UNIT II - BASIC ELECTRICAL PROPERTIES

DRAIN-TO-SOURCE CURRENT I_{ds} versus VOLTAGE V_{ds} RELATIONSHIPS

The whole concept of the MOS transistor evolves from the use of a voltage on the gate to induce a charge in the channel between source and drain, which may then be caused to move from source to drain under the influence of an electric field created by voltage V_{ds} applied between drain and source. Since the charge induced is dependent on the gate to source voltage V_{gs}, then I_{ds} is dependent on both V_{gs} and V_{ds}. Consider a structure, as in Figure 2.1, in which electrons will flow source to drain:

$$I_{ds} = -I_{sd} = \frac{\text{Charge induced in channel } (Q_c)}{\text{Electron transit time } (\tau)} \tag{2.1}$$

First, transit time:

$$\tau_{sd} = \frac{\text{Length of channel } (L)}{\text{Velocity } (v)}$$

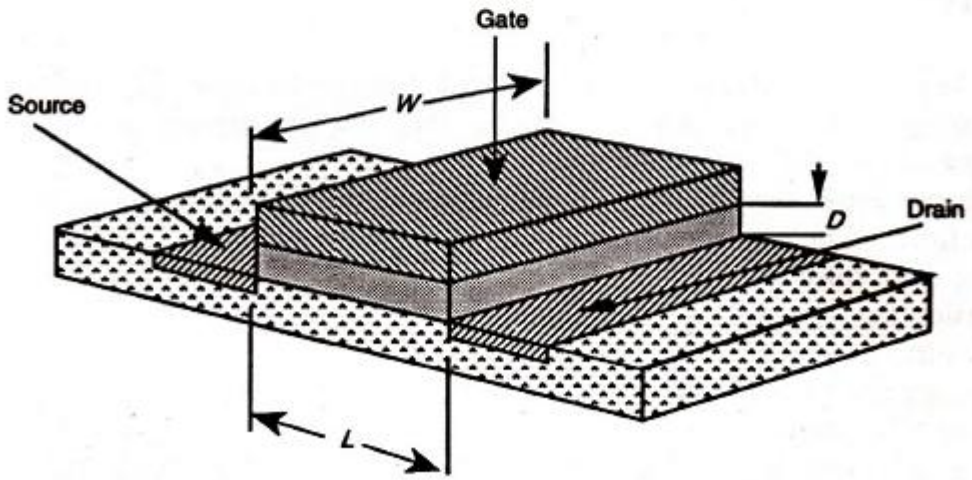


FIGURE 2.1 nMOS transistor structure.

but velocity

$$v = \mu E_{ds}$$

where

μ = electron or hole mobility (surface)

E_{ds} = electric field (drain to source)

Now

$$E_{ds} = \frac{V_{ds}}{L}$$

so that

$$v = \frac{\mu V_{ds}}{L}$$

Thus

$$\tau_{sd} = \frac{L^2}{\mu V_{ds}} \tag{2.2}$$

Typical values of μ at room temperature are:

$$\mu_n \doteq 650 \text{ cm}^2/\text{V sec (surface)}$$
$$\mu_p \doteq 240 \text{ cm}^2/\text{V sec (surface)}$$

The Non-saturated Region

Charge induced in channel due to gate voltage is due to the voltage difference between the gate and the channel, V_{gs} (assuming substrate connected to source). Now note that the voltage along the channel varies linearly with distance X from the source due to the IR drop in the channel and assuming that the device is not saturated then the average value is $V_{ds}/2$.

Furthermore, the effective gate voltage $V_g = V_{gs} - V_t$ where V_t is the threshold voltage needed to invert the charge under the gate and establish the channel.

Note that the charge/unit area = $E_g \varepsilon_{ins} \varepsilon_0$. Thus induced charge

$$Q_c = E_g \varepsilon_{ins} \varepsilon_0 WL$$

where

E_g = average electric field gate to channel
ε_{ins} = relative permittivity of insulation between gate and channel
ε_0 = permittivity of free space

(*Note:* $\varepsilon_0 = 8.85 \times 10^{-14}$F cm^{-1}; $\varepsilon_{ins} \doteq 4.0$ for silicon dioxide)

Now

$$E_g = \frac{\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right)}{D}$$

where D = oxide thickness.

Thus

$$Q_c = \frac{WL\varepsilon_{ins}\varepsilon_0}{D}\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right) \tag{2.3}$$

Now, combining equations (2.2) and (2.3) in equation (2.1), we have

$$I_{ds} = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}\frac{W}{L}\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right)V_{ds}$$

or

$$I_{ds} = K\frac{W}{L}\left((V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right) \tag{2.4}$$

in the non-saturated or resistive region where $V_{ds} < V_{gs} - V_t$ and

$$K = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}$$

The factor W/L is, of course, contributed by the geometry and it is common practice to write

$$\beta = K\frac{W}{L}$$

so that

$$I_{ds} = \beta\left((V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right) \qquad (2.4a)$$

which is an alternative form of equation (2.4).

Noting that gate/channel capacitance

$$C_g = \frac{\varepsilon_{ins}\varepsilon_0 WL}{D} \text{ (parallel plate)}$$

we also have

$$K = \frac{C_g\mu}{WL}$$

so that

$$I_{ds} = \frac{C_g\mu}{L^2}\left((V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right) \qquad (2.4b)$$

which is a further alternative form of equation (2.4).

Sometimes it is convenient to use *gate capacitance per unit area* C_0 (which is often denoted C_{ox}) rather than C_g in this and other expressions. Noting that

$$C_g = C_0WL$$

we may also write

$$I_{ds} = C_0\mu\frac{W}{L}\left((V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right) \qquad (2.4c)$$

The Saturated Region

Saturation begins when $V_{ds} = V_{gs} - V$, since at this point the IR drop in the channel equals the effective gate to channel voltage at the drain and we may assume that the current remains fairly constant as V_{ds} increases further. Thus

$$I_{ds} = K\frac{W}{L}\frac{(V_{gs} - V_t)^2}{2} \qquad (2.5)$$

or, we may write

$$I_{ds} = \frac{\beta}{2}(V_{gs} - V_t)^2 \qquad (2.5a)$$

or

$$I_{ds} = \frac{C_g \mu}{2L^2} (V_{gs} - V_t)^2 \tag{2.5b}$$

We may also write

$$I_{ds} = C_0 \mu \frac{W}{2L} (V_{gs} - V_t)^2 \tag{2.5c}$$

The expressions derived for I_{ds} hold for both enhancement and depletion mode devices, but it should be noted that the threshold voltage for the nMOS depletion mode device (denoted as V_{td}) is *negative*.

Typical characteristics for nMOS transistors are given in Figure 2.2. pMOS transistor characteristics are similar, with suitable reversal of polarity.

ASPECTS OF MOS TRANSISTOR THRESHOLD VOLTAGE V_t

The gate structure of a MOS transistor consists, electrically, of charges stored in the dielectric layers and in the surface to surface interfaces as well as in the substrate itself.

Switching an enhancement mode MOS transistor from the off to the on state consists in applying sufficient gate voltage to neutralize these charges and enable the underlying silicon to undergo an inversion due to the electric field from the gate.

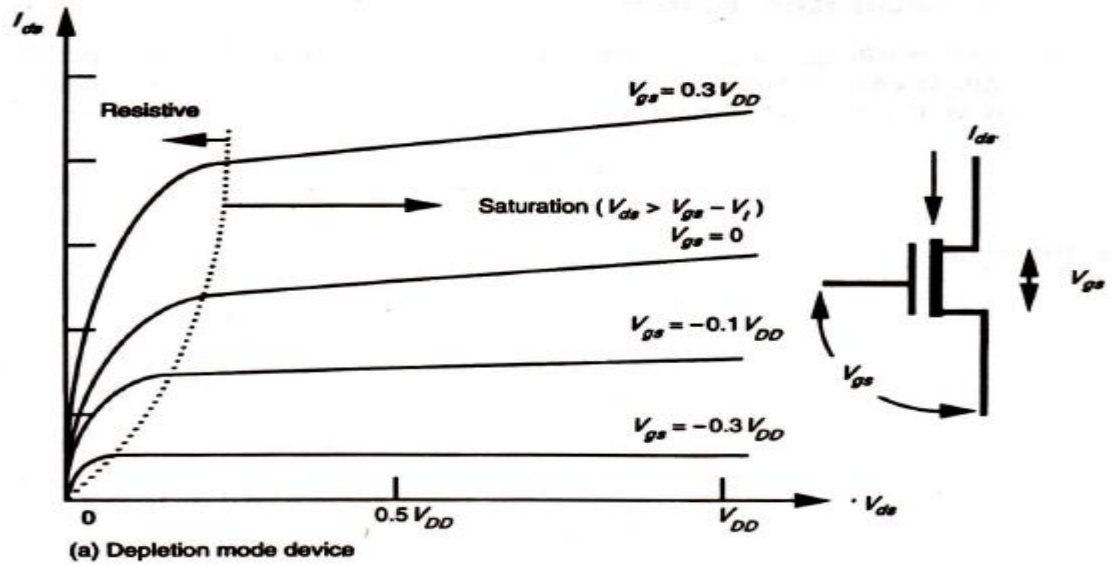
Switching a depletion mode nMOS transistor from the on to the off state consists in applying enough voltage to the gate to add to the stored charge and invert the 'n' implant region to 'p'.

