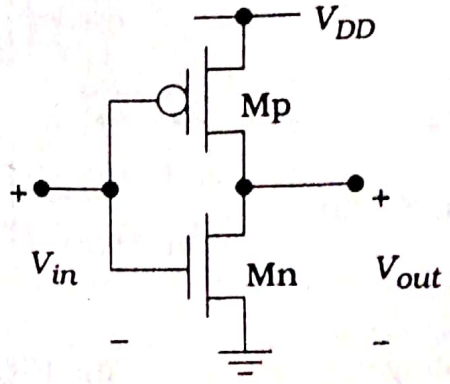
Electronic Analysis of CMOS Logic Gates

7

In the previous chapter we examined the electrical characteristics of MOSFETs. This sets the foundation for analyzing the behavior of transistors in CMOS logic circuits in this chapter. The treatment centers on the important areas of switching speed and layout design, and provides the foundation for much of modern chip design.

DC Characteristics of the CMOS Inverter

The CMOS inverter gives the basis for calculating the electrical characteristics of logic gates. Consider the circuit shown in Figure 7.1. The input voltage V_{in} determines the conduction states of the two FETs Mn and Mp. This produces the output voltage V_{out} of the gate. Two types of calculations are needed to characterize a digital logic circuit. A **DC analysis** determines V_{out} for a given value of V_{in}. In this type of calculation, it is assumed that V_{in} is changed very slowly, and that V_{out} is allowed to stabilize before a measurement is made. A DC analysis provides a direct mapping of the input to the output, which in turn tells us the voltage ranges



pFET: $V_{Tp} < 0$

$$\beta_p = k'_p \left(\frac{W}{L}\right)_p$$

nFET: $V_{Tn} > 0$

$$\beta_n = k'_n \left(\frac{W}{L}\right)_n$$

Figure 7.1 The CMOS inverter circuit

that define Boolean logic 0 and logic 1 values. The second type of characterization is called a **transient analysis**. In this case, the input voltage is an explicit function of time $V_{in}(t)$ corresponding to a changing logic value. The response of the circuit is contained in $V_{out}(t)$. The delay between a change in the input and the corresponding change at the output is the fundamental limiting factor for high-speed design. In this section we will concentrate on the DC analysis. The transient response is analyzed in the next section.

The DC characteristics of the inverter are portrayed in the **voltage transfer characteristic (VTC)**, which is a plot of V_{out} as a function of V_{in}. This is obtained by varying the input voltage V_{in} in the range from 0 V to V_{DD} and finding the output voltage V_{out}. The end point values are easily found with the aid of the circuits in Figure 7.2. If V_{in} is equal to 0 V as in Figure 7.2(a), Mn is off while Mp is on. Since the pFET is on, it connects the output to the power supply and gives $V_{out} = V_{DD}$. This defines the **output high voltage** of the circuit as

$$V_{OH} = V_{DD} \tag{7.1}$$

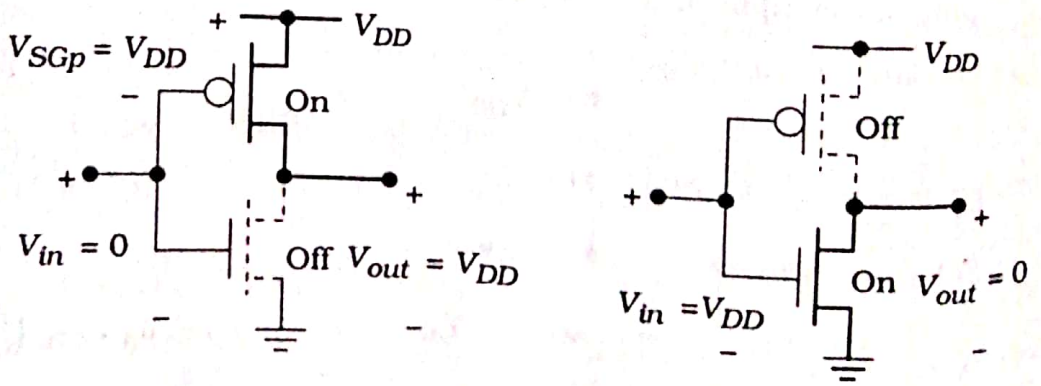
i.e., the highest output voltage is the value of the power supply V_{DD}. The opposite case with $V_{in} = V_{DD}$ is illustrated in Figure 7.2(b). This turns on Mn while Mp is in cutoff. The output node is then connected to 0 V (ground) through the nFET, defining the **output low voltage**

$$V_{OL} = 0 \text{ V} \tag{7.2}$$

The **logic swing** at the output is

$$\begin{aligned} V_L &= V_{OH} - V_{OL} \\ &= V_{DD} \end{aligned} \tag{7.3}$$

Since this is equal to the full value of the power supply, this is called a **full-rail output**.



(a) Low input voltage (b) High input voltage

Figure 7.2 V_{OH} and V_{OL} for the inverter circuit

The VTC for the circuit is obtained by starting with an input voltage of $V_{in} = 0$ V and then increasing it up to a value of $V_{in} = V_{DD}$. This results in the plot shown in Figure 7.3. The details can be understood by writing the device voltages in terms of the input and output voltages:

$$V_{GSn} = V_{in}$$
$$V_{SGp} = V_{DD} - V_{in} \tag{7.4}$$

Mn is in cutoff so long as $V_{in} \leq V_{Tn}$. Since the output voltage is high with a value $V_{out} = V_{DD}$, any input voltage in the range labeled as "0" can be interpreted as a logic 0 input. Increasing V_{in} causes a downward transition in the VTC. This is because the input voltage turns the nFET on while the pFET is still conducting. Note, however, that increasing V_{in} decreases V_{SGp}, so the pFET becomes a less efficient conductor and the output voltage falls. Mp goes into cutoff when

$$V_{in} = V_{DD} - |V_{Tp}| \tag{7.5}$$

For V_{in} greater than this value, $V_{out} = 0$ V since only the nFET is active. This shows that there is a range of input voltages that act as logic 1 input values as indicated by the "1" on the VTC.

The logic 0 and 1 voltage ranges are defined by the changing slope of the VTC. Point 'a' in the drawing is where the slope has a value of -1, and defines the **input low voltage** V_{IL}. By definition, a logic 0 input voltage is defined by

$$0 \leq V_{in} \leq V_{IL} \tag{7.6}$$

The second -1 slope point is labeled as 'b' and defines the **input high voltage** V_{IH}. This is used to define a logic 1 input voltage as

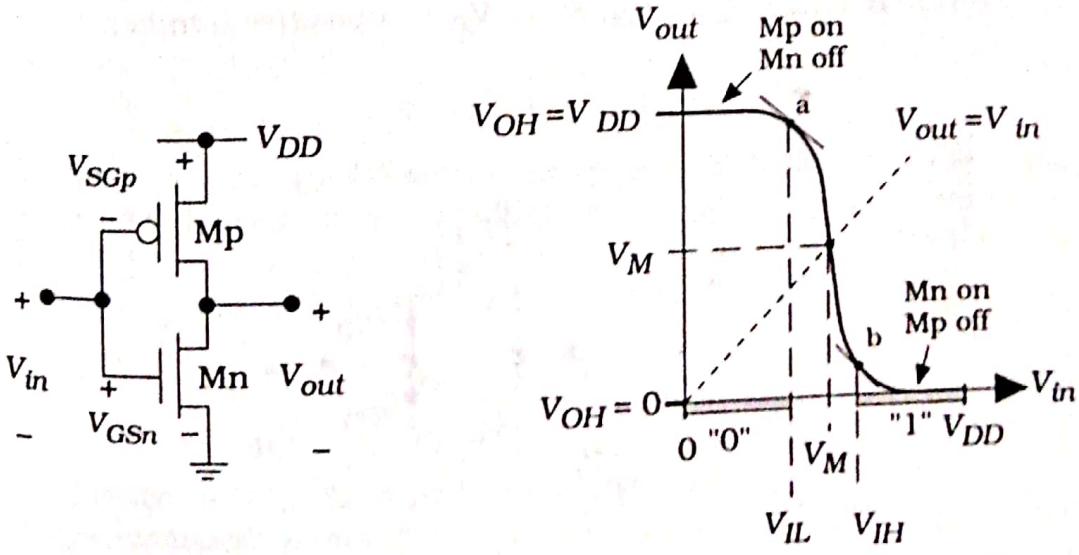


Figure 7.3 Voltage transfer curve for the NOT gate

$$V_{IH} \le V_{in} \le V_{DD} \tag{7.7}$$

The **voltage noise margins** are

$$VNM_H = V_{OH} - V_{IH}$$
$$VNM_L = V_{IL} - V_{OL} \tag{7.8}$$

for high and low states, respectively. The noise margins give a quantitative measure of how stable the inputs are with respect to coupled electromagnetic signal interference.

While it is possible to calculate the exact voltages that define logic 0 and 1 input voltages, it is simpler to introduce the midpoint voltage V_M shown in the VTC. This is defined as the point where the VTC intersects the unity gain line that is defined by $V_{out} = V_{in} = V_M$. A value of $V_{in} = V_M$ is in the transition region and does not represent a Boolean quantity. However, for V_{in} less than V_M the input voltage is toward the logic 0 values while $V_{in} > V_M$ indicates that the input is on the logic 1 side. Knowing the value of V_M thus tells us the center point for input transitions.

To calculate the midpoint voltage we set $V_{in} = V_{out} = V_M$ as shown in Figure 7.4. Equating the drain currents of the FETs gives

$$I_{Dn} = I_{Dp} \tag{7.9}$$

but we need to find the operating region (saturation or non-saturation) of each FET before we can use the expression. Consider first the nFET and recall that the saturation voltage is given by

$$V_{sat} = V_{GSn} - V_{Tn}$$
$$= V_M - V_{Tn} \tag{7.10}$$

where we have used $V_{in} = V_{GSn} = V_M$ in the second line. The drain-source voltage is $V_{DSn} = V_{out} = V_M$. Since V_{Tn} is a positive number,

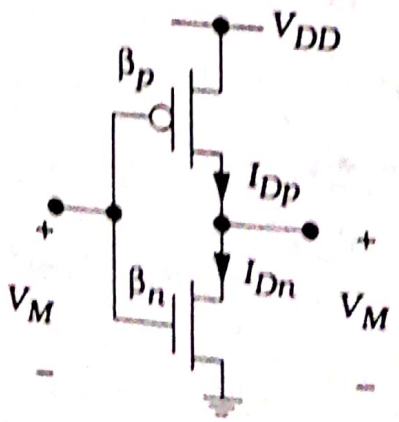


Figure 7.4 Inverter voltages for V_M calculation

$$V_{DSn} > V_{sat} = V_M - V_{Tn} \tag{7.11}$$

which says that Mn must be saturated. The same arguments can be applied to the pFET Mp since $V_{SGp} = V_{SDp}$. Using the saturation current equations from Chapter 6 gives

$$\frac{\beta_n}{2}(V_M - V_{Tn})^2 = \frac{\beta_p}{2}(V_{DD} - V_M - |V_{Tp}|)^2 \tag{7.12}$$

Dividing by β_p and taking the square root gives

$$\sqrt{\frac{\beta_n}{\beta_p}}(V_M - V_{Tn}) = V_{DD} - V_M - |V_{Tp}| \tag{7.13}$$

Simple algebra then gives the midpoint voltage as

$$V_M = \frac{V_{DD} - |V_{Tp}| + \sqrt{\frac{\beta_n}{\beta_p}} V_{Tn}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}} \tag{7.14}$$

This equation shows that V_M is set by the nFET-to-pFET ratio

$$\frac{\beta_n}{\beta_p} = \frac{k'_n \left(\frac{W}{L}\right)_n}{k'_p \left(\frac{W}{L}\right)_p} \tag{7.15}$$

Since k'_n and k'_p are set in the processing, the ratio of the FET sizes establishes the switching point. It is important to remember that nFETs and pFETs have different mobility factors with a typical ratio of

$$\frac{k'_n}{k'_p} \approx 2 \text{ to } 3 \tag{7.16}$$

depending upon the details of the processing. This fact has a significant effect on the choices we make in both the sizing of individual transistors, and the types of circuits that are used in advanced VLSI designs. Note that, since C_{ox} is approximately the same for both FET types,

$$\frac{k'_n}{k'_p} = \frac{\mu_n}{\mu_p} = r \tag{7.17}$$

where r is the mobility ratio introduced in Chapter 5.

A **symmetrical inverter** VTC is one that has equal "0" and "1" input voltage ranges. This can be achieved by choosing

$$V_M = \frac{1}{2}V_{DD} \tag{7.19}$$

in equation (7.12). Rearranging gives us the design equation

$$\frac{\beta_n}{\beta_p} = \left(\frac{\frac{1}{2}V_{DD} - |V_{Tp}|}{\frac{1}{2}V_{DD} - V_{Tn}}\right)^2 \tag{7.19}$$

This allows us to compute the transistor sizes for this particular choice of V_M. Note that if $V_{Tn} = |V_{Tp}|$, then a symmetric design requires that

$$\beta_n = \beta_p \tag{7.20}$$

i.e., the device transconductance values of the two FETs are equal. It is important to remember that β is proportional to the aspect ratio (W/L) of a MOSFET, and that (W/L) is the actual design variable.

Example 7.1

Consider a CMOS process with the following parameters

$$k'_n = 140\ \mu A/V^2 \qquad V_{Tn} = +0.70\ V$$
$$k'_p = 60\ \mu A/V^2 \qquad V_{Tp} = -0.70\ V \tag{7.21}$$

with $V_{DD} = 3.0$ V.

