Metal-Oxide-Semiconductor (MOS)

EL-314N

Dr. M Jawaid Siddiqui
MOS (Metal-Oxide-Semiconductor)

Assume work function of metal and semiconductor are same.
## MOS materials

<table>
<thead>
<tr>
<th>Material</th>
<th>$\Phi_M$</th>
<th>$\chi$</th>
<th>$E_g$</th>
<th>$K_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>4.1 eV</td>
<td>4.05 eV</td>
<td>1.12 eV</td>
<td>11.8</td>
</tr>
<tr>
<td>Al</td>
<td></td>
<td>0.95 eV</td>
<td>$\approx$ 8 eV</td>
<td>$\infty$</td>
</tr>
<tr>
<td>SiO$_2$</td>
<td></td>
<td></td>
<td></td>
<td>3.9</td>
</tr>
</tbody>
</table>
Shown is the semiconductor substrate with a thin oxide layer and a top metal contact, also referred to as the gate.
A second metal layer forms an Ohmic contact to the back of the semiconductor, also referred to as the bulk.
The structure shown has a p-type substrate.
We will refer to this as an n-type MOS capacitor since the inversion layer contains electrons.
To understand the different bias modes of an MOS we consider 3 different bias voltages.

1. below the flatband voltage, $V_{FB}$
2. between the flatband voltage and the threshold voltage, $V_T$, and
3. larger than the threshold voltage.

These bias regimes are called the accumulation, depletion, and inversion mode of operation.
Structure and principle of operation

- Charges in a MOS structure under accumulation, depletion and inversion conditions

\[ V_G < V_{FB} \]

\[ V_{FB} < V_G < V_T \]

\[ V_T < V_G \]
n-MOS transistor
Accumulation Mode

$V_{gs} = 0$

n-MOS transistor
Depletion Mode

$V_{gs} > 0$ and $V_{gs} < V_t$

deployment region

n-MOS transistor
Inversion Mode

$V_{gs} > V_t$

inversion region

deployment region
The four modes of operation of an MOS structure:

- Flatband,
- Depletion,
- Inversion and
- Accumulation.

Flatband conditions exist when no charge is present in the semiconductor so that the Si energy band is flat.

Surface depletion occurs when the holes in the substrate are pushed away by a positive gate voltage.

A more positive voltage also attracts electrons (the minority carriers) to the surface, which form the so-called inversion layer.

Under negative gate bias, one attracts holes from the $p$-type substrate to the surface, yielding accumulation.
MOS capacitor structure

- Polysilicon or metal gate
- Gate oxide
- p-substrate
- Mobile Hole
- $V_g$
MOS capacitor - accumulation

This is called accumulation – the surface is accumulated with majority carriers; it is made more p-type.
Accumulation occurs typically for -ve voltages where the -ve charge on the gate attracts holes from the substrate to the oxide-semiconductor interface.

Depletion occurs for positive voltages.

The +ve charge on the gate pushes the mobile holes into the substrate.

Therefore, the semiconductor is depleted of mobile carriers at the interface and a -ve charge, due to the ionized acceptor ions, is left in the space charge region.
MOS capacitor- flat band

\[ V_g = V_{fb} \sim 0 \]

Flat band – majority carrier density is *constant* and equal to dopant density
The voltage separating the accumulation and depletion regime is referred to as the flatband voltage, \( V_{FB} \).

The flatband voltage is obtained when the applied gate voltage equals the workfunction difference between the gate metal and the semiconductor.

If there is a fixed charge in the oxide and/or at the oxide-silicon interface, the expression for the flatband voltage must be modified accordingly.
MOS capacitor - depletion

\[ 0 < V_g < V_T \]

\[ V_T = \text{Threshold voltage} \]

Depletion:
- Majority carriers pushed away
- Fixed ions end gate charge field lines
- Depletion region | space charge region formed

p-substrate

\[ V > 0 \]

Fixed ions

Depletion

Electrons

\[ E_c \]

\[ E_f \]

\[ E_v \]

\[ V > 0 \]

\[ E_f \]
MOS capacitor - inversion

- Mobile electrons are induced at the surface
- Surface is inverted - it becomes n-type
- Threshold is when mobile charge density equals doping density
MOS capacitor- inversion

