15. Nanoscale Design Issues

Jacob Abraham

Department of Electrical and Computer Engineering
The University of Texas at Austin

VLSI Design
Fall 2020

October 20, 2020

Ideal Transistor I-V

Shockley first-order transistor models

\[ I_{ds} = \begin{cases} 
0 & \text{if } V_{gs} < V_{t} \\
\beta (V_{gs} - V_{t} - \frac{V_{ds}}{2}) V_{ds} & \text{if } V_{ds} < V_{dsat} \\
\frac{\beta}{2} (V_{gs} - V_{t})^2 & \text{if } V_{ds} > V_{dsat}
\end{cases} \]

cutoff
linear
saturation
Ideal nMOS I-V Plot

180 nm TSMC process

- Ideal Models
  - $\beta = 155(W/L) \mu A/V^2$
  - $V_t = 0.4 V$
  - $V_{DD} = 1.8 V$

Simulated nMOS I-V Plot

180 nm TSMC process

BSIM3 3V3 SPICE models

- What differs?
  - Less ON current
  - No square law
  - Current increases in saturation
Velocity Saturation

- We assumed carrier velocity \( \propto \) E-field
  \[ \nu = \mu E_{lat} = \mu V_{ds}/L \]
- Carriers scatter off atoms
- Velocity reaches \( \nu_{sat} \)
  - Electrons: \( 6 - 10 \times 10^6 \) cm/s
  - Holes: \( 4 - 8 \times 10^6 \) cm/s
- Better model
  \[ \nu = \frac{\mu E_{lat}}{1 + \frac{E_{lat}}{E_{sat}}} \Rightarrow \nu_{sat} = \mu E_{sat} \]

Velocity Saturation I-V Effect

- Ideal transistor ON current increases with \( V_{DD}^2 \)
  \[ I_{ds} = \mu C_{ox} \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2} = \frac{\beta}{2} (V_{gs} - V_t)^2 \]
- Velocity-saturated ON current increases with \( V_{DD} \)
  \[ I_{ds} = C_{ox} W(V_{gs} - V_t)\nu_{max} \]
- Real transistors are partially velocity saturated
  - Approximate with \( \alpha \)-power law model
  \[ I_{ds} \propto V_{DD}^\alpha \]
  - \( 1 < \alpha < 2 \) determined empirically
### α-Power Model

\[
I_{ds} = \begin{cases} 
0 & V_{gs} < V_t \\
I_{dsat} \frac{V_{ds}}{V_{dsat}} & V_{ds} < V_{dsat} \\
I_{dsat} & V_{ds} > V_{dsat}
\end{cases}
\]

- \( I_{ds} \) is the drain current.
- \( V_{gs} \) is the gate-source voltage.
- \( V_{ds} \) is the drain-source voltage.
- \( V_{dsat} \) is the saturation voltage.
- \( I_{dsat} \) is the saturation current.

\[
I_{dsat} = P_c \frac{\beta}{2} (V_{gs} - V_t)^\alpha
\]

\[
V_{dsat} = P_\nu (V_{gs} - V_t)^{\alpha/2}
\]

α, β, \( P_c \) and \( P_\nu \) are parameters determined empirically from a curve-fit of I-V characteristics.

### Channel Length Modulation

- Reverse-biased p-n junctions form a **depletion region**
  - Region between n and p with no carriers
  - Width of depletion \( L_d \) region grows with reverse bias
  - \( L_{eff} = L - L_d \)
- Shorter \( L_{eff} \) = more current
  - \( I_{ds} \) increases with \( V_{ds} \)
  - Even in saturation
Channel Length Modulation I-V

\[ I_{ds} = \frac{\beta}{2} (V_{gs} - V_t)^2 (1 + \lambda V_{ds}) \]

\( \lambda = \text{channel length modulation coefficient} \)
- Not feature size
- Empirically fit to I-V characteristics

Body Effect

- \( V_t \): gate voltage necessary to invert channel
- Increases if source voltage increases because source is connected to the channel
- Increase in \( V_t \) with \( V_s \) is called the **body effect**

**Body Effect Model**

\[ V_t = V_{t0} + \gamma \left( \sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s} \right) \]

\( \phi_s = \text{surface potential} \) at threshold
- \( \phi_s = 2\nu_T \ln \left( \frac{N_A}{n_i} \right) \)
  - Depends on doping level \( N_A \)
  - As well as intrinsic carrier concentration \( n_i \)
- \( \gamma = \text{body effect coefficient} \)

\[ \gamma = \frac{t_{ox}}{\varepsilon_{ox}} \sqrt{2q\varepsilon SiN_A} = \frac{\sqrt{2q\varepsilon SiN_A}}{C_{ox}} \]
OFF Transistor Behavior

- What about current in cutoff?
- Simulated results don’t match measurements
- What differs?
  - Current doesn’t go to 0 in cutoff

Leakage Sources

- Subthreshold conduction
  - Transistors can’t abruptly turn ON or OFF
- Junction leakage
  - Reverse-biased PN junction diode current
- Gate leakage
  - Tunneling through ultrathin gate dielectric
- Subthreshold leakage is the biggest source of leakage in modern transistors

