7.1 Technology Scaling – Small is Beautiful

- New technology node every three years or so. Defined by minimum metal line width.
- All feature sizes, e.g. gate length, are ~70% of previous node.
- Reduction of circuit size by 2 – good for cost.
### International Technology Roadmap for Semiconductors, 1999 Edition

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM metal half pitch (nm)</td>
<td>180</td>
<td>130</td>
<td>100</td>
<td>70</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>MPU physical Lg (nm)</td>
<td>140</td>
<td>85</td>
<td>65</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>(T_{ox}) (nm)</td>
<td>1.5-1.8</td>
<td>1.5-1.9</td>
<td>1-1.5</td>
<td>0.8-1.2</td>
<td>0.6-0.8</td>
<td>0.5-0.6</td>
</tr>
<tr>
<td>(V_{DD})</td>
<td>1.5-1.8</td>
<td>1.2-1.5</td>
<td>0.9-1.2</td>
<td>0.6-0.9</td>
<td>0.5-0.6</td>
<td>0.3-0.6</td>
</tr>
<tr>
<td>(I_{on,HP}) (µA/µm)</td>
<td>750/350</td>
<td>750/350</td>
<td>750/350</td>
<td>750/350</td>
<td>750/350</td>
<td>750/350</td>
</tr>
<tr>
<td>(I_{off,HP}) (nA/µm)</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>160</td>
</tr>
<tr>
<td>(I_{on,LP}) (µA/µm)</td>
<td>490/230</td>
<td>490/230</td>
<td>490/230</td>
<td>490/230</td>
<td>490/230</td>
<td>490/230</td>
</tr>
<tr>
<td>(I_{off,LP}) (pA/µm)</td>
<td>7</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>

- \(V_{dd}\) is reduced at each node to contain power consumption in spite of rising transistor density and frequency
- \(T_{ox}\) is reduced to raise \(I_{on}\) for speed consideration

No known solutions
**Excerpt of 2003 ITRS Scaling to 2016**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Node (nm)</td>
<td>90</td>
<td>65</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>HP physical Lg (nm)</td>
<td>37</td>
<td>25</td>
<td>18</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>EOT (nm) (HP/LSTP)</td>
<td>1.2/2.1</td>
<td>0.9/1.6</td>
<td>0.7/1.3</td>
<td>0.6/1.1</td>
<td>0.5/1.0</td>
</tr>
<tr>
<td>Vdd (HP/LSTP)</td>
<td>1.2/1.2</td>
<td>1.1/1.1</td>
<td>1.1/1.0</td>
<td>1.0/0.9</td>
<td>0.9/0.8</td>
</tr>
<tr>
<td>Ion/W,HP (mA/mm)</td>
<td>1100</td>
<td>1510</td>
<td>1900</td>
<td>2050</td>
<td>2400</td>
</tr>
<tr>
<td>Ioff/W,HP (mA/mm)</td>
<td>0.05</td>
<td>0.07</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Ion/W,LSTP (mA/mm)</td>
<td>440</td>
<td>510</td>
<td>760</td>
<td>880</td>
<td>860</td>
</tr>
<tr>
<td>Ioff/W,LSTP (mA/mm)</td>
<td>1e-5</td>
<td>1e-5</td>
<td>6e-5</td>
<td>8e-5</td>
<td>1e-4</td>
</tr>
</tbody>
</table>

- HP: High Performance Logic
- LSTP: Low Standby Power Technology
7.2 Subthreshold Current

- The leakage current that flows at $V_g < V_t$ is called the subthreshold current.

![Graph showing current vs. voltage]

- The current at $V_{gs} = 0$ and $V_{ds} = V_{dd}$ is called $I_{off}$.

