CMOS DEVICE FUNDAMENTALS

- Outline
  - PMOS and NMOS transistors
  - NMOS device operation
  - PMOS device operation
  - Threshold Voltage and Body Effect
  - Short Channel Effects
  - Drain Induced Barrier Lowering
  - Channel Length Modulation
  - Sub-threshold conduction
  - Drawn and Effective Lengths
  - I-V characteristics
NMOS transistor

N Transistor

- Refer to gate, source, drain and bulk voltages as $V_g$, $V_s$, $V_d$ and $V_b$ respectively

- $V_{ab} = V_a - V_b$

- Device is symmetric. Drain and source are distinguished electrically
  - $V_d > V_s$ (reverse for PMOS device)
• All $p-n$ junction diodes must be reverse biased during operation (hence we need well or substrate ties)

• $p$ and $n$ regions are diffused regions
  – $p$ regions have acceptor (Boron) impurities. Majority carrier are holes
  – $n$ regions have donor (Phosphorus/Arsenic) impurities. Majority carriers are electrons.

• $n+$ and $p+$ are heavily doped $n$ and $p$ regions respectively

• Silicon Dioxide used for gate insulator material
  – Field oxide is the other type of oxide between substrate and a wire. It is usually much thicker than gate oxide. Why?

• Static gate current is zero
  – In fact gate acts as a control terminal
  – Current between drain and source modulated by gate voltage
  – Hence the entity that conducts is the area under the gate (ie the channel). The length of this channel is $L$ (since conduction is from source to drain).
PMOS transistor

- Dual of $n$ device
- CMOS (i.e., Complementary MOS) circuits use PMOS and NMOS devices
• Physically, these can be done in a n-well, p-well or twin-tub process

N transistor            P transistor

n+   n+   p+   p+       n   p

Twin-tub process
Enhancement NMOS device operation

- A depletion device conducts when $V_{gs} = 0$

- An enhancement device
  - Does not conduct at $V_{gs} = 0$
  - When $V_{gs}$ is “sufficiently” positive, majority carriers (ie electrons) are induced (enhanced) in the channel. This value of $V_{gs}$ is called the \textbf{threshold voltage}, $V_T$
  - This allows the channel to conduct
  - So the gate voltage modulates this conduction
Enhancement NMOS device operation...

- Assume $V_{ds} > 0$. When $V_{gs} = 0$, there is no current from D to S. Only reverse leakage current through the back-biased diode flows.

- When $V_{gs} > 0$, then positive charge is induced in the oxide. Electrons are attracted to the upper part of the channel.
  - In other words holes are repelled into the substrate
  - The surface of the p-substrate is depleted.
• If $V_{gs}$ is “large enough”, the upper part of the channel changes to n-type due to the enhancement of electrons in it. This is referred to as inversion, and the channel is called an n-channel. This channel is field-induced

– The voltage at which inversion occurs is called the **Threshold Voltage** ($V_T$).
- When $V_{gs} > V_T$, the inversion layer becomes thicker.

- The horizontal electrical field due to $V_{ds}$ moves electrons to the drain from the channel.

- But if $V_{ds} = 0$, then the channel is formed but does not conduct.

- Here is another view of the $n$ transistor in inversion with $V_{ds} = 0$.
source-drain
bulk or body contact
gate
source-drain
gate oxide
p substrate or well
depletion layer
channel (inversion layer)
Enhancement NMOS device operation...

- Consider $V_{gs} > V_T$, and $V_{ds} \leq V_{gs} - V_T$
  
  The inversion layer increases in thickness, and conduction increases. This is called the **linear** or **triode** region

- At the source end, $V_{gs}$ causes some induced inversion layer thickness. At the drain end, $V_{gd}$ causes a smaller induced inversion layer thickness since $V_{ds} > 0$ by definition. But $V_{gd} = V_{gs} - V_{ds} \geq V_T$, so the inversion layer thickness at the drain is not zero
- Hence the inversion layer is present below the entire gate as shown in the figure
- $I_{ds}$ depends on $V_T$, $V_{gs}$ as well as $V_{ds}$ in this region.
Enhancement NMOS device operation.4

- Consider $V_{gs} > V_T$, and $V_{ds} > V_{gs} - V_T$
  - At the source end, $V_{gs}$ causes some induced inversion layer thickness as before, since $V_{gs} - V_T$ is $> 0$
  - At the drain end, $V_{gd} = V_{gs} - V_{ds} < V_T$. So there is no inversion layer (and therefore no channel) under the drain!
  - The inversion layer is said to be pinched off. This is called the saturation region since $I_{ds}$ is independent of $V_d$. 