- Inversion occurs at voltages beyond the threshold voltage.
- In inversion, there exists a negatively charged inversion layer at the oxide-semiconductor interface in addition to the depletion-layer.
- This inversion layer is due to minority carriers, which are attracted to the interface by the positive gate voltage.
MOS Capacitance

- CV measurements of MOS capacitors provide a wealth of information about the structure, which is of direct interest when one evaluates an MOS process.
- Since the MOS structure is simple to fabricate, the technique is widely used.
- To understand CV measurements one must first be familiar with the frequency dependence of the measurement.
- This frequency dependence occurs primarily in inversion since a certain time is needed to generate the minority carriers in the inversion layer.
- Thermal equilibrium is therefore not immediately obtained.
MOS Capacitance

- Capacitance depends on frequency of applied signal.
- If speed of variation is slow enough so that electrons can be generated by thermal generation fast enough to be created in phase with applied signal, then $C_s$ is very large
  \[ C \approx C_{ox} \]
- If variation is too high a frequency, electron concentration remains fixed at the average value and capacitance depends on capacitance of depletion region
  \[ C < C_{ox} \]
Influence of gate on surface potential

\[ \kappa(kappa) = \frac{\partial \psi_s}{\partial V} = \frac{C_{ox}}{C_{ox} + C_{dep}} \]

\[ \psi_s = \text{Surface potential} \]
Gate-depletion capacitive divider

How does changing V change $\psi_s$?
1. $CV = Q$
2. Charge $Q$ on $\psi_s$ is constant
3. Change V, hold $Q$ constant

$$C_{ox} \left( \Delta V - \Delta \psi_s \right) = C_{dep} \Delta \Psi_s$$

$$C_{ox} \Delta V = (C_{ox} + C_{dep}) \Delta \Psi_s$$

$$\Delta \Psi_s = \Delta V \frac{C_{ox}}{C_{ox} + C_{dep}}$$
Capacitance in series

\[ C_{\text{Total}} = \frac{C_{\text{ox}} C_s}{C_{\text{ox}} + C_s} \]
The capacitance-voltage characteristics of this ideal MOS structure varies with the applied bias.

Under accumulation, the MOS structure appears almost like a parallel plate capacitor, dominated by the insulator properties,

\[ C_{ox} = \frac{\varepsilon_{ox}\varepsilon_0}{t_{ox}} \]

As the applied voltage becomes positive, a depletion layer capacitance \( C_d \) is added in series,

\[ C_d = \frac{\varepsilon_s\varepsilon_0}{x_d} \]
The total capacitance is,

$$C = \frac{C_{ox}C_d}{C_{ox} + C_d}$$

A very low frequency, the recombination-generation kinetics of electrons can vary in response to the voltage variations, therefore capacitance resembles that of the parallel plate capacitor.
CV Curve

\[ C = \frac{dQ}{dV} \]

Diagram showing a cross-section of a semiconductor device with labeled parts: gate, conductor, insulator, surface, and depletion region. The graph on the right shows the capacitance \( C \) as a function of gate voltage \( V_g \) with points \( C_{ox} \) and \( C_{dep} \) indicating changes in capacitance.
C-V characteristic of p-type Semiconductor
pMOS

C-V characteristic of n-type Semiconductor
Low frequency capacitance of an nMOS capacitor. Shown are the exact solution for the low frequency capacitance (solid line) and the low and high frequency capacitance obtained with the simple model (dotted lines). $N_a = 10^{17} \text{ cm}^{-3}$ and $t_{ox} = 20 \text{ nm}$. 
(a) The threshold voltage and the ideal MOS structure

(b) In practice, there are several charges in the oxide and at the oxide-semiconduct interface that effect the threshold voltage: $Q_{mi} =$ mobile ionic charge, $Q_{ot} =$ trapped oxide charge, $Q_f =$ fixed oxide charge, $Q_{it} =$ charge trapped at the interface.
Effects of Real Surfaces

- Practical MOS capacitors are usually made up of standard IC materials, poly-Si/SiO$_2$/Si.
- Due to variation in work function and Si-SiO$_2$ interface and trapped charges, the operation of the MOS capacitor deviates from the Ideal case.
- The work function of the semiconductor varies with doping level, therefore the work function potential difference $\Phi_{ms} = \Phi_m - \Phi_s$ also varies.
For negative $\Phi_{ms}$, the metal is positively charge with respect to the semiconductor at equilibrium. As a result, the energy bands bend downward.

