MOS Transistor Theory

- For designing VLSI circuits, designer must understand the performance of MOS technology. Proper understanding of MOS active devices examines various possibilities for configuring inverter circuits.
- After understanding manufacturing process of basic MOS circuits, their electrical characteristics can be studied. VLSI designers must have good knowledge of behavior of circuit they are designing. Even designing with computer aided design processes it is necessary to understand and verify performance specifications.
- Ratio rules, aspects of latch up are important for nMOS and CMOS, BiCMOS respectively. The basic building block using these devices is inverter. Once the expression for nMOS transistor is obtained, the equivalent expression for pMOS can be obtained by reversal of voltage and current polarities in the expression followed by exchange of $\mu_n$ and $\mu_p$ of electrons and hole.

2.1 $I_{ds}$ versus $V_{ds}$ Relationship

- The MOS transistors are voltage controlled device (as against the bipolar transistors are current controlled device). A voltage on the gate terminal induces a charge in the channel that exists between source and drain. The charge then move from source to drain under the influence of electric field generated by voltage $V_{ds}$ applied between drain and source.
- The charge induced is dependent on the gate to source voltage $V_{gs}$ (controlling factor), the current $I_{ds}$ is dependent on both $V_{gs}$ and $V_{ds}$. The relationship between these parameters can be developed.
- Consider a typical structure of nMOS transistor as shown in Fig. 2.1.
Fig. 2.1 nMOS transistor structure

Expression for Transit Time:

- The drain to source current $I_{ds}$ is given by -
  \[ I_{ds} = \frac{Q_c}{\tau} \]  
  ...(2.1)

  where,

  - $Q_c$ is charge induced in channel.
  - $\tau$ is electron transit time.

- The current $I_{ds}$ is due to electrons flowing from source to drain. The transit time for travel of electrons or hole from source to drain ($\tau_{sd}$) is given as -
  \[ \tau_{sd} = \frac{L}{v} \]  
  ...(2.2)

  where,

  - $L$ is length of channel
  - $v$ is velocity of electrons or holes.

- Velocity of electrons is given as -
  \[ v = \mu E_{ds} \]  
  ...(2.3)

  where,

  - $\mu$ is electron or hole mobility (surface)
  - $E_{ds}$ is electric field between drain to source.

- If $V_{ds}$ is voltage between drain and source
\[ E_{ds} = \frac{V_{ds}}{L} \]  

\[ v = \mu \cdot \frac{V_{ds}}{L} \]  

Putting in equation (2.2) we get,

\[ \tau_{sd} = \frac{L^2}{\mu V_{ds}} \]  

- The electron and hole mobilities at room temperature.
  \[ \mu_n = 650 \text{ cm}^2/\text{V sec (surface)} \]
  \[ \mu_p = 240 \text{ cm}^2/\text{V sec (surface)} \]

2.1.1 Non-saturated Region

- When device is operated in non-saturated region, the IR drop in the channel is same throughout the channel and can be taken as average value as \( \frac{V_{ds}}{2} \).

  \[ V_{ds} \] is voltage difference between gate and channel assuming substrate connected to channel.

- In non-saturated region, the effective gate voltage \( V_g \) is given by
  \[ V_g = V_{gs} - V_t \]  

  \[ V_t \] is threshold voltage needed to invert the charge under the gate and to establish the channel.

- The charge gets induced into the channel due to gate voltage and if \( E_g \) is the average electric field from gate to channel.

  The charge per unit area = \( E_g \varepsilon_{irs} \varepsilon_0 \)  

  \[ \varepsilon_{irs} \] is relative permittivity of insulation between gate and channel \( (= 4 \text{ for silicon oxide}) \)

  \[ \varepsilon_0 \] is permittivity of free space \( (8.85 \times 10^{-14} \text{ F/cm}) \)

- The total induced charge for the area of WL is given by,
  \[ Q_c = E_g \varepsilon_{irs} \varepsilon_0 WL \]  

... (2.4)

... (2.5)

... (2.6)

... (2.7)

... (2.8)

... (2.9)
Now,

\[ E_g = \frac{(V_{gs} - V_t) - \frac{V_{ds}}{2}}{D} \] ...(2.10)

where,

\( D \) is oxide thickness.

- Putting the expression for \( E_g \) -

\[ Q_c = \frac{\varepsilon_{\text{ins}} \varepsilon_0 WL}{D} \left[ (V_{gs} - V_t) - \frac{V_{ds}}{2} \right] \] ... (2.11)

- From equations (2.11), (2.6) and (2.1)

\[ I_{ds} = \frac{\varepsilon_{\text{ins}} \varepsilon_0 \mu L W}{D} \left[ (V_{gs} - V_t) - \frac{V_{ds}}{2} \right] V_{ds} \] ... (2.12)

\[ I_{ds} = K \cdot \frac{W}{L} \left[ (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right] \] ... (2.13)

where,

Factor \( K = \frac{\varepsilon_{\text{ins}} \varepsilon_0 \mu}{D} \)

- Factor \( \frac{W}{L} \) decides the geometry of the device and can be combined with factor \( K \) to give,

\[ \beta = K \cdot \frac{W}{L} \]

\[ I_{ds} = \beta \left[ (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right] \] ... (2.14)

- The capacitance formed by gate and channel has a parallel plate geometry.

\[ C_{g-ch} = \frac{\varepsilon_{\text{ins}} \varepsilon_0 WL}{D} \] ... (2.15)

Then in terms of \( C_{g-ch} \)

\[ K = \frac{\mu \cdot C_{g-ch}}{WL} \]

\[ I_{ds} = \frac{\mu \cdot C_{g-ch}}{L^2} \left[ (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right] \]

- Gate capacitance per unit area \( C_0 \) or \( C_{ox} \) is defined as:

\[ C_0 = \frac{C_{g-ch}}{WL} \]
Therefore $I_{ds}$ can be written as:

$$I_{ds} = C_0 \mu \frac{W}{L} \left[ (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right]$$ ...

(2.16)

### 2.1.2 Saturated Region

- The device enters into saturation when $V_{ds} = V_{gs} - V_t$, because at this point the IR drop in the channel equals the effective gate to channel voltage. The current through the channel remains fairly constant for any further increase in $V_{ds}$.

$$I_{ds} = K \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2}$$ ...

(2.17)

$$I_{ds} = \frac{\beta}{2} \left( V_{gs} - V_t \right)^2$$ ...

(2.18)

$$I_{ds} = \frac{\mu C_{g-ch}}{2L^2} (V_{g(mn)} - V_t)^2$$ ...

(2.19)

$$I_{ds} = C_0 \cdot \mu \cdot \frac{W}{2L} \cdot \left( V_{gs} - V_t \right)^2$$ ...

(2.20)

- The expression for $I_{ds}$ hold for both, the depletion and enhancement mode devices. However, the threshold voltage for nMOS depletion mode device is negative and denoted by $V_{th}$.

- Typical characteristics nMOS transistors are shown in Fig. 2.2 pMOS transistor characteristics are similar with suitable reversal of polarity.

- Following expression summarizes current in three regions.

$$I_{ds} = \begin{cases} 
0 & V_{gs} < V_t \quad \text{Cut-off} \\
\beta \left( -V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds} & V_{ds} < V_{dsat} \quad \text{Linear} \\
\frac{\beta}{2} \left( V_{gs} - V_t \right)^2 & V_{ds} > V_{dsat} \quad \text{Saturation}
\end{cases}$$
2.2 MOS Transistor Threshold Voltage

- The gate structure of MOS transistor consists of charges stored in the dielectric layers, surface to surface interfaces and in the substrate itself. The threshold voltage decided by the structural details of gate-channel structure.

- For switching an enhancement mode MOS transistor from OFF to ON state necessitates applying sufficient \( V_{gs} \) voltage to neutralize these charges and enable the underlying silicon to undergo an inversion due to electric field from the gate.

- For switching an depletion mode nMOS transistor from ON to OFF state consists in applying enough voltage to the gate to add to the stored charge and invert the 'n' implant region to 'p' region.

- Threshold voltage is defined as gate voltage for which strong inversion occurs. Threshold voltage depends on several factors.
• The threshold voltage \( (V_t) \) is given by -

\[
V_t = \phi_{ms} \frac{Q_B - Q_{SS}}{C_0} + 2\phi_{FN} \quad \ldots(2.21)
\]

where,

\( Q_B \) is the charge per unit area in the depletion layer beneath the oxide.

\( Q_{SS} \) is the charge density at Si : SiO\(_2\) interface.

\( C_0 \) is capacitance per unit gate area.

\( \phi_{ms} \) is work function difference between gate and Si. (\( -0.9 \) V over p-substrate and \( -0.2 \) V over n-substrate).

\( \phi_{FN} \) is Fermi level potential between inverted surface and bulk Si.

• The value of \( \phi_{ms} \) is negative and negligible in case of polysilicon gate and silicon substrate. Hence the magnitude and sign of \( V_t \) are decided by the balance between the remaining negative term \( -\frac{Q_{SS}}{C_0} \) and the other two terms, both these terms are positive.

