Unit II

MOS Inverters

4.1 Introduction

In Chap. 3, we have seen that a metal–oxide–semiconductor (MOS) transistor can be considered as a voltage-controlled resistor. This basic property can be used to realize digital circuits using MOS transistors. In this chapter, we discuss the realization of various types of MOS inverters. The inverter forms the basic building block of gate-based digital circuits. An inverter can be realized with the source of an n-type metal–oxide–semiconductor (nMOS) enhancement transistor connected to the ground, and the drain connected to the positive supply rail V_{dd} through a pull-up device. The generalized block diagram is shown in Fig. 4.1. The input voltage is ap-plied to the gate of the nMOS transistor with respect to ground and output is taken from the drain. When the MOS transistor is ON, it pulls down the output voltage to the low level, and that is why it is called a *pull-down* device, and the other device, which is connected to V_{dd}, is called the *pull-up* device.

The pull-up device can be realized in several ways. The characteristics of the inverter strongly depend on the pull-up device used to realize the inverter. Theoreti-cally, a passive resistor of suitable value can be used. Although the use of a possible resistor may be possible in realizing an inverter using discrete components, this is not feasible in very-large-scale integration (VLSI) implementation. Instead, an ac-tive pull-up device realized using a depletion-mode nMOS transistor or an enhance-ment-mode nMOS transistor or a p-type metal–oxide–semiconductor (pMOS) transistor could be used. Basic characteristics of MOS inverters are highlighted in Sect. 4.2. The advantages and disadvantages of different inverter configurations are explored in Sect. 4.3 Section 4.3 explores the inverter ratio in different situations. The switching characteristics on MOS inverters are considered in Sect. 4.5 Various delay parameters have been estimated in Sect. 4.6 Section 4.7 presents the different circuit configurations to drive a large capacitive load.

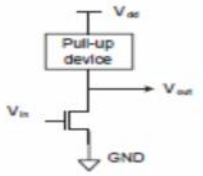


Fig. 4.1 General structure of an nMOS inverter. nMOS n-type metal–oxide–semiconductor

Truth table

V_{in}	V_{out}
0	1
1	0

Fig. 4.2 Truth table and logic symbol of the inverter

4.2 Inverter and Its Characteristics

Before we discuss about the practical inverters realized with MOS transistors, we consider the characteristics of an ideal inverter [1, 2]. The truth table and logic sym-bol of an inverter are shown in Fig. 4.2. The input to the inverter is V_{in} and output is V_{out}.

Figure 4.3 shows how the output of an ideal inverter changes as the input of the inverter is varied from 0 V (logic level 0) to V_{dd} (logic level 1). Initially, output is V_{dd}

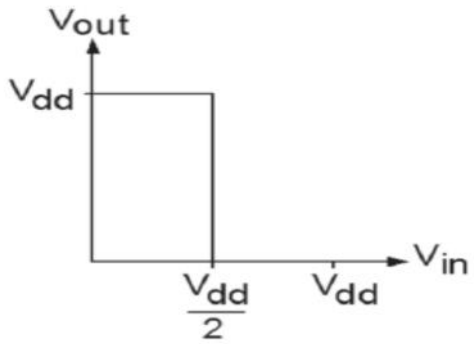


Fig. 4.3 Ideal transfer characteristics of an inverter

when the output is 0 V, and as the input crosses $V_{dd}/2$, the output switches to 0 V, and it remains at this level till the maximum input voltage V_{dd}. This diagram is known as the input–output or *transfer characteristic* of the inverter. The input voltage, $V_{dd}/2$, at which the output changes from high '1' to low '0', is known as *inverter threshold voltage*. For practical inverters realized with MOS devices, the voltage transfer characteristics will be far from this ideal voltage transfer characteristic represented by Fig. 4.3. A more realistic voltage transfer characteristic is shown in Fig. 4.4a. As shown in Fig. 4.4a, because of some voltage drop across the pull-up device, the out-put high voltage level is less than V_{dd} for the low input voltage level. This voltage is represented by V_{OH}, which is the maximum output voltage level for output level '1'.

As the input voltage increases and crosses the threshold voltage of the pull-down transistor, it starts conducting, which leads to a decrease in the output voltage level. However, instead of an abrupt change in the voltage level from logic level '1' to logic level '0', the voltage decreases rather slowly. The unity gain point at which $dV_0 / dV_{in} = -1$ is defined as the *input high voltage* V_{IL}, which is the maximum in-put voltage which can be treated as logic level '0'.

As the input voltage is increased further, the output crosses a point where $V_{in} = V_{out}$.

The voltage at which this occurs is referred to as the *inverter threshold voltage* V_T.

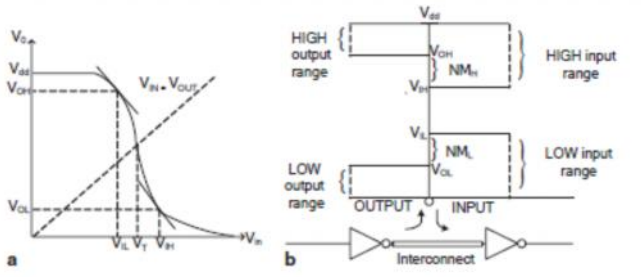


Fig. 4.4 **a** Various voltage levels on the transfer characteristics; **b** low- and high-level noise margins

It may be noted that the inverter threshold voltage may not be equal to $V_{dd}/2$ for practical inverters. Before the output attains the *output low voltage* V_{OL}, which is the minimum output voltage for output logic level '0', the transfer-characteristic curve crosses another important point V_{IH}, the minimum input voltage that can be accepted as logic '1'. This point is also obtained at another unity gain point at which

$dV_0 / dV_{in} = -1$ as shown in Fig. 4.4a.

An important parameter called the *noise margin* is associated with the input–out-put voltage characteristics of a gate. It is defined as the allowable noise voltage on the input of a gate so that the output is not affected. The deviations in logic levels from the ideal values, which are restored as the signal propagates to the output, can be obtained from the DC characteristic curves. The logic levels at the input and output are given by

logic 0 input: $0 \le V_{in} \le V_{IL}$,

logic 1 input: $V_{IH} \le V_{in} \le V_{dd}$,

logic 0 output: $0 \le V_0 \le V_{OL}$,

logic 1 output: $V_{OH} \le V_0 \le V_{dd}$.

The *low-level noise margin* is defined as the difference in magnitude between the minimum low output voltage of the driving gate and the maximum input low volt-age accepted by the driven gate.

$$NM_L = |V_{IL} - V_{OL}| \qquad (4.4)$$

The *high-level noise margin* is defined as the difference in magnitude between the minimum high output voltage of the driving gate and the minimum voltage accept-able as high level by the driven gate:

$$NM_H = |V_{OH} - V_{IH}|. \qquad (4.4)$$

To find out the noise margin, we can use the transfer characteristics as shown in Fig. 4.4a. The noise margins are shown in Fig. 4.4b.

When any of the noise margins is low, the gate is susceptible to a switching noise at the input.

4.3 MOS Inverter Configurations

The various MOS inverter configurations [3] realized using different types of pull-up devices are discussed in this section. In Sect. 4.3.1, the use of a passive re-sistor as the pull-up device is discussed and disadvantages are highlighted. The use of a depletion-mode nMOS transistor as the pull-up device is discussed in Sect. 4.3.2. Section 4.3.3 discusses the use of an enhancement mode of nMOS transistor, whereas Sect. 4.3.4 discusses the use of a pMOS transistor as a pull-up

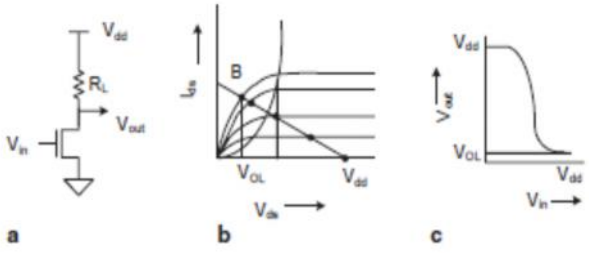


Fig. 4.5 **a** An nMOS inverter with resistive load; **b** voltage–current characteristic; **c** transfer characteristic. nMOS n-type–metal–oxide semiconductor

device in configuration. The pMOS device can also be used to realize the CMOS inverter, where the two transistors are used in complementary mode, as discussed in Sect. 4.3.5. Various inverters introduced in this section are compared in Sect. 4.3.6.

4.3.1 Passive Resistive as Pull-up Device

A passive resistor R_L can be used as the pull-up device as shown in Fig. 4.5a. The value of the resistor should be chosen such that the circuit functionally behaves like an inverter. When the input voltage V_{in} is less than V_{tn}, the transistor is OFF and the output capacitor charges to V_{dd}. Therefore, we get V_{dd} as the output for any input voltage less than V_{tn}. When V_{in} is greater than V_{tn}, the MOS transistor acts as a resistor R_c, where R_c is the channel resistance with $V_{gs} > V_{tn}$. The output capacitor discharges through this resistor and output voltage is given by

$$V_{OL} = V_{dd} \frac{R_c}{R_c + R_L} \tag{4.3}$$

Normally, this output is used to drive other gates. Functionally, this voltage can be accepted as low level provided it is less than V_t. So,

$$V_{OL} = V_{dd} \frac{R_c}{R_c + R_L} < V_{tn}.$$

Assuming the typical value of threshold voltage $V_{tn} = 0.2V_{dd}$, we get

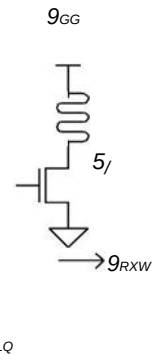
$$R_c \tag{4.4}$$

$$V_{OL} = V_{dd} \frac{}{R + R} \leq 0.2V_{dd} \quad \text{or} \quad R_L > 4R_C$$

cL

This imposes a restriction on the minimum value of load resistance for a successful operation of the circuit as an inverter. The input–output characteristic of the inverter

Fig. 4.6 Realization of a resistive load



is shown in Fig. 4.5b. The circuit operates along the load line as shown in Fig. 4.5b.

For V_{in} = 0 V, the output voltage $V_{out} = V_{dd}$ (point A), and for $V_{in} = V_{dd}$, the output volt-age $V_{out} = V_{OL}$, as shown by point B. The transfer characteristic is shown in Fig. 4.5c, which shows that the output is V_{dd} for V_{in} = OV, but for $V_{in} = V_{dd}$ the output is not OV.

This implementation of this inverter has a number of disadvantages:

•\ As the charging of the output capacitor takes place through the load resistor R_L and discharge through R_c and their values must be different as per Eq. 4.4, there is asymmetry in the ON-to-OFF and OFF-to-ON switching times.

•\ To have higher speeds of operation, the value of both R_c and R_L should be re-duced. However, this increases the power dissipation of the circuit. Moreover, as we shall see later, to achieve a smaller value of R_c, the area of the MOS inverter needs to be increased.

•\ The resistive load can be fabricated by two approaches—using a diffused resis-tor approach or using an undoped poly-silicon approach. In the first case, an n-type or a p-type isolated diffusion region can be fabricated to realize a resistor between the power supply line and the drain of the nMOS transistor. To realize a resistor of the order of few *K*

Ω, as required for proper operation of the circuit, the length to width must be large. To realize this large length-to-width ratio in a small area, a serpentine form is used as shown in Fig. 4.6. However, this re-quires a very large chip area. To overcome the limitation of this approach, the second approach based on undoped poly-silicon can be used. Instead of using doped poly-silicon, which is commonly used to realize the gate and interconnect regions having lower resistivity, undoped poly-silicon is used here to get higher resistivity. Although this approach leads to a very compact resistor compared to the previous approach, the resistance value cannot be accurately controlled leading to large process parameter variations. In view of the above discussion, it is evident that this inverter configuration is not suitable for VLSI realization. Better alternatives for the realization of the pull-up resistor are explored in the following subsections.

4.3.2 nMOS Depletion-Mode Transistor as Pull up

To overcome the limitations mentioned above, MOS transistors can be used as pull-up devices instead of using a passive resistor. There are three possible alternatives for pull-up devices—an nMOS enhancement-mode transistor, a depletion-mode

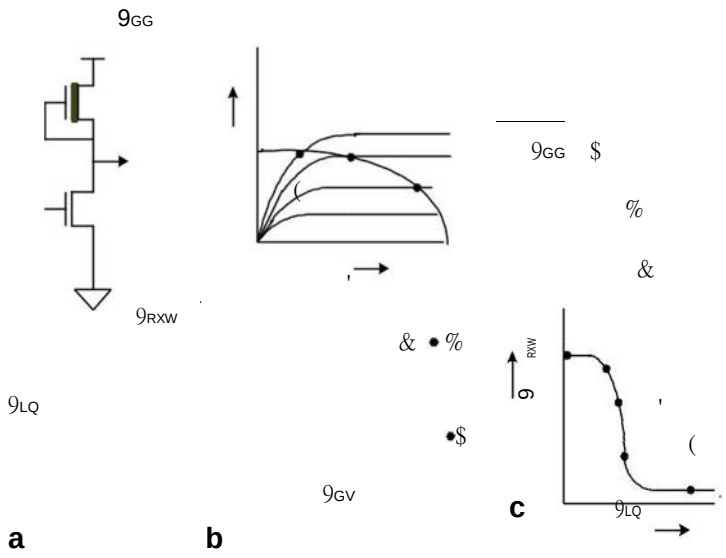


Fig. 4.7 **a** nMOS inverter with depletion-mode transistor as pull-up device; **b** voltage current characteristic; **c** transfer characteristic. nMOS n-type metal–oxide–semiconductor

nMOS transistor, or a pMOS transistor. Any one of the transistors can be used as a pull-up device. First, we consider the use of an nMOS depletion-mode transistor as an active pull-up (pu) device as shown in Fig. 4.7a. As the output of an inverter is commonly connected to the gate of one or more MOS transistors in the next stage, there is no fan-out current, and the currents flowing through both the transistors must be equal. The input voltage is applied to the gate of the pull-down (pd) transis-tor, and the output is taken out from the drain of the pd device.

1.\ *Pull-down device off and pull-up device in linear region:* This corresponds to

point ‘A’ on the curve with the input voltage $V_{in} < V_{tn}$,$V_{out} = V_{dd}$ and $I_{ds} = 0$. In this situation, there is no current flow from the power supply and no current

flows through either of the transistors.

2.\ *Pull-down device in saturation and pull-up device in linear region:* This corre-sponds to point B. Here,

$$I_{pd} = \frac{K_n W_{pd}}{2L_{pd}} (V_{in} - V_{tpd})^2 \qquad (4.5)$$

and

$$I_{pu} = \frac{K_n W_{pu}}{L_{pu}} \left[(V_{out} - V_{tpu}) V_{out} - \frac{V_{out}^2}{2} \right], \qquad (4.6)$$

where V_{tpd} and V_{tpu} are the threshold voltages of the enhancement- and depletion-mode MOS transistors, respectively.

3.\ *Pull-down and pull-up device, both in saturation:* This is represented by point C on the curve. In this situation,

$$I_{pd} = \frac{K_n W_{pd}}{2L_{pd}} (V_{in} - V_{tpd})^2 \qquad (4.7)$$

and

$$I_{pu} = \frac{K_n W_{pu}}{2L_{pu}} V_{tpu}^2 .$$

4.\ *Pull-down device in linear region and pull-up device in saturation*: This situa-tion occurs when input voltage is equal to V_{dd}. Here,

$$I_{pd} = \beta_{pd} \left(V_{in} - V_{tpd} - \frac{V_{OL}}{2} \right) V_{OL}$$

$$I_{pu} = \frac{\beta_{pu}}{2} (V_{tpu})^2, \qquad (4.8)$$

where $\beta_{pd} = K_n \frac{W_{pd}}{L_{pd}}$ and $\beta_{pu} = K_n \frac{W_{pd}}{L_{pd}}$.

Equating the two currents and ignoring V_{OL} / 2 term, we get

$$\beta_{pd} (V_{dd} - V_{tpd}) V_{OL} = \frac{\beta_{pu}}{2} (V_{tpu})^2,$$

$$V_{OL} = \frac{1}{2K V_{dd} - V_{tpd}} (V_{tpu})^2 = \frac{\beta_{pu}}{2\beta_{pd}} \frac{(V_{tpu})^2}{V_{dd} - V_{tpd}} = \frac{1}{2K} \frac{(V_{tpu})^2}{V_{dd} - V_{tpd}},$$

where

$$\frac{\beta_{pd}}{} \quad \frac{(W/L)_{pd}}{}$$

(4.11)

(4.9)

(4.10)

The quantity K is called the *ratio of the inverter*. For successful inverter operation, the low output voltage, V_{OL}, should be smaller than the threshold voltage of the pull-down transistor of the next stage. From the above discussion, we can make the following conclusion:

•\ The output is not ratioless, which leads to asymmetry in switching characteris-tics.

•\ There is static power dissipation when the output logic level is low.

•\ It produces strong high output level, but weak low output level.

4.3.3 nMOS Enhancement-Mode Transistor as Pull up

Alternatively, an enhancement-mode nMOS transistor with gate normally connect-ed to its drain (V_{dd}) can be used as an active pull-up resistor as shown in Fig. 4.8a. Let us consider the output voltage for two situations—when V_{in} = O and $V_{in} = V_{dd}$. In

the first case, the desired output is V_{dd}. But as the output, $V_{out,}$ approaches the voltage ($V_{dd} - V_{tn}$), the pull-up transistor turns off. Therefore, the output voltage cannot

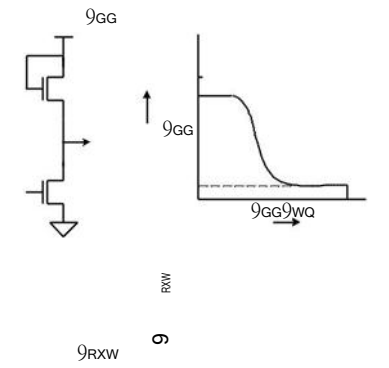


Fig. 4.8 **a** nMOS inverter with enhance-mode transistor as a pull-up device; **b** transfer characteristic. nMOS n-type metal–oxide–semiconductor

Fig. 4.9 **a** A pseudo-nMOS inverter; **b** transfer characteristic. Pseudo-nMOS pseudo-n-type metal–oxide–semiconductor

reach V_{dd}. The maximum output voltage that can be attained is ($V_{dd} - V_{tn}$), where V_{tn} is the threshold voltage of the enhancement-mode pull-up

transistor. The output voltage for $V_{in} = V_{dd}$ is not 0 V, because in this case both the transistors are conduct-ing and act as a voltage divider. The transfer characteristic is shown in Fig. 4.8b. From the above discussion, we can make the following conclusion:

•\ The output is not ratioless, which leads to asymmetry in switching characteris-tics.

•\ There is static power dissipation when the output level is low.

•\ It produces weak low and high output levels.

As a consequence, nMOS enhancement-type transistor is not suitable as a pull-up device for realizing an MOS inverter.

4.3.4 The pMOS Transistor as Pull Up

We can realize another type of inverter with a pMOS transistor as a pull-up device with its gate permanently connected to the ground as shown in Fig. 4.9a. As it is functionally similar to a depletion-type nMOS load, it is called a 'pseudo-nMOS' inverter. Unlike the CMOS inverter, discussed in Sect. 4.2.4, the pull-up transis-tor always remains ON, and there is DC current flow when the pull-down device is ON. The low-level output is also not zero and is dependent on the β_n / β_p ratio like the depletion-type nMOS load. The voltage-transfer characteristic is shown in Fig. 4.9b.

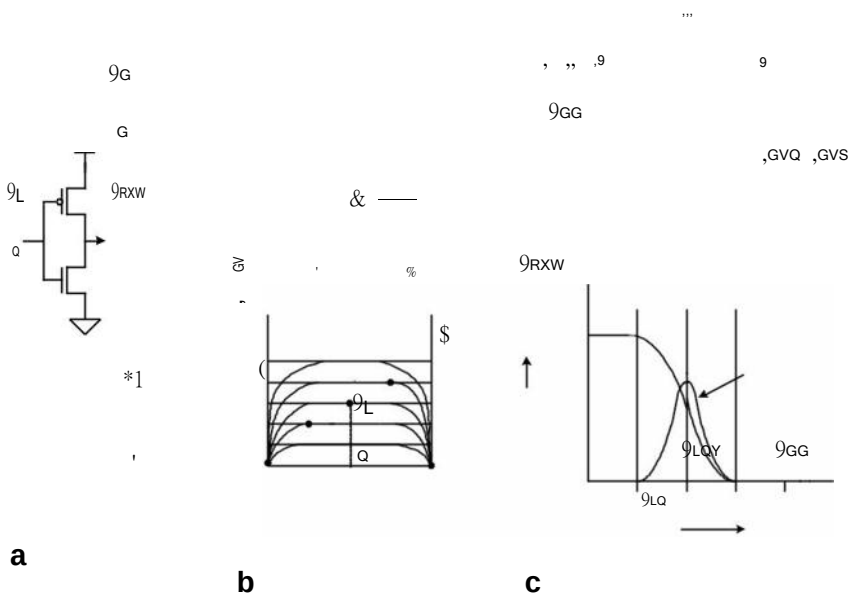


Fig. 4.10 **a** CMOS inverter; **b** voltage–current characteristic; and **c** transfer characteristic

4.3.5 pMOS Transistor as a Pull Up in Complementary Mode

In this case, a pMOS enhancement type transistor is used as a pull-up device. How-ever, here the gates of both the pull-up and pull-down transistors are tied together and used as input as shown in Fig. 4.10a. Output is taken from the drain of the pull-down device as usual. In this case, when the input voltage V_{in} = 0 V, the gate input of the pull-up transistor is below V_{dd} of its source voltage, i.e., $V_{gs} = -V_{dd}$, which makes the pull-up transistor ON, and the pull-down transistor OFF. So, there is no DC current flow between V_{dd} to ground. When the input voltage $V_{in} = V_{dd}$, the gate input of the pull-up transistor is zero with respect to its source, which makes it OFF. The pull-down transistor, however, is ON because the $V_{gs}^{pd} = V_{dd}$. In this situation also, there is no DC current flow between V_{dd} and ground. However, as the gate voltage is gradually increased from '0' to '1', the pull-up transistor switches from ON to OFF and the pull-down transistor switches from OFF to ON. Around the midpoint, both transistors are ON and DC current flows between V_{dd} and ground. Detailed analysis can be made by dividing the entire region of operation into five basic regions as follows:

Region 1: $0 \leq V_{in} < V_{tn}$ The pull-down transistor is off and the pull-up transistor is in the linear (subthreshold) region as shown by a representative point 'A' on the super-imposed nMOS and pMOS transistor's drain-to-source voltage–current character-istic curves. In this region, there is no DC current flow and output voltage remains close to V_{dd}. This corresponds to point 'A' in Fig. 4.10b.