Consider the case where $\beta_n = \beta_p$. We can verify that this is a symmetrical design by calculating

$$V_M = \frac{3 - 0.7 + \sqrt{1}(0.7)}{1 + \sqrt{1}} = 1.5\ V \tag{7.22}$$

so that V_M is one-half the value of the power supply voltage. To achieve this design, we must choose the device aspect ratios such that

$$\frac{\beta_n}{\beta_p} = \frac{k'_n\left(\frac{W}{L}\right)_n}{k'_p\left(\frac{W}{L}\right)_p} = 1 \tag{7.23}$$

where we recall that the process transconductance parameters k' are given by $k' = \mu_n C_{ox}$ and are set by the processing. For the present case we rearrange the expression to read

$$\left(\frac{W}{L}\right)_p = \frac{k'_n}{k'_p}\left(\frac{W}{L}\right)_n \tag{7.24}$$

so that

$$\left(\frac{W}{L}\right)_p = \left(\frac{140}{60}\right)\left(\frac{W}{L}\right)_n = 2.33\left(\frac{W}{L}\right)_n \tag{7.25}$$

This shows that the pFET must be about 2.33 times larger than the nFET.

Let us now examine the case where the nFET and the pFET have the same aspect ratio: $(W/L)_n = (W/L)_p$. With the values provided in the problem statement,

$$\frac{\beta_n}{\beta_p} = \frac{k'_n}{k'_p} = 2.33 \tag{7.26}$$

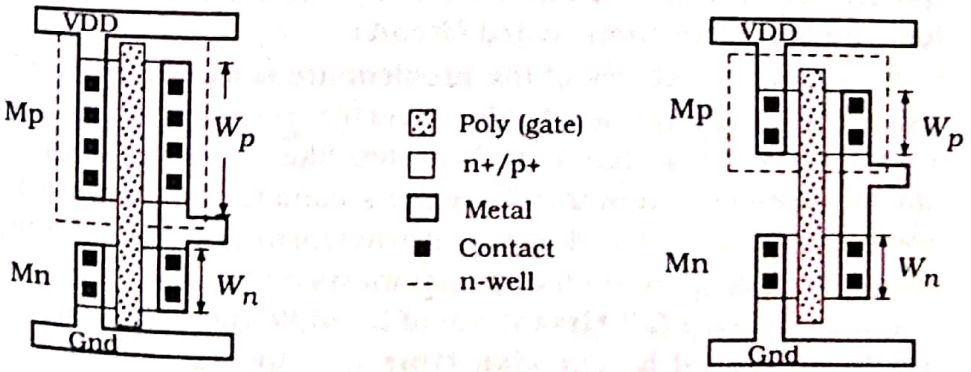
so that the midpoint voltage is given by

$$V_M = \frac{3 - 0.7 + \sqrt{2.33}\ (0.7)}{1 + \sqrt{2.33}} = 1.33\ \text{V} \tag{7.27}$$

This choice shifts V_M to a value that is smaller than $(V_{DD}/2)$.

Figure 7.5 illustrates the difference in the layout between an inverter that uses the two design styles. The channel length is the same for both transistors in the inverter, leaving the channel widths W_p and W_n as the design variables. In Figure 7.5(a), the pFET has a width of about $W_p \approx 2\ W_n$ which gives V_M of about $(V_{DD}/2)$. Equal size transistors are used in the layout of Figure 7.5(b), so that the circuit has $V_M < (V_{DD}/2)$. It is important to remember that we are only dealing with the DC characteristics at the moment. As we will see in the next section, the switching properties of the two designs are also affected by the aspect ratios.

The derivation and examples above illustrate the importance of the FET aspect ratios in the DC behavior of the logic gate. At the physical level, the relative device sizes contained in the ratio (β_n/β_p) determine the switching points. In general, increasing (β_n/β_p) decreases the value of the midpoint voltage V_M. This dependence is illustrated in the plot of Figure



(a) Larger pFET design

(b) Equal aspect ratios

Figure 7.5 Comparison of the layouts for Example 7.1

7.6. With the parameters shown, a symmetrical design with $\beta_n = \beta_p$ gives $V_M = (V_{DD}/2) = 1.5$ V. Increasing the ratio to $(\beta_n/\beta_p) = 1.5$ gives $V_M = 1.42$ V, while $(\beta_n/\beta_p) = 2.5$ decreases the midpoint voltage to $V_M = 1.31$ V. It is also possible to use a ratio of $(\beta_n/\beta_p) < 1$, which shifts the VTC toward the right, i.e., $V_M > (V_{DD}/2)$. However, this is rarely used since the pFET aspect ratios get quite large.

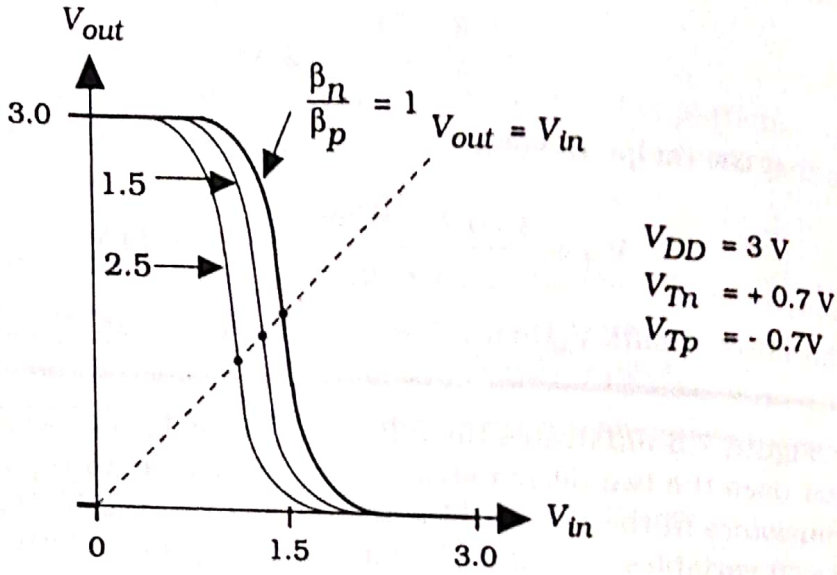


Figure 7.6 Dependence of V_M on the device ratio

7.2 Inverter Switching Characteristics

High-speed digital system design is based on the ability to perform calculations very quickly. This requires that logic gates introduce a minimum amount of time delay when the inputs change. Designing fast logic circuits is one of the more challenging (but critical) aspects of VLSI physical design. As with the DC analysis, analyzing the NOT gate provides a basis for studying more complicated circuits.

The general features of the problem are shown in Figure 7.7. An input voltage $V_{in}(t)$ is applied to the inverter, resulting in an output voltage $V_{out}(t)$. We assume that $V_{in}(t)$ has step-like characteristics and makes an abrupt transition from 0 to 1 (i.e., to a voltage of V_{DD}) at time t_1, and back down to 0 at time t_2. The output waveform reacts to the input, but the output voltage cannot change instantaneously. The output 1-to-0 transition introduces a **fall time** delay of t_f, while the 0-to-1 change at the output is described by the **rise time** t_r. The rise and fall times can be calculated by analyzing the electronic transitions of the circuits.

The rise and fall time delays are due to the parasitic resistance and

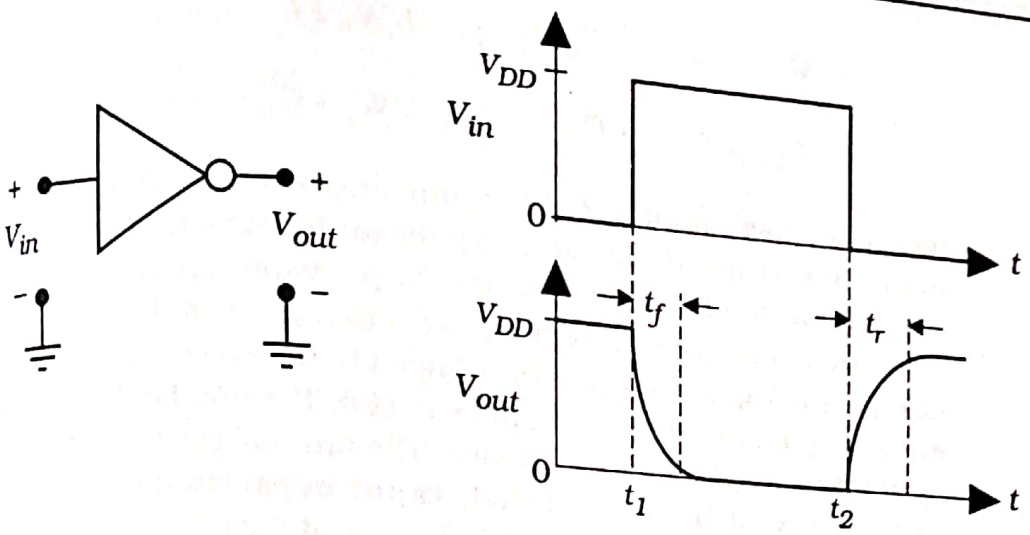
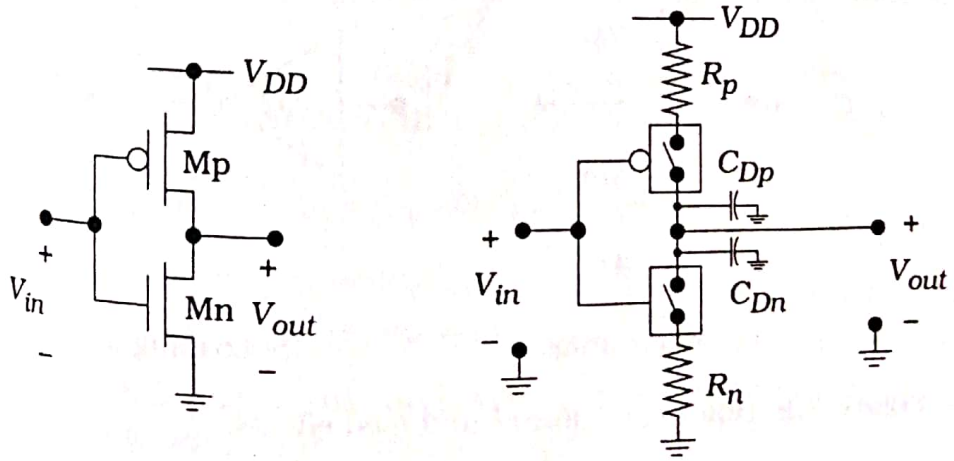


Figure 7.7 General switching waveforms

capacitances of the transistors. Consider the NOT circuit shown in Figure 7.8(a). Both FETs can be replaced by their switch equivalents, which results in the simplified RC model in Figure 7.8(b). It is worth recalling that the actual values of the components depend upon the device dimensions. Once we specify the aspect ratios $(W/L)_n$ and $(W/L)_p$, we can calculate R_n and R_p using

$$R_n = \frac{1}{\beta_n(V_{DD} - V_{Tn})}$$
$$R_p = \frac{1}{\beta_p(V_{DD} - |V_{Tp}|)} \tag{7.28}$$

Knowing the layout dimensions of each FET allows us to find the capacitances C_{Dn} and C_{Dp} at the output node. The formulas are given by



(a) FET circuit (b) RC switch model equivalent

Figure 7.8 RC switch model equivalent for the CMOS inverter

$$C_{Dn} = C_{GSn} + C_{DBn} = \frac{1}{2}C_{ox}L'W_n + C_{jn}A_n + C_{jswn}P_n$$

$$C_{Dp} = C_{GSp} + C_{DBp} = \frac{1}{2}C_{ox}L'W_p + C_{jp}A_p + C_{jswp}P_p \quad (7.2_p)$$

where we have added n and p subscripts to specify the nFET or pFET quantities, respectively.[1] It is significant to remember that increasing the channel width of a FET increases the parasitic capacitance values.

There is one more important point that needs to be included before we obtain a complete model. In a logic chain, every logic gate must drive another gate, or set of gates, to be useful. The number of gates is specified by the **fan-out** (FO) of the circuit. The fan-out gates act as a **load** to the driving circuit because of their **input capacitance** C_{in}. Consider the inverter shown in Figure 7.9(a). The input capacitance of the inverter is just the sum of the FET capacitances

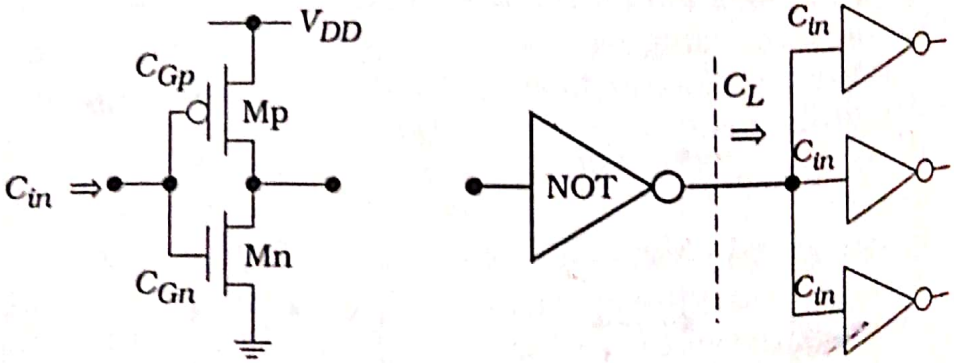
$$C_{in} = C_{Gp} + C_{Gn} \quad (7.30)$$

Figure 7.8(b) shows the effect of input capacitance for a fan-out of FO = 3. The input capacitance to each gate acts as an **external load capacitance** C_L to the driving gate. In this example, it is easily seen that

$$C_L = 3C_{in} \quad (7.31)$$

is the value of the load presented to the NOT gate.