• The expression for \( V_t \) can be evaluated through the following data :

\[
Q_B = \sqrt{2\varepsilon_0 \varepsilon_{si} q N(2\phi_{FN} + V_{SB})} \text{ coulomb / m}^2 \quad \ldots(2.22)
\]

\[
\phi_{FN} = \frac{kT}{q} \ln \frac{N}{n_i} \text{ volts} \quad \ldots(2.23)
\]

\[
Q_{SS} = (1.5 \text{ to } 8) \times 10^{-8} \text{ coulomb / m}^2 \quad \text{depending on crystal orientation.}
\]

Where,

\( V_{SB} \) is substrate bias voltage (negative with respect to source for nMOS, positive for pMOS).

\( q = 1.6 \times 10^{-19} \text{ coulomb} \)

\( N \) is impurity concentration in substrate (\( N_A \) or \( N_D \))

\( \varepsilon_{si} \) is relative permittivity of silicon (\( \approx 11.7 \))

\( n_i \) is intrinsic electron concentration (\( 1.6 \times 10^{10} \text{ /cm}^3 \) at 300 °K)

\( k \) is Boltzmann's constant (\( 1.4 \times 10^{-23} \text{ Joule/°K} \))

• Sometimes a potential may exist between source and substrate as a result of body effects. Then the threshold voltage may get altered. All MOS devices are fabricated on a common substrate. Substrate voltage of all devices is equal. However, in practice several devices may be required to be connected in series for realizing gating functions as shown in Fig. 2.3.
Fig. 2.3 Substrate bias on series connected transistors

- Under normal circumstances ($V_{gs} > V_t$), the depletion layer width remains constant and charge carriers get pulled into the channel from the source. Increasing $V_{SB}$ causes the channel to be depleted of charge carriers and thus the threshold voltage is raised. The density of trapped carriers increases within depletion layer. In order to maintain charge neutrality, the channel charge decreases. The ultimate effect is substrate voltage $V_{SB}$ adds to the junction potential of channel-substrate junction. This increases gate-channel potential drop and hence threshold voltage.

- The change in threshold voltage is given by:
  \[ \Delta V_t = \gamma (V_{SB})^{1/2} \]

where,

\[ \gamma \] is a constant which depends on substrate doping so that more lightly doped substrate smaller will be body effect.

$V_t$ can be written as,

\[ V_t = V_t(0) + \left[ \frac{D \xi d}{\epsilon_{ins} \epsilon_0} \right] \sqrt{2 \epsilon_0 \epsilon_{si} Q N} \left( V_{SB} \right)^{1/2} \]  \hspace{1cm} (2.24)

where,

$V_t(0)$ is threshold voltage for $V_{SB} = 0$.

- Threshold voltage $V_t$ can also be written as,

\[ V_t = V_t(0) + \frac{D \xi d}{\epsilon_{ins} \epsilon_0} \left[ -2 \phi_F + V_{SB} \right] - \sqrt{-2 \phi_F} \]
where, \[ \gamma = \frac{\sqrt{2 q \varepsilon_{si} N}}{C_{ox}} \]

\( \gamma \) is called as body effect coefficient or body factor.

- Expression shows that threshold voltage is affected by material parameters and substrate voltage. The threshold voltage of an n-channel enhancement type MOSFET is positive and p-channel enhancement type MOSFET is negative.

- The magnitude of body effects can be understood by typical values of \( V_t \) which are as under -
  
  For nMOS enhancement mode transistors:
  \[
  \begin{align*}
  V_{SB} &= 0; & V_t &= 0.2 V_{DD} \\
  V_{SB} &= 5 V; & V_t &= 0.3 V_{DD}
  \end{align*}
  \]
  Values for pMOS are similar but negative

  For nMOS depletion mode transistors:
  \[
  \begin{align*}
  V_{SB} &= 0 V; & V_{td} &= -0.7 V_{DD} \\
  V_{SB} &= 5 V; & V_{td} &= -0.6 V_{DD}
  \end{align*}
  \]

2.3 MOS Transistor Transconductance (\( g_m \)) and Output Conductance (\( g_{ds} \))

- Transconductance shows relationship between output current \( I_{ds} \) and input voltage \( V_{gs} \) and is given as -
  \[
  g_m = \frac{\delta I_{ds}}{\delta V_{gs}} \bigg|_{V_{ds}=\text{constant}} \quad \text{(2.25)}
  \]

- By definition of current, the charge in channel \( Q_c \) is such that,
  \[
  \tau = \frac{Q_c}{I_{ds}}
  \]

  where,

  \( \tau \) is transit time for cross-sectional the channel.

- Now change in current is,
  \[
  \delta I_{ds} = \frac{\delta Q_c}{\tau_{ds}} \quad \text{(2.26)}
  \]

  but
  \[
  \tau_{ds} = \frac{L^2}{\mu V_{ds}}
  \]
Therefore,
\[
\delta I_{ds} = \frac{\delta Q_c V_{ds} \mu}{L^2}
\]
\[(2.27)\]

Since
\[
Q = CV, \text{ change in charge}
\]
\[
\delta Q_c = C_g \delta V_{gs}
\]
\[(2.28)\]

From which
\[
\delta I_{ds} = \frac{C_g \delta V_{gs} \mu V_{ds}}{L^2}
\]
\[(2.29)\]

- Transconductance is given by -
\[
g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{C_g \mu V_{ds}}{L^2}
\]
\[(2.30)\]

Under saturation region, \(V_{ds} = V_{gs} - V_t\)

So that,
\[
g_m = \frac{C_g \mu}{L^2} (V_{gs} - V_t)
\]
\[(2.31)\]

From equation (2.15),
\[
C_{g-m} = \frac{\varepsilon_{\text{ins}} \varepsilon_0 WL}{D}
\]

\[
\therefore g_m = \frac{\mu \varepsilon_{\text{ins}} \varepsilon_0 W}{L} (V_{gs} - V_t)
\]
\[(2.32)\]

In terms of \(\beta\),
\[
g_m = \beta (V_{gs} - V_t)
\]
\[(2.33)\]

- Above expressions of \(g_m\) reveals that \(g_m\) of a MOS device can be increased by increasing its width. However, this will also increase the input capacitance and area occupied.

- The resistance of FET in linear region near the origin is known as on state resistance \(R_{on}\) and is given by,

\[
R_{on} = [g_m]^{-1} = \frac{1}{\beta (V_{gs} - V_t)}
\]
where, \[ \beta = K \cdot \frac{W}{L} \]

- The output conductance \( g_{ds} \) can be expressed as -
  \[ g_{ds} = \frac{\delta I_{ds}}{\delta V_{gs}} \]
  \[ = \lambda \cdot I_{ds} = \left( \frac{1}{L} \right)^2 \]

- For the MOS device, strong dependence on the channel length demonstrated as -
  \[ \lambda \propto \frac{1}{L} \quad \text{and} \]
  \[ I_{ds} \propto \frac{1}{L} \]

### 2.4 MOS Transistor Figure of Merit (\( \omega_0 \))

- Figure of merit is a measure for frequency response and switching performance of a MOS transistor. Figure of merit is defined as:
  \[ \omega_0 = \frac{g_m}{C_g} \]

By using equation (2.31),

\[ \omega_0 = \frac{\mu}{L^2} \left( V_{gs} - V_t \right) \]

\[ \omega_0 = \frac{1}{\tau_{sd}} \]

- From equation (2.34), switching speed depends on -
  a) Carrier mobility
  b) Gate voltage (above threshold)
  c) Inversely on square of channel length

- A high speed switching circuit requires a high \( g_m \) as possible.

- The mobility \( \mu \) describes the case with which carriers drift. The mobility may vary in a number of ways. Primarily mobility varies according to the charge carrier, whether electrons or holes. The other factor deciding the mobility is orientation of crystal. For example, electron mobility on a (100) oriented n-type inversion layer surface (\( \mu_n \)) is larger than that of on a (111) oriented
surface. It is almost three times that of hole mobility on a (111) oriented p-type inversion layer.

- Surface mobility is also a function of gate voltage ($V_{gs} - V_t$). Thus a choice of (100) oriented p-type substrate in which the inversion layer will have a surface carrier mobility $\mu_m = 650 \text{ cm}^2 / \text{V - sec}$ at room temperature, would most suited for a fast nMOS circuit. This surface carrier mobility ($\mu_s$) is still quite less than bulk mobilities ($\mu$).

- Typical values of bulk mobilities -
  \[
  \mu_n = 1250 \text{ cm}^2 / \text{V sec} \\
  \mu_p = 480 \text{ cm}^2 / \text{V sec}
  \]

  \[
  \frac{\text{Surface mobility}}{\text{Bulk mobility}} = \frac{\mu_s}{\mu} \approx 0.5
  \]

\textbf{Body Effect / Substrate Bias Effect}

- CMOS devices are consists of common substrate. The substrate bias level is normally equal.