Equilibrium $V = 0$

Flat-band $V = V_{FB} = F_{ms}$
The combined flat band voltage including the effects of difference in work functions and interface charges becomes,

$$V_{FB} = \Phi_{ms} - \frac{Q_i}{C_{ox}}$$

This flat band voltage should be added to the derivation of the threshold voltage earlier,

$$V_{TH} = \Phi_{ms} - \frac{Q_i}{C_{ox}} - \frac{Q_d}{C_{ox}} + 2\phi_F$$

The threshold voltage is the voltage first that is large enough to achieve flat band condition, then depletion and finally to induce the inverted region.
Charge Distribution

\[ \rho(x) \]

- Oxide
- Inversion
- Depletion

Metal
Key Definitions

\[ q \phi(x) = E_{i,\text{bulk}} - E_i(x) \]

\[ q \phi_s = E_{i,\text{bulk}} - E_{i,\text{surf}} \]

\[ q \phi_F = E_{i,\text{bulk}} - E_F \]
Potential Definition

\[ q \phi_s \]

\[ q \phi_F \]

oxide

surface

p-Si
Depletion Width

One-sided junction:

\[
W = \left( \frac{2K_s \varepsilon_o}{q} \frac{N_A + N_D}{N_A N_D} \left( \frac{1}{N_A} \right) \left( \frac{V_{bi}}{\phi_s} - \frac{V_A}{\phi_s} \right) \right)^{1/2}
\]

MOS capacitor:

\[
W = \left( \frac{2K_s \varepsilon_o}{q N_A} \phi_s \right)^{1/2}
\]
Gate Voltage (depletion case)
\[ C = \frac{C_o}{1 + \frac{K_o W}{K_s x_o}} \]

We want \( C \) vs \( V_G \). But \( W \) depends on \( \phi_s \) which is a complicated function of \( V_G \).

\[ C = \frac{C_o}{\sqrt{1 + \frac{v_G}{V_\delta}}} \]

where

\[ V_\delta = \frac{q K_s x_o^2}{2 K_o^2 \epsilon_o} N_A \]
\( \delta \)-depletion capacitance

\[
\frac{C}{C} = \frac{1}{(1-(V_G/V_\delta))^{1/2}}
\]
n-Si

The diagram illustrates the relationship between the gate voltage ($V_G$) and the threshold voltage ($V_T$) for an n-type silicon (n-Si) device. The graph shows three regions:

- **Inversion**: This region is characterized by a constant $V_T$ as $V_G$ increases from negative to positive values.
- **Depletion**: The curve shows the depletion of carriers as $V_G$ becomes more negative, approaching $V_T$.
- **Accumulation**: The graph transitions from depletion to accumulation as $V_G$ becomes more positive than $V_T$, indicating the formation of an inversion layer.
p-Si

The diagram illustrates the accumulation, depletion, and inversion regions of a p-type silicon (p-Si) semiconductor as a function of gate voltage ($V_G$). The curve shows how the voltage ($V_G$) affects the depletion region, leading to the transition from accumulation to depletion to inversion at a threshold voltage ($V_T$).
Exact capacitance

- **Flat band**
- **Inversion**
- **Depletion**

The graph shows the normalized capacitance ($C/C_0$) vs. voltage ($V$) with marked regions for flat band, inversion, and depletion.
C vs f

The diagram shows the variations of capacitance (C) with respect to gate voltage (V_G) for three regions: inversion, depletion, and accumulation. The capacitance is normalized to the oxide capacitance (C_0), with values ranging from 0.2 to 1.0 on the vertical axis.

The x-axis represents the gate voltage (V_G) in volts, with values ranging from -5 to 2.