The threshold voltage V_t may be expressed as:

$$V_t = \phi_{ms} \frac{Q_B - Q_{SS}}{C_0} + 2\phi_{fN} \tag{2.6}$$

where

Q_B = the charge per unit area in the depletion layer beneath the oxide
Q_{SS} = charge density at $Si:SiO_2$ interface



(a) Depletion mode device

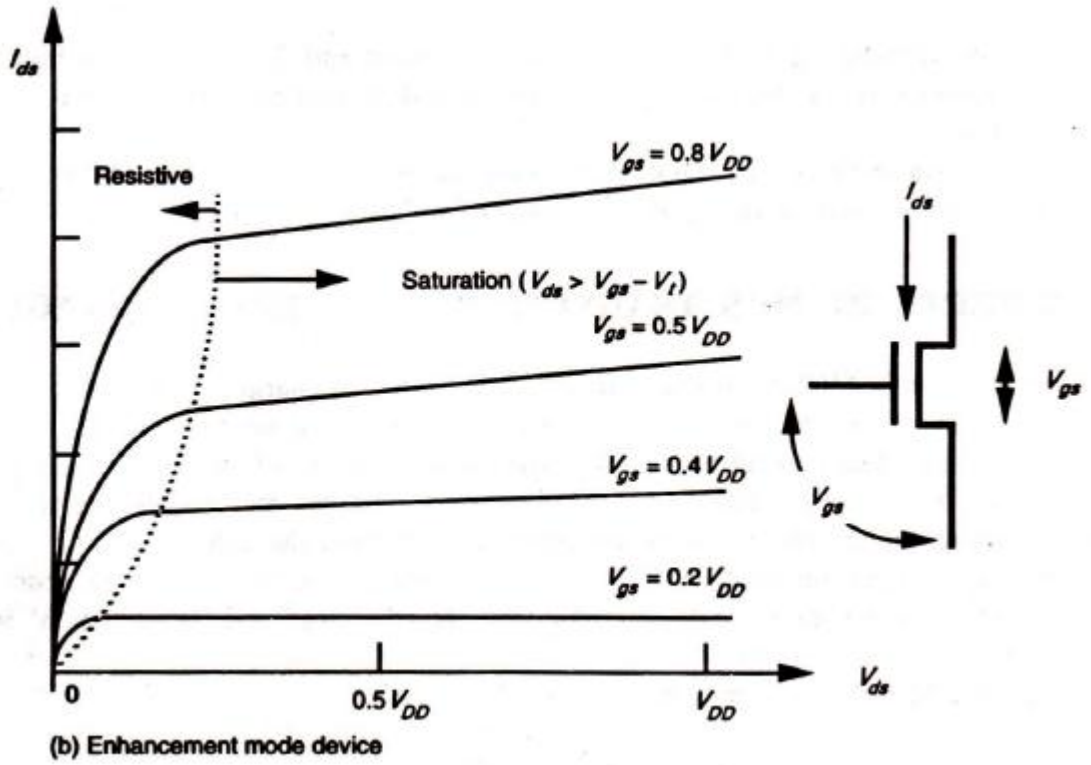


FIGURE 2.2 MOS transistor characteristics.

C_0 = capacitance per unit gate area
ϕ_{ms} = work function difference between gate and Si
ϕ_{fN} = Fermi level potential between inverted surface and bulk Si.

Now, for polysilicon gate and silicon substrate, the value of ϕ_{ms} is negative but negligible, and the magnitude and sign of V_t are thus determined by the balance between the remaining negative term $\frac{-Q_{SS}}{C_0}$ and the other two terms, both of which are positive. To evaluate V_t, each term is determined as follows:

$$Q_B = \sqrt{2\varepsilon_0\varepsilon_{Si}qN(2\phi_{fN} + V_{SB})} \text{ coulomb/m}^2$$

$$\phi_{fN} = \frac{kT}{q} \ln \frac{N}{n_i} \text{ volts}$$

$$Q_{SS} = (1.5 \text{ to } 8) \times 10^{-8} \text{ coulomb/m}^2$$

depending on crystal orientation, and where

V_{SB} = substrate bias voltage (negative w.r.t. source for nMOS, positive for pMOS)
q = 1.6×10^{-19} coulomb
N = impurity concentration in the substrate (N_A or N_D as appropriate)
ε_{si} = relative permittivity of silicon ÷ 11.7
n_i = intrinsic electron concentration (1.6×10^{10}/cm^3 at 300°K)
k = Boltzmann's constant = 1.4×10^{-23} joule/°K

The *body effects* may also be taken into account since the substrate may be biased with respect to the source, as shown in Figure 2.3.

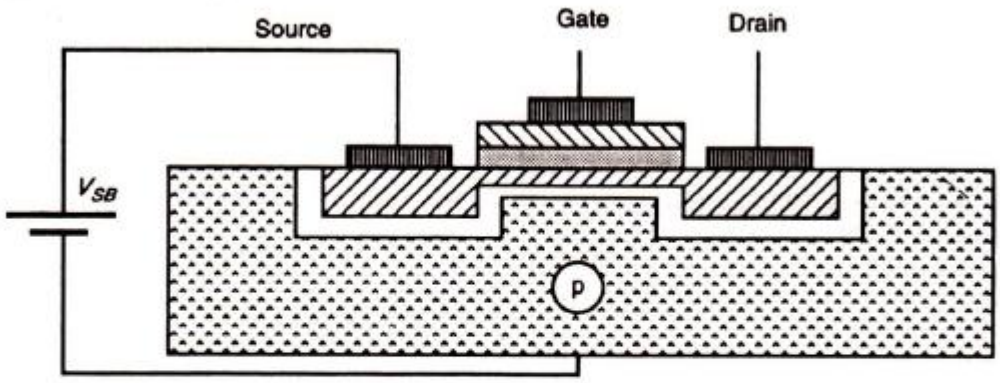


FIGURE 2.3 Body effect (nMOS device shown).

Increasing V_{SB} causes the channel to be depleted of charge carriers and thus the threshold voltage is raised.

Change in V_t is given by $\Delta V_t \doteq \gamma(V_{SB})^{1/2}$ where γ is a constant which depends on substrate doping so that the more lightly doped the substrate, the smaller will be the body effect.

Alternatively, we may write

$$V_t = V_t(0) + \left(\frac{D}{\varepsilon_{ins}\varepsilon_0}\right)\sqrt{2\varepsilon_0\varepsilon_{Si}QN}.\,(V_{SB})^{1/2}$$

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

To establish the magnitude of such effects, typical figures for V_t are as follows:

For nMOS enhancement mode transistors:

$V_{SB} = 0$ V; $V_t = 0.2V_{DD}$ (= +1 V for V_{DD} = +5 V)
$V_{SB} = 5$ V; $V_t = 0.3V_{DD}$ (= +1.5 V for V_{DD} = +5 V)

} Similar but negative values for pMOS

For nMOS depletion mode transistors:

$V_{SB} = 0$ V; $V_{td} = -0.7V_{DD}$ (= −3.5 V for V_{DD} = +5 V)
$V_{SB} = 5$ V; $V_{td} = -0.6V_{DD}$ (= −3.0 V for V_{DD} = +5 V)

MOS TRANSISTOR TRANSCONDUCTANCE g_m AND OUTPUT CONDUCTANCE g_{ds}

Transconductance expresses the relationship between output current I_{ds} and the input voltage V_{gs} and is defined as

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}}\,|V_{ds} = \text{constant}$$

To find an expression for g_m in terms of circuit and transistor parameters, consider that the charge in channel Q_c is such that

$$\frac{Q_c}{I_{ds}} = \tau$$

where τ is transit time. Thus change in current

$$\delta I_{ds} = \frac{\delta Q_c}{\tau_{ds}}$$

Now

$$\tau_{ds} = \frac{L^2}{\mu V_{ds}} \qquad \text{(from 2.2)}$$

Thus

$$\delta I_{ds} = \frac{\delta Q_c V_{ds} \mu}{L^2}$$

but change in charge

$$\delta Q_c = C_g \delta V_{gs}$$

so that

$$\delta I_{ds} = \frac{C_g \delta V_{gs} \mu V_{ds}}{L^2}$$

Now

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{C_g \mu V_{ds}}{L^2}$$

In saturation

$$V_{ds} = V_{gs} - V_t$$

$$g_m = \frac{C_g \mu}{L^2} (V_{gs} - V_t) \qquad (2.7)$$

and substituting for $C_g = \frac{\varepsilon_{ins} \varepsilon_0 WL}{D}$

$$g_m = \frac{\mu \varepsilon_{ins} \varepsilon_0}{D} \frac{W}{L} (V_{gs} - V_t) \qquad (2.7a)$$

Alternatively,

$$g_m = \beta (V_{gs} - V_t)$$

It is possible to increase the g_m of a MOS device by increasing its width. However, this will also increase the input capacitance and area occupied.

A reduction in the channel length results in an increase in ω_0 owing to the higher g_m. However, the gain of the MOS device decreases owing to the strong degradation of the output resistance = $1/g_{ds}$.

