EE241B : Advanced Digital Circuits

Lecture 10 – Variability

Borivoje Nikolić

3nm GAA SRAM bitcell design

SRAM bitcell margin

Samsung’s 3nm GAA SRAM (ISSCC’21)
Announcements

• Quiz today
• New labs posted
• New assignment this week
• Response to project proposals this week
  • Project web pages
Outline

• Variability

• Timing with variations
2. P 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\(^1\), post exposure bake (PEB) temperature\(^2\)
  - Dry etching: microscopic loading effects\(^3\)

---

\[^1\] J.Cain, M.S. Thesis, UC Berkeley
Layout: Proximity Test Structures

• 90nm experiments

  Single gate inverter layout

  Stacked gates

  Dummy poly

  L.T. Pang, VLSI’06

  Ring oscillators and individual transistor leakage currents

  No single gates allowed

  P1 min max2

  P2 mid1

  P3 max1

  P4 max2

  L.T. Pang, CICC’08

• 45nm experiments
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 Stress

- **Poly-Si Gate**
- **Spacer**
- **CESL layer**

- **Vertical compressive strain**
- **Parallel tensile strain**

- **STI**
- **S/D Active**

- **Gate pitch**
- **Active length**

- **45nm STM process: Wafer rotated <100> - higher PMOS mobility**
- **NMOS strained through capping layer**
- **Subatmospheric STI – weak tensile stress**
Impact of Longer Diffusion in 45nm

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

![Graph showing RO Frequency, NMOS $I_{\text{LEAK}}$, and PMOS $I_{\text{LEAK}}$ distributions for different process corners (SS, TT, FF) with longer diffusion and 22 chips from 2 wafers.

- Faster chip
- Slower chip
- 22 chips from 2 wafers

---

EECS241B L10 VARIABILITY
Impact of Shallow Trench Isolation (STI)

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

![Graphs showing RO Frequency, NMOS $I_{\text{LEAK}}$, and PMOS $I_{\text{LEAK}}$ with different conditions: SS, R1, NSTI, R1, and min.]

- R1
- NSTI

- Fastest chip
- Slowest chip

22 chips from 2 wafers
Impact of Correlations
Chip Yield Depends on Inter-Gate Correlation

Yield = Pr (sum of n delays < clock period)

ρ = 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 = \( \Pr (\text{max delay of } K \text{ paths} < \text{clock period}) \)

- \( K = 1 \) gives highest yield

Correlated paths reduce impact of variation

Bowman et al, JSSC, Feb 2002.
2. P Design Variability
Some Random Effects
Random Dopant Fluctuations

- Number of dopants is finite

Frank, IBM J R&D 2002
Random Dopant Fluctuations

$L_g = 17\text{nm}, V_{DS} = 0.7\text{V}$

$\sigma_{VT} = 23\text{mV}$

$L_g = 11\text{nm}, V_{DS} = 0.7\text{V}$

$\sigma_{VT} = 52\text{mV}$
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
Oxide Thickness

- Systematic variations +
- Roughness in the Si./SiO2 interface
- Smaller effect than RDF

Asenov, TED’2002
Transistor Matching

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

![Graph showing $\sigma_{\text{VT}}$ vs. $1/\sqrt{WL}$](image)

- Pelgrom parameter $A_{\text{VT}}$
  - Scales with technology (EOT)

- $A_{\text{VT}}$ in FDSOI technology

- $\sigma_{\text{VT}}$ in nMOSFETs with $|V_D| = 0.9V$

- $A_{\text{VT}} = 1.4\text{mV}\cdot\mu\text{m}$

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/SiO2 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

RTS and Technology Scaling

- RTS exceeds RDF at 3 sigma with 20nm gates

\[ \Delta V_{th, \text{RTS}} \sim \frac{1}{WL} \]

\[ \Delta V_{th, \text{RDF}} \sim \frac{1}{\sqrt{WL}} \]

\[ \Delta V_{th} (\text{mV}) \]

CDF (\(\sigma\))

\[ L/W = 20/45\text{nm} \]

Tega et al., VLSI Tech. 09
3. Design for Performance

3.A Flip-Flop Timing
Example Clock System

**Diagram Description:**

- **SYSCLK**:
  - :M
  - :N

- **PLL**:
  - Connected to SYSCLK

- **Global Clock**:
  - **CLK 1**: Connected to Unit 1
  - **CLK 2**: Connected to Unit 2
  - **CLK 3**: Connected to Unit 3

- **Enable Units**:
  - **Enable_U1_h**:
  - **Enable_U3_h**:

- **Vdd**:
  - Connected to PLL

- **Vdd (always enabled)**:

- **Deskew**:

**Source:** Courtesy of IEEE Press, New York. © 2000
Clock Nonidealities

• Clock skew
  • Spatial variation in temporally equivalent clock edges; deterministic + random, $t_{SK}$

• Clock jitter
  • Temporal variations in consecutive edges of the clock signal; modulation + random noise
  • Cycle-to-cycle (short-term) - $t_{JS}$
  • Long-term - $t_{JL}$