*Intel, T. Ghani et al., IEDM 2003*

90nm technology.
Gate length: 45nm for NMOS, 50nm for PMOS
• Subthreshold current $\propto n_s$ (surface inversion carrier concentration)
• $n_s \propto e^{q\phi_s/kT}$

In subthreshold, $\phi_s = \text{constant} + V_g/\eta$

In a capacitor network:

$$\frac{d\phi_s}{dV_g} = \frac{C_{oxe}}{C_{oxe} + C_{dep}} = \frac{1}{\eta}$$
Subthreshold Leakage Current

\[ I_{ds} \propto n_s \propto e^{q\phi_s / kT} \propto e^{q(\text{constant} + V_{gs} / \eta) / kT} \propto e^{qV_{gs} / \eta kT} \]

\[
\begin{align*}
C_{ox} & \quad V_G \\
C_{dep} & \quad \phi_s \\
\end{align*}
\]

\[ I_{ds} \propto e^{qV_{gs} / \eta kT} \]

\[ \eta = 1 + \frac{C_{dep}}{C_{oxe}} \]

- Subthreshold current changes 10x for \( \eta \cdot 60 \text{mV} \) change in \( V_g \).
  
  Reminder: 60mV is \((\ln 10) \cdot kT/q\)

- Subthreshold swing, \( S \): the change in \( V_{gs} \) corresponding to 10x change in subthreshold current. \( S = \eta \cdot 60 \text{mV} \), typically 80-100mV
**Subthreshold Leakage Current**

- Practical definition of $V_t$: the $V_{gs}$ at which $I_{ds} = 100\text{nA} \times \frac{W}{L}$

$$
I_{\text{subthreshold}}(nA) \approx 100 \times \frac{W}{L} \times e^{q(V_g - V_t) / \eta kT} = 100 \times \frac{W}{L} \times 10^{(V_g - V_t) / S}
$$

$I_{\text{off}}(nA) = 100 \times \frac{W}{L} \times 10^{-V_t / S}$ is determined only by $V_t$ and subthreshold swing.
Subthreshold Swing ($S$)

- Smaller $S$ is desirable (lower $I_{\text{off}}$ for a given $V_t$). Minimum possible value of $S$ is 60mV/dec.

- How do we lower swing?
  - $S = 60 \text{mV} \cdot \left(1 + \frac{C_{\text{dep}}}{C_{\text{oxe}}} \right)$
  - Thinner $T_{\text{ox}}$ => larger $C_{\text{oxe}}$
  - Lower substrate doping => smaller $C_{\text{dep}}$
  - Lower temperature

- Limitations
  - Thinner $T_{\text{ox}}$ – oxide breakdown reliability or oxide leakage current
  - Lower substrate doping – doping is not a free parameter but set by $V_t$. 
7.3 $V_t$ Roll-off

- $V_t$ roll-off: $V_t$ decreases with decreasing $L_g$.
- It determines the minimum acceptable $L_g$ because $I_{off}$ is too large if $V_t$ becomes too small.

*Question:* Why data is plotted against $L_g$, not $L$?

Answer: $L$ is difficult to measure. $L_g$ is. Also, $L_g$ is the quantity that manufacturing engineers can control directly.

---

K. Goto et al., (Fujitsu) IEDM 2003  65nm technology. EOT=1.2nm, $V_{dd}=1$V
Why Does \( V_t \) Decrease with \( L \)? – Potential Barrier Concept

- When \( L \) is small, smaller \( V_g \) is needed to reduce the barrier to 0.2V, i.e. \( V_t \) is smaller.
- \( V_t \) roll-off is greater for shorter \( L \)
Energy-Band Diagram from Source to Drain

- $L$ dependence

- $V_{ds}$ dependence
**$V_t$ Roll-off -- Capacitance Network Model**

As the channel length is reduced, drain to channel distance is reduced; therefore $C_d$ increases

\[ V_{ds} \text{ helps } V_{gs} \text{ to invert the surface, therefore} \]

\[ V_t = V_{t-long} - V_{ds} \cdot \frac{C_d}{C_{oxe}} \]

\[ V_t = V_{t-long} - (V_{ds} + 0.4) \cdot \frac{C_d}{C_{oxe}} \]

Due to built-in potential between N-channel and N$^+$ drain & source

- 2-D Poisson Eq. solution shows that $C_d$ is an exponential function of $L$.