The output conductance g_{ds} can be expressed by

$$g_{ds} = \frac{\delta I_{ds}}{\delta V_{gs}} = \lambda . I_{ds}\ \alpha \left(\frac{1}{L}\right)^2$$

Here the strong dependence on the channel length is demonstrated as

$$\lambda \alpha \left(\frac{1}{L}\right) \text{ and } I_{ds}\ \alpha \left(\frac{1}{L}\right)$$

for the MOS device.

MOS TRANSISTOR FIGURE OF MERIT ω_0

An indication of frequency response may be obtained from the parameter ω_0 where

$$\omega_0 = \frac{g_m}{C_g} = \frac{\mu}{L^2}\ (V_{gs} - V_t) \left(= \frac{1}{\tau_{sd}}\right) \tag{2.8}$$

This shows that switching speed depends on gate voltage above threshold and on carrier mobility and inversely as the square of channel length. A fast circuit requires that g_m be as high as possible.

Electron mobility on a (100) oriented n-type inversion layer surface (μ_n) is larger than that on a (111) oriented surface, and is in fact about three times as large as hole mobility on a (111) oriented p-type inversion layer. Surface mobility is also dependent on the effective gate voltage ($V_{gs} - V_t$).

For faster nMOS circuits, then, one would choose a (100) oriented p-type substrate in which the inversion layer will have a surface carrier mobility $\mu_n \doteqdot 650$ cm^2/V sec at room temperature.

Compare this with the typical bulk mobilities

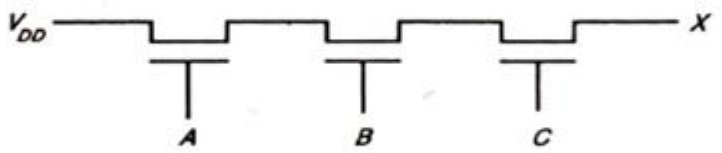
$$\mu_n = 1250 \text{ cm}^2/\text{V sec}$$

$$\mu_p = 480 \text{ cm}^2/\text{V sec}$$

from which it will be seen that $\frac{\mu_s}{\mu} \doteqdot 0.5$ (where μ_s = surface mobility and μ = bulk mobility).

THE PASS TRANSISTOR

Unlike bipolar transistors, the isolated nature of the gate allows MOS transistors to be used as switches in series with lines carrying logic levels in a way that is similar to the use of relay contacts. This application of the MOS device is called the *pass transistor* and switching logic arrays can be formed—for example, an *And* array as in Figure 2.4.



$X = A.B.C$ (Logic 1 = $V_{DD} - V_t$)

$\bar{X} = ?$

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

FIGURE 2.4 Pass transistor And gate.

THE nMOS INVERTER

A basic requirement for producing a complete range of logic circuits is the inverter. This is needed for restoring logic levels, for *Nand* and *Nor* gates, and for sequential and memory circuits of various forms.

The basic inverter circuit requires a transistor with source connected to ground and a load resistor of some sort connected from the drain to the positive supply rail V_{DD}. The output is taken from the drain and the input applied between gate and ground.

Resistors are not conveniently produced on the silicon substrate; even modest values occupy excessively large areas so that some other form of load resistance is required. A convenient way to solve this problem is to use a depletion mode transistor as the load, as shown in Figure 2.5.

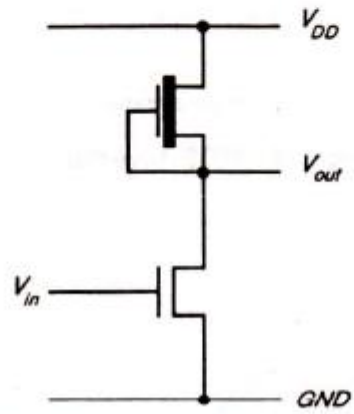
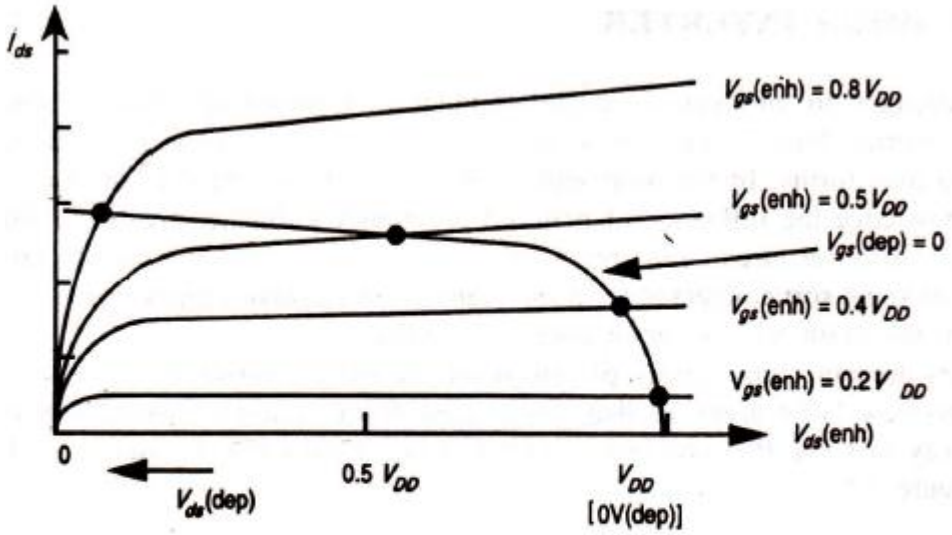


FIGURE 2.5 nMOS inverter.

Now:

- With no current drawn from the output, the currents I_{ds} for both transistors must be equal.
- For the depletion mode transistor, the gate is connected to the source so it is always on and only the characteristic curve $V_{gs} = 0$ is relevant.
- In this configuration the depletion mode device is called the pull-up (p.u.) and the enhancement mode device the pull-down (p.d.) transistor.
- To obtain the inverter transfer characteristic we superimpose the $V_{gs} = 0$ depletion mode characteristic curve on the family of curves for the enhancement mode device, noting that maximum voltage across the enhancement mode device corresponds to minimum voltage across the depletion mode transistor.
- The points of intersection of the curves as in Figure 2.6 give points on the transfer characteristic, which is of the form shown in Figure 2.7.
- Note that as V_{in}(=V_{gs} p.d. transistor) exceeds the p.d. threshold voltage current begins to flow. The output voltage V_{out} thus decreases and the subsequent increases in V_{in} will cause the p.d. transistor to come out of saturation and become resistive. Note that the p.u. transistor is initially resistive as the p.d. turns on.



$V_{ds}(enh) = V_{DD} - V_{ds}(dep) = V_{out}$
$V_{gs}(enh) = V_{in}$ . . . intersection points give transfer characteristic

FIGURE 2.6 Derivation of nMOS inverter transfer characteristic.

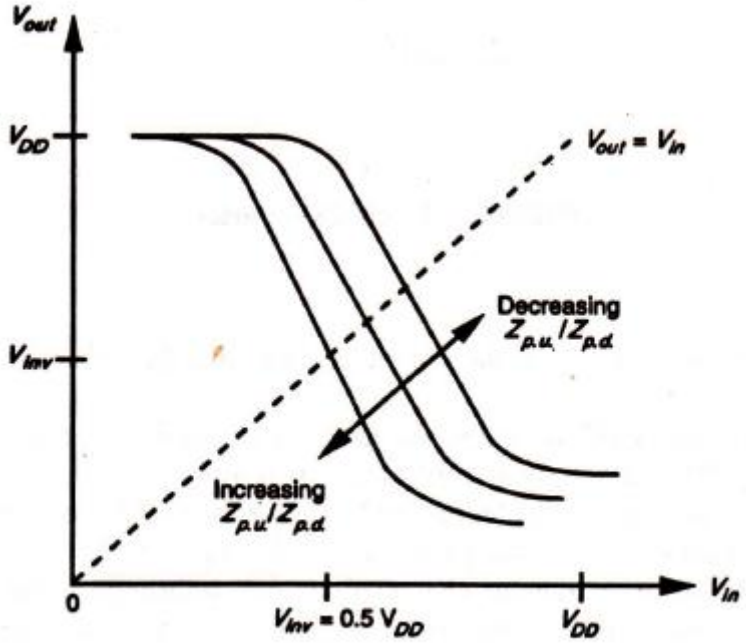


FIGURE 2.7 nMOS inverter transfer characteristic.

- During transition, the slope of the transfer characteristic determines the gain:

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

- The point at which $V_{out} = V_{in}$ is denoted as V_{inv} and it will be noted that the transfer characteristics and V_{inv} can be shifted by variation of the ratio of pull-up to pull-down resistances (denoted $Z_{p.u.}/Z_{p.d.}$ where Z is determined by the length to width ratio of the transistor in question).