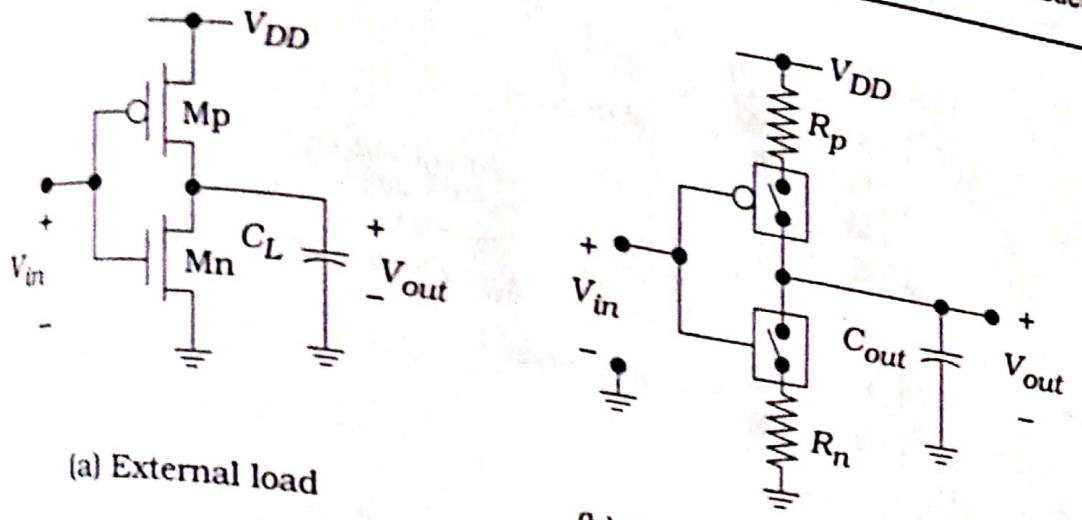
We may now calculate the switching times of the inverter. Figure 7.10 illustrates the general problem. A CMOS NOT gate is used to drive an external load capacitance C_L as in Figure 7.10(a). This gives the complete



(a) Single stage (b) Loading due to fan-out

Figure 7.9 Input capacitance and load effects

[1] Note that the source capacitances C_{Sp} and C_{Sn} do not enter the problem as they are at the power supply and ground, respectively, and have constant voltages.



(a) External load

(b) Complete switching model

Figure 7.10 Evolution of the inverter switching model

switching model shown in Figure 7.10(b) where the total output capacitance is defined as

$$C_{out} = C_{FET} + C_L \tag{7.32}$$

The FET capacitances shown earlier in Figure 7.8 have been merged into the single term

$$C_{FET} = C_{Dn} + C_{Dp} \tag{7.33}$$

and are the parasitic internal contributions that cannot be eliminated. These add with C_L since all elements are in parallel. The total output capacitance C_{out} is the load that the gate must drive; the numerical value varies with the load.

Example 7.2

Let us apply this analysis to find the capacitances in the NOT gate shown in Figure 7.11. It is assumed that all dimensions have units of microns (μm).

First we will find the gate capacitances using

$$C_{Gp} = (2.70)(1)(8) = 21.6 \text{ fF}$$
$$C_{Gn} = (2.70)(1)(4) = 10.8 \text{ fF} \tag{7.34}$$

Next, note that the overlap distance L_o is specified as 0.1 μm, which should be included in the area and perimeter factors in the junction capacitances. For the pFET, the p+ capacitance is

$$C_p = C_j A_{bot} + C_{jsw} P_{sw} \tag{7.35}$$

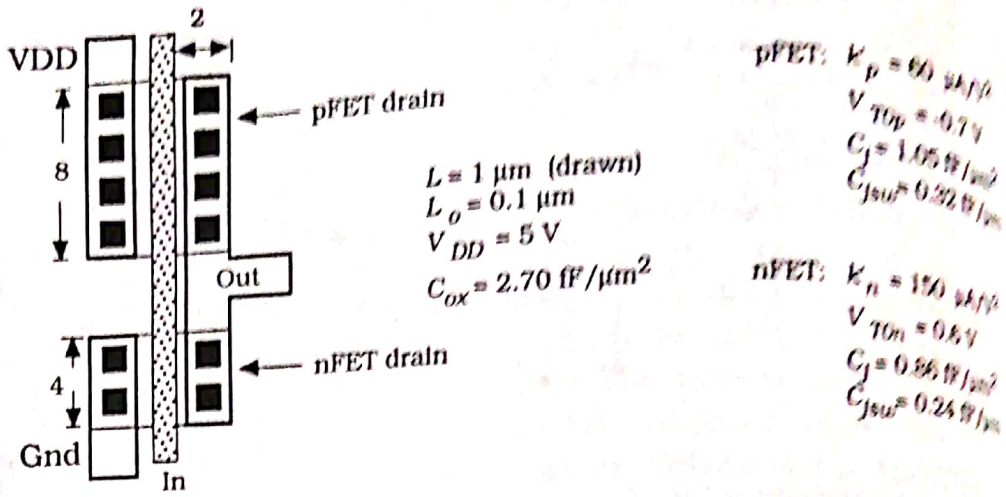


Figure 7.11 Example of capacitance calculations

so

$$C_p = (1.05)(8)(2.1) + (0.32)2(8 + 2.1) = 24.10 \text{ fF} \quad (7.3)$$

The total capacitance at the pFET drain is therefore given by

$$C_{Dp} = \frac{21.6}{2} + 24.10 = 34.9 \text{ fF} \quad (7.3)$$

The nFET drain is analyzed using the same approach. The n+ junction capacitance is

$$C_n = (0.86)(4)(2.1) + (0.24)(2)(4 + 2.1) = 10.15 \text{ fF} \quad (7.3)$$

so that

$$C_{Dn} = \frac{10.8}{2} + 10.15 = 15.55 \text{ fF} \quad (7.3)$$

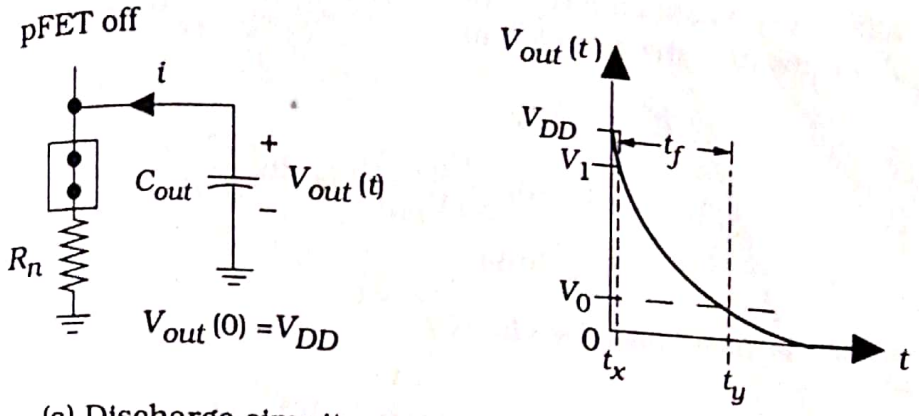
is the total capacitance at the drain of the nFET. Adding gives

$$\begin{aligned} C_{FET} &= C_{Dp} + C_{Dn} \\ &= 34.9 + 15.55 \\ &= 50.45 \text{ fF} \end{aligned} \quad (7.4)$$

as the total internal FET capacitance. The total capacitance at the output is

$$C_{out} = 50.45 + C_L \quad (7.4)$$

in fF, where C_L is the external load (also in fF).



(a) Discharge circuit

(b) Output waveform

Figure 7.12 Discharge circuit for the fall time calculation

7.2.1 Fall Time Calculation

Let us start by calculating the output fall time t_f. We will shift the time origin such that V_{in} changes from 0 to V_{DD} at time $t = 0$. The initial condition at the output is $V_{out}(0) = V_{DD}$. When the input is switched, the nFET goes active while the pFET is driven into cutoff. In terms of the switch models, the nFET switch is closed and the pFET switch is open. This gives us the simplified discharge circuit shown in Figure 7.12(a). The capacitor C_{out} is initially charged to a voltage V_{DD}, and is allowed to discharge to 0 V through the nFET resistance R_n. The current leaving the capacitor is

$$i = -C_{out}\frac{dV_{out}}{dt} = \frac{V_{out}}{R_n} \tag{7.42}$$

which gives the differential equation for the discharge event. Solving with the initial condition $V_{out}(0) = V_{DD}$ results in the well-known form

$$V_{out}(t) = V_{DD}e^{-t/\tau_n} \tag{7.43}$$

where

$$\tau_n = R_n C_{out} \tag{7.44}$$

is the nFET **time constant** with units of seconds. The function is plotted in Figure 7.12(b).

The fall time is traditionally defined to be the time interval from $V_1 = 0.9\,V_{DD}$ to $V_0 = 0.1\,V_{DD}$, which are respectively known as the 90% and the 10% voltages as referenced to the full rail swing of V_{DD}. Rearranging the solution to the form

$$t = \tau_n \ln\left(\frac{V_{DD}}{V_{out}}\right) \tag{7.45}$$

allows us to calculate the time t needed to fall to a particular voltage V_x. From the drawing we see that

$$\begin{aligned} t_f &= t_y - t_x \\ &= \tau_n \ln\left(\frac{V_{DD}}{0.1V_{DD}}\right) - \tau_n \ln\left(\frac{V_{DD}}{0.9V_{DD}}\right) \\ &= \tau_n \ln(9) \end{aligned} \tag{7.47}$$

where we have used the identity

$$\ln(a) - \ln(b) = \ln\left(\frac{a}{b}\right) \tag{7.48}$$

in the last step. Approximating $\ln(9) \approx 2.2$ gives the final result

$$t_f \approx 2.2\tau_n \tag{7.49}$$

as the fall time for the circuit. The output fall time in a generic digital logic gate is usually called the output **high-to-low time** t_{HL} and is identical to the value computed here:

$$t_{HL} = t_f \tag{7.49}$$

The two symbols will be used interchangeably in the discussion.

7.2.2 The Rise Time

The rise time calculation follows in the same manner. Initially, the input voltage is at $V_{in} = V_{DD}$ and is switched to $V_{in} = 0$ V; we time shift this event to occur at $t = 0$ for simplicity. This turns on the pFET while simultaneously driving the nFET into cutoff, so that the simplified charging circuit of Figure 7.13(a) is valid. The output voltage at $t = 0$ is given by $V_{out}(0) = 0$ V.

The charging current is given by

$$i = C_{out}\frac{dV_{out}}{dt} = \frac{V_{DD} - V_{out}}{R_p} \tag{7.50}$$

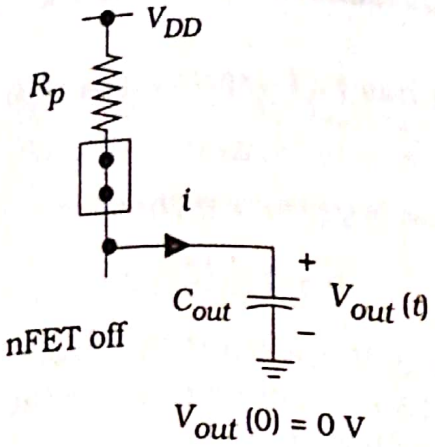
Solving and applying the initial condition gives the exponential form

$$V_{out}(t) = V_{DD}[1 - e^{-t/\tau_p}] \tag{7.51}$$

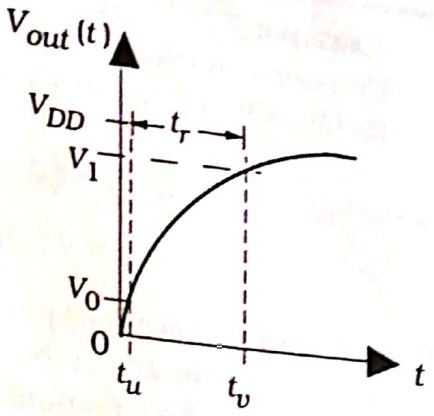
where the pFET time constant is defined by

$$\tau_p = R_p C_{out} \tag{7.52}$$

Figure 7.13(b) shows the output voltage as a function of time. The rise time is taken between 10% and 90% points such that



(a) Charge circuit



(b) Output waveform

Figure 7.13 Rise time calculation

$$t_r = t_v - t_u \tag{7.53}$$

A little algebra yields the expression

$$t_r = \ln(9)\tau_p \approx 2.2\tau_p \tag{7.54}$$

for the rise time t_r. This has the same form as the fall time t_f because of the symmetry of the charge and discharge circuits. The rise time is identical to the output **low-to-high time** t_{LH}; the symbols will be used interchangeably.

The low-to-high time t_{LH} and the high-to-low time t_{HL} represent the shortest amount of time needed for the output to change from a logic 0 to logic 1 voltage, or from a logic 1 to a logic 0 voltage, respectively. Let us assume that the input is a square wave with a period of T sec such that the voltage is 0 for $(T/2)$ and V_{DD} for a $(T/2)$ time interval.[2] We then define the **maximum signal frequency** as

$$f_{max} = \frac{1}{t_{HL} + t_{LH}} = \frac{1}{t_r + t_f} \tag{7.55}$$

since this is the largest frequency that can be applied to the gate and still allow the output to settle to a definable state.[3] If the signal frequency exceeds f_{max}, the output voltage of the gate will not have sufficient time to stabilize to the correct value.