Effect of substrate bias:

- Consider two nMOS transistors connected vertically in series as shown in Fig. 2.4.

\begin{center}
\includegraphics[width=0.5\textwidth]{fig_2_4.png}
\end{center}

Fig. 2.4

- The source to substrate voltage $V_{sb}$ as observed vertically upward i.e. $V_{sb1} = 0$ and $V_{sb2} \neq 0$. It means that $V_{sb2}$ and $V_{sb1}$ are not same. This variation of $V_{sb}$ causes variation in threshold voltages i.e. $V_{t2}$ and $V_{t1}$. Since
$V_{sb2} > V_{sb1}$ therefore, $V_{t2} > V_{t1}$. This variation of threshold voltage due to source to substrate voltage is referred as **body effect**. In complex logic circuit body effect is significantly important.

- The body effect (change in threshold voltage) is because of variation in depletion charge region under oxide as a result of source substrate voltage as shown in Fig. 2.5.

![n-channel MOSFET](image)

**Fig. 2.5 n-channel MOSFET**

Change in threshold voltage $\Delta V_{th}$ is given by:

$$\Delta V_{th} = \sqrt{\frac{2qN_A\varepsilon_{si}}{C_{ox}}} \left[ \sqrt{2\phi_s + V_{sb}} - \sqrt{2\phi_s} \right]$$

where,

$\gamma_n$ is body effect factor depends on gate oxide thickness and substrate doping.

Typical value of $\gamma$ lies between 0.4 to 1.2.

### 2.6 MOS DC Equations

- MOS transistors have three regions of operation:
  1. Cut-off or subthreshold region
  2. Linear or nonsaturation region
  3. Saturation region.

- In cut-off or subthreshold region, where $V_{gs} < V_t$
  $$I_{ds} = 0 \quad \ldots(2.35)$$

- In Linear or nonsaturation region, where $V_{ds} < V_{dsat}$
  $$I_{ds} = \beta \left( V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds} \quad \ldots(2.36)$$
Region (4): \( V_{DD}/2 < V_{in} \leq (V_{DD} + V_{tp}) \)
- In this region pMOS transistor is in saturation and nMOS transistor is in non-saturation region. Output voltage is given by,

\[
V_{out} = (V_{in} - V_{TON}) - \left( \frac{V_{in}}{V_{TON}} \right)^{\frac{2}{\beta_p}} \left( \frac{V_{in} - V_{DD} - V_{tp}}{\beta_n} \right)^{\frac{1}{2}}
\]

\[
V_{out} = V_{OL}
\]

Region (5): \( V_{in} \geq (V_{DD} + V_{tp}) \)
- In this region pMOS transistor is in cut-off region and nMOS transistor is in linear mode. Output voltage is,

\[
V_{out} = 0
\]
- Table 2.3 summarizes operating region of pMOS and nMOS transistors and corresponding \( V_{in} \) and \( V_{out} \).

<table>
<thead>
<tr>
<th>Regions</th>
<th>( V_{in} )</th>
<th>Operating Regions</th>
<th>( V_{out} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pMOS</td>
<td>nMOS</td>
</tr>
<tr>
<td>1</td>
<td>( &lt; V_{ION} )</td>
<td>Linear</td>
<td>Cut-off</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( V_{OH} = \frac{V_{DD}}{2} )</td>
</tr>
<tr>
<td>2</td>
<td>( V_{ION} \leq V_{in} &lt; \frac{V_{DD}}{2} )</td>
<td>Linear</td>
<td>Saturation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( V_{OH} &gt; \frac{V_{DD}}{2} )</td>
</tr>
<tr>
<td>3</td>
<td>( V_{in} = \frac{V_{DD}}{2} )</td>
<td>Saturation</td>
<td>Saturation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0 (Drops sharply)</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{V_{DD}}{2} &lt; V_{in} \leq \frac{V_{DD}}{2} -</td>
<td>V_{tp}</td>
<td>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( - V_{OL} (&lt; \frac{V_{DD}}{2}) )</td>
</tr>
<tr>
<td>5</td>
<td>( \geq \frac{V_{DD}}{2} + V_{tp} )</td>
<td>Cut-off</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.3

2.9.1 Symmetric CMOS Inverter
- Sometimes it is necessary that a CMOS transistor should have symmetric transient response. Therefore CMOS inverter should be designed for symmetric voltage transfer curve.

- For a symmetric voltage transfer curve \( V_{in} = V_{th} = \frac{V_{DD}}{2} \). The input low and input high voltages i.e. \( V_{IL} \) and \( V_{IH} \) are then symmetric about \( V_{TH} \). Also there will be same susceptibility to noise at low and high logic levels.
\[ V_{\text{in}} = V_{\text{th}} = \frac{V_{\text{ON}} + \left( \frac{1}{K_R} \right)^{1/2} (V_{\text{DD}} - V_{\text{tp}})}{1 + \left( \frac{1}{K_R} \right)^{1/2}} \]

\[ \therefore K_R = \left( \frac{V_{\text{DD}} + V_{\text{tp}} - V_{\text{th}}}{V_{\text{th}} - V_{\text{ON}}} \right)^2 \]

But for an ideal inverter,

\[ V_{\text{th}} = \frac{V_{\text{DD}}}{2} \]

\[ \therefore (K_R)_{\text{ideal}} = \left( \frac{0.5 V_{\text{DD}} + V_{\text{tp}}}{0.5 V_{\text{DD}} + V_{\text{ON}}} \right)^2 \]

\[ (K_R)_{\text{ideal}} = 1 \quad \therefore V_{\text{tp}} = V_{\text{ON}} \]

Also \[ K_R = \frac{\mu_n}{\mu_p} \frac{(W / L)_n}{(W / L)_p} \]

and for symmetric inverter \( K_R = 1 \).

\[ \frac{(W/L)_n}{(W/L)_p} = \frac{\mu_p}{\mu_n} \]

\[ \left( \frac{W}{L} \right)_p = \frac{\mu_p}{\mu_n} \left( \frac{W}{L} \right)_n \]

Substituting values of \( \mu_p \) and \( \mu_n \):

\[ \left( \frac{W}{L} \right)_p = \frac{580 \text{ cm}^2 / \text{V} - \text{s}}{230 \text{ cm}^2 / \text{V} - \text{s}} \left( \frac{W}{L} \right)_n \]

\[ \left( \frac{W}{L} \right)_p = 2.5 \left( \frac{W}{L} \right)_n \]

**Input Low Voltage**\( (V_{\text{IL}}) \):

\[ V_{\text{IL}} = \frac{1}{8} (3 V_{\text{DD}} + 2 V_{\text{tp}}) \]

**Input High Voltage**\( (V_{\text{IH}}) \):

\[ V_{\text{IH}} = \frac{1}{8} (5 V_{\text{DD}} - 2 V_{\text{tp}}) \quad \text{As} \ V_{\text{IL}} + V_{\text{IH}} = V_{\text{DD}} \]

**Noise Margin**:\n
\[ \text{NM}_{\text{IH}} = V_{\text{OH}} - V_{\text{IH}} \]
\[ NM_H = V_{DD} - V_{IH} \]
\[ NM_L = V_{IL} - V_{OL} \]
\[ NM_L = V_{IL} - V_{OL} \]

2.9.2 Noise Margin

- Noise margin or noise immunity is allowable input gate voltage so that output will not be corrupted. Two parameters are used to specify noise margin or noise immunity, these are - LOW noise margin \( (NM_L) \) and HIGH noise margin \( (NM_H) \).

- The value of LOW noise margin \( (NM_L) \) is the difference in maximum LOW input voltage recognized by input gate and maximum LOW output voltage of driver gate.

\[ NM_L = V_{IL} - V_{OL} \]  \hspace{1cm} \text{(2.42)}

- The HIGH noise margin \( (NM_H) \) is difference between minimum HIGH output voltage of driving gate and minimum HIGH input voltage recognized by the receiving gate.

\[ NM_H = V_{OH} - V_{IH} \]  \hspace{1cm} \text{(2.43)}

where,

- \( V_{IH} \) is minimum HIGH input voltage
- \( V_{IL} \) is maximum LOW input voltage
- \( V_{OH} \) is minimum HIGH output voltage
- \( V_{OL} \) is maximum LOW output voltage

Noise margin definitions are shown in Fig. 2.18.
- The input range between $V_{IL}$ and $V_{IH}$ are called as intermediate region or forbidden zone. It do not represent any valid logic levels. It is desirable to have $V_{IH}$ and $V_{IL}$ should be close to each other i.e. to switch transfer characteristics abruptly.

- The voltage levels on transfer characteristics are shown in Fig. 2.19. The logic levels are at unity gain point (slope = -1).

![Diagram](image)

**Fig. 2.19 CMOS inverter noise margin**

- Typical values of noise margins are:
  
  $NM_L = 0.46 \ V_{DD}$
  $NM_H = 0.34 \ V_{DD}$

- When input is at its worst legal value, the output is slightly degraded, this is called as noise feed through or propagated noise.