Key markers include:
- Low-frequency region
- High-frequency region
- Inversion
- Depletion
- Accumulation
Calculated dependence of $C_{\text{mos}}$ on the applied voltage for different frequencies. Parameters used: insulator thickness, $2 \times 10^{-8}$ m; semiconductor doping density, $10^{15}$/cm$^3$; generation time, $10^{-8}$ s. Reproduced from Shur M. (1990) *Physics of Semiconductor Devices*, Prentice Hall, Englewood Cliffs, NJ
C vs scan rate

The graph shows the relationship between capacitance (C) and scan rate (R) for different values of R. The x-axis represents the gate voltage (V_G) in volts, and the y-axis represents the normalized capacitance (C/C_0). The graph includes curves for different values of R: 0, 0.035, 0.096, 0.26, and 2.6 V/sec. The curves indicate transitions from depletion to accumulation.
Parallel plate capacitance

- Formula for parallel plate capacitance:
  \[ C_{ox} = \varepsilon_{ox} / x_{ox} \]
  - Permittivity of silicon: \( \varepsilon_{ox} = 3.46 \times 10^{-13} \text{ F/cm}^2 \)
- Gate capacitance helps determine charge in channel which forms inversion region.
- Mobile electrons move if voltage applied between S and D
- \( Q_{\text{channel}} = CV \)
- \( C = C_g = \varepsilon_{ox} WL/x_{ox} = C_{ox} WL \)
- \( V = V_{gc} - V_t = (V_{gs} - V_{ds}/2) - V_t \)
An n-channel MOS transistor. The gate-oxide thickness, $T_{OX}$, is approximately 100 angstroms (0.01 μm). A typical transistor length, $L = 2 \lambda$. The bulk may be either the substrate or a well. The diodes represent pn-junctions that must be reverse-biased.
Parallel plate capacitance

- The channel and the gate form the plates of a capacitor, separated by an insulator - the gate oxide.
- We know that the charge on a linear capacitor, $C$, is
  \[ Q = C \ V \]
- The channel charge, $Q$. 
At lower plate, the channel, is not a linear conductor.
Charge only appears on the lower plate when the voltage between the gate and the channel, $V_{GC}$, exceeds the n-channel threshold voltage.
For nonlinear capacitor we need to modify the equation for a linear capacitor to the following:

$$Q = C(V_{GC} - V_t)$$
Parallel plate capacitance

- The lower plate capacitor is resistive and conducting current, so that the $V_{GC}$ varies.
- In fact, $V_{GC} = V_{GS}$ at the source and $V_{GC} = V_{GS} - V_{DS}$ at the drain.
- What we really should do is find an expression for the channel charge as a function of channel voltage and sum (integrate) the charge all the way across the channel, from $x = 0$ (at the source) to $x = L$ (at the drain).
- Instead we shall assume that the channel voltage, $V_{GC}(x)$, is a linear function of distance from the source and take the average value of the charge, which is

$$Q = C \left[ (V_{GS} - V_t) - 0.5 \ V_{DS} \right]$$
Parallel plate capacitance

- The gate capacitance, $C$, is given by the formula for a parallel-plate capacitor with length $L$, width $W$, and plate separation equal to the gate-oxide thickness, $T_{ox}$.
- Thus the gate capacitance is
  \[ C = \frac{(W L \varepsilon_{ox})}{T_{ox}} = W L C_{ox} \]
  
  where $\varepsilon_{ox}$ is the gate-oxide dielectric permittivity
  
  For SiO$_2$, $\varepsilon_{ox} \approx 3.45 \times 10^{-11}$ Fm$^{-1}$, so that, for a typical gate-oxide thickness of 100 Å ($= 10$ nm), the gate capacitance per unit area, $C_{ox} \approx 3$ fF$\mu$m$^{-2}$.
The channel charge of transistor

- The channel charge in terms of the transistor parameters
  \[ Q = WL \, C_{ox} \left[ (V_{GS} - V_t) - 0.5 \, V_{DS} \right] \]

- The drain–source current is
  \[ I_{DS} = \frac{Q}{t_f} \]
  \[ = \frac{(W/L) \, \mu_n \, C_{ox} \left[ (V_{GS} - V_t) - 0.5 \, V_{DS} \right] \, V_{DS}}{(W/L)k'_n \left[ (V_{GS} - V_t) - 0.5 \, V_{DS} \right] \, V_{DS}} \quad \text{(*)} \]

- The \( t_f \) is time of flight - sometimes called the **transit time** is the time that it takes an electron to cross between source and drain.