\[ V_t = V_{t-long} - (V_{ds} + 0.4) \cdot e^{-L/\lambda_d} \]

where \[ I_d \approx 3 \sqrt{T_{ox} W_{dep} X_j} \]
Vertical dimensions ($T_{ox}$, $W_{dep}$, $X_{j}$) must be scaled to support $L$ reduction

\[ V_t = V_{t-long} - (V_{ds} + 0.4) \cdot e^{-L/L_d} \]

where \( I_d \approx 3 \sqrt{\frac{T_{ox} W_{dep} X_{j}}{}} \)
7.4 Reducing the Gate Insulator Thickness and $T_{oxe}$

- Oxide thickness has been reduced over the years from 300nm to 1.2nm.
- Why reduce oxide thickness?
  - Larger $C_{ox}$ to raise $I_{on}$
  - Reduce subthreshold swing
  - Control $V_t$ roll-off
- Thinner is better. However, if the oxide is too thin
  - Breakdown due to high electric field
  - Leakage current
Tunneling Leakage Current

- For SiO$_2$ films thinner than 1.5nm, tunneling leakage current has become the limiting factor.
- HfO$_2$ has several orders lower leakage for the same EOT.
Replacing $\text{SiO}_2$ with $\text{HfO}_2$---High-$k$ Dielectric

- $\text{HfO}_2$ has a relative dielectric constant ($k$) of ~24, six times large than that of $\text{SiO}_2$.
- For the same EOT, the $\text{HfO}_2$ film presents a much thicker (albeit a lower) tunneling barrier to the electrons and holes.
- $\text{Toxe}$ can be further reduced by introducing metal-gate technology since the poly-depletion effect is eliminated.

(After W. Tsai et al., IEDM'03)
Challenges of High-K Technology

• The challenges of high-k dielectrics are
  – chemical reactions between them and the silicon substrate and gate,
  – lower surface mobility than the Si/SiO₂ system
  – too low a $V_t$ for P-channel MOSFET (as if there is positive charge in the high-k dielectric).

• A thin SiO₂ interfacial layer may be inserted between Si-substrate and high-k film.

Question: How can $T_{inv}$ be reduced?

(Answer is in Sec. 7.4 text)
7.5 How to Reduce $W_{dep}$

- $W_{dep}$ can be reduced by increasing $N_{sub}$

$$V_t = V_{fb} + 2\phi_B + \frac{\sqrt{qN_{sub}2\varepsilon_s2\phi_B}}{C_{ox}} = V_{fb} + 2\phi_B + \frac{2\varepsilon_s2\phi_B}{C_{ox}W_{dep}}$$

  - If $N_{sub}$ is increased, $C_{ox}$ should be increased in order to keep $V_t$ the same.
  - $W_{dep}$ can be reduced in proportion to $T_{ox}$.

- Or use retrograde doping with very thin lightly doped surface layer
  - Also, less impurity scattering in the inversion layer $\Rightarrow$ higher mobility
7.6 Shallow Junction Technology

- The shallow junction extension helps to control $V_t$ roll-off.
- Shallow junction and light doping combine to produce an undesirable parasitic resistance that reduces the precious $I_{on}$.
- Theoretically, metal S/D can be used as a very shallow “junction”.

7.7 Trade-off between $I_{on}$ and $I_{off}$

- Higher $I_{on}$ goes hand-in-hand with larger $I_{off}$ -- think L, $V_t$, $Tox$, $V_{dd}$.
- Figure shows spread in $I_{on}$ (and $I_{off}$) produced by intentional variation in $L_g$ and unintentional manufacturing variances in $L_g$ and other parameters.