Region 2: $V_{tn} < V_{in} < V_{inv}$ Here, the pull-down transistor moves into a saturation region and the pull-up transistor remains in the linear region as represented by point B, when the input is V_{IL}. The pull-down transistor acts as a current source and pull-up transistor acts as a resistor. The current flow through the transistor increases as

V_{in} increases from V_{tn} to V_{inv} and attains a maximum value when $V_{in} = V_{inv}$.
Since the same current flows through both the devices, $I_{dsp} = -I_{dsn}$

For the pMOS devices, the drain current is given by

$$I_{dsp} = -\beta_p \left[(V_{in} - V_{dd} - V_{tp})(V_O - V_{dd}) - \frac{1}{2}(V_O - V_{dd})^2\right], \quad (4.12)$$

where

$$\beta_p = K_p \frac{W_p}{L_p}, \quad V_{gsp} = V_{in} - V_{dd} \quad \text{and} \quad V_{dsp} = V_{Odd} - V_{dd}.$$

The saturation current of the nMOS transistor is given by

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

where $\beta_n = K_n (W/L)_n$ and $V_{gsn} = V_{in}$.

Equating these two, we get

$$V_O = (V_{in} - V_{tp}) + (V_{in} - V_{tp}) - \sqrt{(V_{in} - V_{tp})^2 - \frac{\beta_n}{\beta_p}(V_{in} - V_{tn})^2} \quad (4.13)$$

A plot of this is shown in Region II of Fig. 4.10c.

The V_{IL} corresponds to the point $dV_{out} / dV_{in} = -1$, and, at this point, the nMOS transistor operates in the saturation region, whereas the pMOS transistor operates in the linear region. Equating $I_{dsn} = -I_{dsp}$,

we get

$$\frac{\beta_n}{2} \quad \frac{\beta_p}{2}$$

$$\frac{\beta_n}{2}(V_{gsn} - V_{tn})^2 = \frac{\beta_p}{2}(V_{gsp} - V_{tp})V_{dsp} - V_{dsp}.$$

Substituting

$$V_{gsn} = V_{in},\ V_{gsp} = -(V_{dd} - V_{in}) \quad \text{and} \quad V_{dsp} = -V_{out} - (V_{dd}),$$

we get

$$\frac{\beta_n}{2}(V_{in} - V_{tn})^2 = \frac{\beta_p}{2}\left[2(V_{in} - V_{dd} - V_{tp})(V_{out} - V_{dd}) - (V_{out} - V_{dd})^2\right]. \quad (4.14)$$

Differentiating both sides with respect to V_{in}, we get

$$\frac{\beta_n}{2}(V_{in} - V_{tn}) = \beta_p\left[(V_{in} - V_{dd} - V_{tp})\left(\frac{dV_{out}}{dV_{in}}\right) + (V_{out} - V_{dd}) - (V_{out} - V_{dd})\left(\frac{dV_{out}}{dV_{in}}\right)\right].$$

Substituting, $V_{in} = V_{IL}$ and $dV_{out} / dV_{in} = -1$, we get

$$\beta_n (V_{IL} - V_{tn}) = \beta_p (2V_{out} - V_{IL} + V_{tp} - V_{dd}) + (2V_{out} + V_{tp} - V_{dd} + (\beta_n / \beta_p) V_{tn}) \quad (4.15)$$

or $$V_{IL} = \frac{}{(1 + (\beta_n / \beta_p))}$$

For $\beta_n / \beta_p = 1$, and $V_{out} \cong V_{dd}$ V_{IL} can be found out to be

$$V_{IL} = \frac{1}{8}(3V_{dd} + 2V_{tn}) \quad (4.16)$$

Region 3: $V_{in} = V_{inv}$ At this point, both the transistors are in the saturation condition as represented by the point C on the superimposed characteristic curves. In this regions, both the transistors can be modeled as current sources.

Assuming

$$V_{gs}^{pd} = V_{in} \quad \text{and}$$

$$V_{gs}^{pu} = V_{in} - V_{dd} = V_{inv} - V_{dd},$$

we may equate the saturation currents of the pull-up and pull-down transistors to get

Equating we get,

$$I_{dsn} = \frac{1}{2} K_n \frac{W_n}{L_n} (V_{inv} - V_{tn})^2$$

$$I_{dsp} = -\frac{1}{2} K_p \frac{W_p}{L_p} (V_{inv} - V_{dd} - V_{tp})^2$$

For

$$\frac{\beta_n}{2}(V_{inv} - V_{tn})^2 = \frac{\beta_p}{2}(V_{inv} - V_{dd} - V_{tp})^2$$

$$\text{or} \quad \frac{V_{inv} - V_{dd} - V_{tp}}{V_{inv} - V_{tn}} = -\sqrt{\frac{\beta_n}{\beta_p}}$$

$$\text{or} \quad V_{inv}\left(1 + \sqrt{\frac{\beta_n}{\beta_p}}\right) = V_{dd} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}$$

$$\text{or} \quad V_{inv} = \frac{V_{dd} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}} \qquad (4.17)$$

$\beta_n = \beta_p$ and $V_{tn} = -V_{tp}$, we get $V_{inv} = V_{dd}/2$.

In a CMOS process,

$$K_n = \frac{\mu_n}{\mu_p} \approx 2.5$$

$$K_p$$

To make $\beta_p = \beta_p$, one may choose

$$\frac{W}{L}_n = 2.5 \frac{W}{L}_p, \tag{4.18}$$

which can be realized with pMOS channel 2.5 times wider than the nMOS channel of the same length. The inverter acts as a symmetrical gate when this is realized.

Since both the transistors are in saturation, they act as current sources and at this point the current attains a maximum value as shown in Fig. 4.10c. This leads to power dissipation, known as *short-circuit power dissipation*. Later, we will derive a detailed expression for short-circuit power dissipation.

Region 4: $V_{inv} < V_{in} \leq V_{dd} - V_{tp}$ As the input voltage has been increased above V_{inv}, the nMOS transistor moves from the saturation region to the linear region,

whereas the pMOS transistor remains in saturation. With the increase in input volt-age beyond V_{inv}, the output voltage and also the drain current continue to drop. A representative point in this region is point D. In this region, nMOS transistor acts as a resistor and pMOS transistor acts as a current source.

The drain current for the two transistors are given by

$$I_{dsn} = \beta_n \left(V_{in} - V_{tn}\right)V_O - \frac{V_O^2}{2} \text{ and } I_{dsp} = -\beta_p \left(V_{in} - V_{dd} - V_{tp}\right)^2$$

As

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

we get

$$\beta_n (V_{in} - V_{tn})V_O - \frac{V_O^2}{2} = \beta_p (V_{in} - V_{dd} - V_{tp})^2$$

or

$$\frac{V_O^2}{2} - (V_{in} - V_{tn})V_O + \frac{\beta_p}{\beta_n}(V_{in} - V_{dd} - V_{tp})^2 = 0.$$

Solving for V_O, we get

$$V_O = (V_{in} - V_{tn}) - \sqrt{(V_{in} - V_{tn})^2 - \frac{\beta_p}{\beta_n}(V_{in} - V_{dd} - V_{tp})^2}$$

$$\therefore \quad V_O = (V_{in} - V_{tn}) - \sqrt{(V_{in} - V_{tn})^2 - \beta^p_n (V_{in} - V_{dd} - V_{tp})^2} \quad (4.19)$$

Similarly, we can find out the value of V_{IH} when the nMOS transistor operates in the linear region and pMOS transistor operates in the saturation region. Equating $I_{dsn} = I_{dsp}$, at this point, we get

$$\therefore \quad \frac{\beta_n}{2}2(V_{gsn} - V_{tn})\cdot V_{dsn} - V_{dsn}^2 = \frac{\beta_p}{2}(V_{gsp} - V_{tp})^2 .$$

$$\beta_n (V_{in} - V_{tn})\frac{dV_{out}}{dV_{in}} + V_{out}\frac{dV_{out}}{dV_{in}} - V_{out}$$

Substituting $V_{gsp} = -V_{in}$) and $V_{dsp} = -(V_{dd} - V_{out})$, we get

$$\therefore \quad \frac{\beta_n}{2}2(V_{in} - V_{tn})\cdot V_{out} - V_{out}^2 = \frac{\beta_p}{2}(V_{in} - V_{dd} - V_{tp})^2 .$$

Differentiating both sides with respect to V_{in}, we get

Now, subs

tituting $V_{in} = V_{IH}$ and $dV_{out} / dV_{in} = -1$, we get

$$\beta_n (-V_{IH} + V_{tn} + 2V_{out}) = \beta_p (V_{IH} - V_{dd} - V_{tp}) \quad (4.20)$$

or

$$V_{IH} = \frac{V_{dd} + V_{tp} + \frac{\beta_n}{\beta_p}(2V_{out} + V_{tn})}{1 + \frac{\beta_n}{\beta_p}}. \quad (4.21)$$

(4.22)

This equation can be solved with the Kirchhoff's current law (KCL) equation to obtain numerical values of V_{IH} and V_{out}. For $\beta_n / \beta_p = 1$, the value of $V_{IH} = \frac{1}{8}(5V_{dd} - 2V_{tn})$.

For a symmetric inverter, $V_{IH} + V_{IL} = V_{dd}$.

Noise Margin:

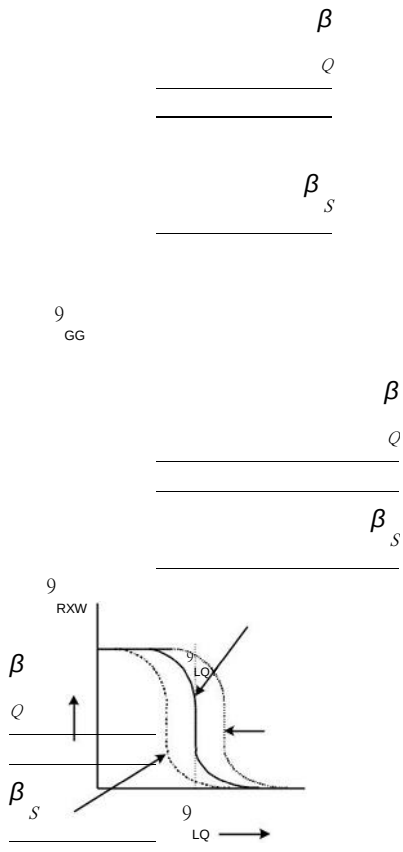


Fig. 4.11 Transfer characteristics for different inverter ratio

$$NM_L = V_{IL} - V_{OL} = V_{IL}$$

$$NM_H = V_{OH} - V_{IH} = V_{dd} - V_{IH}$$

$$= V_{IL}$$

which are equal to each other.

Region 5: $V_{dd} - |V_{tp}| < V_{in} \leq V_{dd}$ In this region, the pull-up pMOS transistor remains

OFF and the pull-down nMOS transistor goes to deep saturation. However, the current flow through the circuit is zero as the p transistor is OFF and the output voltage $V_O = 0$.

Based on the above discussions, key features of the CMOS inverter are high-lighted below:

It may be noted that unlike the use of nMOS enhancement- or depletion-mode transistor as a pull-up device, in this case, there is no current flow either for '0' or '1' inputs. So, there is no static power dissipation. Current flows only during the transition period. So, the static power dissipation is very small. Moreover, for low and high inputs, the roll of the pMOS and nMOS transistors are complementary; when one is OFF, the other one is ON. That is why this configuration is known as the complementary MOS or CMOS inverter. Another advantage is that full high and low levels are generated at the output. Moreover, the output voltage is independent of the relative dimensions of the pMOS and nMOS transistors. In other words, the CMOS circuits are ratioless.

β_n / β_p *Ratio:* As we have mentioned earlier, the low- and high-level outputs of a CMOS inverter are not dependent on the inverter ratio. However, the transfer characteristic is a function of the β_n / β_p ratio. The transfer characteristics for three different ratio values are plotted in Fig. 4.11. Here, we note that the voltage at which

the gate switches from high to low level (V_{inv}) is dependent on the β_n / β_p ratio. V_{inv} increases as β_n / β_p decreases. For a given process technology, the β_n / β_p can be

changed by changing the channel dimensions, i.e., the channel length and width. Keeping L the same, if we increase W_n / W_p ratio, the transition moves towards the left and as W_n / W_p is decreased, the transition moves towards the right as shown in Fig. 4.11. As the carrier mobility depends on temperature, it is expected that the transfer characteristics will be affected with the temperature. However, both β_n and β_p are affected in the same manner ($\beta \propto T^{-1.5}$) and the ratio β_n / β_p remains more or less the same. On the other hand, both the threshold voltages V_{tn} and V_{tp}

Table 4.1 Comparison of the inverters

Inverters	V_{LO}	V_{HI}	Noise margin	Power
Resistor	Weak	Strong	Poor for low	High
nMOS depletion	Weak	Strong	Poor for low	High
nMOS enhancement	Weak	Weak	Poor for both low and high	High
Psuedo-nMOS	Weak	Strong	Poor for low	High
CMOS	Strong	Strong	Good	Low

nMOS n-type metal–oxide–semiconductor, *CMOS* complementary metal–oxide–semiconductor

decrease with increase in temperature leading to some shrinkage of the region I and expansion of region V.

4.3.6 Comparison of the Inverters

Table 4.1 summarizes the performance of the five different types of inverters discussed in this section. As given in column 2, low output level is weak, i.e., the out-put does not go down to 0 V, for all the inverters except CMOS. High-output level is weak only for the inverter with the enhancement-mode nMOS transistor as pull-up devices. For all other inverters, full high-level (V_{dd}) is attained as shown in column 3. As shown in column 4, CMOS gives a good noise margin both for high- and low-logic levels. The inverters with nMOS enhancement-mode transistor as pull up have poor noise margin both for high- and low-logic levels. For all other types of invert-ers, noise margin is poor for low levels. So far, as power consumption is concerned, all inverters except CMOS draw DC current from the supply when the input is high. As a consequence, only CMOS consumes lower power compared to other types of inverters. From this table, we can conclude that CMOS is superior in all respects compared to other types of logic circuits because of these advantages. As a conse-quence, CMOS has emerged as the dominant technology for the present-day VLSI circuits. In case of nMOS depletion mode of transistor as a pull-up device, low-level noise margin is poor, and in case of nMOS enhancement-mode transistor, the noise margin is poor for both

low and high levels. It may be noted that CMOS circuits provide better noise margin compared to other two cases as given in Table 4.1.

4.4 Inverter Ratio in Different Situations

In a practical circuit, an inverter will drive other circuits, and, in turn, will be driven by other circuits. Different inverter ratios will be necessary for correct and satis-factory operation of the inverters. In this section, we consider two situations—an inverter driven by another inverter and an inverter driven through one or more pass transistors—and find out the inverter ratio for these two cases.

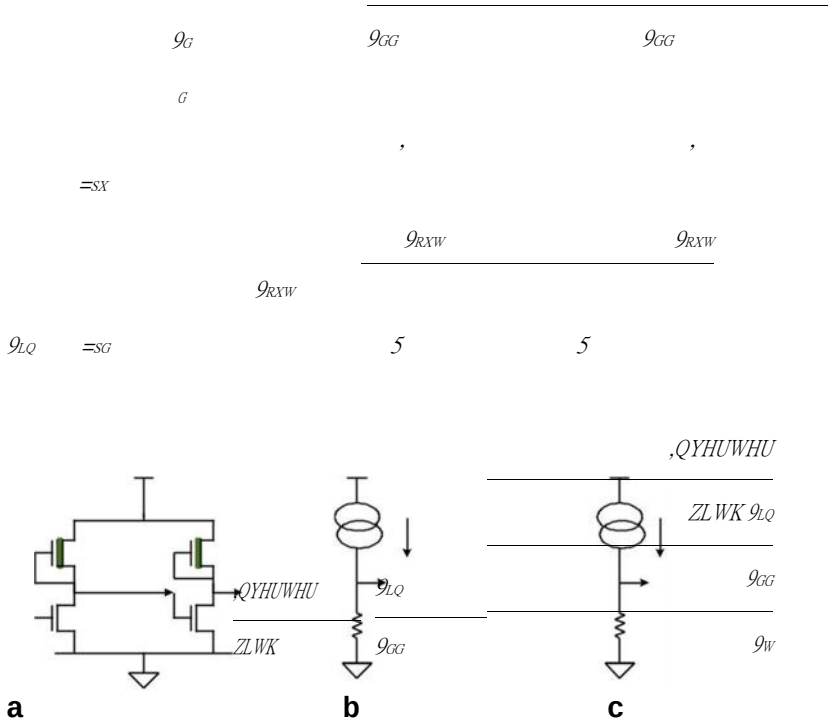


Fig. 4.12 **a** An nMOS inverter driven by another inverter; **b** inverter with $V_{in} = V_{dd}$; and **c** inverter with $V_{in} = V_{dd} - V_t$. nMOS n-type metal–oxide–semiconductor, V_{in} voltage input to the inverter, V_{dd} positive supply rail, V_t inverter threshold voltage

4.4.1 An nMOS Inverter Driven by Another Inverter

Let us consider an nMOS inverter with depletion-type transistor as an active load is driving a similar type of inverter as shown in Fig. 4.12. In order to cascade two or more inverters without any degradation of voltage levels, we have to meet the con-

dition $V_{in} = V_{out} = V_{inv}$; and for equal margins, let us set $V_{inv} = 0.5\ V_{dd}$. This condition is satisfied when both the transistors are in saturation, and the drain current is given by

$$I = K\ W (V_{gs} - V)^2 \qquad (4.23)$$

$$_{ds}\ \frac{}{L}\ \frac{}{2}$$

For the depletion-mode transistor,

$$\therefore \quad I_{ds}^{pu} = K \frac{W_{pu}}{L_{pu}} \frac{(-V_{tdp})^2}{2}, \tag{4.24}$$

where V_{tdp} is the threshold voltage of the depletion-mode transistor and $V_{gs} = 0$. And for the enhancement-mode transistor,

$$\therefore \quad I_{ds} = K \frac{W_{pd}(V_{inv} - V_{tn})^2}{2L_{pd}} \tag{4.25}$$

Equating these currents, we get

$$\therefore \quad \frac{W_{pd}}{L_{pd}}(V_{inv} - V_{tn})^2 = \frac{W_{pu}}{L_{pu}}(-V_{tdp})^2 \tag{4.26}$$

Assuming $Z_{pd} = L_{pd} / W_{pd}$ and $Z_{pu} = L_{pu} / W_{pu}$, where Z is known as the *aspect ratio* of the MOS devices, we have

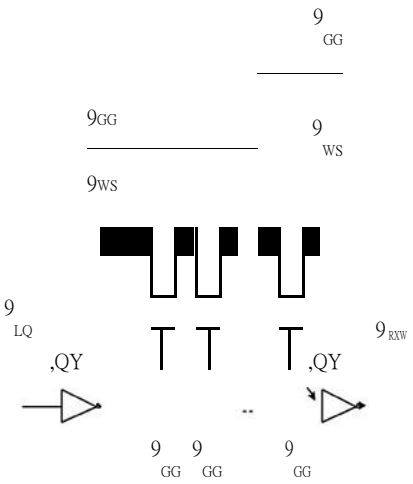


Fig. 4.13 An inverter driven through one or more pass transistors

$$\frac{1}{Z_{pd}} = \frac{(V_{inv} - V_{tn})^2}{} \quad \frac{1}{Z_{pu}} \left(-V_{tdp} \right)^2$$

$$\sqrt{\frac{}{}}$$

$$Z^{pu} (V_{inv} - V_{tn}) = -V_{tdp} \quad (4.27)$$

$$\text{or} \quad V_{inv} = V_{tn} - \frac{V_{tdp}}{\frac{Z_{pu}}{Z_{pd}}}$$

Substituting typical values $= 0.2V_{dd}$ $V_{tdp} = -0.6V_{dd}$ and $V_{inv} = V_{tn}$, $0.5V_{dd}$,

we get

$$\frac{Z_{pu}}{Z_{pd}} = \frac{4}{1}. \quad (4.28)$$

This ratio Z_{pu} / Z_{pd} is known as the inverter ratio (R_{inv}) of the inverter, and it is 4:1 for an inverter directly driven by another inverter.