- \( \mu_n \) is the **electron mobility** (\( \mu_p \) is the **hole mobility**)
The channel charge of transistor

- The constant $k'_n$ is the process transconductance parameter (or intrinsic transconductance):
  \[ k'_n = \mu_n C_{ox} \]
- We also define $\beta_n$, the transistor gain factor (or just gain factor) as
  \[ \beta_n = k'_n \frac{W}{L} \]
- The factor $W/L$ is the transistor shape factor.
The channel charge of transistor

- Equation (*) describes the linear region (or triode region) of operation.
- This equation is valid until $V_{DS} = V_{GS} - V_t$ and then predicts that $I_{DS}$ decreases with increasing $V_{DS}$.
- At $V_{DS} = V_{GS} - V_t = V_{DS(sat)}$ (the saturation voltage) there is no longer enough voltage between the gate and the drain end of the channel to support any channel charge.
- Clearly a small amount of charge remains or the current would go to zero, but with very little free charge the channel resistance in a small region close to the drain increases rapidly and any further increase in $V_{DS}$ is dropped over this region.
- Thus for $V_{DS} > V_{GS} - V_t$ (the saturation region, or pentode region, of operation) the drain current $I_{DS}$ remains approximately constant at the saturation current, $I_{DS(sat)}$, where

$$I_{DS(sat)} = \left(\frac{\beta_n}{2}\right)(V_{GS} - V_t)^2; \quad V_{GS} > V_t \quad \text{..... (**)$$
Figure below shows the n-channel transistor $I-V$ characteristics for a generic 0.5 $\mu$m CMOS process that we shall call G5.

We can fit Eq.(**) to the long-channel transistor characteristics ($W = 60$ $\mu$m, $L = 6$ $\mu$m).

If $I_{DS}(\text{sat}) = 2.5$ mA (with $V_{DS} = 3.0$ V, $V_{GS} = 3.0$ V, $V_{t} = 0.65$ V, $T_{ox} = 100$ Å), the intrinsic transconductance is
MOS n-channel transistor characteristics for a generic 0.5 μm process (G5). (a) A short-channel transistor, with $W=6$ μm and $L=0.6$ μm (drawn) and a long-channel transistor ($W=60$ μm, $L=6$ μm) 
(b) The 6/0.6 characteristics represented as a surface. (c) A long-channel transistor obeys a square-law characteristic between $I_{DS}$ and $V_{GS}$ in the saturation region ($V_{DS} = 3$ V). A short-channel transistor shows a more linear characteristic due to velocity saturation. Normally, all of the transistors used on an ASIC have short channels.
MOS n-channel transistor characteristics for a generic 0.5 μm process (G5). (a) A short-channel transistor, with \( W=6 \) μm and \( L=0.6 \) μm (drawn) and a long-channel transistor (\( W=60 \) μm, \( L=6 \) μm) (b) The 6/0.6 characteristics represented as a surface. (c) A long-channel transistor obeys a square-law characteristic between \( I_{DS} \) and \( V_{GS} \) in the saturation region (\( V_{DS} = 3 \) V). A short-channel transistor shows a more linear characteristic due to velocity saturation. Normally, all of the transistors used on an ASIC have short channels.
MOS n-channel transistor characteristics for a generic 0.5 μm process (G5). (a) A short-channel transistor, with W=6 μm and L=0.6 μm (drawn) and a long-channel transistor (W=60 μm, L=6 μm) (b) The 6/0.6 characteristics represented as a surface. (c) A long-channel transistor obeys a square-law characteristic between $I_{DS}$ and $V_{GS}$ in the saturation region ($V_{DS} = 3$ V). A short-channel transistor shows a more linear characteristic due to velocity saturation. Normally, all of the transistors used on an ASIC have short channels.
The channel charge of transistor