[2] This defines what is known as a 50% duty cycle.

[3] This definition assumes that t_{HL} and t_{LH} have the same order of magnitude to be useful.

Example 7.3

Consider an inverter circuit that has FET aspect ratios of $(W/L)_n = 6$ and $(W/L)_p = 8$ in a process where

$$k'_n = 150\ \mu A/V^2 \qquad V_{Tn} = +0.70\ V$$
$$k'_p = 62\ \mu A/V^2 \qquad V_{Tp} = -0.85\ V \tag{7.56}$$

and uses a power supply voltage of $V_{DD} = 3.3$ V. The total output capacitance is estimated to be $C_{out} = 150$ fF. Let us compute the rise and fall times using the equations derived above.

Consider first the fall time. The pFET resistance is given by

$$R_p = \frac{1}{\beta_p(V_{DD} - |V_{Tp}|)} = \frac{1}{(62\times10^{-6})(8)(3.3-0.85)} = 822.9\ \Omega \tag{7.57}$$

The time constant for the charging event is computed using the RC product R_pC_{out} to find

$$\tau_p = (822.9)(150\times10^{-15}) = 123.43 \quad ps \tag{7.58}$$

where 1 ps (picosecond) is 10^{-12} sec. The rise time is

$$t_r = 2.2\tau_p = 271.55 \quad ps \tag{7.59}$$

The fall time is calculated in a similar manner. First, we find the nFET resistance

$$R_n = \frac{1}{\beta_n(V_{DD} - V_{Tn})} = \frac{1}{(150\times10^{-6})(6)(3.3-0.70)} = 427.35\ \Omega \tag{7.60}$$

so that the discharge time constant is

$$\tau_p = (427.35)(150\times10^{-15}) = 64.1 \quad ps \tag{7.61}$$

The fall time is

$$t_f = 2.2\tau_n = 141.0 \quad ps \tag{7.62}$$

Combining these results, the maximum signal frequency is

$$f_{max} = \frac{1}{t_r + t_f} = \frac{1}{(271.55 + 141.0)\times 10^{-12}} = 2.42 \quad \text{GHz} \tag{7.63}$$

where 1 GHz = 10^9 Hz. Although this is a very high frequency, it is important to remember that this refers only to a single inverter.

7.2.3 The Propagation Delay

The propagation delay time t_p is often used to estimate the "reaction" delay time from input to output. When we use step-like input voltages, the propagation delay is defined by the simple average of the two time intervals shown in Figure 7.14 by

$$t_p = \frac{(t_{pf} + t_{pr})}{2} \tag{7.64}$$

In this expression, t_{pf} is the output fall time from the maximum level to the "50%" voltage line, i.e., from V_{DD} to $(V_{DD}/2)$; t_{pr} is the propagation rise time from 0 V to $(V_{DD}/2)$. Using the exponential equations for V_{out} we obtain

$$\begin{aligned} t_{pf} &= \ln(2)\tau_n \\ t_{pr} &= \ln(2)\tau_p \end{aligned} \tag{7.65}$$

Approximating $\ln(2) \approx 0.693$ then gives

$$t_p \approx 0.35(\tau_n + \tau_p) \tag{7.66}$$

The propagation delay time is a useful estimate of the basic delay, but does not provide detailed information on the rise and fall times as individual quantities. Propagation delays are commonly used in basic logic simulation programs.

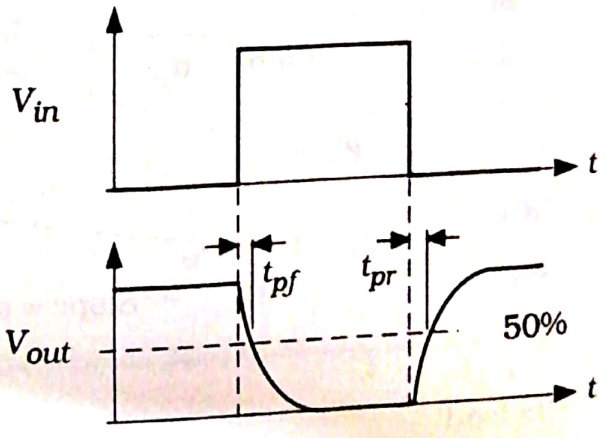


Figure 7.14 Propagation time definitions

7.2.4 General Analysis

The rise and fall time equations provide the basis for high-speed CMOS design. We can manipulate them to show us how to design single logic gates and then characterize the behavior of the gates when used in logic cascades.

To see the important factors, recall that the total output capacitance consists of two terms such that

$$C_{out} = C_{FET} + C_L \tag{7.67}$$

C_{FET} represents the parasitic capacitances of the transistors, while C_L is the external load. The layout geometry establishes the value of C_{FET}, but the load capacitance C_L varies with the application. Substituting this expression into the rise and fall time equations gives

$$\begin{aligned} t_r &\approx 2.2R_p(C_{FET} + C_L) \\ t_f &\approx 2.2R_n(C_{FET} + C_L) \end{aligned} \tag{7.68}$$

which can be cast into the forms

$$\begin{aligned} t_r &= t_{r0} + \alpha_p C_L \\ t_f &= t_{f0} + \alpha_n C_L \end{aligned} \tag{7.69}$$

These show that the rise and fall times are linear functions of the load capacitance C_L. The general behavior of both quantities is shown in Figure 7.15. Under zero-load conditions ($C_L = 0$), the inverter drives its own capacitances such that

$$\begin{aligned} t_r &= t_{r0} \approx 2.2R_p C_{FET} \\ t_f &= t_{f0} \approx 2.2R_n C_{FET} \end{aligned} \tag{7.70}$$

are determined solely from the inverter parameters. When an external

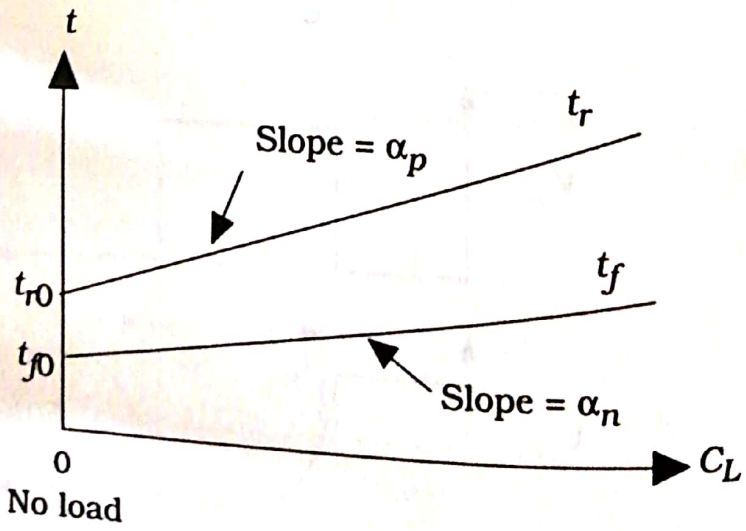


Figure 7.15 General behavior of the rise and fall times

load C_L is added, the switching times increase in a linear fashion. Large capacitive loads may cause problems because of longer delays. The dependence is described by the slope values

$$\alpha_p = 2.2R_p = \frac{2.2}{\beta_p(V_{DD} - |V_{Tp}|)} \tag{7.71}$$

and

$$\alpha_n = 2.2R_n = \frac{2.2}{\beta_n(V_{DD} - V_{Tn})} \tag{7.72}$$

Note that these are inversely proportional to the aspect ratios since

$$\beta_p = k_p'\left(\frac{W}{L}\right)_p, \qquad \beta_n = k_n'\left(\frac{W}{L}\right)_n \tag{7.73}$$

For a given load capacitance C_L, t_r and t_f can be reduced by using large FETs. However, increasing the aspect ratio of a transistor implies that it will consume more area on the chip, which in turn decreases the number of devices that can be placed on the die area allocated for the circuit. Designing for speed thus decreases the integration density of the circuit. This is called the **speed versus area trade-off** which says that

Fast circuits consume more area than slow circuits

Chip designers regularly face the problem of minimizing the switching delays without requiring excessive amounts of silicon "real estate," which is slang for chip area.

Example 7.4

Let us use the results of Example 7.3 to find the general delay equations for the case where the internal FET capacitance is $C_{FET} = 80$ fF.

The rise time t_r is controlled by the pFET that has a resistance of $R_p = 822.9\ \Omega$. The slope is given by

$$\alpha_p = 2.2R_p = 1,810.4\Omega \tag{7.74}$$

while

$$\begin{aligned} t_{r0} &\approx 2.2R_pC_{FET} \\ &= 2.2(822.9)(80\times10^{-15}) \\ &= 144.9 \text{ ps} \end{aligned} \tag{7.75}$$

The rise time can thus be written in the form

$$\begin{aligned} t_r &= t_{r0} + \alpha_pC_L \\ &= 144.9 + 1.810C_L \text{ ps} \end{aligned} \tag{7.76}$$

which requires that C_L be in units of fF.

For the fall time equation, we calculate

$$\alpha_n = 2.2(427.35) = 940.2\Omega \tag{7.7}$$

and

$$t_{f0} \approx 2.2(940.2)(80\times10^{-15}) = 165.5 \text{ ps} \tag{7.7}$$

yielding

$$t_f = 165.5 + 0.940C_L \text{ ps} \tag{7.7}$$

as the general expression.

As an example of using these equations, suppose that the load is specified as C_L = 150 fF. We compute

$$t_r = 144.9 + 1.810(150) = 416.4 \text{ ps}$$
$$t_f = 165.5 + 0.940(150) = 306.5 \text{ ps} \tag{7.8}$$

for the rise and fall times at the output. This corresponds to a maximum switching frequency for the gate of $f_{max} \approx 1.38$ GHz.

The relative values of $(W/L)_n$ and $(W/L)_p$ determine the shape of the output waveform. For example, if we design the circuit such that

$$R_p = R_n \tag{7.8}$$

then the output waveform is symmetrical with

$$t_r = t_f \tag{7.8}$$

To equalize the resistances we must design the circuit such that

$$\beta_p(V_{DD} - |V_{Tp}|) = \beta_n(V_{DD} - V_{Tn}) \tag{7.8}$$

is satisfied. If $V_{Tn} = |V_{Tp}|$, then the requirement reduces to

$$\beta_p = \beta_n \tag{7.8}$$

which gives the DC midpoint voltage at $V_M = (V_{DD}/2)$. This illustrates the fact that the nFET/pFET ratio (β_n/β_p) determines the DC midpoint voltage, while the individual values of β_n and β_p establish the switching times t_f and t_r, respectively.

7.2.5 Summary of the Inverter Circuit

It is worth taking the time to summarize the results of our study to this point. The electrical characteristics of an isolated CMOS inverter are established by two sets of parameters:

- The processing variables, such as k' and V_T values, and parasitic capacitances,

and,

- The transistor aspect ratios $(W/L)_n$ and $(W/L)_p$.

VLSI designers do not have any control over the processing parameters, as they are set by the details of the manufacturing sequence. Device sizing thus becomes the critical issue in high-speed circuit design.

System design is accomplished by using cascades of logic gates to perform the necessary binary operations. In electrical terms, the logic flow path establishes the load capacitance C_L seen by each gate. The choice of aspect ratios is the key to achieving the desired transient response of a chain of gates.

7.3 Power Dissipation

An important characteristic of CMOS integrated circuits is the power dissipated by a particular design technique. The general problem is shown in Figure 7.16. The current I_{DD} flowing from the power supply to ground gives a dissipated power of

$$P = V_{DD} I_{DD} \tag{7.85}$$

Since the value of the voltage supply V_{DD} is assumed to be a constant, we can find the value of P by studying the nature of the current flow. We usually divide the currents into DC and dynamic (or switching) contributions, so let us write

$$P = P_{DC} + P_{dyn} \tag{7.86}$$

where P_{DC} is the DC term and P_{dyn} is due to dynamic switching events.