Intel, T. Ghani et al., IEDM 2003
Circuit Techniques to Relax the Conflict between $I_{on}$ and $I_{off}$

- Substrate (well) bias
  - Only some circuit blocks need to operate at high speed.
  - Can use reverse well bias to raise the $V_t$ for the rest.
  - This techniques can also reduce the chip-to-chip and block-to-block variations with intelligent control circuitry.
- Multiple $V_t$ or $V_{dd}$
  - Lower $V_t$ or higher $V_{dd}$ are used only in the blocks that need speed
- Alternative MOSFET structures.
7.8 More Scalable Device Structures

- Vertical Scaling is important. For example, reducing $T_{ox}$ gives the gate excellent control of Si surface potential.
- But, the drain could still have more control than the gate along another leakage current path that is some distance below the Si surface. (Right figure.)
2-D Potential Profile

- The drain voltage can pull the potential barrier down and allow leakage current to flow along a submerged path.
Ultra-Thin-Body (UTB) MOSFET

- MOSFET built on very thin silicon film on an insulator (SiO$_2$).
- Since the silicon film is very thin, perhaps less than 10nm, no leakage path is very far from the gate.
New Structure I--Ultra-Thin-Body MOSFET

- The subthreshold leakage is reduced as the silicon film is made thinner.

\[ T_{\text{ox}} = 1.5 \text{nm}, \quad N_{\text{sub}} = 1 \times 10^{15} \text{cm}^{-3}, \quad V_{\text{dd}} = 1 \text{V}, \quad V_{\text{gs}} = 0 \]
Preparation of Silicon-on-Insulator (SOI) Substrate

- Initial Silicon wafer A and B
- Oxidize wafer A to grow SiO2
- Implant hydrogen into wafer A
- Place wafer A, upside down, over wafer B.
- A low temperature annealing causes the two wafers to fuse together.
- Apply another annealing step to form H2 bubbles and split wafer A.
- Polish the surface and the SOI wafer is ready for use.
- Wafer A can be reused.
Due to the high cost of SOI wafers, only some microprocessors, which command high prices and compete on speed, have embraced this technology.

In order to benefit from the UTB concept, Si film thickness must be aggressively reduced to ~ Lg/4
New Structure II--Multi-Gate MOSFET and FinFET

- The second way of eliminating deep leakage paths is to provide gate control from more than one side of the channel.
- The Si film is very thin so that no leakage path is far from one of the gates.
- Because there are more than one gates, the structure may be called multi-gate MOSFET.
FinFET

- One multi-gate structure, called FinFET, is particularly attractive for its simplicity of fabrication.
- Called FinFET because its silicon body resembles the back fin of a fish.
- The channel consists of the two vertical surfaces and the top surface of the fin.

Question: What is the channel width, W?
Answer: The sum of twice the fin height and the width of the fin.
**FinFET Process Flow**

1. **SOI Substrate**
2. **Fin Patterning**
3. **Si Fin**
4. **BOX**
5. **Si$_3$N$_4$ Spacer**
6. **Gate Etch**
7. **Spacer Formation**
8. **Poly Gate Deposition/Litho**
9. **Resist**
10. **Poly**
11. **NiSi**
12. **S/D Implant + RTA**
13. **Silicidation**
Variations of FinFET

- **Tall FinFET** has the advantage of providing a large $W$ and therefore large $I_{on}$ while occupying a small footprint.
- **Short FinFET** has the advantage of less challenging lithography and etching.
- **Nanowire FinFET** gives the gate even more control over the silicon wire by surrounding it.
Tall FinFET with $L_g=10\text{nm}$

B. Yu et al., IEDM 2002
Nanowire FinFET

5nm Gate Length
F. Yang et al., 2004 VLSI Tech Symp.
**Device Simulation and Process Simulation**

**Device Simulation**
- Commercially available computer simulation tools can solve all the equations presented in this book simultaneously with few or no approximations.
- Device simulation provides quick feedback about device design before long and expensive fabrication.

**Process Simulation**
- Inputs to process simulation: lithography mask pattern, implantation dose and energy, temperatures and times for oxidization and annealing steps, etc.
- The process simulator generates a 2-D or 3-D structures with all the deposited or grown and etched thin films and doped regions.
- This output may be fed into a device simulator as input together with applied voltages.
Example of Device Simulation---

Density of Inversion Charge in the Cross-Section of a FinFET Body

- The inversion layer has a significant thickness ($T_{\text{inv}}$).
- There are more subthreshold inversion electrons at the corners.