DETERMINATION OF PULL-UP TO PULL-DOWN RATIO ($Z_{p.u}/Z_{p.d.}$) FOR AN nMOS INVERTER DRIVEN BY ANOTHER nMOS INVERTER

Consider the arrangement in Figure 2.8 in which an inverter is driven from the output of another similar inverter. Consider the depletion mode transistor for which $V_{gs} = 0$ under all conditions, and further assume that in order to cascade inverters without degradation of levels we are aiming to meet the requirement

$$V_{in} = V_{out} = V_{inv}$$

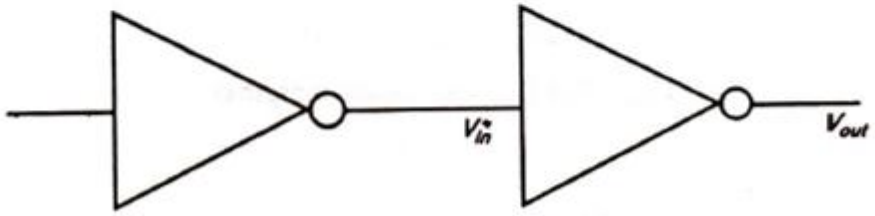


FIGURE 2.8 nMOS inverter driven directly by another inverter.

For equal margins around the inverter threshold, we set $V_{inv} = 0.5V_{DD}$. At this point both transistors are in saturation and

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

In the depletion mode

$$I_{ds} = K\frac{W_{p.u.}}{L_{p.u.}}\frac{(-V_{td})^2}{2} \text{ since } V_{gs} = 0$$

and in the enhancement mode

$$I_{ds} = K\frac{W_{p.d.}}{L_{p.d.}}\frac{(V_{inv} - V_t)^2}{2} \text{ since } V_{gs} = V_{inv}$$

Equating (since currents are the same) we have

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

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

Now write

$$Z_{p.d.} = \frac{L_{p.d.}}{W_{p.d.}};\ Z_{p.u.} = \frac{L_{p.u.}}{W_{p.u.}}$$

we have

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

whence

$$V_{inv} = V_t - \frac{V_{td}}{\sqrt{Z_{p.u.}/Z_{p.d.}}} \tag{2.9}$$

Now we can substitute typical values as follows:

$$V_t = 0.2V_{DD};\ V_{td} = -0.6V_{DD}$$

$$V_{inv} = 0.5V_{DD} \text{ (for equal margins)}$$

thus, from equation (2.9)

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

whence

$$\sqrt{Z_{p.u.}/Z_{p.d.}} = 2$$

and thus

$$Z_{p.u.}/Z_{p.d.} = 4/1$$

for an inverter directly driven by an inverter.

PULL-UP TO PULL-DOWN RATIO FOR AN nMOS INVERTER DRIVEN THROUGH ONE OR MORE PASS TRANSISTORS

Now consider the arrangement of Figure 2.9 in which the input to inverter 2 comes from the output of inverter 1 but passes through one or more nMOS transistors used as switches in series (called *pass transistors*).

We are concerned that connection of pass transistors in series will degrade the logic 1 level into inverter 2 so that the output will not be a proper logic 0 level. The critical condition is when point A is at 0 volts and B is thus at V_{DD}, but the voltage into inverter 2 at point C is now reduced from V_{DD} by the threshold voltage of the series pass transistor. With all pass transistor gates connected to V_{DD} (as shown in Figure 2.8), there is a loss of

V_{tp}, however many are connected in series, since no static current flows through them and there can be no voltage drop in the channels. Therefore, the input voltage to inverter 2 is

$$V_{in2} = V_{DD} - V_{tp}$$

where

V_{tp} = threshold voltage for a pass transistor.

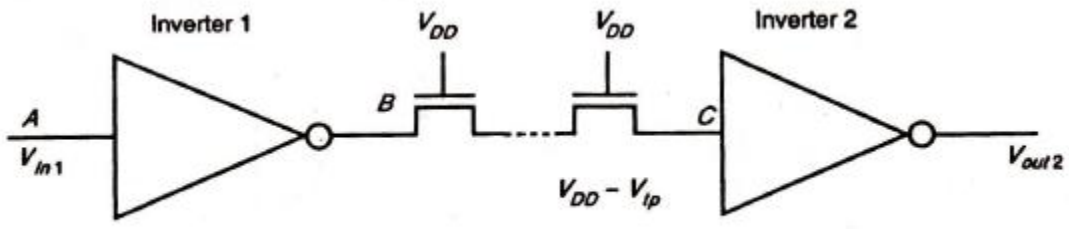


FIGURE 2.9 Pull-up to pull-down ratios for inverting logic coupled by pass transistors.

We must now ensure that for this input voltage we get out the same voltage as would be the case for inverter 1 driven with input = V_{DD}.

Consider inverter 1 (Figure 2.10(a)) with input = V_{DD}. If the input is at V_{DD}, then the p.d. transistor T_2 is conducting but with a low voltage across it; therefore, it is in its resistive region represented by R_1 in Figure 2.10. Meanwhile, the p.u. transistor T_1 is in saturation and is represented as a current source.

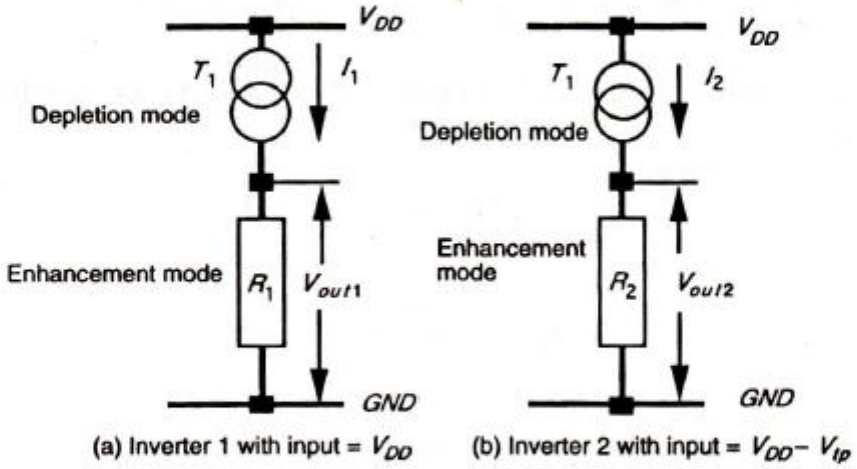


FIGURE 2.10 Equivalent circuits of inverters 1 and 2.

For the p.d. transistor

$$I_{ds} = K\frac{W_{p.d.1}}{L_{p.d.1}}\left((V_{DD} - V_t)\,V_{ds1} - \frac{V_{ds1}^2}{2}\right) \qquad \text{(from 2.4)}$$

Therefore

$$R_1 = \frac{V_{ds1}}{I_{ds}} = \frac{1}{K}\frac{L_{p.d.1}}{W_{p.d.1}}\left(\frac{1}{V_{DD} - V_t - \dfrac{V_{ds1}}{2}}\right)$$

Note that V_{ds1} is small and $V_{ds1}/2$ may be ignored.

Thus

$$R_1 \doteq \frac{1}{K}Z_{p.d.1}\left(\frac{1}{V_{DD} - V_t}\right)$$

Now, for depletion mode p.u. in saturation with $V_{gs} = 0$

$$I_1 = I_{ds} = K\frac{W_{p.u.1}}{L_{p.u.1}}\frac{(-V_{td})^2}{2} \qquad \textbf{(from 2.5)}$$

The product

$$I_1R_1 = V_{out\,1}$$

Thus

$$V_{out1} = I_1 R_1 = \frac{Z_{p.d.1}}{Z_{p.u.1}} \left(\frac{1}{V_{DD} - V_t} \right) \frac{(V_{td})^2}{2}$$

Consider inverter 2 (Figure 2.10(b)) when input = $V_{DD} - V_{tp}$. As for inverter 1

$$R_2 \doteqdot \frac{1}{K} Z_{p.d.2} \frac{1}{((V_{DD} - V_{tp}) - V_t}$$

$$I_2 = K \frac{1}{Z_{p.u.2}} \frac{(-V_{td})^2}{2}$$

whence

$$V_{out2} = I_2 R_2 = \frac{Z_{p.d.2}}{Z_{p.u.2}} \left(\frac{1}{V_{DD} - V_{tp} - V_t} \right) \frac{(-V_{td})^2}{2}$$

If inverter 2 is to have the same output voltage under these conditions then $V_{out\,1} = V_{out\,2}$. That is

$$I_1 R_1 = I_2 R_2$$

Therefore

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

Taking typical values

$$V_t = 0.2 V_{DD}$$

$$V_{tp} = 0.3 V_{DD}^*$$

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

Therefore

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

Summarizing for an nMOS inverter:

- An inverter driven directly from the output of another should have a $Z_{p.u.}/Z_{p.d.}$ ratio of $\geq 4/1$.
- An inverter driven through one or more pass transistors should have a $Z_{p.u.}/Z_{p.d.}$ ratio of $\geq 8/1$.

Note: It is the driven, *not* the driver, whose ratio is affected.

ALTERNATIVE FORMS OF PULL-UP

Up to now we have assumed that the inverter circuit has a depletion mode pull-up transistor as its load. There are, however, at least four possible arrangements:

1. *Load resistance* R_L (Figure 2.11). This arrangement is not often used because of the large space requirements of resistors produced in a silicon substrate.