The DC contribution can be calculated by examining the voltage transfer curve reproduced in Figure 7.17(a). When the input voltage V_{in} is stable at a low logic 0 value, the nFET Mn is off; as seen earlier in Figure 7.2, there is no direct current flow path between V_{DD} and ground. Ideally, the

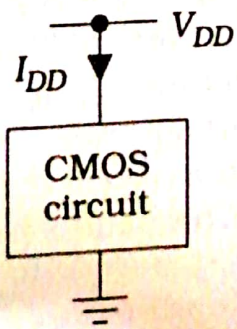


Figure 7.16 Origin of power dissipation calculation

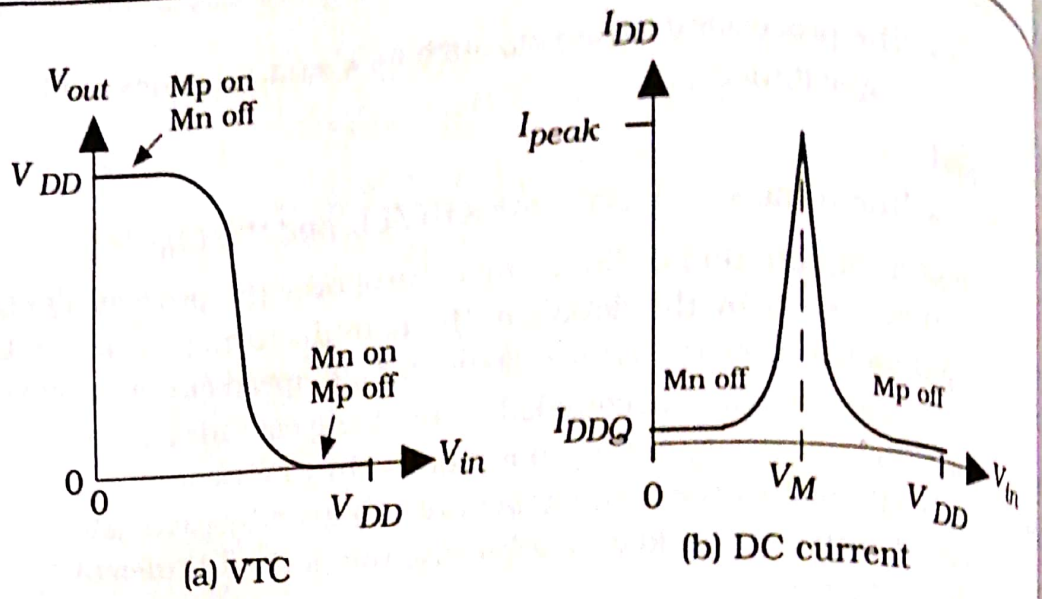


Figure 7.17 DC current flow

DC current flow for this case would be $I_{DD} = 0$, but in a realistic circu small **leakage currents** exist.[4] The value is denoted as I_{DDQ} and is call the **quiescent** leakage current. When V_{in} is switched, the current flo reaches a peak value I_{peak} at V_M as shown in Figure 7.17(b). Howeve when the input reaches a logic 1 voltage, then the pFET Mp turns o once again preventing a direct current flow path. If we assume that th inputs are in stable 0 or 1 states as in an idle system, the DC power diss pation is given by

$$P_{DC} = V_{DD} I_{DDQ} \tag{7.8}$$

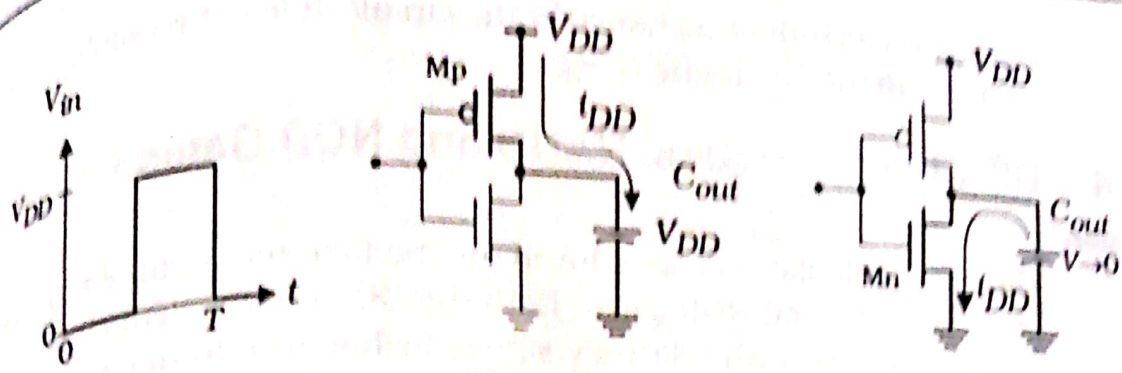
The leakage current I_{DDQ} is usually quite small, with a typical value o the order of a picoampere per gate. The value of P_{DC} is thus quite smal This consideration was a major factor in the move to CMOS in the mid 1990's.

To find the dynamic power dissipation P_{dyn}, we use a square-wav input voltage $V_{in}(t)$ as shown in Figure 7.18(a). The waveform has a perio T corresponding to a switching frequency of

$$f = \frac{1}{T} \tag{7.8}$$

with units of Hertz; the frequency is the number of cycles completed i one second. During the first half-cycle, the input voltage is at a value V_{in} = 0. This turns on the pFET Mp as shown in Figure 7.18(b). Since th nFET is off, the current i_{DD} flows through Mp and charges C_{out} to a voltag of $V_{out} = V_{DD}$. During the second half-cycle, the input voltage is high, turn ing on the nFET Mn. This causes the discharge event illustrated in Figur

[4] These are discussed in more detail in Chapter 9.



(a) Input voltage (b) Charge (c) Discharge

Figure 7.18 Circuit for finding the transient power dissipation

7.18(c) where V_{out} decays to 0 V. The dynamic power P_{dyn} arises from the observation that a complete cycle effectively creates a path for current to flow from the power supply to ground: during the charge event, current flows to the capacitor C_{out} while the discharge path to ground completes the circuit.

To calculate P_{dyn}, we note that the charging event leaves C_{out} with a voltage of $V_{out} = V_{DD}$. This corresponds to a stored electric charge on the capacitor of

$$Q_e = C_{out} V_{DD} \tag{7.89}$$

which has units of coulombs. When the capacitor is discharged through the nFET, the same amount of charge is lost. The average power dissipated over a single cycle with a period T is

$$P_{av} = V_{DD} I_{DD} = V_{DD}\left(\frac{Q_e}{T}\right) \tag{7.90}$$

Substituting for Q_e gives

$$P_{sw} = C_{out} V_{DD}^2 f \tag{7.91}$$

as the switching power. Combining the DC and dynamic power terms gives the total power as

$$P = V_{DD} I_{DDQ} + C_{out} V_{DD}^2 f \tag{7.92}$$

which will usually be dominated by the dynamic term. This illustrates an extremely important point:

- The dynamic power dissipation is proportional to the signal frequency

In other words, a fast circuit dissipates more power than a slow circuit. If we double the switching speed, then the dynamic power dissipation doubles. These are simply statements of the physical law that we must pro-

vide energy to induce a change in the circuit. It is not possible to switch a circuit without expending energy.

7.4 DC Characteristics: NAND and NOR Gates

The basic calculations introduced for the inverter circuit can be used to analyze NAND and NOR gates. Both the DC and transient characteristics can be obtained with relatively simple techniques. In this section we will examine the relationship between device sizes and the transitions described by the VTC.

7.4.1 NAND Analysis

Let us start with the NAND2 gate illustrated in Figure 7.19. We will analyze the case where like-polarity FETs have the same aspect ratio. This means that both pFETs are described by β_p and both nFETs have the same β_n. Since the pFETs are in parallel while the nFETs are in series, the circuit behaves quite differently from the simple inverter.

The presence of two independent inputs implies that more than one VTC curve is needed to describe the circuit. Suppose that we look for transitions where V_{out} is initially high at V_{DD} and then falls to 0 V when inputs are changed. Figure 7.20(a) summarizes the possible starting points that can lead to this situation. In case (i), both V_A and V_B are at 0 V and then switched to the bottom line condition where $V_A = V_B = V_{DD}$ such that $V_{out} = 0$ V. Since both inputs are increased at the same time, this describes the case for simultaneous input switching. The other two possibilities (ii) and (iii) describe cases where only a single input is changed. For example, in (ii) V_A is changed from 0 V to V_{DD} while V_B is held constant at V_{DD}. These three possibilities lead to the three distinct transitions shown in the plot of Figure 7.20(b). This shows that the simultaneous switching case is "pushed to the right" compared to the single-switched input cases.

It is instructive to calculate the value of the midpoint voltage V_M for the

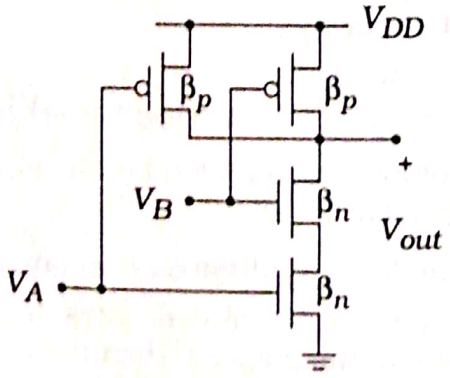
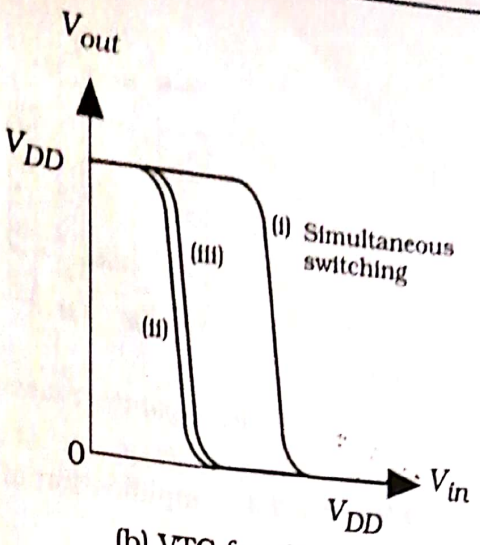


Figure 7.19 NAND2 logic circuit

	V_A	V_B	V_{out}
(i)	0	0	V_{DD}
(ii)	0	V_{DD}	V_{DD}
(iii)	V_{DD}	0	V_{DD}
	V_{DD}	V_{DD}	0

(a) Transition table



(b) VTC family

Figure 7.20 NAND2 VTC analysis

case of simultaneous switching using layout drawings. The circuit problem is illustrated in Figure 7.21, where W_n and W_p are the nFET and pFET channel widths, respectively. All transistors are assumed to have the same channel length L. Now then, for this case both input voltages V_A and V_B are equal to V_M. On the layout plot, both gates are thus at the same potential and can be connected to simplify the calculations.

Consider the nFETs first. In Figure 7.22(a), the layout is shown in its original form with two separate series-connected transistors. Let us "merge" the two gates together into one to obtain the patterning shown in Figure 7.22(b). If we ignore the n+ region that separates the two gates, then the structure can be approximated as a single nFET with an aspect ratio of $(W_n/2L)$ as shown. Since the original nFETs each had a device transconductance of β_n, the single equivalent transistor is described by the value $(\beta_n/2)$.

The pFETs can be combined in a similar manner. The original parallel-connected transistors are illustrated in Figure 7.23(a). Owing to the paral-

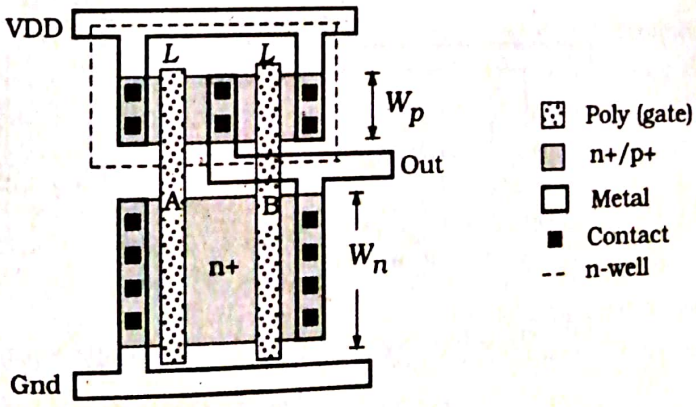
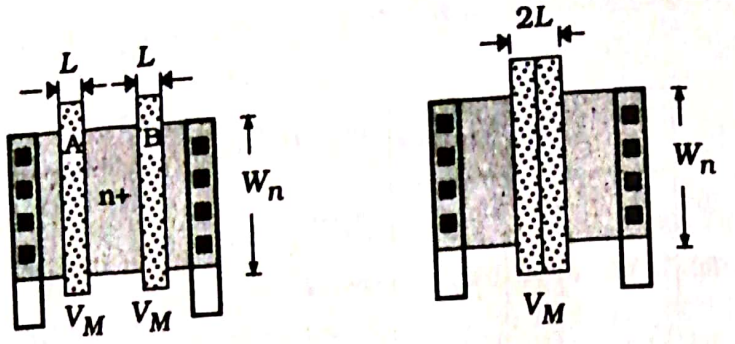


Figure 7.21 Layout of NAND2 for V_M calculation



(a) Separate transistors (b) Single equivalent FET

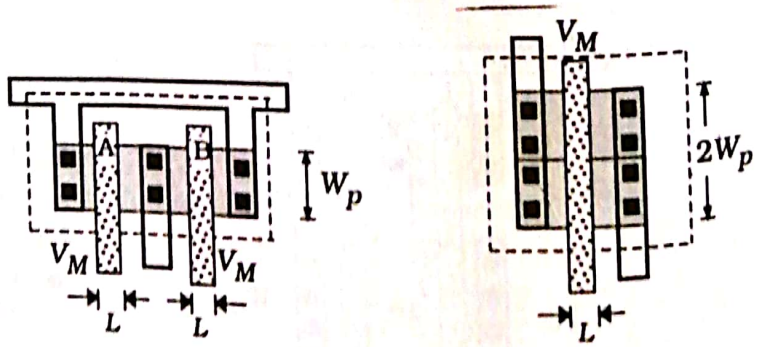
Figure 7.22 Simplification of the series-connected nFETs

lel wiring, the left and right sides are electrically the same point, so that the two may be simplified into the single gate structure shown in Figure 7.23(b). In this case, the two combine to act as a single pFET with an aspect ratio of $(2W_p/L)$. If the original devices each have β_p, then the equivalent structure acts as a pFET with $2\beta_p$.

