Scaling of MOSFETs

Siva Theja M
Department of Electrical Engineering
IIT Madras
Advisor: Prof Amitava Das Gupta

Indo German Winter Academy – 2006
11th December 2006, Digha
Overview

• MOS Capacitor and MOSFET
• Mobility Models
• Subthreshold Conduction
• Scaling
• Problems arising during Scaling
MOS Capacitor

- Accumulation
- Depletion
- Inversion
- Strong inversion
MOS Capacitor

• Threshold voltage

\[ V_T = \Phi_{ms} - \frac{Q_i}{C_i} - \frac{Q_d}{C_i} + 2\phi_F \]

• Substrate bias effect or body effect

\[ \Delta V_{th} = \gamma (\sqrt{2\phi_B} - V_{BS} - \sqrt{2\phi_B}) \]

\[ \gamma = \frac{\sqrt{2\varepsilon_s qN_A}}{C_{ox}} \]

Gamma is called body effect parameter.
MOSFET
MOSFET

• Gradual channel approximation

\[ I_{D, \text{sat}} = \frac{\varepsilon_{OX} \mu Z}{2 d_{OX} L_G} (V_{GS} - V_{th})^2 \]

\[ g_{m, \text{sat}} = \frac{dI_{D, \text{sat}}}{dV_{GS}} = \frac{\varepsilon_{OX} \mu Z}{d_{OX} L_G} (V_{GS} - V_{th}) \]
Mobility and Electric Fields

- Dependence of mobility on gate voltage, Transverse electric field
- Universal mobility Degradation curve

\[ \mu = \frac{\mu_0}{1 + \theta(V_G - V_T)} \]
Mobility and Electric Fields

• Dependence of mobility on drain voltage or longitudinal electric field
• Saturation of velocity
• $I_D$ would be linearly dependent on $V_G$ and not quadratically for short channel MOSFETs

\[
I_{D,\text{sat}} = \frac{\varepsilon_{OX}}{d_{OX}} \left( V_{GS} - V_{th} \right) v_{sat} Z
\]

\[
g_{m,\text{sat}} = \frac{dI_{D,\text{sat}}}{dV_{GS}} = \frac{\varepsilon_{OX}}{d_{OX}} v_{sat} Z
\]
Subthreshold Characteristics

\[ I_D = \mu(C_d + C_{it}) \frac{Z \left( \frac{kT}{q} \right)^2}{L} \left( 1 - e^{-\frac{qV_D}{kT}} \right) \left( e^{\frac{q(V_G - V_T)}{c_r kT}} \right) \]

- \(C_i\) is gate capacitance
- \(C_d\) is depletion capacitance in the channel
- \(C_{it}\) is the interface state capacitance

\[ c_r = \left[ 1 + \frac{C_d + C_{it}}{C_i} \right] \]
Subthreshold Characteristics

$S$: Subthreshold slope

Small $S$, better switch

$$S = 2.3 \frac{kT}{q} \left[ 1 + \frac{C_d + C_{it}}{C_i} \right]$$

$S$ depends on

- Gate thickness
- Channel doping
- No. of interface states

Trade off in choosing $V_{th}$
Scaling

- Introduction to scaling
- Scaling in MOSFETS
- Why scaling?
- Moore’s Law
Why Scaling?

<table>
<thead>
<tr>
<th>Processor name</th>
<th>Year of introduction</th>
<th>Transistors</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004</td>
<td>1971</td>
<td>2,250</td>
</tr>
<tr>
<td>8008</td>
<td>1972</td>
<td>2,500</td>
</tr>
<tr>
<td>8080</td>
<td>1974</td>
<td>5,000</td>
</tr>
<tr>
<td>8086</td>
<td>1978</td>
<td>29,000</td>
</tr>
<tr>
<td>286</td>
<td>1982</td>
<td>120,000</td>
</tr>
<tr>
<td>386</td>
<td>1985</td>
<td>275,000</td>
</tr>
<tr>
<td>486 DX</td>
<td>1989</td>
<td>1,180,000</td>
</tr>
<tr>
<td>Pentium</td>
<td>1993</td>
<td>3,100,000</td>
</tr>
<tr>
<td>Pentium II</td>
<td>1997</td>
<td>7,500,000</td>
</tr>
<tr>
<td>Pentium III</td>
<td>1999</td>
<td>24,000,000</td>
</tr>
<tr>
<td>Pentium 4</td>
<td>2000</td>
<td>42,000,000</td>
</tr>
</tbody>
</table>