• Variation of the pulse width
  • for level-sensitive clocking
Clock Skew and Jitter

- Both skew and jitter affect the effective cycle time
- Only skew affects the race margin, if jitter is from the source
  - Distribution-induced jitter affects both
Sources of clock uncertainty

1. Clock Generation
2. Devices
3. Interconnect
4. Power Supply
5. Temperature
6. Capacitive Load
7. Coupling to Adjacent Lines
Flip-Flop Parameters

Delays can be different for rising and falling data transitions
Latch Parameters

Delays can be different for rising and falling data transitions

Unger and Tan
Trans. on Comp.
10/86
Clock Constraints in Edge-Triggered Systems

\[ t_{CL} \leq (t_{CY} - t_{SK} - t_{JS}) - (t_{SU} + t_{CQ}) \]

\[ t_{CL} \geq t_{SK} + (t_{H} - t_{CQ}) \]

3. B Timing with Uncertainty/Variations
Pictorial View of Setup and Hold Tests

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time

Actual early AT

Actual late AT

Early slack

Late slack

0 or more switching(s) allowed

Data must be stable

Early RAT

Late RAT

Latest clock arrival time

Earliest clock arrival time (next cycle)

Hold time

Setup time
Handling of Across-Chip Variation

• Each gate has a range of delay: $[lb, ub]$
  • The lower bound is used for **early timing**
  • The upper bound is used for **late timing**

• This is called an **early/late split**

• Static timing obtains bounds on timing slacks
  • Timing is performed as one forward pass and one backward pass

![Setup test](image1)
![Hold test](image2)

ICCAD '07 Tutorial
Chandu Visweswariah
How is the Early/Late Split Computed?

- The best way is to take known effects into account during characterization of library cells
  - History effect, simultaneous switching, pre-charging of internal nodes, etc.
  - This drives separate characterization for early and late; this is the most accurate method
- Failing that, the most common method is derating factors
  - Example: Late delay = library delay * 1.05
    Early delay = library delay * 0.95
- The IBM way of achieving derating is LCD factors (Linear Combination of Delay) (FC=fast chip, SC=slow chip, see next page)
  - Late delay = $\alpha_L \times FC\_delay + \beta_L \times NOM\_delay + \gamma_L \times SC\_delay$
  - Early delay = $\alpha_E \times FC\_delay + \beta_E \times NOM\_delay + \gamma_E \times SC\_delay$
  - Across-chip variation is therefore assumed to be a fixed proportion of chip-to-chip variation for each cell type
IBM Delay Modeling*

At a given corner
late delay = intrinsic + systematic + random
early delay = intrinsic – systematic – random

*P. S. Zuchowski, ICCAD’04
The Problem with an Early/Late Split

- The early/late split is very useful
  - Allows bounds during delay modeling
  - Any unknown or hard-to-model effect can be swept under the rug of an early/late split
- But, it has problems
  - Additional pessimism (which may be desirable)
  - Unnecessary pessimism (which is never desirable)
How to Have Less Pessimism?

• Common path pessimism removal
• Account for correlations
• Credit for statistical averaging of random
Statistical Timing

• Deterministic

• Statistical
Statistical Max Operation

\[
A = a_0 + \sum_{i=1}^{n} a_i \Delta X_i + a_{n+1} \Delta R_a
\]

\[
B = b_0 + \sum_{i=1}^{n} b_i \Delta X_i + b_{n+1} \Delta R_b
\]

\[
\sigma_A = \sqrt{\sum_{i=1}^{n+1} a_i^2}
\]

\[
\sigma_B = \sqrt{\sum_{i=1}^{n+1} b_i^2}
\]

\[
\rho = \frac{\sum_{i=1}^{n} a_i b_i}{\sigma_A \sigma_B}
\]

\[
\theta \equiv \left( \sigma_A^2 + \sigma_B^2 - 2\rho \sigma_A \sigma_B \right)^{1/2}
\]

\[
t = \Phi \left[ \frac{a_0 - b_0}{\theta} \right]
\]

\[
E[max(A, B)] = a_0 t + b_0 (1 - t) + \theta \phi \left[ \frac{a_0 - b_0}{\theta} \right]
\]

\[
E[max(A, B)]^2 = (\sigma_A^2 + a_0^2) t + (\sigma_B^2 + b_0^2) (1 - t) + (a_0 + b_0) \theta \phi \left[ \frac{a_0 - b_0}{\theta} \right]
\]

** M. Cain, “The moment-generating function of the minimum of bivariate normal random variables,” American Statistician, May ’94, 48(2)
Unified View of Correlations

Independently random part

Spatially correlated part: within-chip distance-related correlation

Globally correlated part: chip-to-chip, wafer-to-wafer, batch-to-batch variation

\[ D = a_0 + \sum a_i \Delta X_i + a_r \Delta X_R \]
Spatial Correlation vs. Early/Late Split

Dependence on common virtual variables cancels out at the timing test

ICCAD '07 Tutorial  
Chandu Visweswaraniah
Next Lecture

• Latch timing
• Flip-flops