Let us now use these results to find V_M for the case of simultaneous switching. Replacing the transistor pairs by their single-FET equivalents gives the inverter circuit in Figure 7.24, where the nFET and pFET transconductances are ($\beta_n/2$) and $2\beta_p$, respectively. The calculation then proceeds in the same manner as for the "normal" NOT gate. Both transistors are saturated, so equating currents gives

$$\frac{(\beta_n/2)}{2}(V_M - V_{Tn})^2 = \frac{(2\beta_p)}{2}(V_{DD} - V_M - |V_{Tp}|)^2 \qquad (7.93)$$

Taking square roots of both sides and solving for the midpoint voltage results in the expression



(a) Separate transistors (b) Single equivalent FET

Figure 7.23 Simplification of parallel-connected pFETs

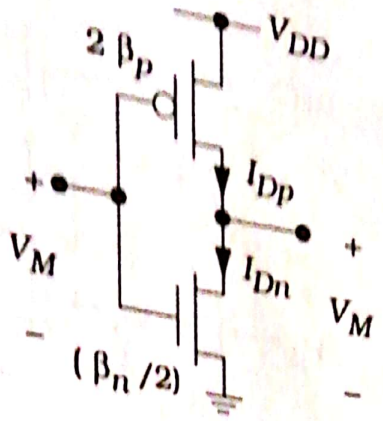


Figure 7.24 Simplified V_M circuit for the NAND2 gate

$$V_M = \frac{V_{DD} - |V_{Tp}| + \frac{1}{2}\sqrt{\frac{\beta_n}{\beta_p}}V_{Tn}}{1 + \frac{1}{2}\sqrt{\frac{\beta_n}{\beta_p}}} \tag{7.94}$$

This has the same form as the NOT gate in equation (7.14), except that the square root term is multiplied by a factor of (1/2). This reduces the denominator, which is why the VTC curve is shifted toward the right. If we apply the same reasoning to an *N*-input NAND gate, the simultaneous switching point is found to be

$$V_M = \frac{V_{DD} - |V_{Tp}| + \frac{1}{N}\sqrt{\frac{\beta_n}{\beta_p}}V_{Tn}}{1 + \frac{1}{N}\sqrt{\frac{\beta_n}{\beta_p}}} \tag{7.95}$$

The right shift is due to the series-connected nFETs, since their resistances add.

7.4.2 NOR Gate

The NOR2 gate can be analyzed using the same techniques. We assume that the nFETs have the same β_n and that both pFETs are described by β_p as shown in the basic circuit of Figure 7.25. To construct VTC, note that $V_{out} = V_{DD}$ requires that $V_A = V_B = 0$ V. If either input (or both) are switched to logic 1 values, then the output will fall to $V_{out} = 0$ V. The three combinations are listed in the function table of Figure 7.26(a). As with the NAND2 gate, there are three distinct transitions shown in the VTC family of Figure 7.26(b). Case (i) describes the simultaneous switching event where both V_A and V_B are increased from 0 V toward V_{DD}. This case is the leftmost plot in the VTC family, exactly opposite to that found for the NAND2. Single-input switching cases (ii) and (iii) are distinct, but are close to each other.

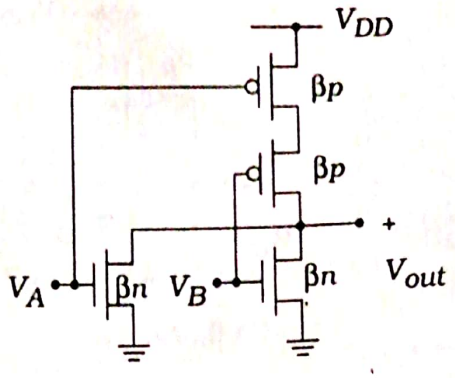


Figure 7.25 NOR2 circuit

The techniques of combining series and parallel transistors may be used to compute V_M for the simultaneous switching case. Since the nFETs are in parallel, they may be combined to a single equivalent nFET with a transconductance of $2\beta_n$. The series-connected pFETs act as a single pFET with $(\beta_p/2)$ which gives rise to the simplified equivalent circuit in Figure 7.27. Equating the saturation currents using the effective transconductance values gives us

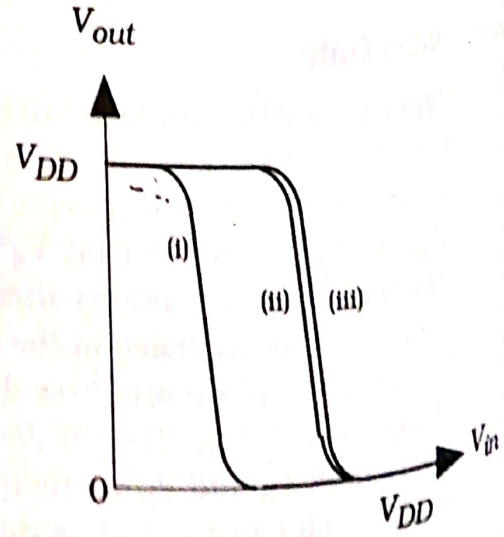
$$\frac{(2\beta_n)}{2}(V_M - V_{Tn})^2 = \frac{(\beta_p/2)}{2}(V_{DD} - V_M - |V_{Tp}|)^2 \tag{7.96}$$

This may be solved to give

$$V_M = \frac{V_{DD} - |V_{Tp}| + 2\sqrt{\frac{\beta_n}{\beta_p}}V_{Tn}}{1 + 2\sqrt{\frac{\beta_n}{\beta_p}}} \tag{7.97}$$

	V_A	V_B	V_{out}
	0	0	V_{DD}
(iii)	0	V_{DD}	0
(ii)	V_{DD}	0	0
(i)	V_{DD}	V_{DD}	0

(a) Transition table



(b) VTC family

Figure 7.26 NOR2 VTC construction

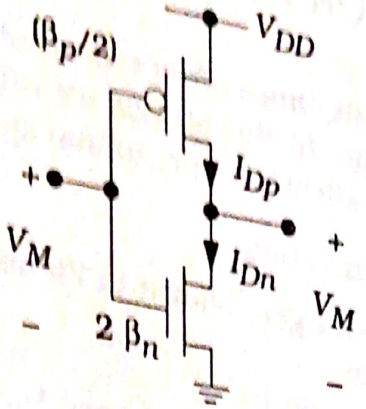


Figure 7.27 NOR2 V_M calculation for simultaneous switching

Comparing this with the NOT and NAND expressions shows that the only difference is the factor of 2 multiplying the square root term. This increases the denominator, which decreases the value of V_M from that of an inverter with a device ratio of (β_n/β_p). The midpoint voltage for an N-input NOR gate is

$$V_M = \frac{V_{DD} - |V_{Tp}| + N\sqrt{\frac{\beta_n}{\beta_p}}V_{Tn}}{1 + N\sqrt{\frac{\beta_n}{\beta_p}}} \tag{7.98}$$

It is worthwhile noting that the NAND and NOR gates tend have opposite behaviors with respect to the reference NOT gate VTC.

As a final comment, we note that both the NAND and NOR gates exhibit low DC power dissipation values of

$$P_{DC} = V_{DD}I_{DDQ} \tag{7.99}$$

since there is no direct current flow path from the power supply to ground when the inputs are stable logic 0 or logic 1 values. The low power characteristic of the gates is due to the use of complementary pairs and series-parallel structuring of the transistor arrays. Dynamic power is still present in the general form

$$P_{sw} = C_{out}V_{DD}^2 f_{gate} \tag{7.100}$$

which shows the dependence on gate switching frequency f_{gate}. Since it takes more than a single input to switch the gate, f_{gate} is different from the basic switching frequency used for the inverter. This is discussed in more detail later.

7.5 NAND and NOR Transient Response

Transient switching times often represent the limiting factor in designing a digital logic chain. In this section we will examine how the FET topology and device sizing affect the operational speed of the gate.

7.5.1 NAND2 Switching Times

Consider the NAND2 gate shown in Figure 7.28. The total output capacitance is denoted as

$$C_{out} = C_{FET} + C_L \tag{7.101}$$

where C_L is the external load and

$$C_{FET} = C_{Dn} + 2C_{Dp} \tag{7.102}$$

represents the parasitic internal FET capacitances. Note that there are two contributions of C_{Dp} since two pFETs are connected to the output node. The drawing identifies the transistors by their resistance values

$$R_p = \frac{1}{\beta_p(V_{DD} - |V_{Tp}|)}, \qquad R_n = \frac{1}{\beta_n(V_{DD} - V_{Tn})} \tag{7.103}$$

The transient calculations are based on finding RC time constants for the charge time (t_r or t_{LH}) and fall time (t_f or t_{HL}) for the transitions. The procedure is complicated by the presence of two inputs. We will concentrate on estimating the worst-case values of the switching times.

Let us consider the rise time t_r first. The output voltage is initially at a value $V_{out}(0) = 0$ V and is then charged to V_{DD}. If only one pFET is conducting, we obtain the simplified charging circuit shown in Figure 7.29(a) where C_{out} charges through a pFET resistance R_p. Since this looks like the charging circuit for a simple inverter, we can write

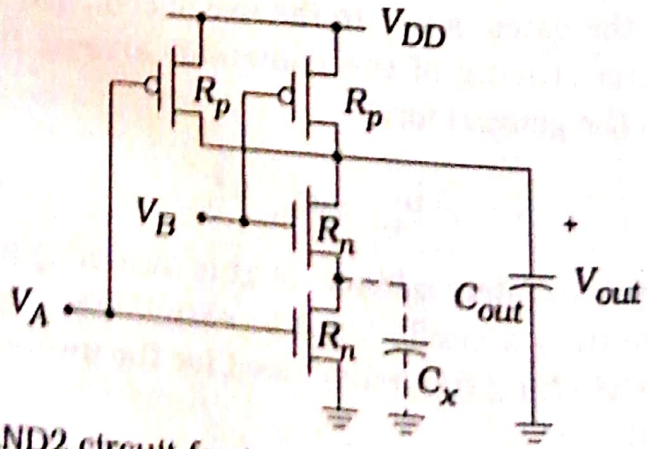
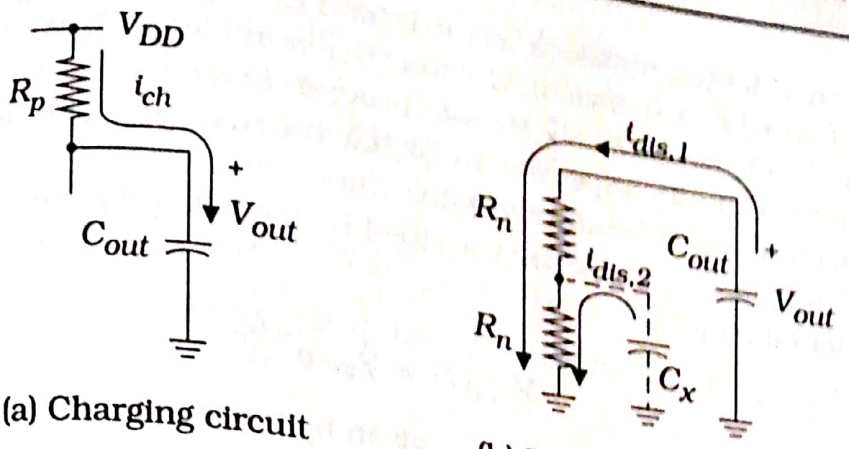


Figure 7.28 NAND2 circuit for transient calculations



(a) Charging circuit

(b) Discharging circuit

Figure 7.29 NAND2 subcircuits for estimating rise and fall times

$$V_{out}(t) = V_{DD}[1 - e^{-t/\tau_p}] \tag{7.104}$$

where

$$\tau_p = R_p C_{out} \tag{7.105}$$

is the time constant. The rise time is thus given by

$$t_r \approx 2.2\tau_p \tag{7.106}$$

This is considered to be a "worst-case" situation since only one pFET is charging C_{out}. Note that this can be cast into the linear form

$$t_r = t_0 + \alpha_0 C_L \tag{7.107}$$

where

$$t_0 = 2.2 R_p C_{FET} \tag{7.108}$$

is the zero-load value, and

$$\alpha_0 = 2.2 R_p \tag{7.109}$$

is the slope of t_r as a function of the load capacitance C_L. If both pFETs are conducting, then the equivalent resistance is lowered to $(R_p/2)$ since the two are in parallel; this would be the "best-case" event, i.e., the one with the shortest charging time. Design is usually based on worst-case analysis since we want to insure that the circuit operates under all conditions.

The situation is more complicated when we analyze the fall time t_f where C_{out} discharges through the series-connected nFET chain. RC modeling of each device leads to the "ladder" network shown in Figure 7.29(b). While the main item of interest is discharging C_{out}, the situation is com-

plicated by the presence of the inter-FET capacitance C_X between the two n-channel transistors. In the worst-case analysis, C_X will have charge that will flow through nFET MnA to ground. Since the current through a FET is limited by its aspect ratio (W/L), the discharge rate is limited by the current that MnA can maintain.

The discharge can be described by modeling the output voltage in the exponential form

$$V_{out}(t) = V_{DD}e^{-t/\tau_n} \tag{7.110}$$

such that the time constant is given by the **Elmore formula** as

$$\tau_n = C_{out}(R_n + R_n) + C_X R_n \tag{7.111}$$

This estimates the time constant as the superposition of time constants

$$\tau_n = \tau_{n1} + \tau_{n2} \tag{7.112}$$

where

$$\tau_{n1} = C_{out}(R_n + R_n) \tag{7.113}$$

is the time constant for C_{out} discharging through two nFETs, each with a resistance R_n; this is shown by the current $i_{dis,1}$ in the drawing. The other term

$$\tau_{n2} = C_X R_n \tag{7.114}$$

is the time constant for C_X discharging through one nFET with a resistance R_n. This is corresponds to the discharge current $i_{dis,2}$. The fall time t_f is then given by

$$t_f \approx 2.2\tau_n \tag{7.115}$$

Substituting the time constant expression transforms this into

$$t_f \approx 2.2[(C_{FET} + C_L)(2R_n) + C_X R_n] \tag{7.116}$$

Grouping terms results in the linear expression

$$t_f = t_1 + a_1 C_L \tag{7.117}$$

with a zero-load delay of

$$t_1 = 2.2R_n(2C_{FET} + C_X) \tag{7.118}$$

and a slope of

$$a_1 = 4.4R_n \tag{7.119}$$

where the multiplier is from (2×2.2). Although we are able to write $t_f = a_0 + a_1 C_L$

linear function of C_L, both the zero-load delay and the slope are affected by the series-connected nFETs in the discharge circuitry.