Subthreshold Leakage

- Subthreshold leakage is exponential with $V_{gs}$

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{n V_T}} \left( 1 - e^{-\frac{V_{ds}}{V_T}} \right), \quad I_{ds0} = \beta V_T^2 e^{1.8}$$

- $n$ is process dependent, typically 1.4 – 1.5
Other Leakage Sources

Drain-Induced Barrier Lowering (DIBL)
- Drain Voltage also affects $V_t$ ($V'_t = V_t - \eta V_{ds}$)
- High drain voltage causes subthreshold leakage to increase

Junction Leakage
- Reverse-biased p-n junctions have some leakage
  
  $$I_D = I_s \left( \frac{V_D}{e^{\frac{V_D}{V_T}} - 1} \right)$$
  
  $I_s$ depends on doping levels
  - As well as area and perimeter of diffusion regions
  - Typically $< 1 \text{ fA/\mu m}^2$

Gate Leakage
- Carriers may tunnel through very thin gate oxides

Preamted tunneling current (from Song, 2001)
Temperature Sensitivity

- Increasing temperature
  - Reduces mobility
  - Reduces $V_{t}$
- $I_{ON}$ decreases with temperature
- $I_{OFF}$ increases with temperature

So What?

- So what if transistors are not ideal?
  - They still behave like switches, and isn’t that enough for digital logic?
- But these effects matter for . . .
  - Supply voltage choice
  - Logical effort
  - Quiescent power consumption
  - Pass transistors
  - Temperature of operation
Parameter Variations

- Transistors have uncertainty in parameters
  - Process: $L_{eff}$, $V_t$, $t_{ox}$ of nMOS and pMOS
  - Vary around typical (T) values
- Fast (F)
  - $L_{eff}$: short
  - $V_t$: low
  - $t_{ox}$: thin
- Slow (S): opposite
- Not all parameters are independent for nMOS and pMOS

Environmental Variation

- $V_{DD}$ and Temperature also vary in time and space
- Fast:
  - $V_{DD}$: high
  - Temperature: low

<table>
<thead>
<tr>
<th>Corner</th>
<th>Voltage</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>1.98</td>
<td>0°C</td>
</tr>
<tr>
<td>T</td>
<td>1.8</td>
<td>70°C</td>
</tr>
<tr>
<td>S</td>
<td>1.62</td>
<td>125°C</td>
</tr>
</tbody>
</table>
Process Corners

- Process corners describe worst case variations
  - If a design works in all corners, it will probably work for any variation
- Describe corner with four letters (T, F, S)
  - nMOS speed
  - pMOS speed
  - Voltage
  - Temperature

Important Corners

Some critical simulation corners include

<table>
<thead>
<tr>
<th>Purpose</th>
<th>nMOS</th>
<th>pMOS</th>
<th>$V_{DD}$</th>
<th>Temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle time</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Power</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Subthreshold leakage</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>S</td>
</tr>
<tr>
<td>Pseudo-nMOS</td>
<td>S</td>
<td>F</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Variability

- **Variability**: Statistical relationship between design parameters and process parameters
  - Need the ability to accurately model the relationship and incorporate the behavior into simulation tools
  - Possible to compensate for the variability
  - Example: $L_{eff}$, $V_t$
  - Conductor thickness as a function of interconnect density
- Modeling deficiencies may make variability look like uncertainty
  - Example: circuit switching activity factor
Sources of Variations

- **Process Variations**
  - Dopant fluctuation
  - Oxide thickness
  - Gate work function
  - Line edge roughness
  - Chemical-Mechanical Polish (CMP)
  - Rapid-Thermal Anneal (RTA)
  - Layout proximity effect

- **Environment Variations**
  - Temperature
  - Voltage
  - Package stress, 3D Technologies

- **Temporal Variations**
  - Transistor Aging (e.g., BTI)
  - RTN induced Vmin fluctuation

Causes of Variations

- Minimum Feature Size
  - Human hair, 100 μm
  - Amoeba, 15 μm
  - Red blood cell, 7 μm
  - AIDS virus, 0.1 μm
  - Buckyball, 0.001 μm

Source: J. Kulkarni
Features Smaller than Wavelengths

What is drawn is not what is printed on silicon

Source: Raul Camposano, Synopsys

Optical Proximity Correction (OPC)

What you see is NOT what you get

Department of Electrical and Computer Engineering, The University of Texas at Austin
J. A. Abraham, October 20, 2020
Imperfect Process Control

- Neighboring shapes interfere with the desired shape at some location: results in pattern sensitivity
- This is predominantly in the same plane
- There will be some interference from buried features for interconnect

Source: T. Brunner, ICP 2003

Increasing Mask Complexity

Exacerbated by increasing use of resolution enhancement techniques (RETs)

Source: K. Nowka, IBM
**Line Edge Roughness**

- In the lithography process, dose of photons will fluctuate due to finite quanta
  - Shot noise
- There will be fluctuations in the photon absorption positions
  - Due to nanoscale impurities in the resist composition