This table shows the year of introduction and the number of transistors on the processors of Intel Corporation.
Let’s Start

Scaling of a MOSFET:

**Side view**

**Top view**

Scaling factor $\xi$
Types of Scaling

- Constant voltage scaling
- Purely geometrical
- Only lateral dimensions are reduced
- Longitudinal electric field increases
- Can cause break down
Consequences of Constant Voltage Scaling

• Drain current $I_D$ is unchanged
• Transconductance is unaffected
• Capacitance $C_G$ decreases as $k^2$
• Charging time $\tau$ decreases as $k^2$
• So, higher clock speeds can be obtained

$$I_{D,\text{sat}} = \frac{\varepsilon_{OX} \mu Z}{2d_{OX} L_G} (V_{GS} - V_{th})^2$$

$$g_{m,\text{sat}} = \frac{dI_{D,\text{sat}}}{dV_{GS}} = \frac{\varepsilon_{OX} \mu Z}{d_{OX} L_G} (V_{GS} - V_{th})$$
Consequences of Constant Voltage Scaling

The following table illustrates performance increase as linewidths shrink.

<table>
<thead>
<tr>
<th>Year</th>
<th>Linewidth</th>
<th>Clockspeed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning of 1970s</td>
<td>10 µm</td>
<td>1 MHz</td>
</tr>
<tr>
<td>End of 1970s</td>
<td>3 µm</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Beginning of 1980s</td>
<td>2 µm</td>
<td>20 MHz</td>
</tr>
<tr>
<td>End of 1980s</td>
<td>0.8 µm</td>
<td>50 MHz</td>
</tr>
<tr>
<td>Beginning of 1990s</td>
<td>0.5 µm</td>
<td>100 MHz</td>
</tr>
<tr>
<td>End of 1990s</td>
<td>0.25 µm</td>
<td>750 MHz</td>
</tr>
<tr>
<td>Beginning of 2000s</td>
<td>0.13 µm</td>
<td>2 GHz</td>
</tr>
</tbody>
</table>

Table shows typical linewidths and clock speeds versus year.
Short Channel Effects

- No longer resembles planar capacitor
Short Channel Effects

- Change in threshold voltage
- Clear pinch-off of channel
- Increased output impedance
- Increased leakage current (why and it’s consequences)
Channel Length Modulation

\[ I_D = I_{Do} \left( 1 + \lambda V_{DS} \right) \]

\( \lambda \): Channel length modulation factor
Solution?

- Scale all dimensions (lateral dimensions also) so that initial field is ‘recreated’
Constant Field Scaling

- All the lengths ($L_G, Z, d_{OX}$) and voltages ($V_{DS}, V_{GS}, V_{th}$) are scaled by same factor $k$
- Electric field unchanged
- Punch through effect
- Sol: Increase doping of acceptor by same factor, $k$ (Scaling of Depletion widths)
Short Channel Effect

• $V_T$ decreases with length for very small geometries
• Charge sharing between source/drain and gate
• Roll off
• Change junction depth
Short Channel Effect

Figure 6-45
Short channel effect in a MOSFET. Cross-sectional view of MOSFET along the length showing depletion charge sharing (colored regions) between the gate, source and drain.
Narrow width Effect

- $V_T$ goes up as width is reduced
- Effective depletion charge increased rather than decreased
Narrow width Effect

Figure 6-46
Roll-off of $V_T$ with decreasing channel length, and increase of $V_T$ with decreasing width.