2.9.3 Rise Time

- The delay associated with the CMOS inverter can be more precisely estimated by splitting the output transitions into full time $\tau_f$ and rise-time $\tau_r$, corresponding to the charging and discharging of the capacitive load $C_L$. These two phenomenon can be independently estimated more precisely to arrive at the overall delay.
Estimation of Rise-time:

- The pull-up p-device drives the capacitive load and can be assumed to be in saturation for the entire charging period of the load capacitor $C_L$. The equivalent circuit for this condition is shown in Fig. 2.20. Now the saturation current for the p-transistor is given by:

$$I_{dsp} = \frac{(V_{gs} - |V_{ip}|)^2 \beta_p}{2}$$

![Equivalent circuit diagram]

Fig. 2.20 Rise-time model

- This current being constant, output voltage,

$$V_{out} = \frac{t I_{dsp}}{C_L}$$

Putting expression for $I_{dsp}$, we get,

$$t = \frac{2V_{out} C_L}{(V_{gs} - |V_{ip}|)^2 \beta_p}$$

- The rise time $\tau_r$ corresponds to time taken by $V_{out}$ to reach $V_{DD}$ (approximately), so that,

$$\tau_r = \frac{2C_L V_{DD}}{(V_{DD} - |V_{ip}|)^2 \beta_p}$$
But $|V_T| = 0.2 V_{DD}$ from which,

$$\tau_r = \frac{3C_L}{V_{DD} \beta_p} \quad \text{(30)}$$

2.9.4 Fall Time

- Fall-time is associated with the discharging of $C_L$ through the pull-down n-type device. The equivalent circuit model for fall-time estimation is shown in Fig. 2.21 which shows a constant discharge current.

![Fig. 2.21 Fall-time model](image)

Hence fall-time comes to,

$$\tau_f = \frac{3C_L}{V_{DD} \beta_n} \quad \text{(2.45)}$$

From equations 2.44 and 2.45 and

$$\frac{\tau_r}{\tau_f} = \frac{\beta_n}{\beta_p}$$

- However, $\mu_n$ and $\mu_p$ are not same and $\mu_n = 2.5 \mu_p$ because of which $\beta_n = 2.5 \beta_p$. This shows that rise time is slower by a factor of 2.5 when both the n and p-devices are minimum size devices.
• Keeping the channel length minimum, symmetrical operation can be achieved by making \( W_p = 2.5W_n \) where \( W_p \) is channel width of p-device and \( W_n \) that of n-device. However, with the other geometries as per the minimum size lambda-based rules, this would result in the inverter having an input capacitance of \( 2.5 C_g \) for p-device plus \( C_g \) for n-device giving a total capacitance of \( 3.5 C_g \).

• The analytical models used above for the estimation of rise and fall-times are adequate enough to get optimistic results. However, they do not consider certain factors affecting the rise and fall-times such as,
  i) \( \tau_f \) and \( \tau_r \) are proportional to \( C_L \).
  ii) \( \tau_f \) and \( \tau_r \) are proportional to \( 1/V_{DD} \).
  iii) for equal n and p-transistor geometries, \( \tau_r = 2.5 \tau_f \).

• Whenever output signals are required to propagate from the chip to off-chip destinations, the output experiences a comparatively large capacitive load. Such off-chip capacitances may be several orders higher than on-chip \( C_g \) values and typically the off-chip load \( C_L \geq 10^4 C_g \). Such capacitances must clearly be driven through low resistances to avoid excessively long delays.

2.9.5 Power Dissipation in CMOS Inverter

• Power dissipation in a CMOS inverter is given as -

\[
P = V_{DD} \cdot I_{DC}
\]

where,

\( V_{DD} \) is power supply voltage.

\( I_{DC} \) is current drawn from power supply in steady state.

• In CMOS inverter power dissipation is caused due to various factors, depending on which it is categorized into two types.

1. Static power dissipation : Static power dissipation is due to the leakage current drawn from the power supply.

2. Dynamic power dissipation : Dynamic power dissipation is due to switching transient current and charging and discharging of load capacitance.

2.9.5.1 Static Power Dissipation

• Static power dissipation is also known as steady state power dissipation. The static power dissipation is given by -

\[
P_{static} = I_{static} \cdot V_{DD}
\]

where,
\( I_{\text{static}} \) is the current flowing from supply to ground in absence of switching.

- Static power dissipation is due to leakage current flowing through the reverse biased junctions of transistors. The leakage current per unit drain area typically ranging between 10 to 100 pA/\( \mu \text{m}^2 \) at room temperature.

- Junction leakage current is also due to thermally generated carriers. These carriers increases with increase in temperature. The power dissipation increases with increase in package density or number of devices. The total static power dissipation is given as:

\[
P_{\text{Total static}} = \sum_{i=1}^{n} I_{\text{static}} \cdot V_{DD}
\]

2.9.5.2 Dynamic Power Dissipation

- During switching of transistors current pulses of small duration are produced. Also the load capacitance \( C_L \) gets charged through the pMOS transistor, its voltage rises from 0 to \( V_{DD} \) and draws energy from power supply.

- If capacitive load increases its charge and discharging dominates the current drawn from power supply.

- For a square wave input of frequency \( f_p \) and voltage \( V_{in} \), dynamic power \( P_d \) during switching is given by:

\[
P_{\text{dynamic}} = C \cdot V_{DD}^2 \cdot f_p
\]

2.9.5.3 Short Circuit Power Dissipation

- It indicates the power dissipation due to short circuit current in an unloaded inverter. Short circuit power dissipation is given as:

\[
P_{sc} = \frac{K}{12} \left( V_{DD} - V_t \right) \frac{3 t_{rf}}{t_p}
\]

where,

\( t_p \) is period of input waveform.
\( t_{rf} \) is rise time of input waveform.
\( K \) is a constant.

2.9.5.4 Total Power Dissipation

- Total power dissipation of CMOS inverter is expressed as sum of all power dissipation.

\[
P_{\text{total}} = P_{\text{static}} + P_{\text{dynamic}} + P_{sc}
\]
2.9.6 W/L Calculations

- The switching speed of CMOS logic device is mainly decided by -
  a) Load Capacitance (C_L)
  b) Gain Factor β (W/L)

c) Supply Voltage (V_{DD})

- The rise time (t_r) and fall time (t_f) delays are inversely proportional to gain factor β of CMOS transistor. And the gain factor β is directly proportional to channel width (W). In other words lowering channel width W decreases gain factor β, increases delay time t_d.

- The gain factor β for nMOS and pMOS transistors are abbreviated as β_n and β_p respectively. For equal sized nMOS and pMOS transistors β_n is 2 to 3 times greater than β_p, i.e. β_n = 2 to 3 β_p.

- For nMOS transistor -
- For nMOS transistor -

\[ R_n = \frac{L_n}{W_n \beta_n} \]  \hspace{1cm} \text{...(2.46)}

where,

- \( R_n \) is resistance of nMOS
- \( L_n \) is length of channel.

- For pMOS transistor -

\[ R_p = \frac{L_p}{W_p \beta_p} \]  \hspace{1cm} \text{...(2.47)}

where,

- \( R_p \) is resistance of pMOS
- \( L_p \) is length of channel.

- When current through nMOS and pMOS are equal then the resistance of channel must be equal i.e.

\[ R_p = R_n \]  \hspace{1cm} \text{...(2.48)}

\[ \frac{L_p}{W_p \beta_p} = \frac{L_n}{W_n \beta_n} \]

\[ \therefore \frac{\beta_n}{\beta_p} = \frac{W_p}{W_n} \]

Assuming channel lengths \( L_p \) and \( L_n \) are equal.

Then

\[ \frac{\beta_n}{\beta_p} = \frac{W_p}{W_n} \]
Also let \[ \beta_n = 2.5 \beta_p \]
\[ \frac{W_p}{W_n} = \frac{2.5 \beta_p}{\beta_p} \]
\[ \therefore W_p = 2.5 W_n \]

Therefore channels of equal length in pMOS and nMOS, for driving equal current the width of pMOS must be 2.5 times the width of nMOS transistor.

2.9.7 Transmission Gate

- A parallel combination of MOSFET is known as transmission gate. It is also known as pass gate. Fig. 2.22 (a) shows structure of CMOS transmission gate.

![Diagram of CMOS transmission gate](image)

- A transmission gate consists of pMOS and nMOS transistors. A control signal is applied to nMOS transistor and its complement is applied to pMOS transistor.

- The transmission gate is used as bidirectional switch controlled by control signal. When control signal S is high both pMOS and nMOS transistors are ON and provides a low resistance path between terminals A and B vice-versa. CMOS TG are compact circuit requires minimum transistors.
2.9.7.1 Applications of TG

1. OR Operation

- OR Gate using TG is shown in Fig. 2.23.

![Fig. 2.23 OR operation]

2. XOR Operation

- XOR operation using transmission gate is shown in Fig. 2.24.

![Fig. 2.24 XOR operation]

3. Two Input Multiplexer
2.9.8 Tristate Inverter

- Cascade arrangement of transmission gate and inverter is called tristate inverter. The tristate inverter is approximately half the speed of complementary CMOS inverter.
- The tristate inverter is the basic for various types of clocked logic, latches, bus drivers and multiplexers.