The Elmore formulation of time constants for RC ladder-type networks illustrates that series-connected FETs lead to longer delays in CMOS circuits. To understand this comment, let us rewrite equation (7.111) as

$$\tau_n = R_n(2C_{out} + C_X) \tag{7.120}$$

In this form, we can interpret the time constant as R_n multiplying an effective capacitance with a value

$$C_{eff} = 2C_{out} + C_X \tag{7.121}$$

which is larger than twice the output capacitance. Alternately, we may write

$$\tau_n = C_{out}(2R_n) + C_X R_n \tag{7.122}$$

which clearly shows the effect of the series-connected FETs in the term $2R_n$ and the increase due to the parasitic capacitance C_X. Regardless of the interpretation one chooses, it is important to remember that series-connected FET chains can lead to excessive logic delays.

52 NOR2 Switching Times

The analysis of the NOR2 transients proceeds in the same manner. Figure 7.30 shows the circuit with FET resistances and the capacitances. The output capacitance for any gate is given by the general form

$$C_{out} = C_{FET} + C_L \tag{7.123}$$

For the NOR2 circuit, the internal capacitance can be broken down into components as

$$C_{FET} = 2C_{Dn} + C_{Dp} \tag{7.124}$$

since there are two nFETs connected to the output node but only one

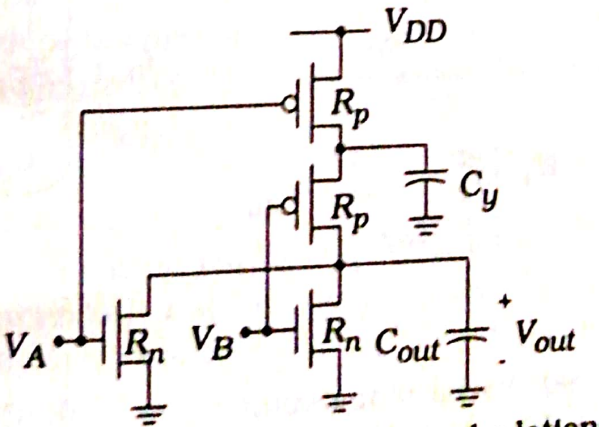


Figure 7.30 NOR2 circuit for switching time calculations

pFET. The inter-FET capacitance C_y represents the parasitic contribu-
tions between the two pFETs.

Figure 7.31 shows the subcircuits for the output transients. The fall
time t_f may be computed using the worst-case circuit in Figure 7.31(a)
where only one nFET acts to discharge the output capacitance. We then
write the output voltage as

$$V_{out}(t) = V_{DD}e^{-t/\tau_n} \tag{7.125}$$

with

$$\tau_n = R_n C_{out} \tag{7.126}$$

as the time constant. The fall time is then given by

$$t_f \approx 2.2\tau_n \tag{7.127}$$

which is identical to that for a simple inverter. Expanding C_{out} gives the
linear dependence

$$t_f = t_1 + \alpha_1 C_L \tag{7.128}$$

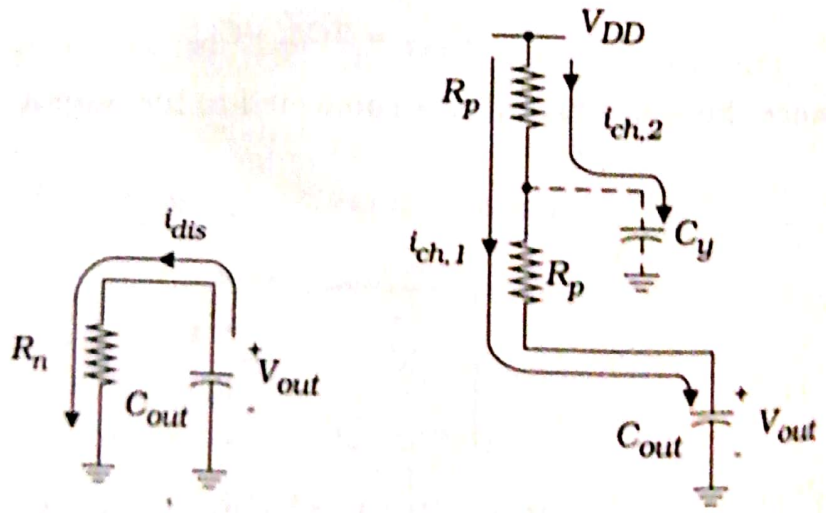
where the zero-load delay is

$$t_1 = 2.2R_n C_{FET} \tag{7.129}$$

and the slope is

$$\alpha_1 = 2.2R_n \tag{7.130}$$

These results are similar to the NOT gate, but it is important to remember
that C_{FET} is larger for the NOR2 gate.



(a) Discharging circuit (b) Charging circuit

Figure 7.31 Subcircuits for the NOR2 transient calculations

The charging circuit for finding the rise time t_r is shown in Figure 7.31(b). We will write the output voltage in the exponential form

$$V_{out}(t) = V_{DD}[1 - e^{-t/\tau_p}] \tag{7.131}$$

However, since C_y will be charged during this event, we must use the Elmore formula to find the time constant. The two paths are shown as $i_{ch,1}$ and $i_{ch,2}$ in the drawing. The primary charge path due to $i_{ch,1}$ is described by a time constant

$$\tau_1 = C_{out}(R_p + R_p) \tag{7.132}$$

while that associated with $i_{ch,2}$ is

$$\tau_2 = C_y R_p \tag{7.133}$$

Superposing gives the total effective time constant in the form

$$\begin{aligned} \tau_p &= \tau_1 + \tau_2 \\ &= C_{out}(2R_p) + C_y R_p \end{aligned} \tag{7.134}$$

such that the rise time is

$$t_r = 2.2\tau_p \tag{7.135}$$

Since the series-connected pFETs introduce a large time constant, the rise time may be quite large compared to the fall time. Substituting for C_{out} gives the linear equation

$$t_r = t_0 + \alpha_0 C_L \tag{7.136}$$

where

$$t_0 = 2.2R_p(2C_{FET} + C_y) \tag{7.137}$$

and

$$\alpha_0 = 4.4R_p \tag{7.138}$$

characterize the dependence of t_r on C_L. As with the NAND2 gate, the presence of series-connected FETs slows down the associated switching time.

7.5.3 Summary

The analyses above illustrate that the NAND and NOR gates exhibit complementary characteristics at both the DC and transient levels. This arises because they are constructed using complementary series-parallel transistor arrangements.

While the DC characteristics are important, most design effort is

directed toward minimizing delays through logic chains. The study abo allows us to make some general statements about NAND and NOR gate as compared to the simpler NOT circuit. First, we have seen that the ris time can be written in the form

$$t_r = t_0 + \alpha_0 C_L \tag{7.13}$$

while the fall time has the same structure with

$$t_f = t_1 + \alpha_1 C_L \tag{7.14}$$

The constants (t_0 and α_0 for the rise time, and t_1 and α_1 for the fall time) depend upon the parasitic transistor resistances and capacitances. The constants are the smallest for a NOT gate, so we often use it as a reference. This, of course, is because the inverter consists of only two FETs. In general, adding complementary transistor pairs increases the delay times because C_{FET} is increased. The number of inputs to a logic gate is called the **fan-in** (FI). Since every input is connected to a complementary pair, we can state that

- Switching delays increase with the fan-in.

This says, for example, a NAND3 gate will be slower than a NAND2 gate if the two use the same size transistors. Of course, the actual delay depends upon the value of the load capacitance C_L such that

- Switching delays increase with the external load.

Since logic functions are implemented using cascades of gates, the effect of this dependence varies with the circuit.

Let us summarize the results of the NAND and NOR analysis. As with the inverter, the electrical characteristics of these gates are set by

- The processing variables and
- The aspect ratios $(W/L)_n$ and $(W/L)_p$ of every FET

Furthermore, series transistors introduced us to the problem of parasitic capacitance between the two devices. This factor leads us to make one additional statement

- The details of the layout geometries affect the transient response of the logic gate.

We thus conclude that the physical layout and structure of the circuitry is a critical factor in designing high-speed logic networks.

7.6 Analysis of Complex Logic Gates

The analysis techniques developed for the NAND and NOT circuits may be extended to analyze complex CMOS logic gates with AOI and OAI structuring. The most important problem is the transient delay associated with

series-connected FETs.

Consider the complex logic gate shown in Figure 7.32. This implements the logic function

$$f = \overline{x \cdot (y + z)} \tag{7.141}$$

with series-parallel FET arrays. The aspect ratio values shown in the drawing are the critical parameters that affect the rise and fall times. The fall time is governed by the nFETs. If we assume that they are all the same size with

$$\left(\frac{W}{L}\right)_{nx} = \left(\frac{W}{L}\right)_{ny} = \left(\frac{W}{L}\right)_{nz} \tag{7.142}$$

then the nFET resistance R_n can be used to describe each one. The worst-case fall time will occur when $x = 1$, but only one of the ORed inputs y or z is 1. This results in a 2-FET series pair that must handle the discharge of the output capacitor

$$C_{out} = C_{FET} + C_L \tag{7.143}$$

With the capacitance C_n in the chain, the time constant is

$$\tau_n = R_n C_n + 2R_n C_{out} \tag{7.144}$$

which gives a fall time of

$$\begin{aligned} t_f &= 2.2\tau_n \\ &= 2.2R_n[C_n + 2(C_{FET} + C_L)] \\ &= t_1 + \alpha_1 C_L \end{aligned} \tag{7.145}$$

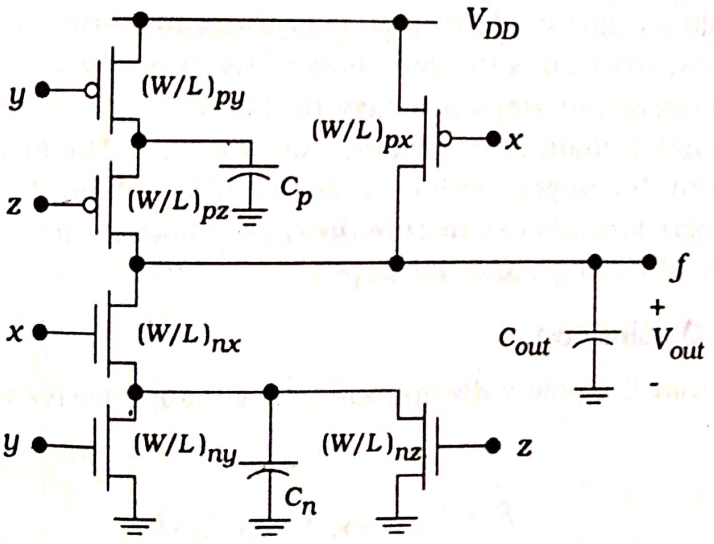


Figure 7.32 Complex logic gate circuit

where

$$t_1 = 2.2R_n(C_n + 2C_{FET}) \tag{7.14}$$

is the zero-load time, and

$$\alpha_1 = 2.2R_n \tag{7.14}$$

is the slope.

The rise time t_r is determined by the pFETs. If these are chosen with equal aspect ratios

$$\left(\frac{W}{L}\right)_{px} = \left(\frac{W}{L}\right)_{py} = \left(\frac{W}{L}\right)_{pz} \tag{7.14}$$

then we can use the same R_p for each device. The limiting series chain is with the y and z input p-channel transistors; the x-input pFET provides the fast switching, and could be decreased to half-size without affecting the results. The series chain gives a time constant of

$$\tau_p = R_pC_p + 2R_pC_{out} \tag{7.149}$$

where C_p is the parasitic capacitance between the pFETs. The worst-case rise time is thus of the form

$$t_r = t_0 + \alpha_0 C_L \tag{7.150}$$

where the zero-load delay is

$$t_0 = 2.2R_p(C_p + 2C_{FET}) \tag{7.151}$$

and the slope is

$$\alpha_0 = 2.2R_p \tag{7.152}$$

An arbitrary gate yields equations of the same form for both the rise and fall times, illustrating the generality of the procedure.

The important steps are easy to follow. Find the longest series-connected nFET chain for the worst-case fall time. The longest rise time will be due to the longest series-connected pFET chain. For both cases, use the Elmore formula to compute the time constant, then separate terms for the zero bias delays and the slopes.

7.6.1 Power Dissipation

Recall that the power dissipation in a simple inverter was written in the form

$$P = V_{DD}I_{DDQ} + C_{out}V_{DD}^2 f \tag{7.153}$$

When we analyze a general static CMOS logic gate, the DC term is still small, but the dynamic switching power P_{dyn} becomes important in high-speed, high-density designs.

To model the dynamic power dissipation of an arbitrary gate we recall that P_{dyn} originates from an output switching event. First, the output capacitor C_{out} is charged from 0 V to V_{DD}, corresponding to an output logic 0→1 transition. Then, C_{out} discharges to give a 1→0 transition, completing the cycle. To model the number of transitions that take place over a switching period T we introduce the **activity coefficient** a that represents the probability that an output 0→1 transition takes place during one period. The dynamic power is then modified to read

$$P_{dyn} = aC_{out}V_{DD}^2 f \tag{7.154}$$

For a network that consists of N gates, the total dynamic power is more generally written in the form

$$P_{dyn} = \sum_{i=1}^{N} a_i C_i V_i V_{DD} f \tag{7.155}$$

where, for the i-th gate, a_i is the activity coefficient and C_i is the node capacitance that charges to a maximum value of V_i.