4.4.2 An nMOS Inverter Driven Through Pass Transistors

Here, an nMOS inverter is driven through one or more pass transistors as shown in Fig. 4.13. As we have seen in Sect. 4.1, a pass transistor passes a weak high level. If V_{dd} is applied to the input of a pass transistor, at the output, we get (V_{dd}–V_{tp}), where V_{tp} is the threshold voltage of the pass transistor. Therefore, instead of V_{dd}, a degraded high level (V_{dd}–V_{tp}) is applied to the second inverter. We have to ensure that the same voltage levels are produced at the outputs of the two Inverters in spite of different input voltage levels.

First, let us consider inverter 1 with input voltage V_{dd}. In this situation, the pull-down transistor is in active mode and behaves like a resistor, whereas the pull-up transistor is in saturation, and it represents a current source.

For the pull-down transistor

$$I_{ds} = K \frac{W}{L}_{pd1} \left(V - \frac{}{2} \right)^2 \quad (4.29)$$

Therefore,

$$R = \frac{V}{I} = \frac{1}{K} \frac{L_{pd1}}{W} \frac{1}{V_{dd} - V_{tn} - \frac{V}{2}}. \quad (4.30)$$

Ignoring

$$\frac{V}{2} \text{ factor, } R_1 = \frac{Z_{pd1}}{K} \frac{1}{(V_{dd} - V_{tn})}. \quad (4.31)$$

Now, for the depletion-mode pull-up transistor, $V_{gs} = 0$,

$$I_1 = I_{ds} = K \frac{W_{pu1}}{L_{pu1}} \frac{(-V_{tdp})^2}{2}. \quad (4.32)$$

The product

$$I_1 R_1 = V_{out1} = \frac{Z_{pd1}}{Z_{pu1}} \frac{1}{V_{dd} - V_{tn}} \frac{(-V_{tdp})^2}{2}. \quad (4.33)$$

In a similar manner, we can consider the case of inverter 2, where the input voltage is $(V_{dd} - V_{tp})$. As in case of inverter 1, we get

$$\therefore \quad R_2 \approx \frac{Z_{pd2}}{K\,(V_{dd} - V_{tp} - V_{tn})^2} \quad \text{and} \quad I_2 = K\,\frac{1}{Z_{pu2}}\,\frac{(-V_{td})^2}{2} \tag{4.34}$$

Therefore,

$$\therefore \quad V_{out2} = I_2 R_2 = \frac{Z_{pd2}}{Z_{pu2}}\,\frac{1}{(V_{dd} - V_{tp} - V_{tn})^2}\,\frac{(-V_{td})^2}{2} \tag{4.35}$$

If we impose the condition that both the inverters will have the same output,

$$\therefore \quad V_{out1} = V_{out2} \text{ then, } I_1 R_1 = I_2 R_2 \quad \text{or} \quad \frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}}\,\frac{(V_{dd} - V_{tn})}{(V_{dd} - V_{tp} - V_{tn})} \tag{4.36}$$

Substituting typical values $V_{tn} = 0.2V_{dd}$ and $V_{tp} = 0.3V_{dd}$,

we get

$$\therefore \quad \frac{Z_{pu2}}{Z_{pd2}} = \frac{4.0}{2.5}\,\frac{Z_{pu1}}{Z_{pd1}} \quad \text{or} \quad \frac{Z_{pu2}}{Z_{pd2}} \approx 2.\,\frac{Z_{pu1}}{Z_{pd1}} = 1.8 \tag{4.37}$$

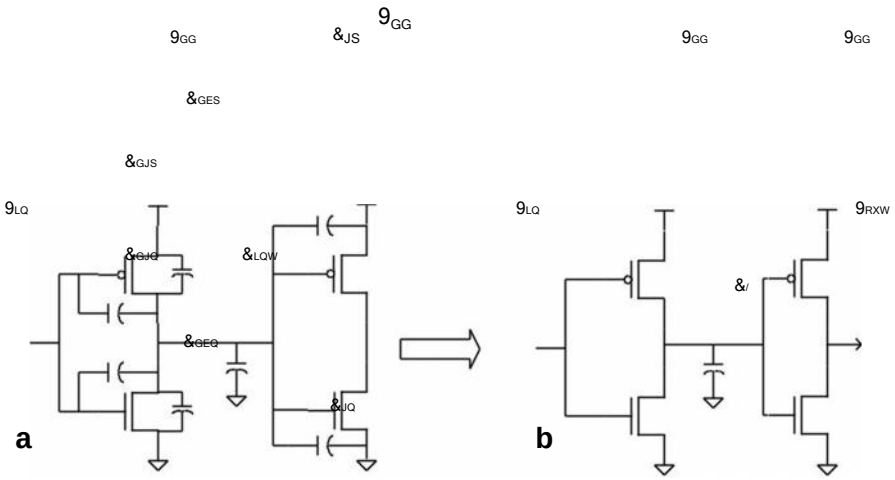
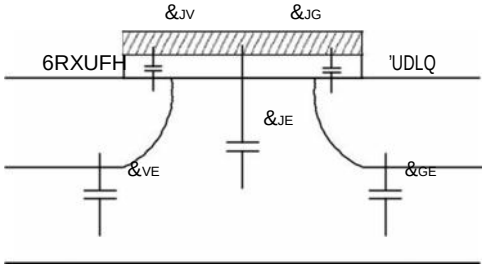


Fig. 4.14 **a** Parasitic capacitances of a CMOS inverter. **b** CMOS complementary metal–oxide–semiconductor

Fig. 4.15 Internal para-sitic capacitances of an MOS transistor. *MOS* metal–oxide–semiconductor



It may be noted that, if there are more than one-pass transistors before the second inverter, the degradation in the high voltage level is the same. Therefore, we may conclude that, if an inverter is driven through one or more pass transistors, it should have inverter ratio $Z_{pu} / Z_{pd} \geq 8 / 1$.

4.5 Switching Characteristics

So far, we have discussed the DC characteristics of an inverter, which gives us the behaviour of the inverter in static conditions when inputs do not change. To understand the switching characteristics, it is necessary to understand the parameters that affect the speed of the inverters. In this section, we consider the dynamic behav-iour of a CMOS inverter. A CMOS inverter driving a succeeding stage is shown in Fig. 4.14a. As shown in the figure, various parasitic capacitances exist (Fig. 4.15).

The capacitances C_{gd} and C_{gs} are mainly due to gate overlap with the diffusion regions, whereas C_{db} and C_{gb} are voltage-dependent junction capacitances. The capacitance C_{out} is the lumped value of the distributed capacitances due to interconnection and C_{gn} and C_{gp} are due to the thin oxide capacitances over the gate area of the nMOS and pMOS transistors, respectively. For the sake of simplicity, all the capacitances are combined into an equivalent lumped capacitance C_L, which is connected as the load capacitance of the inverter as shown in Fig. 4.14b. Here,

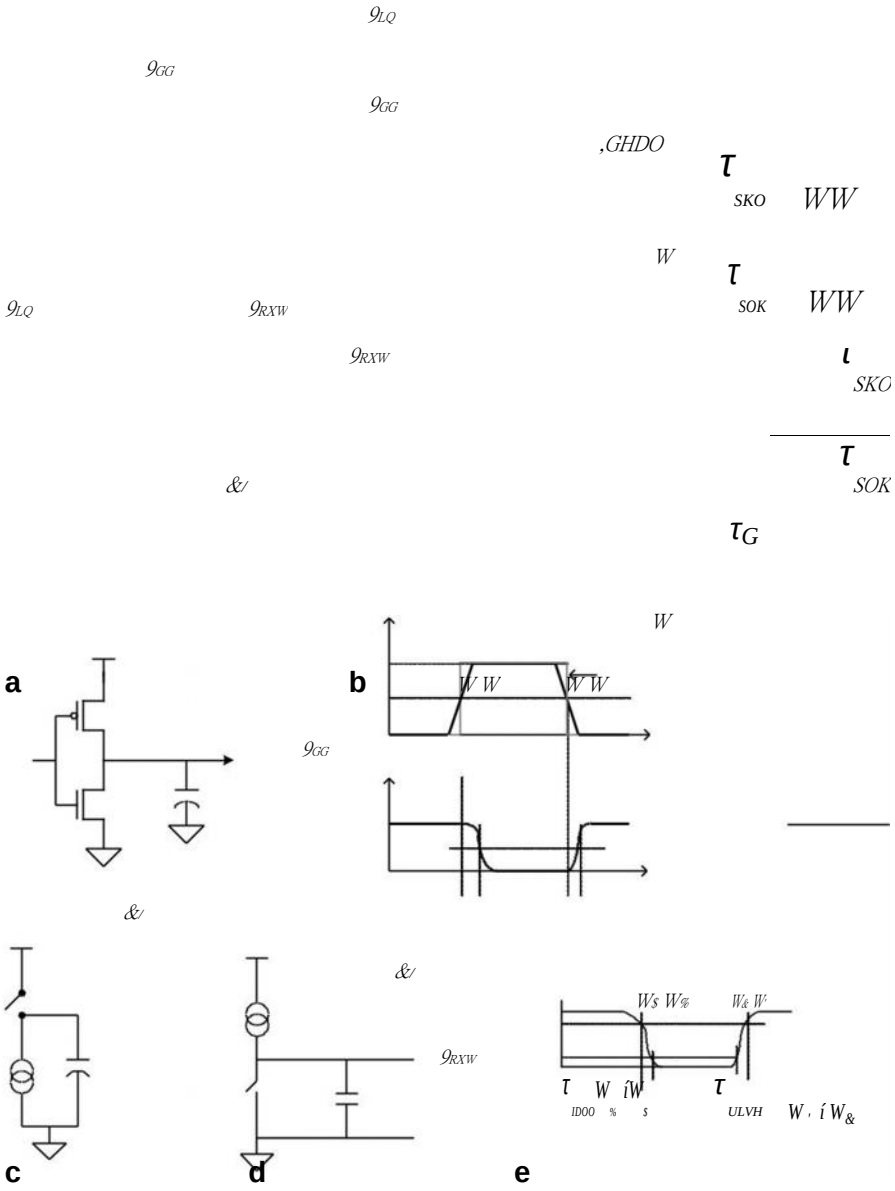


Fig. 4.16 **a** CMOS inverter; **b** delay-time timings; **c** fall-time model; **d** rise-time model; **e** Rise time and fall times. *CMOS* complementary metal–oxide–semiconductor

$$C_L = C_{dgn} + C_{dgp} + C_{dbn} + C_{dbp} + C_{int} + C_{gn} + C_{gp}.$$

We assume that an ideal step waveform, with zero rise and fall time is applied to the input as shown in Fig. 4.14b. The delay t_d is the time difference between the midpoint of the input swing and the midpoint of the swing of the output signal. The load capacitance shown at the output of the inverter represents the total of the input capacitance of driven gates, the parasitic capacitance at the output of the gate itself, and the wiring capacitance. In the following section, we discuss the estimation of the load capacitances.

4.5.1 Delay-Time Estimation

The circuit for high-to-low propagation delay time t_{PHL} estimation can be modeled as shown in Fig. 4.16c. It is assumed that the pull-down nMOS transistor remains in saturation region and the pMOS transistor remains off during the entire discharge period. The saturation current for the nMOS transistor is given by

$$I_{dsn} = \frac{\beta_n}{2}(V_{gs} - V_{tn})^2 \tag{4.38}$$

The capacitor C_L discharges from voltage V_{dd} to $V_{dd}/2$ during time t_0 to t_1.

$$\frac{V_{dd}}{2} = V_{dd} - \frac{\beta_n}{2C_L}(V_{gs} - V_{tn})^2 \times (t_1 - t_0). \quad (4.39)$$

Substituting $V_{gs} = V_{dd}$ and $t_{phl} = t_1 - t_0$,

we get the fall delay

$$t_{phl} = \frac{V_{dd}}{\beta_n} \times \frac{C_L}{(V_{dd} - V_{tn})^2}. \quad (4.40)$$

When the input goes from high (V_{dd}) to low, initially the output is at a low level. The pull-up pMOS transistor operates in the saturation region. In a similar manner, based on the model of Fig. 4.16d, the rise time delay is given by

$$t_{plh} = \frac{C_L}{\beta_p} \frac{1}{\left(1 - \left|\frac{V_{tp}}{V_{dd}}\right|\right)^2}. \quad (4.41)$$

For equally sized MOS transistors, the rise delay is greater than the fall delay be-cause of lower mobility of p-type carriers. So, the pMOS

transistor should be sized by increasing the width W_p in order to get equal size and fall delays.

Delay Time By taking average of the rise and full delay time, we get the delay time

$$t_d = \frac{1}{2}\left(t_{phl} + t_{plh}\right)$$

$$= \frac{C_L}{2V_{dd}}\left[\frac{1}{\beta_n\left(1 - \frac{V_{tn}}{V_{dd}}\right)^2} + \frac{1}{\beta_p\left(1 - \frac{V_{tp}}{V_{dd}}\right)^2}\right] \quad (4.42)$$

Assuming $V_{tn} = -V_{tp} = V_t$, the delay is given by

$$t_d = \frac{L_n}{k_n W_n} + \frac{L_p}{k_p W_p} \frac{C_L}{V_{dd}\left(1 - \frac{V_t}{V_{dd}}\right)^2} \quad (4.43)$$

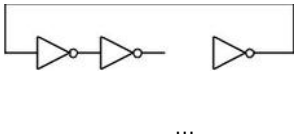


Fig. 4.17 Ring oscillator realized using odd number of inverters

This expression gives us a simple analytical expression for the delay time. It is observed that the delay is linearly proportional to the total load capacitance C_L. The delay also increases as the supply voltage is scaled down, and it increases drasti-cally as it approaches the threshold voltage. To overcome this problem, the thresh-old voltage is also scaled down along with the supply voltage. This is known as the *constant field scaling*. Another alternative is to use *constant voltage scaling*, in which the supply voltage is kept unchanged because it may not be always possible to reduce the supply voltage to maintain electrical compatibility with other sub-systems used to realize a complete system. The designer can control the following parameters to optimize the speed of CMOS gates.

- The width of the MOS transistors can be increased to reduce the delay. This is known as *gate sizing*, which will be discussed later in more detail.

- The load capacitance can be reduced to reduce delay. This is achieved by using transistors of smaller and smaller dimensions as provided by future-generation devices.

- Delay can also be reduced by increasing the supply voltage V_{dd} along and/or reducing the threshold voltage V_t of the transistors.

4.5.2 Ring Oscillator

To characterize a particular technology generation, it is necessary to measure the gate delay as a measure of the performance [1]. It is difficult to measure delays of the order of few nanoseconds with the help of an oscilloscope. An indirect approach is to use a ring oscillator to measure the delay. A ring oscillator is realized by a cascade connection of odd number of a large number of inverters, where the output of the last stage is connected to the input of the first stage, as shown in Fig. 4.17. The circuit oscillates because the phase of the signal fed to the input stage from the output stage leads to positive feedback. The frequency of oscillation for this closed-loop cascade connection is determined by the number of stages and the de-lay of each stage. The output waveform of a three-stage ring oscillator is shown in Fig. 4.18. The time period can be expressed as the sum of the six delay times

$$
\begin{aligned}
T &= t_{phl1} + t_{plh2} + t_{phl3} + t_{plh1} + t_{phl2} + t_{plh3} \\
&= (t_{phl1} + t_{plh1}) + (t_{phl2} + t_{plh2}) + t_{phl3} + t_{plh3} \\
&= 2t_d + 2t_d + 2t_d \\
&= 6t_d \\
&= 2.3t_d .
\end{aligned}
$$

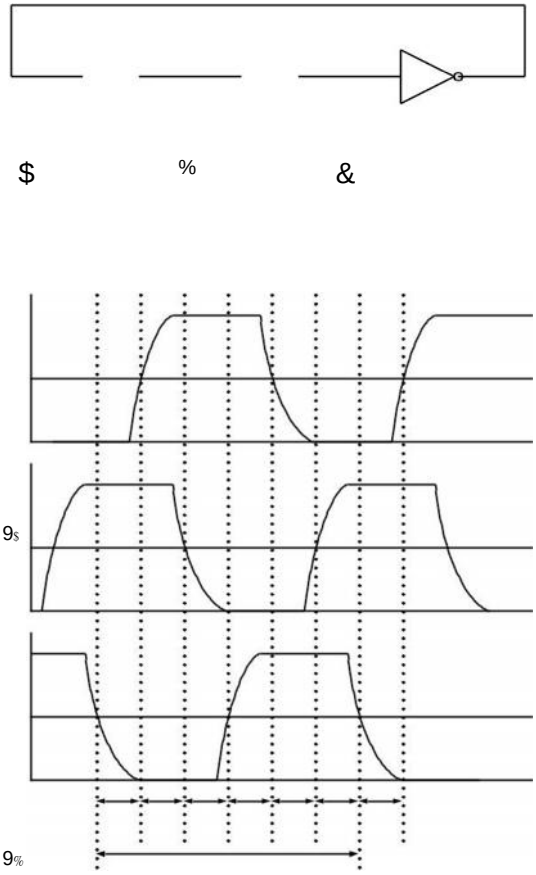


Fig. 4.18 Output waveform of a three-stage ring oscillator

For an *n*-stage (where n is an odd number) inverter, the time period $T = 2 \bullet n\tau_d$. Therefore, the frequency of oscillation $f = 1 / 2nt_d$ or $t_d = 1 / 2nf$. It may be noted that the delay time can be obtained by measuring the frequency of the ring oscil-lator. For better accuracy of the measurement of frequency, a suitable value of *n* (say 151) can be chosen. This can be used for the characterization of a fabrication process or a particular design. The ring oscillator can also be used for on-chip clock generation. However, it does not provide a stable or accurate clock frequency due to dependence on temperature and other parameters. To generate stable and accurate clock frequency, an off-chip crystal is used to realize a crystal oscillator.

4.6 Delay Parameters

In order to understand the delay characteristics of MOS transistors, we have to con-sider various parameters such as resistance and capacitances of the transistors along with wiring and parasitic capacitances. For ease of understanding and simplified treatment, we will introduce simplified circuit parameters as follows:

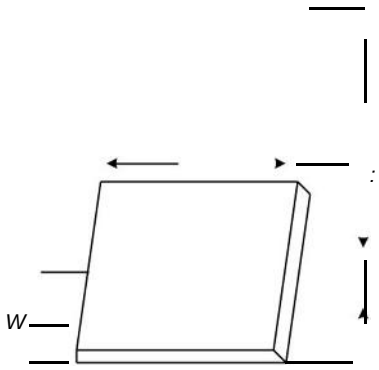


Fig. 4.19 One slab of conducting material

4.6.1 Resistance Estimation

Let us consider a rectangular slab of conducting material of resistivity ρ, of width *W*, of thickness *t* and length *L* as shown in Fig. 4.19. The resistance between A and B, R_{AB} between the two opposite sides is given by

$$\therefore \quad R_{AB} = \frac{\rho L}{t \cdot W} = \frac{\rho L}{A}\Omega, \tag{4.44}$$

where A is the cross section area. Consider the case in which $L = W$, then

$$\therefore \quad R_{AB} = \frac{\rho}{t} = R\Omega \tag{4.45}$$

where R_s is defined as the resistance per square or the sheet resistance.

Thus, it may be noted that R_s is independent of the area of the square. The actual value of R_s will be different for different types of layers, their thickness, and resis-tivity. For poly-silicon and metal, it is very easy to

envisage the thickness, and their resistivities are also known. But, for diffusion layer, it is difficult to envisage the depth of diffusion and impurity concentration level. Although the voltage–current characteristics of an MOS transistor is nonlinear in nature, we may, approximate its behaviour in terms of a 'channel' resistance to estimate delay performance. We know that in the linear region

$$I_{ds} = \beta (V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2} \quad (4.46)$$

Assuming, $V_{ds} \ll (V_{gs} - V_t)$, we may ignore the quadratic term to get

$$I_{ds} = \beta (V_{gs} - V_t).V_{ds}$$

$$\text{or } R_c = \frac{V_{ds}}{I_{ds}} = \frac{1}{\beta (V_{gs} - V_t)} = \frac{1}{\mu C_g (V_{gs} - V_t)} \frac{L}{W} \quad (4.47)$$

$$= K\frac{L}{W} \text{ where } K = \frac{1}{\mu C_g (V_{gs} - V_t)}.$$

Table 4.2 Sheet resistances of different conductors

	Sheet resistance in ohm/sq.		
Layer	Min.	Typical	Max.
Metal	0.03	0.07	0.1
Diffusion (n^t, p^t)	10	25	100
Silicide	2	3	6
Poly-silicon	15	20	30
n-channel		10^4	
p-channel		2.5×10^4	

K may take a value between 1000 to 3000 Ω/sq. Since the mobility and the threshold voltage are functions of temperature, the channel resistance will vary with temperature. Typical sheet resistances for different conductors are given in Table 4.2.

Sheet resistance concept can be conveniently applied to MOS transistors and inverters. A diffusion and a poly-silicon layer are laid orthogonal to each other with overlapping areas to represent a transistor. The thinox mask layout is the hatched overlapped region. The poly-silicon and the thinox layer prevent diffusion to occur below these layers. Diffusion takes place in the areas where these two layers do not overlap. Assuming a channel length of 2λ and channel width of 2λ, the resistance of this channel is given by

$$R = 1 \text{ square} \times R_s \ \Omega \text{ per square where } R_s = 10^4\ \Omega.$$

In this case, the length to width ratio is 1:1.