2.9.9 CMOS Combinational Logic Design

- Combinational logic circuit can be implemented by using CMOS transistor. AOI operation using CMOS transistor is shown.
2.10 Pass Transistor

- MOS transistors can be used as switches by virtue of the isolated nature of the gate. When the transistor is operating in the linear region, the device acts as a linear resistance under gate voltage control. In this mode the transistor can be used as an on-off switch as shown in Fig. 2.26.

![Fig. 2.26 The MOS transistor as a switch](image)

- The switch is turned off by setting $V_{GS} = 0$, the channel disappears and only a small amount of leakage current flows at the drain end. By setting $V_{GS} = V_{DD}$, the switch is turned on and the current path provides a resistance $R_{on}$. In case of a p-channel transistor, the switch is turned on if $V_{GS} = -V_{DD}$. If the transistor is connected in series with a high impedance circuit the total current flowing through the transistor will be very small and so will be the voltage drop across the channel. When used as a logical element, an input voltage source signal, $V_{in}$, which is either zero or $V_{DD}$ representing logical 0 to 1, respectively, is applied as shown in the Fig. 2.27.

![Fig. 2.27 Pass transistor](image)
The output node $V_{out}$ is either held at its previous logic value of the switch is open or its capacitance $C$ is charged up to $V_{in}$ if the gate voltage allows the transistor path to be closed, forcing a current for a short period of time. The transistor used in this fashion is called a pass transistor and the process of transferring the charge from the input node to the output under the control of gate voltage is called charge steering.

A number of transistors can be also used as switches in series in switching logic arrays. An AND array is shown in Fig. 2.28.

![Fig. 2.28 Pass transistor AND gate](image)

Note: Means must exist so that $X$ assumes ground potential when $A + B + C = 0$.

### 2.11 nMOS Inverter

The simplest form of an nMOS logic circuit is an inverter as shown in Fig. 2.29. It consists of a load resistance $R$ called the pull-up resistor and a pull-down transistor $T$ connected in series between supply voltage $V_{DD}$ and ground. The value of the resistor is so decided as to limit the pull-up current to some fraction of maximum pull-down current provided by the transistor. For $V_{DD}$ equal to +5 V, value of $R$ would typically be 40 kΩ which limits the pull-up current to about 125 μA. The input voltage is applied to the gate of transistor $T$. The transistor offers a very high input impedance through the gate capacitance $C_g$. In a typical application, the output drives a similar load capacitance $C$.

If $V_{in}$ is less than $V_{th}$, the threshold voltage of $T$, transistor $T$ will be turned off and the load capacitance $C$ will be charged to $V_{DD}$, which will be the output voltage $V_{out}$. However, due to a small leakage current flowing through $T$, the output voltage $V_{out}$ never equals $V_{DD}$ but is slightly less than $V_{th}$. If $V_{in}$ is increased beyond $V_{th}$, transistor $T$ will initially conduct in the
saturation region; the capacitance C will start discharging, pulling down $V_{out}$ to a lower voltage and the current $I_{ds}$ will increase due to the larger value of $V_{gs}$ which equals $V_{in}$. This continues until the transistor goes into the linear region of the $I_{ds} - V_{ds}$ curve. When $V_{in}$ equals $V_{DD}$, T will conduct heavily and the output voltage will be given by,

$$V_{out} = V_{DD} \frac{R_{ch}}{R + R_{dl}}$$

- Where $R_{ch}$ is the channel resistance with $V_{gs} = V_{DD}$. Since $R_{ch}$ is very small compared to R, the output voltage will be almost the ground potential. In a typical application, the output of the inverter is used to drive another inverter and $V_{out}$ must be less than $V_{th}$ to ensure that the pull-down transistor of the driven inverter is turned off. Since $V_{th} = 0.2V_{DD}$, it is necessary that

$$R \geq 4R_{ch}$$

- This means that the output capacitance C discharges through R, at least four times faster during the pull-down phase than during its pull-up phase through R. This basic asymmetry of switching times is a fundamental limitation of 'ratioed' logic which this inverter represents. To have a higher speed of operation, R must be made small. This not only requires a smaller value of $R_{ch}$ (requiring a larger silicon area for the pull-down transistor) to satisfy the ratio criteria, but also involves more power dissipation in the load resistor $[(V_{DD} - V_{out})^2 / R]$ watts due to increased current. An effect same as that due to lowering R can be achieved by lowering the value $V_{th}$, which results in a larger source-to-drain current. This would certainly give higher speed, but to produce a lower value of the output voltage $V_{out}$, $R_{ch}$ should be accordingly lower. For example, if $V_{th} = 0.1V_{DD}$, the ratio becomes $R \geq 9R_{ch}$.
Lower channel resistance implies larger silicon area for the pull-down. Fabricating a large resistive load on the silicon surface takes up quite a bit of area. A better solution is to use a depletion mode transistor as a pull-up load with source and gate connected together so that \(V_{GS} = 0\). Such an arrangement is shown in Fig. 2.30.

![Fig. 2.30 nMOS inverter](image)

- With such a configuration, with no current drawn from the output, the currents \(I_{ds}\) for both transistors are equal. For the depletion mode transistor \(T_2\), the gate is connected to the source so it is always on and only the characteristic curve \(V_{gs} = 0\) is relevant. In such a configuration the depletion mode device is called the pull-up (P.U.) and the enhancement mode device the pull-down (P.D.) transistor. The transfer characteristic of such an inverter can be obtained by superimposing the \(V_{gs} = 0\) depletion mode characteristic curve on the family of curves for the enhancement mode device with the condition that the maximum voltage across the enhancement mode device corresponds to minimum voltage across the depletion mode device. The points where these curves intersect are the points on the transfer characteristics of the nMOS inverter in Fig. 2.30.

- The superimposition of the two characteristics is shown in Fig. 2.31.
  - The superimposition of the two characteristics is shown in Fig. 2.31.
Fig. 2.31 Derivation of nMOS inverter transfer characteristic

- The intersection results into the transfer characteristic as shown in Fig. 2.32.

Fig. 2.32 nMOS inverter transfer characteristic

- Now the input voltage $V_{in}$ is the gate to source voltage $V_{gs}$ for the pull-down transistor. As this exceeds the threshold voltage of the p.d. transistor, current begins to flow. This causes the output voltage $V_{out}$ to decrease and further increase in $V_{in}$ causes the p.d. transistor to come out of saturation and become resistive. Initially, the pull-up transistor is resistive till the p.d. device turns on. The gain is defined as the slope of the transfer characteristic during transition.

$$\text{Gain} = \frac{\delta V_{out}}{\delta V_{in}}$$

The point on the transfer characteristic at which $V_{out} = V_{in}$ is denoted as $V_{inv}$. 
2.12 Estimation of Pull-up to Pull-down Ratio \((Z_{p.u.} / Z_{p.d.})\) of an nMOS Inverter Driven by Another nMOS Inverter

- Fig. 2.33 shows an inverter driven from the output of another similar inverter. Let \(V_{gs} = 0\) for the depletion mode transistor under all conditions. Also, in order to cascade inverters without having any inverse effect on levels, our target is to meet the requirement.

![Fig. 2.33 nMOS inverter driven directly by another inverter](image)

\[ V_{in} = V_{out} = V_{inv} \]

- In order to have equal margins around the inverter threshold, we select \(V_{inv} = 0.5V_{th}\). Then both the transistors are in saturation. The drain-to-source current under saturation is given by,

\[ I_{ds} = \frac{W}{L} K \left( \frac{V_{gs} - V_t}{2} \right)^2 \]

**Enhancement mode**

\( I_{ds} \) for the pull-down device which is in the enhancement mode and has \(V_{gs} = V_{inv}\) is given by,

\[ I_{ds} = \frac{W_{p.d.}}{L_{p.d.}} K \left( \frac{V_{inv} - V_t}{2} \right)^2 \]

**Depletion mode**

The pull-up device which is in the depletion mode and has \(V_{gs} = 0\) is given by,

\[ I_{ds} = \frac{W_{p.u.}}{L_{p.u.}} K \left( \frac{V_{inv} - V_t}{2} \right)^2 = \frac{W_{p.u.}}{L_{p.u.}} K \left( -V_t \right)^2 \]

Since the two currents are same (p.u. and p.d. devices are in series), we have

\[ \frac{W_{p.u.}}{L_{p.u.}} (-V_t)^2 = \frac{W_{p.d.}}{L_{p.d.}} (V_{inv} - V_t)^2 \]

where \(W_{p.u.}, L_{p.u.}, W_{p.d.},\) and \(L_{p.d.}\) are the widths and lengths of the pull-up and pull-down transistors respectively.