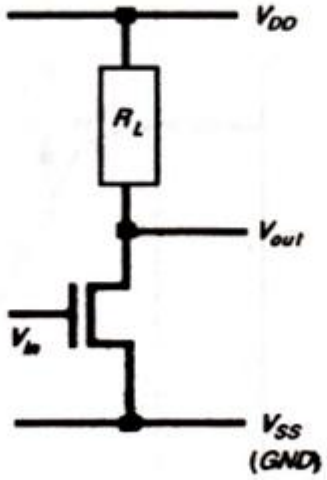


FIGURE 2.11 Resistor pull-up.

2. *nMOS depletion mode transistor pull-up* (Figure 2.12).
   (a) Dissipation is high since rail to rail current flows when V_{in} = logical 1.
   (b) Switching of output from 1 to 0 begins when V_{in} exceeds V_t of p.d. device.
   (c) When switching the output from 1 to 0, the p.u. device is non-saturated initially and this presents lower resistance through which to charge capacitive loads.

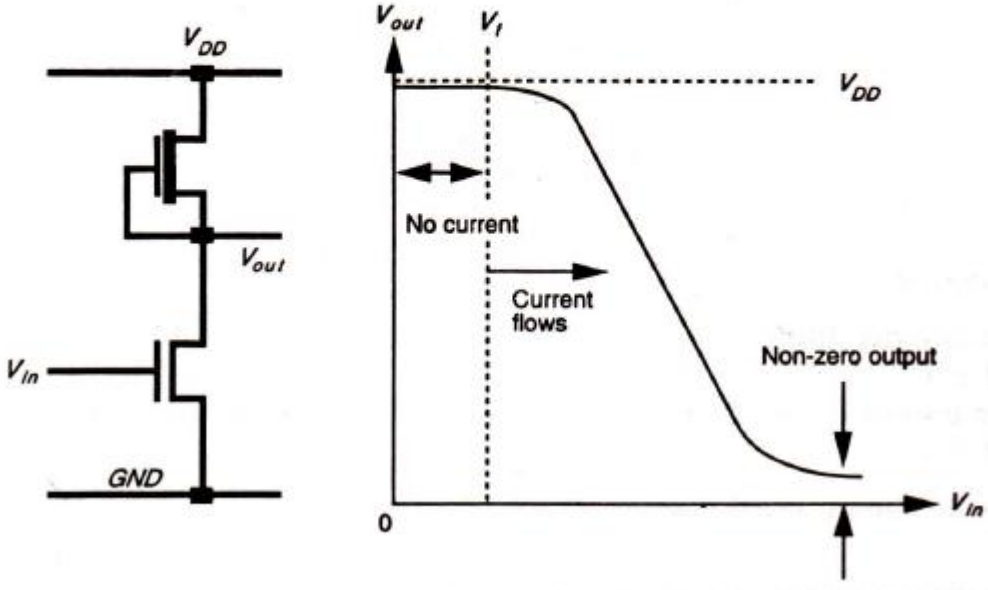


FIGURE 2.12 nMOS depletion mode transistor pull-up and transfer characteristic.

3. *nMOS enhancement mode pull-up* (Figure 2.13).
   (a) Dissipation is high since current flows when V_{in} = logical 1 (V_{GG} is returned to V_{DD}).
   (b) V_{out} can never reach V_{DD} (logical I) if $V_{GG} = V_{DD}$ as is normally the case.

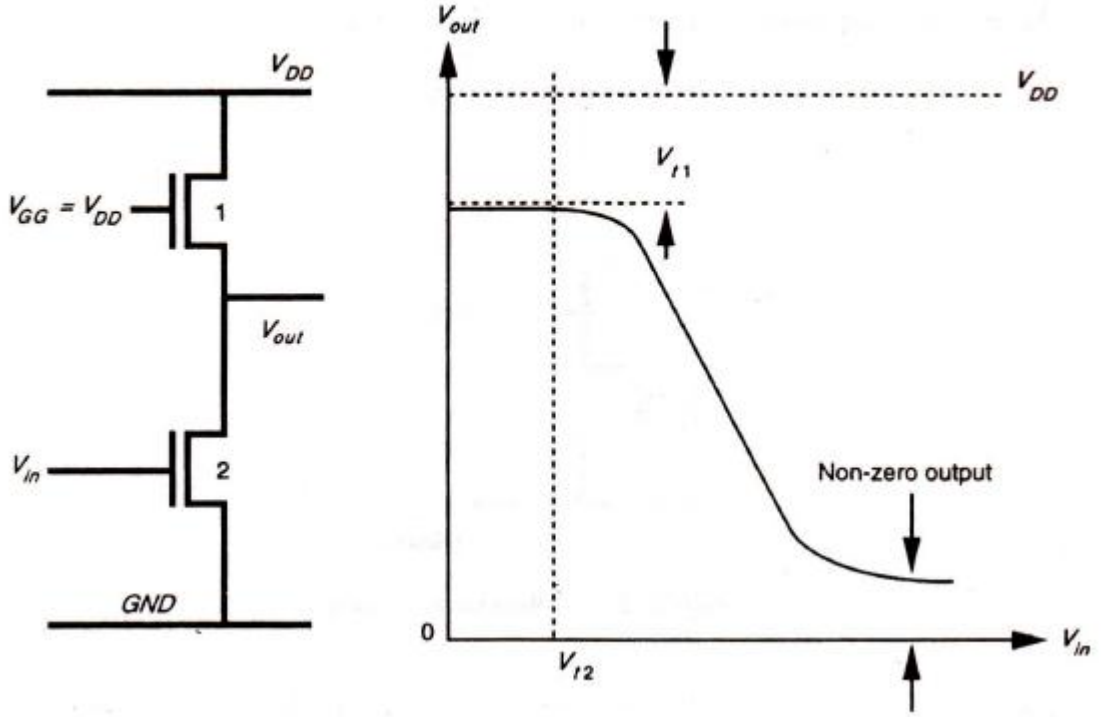
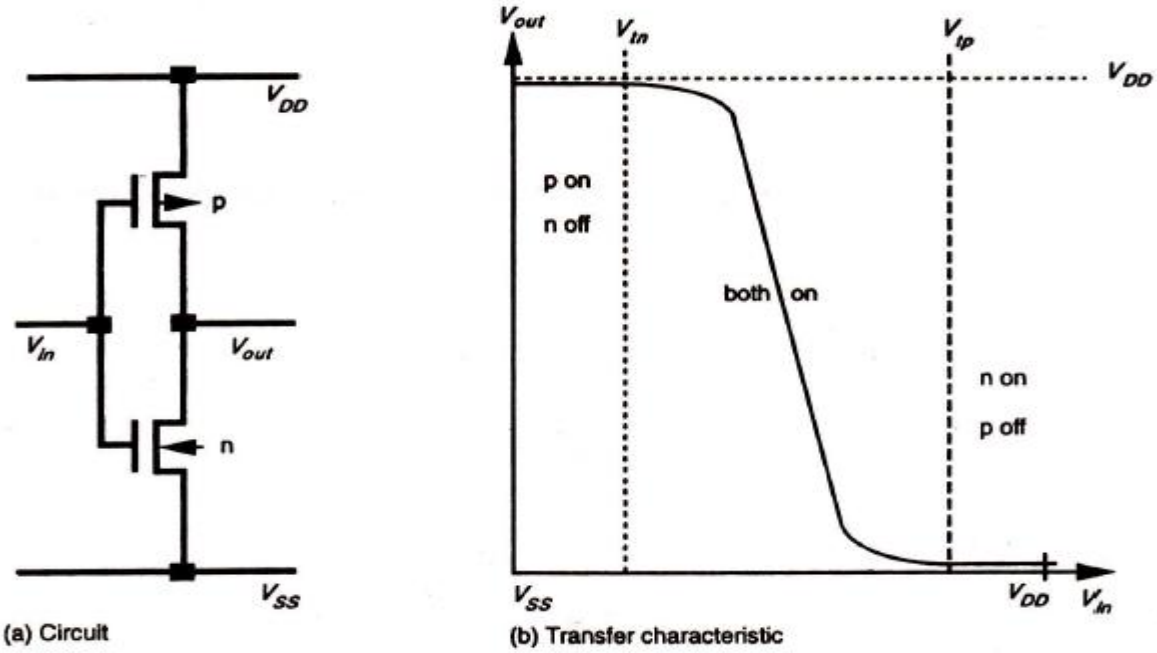
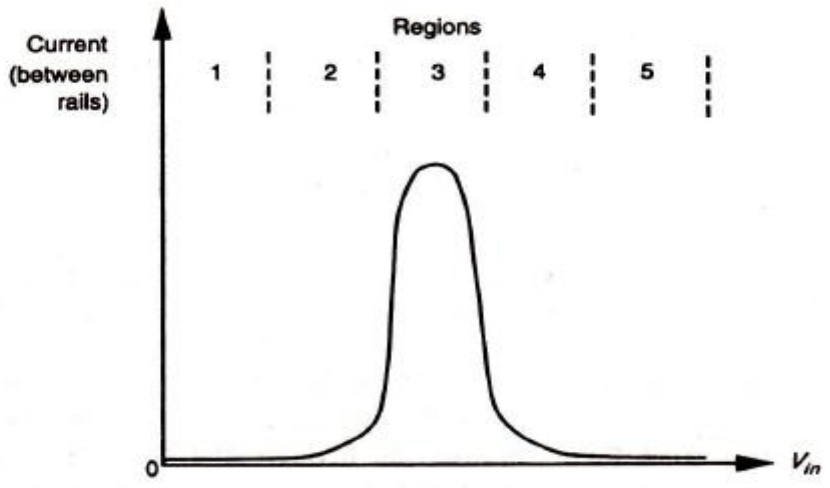


FIGURE 2.13 nMOS enhancement mode pull-up and transfer characteristic.