C.-H. Lin et al., 2005 SRC TECHCON
I-V of a Nanowire “Multi-Gate” MOSFET

C.-H. Lin et al., 2005 SRC TECHCON

Semiconductor Devices for Integrated Circuits (C. Hu)
Example of Process Simulation

• FinFET Process

The small figures only show 1/4 of the complete FinFET - the quarter farthest from the viewer.

Manual, Taurus Process, Synoposys Inc.
7.9 Output Conductance

- \( I_{\text{dsat}} \) does NOT saturate in the saturation region, especially in short channel devices!
- The slope of the \( I_{\text{ds}} - V_{\text{ds}} \) curve in the saturation region is called the **output conductance** \( (g_{\text{ds}}) \),

\[
g_{\text{ds}} = \frac{dI_{\text{dsat}}}{dV_{\text{ds}}}\]

- A smaller \( g_{\text{ds}} \) is desirable for a large voltage gain, which is beneficial to analog and digital circuit applications.

![Graph showing the relationship between \( I_{\text{ds}} \) and \( V_{\text{ds}} \) for different \( V_{\text{gs}} \) values.](image)
Example of an Amplifier

- The bias voltages are chosen such that the transistor operates in the saturation region. A small signal input, $v_{in}$, is applied.

$$i_{ds} = g_{msat} \cdot v_{gs} + g_{ds} \cdot v_{ds}$$

$$= g_{msat} \cdot v_{in} + g_{ds} \cdot v_{out}$$

$$i_{ds} = -v_{out} / R.$$ 

$$v_{out} = \frac{-g_{msat}}{(g_{ds} + 1/R)} \times v_{in}$$

- The voltage gain is $g_{msat}/(g_{ds} + 1/R)$.
- A smaller $g_{ds}$ is desirable for large voltage gain.
- Maximum available gain is $g_{msat}/g_{ds}$
What Parameters Determine the $g_{ds}$?

$$g_{ds} = \frac{dI_{dsat}}{dV_{ds}} = \frac{dI_{dsat}}{dV_{t}} \cdot \frac{dV_{t}}{dV_{ds}}$$

$$\frac{dI_{dsat}}{dV_{t}} = -\frac{dI_{dsat}}{dV_{gs}} = -g_{msat} \quad \text{and} \quad \frac{dV_{ds}}{dV_{ds}} = e^{-L/l_{d}}$$

$I_{dsat}$ is a function of $V_{gs} - V_{t}$ (From Eq. 7.3.3, $V_{t} = V_{t-long} - V_{ds} \cdot e^{L/l_{d}}$)

$$g_{ds} = g_{msat} \times e^{-L/l_{d}}$$

Max voltage gain ($R \rightarrow \infty$) = $\frac{g_{msat}}{g_{ds}} = e^{L/l_{d}}$

- A larger $L$ or smaller $l_{d}$, i.e. smaller $Tox$, $W_{dep}$, $X_{j}$, can increase the maximum voltage gain.
- The cause is “$V_{t}$ dependence on $V_{ds}$” in short channel transistors.
For large $L$ and $V_{ds}$ close to $V_{dsat}$, another mechanism may be the dominant contributor to $g_{ds}$. That is **channel length modulation**.

A voltage drop, $V_{ds} - V_{dsat}$, is dissipated over a finite distance next to drain, causing the “channel length” to decrease. More with increasing $V_{ds}$.

$$g_{ds} = \frac{l_d \cdot l_{dsat}}{L(V_{ds} - V_{dsat})}$$

$$l_d \approx 3 \sqrt{T_{ox} W_{dep} X_j}$$

$V_{c} = V_{dsat}$
7.10 MOSFET Modeling for Circuit Simulation

- For circuit simulation, MOSFETs are modeled with analytical equations.
- Device model is the link between technology/manufacturing and design/product. The other link is design rules.
- Circuits are designed A. through circuit simulations or B. using cell libraries that have been carefully designed beforehand using circuit simulations.
- **BSIM (Berkeley Short-channel IGFET Model)** is the world’s first industry standard MOSFET model. It contains all the equations presented in these chapters.
Examples of BSIM Model Results

![Graphs showing semiconductor device characteristics](image)
Example of BSIM Model Results
Example of BSIM Model Results