Denoting by,

\[ Z_{p.d.} = \frac{L_{p.d.}}{W_{p.d.}} \quad \text{and} \quad Z_{p.u.} = \frac{L_{p.u.}}{W_{p.u.}} \]

We get,

\[ \frac{1}{Z_{p.u.}} (-V_t)^2 = \frac{1}{Z_{p.d.}} (V_{inv} - V_t)^2 \]
from which

\[ V_{\text{inv}} = V_t - \frac{V_{id}}{\sqrt{Z_{p.u.} / Z_{p.d.}}} \]

The typical values for the voltages are \( V_{id} = -0.6V_{DD} \), \( V_t = 0.2V_{DD} \) and \( V_{\text{inv}} = 0.5V_{DD} \) (to have equal margins). Putting these values into equation we get,

\[ 0.5 = 0.2 + \frac{0.6}{\sqrt{Z_{p.u.} / Z_{p.d.}}} \]

from which

\[ \frac{Z_{p.u.}}{Z_{p.d.}} = 2 \]

or

\[ \frac{Z_{p.u.}}{Z_{p.d.}} = 4 \]

For an inverter driven directly by an inverter.

2.13 Determination of Pull-up to Pull-down Ratio for an nMOS Inverter Driven through One or More Pass Transistors

- Sometimes the input to an inverter may come from the output of an inverter but after passing through one or more nMOS transistors that are used as pass transistors. Such an arrangement is shown in Fig. 2.34.

![Fig. 2.34 Pull-up to pull-down ratios for inverting logic coupled by pass transistor](image)

- The point of concern here is that the connection of pass transistors in series may degrade the logic 1 level into inverter 2 so that the output may not be a proper logic 0 level. Of special concern is the condition when point A in Fig. 2.34 is at 0 volts and B is thus at \( V_{DD} \), but the voltage into inverter 2 at point C has got reduced from \( V_{DD} \) by the threshold voltage of the series pass transistor. With the gates of all pass transistors connected to \( V_{DD} \), as shown in the Fig. 2.34, there is a reduction in voltage by \( V_{tp} \) where \( V_{tp} \) is the threshold voltage of a pass transistor. Although many devices are connected in series, there can be no voltage drop in the channels since no static current flows through them. Hence the input voltage to inverter 2 is
\[ V_{in2} = V_{DD} - V_{ip} \]

We now aim at getting the same output voltage as would be the case for inverter 1 driven with input \( V_{DD} \).

- When input to inverter 1 is \( V_{DD} \), its pull down transistor \( T_2 \) is conducting but with a low voltage across it. Since it is in its resistive region of operation, we represent it by resistance \( R_1 \) as shown in Fig. 2.35. At the same time, the pull-up transistor \( T_1 \) is in saturation and can be represented as a current source. The current in the pull-down transistor, which in its linear region of operation is given by,

\[
I_{ds} = K \frac{W_{p,d,1}}{L_{p,d,1}} \left[ (V_{DD} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right]
\]

Now,
\[
R_1 = \frac{V_{ds}}{I_{ds}} = \frac{1}{K} \frac{L_{p,d,1}}{W_{p,d,1}} \left[ \frac{1}{V_{DD} - V_t - \frac{V_{ds}}{2}} \right]
\]

Since \( V_{ds} \) is small and hence \( V_{ds} / 2 \) can be neglected.

Then,
\[
R_1 = \frac{1}{K} \frac{Z_{p,d,1}}{V_{DD} - V_t}
\]

The pull-up device is in depletion mode with \( V_{gs} = 0 \).

Its current
\[
I_1 = I_{ds} = K \frac{W_{p,u,1}}{L_{p,u,1}} \frac{(-V_{td})^2}{2}
\]

The output of inverter 1,
\[
V_{out1} = I_1 R_1 = \frac{Z_{p,d,1}}{Z_{p,u,1}} \left[ \frac{1}{V_{DD} - V_t} \right] \frac{(V_{td})^2}{2}
\]
Now consider inverter 2 with input as $V_{DD} - V_{tp}$.

Similar to that for inverter 1,

$$R_2 = \frac{1}{K} \frac{Z_{p.d.2}}{[V_{DD} - V_{tp} - V_t]} \cdot \frac{1}{2}$$

and

$$I_2 = \text{Current for depletion mode pull-up device in saturation with } V_{gs} = 0$$

$$= K \frac{1}{Z_{p.u.2}} \frac{(-V_{td})^2}{2}$$

Output voltage of inverter 2 $V_{out2}$ is given by,

$$V_{out2} = I_2 R_2$$

$$= \frac{Z_{p.d.2}}{Z_{p.u.2}} \left[ \frac{1}{V_{DD} - V_{tp} - V_t} \right] \frac{(-V_{td})^2}{2}$$

- If the output of inverter 2 should be the same as that of inverter 1 under these conditions, then $V_{out1} = V_{out2}$ or $I_1 R_1 = I_2 R_2$.

$$\frac{Z_{p.u.1}}{Z_{p.d.1}} \frac{(V_{DD} - V_t)}{(V_{DD} - V_{tp} - V_t)} = \frac{Z_{p.u.2}}{Z_{p.d.2}}$$

The typical value of voltages are

$$V_t = 0.2 \ V_{DD} \quad \text{and} \quad V_{tp} = 0.3 \ V_{DD}$$

Using these values, we get,

$$\frac{Z_{p.u.1}}{Z_{p.d.1}} \frac{0.8}{0.5} = \frac{Z_{p.u.2}}{Z_{p.d.2}}$$

Thus,

$$\frac{Z_{p.u.2}}{Z_{p.d.2}} = 2 \frac{Z_{p.u.1}}{Z_{p.d.1}} \cdot \frac{8}{1}$$

Therefore an inverter driven through one or more pass transistors has a $Z_{p.u.}/Z_{p.d.}$ ratio of $\geq 8 / 1$.

2.14 Alternative Forms of Pull-up

- So far we have seen two forms of pull-up; first in the form of a resistance and the second realized by a depletion mode pull-up device. Two other possible forms of pull-up are as follows:

  a) Complementary transistor pull-up (CMOS):

  - This forms of pull-up is shown in Fig. 2.36. The output gives full logical 1
and 0 levels. No current flows either for logical 1 or for logical 0 inputs. For similar dimension devices, the n-channel device is faster than the p-channel device.

Fig. 2.36 Complementary transistor pull-up (CMOS)

b) nMOS enhancement mode pull-up:
- This configuration is shown in Fig. 2.37. In this circuit for output of logical 1, the $V_{out}$ never reaches $V_{DD}$. Dissipation in the circuit is also high since current flows when $V_{in} =$ logical 1. Dissipation can be considerably reduced if $V_{GG}$ is obtained from a switching source such as from one phase of a clock. In this mode of pull-up, if $V_{GG}$ is higher than $V_{DD}$, an extra supply rail is required.
2.15 CMOS Inverter

- In CMOS, both p- and n-channel transistors are used. The circuit is fabricated on an n-type silicon substrate in which a p-type "well" or "tub" is created by diffusion. The n-channel transistors are created in the p-well region. The p-channel devices are made in the n-substrate under an ion-implanted layer called the p⁺-layer. Similar to the depletion-mode implantation used for nMOS. In this p-well process, special "p-plugs" are used to connect the p-well substrate and the source of the n-channel transistor to ground. Similarly, an "n-plug" connects the n-substrate and the source of the p-type transistor to $V_{DD}$, the positive supply voltage.

- A schematic circuit representation of the CMOS inverter is shown in Fig. 2.38 along with its transfer characteristics.

- The operation of the circuit on an inverter can be explained as follows. All voltages are referenced with respect to $V_{SS}$, the ground potential. When the input voltage $V_{in}$ is zero, the gate of the p-channel transistor is at $V_{DD}$ below the source potential, that is, $V_{gs} = V_{DD}$. This turns on this transistor, offering a low-resistance path to load capacitance C which will be charged up to $V_{DD}$. No current flows through the n-channel transistor, which is turned off since $V_{gs} = 0$ for this transistor. Now if the input voltage is raised to the threshold voltage level of the n-channel transistor and then to $V_{DD}$, the n-channel transistor will conduct while the p-channel transistor gets turned off, discharging the load capacitance C to ground potential. Note that the current flows until the output node reaches $V_{DD}$ (when charging) or ground (when discharging), the transistors providing either the charge or the discharge current.
Fig. 2.38 CMOS Inverter

- A detailed analysis of the circuit now follows on the basis of its transfer characteristics shown in Fig. 2.38 (b). The current/voltage relationships for the MOS transistor is given by the expression for \( I_{ds} \).

\[
I_{ds} = \frac{W}{L} (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2}
\]

in the resistive region and for saturation region by,

\[
I_{ds} = \frac{W}{L} K \frac{(V_{gs} - V_t)^2}{2}
\]

The factor \( K \) depends on the geometry of the technology involved since

\[
K = \frac{\varepsilon_0 \varepsilon_{ins} \mu}{D}
\]

The factor \( W/L \) is also attributed to the geometry and it is common practice to write

\[
\beta = K \frac{W}{L}
\]

which gives for example,

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

in saturation. The factor \( \beta \) is applicable to both nMOS and pMOS transistors as follows,

\[
\beta_p = \frac{\varepsilon_0 \varepsilon_{ins} \mu_p}{D} \frac{W_p}{L_p}
\]

\[
\beta_n = \frac{\varepsilon_0 \varepsilon_{ins} \mu_n}{D} \frac{W_n}{L_n}
\]

- where \( W_p \) and \( L_p \), \( W_n \) and \( L_n \) are the p- and n-transistor dimensions respectively and \( \mu_p \) and \( \mu_n \) are the hole and electron mobilities respectively. From Fig. 2.38 (b) and (c), we find that the CMOS inverter has five distinct regions of operation.
- The region 1 corresponds to operation when $V_{in} = \text{logic 0}$, the p-transistor turned fully on while the n-transistor is fully turned off. Hence no current flows through the inverter circuit and the output is directly connected to $V_{DD}$ through the pull-up p-transistor. The output has a good logic 1 level. The inverted output corresponds to region 5 when $V_{in} = \text{logic 1}$, the n-transistor turned fully on while the p-device is fully off. In this region, again, no current flows through the circuit. These two regions viz. region 1 and 5 are the static conditions.