   (c) V_{GG} may be derived from a switching source, for example, one phase of a clock, so that dissipation can be greatly reduced.
   (d) If V_{GG} is higher than V_{DD} then an extra supply rail is required.

4. *Complementary transistor pull-up* (CMOS) (Figure 2.14).
   (a) No current flow either for logical 0 or for logical 1 inputs.
   (b) Full logical 1 and 0 levels are presented at the output.
   (c) For devices of similar dimensions the p-channel is slower than the n-channel device.





(c) CMOS inverter current versus V_{in}

FIGURE 2.14 Complementary transistor pull-up (CMOS).

THE CMOS INVERTER

The general arrangement and characteristics are illustrated in Figure 2.14. We have seen (equations 2.4 and 2.5) that the current/voltage relationships for the MOS transistor may be written

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

in the resistive region, or

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

in saturation. In both cases the factor K is a technology-dependent parameter such that

$$K = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}$$

The factor W/L is, of course, contributed by the geometry and it is common practice to write

$$\beta = K\frac{W}{L}$$

so that, for example

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

in saturation, and where β may be applied to both nMOS and pMOS transistors as follows:

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

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

where W_n and L_n, W_p and L_p are the n- and p-transistor dimensions respectively. With regard to Figures 2.14(b) and 2.14(c), it may be seen that the CMOS inverter has five distinct regions of operation.

Considering the static conditions first, it may be seen that in *region 1* for which V_{in} = logic 0, we have the p-transistor fully turned on while the n-transistor is fully turned off. Thus no current flows through the inverter and the output is directly connected to V_{DD} through the p-transistor. A good logic 1 output voltage is thus present at the output.

In *region 5* V_{in} = logic 1, the n-transistor is fully on while the p-transistor is fully off. Again, no current flows and a good logic 0 appears at the output.

In *region 2* the input voltage has increased to a level which just exceeds the threshold voltage of the n-transistor. The n-transistor conducts and has a large voltage between source and drain; so it is in saturation. The p-transistor is also conducting but with only a small voltage across it, it operates in the unsaturated resistive region. A small current now flows through the inverter from V_{DD} to V_{SS}. If we wish to analyze the behavior in this region, we equate the p-device resistive region current with the n-device saturation current and thus obtain the voltage and current relationships.

Region 4 is similar to region 2 but with the roles of the p- and n-transistors reversed. However, the current magnitudes in regions 2 and 4 are small and most of the energy consumed in switching from one state to the other is due to the larger current which flows in region 3.

Region 3 is the region in which the inverter exhibits gain and in which both transistors are in saturation.

The currents (with regard to Figure 2.14(c)) in each device must be the same since the transistors are in series, so we may write

$$I_{dsp} = - I_{dsn}$$

where

$$I_{dsp} = \frac{\beta_p}{2} (V_{in} - V_{DD} - V_{tp})^2$$

and

$$I_{dsn} = \frac{\beta_n}{2} (V_{in} - V_{tn})^2$$

from whence we can express V_{in} in terms of the β ratio and the other circuit voltages and currents

$$V_{in} = \frac{V_{DD} + V_{tp} + V_{tn}(\beta_n/\beta_p)^{1/2}}{1 + (\beta_n/\beta_p)^{1/2}} \qquad (2.10)$$

Since both transistors are in saturation, they act as current sources so that the equivalent circuit in this region is two current sources in series between V_{DD} and V_{SS} with the output voltage coming from their common point. The region is inherently unstable in consequence and the changeover from one logic level to the other is rapid.

If $\beta_n = \beta_p$ and if $V_{tn} = -V_{tp}$, then from equation (2.10).

$$V_{in} = 0.5\ V_{DD}$$

This implies that the changeover between logic levels is symmetrically disposed about the point at which

$$V_{in} = V_{out} = 0.5\ V_{DD}$$

since only at this point will the two β factors be equal. But for $\beta_n = \beta_p$ the device geometries must be such that

$$\mu_p W_p/L_p = \mu_n W_n/L_n$$

Now the mobilities are inherently unequal and thus it is necessary for the width to length ratio of the p-device to be two to three times that of the n-device, namely

$$W_p/L_p \doteq 2.5\ W_n/L_n$$

However, it must be recognized that mobility μ is affected by the transverse electric field in the channel and is thus dependent on V_{gs} (and thus on V_{in} in this in case). It has been shown empirically that the actual mobility is

$$\mu = \mu_z\ (1 - \phi\ (V_{gs} - V_t))^{-1}$$

ϕ is a constant approximately equal to 0.05, V_t includes any body effect, and μ_z is the mobility with zero transverse field. Thus a β ratio of 1 will only hold good around the point of symmetry when $V_{out} = V_{in} = 0.5V_{DD}$.

The β ratio is often unimportant in many configurations and in most cases minimum size transistor geometries are used for both n- and p-devices. Figure 2.15 indicates the trends in the transfer characteristic as the ratio is varied. The changes indicated in the figure would be for quite large variations in β ratio (e.g. up to 10:1) and the ratio is thus not too critical in this respect.

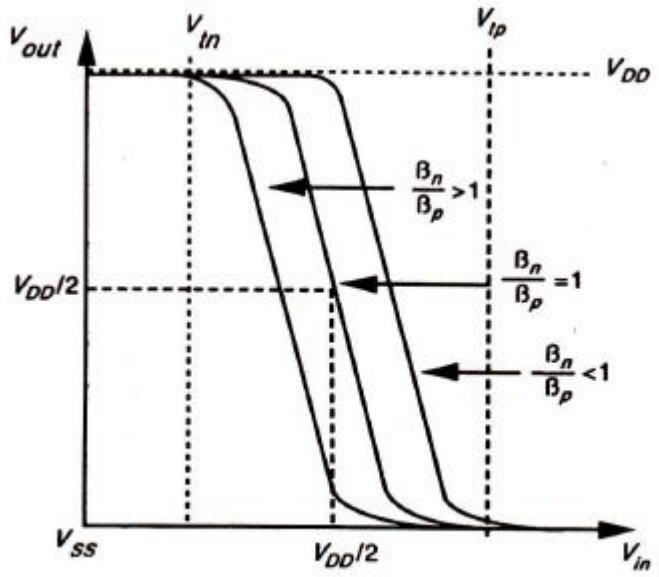


FIGURE 2.15 Trends in transfer characteristic with β ratio.

MOS TRANSISTOR CIRCUIT MODEL

The MOS transistor can be modeled with varying degrees of complexity. However, a consideration of the actual physical construction of the device (as in Figure 2.16) leads to some understanding of the various components of the model.

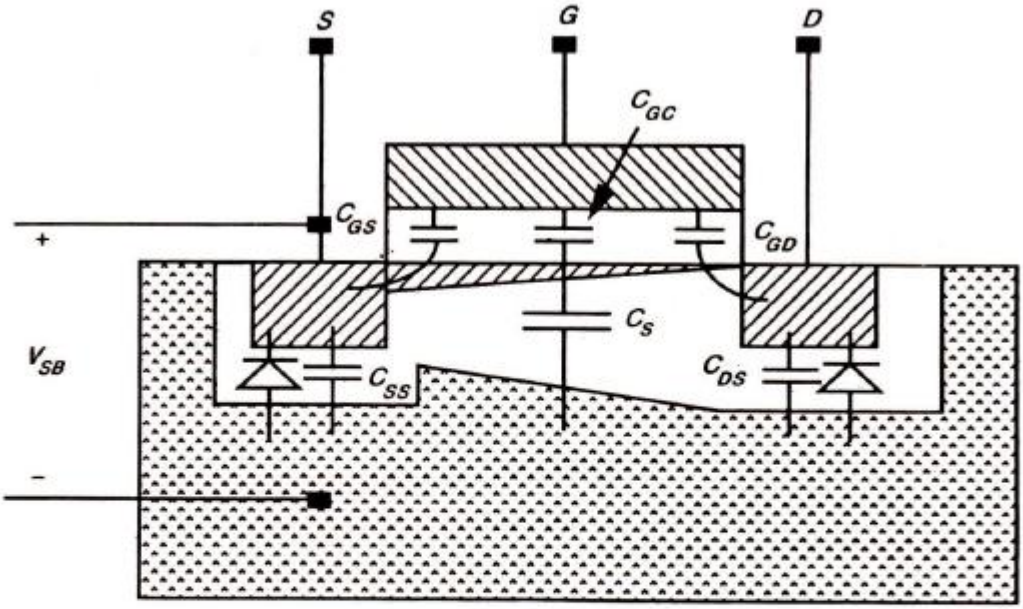


FIGURE 2.16 nMOS transistor model.

Notes: C_{GC} = gate to channel capacitance
C_{GS} = gate to source capacitance
C_{GD} = gate to drain capacitance
} Small for self-aligning nMOS process

Remaining capacitances are associated with the depletion layer and are voltage dependent. Note that C_{SS} indicates source-to-substrate, C_{DS} drain-to-substrate, and C_S channel-to-substrate capacitances.