Activity coefficients can be determined from truth tables. Figure 7.33 provides the truth tables for the NOR2 and NAND2 functions. We will assume that each input combination has equal probability of occurring. Let us analyze the NOR2 transitions first. Since the activity factor a_{NOR2} is the probability that the gate makes a 0 → 1 transition, it can be calculated by

$$a = p_0 p_1 \tag{7.156}$$

where p_0 is the probability that the output is initially at 0, and p_1 the probability that it makes a transition to 1. The truth table shows us that $p_0 = (3/4)$ and $p_1 = (1/4)$, so

A	B	$\overline{A+B}$	$\overline{A \cdot B}$
0	0	1	1
0	1	0	1
1	0	0	1
1	1	0	0

Figure 7.33 Truth tables for determining activity coefficients

$$a_{NOR2} = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16} \tag{7.157}$$

The NAND2 gate can be analyzed in the same manner. For this gate, the truth table shows that $p_0 = (1/4)$ and $p_1 = (3/4)$ so

$$a_{NAND2} = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16} \tag{7.158}$$

has the same value as the NOR2 gate. If we look at 3-input gates, the truth tables give

$$a_{NOR3} = \frac{7}{64} = a_{NAND3} \tag{7.159}$$

Similarly, we can calculate

$$a_{XNOR2} = \frac{1}{4} = a_{XOR2} \tag{7.160}$$

since $p_0 = (1/4) = p_1$. The technique can be applied to an arbitrary gate.

The limit on this simple treatment is that, in practice, we rarely have input combinations that occur with equal probability. More advanced techniques have been developed to handle these situations. The interested reader is directed to Reference [2] for an excellent discussion of the details. Reference [8] is a very thorough analysis of power dissipation and low-power design.

7.7 Gate Design for Transient Performance

High-speed circuits are limited by the switching time of individual gates. Logic formation determines the series and parallel connections of the transistors. The aspect ratios are the critical design parameters for both the DC and transient switching times. Once these are specified and the transistors are created in the layout, all of the parasitics are set.

The DC switching characteristics are often considered less important than the switching speed. It is common to design a gate to have the desired transient times, and then check the DC VTC to insure that it is acceptable. This approach is based on the fact that the individual nFET and pFET aspect ratios determine the switching response, while the DC transition point is a result of the ratio of the nFET to pFET values. For example, the value of β_n/β_p gives V_M for an inverter, while t_r depends primarily on β_p and t_f is established by β_n.

The design philosophy used to select aspect ratios varies with the situation. A straightforward approach is to use the inverter as a reference and then attempt to design other gates that have approximately the same switching times. Since the NOT gate is the simplest, it can be built using

relatively small transistors. We will use the device transconductance

$$\beta = k'\left(\frac{W}{L}\right) \tag{7.161}$$

as being equivalent to the aspect ratio.

Figure 7.34(a) shows an inverter with device sizes specified by β_p and β_n, which we will assume are known. These set the rise and fall times t_r and t_f for the circuit, which serve as the reference switching times. Since both transistors drive the same capacitance, the difference is in the resistance values

$$R_p = \frac{1}{\beta_p(V_{DD} - |V_{Tp}|)}, \qquad R_n = \frac{1}{\beta_n(V_{DD} - V_{Tn})} \tag{7.162}$$

Recall that a symmetrical inverter has

$$\beta_n = \beta_p \tag{7.163}$$

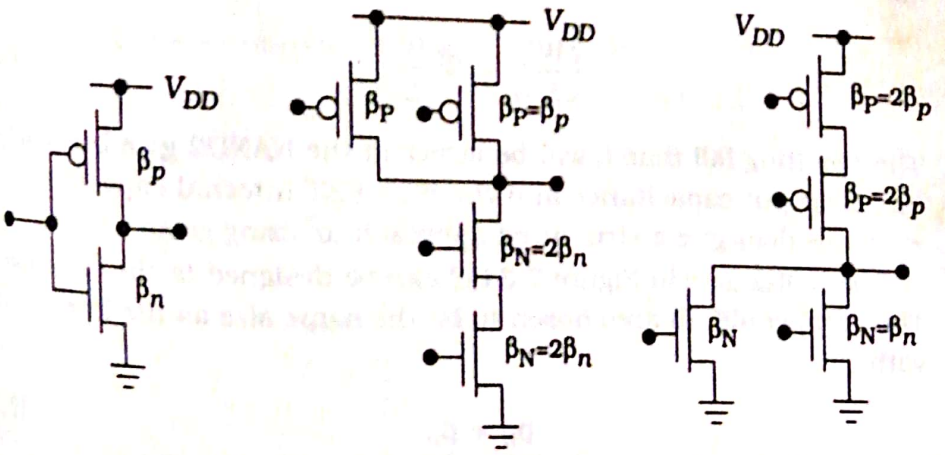
and requires the device sizes to be related by

$$\left(\frac{W}{L}\right)_p = r\left(\frac{W}{L}\right)_n \tag{7.164}$$

where

$$r = \frac{k'_n}{k'_p} \tag{7.165}$$

is the process transconductance ratio. A nonsymmetrical design that uses equal size transistors such that $\beta_n > \beta_p$ is also commonly used as a reference.



(a) Inverter (b) NAND2 (c) NOR2

Figure 7.34 Relative FET sizing

Let us use these values to find the device sizes β_P and β_N for the NAND2 gate in Figure 7.34(b) with the philosophy that we want to achieve symmetrical rise and fall times. Consider first the parallel pFETs. Since the worst-case situation is where only one transistor contributes to the rise time, we shall select the same size as the inverter:

$$\beta_P = \beta_p \tag{7.168}$$

The actual rise time t_r will be longer than that of the inverter because C_{out} is larger. The series-connected nFET chain has to be modeled as having series-connected resistors between the output and ground, with a total value of

$$R = R_N + R_N \tag{7.167}$$

where

$$R_N = \frac{1}{\beta_N(V_{DD} - V_{Tn})} \tag{7.169}$$

Using the inverter as a reference, we set

$$R = R_n = 2R_N \tag{7.169}$$

Substituting,

$$\frac{1}{\beta_n(V_{DD} - V_{Tn})} = \frac{2}{\beta_N(V_{DD} - V_{Tn})} \tag{7.170}$$

which has the solution

$$\beta_N = 2\beta_n \tag{7.171}$$

i.e., the series-connected nFETs are twice as large as the inverter transistor:

$$\left(\frac{W}{L}\right)_N = 2\left(\frac{W}{L}\right)_n \tag{7.172}$$

The resulting fall time t_f will be larger in the NAND2 gate because of the larger output capacitance and the FET-FET internal capacitance. However, this does give a structured approach to sizing gates.

The NOR2 gate in Figure 7.34(c) can be designed in the same manner. The parallel nFETs are chosen to be the same size as the inverter device with

$$\beta_N = \beta_n \tag{7.173}$$

since this gives the worst-case discharge. The series-connected pFET resistances add to a total of $2R_p$. Equating this to the inverter resistance

k_p gives

$$\frac{1}{\beta_p(V_{DD}-|V_{Tp}|)} = \frac{2}{\beta_P(V_{DD}-|V_{Tp}|)} \tag{7.174}$$

so that

$$\beta_P = 2\beta_p \tag{7.175}$$

indicating that the pFETs are twice as large as the inverter transistors:

$$\left(\frac{W}{L}\right)_P = 2\left(\frac{W}{L}\right)_p \tag{7.176}$$

The main problem is that pFETs are intrinsically slow, so that the value of $(W/L)_p$ may be large to begin with.

This technique can be extended to larger chains. For n series-connected FETs, the size must be n times larger than the inverter value. The NAND3 gate in Figure 7.35(a) would thus be designed with

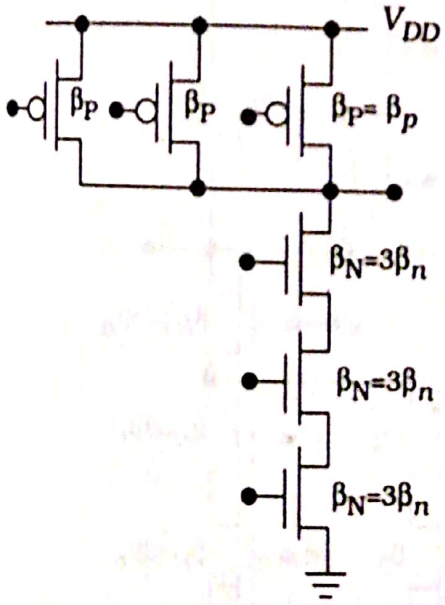
$$\beta_N = 3\beta_n, \qquad \beta_P = \beta_p \tag{7.177}$$

such that

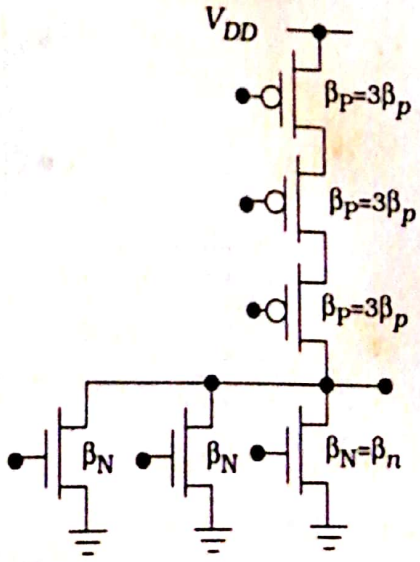
$$\left(\frac{W}{L}\right)_N = 3\left(\frac{W}{L}\right)_n, \qquad \left(\frac{W}{L}\right)_P = \left(\frac{W}{L}\right)_p \tag{7.178}$$

while the NOR3 gate in Figure 7.35(b) would have

$$\beta_N = \beta_n, \qquad \beta_P = 3\beta_p \tag{7.179}$$



(a) NAND3



(b) NOR3

Figure 7.35 Sizing for 3-input gates

with

$$\left(\frac{W}{L}\right)_N = \left(\frac{W}{L}\right)_n, \qquad \left(\frac{W}{L}\right)_P = 3\left(\frac{W}{L}\right)_p \tag{7.1}$$

Since the reference values β_n and β_p are arbitrary, the sizes can be adjusted as needed to accommodate reasonable values. Also note that if we select a symmetric inverter design with $\beta_n = \beta_p$, then the resulting gates will also be approximately symmetric.

Complex logic gates can be designed in the same manner. Consider the gate in Figure 7.36 that has an output of

$$f = \overline{(a \cdot b + c \cdot d) \cdot x} \tag{7.1}$$

using series-parallel structuring. Consider the nFET array first. Any discharge event will have current flow through a minimum of three series-connected nFETs. The device sizes would all be the same with the value

$$\beta_N = 3\beta_n = \beta_{N1} \tag{7.182}$$

The pFET array is a little different. The worst-case charge path is through two series-connected transistors on the left side of the circuit. The sizes would be

$$\beta_P = 2\beta_p \tag{7.183}$$

for the pFETs in the inputs a, b, c, and d. The x-input pFET is alone, so

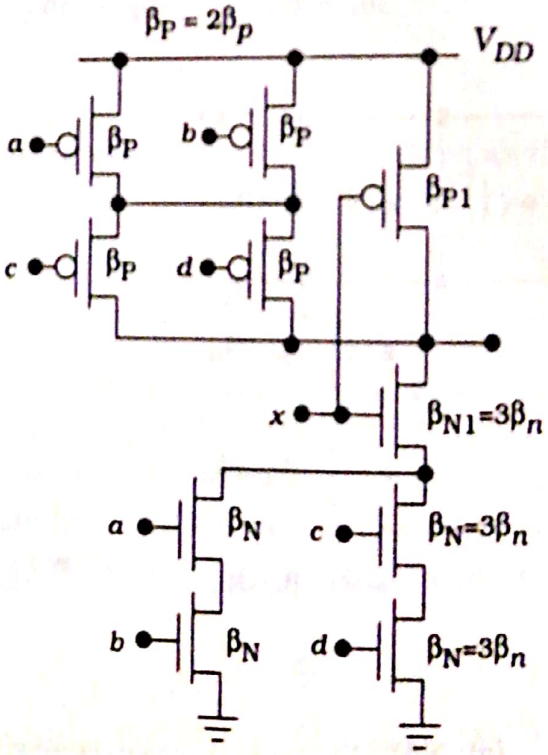


Figure 7.36 Sizing of a complex logic gate

that we can select its size as being the same as for an inverter:

$$\beta_{P1} = \beta_p \tag{7.184}$$

Alternately, the choice

$$\beta_{P1} = \beta_P = 2\beta_p \tag{7.185}$$

may lead to simpler layout since only a single size pFET would be used. Note that the two options for β_{P1} result in different input capacitances for the x-input.

Although this approach provides a nice structured methodology, it leads to large transistors. The designer must decide whether the real estate consumption is worth the added speed. This becomes more complicated as the number of FETs increases since the FET-to-FET parasitic capacitance terms in the Elmore time constant formula will also increase. In practice, we may just select a standard cell that meets the area allocation and then find the overall speed of the logic cascade. If the design is not fast enough, we can apply some of the techniques in the next chapter to find a better design.