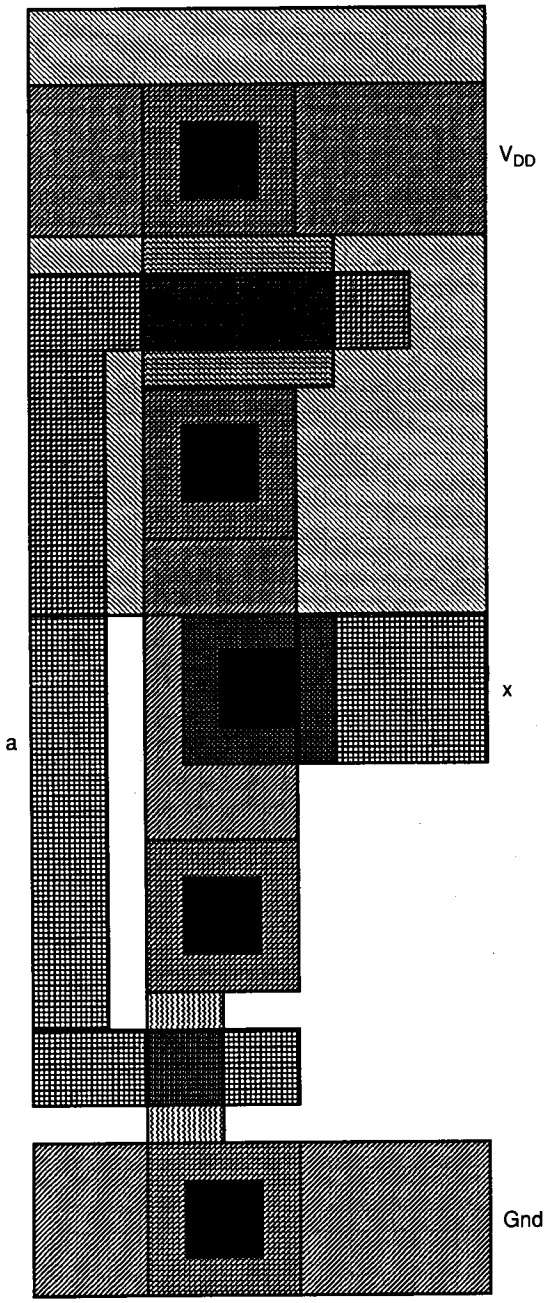

FIGURE 7.5-5
Layout for CMOS inverter with input a and output x.

7.6 CMOS LOGIC GATES

Classical CMOS logic gates are described in this section. As their characteristics are developed, a comparison with the corresponding NMOS logic gates will be given. Two-input NOR and NAND logic gates are used to develop the unique characteristics of CMOS logic. A discussion of multi-input CMOS gates will conclude the section.

7.6.1 CMOS NOR Logic Gate

Once the reference inverter circuit for a particular logic family is defined, the design of circuits to implement other logic functions can be based on that inverter circuit. Figure 7.6-1 shows the circuit schematic for a CMOS two-input positive logic NOR gate based on the CMOS inverter of Fig. 7.5-1. This circuit is obviously different from the NMOS NOR logic circuit of Fig. 7.4-1 because four transistors are required to implement the logic function rather than three. The two pulldown transistors, M1 and M2, are in parallel as they were for the corresponding NMOS logic gate. However, two pullup transistors, M3 and M4, are required in a series connection to complete the CMOS NOR circuit. The layout of a two-input CMOS NOR gate with all minimum-size transistors is shown in Fig. 7.6-2. From this figure, it can be observed that greater silicon area is required compared to an equivalent layout of the NMOS NOR gate of Fig. 7.4-1 given a common basic process resolution. The slight decrease in pullup transistor area for the CMOS gate—the longer depletion pullup of NMOS is unnecessary—is more than offset by the additional circuit connections, the second pullup transistor, and the n-well isolation required for the CMOS pullup transistors.

The operation of the CMOS NOR gate of Fig. 7.6-1 can be explained as follows. When both inputs a and b are below V_{TN}, the parallel n-channel pulldown transistors are off and the series p-channel pullup transistors conduct. This condition provides an active pullup through the series transistors for the output of the CMOS gate. There is no competition from the pulldown transistors.

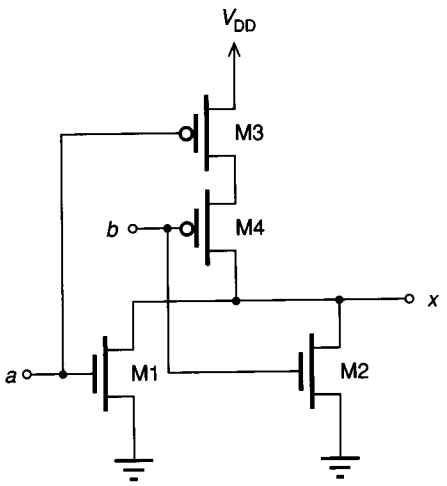

FIGURE 7.6-1
Two-input CMOS NOR gate.