7
— If $V_{ds} = V_{gs} - V_T$, then the inversion layer is barely pinched off (at the drain end).

— Conduction occurs by drift. Electrons leave the channel and are accelerated towards the drain terminal, through the depletion region.

— $I_{ds}$ depends on $V_{gs}$ alone in this region.

— The effective channel has a voltage $V_{gs} - V_T$ applied across it. This is because if $V_{ds} = V_{gs} - V_T$, then the inversion layer is barely pinched off (at the drain end). If $V_{ds} > V_{gs} - V_T$, then the channel is pinched off somewhere between the drain and source ends. Therefore the effective channel has a voltage of $V_{gs} - V_T$ applied across it.
Enhancement NMOS device operation...

- So, there are three regions of conduction
  - **Cut-off**: \( V_{gs} < V_T \)
  - **Linear**: \( 0 < V_{ds} < V_{gs} - V_T \)
  - **Saturation**: \( 0 < V_{gs} - V_T < V_{ds} \)

- What does \( V_T \) depend on?
  - Gate and insulator materials, thickness of insulator, channel doping density - process dependent factors
  - \( V_{sb} \) (body effect), temperature (inversely) - operational factors
  - More on this later
Models for NMOS device operation

- **Cut-off**: $I_{ds} = 0$ (for now)
  when $V_{gs} < V_T$

- **Linear**: $I_{ds} = \beta([V_{gs} - V_T] V_{ds} - \frac{V_{ds}^2}{2})$
  when $0 < V_{ds} < V_{gs} - V_T$

- **Saturation**: $I_{ds} = \frac{\beta}{2}(V_{gs} - V_T)^2$ (for now)
  when $0 < V_{gs} - V_T < V_{ds}$
  This is obtained by using $V_{ds} = V_{gs} - V_T$ in the equation for linear $I_{ds}$ (see comment two pages prior to this one)

- Where $\beta = \frac{\mu \varepsilon}{t_{ox}} \left( \frac{W}{L} \right)$

- Note that $\beta$ depends on process dependent factors
  
  - $\beta \propto \mu$, the surface mobility of electrons in the channel. $\mu = \frac{\text{avg carrier drift velocity}}{\text{Electric field}} = \frac{v}{E}$. For PMOS, it is the surface mobility of holes in the channel
  
  - $\beta \propto \varepsilon$, the permittivity of the gate oxide $= k \cdot \varepsilon_0$, where $k \equiv$ dielectric constant and $\varepsilon_0 \equiv$ permittivity of free space
— $\beta \propto \frac{1}{t_{ox}}$, the oxide thickness. For a given process, this value is close to the minimum value so that the oxide does not break down due to excessive electric field. In other words $V_{DD}/t_{ox}$ is close to $E_{breakdown}$, the maximum sustainable electric field in the oxide material.

- Note that $\beta$ depends on geometric factors too (these are in a circuit designer’s control)
  
  — $\beta \propto W$, the width of the device. This is usually called device “size”. So we increase $W$ for a larger drive capability
  
  — $\beta \propto \frac{1}{L}$, the channel length. Due to processing limitations this has a minimum (minimum process feature size). Current minimums are $\sim 0.2\mu m$
  
  — Sometimes we intentionally choose $L > L_{min}$. A latch is a common example
Enhancement PMOS device operation

- Similar conceptual operation - just flip the signs

- In practice, $2\mu_{PMOS} \simeq \mu_{NMOS}$.