Figure 6-47
Narrow width effect in a MOSFET. Cross-sectional view of MOSFET along the width, showing additional depletion charge (colored regions) underneath the field or the LOCOS isolation regions.
Consequences of Constant Field Scaling

• Assuming $V_{GS} = 0$

\[
I_{D, \text{sat}} = \frac{\varepsilon_{OX} \mu Z}{d_{OX} L_G} \frac{1}{2} V_{DS, \text{sat}}^2 = \frac{\varepsilon_{OX} \mu Z}{2 d_{OX} L_G} (V_{GS} - V_{th})^2
\]

\[
g_{m, \text{sat}} = \frac{d I_{D, \text{sat}}}{d V_{GS}} = \frac{\varepsilon_{OX} \mu Z}{d_{OX} L_G} (V_{GS} - V_{th})
\]
Consequences of Constant Field Scaling

- $I_D$ decreases by the factor $k$
- Transconductance is unchanged
- Static power dissipation of one transistor (IV) is scaled by $k^2$
- Dynamic power consumption ($P \tau f$) is scaled by $k^2$
- Power per unit area is unchanged
- Dynamic power per transistor area is unchanged
- Delay ($C_G / I_D$) is unchanged
- PDP is scaled by $k^2$
But

- We want newer generation of scaled down transistors to work faster
- So, supply voltage is scaled by and $\sqrt{k}$ not $k$
- Higher electric fields, Currents and so, lower delay times
- This increases dynamic power consumption per unit area
Yet another problem

- Power per device is actually less than $k^2$
- Consequently, power per unit area increases
- Circuits operate slower at high temperatures, and have reduced reliability and shorter lifetimes

Solution: Cooling fins and chip fans.
Saturated Velocity Model

- Saturated velocity model for current
- Same results
Hot Electron Effects

- High electric field
- High kinetic energy
- Hot electrons
- Gate current, input impedance
- Oxide charges, interface states
- Increase of $V_T$
- Substrate current, impact ionization, output impedance
- Lightly Doped Drain (LDD)
Hot Electron Effects

Figure 6-42
Hot carrier degradation in MOSFETs. The linear region transfer characteristics before and after hot carrier stress indicate an increase in $V_t$ and decrease of transconductance (or channel mobility) due to hot electron damage. The damage can be due to hot electron injection into the gate oxide which increases the fixed oxide charge, and increasing last interface state densities at the oxide-silicon interface (indicated by x).
Drain-Induced Barrier Lowering
Drain-Induced Barrier Lowering

- DIBL and punch-through
- Reverse-bias to substrate
- High channel doping
- Anti-punch through implant through out the channel
- Halo or pocket implants
Gate-Induced Drain Leakage

- Increase of sub-threshold currents for negative gate bias
- Narrow depletion regions
- Tunneling
- Drain doping should be $\sim 10^{18}$ cm$^{-3}$
Other factors

• Interconnect capacitance
• Process variations
• The transistor characteristics become less deterministic, but more statistical. This statistical variation increases design difficulty
## Other factors

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Constant Field Scaling</th>
<th>Constant Voltage Scaling</th>
<th>Constant Voltage Scaling with velocity saturation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate length</td>
<td>$L$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Gate width</td>
<td>$W$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Field</td>
<td>$E$</td>
<td>$1$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Oxide thickness</td>
<td>$t_{ox}$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Substrate doping</td>
<td>$N_a$</td>
<td>$\alpha^2$</td>
<td>$\alpha^2$</td>
<td>$\alpha^2$</td>
</tr>
<tr>
<td>Gate capacitance</td>
<td>$C_\sigma$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Oxide capacitance</td>
<td>$C_{ox}$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Transit time</td>
<td>$t_r$</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Transit frequency</td>
<td>$f_T$</td>
<td>$\alpha$</td>
<td>$\alpha^2$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Voltage</td>
<td>$V$</td>
<td>$1/\alpha$</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>Current</td>
<td>$I$</td>
<td>$1/\alpha$</td>
<td>$\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Power</td>
<td>$P$</td>
<td>$1/\alpha^2$</td>
<td>$\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Power-delay</td>
<td>$P \triangle t$</td>
<td>$1/\alpha^3$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
</tbody>
</table>
Acknowledgements and References

- Solid State Electronic Devices – Ben G. Streetman
- Semiconductor Devices – Amitava Das Gupta and Nandita Das Gupta
- The Omniscient Google
- Wikipedia
- Presentation by E.F. Schubert, Rensselaer Polytechnic Institute
- Special thanks to my adviser, Prof. A Das Gupta
Thank You
Questions?