For a transistor with $L = 8\lambda$ and $W = 2\lambda$,

$$Z = \frac{L}{W} = \frac{4}{1}.$$

Thus, the channel resistance = $4 \times R_s = 4 \times 10^4\ \Omega$. It can be regarded as four-unit squares in series. It is also possible to consider an inverter with a given inverter ratio of z_{pu}:z_{pd}. The most common ratio of 4:1 can be realized in one of the two ways as shown in Fig. 4.20.

4.6.2 Area Capacitance of Different Layers

From the structure of MOS transistors, it is apparent that each transistor can be rep-resented as a parallel-plate capacitor. The area capacitance can be calculated based on the dielectric thickness (silicon dioxide) between the conducting layers.

Here, area capacitance

$$C = \frac{\varepsilon_0 \varepsilon_{ins} A}{D} \text{ Farads}, \tag{4.48}$$

Table 4.3 Capacitance of different materials

Capacitance	Value in pF/μm^2	Relative value
Gate to channel	4×10^{-4}	1
Diffusion	1×10^{-4}	0.25
Poly-silicon	4×10^{-4}	0.1
Metal 1	0.3×10^{-4}	0.075
Metal 2	0.2×10^{-4}	0.50
Metal 2 to metal	0.4×10^{-4}	0.15
Metal 2 to poly	0.3×10^{-4}	0.075

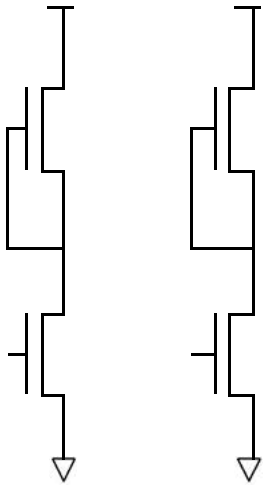


Fig. 4.20 Two different inverter configurations with inverter ratio 4:1

where D is the thickness of the silicon dioxide, A is the Area of place, ε_{ins} is the relative permittivity of SO_2 = 3.9, and ε_0 = 8.85 × 10–^{14} F/cm, permittivity of free space

For 5-μm MOS technology, typical area capacitance is given in Table 4.3:

4.6.3 Standard Unit of Capacitance C_g

The standard unit of capacitance C_g is defined as the gate-to-channel capacitance of a minimum-size MOS transistor. It gives a value approximate to the technology and can be conveniently used in calculations without associating with the absolute value.

Considering 5 μm technology, where gate area = 5 μm x 5 μm = 25 μm^2, area ca-pacitance = 4×10^{-4} pF/cm^2.

Therefore, standard value of C_g = 25 × 4 x 10^{-4} pF = .01 pF.

Example 2.2

Let us consider a rectangular sheet of length *L* and width *W*. Assuming *L* = 20 and

W = 3, the relative area and capacitances can be calculated as follows:

$$\text{Relative area} = \frac{20 \times 3}{2 \times 3} = 10$$

a.\ Consider the area in metal

\ capacitance to substrate = 15 × 0.075 C_g = 1.125 C_g

b.\ Consider the same are in poly

\ capacitance to substrate = 15 × 0.20 = 1.5 C_g

c.\ Considering it in diffusion

\ capacitance to substrate = 15 × 0.25 = 3.70 C_g

A structure occupying different layers can be calculated in the same manner.

4.6.4 The Delay Unit

Let us consider the situation of one standard gate capacitance being charged through one square of channel resistance.

Time constant = 1 R_s × 1 C_g s

For 5-μm technology

$\tau = 10^4\ \Omega \times 0.01$ pF = 0.1 ns

when circuit wiring and parasitic capacitances are taken into account, this figure increases by a factor of 2–3.

So, in practice τ = 0.2–0.3 ns.

The value of τ obtained from transit time calculations is

$$I_{sd} = \mu \frac{L_2}{\cdot V_{ds}}$$

Substituting the values corresponding to 5 μm technology

$$\tau_{sd} = \frac{25^2\ \mu m^2\ Vs \times 10^9\ ns 10^4 =}{0.13ns.\ 650 cm^2\ 3V\ 10^{10}\ s\mu m^2}$$

(Assuming C_g is charged from 0 V to 63 % of V_{dd})

It may be noted that this is very close to the theoretical estimate, and it is used as the fundamental time unit and all timings can be evaluated in terms of τ.

4.7 Driving Large Capacitive Loads

There are situations when a large load capacitance such as, long buffers, off-chip capacitive load or I/O buffer are to be driven by a gate [1]. In such cases, the delay can be very high if driven by a standard gate. A super buffer or a BiCMOS inverter is used or cascade of such gates of increasing can be used to achieve smaller delays as discussed in the following subsections.

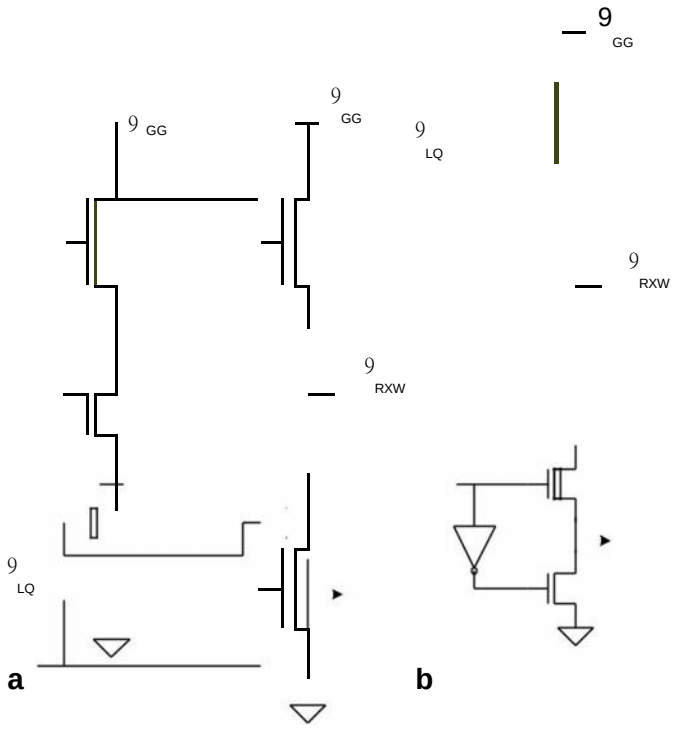


Fig. 4.21 **a** Inverting super buffer; **b** noninverting super buffer

4.7.1 Super Buffers

We have seen that one important drawback of the basic nMOS inverters (because of ratioed logic) in driving capacitive load is asymmetric drive capability of pull-up and pull-down devices. This is because of longer channel length (four times) of the pull-up device. Moreover, when the pull-down transistor is ON, the pull-up transis-tor also remains ON. As a consequence, a complete pull-down current is not used to discharge the load capacitance, but part of it is neutralized by the current pass-ing through the pull-up transistors. This asymmetry can be overcome in especially designed circuits known as super buffers. It is possible to realize

both inverting and noninverting super buffers as shown in Fig. 4.21. In standard nMOS inverter, the gate of the depletion-mode pull-up device is tied to the source. Instead, the gate of the pull-up device of the super buffer is driven by another inverter with about twice the gate drive. Thus, the pull-up device is capable of sourcing about four times the current of the standard nMOS inverter. This is the key idea behind both the inverting and noninverting types of super buffers. This not only overcomes the asymmetry but also enhances the drive capability.

A schematic diagram of nMOS super buffers, both inverting and noninverting types, are shown in Fig. 4.21. As shown, the output stage is a push–pull stage. The gate of the pull-up device is driven by a signal of opposite level of the pull-down de-vice, generated using a standard inverter. For the inverting-type super buffer, when the input voltage is low, the gates of both the pull-down transistors are low, the gates of both the pull-up devices are high, and the output is high. When the input voltage is high, the gates of both pull-down devices switch to high, the gates of both pull-up devices switch to low, and the output switches from high to low. The small-channel resistance of the pull-down device discharges the output load capacitor quickly. When the input voltage goes from high to low, the gate of the pull-up device quickly switches to high level. This drives the pull-down stage very quickly, and the output switches quickly from low to high. It can be shown that the high-to-low and low-to-high switching times are comparable.

In order to compare the switching speed of a standard inverter and a super buffer, let us compare the current sourcing capability in both the situations.

For a standard inverter, the drain current of the pull-up device in saturation ($0 < V_0 < 2$ V) and linear region are given as follows:

$$I_{ds}(\text{sat}) = \frac{\beta_{pu}}{2}(V_{gs} - V_{tdep})^2 \quad \text{where} \quad V_{tdep} = -3V$$
$$= \frac{\beta_{pu}}{2}(0+3)^2 = 4.5\,\beta_{pu}\,. \tag{4.49}$$

$$I_{ds}(\text{lin}) = \beta_{pu}\left[(V_{gs} - V_{tdep})V_{ds} - \frac{V_{ds}^2}{2}\right]$$
$$= \frac{\beta_{pu}}{2}\left[2(0+3)2.5 - (2.5)^2\right] = 4.38\,\beta_{pu}\,. \tag{4.50}$$

The average of the saturation current for $V_{ds} = 5$ V and linear current for $V_{ds} = 2.5$ V is approximately $4.4\,\beta_{pu}$.

For the super buffer, the pull-up device is always in linear region because $V_{gs} = V_{ds}$. The average pull-up current can be assumed to be an average of the linear currents at $V_{ds} = 5$ V and $V_{ds} = 2.5$ V.

$$I_{ds}(5v) = \frac{\beta_{pu}}{2}\left[2(5+3)5 - 5^2\right]$$

$$= 27.5\,\beta_{pu},$$

$$I_{ds}(2.5V) = \frac{\mu_{pu}}{2}2(2.5 - 3)2.8 + -(2.5^2)$$

$$= 10.62\beta_{pu}.$$

The average source current

$$= \frac{(27.5 + 10.62)\beta_{pu}}{2} = 19.06\beta_{pu}.$$

Therefore, the current drive is = 19.06 / 4.44 = 4.3 times that of standard inverter for the super buffer using totem pole output configuration. This high drive also al-leviates the asymmetry in driving capacitive loads and makes the nMOS inverter behave like a ratioless circuit.

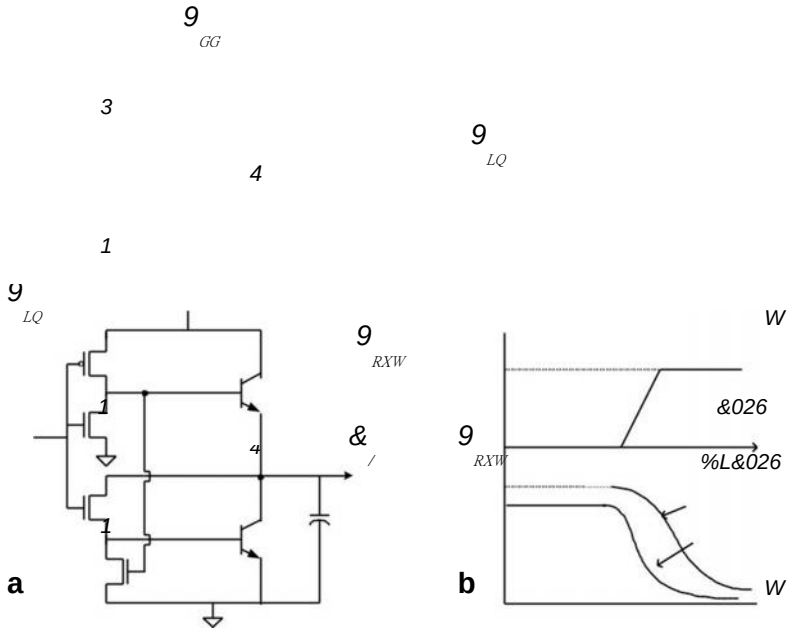


Fig. 4.22 **a** A conventional BiCMOS inverter; **b** output characteristics of static CMOS and BiCMOS. *CMOS* complementary metal–oxide–superconductor

4.7.2 BiCMOS Inverters

Higher current drive capability of bipolar NPN transistors is used in realizing BiCMOS inverters. Out of the several possible configurations, the conventional BiC-MOS inverter, shown in Fig. 4.22, is considered here. The operation of the inverter is quite straightforward. The circuit requires four MOS transistors and two bipolar NPN transistors Q_1 and Q_2. When the input V_{in} is low, the pMOS transistor P_1 is ON, which drives the base of the bipolar transistor Q_1 to make it ON.The nMOS transistor N_1 and N_2 are OFF, but transistor N_3 is ON, which shunts the base of Q_2 to turn it OFF. The current through Q_1 charges capacitor C_L and at the output V_{out}, we get a voltage ($V_{dd}-V_{be}$), where V_{be} is the base–emitter voltage drop of the transistor Q_1. If the input is high (V_{dd}), transistors P_1 and N_3 are OFF and N_1 and N_2 are ON. The drain current of N_2 drives the base of the transistor Q_2, which turns ON and the transistor N_1 shunts the base of Q_1 to turn it OFF. The capacitor C_L discharges through Q_2. The conventional BiCMOS gate gives high-current-drive capability, zero static power dissipation and high input impedance. The DC characteristic of the conventional BiCMOS inverter is shown in Fig. 4.22a. For a zero input voltage, the pMOS transistor

operates in the linear region. This drives the NPN transistor Q_1, and the output we get is $V_{dd} - V_{be}$, where V_{be} is the base–emitter voltage drop of Q_1. As input voltage increases, the subthreshold leakage current of the transistor

N_3 increases leading to a drop in the output voltage. For $V_{in} = V_{inv}$ ($V_{dd}/2$), both P_1 and N_2 transistors operate in the saturation region driving both the NPN transistors

(Q1 and Q2) ON. In this region, the gain of the inverter is very high, which leads to a sharp fall in the output voltage, as shown in Fig. 4.22b. As the input voltage is further increased, the output voltage drops to zero. The output voltage characteris-tics of CMOS and BiCMOS are compared in Fig. 4.22b. It may be noted that the BiCMOS inverter does not provide strong high or strong low outputs. High output

is $V_{dd}-V_{CE1}$, where V_{CE1} is the saturation voltage across Q1 and the low-level output is V_{CE2}, which is the saturation voltage across Q2.

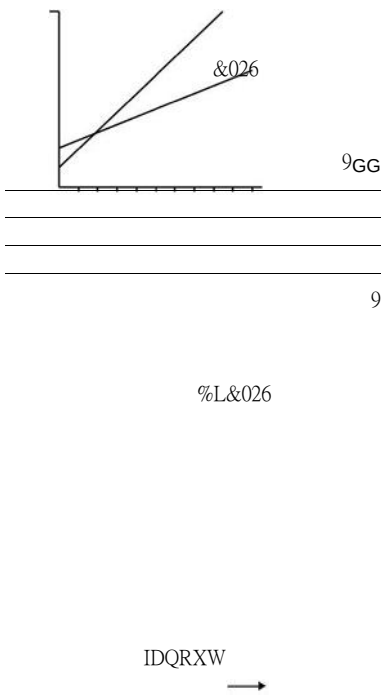


Fig. 4.23 Delay of static CMOS and BiCMOS for different fan-out. *CMOS* complementary metal–oxide–superconductor

The delays of CMOS and BiCMOS inverters are compared in Fig. 4.23 for different fan-outs. Here, it is assumed that for CMOS, W_p = 15 μm and W_n = 7 μm and for the BiCMOS, $W_p = W_n$ = 10 μ m and $W_{Q1} = W_{Q2}$ = 2 μ m. It may be noted that for fan-out of 1 or 2, CMOS provides smaller delay compared to BiCMOS due to longer delay through its two stages. However, as fan-out increases further, BiCMOS performs better. So, BiCMOS inverters should be used only for larger fan-out.

4.7.3 Buffer Sizing

It may be observed that an MOS transistor of unit length (2λ) has gate capacitance proportional to its width (W), which may be multiple of λ. With the increase of the width, the current driving capability is increased. But this, in turn, also increases the gate capacitance. As a consequence, the delay in driving a load capacitance C_L by a transistor of gate capacitance C_g is given by the relationship $(c_L / c_g)\tau$, where τ is the unit delay, or delay in driving an inverter by another of the same size.

Let us now consider a situation in which a large capacitive load, such as an out-put pad, is to be driven by an MOS gate. The typical value of such load capacitance is about 100 pF, which is several orders of magnitude higher than C_g. If such a load is driven by an MOS gate of minimum dimension (2λ x 2λ), then the delay will be $10^3\tau$. To reduce this delay, if the driver transistor is made wider, say $10^3 \times 2\lambda$, the delay of this stage becomes τ, but the delay in driving this driver stage is 1000τ, so, the total delay is 1001τ, which is more than the previous case. It has been observed that the overall delay can be minimized by using a cascaded stage of inverters of increasing size as shown in Fig. 4.24. For example, considering each succeeding stage is ten times bigger than that of the preceding stage, each stage gives a delay of 10τ. This results in a total delay of 31τ instead of 1001τ. Now, the question arises about the relative dimension of two consecutive stages. If relative dimension is large, fewer stages are needed, and if relative dimension is small, a large number of driver stages is needed. Let us assume that there are n stages and each buffer is scaled up by a factor of f, which gives stage delay of $f\tau$. For n stages, the delay is $nf\tau$. Let y be the ratio of the load capacitance to one gate capacitance. Then, for n buffer stages in cascade

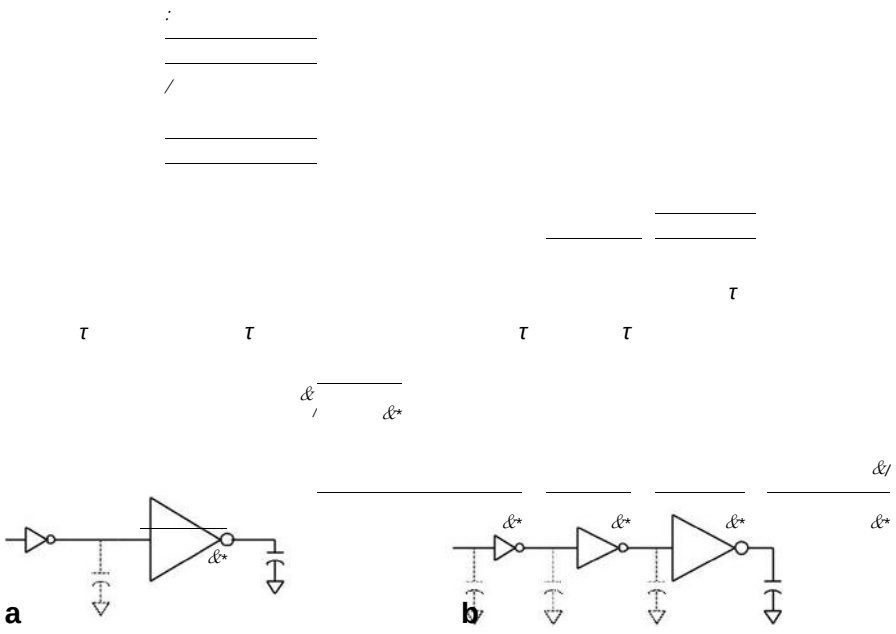


Fig. 4.24 **a** Using a single driver with *W* to *L* ratio of 1000:1; **b** using drivers of increasing size with stage ratio of 10. *W* width; *L* length

$$y = \frac{c_n}{c_g} = \frac{c_{g1}}{c_g}\cdot\frac{c_{g2}}{c_{g1}}\cdot\frac{c_{g3}}{c_{g2}}\cdots\frac{c_{gn}}{c_{g(n-1)}} = f^n$$

or $\ln y = n \ln f$

or $n = \dfrac{\ln y}{\ln f}$.

As each stage gives a delay of $f\tau$, the total delay of n stages is

(4.52)

(4.51)

The delay is proportional to ln(y), which is dependent on the capacitive load C_L and for a given load ln(y) is constant. To find out the value of f at which the delay is minimum, we can take the derivative of f/ln(f) with respect to f and equate it to zero.

$$\frac{d\frac{f}{\ln(f)}}{dt} = \frac{\ln f - f\frac{1}{f}}{\ln^2 f} = 0.$$

or ln(f) = 1 or f = e, where e is the base of natural logarithm.

The minimum total delay is

$$t_{min} = nf\tau = e\tau \ln\frac{C_L}{}$$

(4.54)

(4.53)

Therefore, the total delay is minimized when each stage is larger than the previous one by a factor of *e*. For 4:1 nMOS inverters, delay per pair is 5τ and overall delay is

$$\frac{n}{2}e5\tau = \frac{5}{2}ne\tau = 2.5en\tau \ (n \text{ even}).$$

Fig. 4.25 Variation of delay with stage ratio

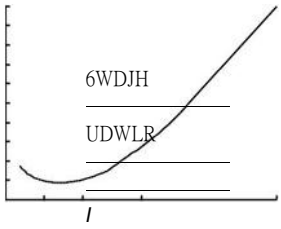


Table 4.4 Variation of delay with buffer sizing

f	f/e ln(f)
2	1.062
e	1.000
3	1.005
4	1.062
5	1.143
6	1.232

For CMOS, the overall delay is = 3.5$en\tau$ (n even)

It may be noted that the minimum total delay is a slowly varying function of f as shown in Fig. 4.25. Therefore, in practice, it may be worthwhile to scale the buffers by 4 or 5 to 1. This allows saving of the real estate at the cost small speed penalty, say about 10 % that the minimum possible delay (Table 4.4).