- So PMOS devices are usually twice the size (size usually means $W$, since $L$ is usually kept at a minimum anyway) of NMOS devices for the same drive capacity

- For this reason, we avoid PMOS devices as far as possible. Another reason is that $p$-type and $n$-type diffusion regions have stricter spacing requirements. So if we use only one type of diffusion, it results in smaller circuit area
Threshold Voltage and Body Effect

- Depends on factors like gate material, insulator material and thickness, channel doping.

- To modify it, change doping, or insulator material

- Our model assumed $V_{sb} = 0$

- In general, $V_T \simeq V_{T(0)} + \gamma \sqrt{V_{sb}}$, where $V_{T(0)}$ is the threshold voltage when $V_{sb} = 0$.

  where $\gamma = \left( \frac{t_{ox}}{\varepsilon_{ox}} \right) \sqrt{2q\varepsilon_SiN}$

  - $t_{ox}$ : oxide thickness
  - $\varepsilon_{ox}$ : dielectric constant of SiO$_2$
  - $q$ : electron charge
  - $\varepsilon_{Si}$ : dielectric constant of Si substrate
  - $N$ : doping density of substrate

  - Note that $\frac{\varepsilon_{ox}}{t_{ox}}$ is often called $C_{ox}$, the capacitance per unit area of an oxide capacitor.

- Circuits get slower due to body effect.
- Commonly occurs in “series-connected” devices since several devices in the series-connected stack may have $V_{sb} > 0$.

- Therefore, these devices exhibit a larger $V_T$, resulting in slower turn-on.

- Because of body effect, no more than 3-4 devices are connected in series in a digital design. Beyond this, the entire series connected chain slows down significantly.

- This restriction is alleviated in Silicon-on-Insulator (SOI) design. SOI processes exhibit close to zero $\gamma$ (more on this later).
Short Channel Effects

- For long channel lengths, $V_T$ depends on technology, and $V_{sb}$

- When $L$ is very small some earlier assumptions are not valid

- The depletion charge under the gate was assumed to be caused entirely by the vertical field

- Depletion regions near the source and drain (which are depleted due to the source and drain fields) were ignored

- So the region below the gate is already partially depleted

- Hence a smaller $V_{gs}$ suffices to create the inversion layer, and $V_T$ decreases with $L$ for short-channel devices

- $\Delta V_T = \frac{qNW_{m}X_J(\sqrt{1+2W_{m}/X_J}-1)}{C_{ox}L}$
  - $N$ is the doping density of the substrate
- $W_m$ is the maximum depletion thickness
- $X_J$ is the junction depth (this is a physical quantity)
- $C_{ox}$ is the oxide capacitance per unit area
- $L$ is the channel length

\[
\begin{align*}
\text{Long Channel } V_T & \\
\text{assumes low } V_{ds} &
\end{align*}
\]

- Lower this variation by using shallow junctions, thinner oxides and larger channel lengths.

- Usually it is hard to control, so designers are often forced to design around this problem.
Drain Induced Barrier Lowering

- When $V_{ds}$ is large, then the depletion region near the drain junction is wider.

- By the same argument as we made for short channel effects, a smaller $V_{gs}$ is required to create the inversion layer, and therefore $V_T$ decreases with $V_{ds}$.

- This is called Drain Induced Barrier Lowering (DIBL).

- If $V_d$ is large the source and drains can be shorted (punch-through).

\[ V_T = V_T^{V_{ds}=0} - \sigma V_{ds} \]

where $\sigma = \frac{6t_{ox}}{W_M} \exp\left(\frac{-\pi L}{4W_M}\right)$
• Bigger problem than short channel effects - since most devices in a design have same $L$, but $V_{ds}$ varies

• Example - DRAM cell exhibits increased leakage

- This is a single transistor (1-T) DRAM cell
- It is used in most DRAMs
- Write : put date on BL, assert WL
- Read : precharge BL to a intermediate voltage $V_{pch}$. Now assert WL. This causes charge redistribution between $C_T$ and $C_{BL}$, which is sensed or amplified using a circuit called a sense amplifier.