BiCMOS Inverters

As in nMOS and CMOS logic circuitry, the basic logic element is the inverter circuit.

When designing with BiCMOS in mind, the logical approach is to use MOS switches to perform the logic function and bipolar transistors to drive the output loads. The simplest logic function is that of inversion, and a simple BiCMOS inverter circuit is readily set out as shown in Figure 2.17.

It consists of two bipolar transistors T_1 and T_2 with one nMOS transistor T_3, and one pMOS transistor T_4, both being enhancement mode devices. The action of the circuit is straightforward and may be described as follows:

- With V_{in} = 0 volts (*GND*) T_3 is off so that T_1 will be non-conducting. But T_4 is on and supplies current to the base of T_2 which will conduct and act as a current source to charge the load C_L toward +5 volts(V_{DD}). The output of the inverter will rise to +5 volts less the base to emitter voltage V_{BE} of T_2.
- With V_{in} = +5 volts (V_{DD}) T_4 is off so that T_2 will be non-conducting. But T_3 will now be on and will supply current to the base of T_1 which will conduct and act as a current sink to the load C_L discharging it toward 0 volts (*GND*). The output of the inverter will fall to 0 volts plus the saturation voltage V_{CEsat} from the collector to the emitter of T_1.
- T_1 and T_2 will present low impedances when turned on into saturation and the load C_L will be charged or discharged rapidly.

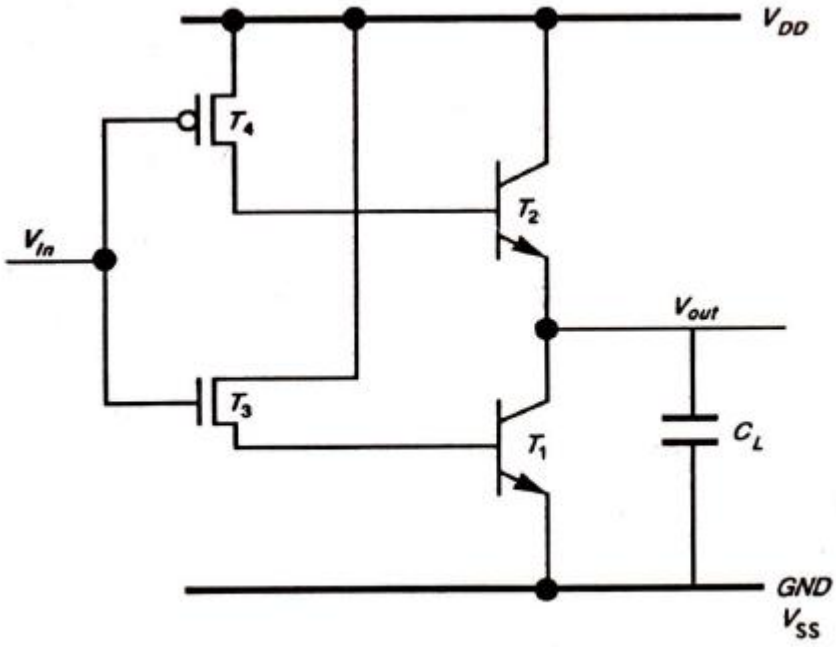


FIGURE 2.17 A simple BiCMOS inverter.

- The output logic levels will be good and will be close to the rail voltages since V_{CEsat} is quite small and V_{BE} is approximately + 0.7 volts.
- The inverter has a high input impedance.
- The inverter has a low output impedance.
- The inverter has a high current drive capability but occupies a relatively small area.
- The inverter has high noise margins.

However, owing to the presence of a DC path from V_{DD} to *GND* through T_3 and T_1, this is not a good arrangement to implement since there will be a significant static current flow whenever V_{in} = logic 1. There is also a problem in that there is no discharge path for current from the base of either bipolar transistor when it is being turned off. This will slow down the action of this circuit.

An improved version of this circuit is given in Figure 2.18, in which the DC path through T_3 and T_1 is eliminated, but the output voltage swing is now reduced, since the output cannot fall below the base to emitter voltage V_{BE} of T_1.

An improved inverter arrangement, using resistors, is shown in Figure 2.19. In this circuit resistors provide the improved swing of output voltage when each bipolar transistor is off, and also provide discharge paths for base current during turn-off.

The provision of on chip resistors of suitable value is not always convenient and may be space-consuming, so that other arrangements—such as in Figure 2.20—are used. In this circuit, the transistors T_5 and T_6 are arranged to turn on when T_2 and T_1 respectively are being turned off.

In general, BiCMOS inverters offer many advantages where high load current sinking and sourcing is required. The arrangements lead on to the BiCMOS gate circuits

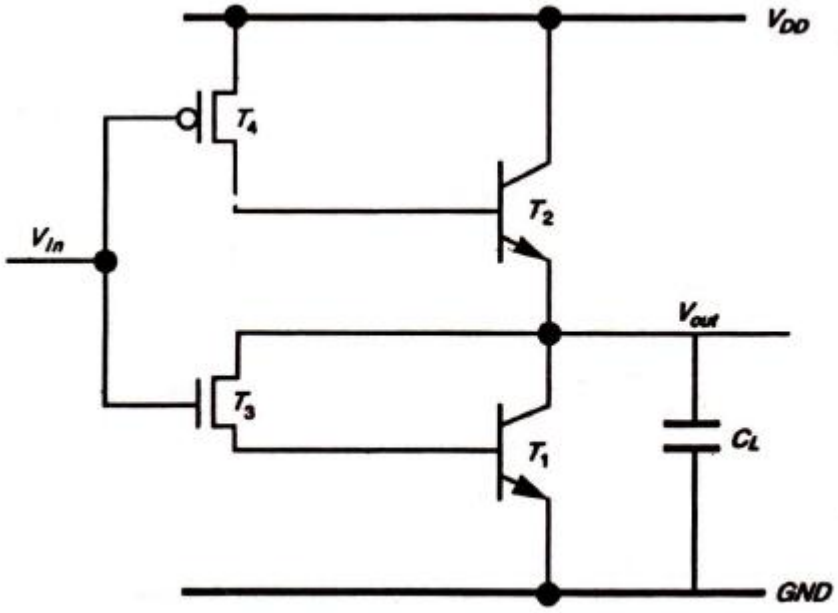


FIGURE 2.18 An alternative BiCMOS inverter with no static current flow.

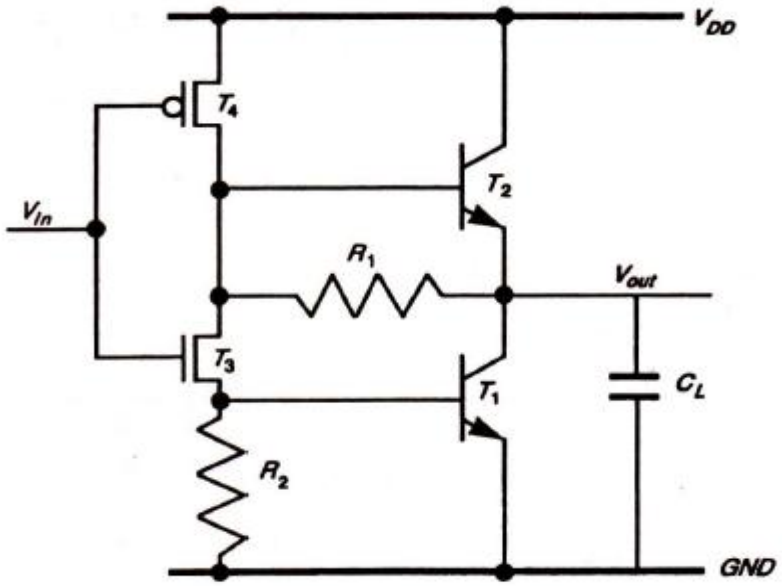


FIGURE 2.19 An improved BiCMOS inverter with better output logic levels.

LATCH-UP IN CMOS CIRCUITS

A problem which is inherent in the p-well and n-well processes is due to the relatively large number of junctions which are formed in these structures and, as mentioned earlier, the consequent presence of parasitic transistors and diodes. Latch-up is a condition in which the parasitic components give rise to the establishment of low-resistance conducting paths between V_{DD} and V_{SS} with disastrous results. Careful control during fabrication is necessary to avoid this problem.

Latch-up may be induced by glitches on the supply rails or by incident radiation. The mechanism involved may be understood by referring to Figure 2.21, which shows the key parasitic components associated with a p-well structure in which an inverter circuit (for example) has been formed.