MOS Combinational Circuits

Abstract This chapter deals with metal–oxide–semiconductor (MOS) combina-tional circuits. The operation of pass-transistor logic circuits based on switch logic is explained and advantages and limitations of pass-transistor logic circuits are highlighted. The different members of pass-transistor logic family are introduced. The static and switching characteristics of multi-input NOR and NAND gates based on gate logic are discussed in detail. The operation of MOS dynamic circuits are explained and charge sharing and charge leakage problems associated with MOS dynamic circuits are introduced. The clock skew problem of MOS dynamic circuits is also discussed. The operation of domino-complementary metal–oxide–semi-conductor (CMOS) and (NO Race) NORA-CMOS dynamic circuits is explained. Realization of several example functions, such as full adder, parity generator, and priority encoder, using the three logic styles is considered and their area and delay are compared.

Keywords Pass-transistor logic · MOS dynamic circuits · Complementary pass-transistor logic · Swing-restored pass-transistor logic · Double pass-transistor logic · Charge sharing problem · Charge leakage problem · Clock skew · Domino CMOS · NORA CMOS · Fan-in · Fan-out · Switching characteristic · Pre-charge logic · Priority encoder · Parity generator

5.1 Introduction

There are two basic approaches of realizing digital circuits by metal–oxide–semi-conductor (MOS) technology: switch logic and gate logic. A switch logic is based on the use of "*pass transistors*" or transmission gates, just like relay contacts, to steer logic signals through the device. On the other hand, *gate logic* is based on the realization of digital circuits using inverters and other conventional gates, as it is typically done in transistor–transistor logic (TTL) circuits. Moreover, depending on how circuits function, they can also be categorized into two types: *static* and *dynamic* gates. In case of static gates, no clock is necessary for their operation and the output remains steady for as long as the supply voltage is maintained. Dynamic circuits are realized by making use of the information storage capability of the in-trinsic capacitors present in the MOS circuits.

DOI 10.1007/978-81-322-1937-8_5,

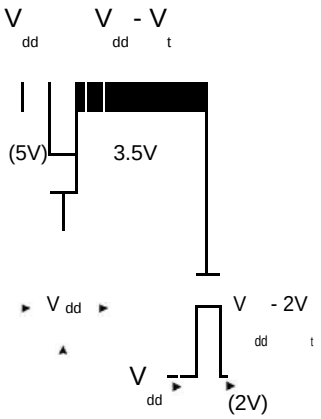


Fig. 5.1 Pass-transistor output driving another pass-transistor stage

A basic operation of pass-transistor logic circuits is considered in Sect. 5.2. Re-alization of pass-transistor logic circuits, overcoming the problems highlighted in Sect. 5.2.1, has been presented in Sect. 5.2.2. The advantages and disadvantages of pass-transistor logic circuits are considered in Sect. 5.2.2. Pass-transistor logic families are introduced in Sect. 5.2.3. Logic circuits based on gate logic are consid-ered in Sect. 5.3. The operation of n-type MOS (nMOS), NAND, and NOR gates has been presented in Sect. 5.3.2. Realization of complementary metal–oxide–semi-conductor (CMOS) NAND and NOR gates has been discussed in Sect. 5.3.3. The switching characteristic of CMOS gates is considered in Sect. 5.3.4. Section 5.4 introduces MOS dynamic logic circuits. Section 5.5 presents the realization of some example circuits.

5.2 Pass-Transistor Logic

As pass transistors functionally behave like switches, logic functions can be re-alized using pass transistors in a manner similar to relay contact networks. Re-lay contact networks have been presented in the classical text of Caldwell (1958). However, a closer look will reveal that there are some basic differences between relay circuits and pass-transistor networks. Some of the important differences are discussed in this section [1]. In our discussion, we assume that the pass transistors are nMOS devices.

As we have seen in Sect. 3.6, a high-level signal gets degraded as it is steered through an nMOS pass transistor. For example, if both drain and gate are at some high voltage level, the source will rise to the lower of the two potentials: V_{dd} and ($V_{gs}-V_t$). If both drain and gate are at V_{dd}, the source voltage cannot exceed ($V_{dd}-V_t$). We have already seen that when a high-level signal is steered through a pass tran-sistor and applied to an inverter gate, the inverter has to be designed with a high inverter ratio (8:1) such that the gate drive (3.5 V assuming the threshold voltage equal to 1.5 V) is sufficient enough to drive the inverter output to an acceptable low level. This effect becomes more prominent when a pass-transistor output is allowed to drive the gate of another pass-transistor stage as shown in Fig. 5.1. As the gate voltage is 3.5 V, the high-level output from the second pass transistor can never exceed 2.0 V, even when the drain voltage is 5 V as shown in Fig. 5.1. In such a situation, the gate potential will not be sufficient enough to drive the output low of an inverter with an aspect ratio of 8:1. Therefore, while synthesizing nMOS pass-transistor logic, one must not drive the gate of a pass transistor by the output of another pass transistor. This problem does not exist either in relay logic or in case of transmission gate.

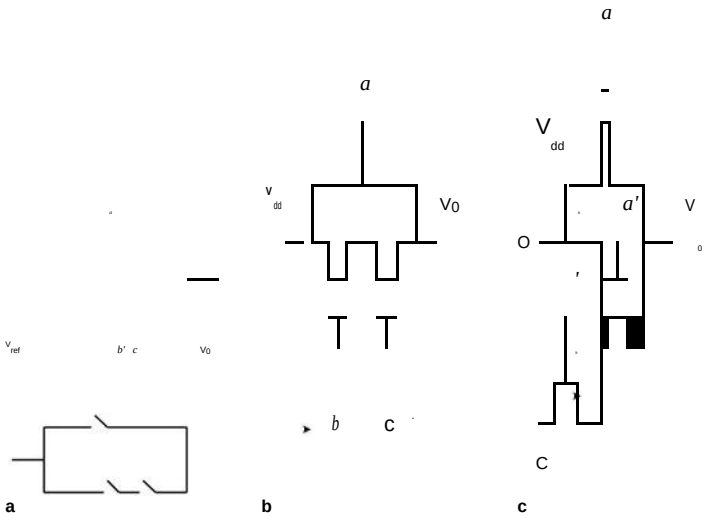


Fig. 5.2 **a** Relay logic to realize $f = a + b'c$. **b** Pass-transistor network corresponding to relay logic. **c** Proper pass-transistor network for $f = a + b\,'\,c$

In the synthesis of relay logic, the network is realized such that high-level signal reaches the output for which the function is “1.” Absence of the high voltage level is considered to be “0.” In pass-transistor logic, however, this does not hold good. For example, to realize the function $f = a + b\,'\,c$ the relay logic network is shown in Fig. 5.2a. The pass-transistor realization based on the same approach is shown in Fig. 5.2b. It can be easily proved that the circuit shown in Fig. 5.2b does not realize the function $f = a + b\,'c$. Let us consider an input combination (say 100) for which the function is “1,” the output will rise to high level by charging the output load capacitance. Now, if we apply an input combination 010, for which the function is “0,” then the output remains at “1” logic level because the output capacitance does not get a discharge path when input combination of 010 is applied. A correct nMOS pass-transistor realization of $f = a + b\,'c$ is shown in Fig. 5.2c. The difference be-tween the circuit of Fig. 5.2c and that of Fig. 5.2b is that it provides a discharge path of the output capacitor when the input combination corresponds to a low-load output. Therefore, in synthesizing pass-transistor logic, it is essential to provide both charging and discharging path for the load capacitance. In other words, charg-ing path has to be provided for all input combinations requiring a high output level and discharge path has to be provided for all input combinations requiring a low output level.

Finally, care should be taken such that advertently or inadvertently an output point is not driven by signals of opposite polarity. In such a situation, each transis-tor acts as a resistor and a voltage about half of the high level is generated at the output. Presence of this type of path is known as *sneak path*. In such a situation, an undefined voltage level is generated at a particular node. This should be avoided.

5.2.1 Realizing Pass-Transistor Logic

Pass transistors can be used to realize multiplexers of different sizes. For example, 2-to-1 and 4-to-1 multiplexer realization using pass transistor is shown in Fig. 5.3a, b, respectively. Now, any two-variable function can be expanded around the vari-ables using Shannon's expansion thereon. The expanded function can be represent-ed by the following expression:

\

$$f(a, b) = g_0\, a'b' + g_1 a'b + g_2\, ab' + g_3\, ab. \quad (5.1)$$

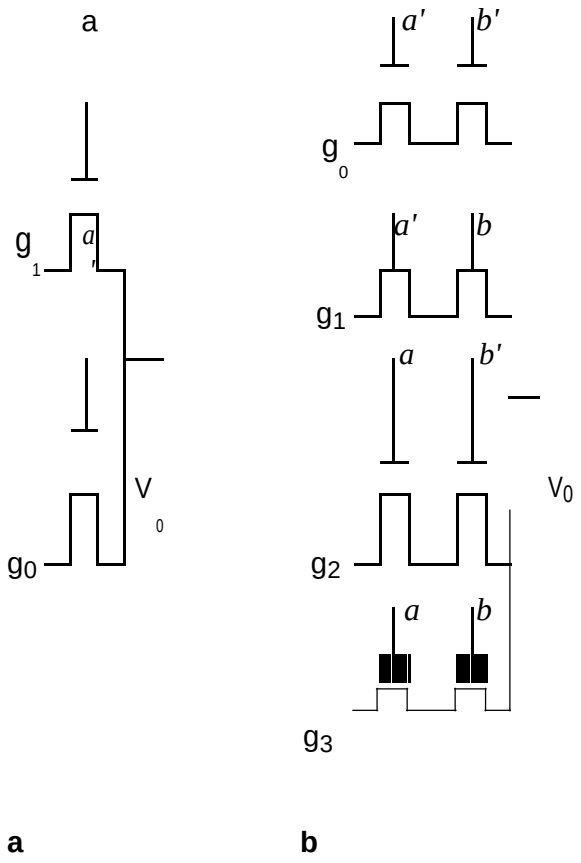
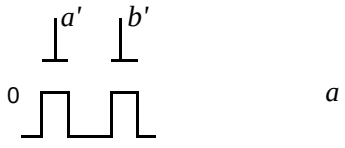


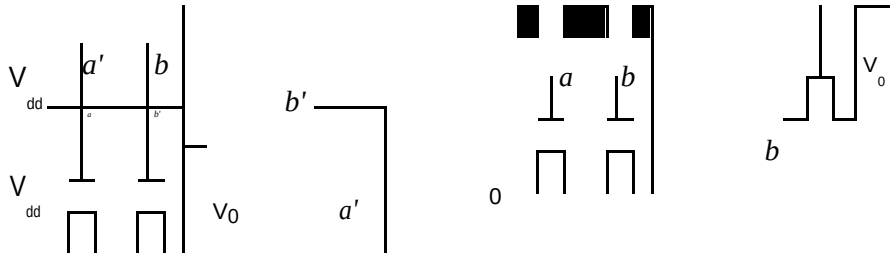
Fig. 5.3 **a** A 2-to-1 multiplexer. **b** A 4-to-1 multiplexer circuit using pass-transistor network

Fig. 5.4 **a** Multiplexer real-ization of $f = a'b + ab'$.

b Minimum transistor pass-transistor realization of

$f = a'b + ab'$





Each g_i can assume a binary value, either 0 or 1, depending on the function. By ap-plying the variables to control the gate potential and applying the logic value g_i to an input a_i the function can be realized. In this way, any function of two variables can be realized.

The realization of the function $f = a'b + ab'$ is shown in Fig. 5.4a. When a func-tion is realized in the above manner, none of the problems mentioned earlier arises. This approach may be termed as universal logic module (ULM)-based approach because of the ability of multiplexers to realize any function up to a certain number of input variables.

However, the circuit realized in the above manner may not be optimal in terms of the number of pass transistors. That means, there may exist another realization with lesser number of pass transistors. For example, optimal realization of $f = a'b + ab'$ is shown in Fig. 5.4b. Optimal network may be obtained by expanding a function iteratively using Shannon's expansion theorem. The variable around which expan-sion is done has to be suitably chosen and each reduced function is further expanded until it is a constant or a single variable.

Example 5.1 Consider the function $f = a + b'c$. Realize it using pass-transistor logic.

By expanding around the variable "a," we get $f = a.1 + a'.b'c$.

Now, by expanding around the variable "b," we get $f = a \square 1 + a'(b \square 0 + b' \square c)$

No further expansion is necessary because there does not exist any reduced function, which is a function of more than one variable. Realization of the function

$f = a + b'c$ based on this approach is shown in Fig. 5.2c. Circuit realization based on the above approach is not only area efficient but also satisfies all the require-ments of pass-transistor logic realization mentioned earlier.

5.2.2 Advantages and Disadvantages

Realization of logic functions using pass transistors has several advantages and disadvantages. The advantages are mentioned below:

a.\ Ratioless: We have seen that for a reliable operation of an inverter (or gate logic), the width/length (W/L) ratio of the pull-up device is four times (or more) that of the pull-down device. As a result, the geometrical dimension of the transistors is not minimum (i.e., $2\lambda \times 2\lambda$). The pass transistors, on the other hand, can be of minimum dimension. This makes pass-transistor circuit realization very area efficient.

b.\ Powerless: In a pass-transistor circuit, there is no direct current (DC) path from supply to ground (GND). So, it does not require any standby power, and power dissipation is very small. Moreover, each additional input requires only a mini-mum geometry transistor and adds no power dissipation to the circuit.

c.\ Lower area: Any one type of pass-transistor networks, nMOS or p-type MOS (pMOS), is sufficient for the logic function realization. This results in a smaller number of transistors and smaller input loads. This, in turn, leads to smaller chip area, lower delay, and smaller power consumption.

However, pass-transistor logic suffers from the following disadvantages:

1.\ When a signal is steered through several stages of pass transistors, the delay can be considerable, as explained below.

When the number of stages is large, a series of pass transistors can be modeled as a network of resistance capacitance (RC) elements shown in Fig. 5.5. The ON

resistance of each pass transistor is R_{pass} and capacitance C_L. The value of R_{pass} and C_L depends on the type of switch used. The time constant $R_{pass}\ C_L$ approximately

gives the time constant corresponding to the time for C_L to charge to 63 % of its final value. To calculate the delay of a cascaded stage of n transistors, we can simplify the equivalent circuit by assuming that all resistances and capacitances are equal and

can be lumped together into one resistor of value $n \times R_{pass}$ and one capacitor of value $n \times C_L$. The equivalent time constant is $n^2\ R_{pass}\ C_L$. Thus, the delay increases as the square of number of pass transistors in series. However, this simplified analysis

leads to overestimation of the delay. A more accurate estimate of the delay can be obtained by approximating the effective time constant to be

$$\tau_{eft} = \sum_{k} R_{k0}\ C_k, \qquad (5.2)$$

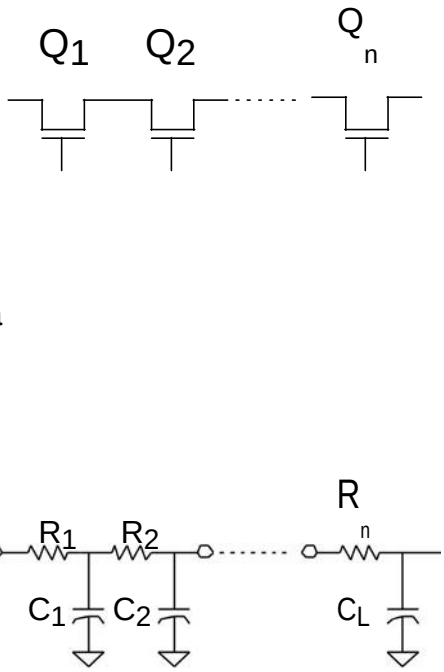


Fig. 5.5 **a** Pass-transistor network. **b** RC model for the pass-transistor network. RC resistance capacitance

where R_k stands for the resistance for the path common to k and input. For $k = 4$,

$$\tau_{\text{eft}} = R_1C_1 + (R_1 + R_2)C_2 + (R_1 + R_2 + R_3)C_3 + (R_1 + R_2 + R_3 + R_n)C_L. \quad (5.3)$$

Time constant $\tau = CR_{\text{eq}} \frac{n(n+1)}{2}$.

Assuming all resistances are equal to R_{pass} and all capacitances are equal to C, the delay of the n-stage pass-transistor network can be estimated by using the

Elmore

approximation:

\

$$t_p = 0.69 \frac{n(n+1)}{2} R_{pass} C. \quad (5.4)$$

For $n = 4$, the delay is

$$t_p = 6.9 R_{pass} C,$$

which are ten times that of single-stage delays.

The above delay estimate implies that the propagation delay is proportional to n^2 and increases rapidly with the number of cascaded stages. This also sets the limit on the number of pass transistors that can be used in a cascade. To overcome this problem, the buffers should be inserted after every three of four pass-transistor stages.

For a large value of n, let buffers are introduced after each k stages, as shown in Fig. 5.6 for $k = 3$. Assuming a propagation delay of each buffer is t_{buf}, the overall propagation delay can be computed as follows:

$$t_p = 0.69 \left[C_L R_{pass} \frac{k(k+1)}{2} \frac{n}{k} + \frac{n}{k} t_{buf} \right] = 0.69 \left[C_L R_{pass} \frac{n(k+1)}{2} + \frac{n}{k} t_{buf} \right].$$

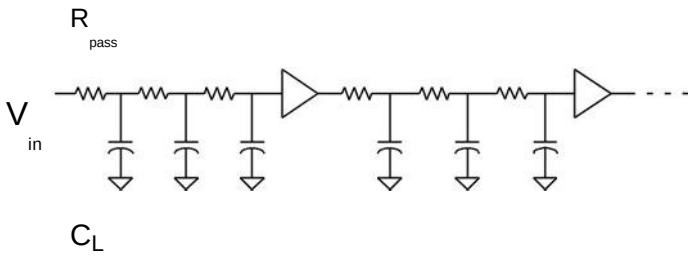


Fig. 5.6 Buffers inserted after every three stages

In contrast to quadratic dependence on n in a circuit without buffer, this expression shows a linear dependence on the number n. An optimal value of k can be obtained from

$$\frac{\delta t_p}{\delta k} = 0 \Rightarrow k_{opt} = 1.7\sqrt{\frac{t_{buf}}{R_{pass} C_L}},$$

for R_{pass} = 10kΩ, C = 10fF, t_{buf} = 0.5n sec, we get k = 3.8 .

Therefore, a cascade of more than four pass-transistor stages is not recommend-ed to restrict the delay within a reasonable limit.

2.\ As we have mentioned earlier, there is a voltage drop ($V_{out} = V_{dd} - V_{tn}$) as we steer the signal through nMOS transistors. This reduced level leads to high static currents at the subsequent output inverters and logic gates. In order to avoid this, it is necessary to use additional hardware known as the *swing restoration logic* at the gate output.

3.\ Pass-transistor structure requires complementary control signals. Dual-rail logic is usually necessary to provide all signals in the complementary form. As a con-sequence, two MOS networks are again required in addition to the swing restora-tion and output buffering circuitry.

The required double inter-cell wiring increases wiring complexity and capacitance by a considerable amount.

5.2.3 Pass-Transistor Logic Families

Several pass-transistor logic styles have seen proposed in recent years to overcome the limitations mentioned above. The most common ones are: the conventional nMOS pass-transistor logic or complementary pass-transistor logic (CPL), the dual pass-transistor logic (DPL), and the swing-restored pass-transistor logic (SRPL) [2]. In this section, they are introduced and compared.