- $k'_n = \frac{[2(L/W) I_{DS(sat)}]}{(V_{GS} - V_t)^2}$
  
  $= \frac{[2(6/60) (2.5 \times 10^{-3})]}{[(3.0 - 0.65)^2]}$
  
  $= 9.05 \times 10^{-5} \text{ AV}^{-2} \sim 90 \mu\text{AV}^{-2}$

- This value of $k'_n$, calculated in the saturation region, will be different (typically lower by a factor of 2 or more) from the value of $k'_n$ measured in the linear region.

- We assumed the mobility, $\mu_n$, and the threshold voltage, $V_t$, are constants.
The channel charge of transistor

For the p-channel transistor in the G5 process, \( I_{DS\text{(sat)}} = -850 \, \mu A \) (\( V_{DS} = -3.0 \, V \), \( V_{GS} = -3.0 \, V \), \( V_t = -0.85 \, V \), \( W = 60 \, \mu m \), \( L = 6 \, \mu m \)). Then

\[ k'_{\rho} = [2(L/W) \times I_{DS\text{(sat)}}]/(V_{GS} - V_t)^2 \]

\[ = [2(6/60) \times (850 \times 10^{-6})]/[(-3.0 - (-0.85))^2] \]

\[ = 3.68 \times 10^{-5} \, AV^{-2} \]
The source and drain of CMOS transistors look identical.

The source of an n-channel transistor is lower in potential than the drain and vice versa for a p-channel transistor.

In an n-channel transistor the threshold voltage, $V_t$, is normally positive, and the terminal voltages $V_{DS}$ and $V_{GS}$ are also usually positive.

In a p-channel transistor $V_t$ is normally negative and we have a choice: We can write everything in terms of the magnitudes of the voltages and currents or we can use negative signs in a consistent fashion.
P-channel MOS transistor

n-channel enhancement MOS

Drain Current ($I_{ds}$)

Assume source-to-drain voltage ($V_{ds}$) is fixed

$V_{tn}$

gate-to-source voltage ($V_{gs}$)

p-channel enhancement MOS

Drain Current ($I_{ds}$)

$-V_{tp}$

gate-to-source voltage ($V_{gs}$)
Here are the equations for a p-channel transistor using negative signs:

\[ I_{DS} = -k' \rho \frac{W}{L} \left[ (V_{GS} - V_t) - 0.5 V_{DS} \right] V_{DS} ; \quad V_{DS} > V_{GS} - V_t \]

\[ I_{DS(sat)} = -\frac{\beta}{2} \frac{V_{GS} - V_t}{2}^2 ; \quad V_{DS} < V_{GS} - V_t \]

In these two equations \( V_t \) is negative, and \( V_{DS} \) & \( V_{GS} \) are also normally negative (\( -3 \, \text{V} < -2 \, \text{V}, \) for example). The \( I_{DS} \) is then negative, corresponding to conventional current flowing from source to drain of a p-channel transistor (and hence the negative sign for \( I_{DS(sat)} \)).
MOSFET Capacitances

- Intrinsic capacitance: \( C_g = WLC_{ox} \quad C_g = \frac{2}{3} WLC_{ox} \)

- Parasitic capacitances:
  - Depletion capacitance \( C_d = \varepsilon_s WL / W_{dm} \)
  - Overlap capacitance
  - Junction capacitance \( C_j = \varepsilon_s Wd / W_{dj} = Wd \sqrt{\frac{\varepsilon_s q N_a}{2(\psi_{bi} + V_j)}} \)
Overlap Capacitance

$$C_{do} = WL_{ov} C_{ox} = \frac{\varepsilon_{ox} W L_{ov}}{t_{ox}}$$

Outer fringe

$$C_{of} = \frac{2 \varepsilon_{ox} W}{\pi} \ln \left(1 + \frac{t_{gate}}{t_{ox}}\right)$$

Inner fringe

$$C_{if} = \frac{2 \varepsilon_{si} W}{\pi} \ln \left(1 + \frac{x_j}{2t_{ox}}\right)$$