- Now consider region 2 of operation. The input voltage has increased to a level that is just above the threshold voltage of the n-transistor. The n-transistor conducts and has a large voltage difference between drain and source and is in saturation. The p-transistor is also conducting, but with only a small voltage difference between its drain and source and hence it operates in the unsaturated resistive region. The inverter circuit draws a small current from the $V_{DD}$ supply.
• In the region 4, conditions are same as those in region 2 but with the roles of the p- and n-transistors reversed. That is, p-transistor has a large voltage across it while the n-transistor has a small drop across it. The current drain in both the regions 2 and 4 is small.

• Most of the energy consumed in switching from one state to the other is attributed to the large current flow in region 3. This is the region of operation in which the inverter exhibits gain and in which both the devices are in saturation. Since the two transistors are in series, the current through them is same and we can write.

\[
I_{d_{p}} = -I_{d_{n}}
\]

where, \[
I_{d_{p}} = \frac{\beta_{p}}{2} (V_{in} - V_{DD} - V_{tp})^2
\]

is the drain to source current for the p-transistor.

and \[
I_{d_{n}} = \frac{\beta_{n}}{2} (V_{in} - V_{in})^2
\]

is the drain-to-source current for the n-transistor. Writing for \(V_{in}\) in terms of the \(\beta\) ratio and other circuit parameters, we get,

\[
V_{in} = \frac{V_{tp} + V_{DD} + V_{th}(\beta_{n} / \beta_{p})^{1/2}}{1 + (\beta_{n} / \beta_{p})^{1/2}}
\]

• In region 3, both transistors are in saturation. Here they act as sources which can be shown on the equivalent circuit as in Fig. 2.39.
The region 3 is very unstable and changeover from one logic level to the other is rapid.

If \( V_{in} = -V_{tp} \) and \( \beta_p = \beta_n \), equation reduces to
\[
V_{in} = 0.5V_{DD}
\]

This indicates that the changeover between logic levels is symmetrically set about the point corresponding to \( V_{in} = V_{out} = 0.5V_{DD} \) because only at this point the two \( \beta \) factors will be equal. However, in order that \( \beta_n = \beta_p \), the device geometries should satisfy the condition that
\[
\frac{\mu_n W_n}{L_n} = \frac{\mu_p W_p}{L_p}
\]

The mobilities of electrons and holes are inherently unequal.

Thus it is necessary that the width to length ratio \( (W/L) \) of the p-device be two to three times that of the n-device i.e.
\[
\frac{W_p}{L_p} = 2.5 \frac{W_n}{L_n}
\]

However mobility \( \mu \) is not constant and is affected by the transfer electric field in the channel which is a function of \( V_{GS} \). Thus mobility is dependent on \( V_{in} \). The actual mobility is given by the empirical relation
\[
\mu = \frac{\mu_z}{1 - \phi(V_{GS} - V_t)}
\]

where \( \mu_z \) is the mobility with zero transverse field, \( \phi \) is a constant approximately equal to 0.5. Thus equal \( \beta \) values exist only around the point of symmetry when \( V_{in} = V_{out} = 0.5V_{DD} \).

By keeping minimum size geometries for both p- and n-devices, effect \( \beta \) ratio is minimized. Variation of \( \beta \) causes transfer characteristic of the inverter to change as indicated in Fig. 240. However, the changes indicated are highly exaggerated and would apply for quite large \( \beta \) ratio variations (e.g. upto 10 : 1).
2.16 BiCMOS Inverter

- The very approach of BiCMOS is to exploit the advantageous characteristics of bipolar and CMOS technologies. Hence in BiCMOS design, the rational approach is to use MOS switches to perform the logic function and bipolar transistors to drive the output loads. The basic logic element here too is the inverter and a simple BiCMOS inverter circuit can be imagined like the one shown in Fig. 2.41.

![Fig. 2.41 A simple BiCMOS inverter](image-url)
• The inverter circuit consists of two bipolar transistors $T_1$ and $T_2$, one nMOS transistor $T_3$ and one pMOS transistor $T_4$. Both the MOS devices are enhancement mode devices. The functioning of the circuit is as follows:
  a) With $V_{in}$ at logic 0 i.e. 0 volts (GND), $T_3$ is off which keeps $T_1$ non-conducting. However $T_4$ is on and supplies base current to $T_2$ which conducts and acts as a current source to charge the load $C_L$ toward +5 volts ($V_{DD}$). The output $V_{out}$ goes to +5 V less the base to emitter drop $V_{BE}$ of $T_2$.
  b) When $V_{in} =$ logic 1 i.e. +5 V ($V_{DD}$), $T_4$ is off so that $T_2$ will be non-conducting. But $T_3$ is on and supplies current to base of $T_1$ which conducts and acts as a current sink to the load $C_L$ which discharges through it to 0 volts (GND). The $V_{out}$ falls to 0 volts plus the saturation voltage $V_{CESat}$ between collector and emitter of $T_1$.
  c) Charging and discharging of the load $C_L$ is very fast because transistors $T_1$ and $T_2$ present low impedances when turned on into saturation.
  • The output logic levels will approximate the rail voltages since $V_{CESat}$ is quite small and $V_{BE}$ equals 0.7 volts approximately. The inverter offers a low output impedance and a high input impedance. It occupies a relatively small area but still has a high current drive capability. The inverter circuit has high noise margins.
  • However there is a constant D.C. path between the rails through $T_3$ and $T_1$ which allows a significant static current flow whenever $V_{in} =$ logic 1. This is not a desirable arrangement. Also, there is another problem, that there is no discharge path for current from the base of either npn transistor when it is being turned off. This adversely affects the speed of action of the circuit.
  • The problem of the D.C. path through $T_1$ and $T_3$ is eliminated in an improved inverter circuit shown in Fig. 2.42. However the output voltage swing gets reduced because the output cannot go below the base to emitter voltage $V_{BE}$ of transistor $T_1$.

![Diagram of BiCMOS inverter with no static current flow](image_url)

**Fig. 2.42** An alternative BiCMOS inverter with no static current flow
- A further improvement in inverter circuit can be achieved using resistors as shown in Fig. 2.43. Here, the resistors provide an improved swing of output voltage when either bipolar transistor is off.

Fig. 2.43 An improved BiCMOS inverter with better output logic levels

- They also provide discharge paths for the base currents during turn-off. However, fabricating resistors of suitable values in not always convenient and may occupy larger space. Hence other arrangements, like the one shown in Fig. 2.44 are used.

Fig. 2.44 An improved BiCMOS inverter using MOS transistors for base current discharge

- In the circuit shown in Fig. 2.44, arrangement is made to turn on transistors $T_5$ and $T_6$ when $T_2$ and $T_1$ respectively are being turned off. That is when $T_2$ is to be tuned off, $T_5$ gets turned on and provides discharge path for base current of $T_2$. Thus we observe that BiCMOS inverters are more suitable where high load current sinking and sourcing is required.
Solved Examples

Example 2.1: An nMOS transistor is operated in triode region with following parameters: \( V_{gs} = 4 \text{V}, \ V_{tn} = 1 \text{V}, \ V_{ds} = 2 \text{V}, \ \frac{W}{L} = 100, \ \mu_n \cdot C_{ox} = -90 \mu\text{A}/\text{V}^2 \). Find its drain current and drain to source resistance.

Solution: nMOS transistor is operating in triode region
\[
V_{ds} < \left( V_{gs} - V_{tn} \right)
\]

Given: \( V_{gs} = 4 \text{V} \)
\( V_{tn} = 1 \text{V} \)
\( V_{ds} = 2 \text{V} \)
\( \frac{W}{L} = 100 \)

\( \mu_n \cdot C_{ox} = 90 \mu\text{A}/\text{V}^2 \)

\( I_D = ? \) (to find)

\( R_{ds} = ? \) (to find)

Drain current of nMOS transistor in triode region is given by,
\[
I_D = \mu_n \cdot C_{ox} \cdot \frac{W}{L} \left[ \left( V_{gs} - V_{tn} \right) V_{ds} - \frac{V_{ds}^2}{2} \right]
\]

\[
I_D = (90 \times 10^{-6}) 100 \left[ (4 - 1) 2 - \frac{2^2}{2} \right]
\]

\( I_D = 9 \times 10^{-3} \left[ 6 - 2 \right] \)

\( I_D = 0.036 \text{ Amp} \)

\( I_D = 36 \text{ mA} \)

... Ans.