Complementary Pass-Transistor Logic A basic structure of a CPL gate is shown in Fig. 5.7a. It comprises two nMOS pass-transistor logic network (one for each rail), two small pull-up pMOS transistors for swing restoration, and two output inverters for the complementary output signals. CPL realization for the 2-to-1

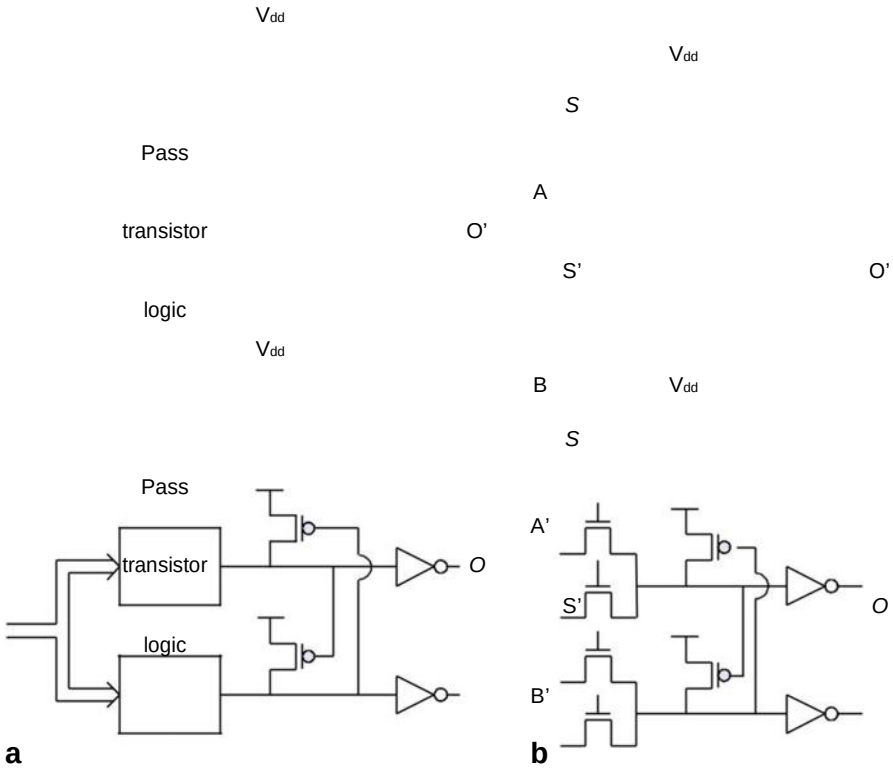
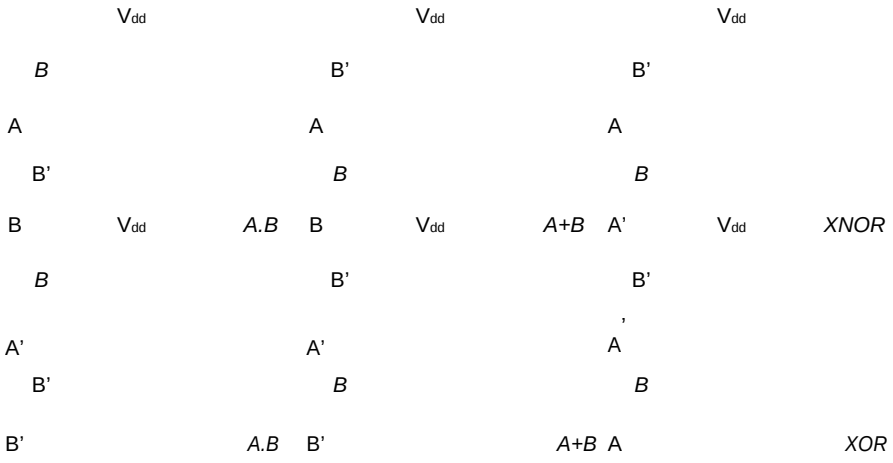


Fig. 5.7 **a** Basic complementary pass-transistor logic (CPL) structure; and **b** 2-to-1 multiplexer realization using CPL logic



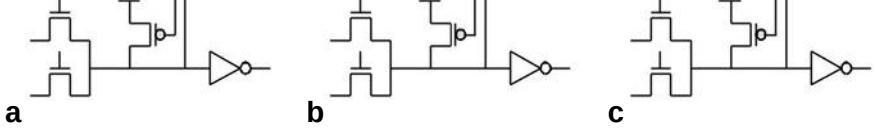


Fig. 5.8 Complementary pass-transistor logic (CPL) logic circuit for **a** 2-input AND/NAND, **b** 2-input OR/NOR, and **c** 2-input EX-OR

multiplexer is shown in Fig. 5.7b. This is the basic and minimal gate structure with ten transistors. All two-input functions can be implemented by this basic gate struc-ture. Realizations of 2-input NAND, NOR, and EX-OR functions are shown in Fig. 5.8a–c, respectively.

Swing-Restored Pass-Transistor Logic The SRPL logic style is an extension of CPL to make it suitable for low-power low-voltage applications. The basic SRPL logic gate is shown in Fig. 5.9a, where the output inverters are cross-coupled with a latch structure, which performs both swing restoration and output buffering. The pull-up pMOS transistors are not required anymore and that the output nodes of the nMOS networks are the gate outputs. Figure 5.9b shows an example of AND/ NAND gate using SRPL logic style. As the inverters have to drive the outputs and must also be overridden by the nMOS network, transistor sizing becomes very dif-ficult and results in poor output-driving capacity, slow switching, and larger short-circuit currents. This problem becomes worse when many gates are cascaded.

Double Pass-Transistor Logic The DPL is a modified version of CPL, in which both nMOS and pMOS logic networks are used together to alleviate the problem of the CPL associated with reduced high logic level. As this provides full swing on

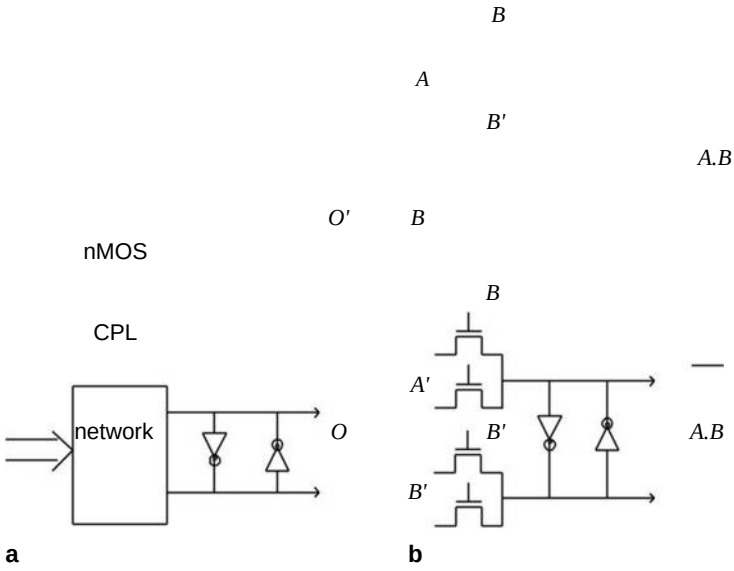
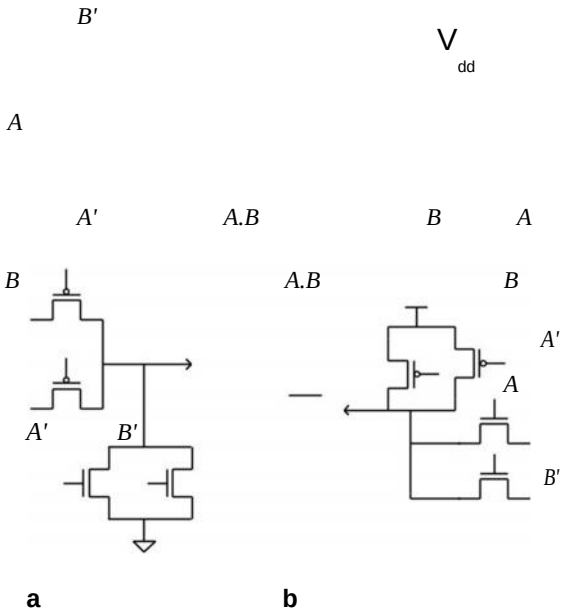


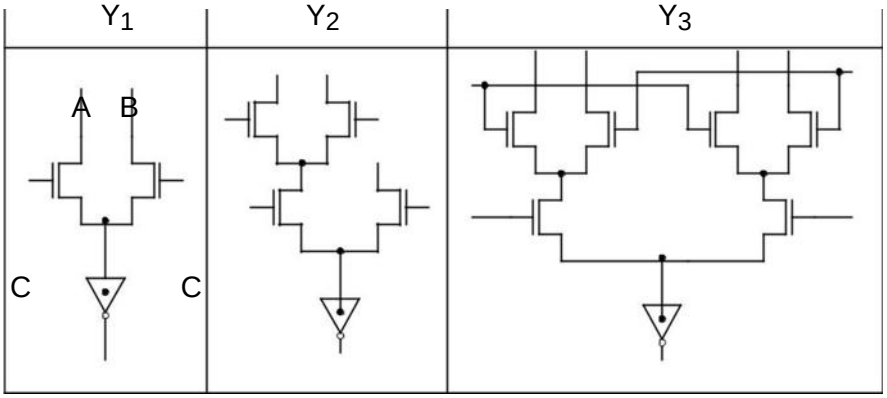
Fig. 5.9 **a** Basic swing-restored pass-transistor logic (SRPL) configuration; and **b** SRPL realization of 2-input NAND gate

Fig. 5.10 Double pass-transistor logic (DPL) realization of 2-input AND/NAND function



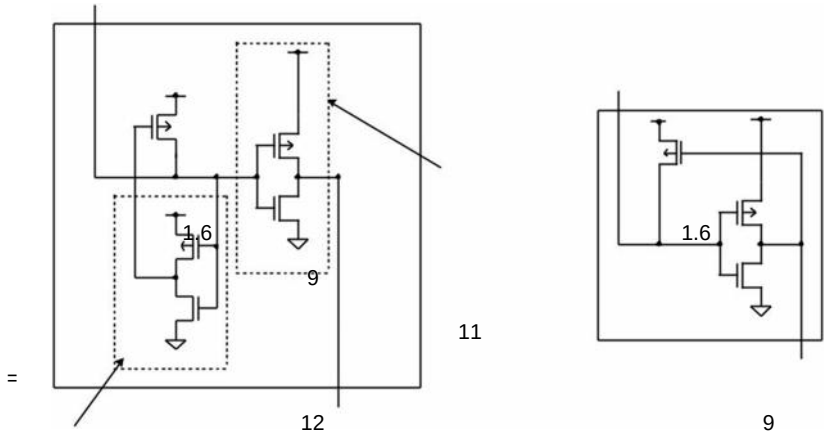
the output, no extra transistors for swing restoration is necessary. Two-input AND/ NAND DPL realization is shown in Fig. 5.10. As shown in the figure, both nMOS and pMOS pass transistors are used. Although, owing to the addition of pMOS transistor, the output of the DPL is full rail-to-rail swing, it results in increased input capacitance compared to CPL. It may be noted that DPL can be regarded as dual-rail pass-gate logic. The DPL has a balanced input capacitance, and the delay is independent of the input delay contrary to the CPL and conventional CMOS pass-transistor logic, where the input capacitance for different signal inputs is the same. As two current paths always drive the output in DPL, the reduction in speed due to the additional transistor is compensated.

Single-Rail Pass-Transistor Logic A single-rail pass-transistor logic style has been adopted in the single-rail pass-transistor logic (LEAP; LEAn integration with pass transistors) [7] design which exploits the full functionality of the multiplexer structure scheme. Swing restoration is done by a feedback pull-up pMOS transistor. This is slower than the cross-coupled pMOS transistors of CPL working in dif-ferential mode. This swing restoration scheme works for $V_{dd} > V_{tn} + V_{tp}$, because the threshold voltage drop through the nMOS network for a logic “1” prevents the nMOS of the inverter and with that the pull-up pMOS from turning ON. As a conse-quence, robustness at low voltages is guaranteed only if the threshold voltages are appropriately selected. Complementary inputs are generated locally by inverters.



out out out

a



1

b 12

Fig. 5.11 Single-rail pass-transistor logic (LEAP) cells

Three basic cells; Y_1, Y_2, and Y_3, used in LEAP logic design is shown in Fig. 5.11. It may be noted that Y_1 cell corresponds to a 2-to-1 multiplexer circuit, whereas Y_3 corresponds to 4-to-1 multiplexer circuit realized using three 2-to-1 multiplexers. Y_2 is essentially a 3-to-1 multiplexer realized using two 2-to-1 multiplexers. Any complex logic circuit can be realized using these three basic cells.

Comparison of the pass-transistor logic styles based on some important circuit parameters, such as number of MOS transistors, the output driving capabilities, the presence of input/output decoupling, requirement for swing restoration circuits, the number of rails, and the robustness with respect to voltage scaling and transistor sizing, is given in Table 5.1.

Table 5.1 Qualitative comparisons of the logic styles

Logic style	#MOS networks	Output driving	I/O decoupling	Swing restoration	# Rails	Robustness
CMOS	2*n*	Med/good	Yes	No	Single	High
CPL	2*n* + 6	Good	Yes	Yes	Dual	Medium
SRPL	2*n* + 4	Poor	No	Yes	Dual	Low
DPL	4*n*	Good	Yes	No	Dual	High
LEAP	*n* + 3	Good	Yes	Yes	Single	Medium
DCVSPG	2*n* + 2	Medium	Yes	No	Dual	Medium

MOS metal–oxide–semiconductor, *I/O* input/output, *CMOS* complementary metal–oxide–semiconductor, *CPL* complementary pass-transistor logic, *SRPL* swing-restored pass-transistor logic, *DPL* double pass-transistor logic, *LEAP* single-rail pass-transistor logic, *DCVSPG* differential cas-code voltage switch pass gate

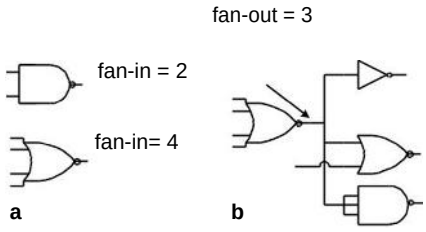


Fig. 5.12 **a** Fan-in of gates; and **b** fan-out of gates

5.3 Gate Logic

In gate logic, conventional gates, such as inverters, NAND gates, NOR gates, etc., are used for the realization of logic functions. In Chap. 4, we have already discussed the possible realization of inverter circuits by using both nMOS and CMOS tech-nology. Realization of NAND, NOR, and other types of gates are considered in this section.

A basic inverter circuit can be extended to incorporate multiple inputs leading to the realization of standard logic gates such as NAND and NOR gates [4]. These gates along with the inverters can be used to realize any complex Boolean function. In this section, we shall consider various issues involved in the realization of these multi-input gates. But, before that we introduce the concept of fan-in and fan-out, which are important parameters in the realization of complex functions using these gates.

5.3.1 Fan-In and Fan-Out

Fan-in is the number of signal inputs that the gate processes to generate some out-put. Figure 5.12a shows a 2-input NAND gate and a four-input NOR gate with fan-in of 2 and 4, respectively. On the other hand, the fan-out is the number of logic inputs driven by a gate. For example, in Fig. 5.12b a four-input NOR gate is shown with fan-out of 3.

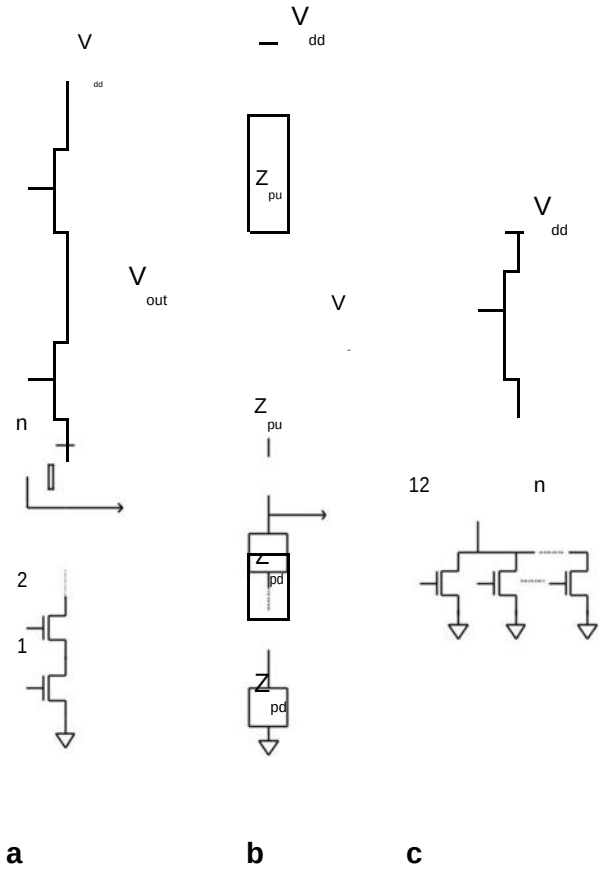


Fig. 5.13 **a** *n*-input nMOS NAND gate; **b** equivalent circuits; and **c** n-input nMOS NOR gate. nMOS n-type MOS

5.3.2 nMOS NAND and NOR Gates

Realizations of nMOS NAND gate with two or more inputs are possible. Let us consider the generalized realization with *n* inputs as shown in Fig. 5.13 with a de-pletion-type nMOS transistor as a pull-up device and *n* enhancement-type nMOS transistors as pull-down devices. In this kind of realization, the length/width (L/W) ratio of the pull-up and pull-down transistors should be carefully chosen such that the desired logic levels are maintained. The critical factor here is the low-level output voltage, which should be sufficiently low such that it turns off the transistors of the following stages. To satisfy this, the output voltage should be less than the

threshold voltage, i.e., $V_{out} \leq V_t = 0.2\ V_{dd}$. To get a low-level output, all the pull-down transistors must be ON to provide the GND path. Based on the equivalent circuit shown in Fig. 5.13b, we get the following relationship:

\

$$\frac{V_{dd} \wedge nZ_{pd}}{nZ_{pd} + Z_{pu}} 0.2V_{dd}\ , \qquad (5.5)$$

for the boundary condition

$$\frac{nZ_{pd}}{nZ_{pd} + Z_{pu}} 0.2$$

$$\text{or } nZ_{pd}\ 0.2nZ_{pd} + 0.2Z_{pu}$$

$$\text{or } \frac{Z_{pu}}{nZ_{pd}} 4.$$

Therefore, the ratio of Z_{pu} to the sum of all the pull-down Z_{pd}s must be at least 4:1. It may be noted that, not only one pull-down transistor is required per input of the NAND gate stage but also the size of the pull-up transistor has to be adjusted

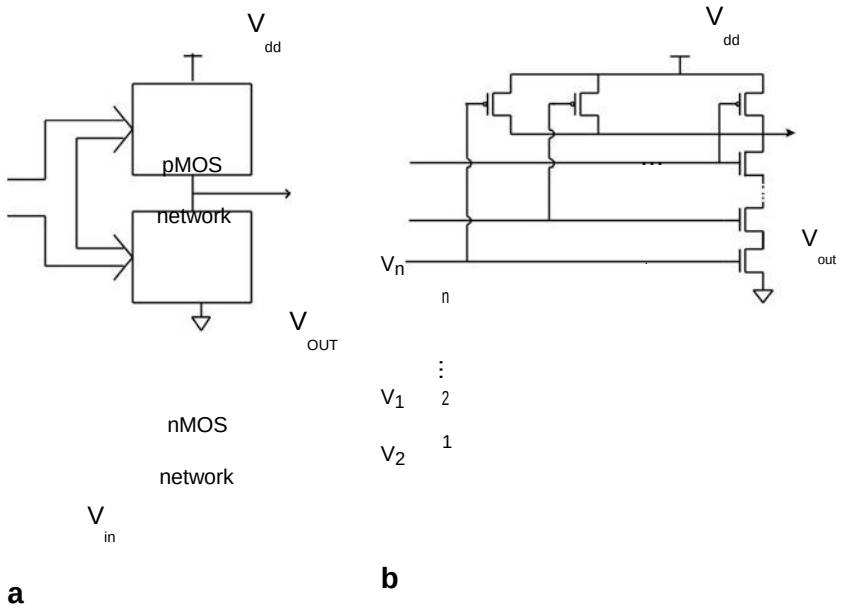


Fig. 5.14 **a** General CMOS network; and **b** *n*-input CMOS NAND gate. CMOS complementary MOS, p-type MOS, n-type MOS

to maintain the required overall ratio. This requires a considerably larger area than those of the corresponding nMOS inverter.

Moreover, the delay increases in direct proportion to the number of inputs. If each pull-down transistor is kept of minimum size, then each will represent one gate capacitance at its input and resistance of all the pull-down transistors will be in series. Therefore, for an *n*-input NAND gates, we have a delay of *n*-times that of an inverter, i.e.,

$$\tau_{NAND} = n\tau_{inv}.$$

As a consequence, nMOS NAND gates are only used when absolutely necessary and the number of inputs is restricted so that area and delay remain within some limit.

An *n*-input NOR gate is shown in Fig. 5.13c. Unlike the NAND gate, here the output is low when any one of the pull-down transistors is on, as it happens in the case of an inverter. As a consequence, the aspect ratio of the pull-up to any pull-down transistor will be the same as that of an inverter, irrespective of the number of inputs of the NOR gate.

The area occupied by the nMOS NOR gate is reasonable, because the pull-up transistor geometry is not affected by the number of inputs. The worst-case delay of an NOR gate is also comparable to the corresponding inverter. As a consequence, the use of NOR gate in circuit realization is preferred compared to that of NAND gate, when there is a choice.

5.3.3 CMOS Realization

The structure of a CMOS network to realize any arbitrary Boolean function is shown in Fig. 5.14a. Here, instead of using a single pull-up transistor, a network of pMOS transistors is used as pull up. The pull-down circuit is a network of nMOS transistors, as it is done in the case of an nMOS realization.

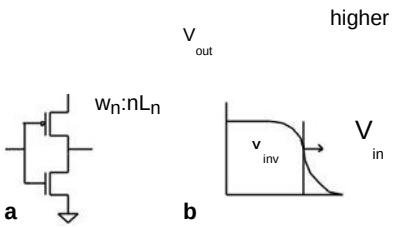


Fig. 5.15 **a** Equivalent circuit of *n*-input complementary MOS (CMOS) NAND gate; and **b** transfer characteristics of *n*-input CMOS NAND gate

5.3.3.1 CMOS NAND Gates

Let us consider an *n*-input NAND gate as shown in Fig. 5.14b. It has fan-in of *n* having *n* number of nMOS transistor in series in the pull-down path and *n* number of pMOS transistors in parallel in the pull-up network. We assume that all the tran-sistors are of the same size having a width $W = W_n = W_p$ and length $L = L_n = L_p$. If all inputs are tied together, it behaves like an inverter. However, static and dynamic characteristics are different from that of the standard inverter. To determine the inversion point, pMOS transistors in parallel may be equated to a single transistor with the width *n* times that of a single transistor. And nMOS transistors in series may be considered to be equivalent to have a length equal to *n* times that of a single transistor as shown in Fig. 5.15a. This makes the transconductance ratio = $\beta_n / n^2\beta_p$.