- Reading is destructive, so reassert value on BL after reading it
- Problem: when a transistor is not addressed, say that BL is high. This causes $V_T$ to drop, causing larger leakage of the charge in the cell,
or possible punch-through (i.e. destruction of charge in the cell).
Channel Length Modulation

- We earlier stated that in saturation,
  \[ I_{ds} = \frac{\beta}{2} (V_{gs} - V_T)^2 \]
  for \( 0 < V_{gs} - V_T < V_{ds} \)

- Actually, the saturation \( I_{ds} \) has a **weak** dependence on \( V_{ds} \)

- More accurately, in saturation,
  \[ I_{ds} = \frac{\beta}{2} (V_{gs} - V_T)^2 (1 + \lambda V_{ds}) \]
  for \( 0 < V_{gs} - V_T < V_{ds} \)

- \( \lambda \) is the **channel length modulation factor**. Typical value is 0.02 \( V^{-1} \)
Sub-threshold conduction

- We earlier stated that in cut-off, 
  \[ I_{ds} = 0 \]
  for \( V_{gs} < V_T \)

- Actually, there is a small electric field in the channel, resulting in a small diffusion current

- More accurately, in cut-off, 
  \[ I_{ds} \approx \frac{W}{T} I_{D0} e^{\frac{qV_g}{nkT}} \left[ e\left( -\frac{qV_s}{kT} \right) - e\left( -\frac{qV_d}{kT} \right) \right] \]
  for \( V_{gs} < V_T \), where

  - \( I_{D0} \) is experimentally derived

  - \( n \) is the subthreshold slope factor, derived experimentally

  - \( k \) is Boltzmann’s constant \((1.38 \times 10^{-23} \text{ J/}^o\text{K})\)

- This can be simplified since \( \frac{qV_s}{kT} \gg 1 \) and \( n \approx 1.5 \) in general
  \[ I_{ds} \approx \frac{W}{T} I_{D0} e^{\frac{qV_{gs}}{nkT}} \]
  for \( V_{gs} < V_T \)

- \( I_{ds} \) in this mode has an exponential dependence on \( V_{gs} \) actually, though the current value is small
• This is the source of standby power consumption in portable devices

• Some extremely low-power circuits exploit this region of operation (digital watches)
Drawn and Effective Channel lengths

- When a transistor is fabricated, its **effective** channel length $L_{\text{eff}}$ is different from the **drawn** length $L_{\text{drawn}}$

  - $L_{\text{eff}} = L_{\text{drawn}} - 2\Delta L_{\text{diff}} - 2\Delta L_{\text{poly}}$
    - The $\Delta L_{\text{diff}}$ factor is due to lateral diffusion of the source/drain junctions **under** the gate
    - $\Delta L_{\text{diff}} \sim 0.7X_I$, the source/drain **junction depth**!
    - The $\Delta L_{\text{poly}}$ factor is half the difference between $L_{\text{drawn}}$ and $L_{\text{final}}$

- The gate-diffusion overlaps results in a fixed capacitance (more on this later)
I-V Characteristics of NMOS/PMOS devices

\[ |V_{gs} - V_T| = |V_{ds}| \]

- Note that the use of absolute value of voltages and currents in the graph above accounts for the fact that PMOS signs are negative.
• Subthreshold (cut-off) region conduction not shown

• Qualitatively, the first quadrant graph corresponds to the NMOS device characteristic. The third quadrant graph corresponds to the PMOS device characteristic. Both these graphs “fold” into the graph on the previous page by taking the absolute values of the voltages and currents.
I-V Characteristics of NMOS device...

- In linear region:
  - Channel resistance = $R_{\text{cllin}} =$
    \[ \lim_{V_{ds} \to 0} \left( \frac{dI_{ds}}{dV_{ds}} \right)^{-1} \]
    \[ = \frac{1}{\beta(V_{gs} - V_T)} \]
  - Depends on $V_{gs}$
  - Transconductance = $g_m = \left( \frac{dI_{ds}}{dV_{gs}} \right) = \beta V_{ds}$
  - Higher current gain with higher $V_{ds}$

- In saturation region:
  - Transconductance = $g_m = \left( \frac{dI_{ds}}{dV_{gs}} \right) = \beta (V_{gs} - V_T)$
- Depends on $V_{gs}$ itself