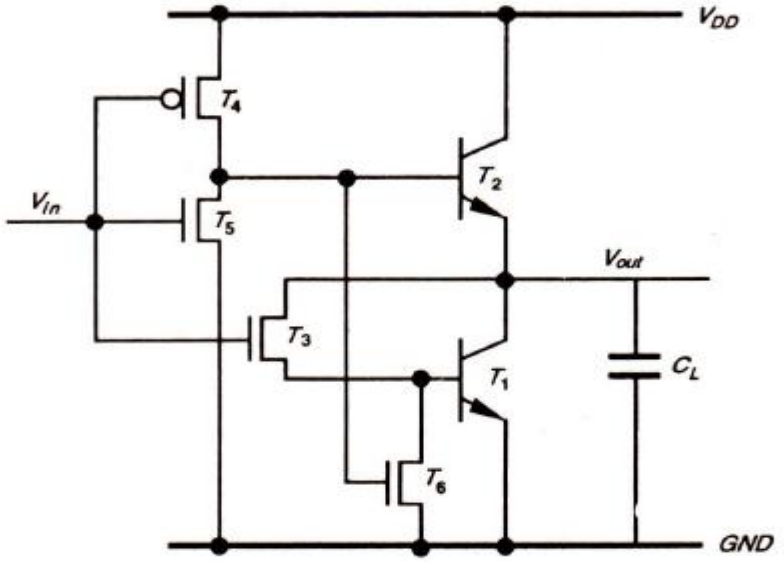


FIGURE 2.20 An improved BiCMOS inverter using MOS transistors for base current dischange.

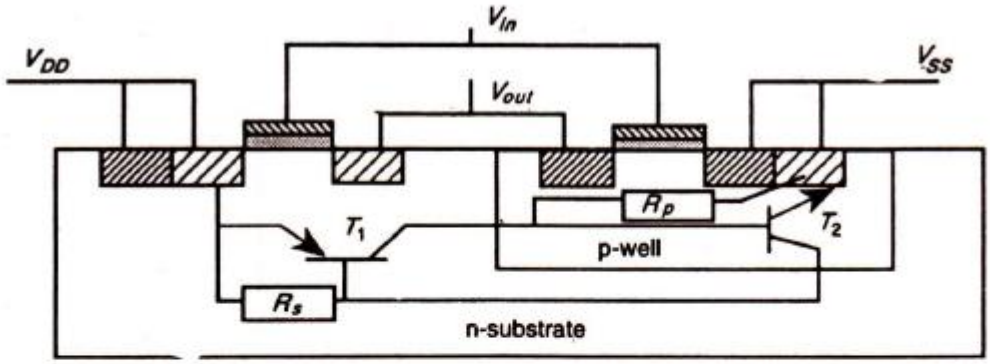


FIGURE 2.21 Latch-up effect in p-well structure.

There are, in effect, two transistors and two resistances (associated with the p-well and with regions of the substrate) which form a path between V_{DD} and V_{SS}. If sufficient substrate current flows to generate enough voltage across R_s to turn on transistor T_1, this will then draw current through R_p and, if the voltage developed is sufficient, T_2 will also turn on, establishing a self-sustaining low-resistance path between the supply rails. If the current gains of the two transistors are such that $\beta_1 \times \beta_2 > 1$, latch-up may occur. Equivalent circuits are given in Figure 2.22.

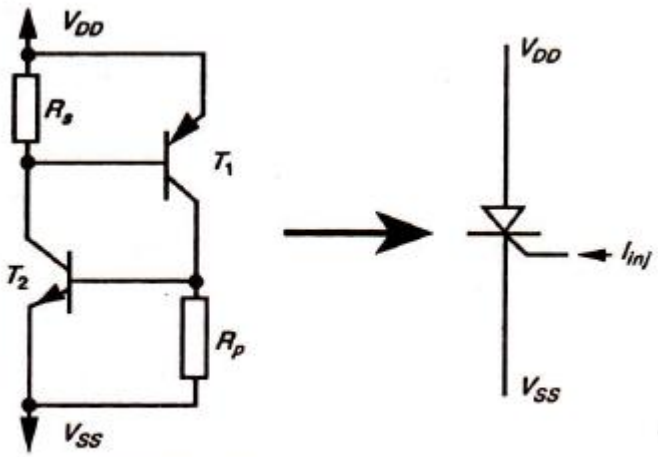


FIGURE 2.22 Latch-up circuit model.

With no injected current, the parasitic transistors will exhibit high resistance, but sufficient substrate current flow will cause switching to the low-resistance state as already explained. The switching characteristic of the arrangement is outlined in Figure 2.23.

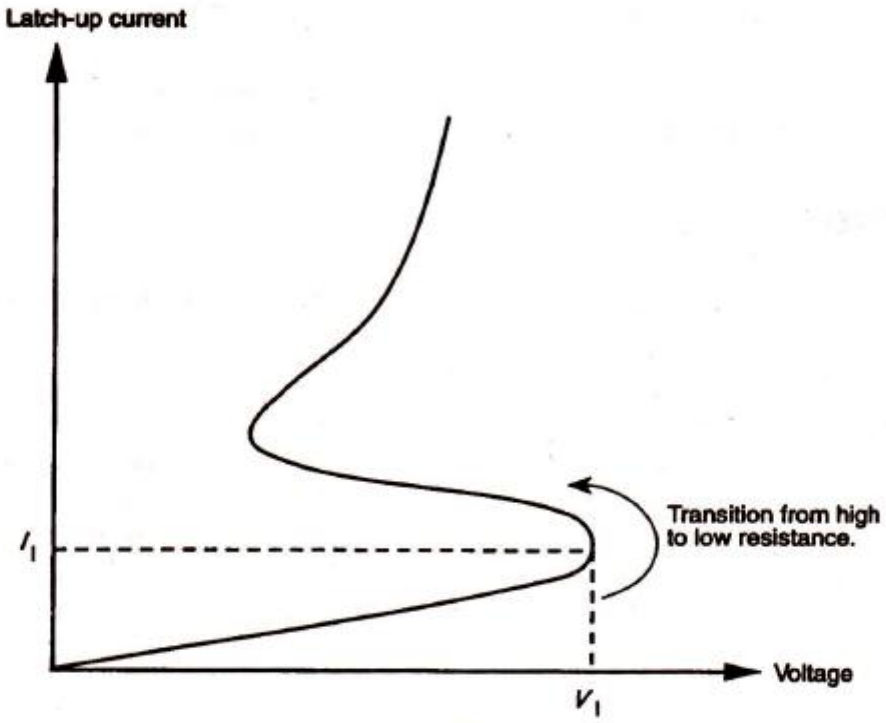


FIGURE 2.23 Latch-up current versus voltage.

Once latched-up, this condition will be maintained until the latch-up current drops below I_l. It is thus essential for a CMOS process to ensure that V_l and I_l are not readily achieved in any normal mode of operation.

Remedies for the latch-up problem include:

1. an increase in substrate doping levels with a consequent drop in the value of R_s;
2. reducing R_p by control of fabrication parameters and by ensuring a low contact resistance to V_{SS};
3. other more elaborate measures such as the introduction of guard rings.

For completeness, the latch-up configuration for an n-well structure is given in Figure 2.24.

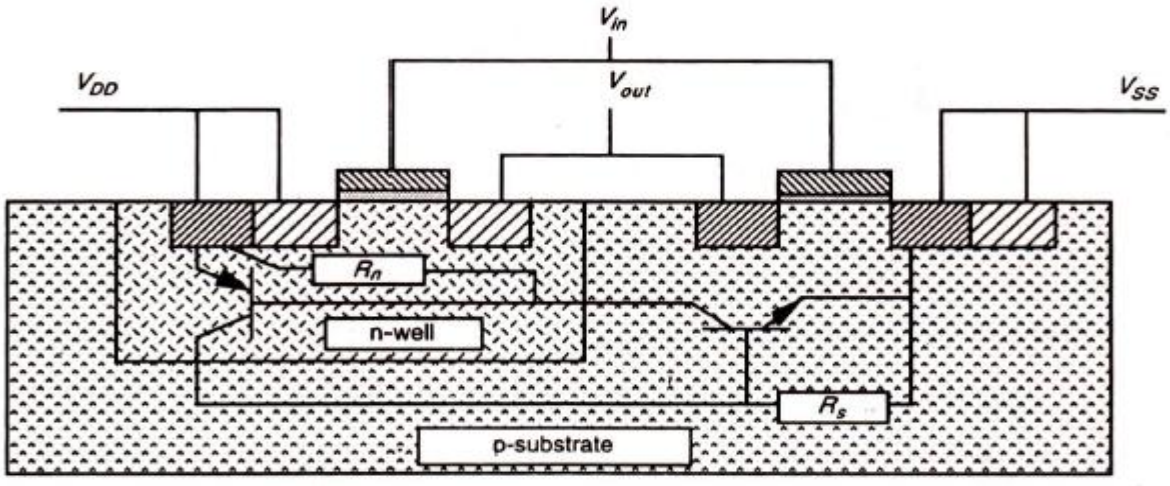


FIGURE 2.24 Latch-up circuit for n-well process.

BiCMOS LATCH-UP SUSCEPTIBILITY

One benefit of the BiCMOS process is that it produces circuits which are less likely to suffer from latch-up problems. This is due to several factors:

- A reduction of substrate resistance R_s.
- A reduction of n-well resistance R_w.
- A reduction of R_s and R_w means that a larger lateral current is necessary to invite latch-up and a higher value of holding current is also required.
- The parasitic (vertical) pnp transistor which is part of the n-well latch-up circuit has its beta reduced owing to the presence of the buried n+ layer. This has the effect of reducing carrier lifetime in the n-base region and this contributes the reduction in beta.