Substituting these values in Eq. 4.17, we get

\

$$V_{inv} = \frac{V_{dd} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{n^2\beta_p}}}{1 + \sqrt{\frac{\beta_n}{n^2\beta_p}}} \quad (5.6)$$

$$n^2\beta_p$$

or

$$V_{inv} = \frac{V_{dd} + V_{tp}\sqrt{\frac{\beta_n}{n^2\beta_p}} + V_{th}}{1 + \sqrt{\frac{\beta_n}{n^2\beta_p}}} \quad (5.7)$$

As the fan-in number n increases, the ratio (transconductance ratio) $\frac{\beta_n}{\beta_p}$ decreases

leading to increase in V_{inv}. In other words, with the increase in fan-in, the switching point (inversion point) moves towards the right. It may be noted that in our analysis, we have ignored the body effect.

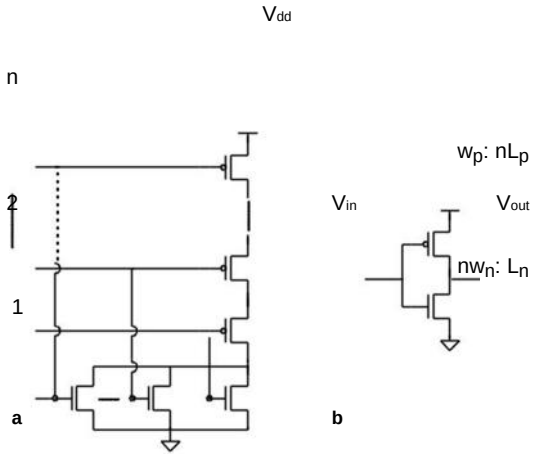


Fig. 5.16 **a** *n*-input complementary MOS (CMOS) NOR gate and **b** the equivalent circuit

5.3.3.2 CMOS NOR Gates

The realization of the *n*-input CMOS NOR gate is shown in Fig. 5.16a. Its equiva-lent circuit is shown in Fig. 5.16b. In this case, the transconductance ratio is equal

to $\frac{n_2 \beta_n}{\beta_p}$.

Substituting this ratio in Eq. 5.6, we get

\

$$V_{inv} = \frac{V_{dd} + V_{tp} + nV_{th}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + n\sqrt{\frac{\beta_n}{\beta_p}}}. \quad (5.8)$$

$\sqrt{\beta_p}$

As the fan-in number n increases, the $\frac{\beta_n}{}$ ratio (transconductance ratio) increases β_p

leading to a decrease in V_{inv}. In other words, with the increase in fan-in, the switch-ing point (inversion point) moves towards the left. Here, also, we have ignored

the body effect. Therefore, with the increase in fan-in the inversion point moves towards left for NOR gates, whereas it moves towards right in case of NAND gates.

5.3.4 Switching Characteristics

To study the switching characteristics, let us consider the series and parallel tran-sistors separately. Figure 5.17a shows n pull-up pMOS transistors with their gates tied together along with a load capacitance C_L. An equivalent circuit is shown in Fig. 5.17b. Intrinsic time constant of this network is given by

$$t_{dr} = \frac{R_p}{n}(nC_{outp}) + \frac{R_p}{n} \tag{5.9}$$

$$C_L = \frac{R_p}{n}(nC_{outp} + C_L).$$

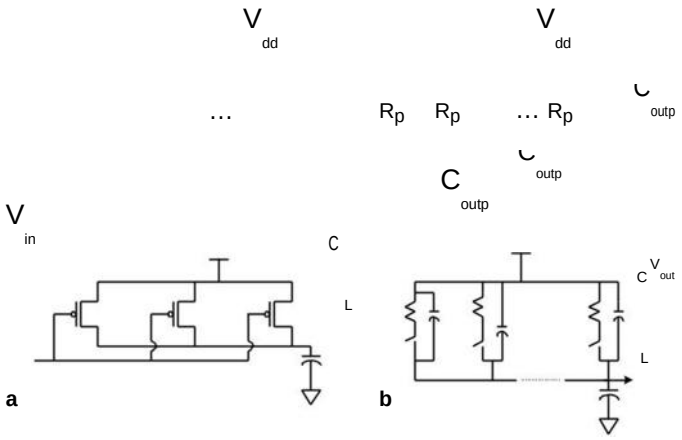


Fig. 5.17 **a** Pull-up transistor tied together with a load capacitance; and **b** equivalent circuit

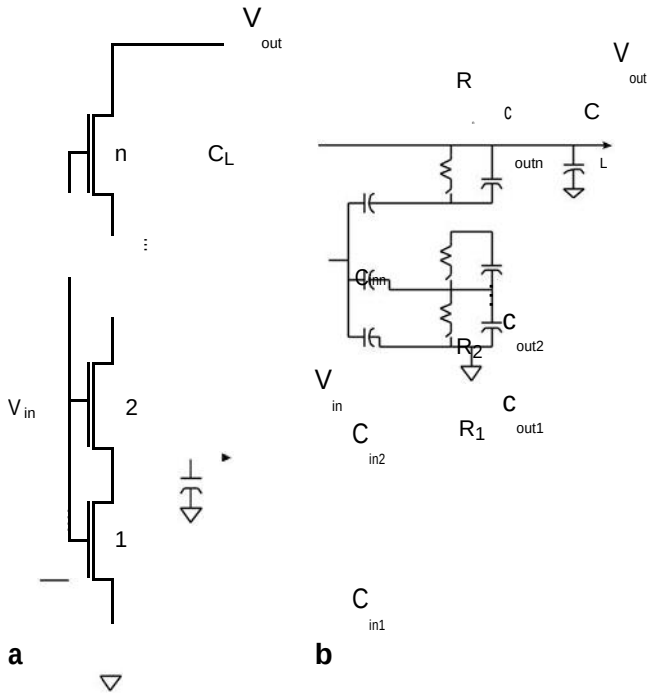


Fig. 5.18 **a** Pull-down transistors along with load capacitance C_L, and **b** equivalent circuit

Here, the load capacitance C_L includes all capacitances on the output node except the output capacitances of the transistors in parallel. The worst-case rise time occurs when only one pMOS transistor is ON and remaining pMOS transistors are OFF. This rise time is close to the rise time of an inverter.

To find out the fall-time (high-to-low time), let us consider the n numbers of series-connected-nMOS transistors as shown in Fig. 5.18a. The intrinsic switching time for this network based on Fig. 5.18b is given by

\

$$t_{df} = nR_n \left(C_{load} + 0.35 R_n C_{inn} (n-1) + \frac{C_{outn}}{n} \right)^2 \quad (5.10)$$

The first term represents the intrinsic switching time of the n-series-connected tran-sistors and the second term represents the delay caused by R_n charging C_{inn}. For an n-input NAND gate

\

$$t_{dr} = \frac{R_p}{n} \left(nC_{outp} + \frac{C_{outn}}{n} + C_{load} \right) \quad (5.11)$$

and

\

$$t_{df} = nR_n \left(\frac{C_{outn}}{n} + nC_{outp} + C_L + 0.35 R_n C_{inn} (n-1) \right)^2 . \quad (5.12)$$

Thus, the delay increases linearly with the increase of fan-in number n.

5.3.5 CMOS NOR Gate

In a similar manner, the rise and fall times of an *n*-input NOR gate can be obtained as

\

$$t_{df} = \frac{R_n}{n}(nC_{outp} + C_L) \tag{5.13}$$

and

\

$$t_{dr} = nR_p\frac{C_{outp}}{n} + nC_{outn} + C_L + 0.35R_p C_{inn}(n-1)^2 . \tag{5.14}$$

It may be noted that in case of NAND gate, the discharge path is through the series– connected-nMOS transistors. As a consequence, the high-to-low delay increases with the number of fan-in. If the output load capacitance is considerably larger than other capacitances then the fall time reduces to $t_{df} = nR_n C_L$ and $t_{dr} = R_p C_L$. On the other hand, in case of NOR gate the charging of the load capacitance take place through the series connected pMOS transistors giving rise time $t_{dr} = nR_p C_L$ and $t_{df} = R_n C_L$. As $R_p > R_n$, the rise-time delay for NOR gates is more than the fall-time delay of NAND gates with the same fan-in. This is the reason why NAND gates are generally a better choice than NOR gates in complementary CMOS logic. NOR gates may be used for limited fan-out.

5.3.6 CMOS Complex Logic Gates

By combining MOS transistors in series–parallel fashion, basic CMOS NAND/NOR gates can be extended to realize complementary CMOS complex logic gates. Basic concept of realizing a complex function f by a single gate is shown in Fig. 5.19a. Here, the nMOS network corresponds to the function f' and the pMOS network is complementary of the nMOS network. For example, the realization of the function

$f = A' + BC$ is shown in Fig. 5.19b.

Here, $f = A' + BC \Rightarrow f' = (A' + BC)' = A(B' + C')$.

The pull-down nMOS network realizes $A(B' + C')$, whereas pull-up pMOS net-work realizes $A' + BC$. In this realization, both nMOS and pMOS network are, in general, series–parallel combination MOS transistors as shown in Fig. 5.19b. Full-complementary CMOS realization of half-adder function is shown in Fig. 5.19c.

Here, $S = A \oplus B = A'B + AB' \Rightarrow S' = (A'B + AB')' = (A + B')(A' + B)$.

In realizing complex functions using full-complementary CMOS logic, it is nec-essary to limit the number of MOS transistors in both the pull-up and pull-down network, so that the delay remains within the acceptable limit. Typically, the limit is in the range of 4–6 MOS transistors.

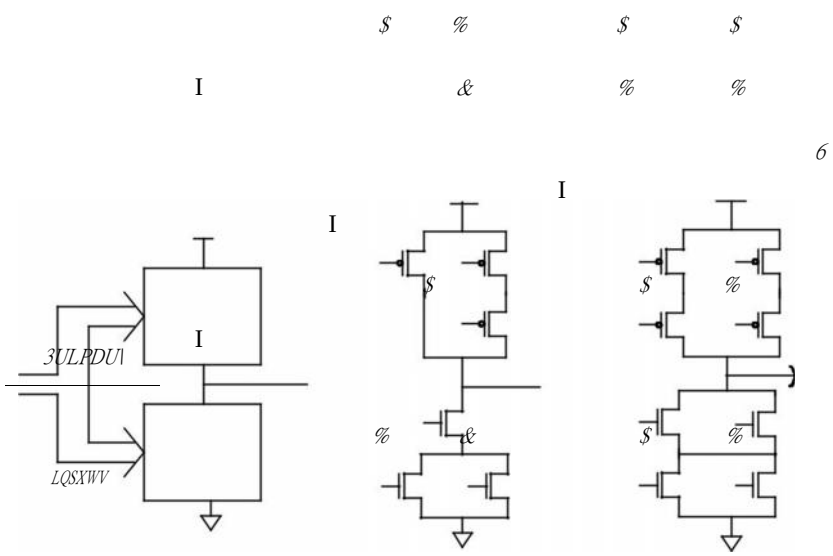


Fig. 5.19 **a** Realization of a function *f* by complementary MOS (CMOS) gate; **b** realization of

$f = A' + BC$; and **c** realization of $S = A \oplus B$

5.4 MOS Dynamic Circuits

The MOS circuits that we have discussed so far are static in nature [2]. In static circuits, the output voltage levels remain unchanged as long as inputs are kept the same and the power supply is maintained. nMOS static circuits have two disad-vantages: they draw static current as long as power remains ON, and they require larger chip area because of “ratioed” logic. These two factors contribute towards slow operation of nMOS circuits. Although there is no static power dissipation in a full-complementary CMOS circuit, the logic function is implemented twice, one in the pull-up p-network and the other in the pull-down n-network. Due to the extra area and extra number of transistors, the load capacitance on gates of a full-complementary CMOS is considerably higher. As a consequence, speeds of operation of the CMOS and nMOS circuits are comparable. The CMOS not only has twice the available current drive but also has twice the capacitance of nMOS. The trade-off in choosing one or the other is

between the lower power of the CMOS and the lower area of nMOS (or pseudo nMOS).

In the static combinational circuits, capacitance is regarded as a parameter re-sponsible for poor performance of the circuit, and therefore considered as an un-desirable circuit element. But, a capacitance has the important property of holding charge. Information can be stored in the form of charge. This information storage capability can be utilized to realize digital circuits. In MOS circuits, the capaci-tances need not be externally connected. Excellent insulating properties of silicon dioxide provide very good quality gate-to-channel capacitances, which can be used for information storage. This is the basis of MOS dynamic circuits. In Sect. 5.4.1, we discuss how dynamic MOS circuits are realized utilizing these capacitances us-ing single-phase and two-phase clocks.

We shall see that the advantage of low power of full-complementary CMOS circuits and smaller chip area of nMOS circuits are combined in dynamic circuits leading to circuits of smaller area and lower power dissipation. MOS dynamic circuits

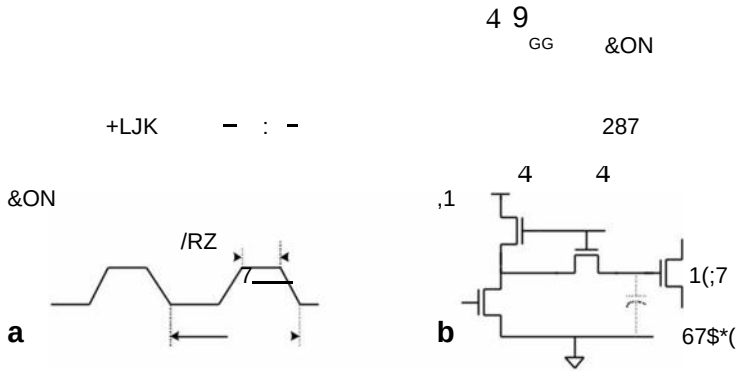


Fig. 5.20 **a** Single-phase clock; and **b** single-phase n-type MOS (nMOS) inverter

are also faster in speed. However, these are not free from disadvantages. Like any other capacitors, charge stored on MOS capacitors also leak. To retain information, it is necessary to periodically restore information by a process known as *refreshing*. There are other problems like *charge sharing* and *clock skew* leading to hazards and races. Suitable measures should be taken to overcome these problems. These problems are considered in Sect. 5.4.3.

5.4.1 Single-Phase Dynamic Circuits

To realize dynamic circuits, it is essential to use a clock. Two types of clocks are commonly used; *single-phase* clock and nonoverlapping *two-phase* clock. The single-phase clock consists of a sequence of pulses having high (logic 1) and low (logic 0) levels with width W and time period T as shown in Fig. 5.20a. A single-phase clock has two states (low and high) and two edges per period. The schematic diagram of a single-phase dynamic nMOS inverter is shown in Fig. 5.20b. Opera-tion of a single-phase inverter circuit is explained below.

When the clock is in the high state, both transistors Q_2 and Q_3 are ON. Depend-ing on the input, Q_1 is either ON or OFF. If the input voltage is low, Q_1 is OFF and the output capacitor (gate capacitor of the next stage) charges to V_{dd} through Q_2 and Q_3. When the input voltage is high, Q_1 is ON and the output is discharged through it to a low level. When the clock is in the low state, the transistors Q_2 and Q_3 are OFF, isolating the output capacitor. This voltage is maintained during the OFF period of the clock, provided the period is not too long. During this period, the power supply is also disconnected and no current flows through the circuit. As current flows only when the clock is high, the power consumption is small, and it depends on the duty cycle (ratio of high time to the time period *T*). It may be noted that the output of the circuit is also ratioed, because the low output voltage depends on the ratio of the ON resistance of Q_1 to that of Q_2. As we know that, this ratio is related to the physical dimensions of Q_1 to Q_2 (L:W ratio) and is often referred to as the *inverter ratio*. The idea of the MOS inverter can be extended to realize dynamic MOS NAND, NOR, and other gates as shown in Fig. 5.21.

The circuits realized using the single-phase clocking scheme has the disadvan-tage that the output voltage level is dependent on the inverter ratio and the number of transistors in the current path to GND. In other words, single-phase dynamic

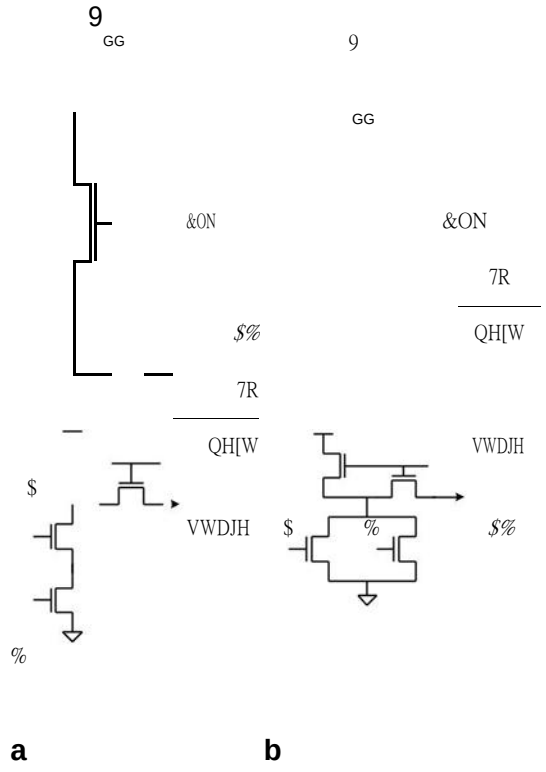


Fig. 5.21 **a** 2-input single-phase NAND; and **b** 2-input single-phase NOR gate

circuits are ratioed logic. Moreover, as we have mentioned above, the circuit dis-sipates power when the output is low and the clock is high.

Another problem arising out of single-phase clocked logic is known as *clock skew* problem. This is due to a delay in a clock signal during its journey through a number of circuit stages. This results in undesired signals like glitch, hazards, etc. Some of the problems can be overcome using two-phase clocking scheme as dis-cussed in the following subsection.

5.4.2 Two-Phase Dynamic Circuits

A two-phase nonoverlapping clock is shown in Fig. 5.22a. As the two phases (φ_1 and φ_2) are never high simultaneously, the clock has three states and four edges and satisfies the property $\varphi_1.\varphi'_2 = 0$. There is a dead time, Δ, between transitions of the clock signals as shown in Fig. 5.22a. The schematic diagram of a circuit that generates two-phase clock is shown in Fig. 5.22b. The circuit takes a single-phase clock as an input and generates two-phase clock as the output. The dead time, Δ, is decided by the delay through the NAND gates and the two inverters. As the clock (clk) signal goes low, φ_1 also goes low after four gate delays. φ_2 can go high after seven gate delays. In a similar manner, as clk goes high, φ_2 goes low after five gate delays and φ_1 goes high after eight gate delays. So φ_2 (φ_1) goes high after three gate delays and φ_1 (φ_2) goes low. Here the dead time is three gate delays. A longer dead time can be obtained by inserting more number of inverters in the feedback path.

This two-phase clocking gives a great deal of freedom in designing circuits. An inverter based on two-phase clock generator is shown in Fig. 5.23. When the clock Φ_2 is high, the intrinsic capacitor charges to V_{dd} through Q_1. And clock Φ_1, which comes after Φ_2 performs the evaluation. If V_{in} is high, Q_2 is turned ON and since Q_3 is ON, the capacitor discharges to the GND level and the output V_0 attains low logic level. If V_{in} is low, the Q_2 is OFF and there is no path for the capacitor to discharge. Therefore, the output V_0 remains at high logic level. It may be noted that the pull-up and pull-down transistors are never simultaneously ON. The circuit has no DC current path regardless of the state of the clocks or the information stored on the parasitic capacitors. Moreover, the output is not ratioed, i.e., the low-level output is independent of the relative value of the aspect ratio of the transistors. That

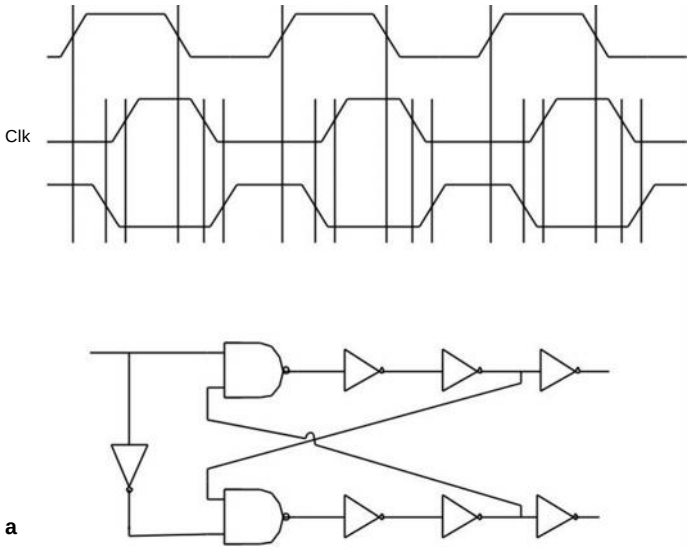
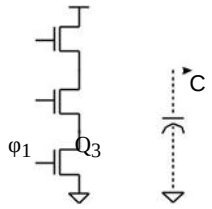


Fig. 5.22 **a** Two-phase clock; and **b** a two-phase clock generator circuit

Fig. 5.23 Two-phase n-type MOS (nMOS) inverter



is why the circuits based on two-phase clocking are often termed as *ratioless* and *powerless*. Moreover, in dynamic circuits, there can be at the most one transition per clock cycles, and therefore there cannot be multiple spurious transitions called *glitches*, which can take place in static CMOS circuits. Like inverter, NAND, NOR, and other complex functions can be realized by replacing Q_2 with a network of two or more nMOS transistors.