- Poly lines subject to increasing line edge roughness (LER)
  - Impact: circuit delay and leakage power

**Random Dopant Fluctuations**

![Graph showing the mean number of dopant atoms versus technology node (nm)]
Dopant Atoms in Channel

Source: D. Frank et al., VLSI Tech. 1999
D. Frank, H. Wong, IWCE, 2000

> 200 mV $V_t$ shift

Leakage Variation due to Dopant Fluctuations

Source: K. Agarwal, VLSI 2006

> 200 mV $V_t$ shift translates to $\approx 100X$ increase in leakage
Other Sources of Variability
N. Rohrer, ISSCC 2006

- **Negative Bias Temperature Instability (NBTI)**
  - At high negative bias and elevated temperature, the p-MOS $V_t$ gradually becomes more and more negative – reducing p-channel current
  - Mechanism thought to be the breakdown of H-Si bonds at the Si/SiO$_2$ interface, creating surface traps and injecting positive H-related species into the oxide
  - Associated with the average NBTI shift, there are also random shifts – even identical use conditions result in mismatch shifts, due to random variations in the number and spatial distribution of the charges/interface states formed
- **Charge trapping and hot-carrier defect generation mechanisms**
  - Result in long-term $V_t$ shifts in both n- and p-channel devices
  - The long-term $V_t$ shifts are parameter variations which must be accounted for during circuit design

Fluctuation in Gate Oxide Thickness

- Gate oxide variations have an exponential effect on gate tunneling currents
- Impact on device threshold, but significantly less $V_t$ variation than due to random dopant fluctuations
- Recent advances in high-k gate dielectrics (Hafnium oxides) with metal gates have alleviated this problem

1.1 nm oxide: $\approx$ 6 atomic layers

*Source: K. Nowka, IBM*
**Variability due to Back-End Processing**

- Chemical/Mechanical Polishing (CMP)
- Introduces large systematic intra-layer interconnect thickness
- Additional inter-layer interconnect thickness effects as well
- Topography variations result in focus variation for lines – leading to width variations

![Chemical/Mechanical Polishing Diagram](image)

Source: K. Nowka, IBM

**Dynamic Temperature Variations**

Thermal Map – 1.5 GHz Itanium Chip

![Thermal Map](image)

Source: Intel Corporation, ISSCC 2005
Dynamic Voltage and Power Variations

Effect of Variations on Circuit Performance

- Ring oscillators used for performance monitoring
- Variations of 11% slower to 13% faster than mean on the same die

Source: Anne Gattiker, IBM
**Variation Effects in Real Chips**

Source: Kevin Nowka, IBM

- Multicore chip from IBM
  - Core-0 was found to be $\approx 15\%$ slower than other parts
  - Models predicted that all parts of the design are identical

---

**Variation in Other Circuit Elements**

Normalized capacitance distribution on a single layer

Source: C. Visweswariah

- This enormous variation has a significant impact on analog/RF design
- Industry “sweet spots” for analog design are $0.25\mu - 0/18\mu$
- High frequency RF designs forced to use much smaller dimensions
### Delay Impact of Variations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Delay Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEOL metal (Metal mistrack, thin/thick wires)</td>
<td>-10% → +25%</td>
</tr>
<tr>
<td>Environmental (Voltage islands, IR drop, temperature)</td>
<td>± 15%</td>
</tr>
<tr>
<td>Device fatigue (NBTI, hot electron effects)</td>
<td>± 10%</td>
</tr>
<tr>
<td>$V_t$ and $T_{ox}$ device family tracking</td>
<td>± 5%</td>
</tr>
<tr>
<td>Model/hardware uncertainty (Per cell type)</td>
<td>± 5%</td>
</tr>
<tr>
<td>N/P mistrack (Fast rise/slow fall, fast fall/slow rise)</td>
<td>± 10%</td>
</tr>
<tr>
<td>PLL/Clock Tree (Jitter, duty cycle, phase error)</td>
<td>± 10%</td>
</tr>
</tbody>
</table>

Requires $2^{20}$ timing runs or [-65%, +80%] guard band

Source: K. Kalafala, C. Visweswarah

---

### 6-T SRAM Bitcell Scaling

Source: Class notes from J. Kulkarni

---

Department of Electrical and Computer Engineering, The University of Texas at Austin
J. A. Abraham, October 20, 2020
Study of Variations in SRAMs

75% of the die area occupied by SRAMs
Key enabler for logic technology scaling

Intel EX Xeon server processor
- 18 Cores, 22nm
- 5.6 Billion transistors
- 45 MBytes of L3 cache
- 2.26 B transistors for 6T SRAM bitcells
- \( \approx 40\% \) total transistors in just L3 bit cells

Source: Class notes from J. Kulkarni

Statistical Static Timing Analysis (SSTA)

- Determine the circuit timing from the delays of components
- Path-based SSTA
  - Select representative set of critical paths from normal (static) timing analysis
  - Model the delay of each path as a function of random variables (the underlying sources of variation)
  - Predict the parametric yield curve, as well as generate diagnostics
- Generate set of path delay tests for manufacturing screen