell Trends

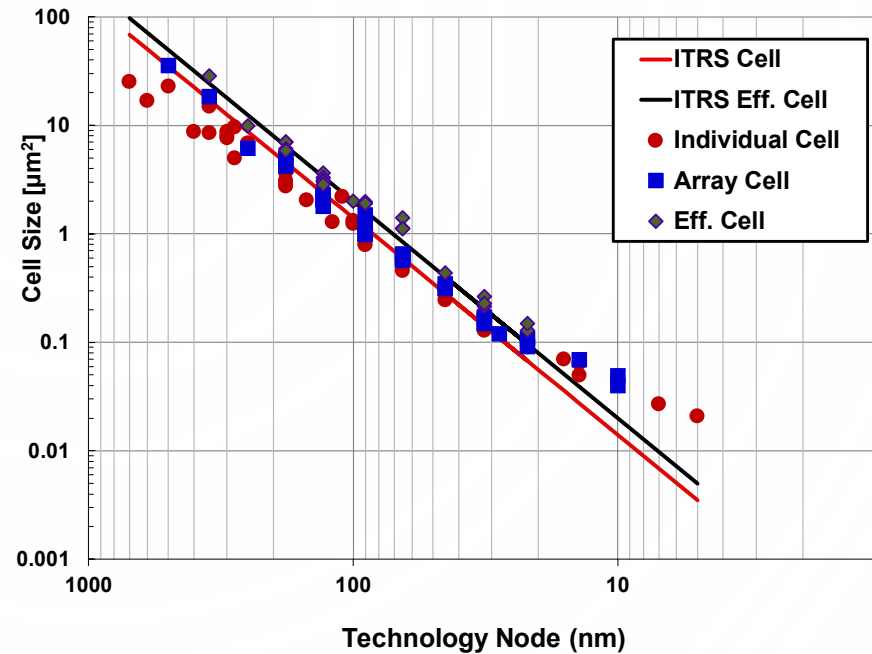


# SRAM Scaling or Not?

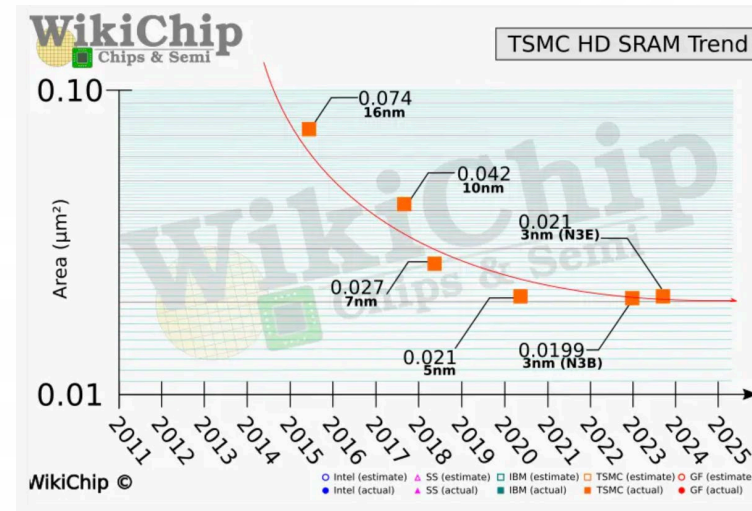
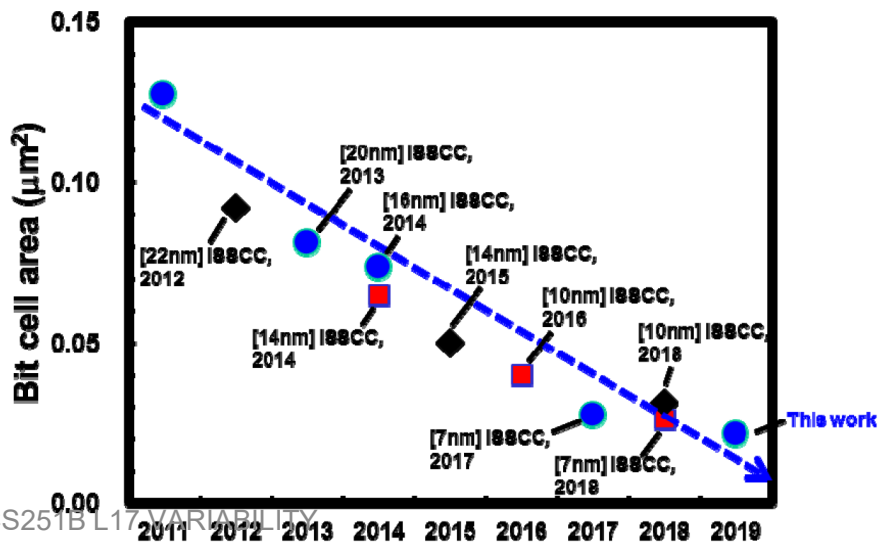
- TSMC at IEDM'19



- Bora's spreadsheet/WikiChip



- TSMC at ISSCC'20



# Summary

- Variability: Systematic and random
- Random, uncorrelated variations average out
- Identified random and systematic sources of variability

# Next Lecture

- **Memory**