5.4.3 CMOS Dynamic Circuits

Dynamic CMOS circuits avoid area penalty of static CMOS and at the same time retains the low- power dissipation of static CMOS, and they can be con-sidered as extension of pseudo-nMOS circuits. The function of these circuits can be better explained using the idea of *pre-charged* logic. Irrespective of

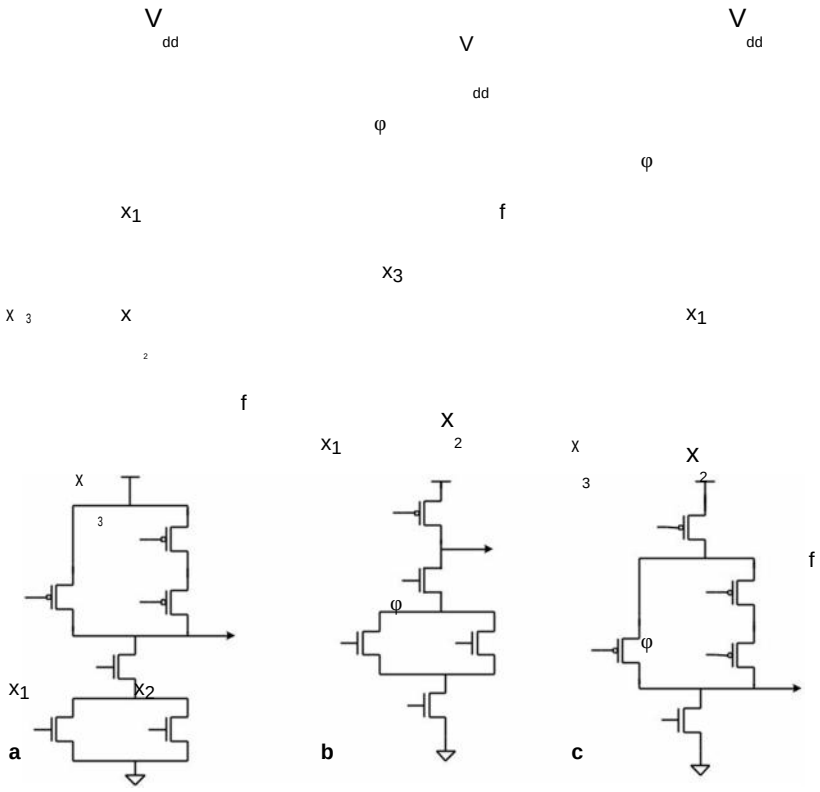


Fig. 5.24 Realization of function $f = x_3 (x_1 + x_2)$ using **a** static complementary MOS (CMOS), **b** dynamic CMOS with n-block, and **c** dynamic CMOS with p-block

several alternatives available in dynamic circuits, in all cases the output node is pre-charged to a particular level, while the current path to the other level is turned OFF. The charging of inputs to the gate must take place during this phase. At the end of this *pre-charge phase*, the path between the output node and the pre-charge source is turned OFF by a clock, while the path to the other level is turned ON through a network of MOS transistors, which represents this func-tion. During this phase, known as *evaluation phase*, the inputs remain stable and depending on the state of the inputs, one of the two possibilities occurs. The network of transistors either connects the output to the other level discharging the output or no path is established between the other level and output, thereby retaining the charge. This operation is explained with the help of a circuit re-alization

for the function $f = x_3 + x_1\, x_2$. The realization of this function using full-complementary CMOS approach is shown in Fig. 5.24a. There are two alternative approaches for realizing it using dynamic CMOS circuits. As shown in Fig. 5.23b, the circuit is realized using the nMOS block scheme, where there is only one pMOS transistor, which is used for pre-charging. During $\varphi = 0$, the out-put is pre-charged to a high level through the pMOS transistor. And during the evaluation phase ($\varphi = 1$), the output is evaluated through the network of nMOS transistors. At the end of the evaluation phase, the output level depends on the input states of the nMOS transistors. In the second approach, the circuit uses the pMOS block scheme, where there is only one nMOS transistor used for pre-discharging the output as shown in Fig. 5.24c. Here, during pre -charge phase, the output is pre-discharged to a low level through the nMOS transistor and during evaluation phase the output is evaluated through the network of pMOS transistors. In the evaluation phase, the low-level output is either retained or charged to a high level through the pMOS transistor network.

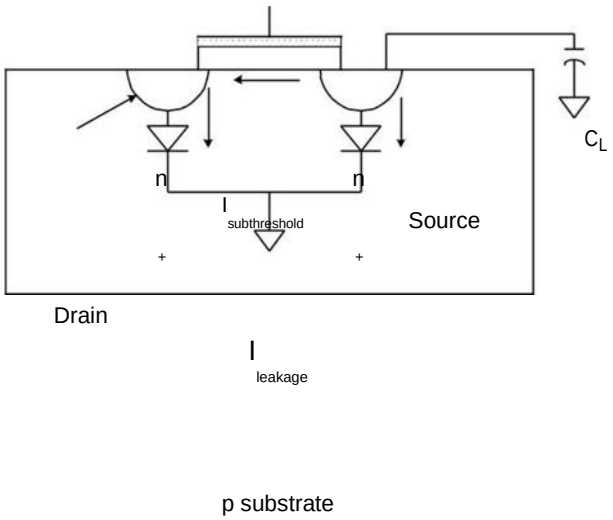


Fig. 5.25 Reverse-biased parasitic diode and subthreshold leakage

5.4.4 Advantages and Disadvantages

The dynamic CMOS circuits have a number of advantages. The number of transis-tors required for a circuit with fan-in N is $(N + 2)$, in contract to 2 N in case of state CMOS circuit. Not only dynamic circuits require $(N + 2)$ MOS transistors but also the load capacitance is substantially lower than that for static CMOS circuits. This is about 50 % less than static CMOS and is closer to that of nMOS (or pseudo nMOS) circuits. But, here full pull-down (or pull-up) current is available for discharging (or charging) the output capacitance.

Therefore, the speed of the operation is faster than that of the static CMOS cir-cuits. Moreover, dynamic circuits consume static power closer to the static CMOS. Therefore, dynamic circuits provide superior (area-speed product) performance compared to its static counterpart. For example, a dynamic NOR gate is about five times faster than the static CMOS NOR gate. The speed advantage is due to smaller output capacitance and reduced overlap current. However, dynamic circuits are not free from problems, as discussed in the following subsections.

5.4.4.1 Charge Leakage Problem

However, the operation of a dynamic gate depends on the storage of information in the form of charge on the MOS capacitors. The source–drain diffusions form para-sitic diodes with the substrate. If the source (or drain) is connected to a capacitor C_L with some charge on it, the charge will be slowly dissipated through the reverse-biased parasitic diode. Figure 5.25 shows the model of an nMOS transistor. The leakage current is expressed by the diode equation:

$$i_D = I_s (e^{-qV/kT} - 1),$$

where I_s is the reverse saturation current, V is the diode voltage, q is the electronic charge (1.602×10^{-19} °C), k is the Boltzman in constant (1.38×10^{-23} J/K), and T is the temperature in Kelvin.

It may be noted that the leakage current is a function of the temperature. The current is in the range 0.1–0.5 nA per device at room temperature. The current doubles

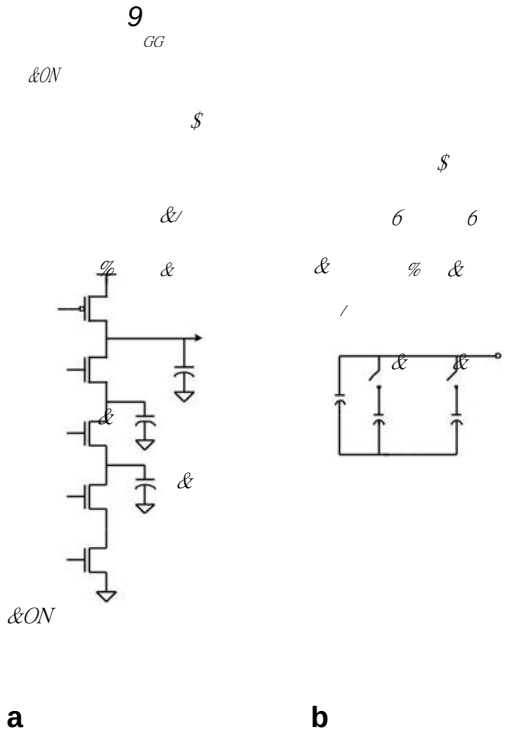


Fig. 5.26 **a** Charge sharing problem; and **b** model for charge sharing

for every 10 °C increase in temperature. Moreover, there will be some subthreshold leakage current. Even when the transistor is OFF and the current can still flow from the drain to source. This component of current increases when the gate voltage is above zero and as it approaches the threshold voltage, this effect becomes more pronounced. A sufficiently high threshold voltage V_t is recommended to alleviate this problem. The charge leakage results in gradual degradation of high-level volt-age output with time, which prevents it to operate at a low clock rate. This makes it unattractive for many applications, such as toys, watches, etc., which are operated at lower frequency to minimize power dissipation from the battery. As a consequence, the voltage on the charge storage node slowly decreases. This needs to be compen-sated by refreshing the charge at regular intervals.

5.4.4.2 Charge Sharing Problem

Dynamic circuits suffer from charge sharing problem because of parasitic capaci-tances at different nodes of a circuit. For example, consider the circuit of Fig. 5.26a. The node A charges to V_{dd} during the pre-charge phase with a charge of $C_L V_{dd}$ stored on the capacitor C_L. Assume that the parasitic capacitances of nodes B and C are C_1 and C_2, respectively. To start with these capacitors are assumed to have no charges. As there is no current path from node A to GND, it should stay at V_{dd} during the evaluation phase. But because of the presence of C_1 and C_2, redistribution of charge will take place leading to the so-called *charge-sharing* problem. This charge-shar-ing mechanism can be modeled as shown in Fig. 5.26b, where C_1 and C_2 are con-nected to node A through two switches representing the MOS transistors. Before the switches are closed, the charge on C_L is given by $Q_A = V_{dd}C_L$ and charges at node B and C are $Q_B = 0$ and $Q_C = 0$, respectively.

After the switches are closed, there will be a redistribution of charges based on the charge conservation principle, and the voltage V_A at node A is given by

$$C_L V_{dd} = (C_L + C_1 + C_2)V_A. \quad (5.15)$$

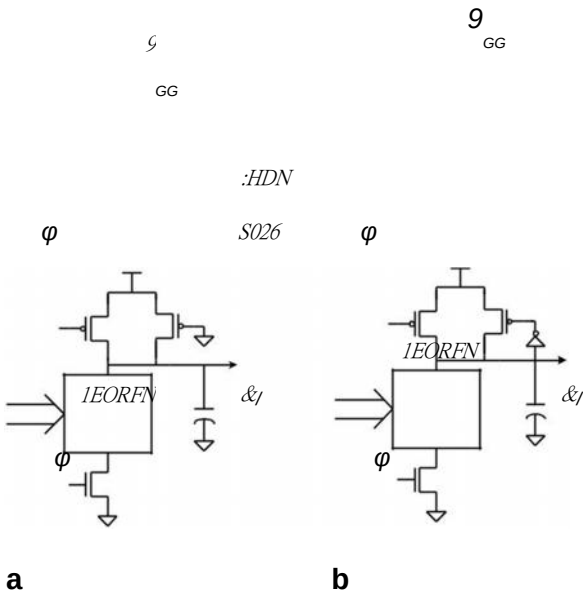


Fig. 5.27 A weak p-type MOS (pMOS) transistor to reduce the impact of charge leakage and charge sharing problem

Therefore, the new voltage at node A, at the end of the evaluation phase, will be \

$$V_A = \frac{C_L}{(C_L + C_1) + C_2} V_{dd} . \quad (5.16)$$

If $C_1 = C_2 = 0.5\ C_L$, then $V_A = 0.5\ V_{dd}$. This may not only lead to incorrect interpreta-tion of the output but also results in the high-static-power dissipation of the suc-ceeding stage.

To overcome the charge sharing and leakage problem, some techniques may be used. As shown in Fig. 5.27, a weak pMOS (low W/L) is added as a pull-up tran-sistor. The transistor always remains ON and behaves like a pseudo-nMOS circuit during the evaluation phase. Although there is static power dissipation due to this during evaluation phase, it helps to maintain the voltage by replenishing the charge loss due to the leakage current or charge sharing.

5.4.4.3 Clock Skew Problem

It is not uncommon to use several stages of dynamic circuits to realize a Boolean function. Although same clock is applied to all these stages, it suffers delay due to resistance and parasitic capacitances associated with the wire that carry the clock pulse and this delay is approximately proportional to the square of the length of the wire. As a result, different amounts of delays are experienced at different points in the circuit and the signal-state changes that are supposed to occur simultaneously may never actually occur at the same time. This is known as *clock skew* problem and it results in hazard and race conditions. For example, let us consider the two stages of CMOS dynamic circuits. Timing relationship of the two clocks of two consecutive stages is shown in Fig. 5.28a. The two clocks are not in synchroniza-tion. When pre-charging is in progress for the first stage, evaluation phase has al-ready started in the second stage. As the output is high at the time of the pre-charge phase, this will discharge the output of the second stage through the MOS transistor Q_1, irrespective of what will be the output of the first stage during the evaluation phase. The charge loss leads to reduced noise margins and even malfunctioning

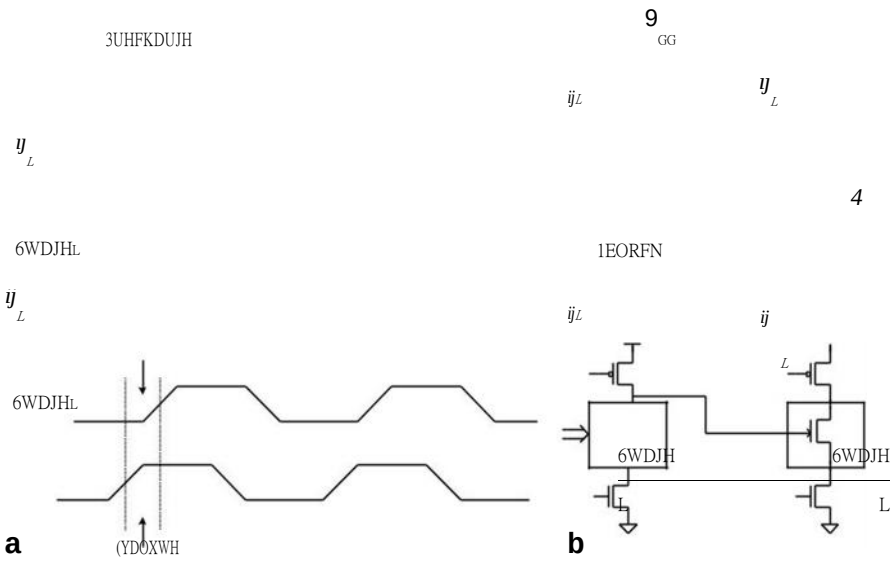
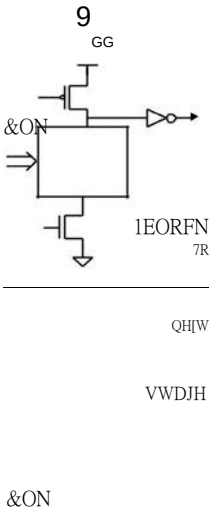


Fig. 5.28 **a** Evaluate phase of a particular stage overlapping with the pre-charge phase of the preceding stage

Fig. 5.29 Domino logic and low levels, respectively



of the circuit; because, if the output depends on the difference in delay time of the two clocks, then the charge on the load capacitance C_L of the second stage may discharge to the extent that the voltage is less than the threshold voltage of the MOS transistor of the next stage. As a consequence, this problem may lead to incorrect output as a cascading problem arises because the output is pre-charged to a high level, which leads to inadvertent discharge of the next stage. This problem can be overcome if the output can be set to low during pre-charge. This basic concept is used in realizing domino CMOS.

There are several approaches to overcome this problem. One straightforward approach to deal with this problem is to introduce a delay circuit in the path of the clock that will compensate for the discharge delay between the first and the second stage of the dynamic circuit. It may be noted that the delay is a function of the complexity of the discharge path. This "self-timing" scheme is fairly simple and often used is CMOS programmable logic array (PLA) and memory design. How-ever, there exist other approaches by which this problem can be overcome. Special type of circuits, namely domino and (no race) NORA CMOS circuits, as discussed below, can also be used.

5.4.5 Domino CMOS Circuits

A single-domino CMOS gate is shown in Fig. 5.29. It consists of two distinct components. The first component is a conventional dynamic pseudo-nMOS gate, which

works based on the pre-charge technique. The second component is a static invert-ing CMOS buffer. Only the output of the static buffer is fed as an input to the subse-quent stage. During the pre-charge phase, the dynamic gate has a high output, so the buffer output is low. Therefore, during pre-charge, all circuit inputs which connect the output of one domino gate to the input of another are low, and the transistors which are driven by these outputs are OFF. Moreover, during the evaluation phase, a domino gate can make only one transition namely a low to high. This is due to the fact that the dynamic gate can only go from high to low during the evaluation phase. As a result, the buffer output could only go from low to high during evalua-tion phase, which makes the circuits glitch-free.

During the evaluation phase, output signal levels change from low to high from the input towards the output stages, when several domino logic gates are cascaded. This phenomenon resembles the chain action of a row of falling dominos, hence the name *domino logic* [5].

Domino CMOS circuits have the following advantages:

- Since no DC current path is established either during the pre-charge phase or during the evaluation phase, domino logic circuits have lower power consump-tion.

- As n-block is only used to realize the circuit, domino circuits occupy lesser chip area compared to static CMOS circuits.

- Due to lesser number of MOS transistors used in circuit realization, domino CMOS circuits have lesser parasitic capacitances and hence faster in speed com-pared to static CMOS.

Full pull-down current is also available to drive the output nodes. Moreover, the use of single clock edge to activate the circuit provides simple operation and maxi-mum possible speed. One limitation of domino logic is that each gate requires an inverting buffer. However, this does not pose a serious problem, because buffers are needed anyway to achieve higher speed and higher fan-out. Another limitation of the domino logic is

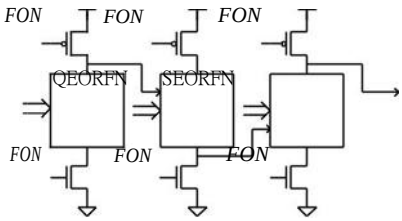
that all the gates are non-inverting in nature, which calls for new logic design paradigm which is different from the conventional approach.

5.4.6 NORA Logic

Another alternative is to use the duality property of the n-blocks and p-blocks used in the realization CMOS circuits. The pre-charge output of an n-block is “1,” where-as the pre-charge output of a p-block is 0. By alternatively using p-blocks and n-blocks, as shown in Fig. 5.30, the clock skew problem is overcome in NORA logic circuits.

NORA stands for NO RAce [6]. Here, both the n- and p-blocks are in pre-charge phase when $clk = 1$ or $(clk = 0)$. The outputs of n-mostly and p-mostly blocks are pre-charged and pre-discharged to 1 and 0, respectively. During this phase, the in-puts are set up. The n-mostly and p-mostly blocks are in evaluation phase when

Fig. 5.30 NORA logic style



clk = 1. During this phase, inputs are held constant and outputs are evaluated as a function of the inputs.

As pre-charged condition output is 1 (0) for an n-block (p-block), it cannot lead to any discharge of the outputs of a p-block (n-block). As a consequence, the inter-nal delays cannot result in a race condition, and we may conclude that the circuits designed based on the NORA logic are internal-delay race-free.

Compared to the domino technique, this approach gives higher flexibility. As both inverted and non-inverted signals are available from the NORA technique, alternating n-mostly and p-mostly blocks can be used to realize circuits of arbitrary logical depth. When connection between same types of blocks is necessary, domino like inverters may be used between them. The NORA logic style has the disadvan-tage that p-block is slower than n-blocks, due to lower mobility of the current carrier of the pMOS circuits. Proper sizing, requiring extra area, may be used to equalize the delays. Compared to domino logic circuits, the NORA logic circuits are faster due to the elimination of the static inverter and the smaller load capacitance. The Digital Equipment Corporation (DEC)-alpha processor, the first 250 MHz CMOS microprocessor, made extensive use of NORA logic circuits.

NORA technique can be also used to realize pipelined circuit in a convenient manner. A single stage of a pipeline block is shown in Fig. 5.30. It consists of one n-mostly, one p-mostly, and a clocked CMOS (C^2MOS) block, which is used as a latch stage for storing information. The pipeline circuits can be realized by apply-ing φ and φ' to consecutive pipeline blocks. For $\varphi = 0$ and $\varphi' = 1$, the φ-sections are pre-charged; while, the φ-sections outputs are held constant by the C^2MOS latch stages. Then for phase $\varphi = 1$ and $\varphi' = 0$, the φ-sections are in evaluation phase and φ-sections are in pre-charge phase. Now, the φ-section outputs, evaluated in the previous phase, are held constant and the φ-sections can use this information for computation. In this way, information flows through the pipeline of alternating φ and φ' sections.