Drain to source resistance \( R_{ds} \) is given by,
\[
R_{ds} = \frac{V_{ds}}{I_D}
\]

\[
R_{ds} = \frac{2}{(36 \times 10^{-3})}
\]

\( R_{ds} = 55.56 \text{\Omega} \).
Example 2.4: An nMOS transistor is operating in active region with the following parameters: \( V_{gs} = 3.9 \, V \), \( V_{in} = 1 \, V \), \( \frac{W}{L} = 100 \), \( \mu_n \cdot C_{ox} = 90 \mu A/V^2 \). Find \( I_D \) and \( R_{ds} \).

Solution: nMOS transistor is operating in active region.

\[
V_{ds} = V_{gs} - V_{tn}
\]

Given: \( V_{gs} = 3.9 \, V \)

\[
V_{tn} = 1 \, V
\]

\[
\frac{W}{L} = 100
\]

\[
\mu_n \cdot C_{ox} = 90 \mu A/V^2
\]

\[
I_D = ? \text{ (to find)}
\]

\[
R_{ds} = ? \text{ (to find)}
\]

i) Drain current \( (I_D) \): Drain current of nMOS transistor operating in active region is

\[
I_D = \frac{\mu_n \cdot C_{ox} \cdot W}{2 \cdot L} \cdot (V_{gs} - V_{tn})^2
\]

\[
I_D = \frac{90 \times 10^{-6}}{2} \times 100 (3.9 - 1)^2
\]

\[
I_D = 37.845 \, mA \quad \text{... Ans.}
\]

ii) Drain to source resistance \( (R_{ds}) \):

\[
R_{ds} = \frac{V_{ds}}{I_D}
\]

\[
R_{ds} = \frac{V_{gs} - V_{tn}}{I_D} \quad \because \text{transistor is in active region i.e.}
\]

\[
V_{ds} = V_{gs} - V_{tn}
\]

\[
R_{ds} = \frac{V_{gs} - V_{tn}}{I_D}
\]

\[
R_{ds} = \frac{(3.9 - 1)}{(37.845 \times 10^{-3})} \quad \text{(Ans.)}
\]

\[
R_{ds} = 76.628 \, \Omega
\]
**Example 2.5**: An nMOS transistor is operating in saturation region with following parameters:

\[ V_{gs}^{(n)} = 5 \text{ V}, \ V_{tn} = 1.2 \text{ V}, \ \frac{W}{L} = 110, \ \mu_n \cdot C_{ox} = 110 \ \mu\text{A}/\text{V}^2. \] Find transconductance of the device.

**Solution**: nMOS transistor is operating in saturation region.

Given:

\[
\begin{align*}
V_{ds} &= V_{gs} - V_t \\
V_{gs} &= 5 \text{ V} \\
V_{tn} &= 1.2 \text{ V} \\
\frac{W}{L} &= 110 \\
\mu_n \cdot C_{ox} &= 110 \ \mu\text{A}/\text{V}^2 \\
g_m &= ? \ (\text{to find})
\end{align*}
\]

Transconductance of nMOS transistor is given by,

\[ g_m = \frac{C_g \cdot \mu_n}{L^2} \cdot V_{ds} \]

\[
g_m = \frac{c_g \cdot \mu_n \ (V_{gs} - V_t)}{L^2} \quad \because \ V_{ds} = V_{gs} - V_t
\]

\[
g_m = \frac{(C_{ox} \cdot WL) \mu_n \ (V_{gs} - V_t)}{L^2} \quad \because \ C_g = C_{ox} \cdot WL
\]

\[
g_m = \mu_n \cdot C_{ox} \cdot \frac{W}{L} \ (V_{gs} - V_t) \quad \text{Substituting values}
\]

\[
g_m = (110 \times 10^{-6}) \ 110 (5 - 1.2)
\]

\[
g_m = 45.98 \times 10^{-3} \ \mu\text{A} \quad \ldots \ \text{Ans.}
\]

**Example 2.6**: Calculate the threshold voltage for a transistor at 300 K for a process with a Si substrate with \( N_A = 1.80 \times 10^{16} \) and SiO2 gate oxide with thickness 200\( \text{Å} \). Assume \( \phi_m = -0.9 \text{ V} \); \( Q_f = 0 \).

**Solution**: Given:

\[
N_A = 1.80 \times 10^{16}
\]

For Si:

\[
\begin{align*}
N_i &= 1.45 \times 10^{10} \ \text{cm}^{-3} \ \text{at} \ 300 \ \text{°K} \\
t_{ox} &= 200 \text{ Å} = 0.2 \times 10^{-5}
\end{align*}
\]
\[ \phi_{ms} = -0.9 \text{ V} \]
\[ \varepsilon_{si} = 1.06 \times 10^{-12} \text{ farad/cm} \]

Bulk potential \( \phi_b \) is given by,

\[
\phi_b = \frac{K}{q} \ln \left( \frac{N_A}{N_i} \right)
\]

\[ \phi_b = 0.02586 \ln \left( \frac{1.8 \times 10^{16}}{1.45 \times 10^{10}} \right) \]

\[ \frac{K}{q} = 0.02586 \text{ at 300 } \degree \text{K} \]

\[ \phi_b = 0.36 \text{ volts} \]

Oxide capacitance \( C_{ox} \) is given by,

Oxide capacitance \( C_{ox} \) is given by,

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

\[ C_{ox} = \frac{3.9 \times 8.85 \times 10^{-14}}{0.2 \times 10^{-5}} \]

\[ C_{ox} = 1.726 \times 10^{-7} \text{ F/cm}^2 \]

Threshold voltage \( V_t \) is given by

\[
V_t = \phi_{ms} + \frac{\sqrt{2 \varepsilon_{si} q N_A}}{C_{ox}} \frac{2 \phi_b}{C_{ox}} + 2 \phi_b
\]

\[ V_t = -0.9 + \frac{\sqrt{2 \times (1.06 \times 10^{-12}) \times (1.6 \times 10^{-19}) \times (1.8 \times 10^{16}) \times (2 \times 0.36)}}{(1.726 \times 10^{-7})} + (2 \times 0.36) \]

\[ V_t = -0.9 + 0.384 + 0.72 \]

\[ V_t = 0.16 \text{ volts} \]

... Ans.
Points to Remember

1. MOS transistor is a majority-carrier device in which the current in a conducting channel between source and drain is controlled by gate voltage.

2. In nMOS transistors, the majority carriers are electrons and in pMOS transistors, the majority carriers are holes.

3. With negative gate voltage, the mobile positively charged holes are attracted to the region beneath the gate. This is called accumulation mode.

4. With small positive gate voltage, the holes in the body are repelled from the region directly beneath the gate forming depletion region below the gate.

5. With high positive gate voltage (greater than threshold voltage $V_t$), attracting more positive charges to gate. The holes are repelled further and free electrons attracted near to gate. This conductive layer of electrons in the p-type body is called inversion layer.

6. A CMOS transistor can operate in three modes: Cut-off, Linear and Saturation.

7. The delay of MOS circuits is determined by the time required to charge or discharge the capacitance of the circuit.

8. The transit time for travel of electrons or hole from source to drain ($\tau_{sd}$) is given by:

$$\tau_{sd} = \frac{L^2}{\mu V_{ds}}$$

9. $I_{ds}$ in non-saturated region is given by:

$$I_{ds} = C_{ox} \mu \frac{W}{L} \left[ (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right]$$

10. $I_{ds}$ in saturation region is given by:

$$I_{ds} = C_{ox} \mu \frac{W}{2L} (V_{gs} - V_t)^2$$

11. Different MOS capacitance models are:

1. Simple MOS capacitance models.
2. Detailed MOS gate capacitance model.
3. Detailed MOS Diffusion capacitance model.

12. The parasitic capacitance arise from reverse biased p-n junction and are called as diffusion capacitance.
13. There are three types of diffusion regions.
   1) Isolated diffusion region.
   2) Shared contacted diffusion region.
   3) Merged into uncontacted region.
14. Gate capacitance has two components.
   1) Intrinsic capacitance
   2) Overlap capacitance
15. The parasitic capacitance exists in the reverse biased p-n junction between source diffusion and the body. The capacitance is dependent on area AS and sidewall parameter PS of the source diffusion region geometry.

16. Total capacitance seen from gate terminal of CMOS transistor is,
    \[ C_g = C_{gs} + C_{gb} + C_{gb} \]
17. The gate resistance along with gate parasitic capacitance introduces RC delay that causes oscillations in the circuit.
18. The variation of threshold voltage due to source to substrate voltage is referred as body effect.
19. CMOS voltage transfer characteristics is dived into five regions.
20. Power dissipation in a CMOS inverter is categorized in static and dynamic power dissipation.
21. The parallel combination of MOSFET and inverter is known as transmission gate.