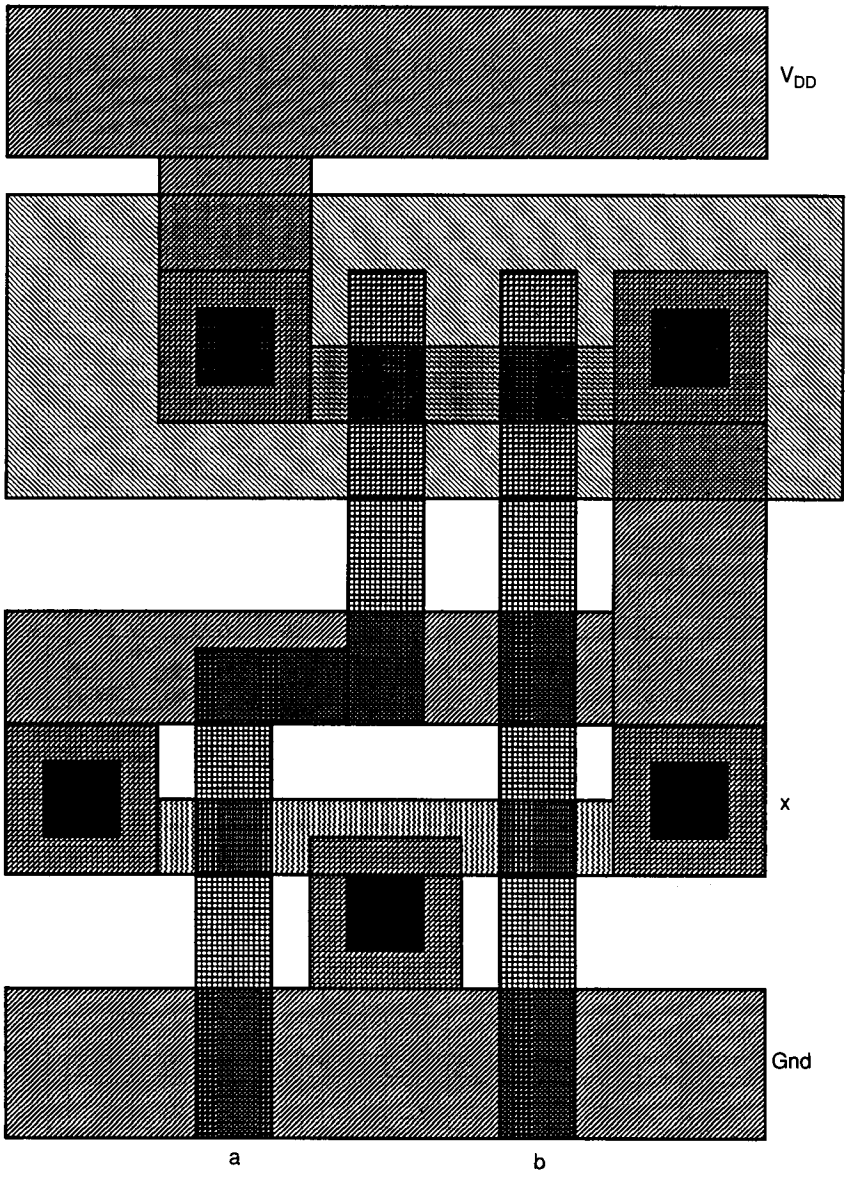

FIGURE 7.6-2
CMOS NOR gate layout with inputs a and b and output x.

Using the resistive model of Sec. 7.5 for the p-channel transistors, it follows that

$$R_{UP} = R_{P3} + R_{P4} \propto \frac{2L_P}{W_P K'_P} \tag{7.6-1}$$

Thus, for equally sized pullup transistors, the equivalent pullup resistance is twice that of either pullup alone.

If input a is connected to a voltage greater than $V_{DD} - |V_{TP}|$, p-channel pullup transistor M3 is off and n-channel pulldown transistor M1 conducts. Now the series path to V_{DD} is broken, and the output node is pulled to 0 V by transistor M1. This gives a pulldown resistance

$$R_{N1} \propto \frac{L_1}{W_1 K'_N} \tag{7.6-2}$$

A corresponding analysis for input b high, which turns M4 off, shows that the output node is pulled to 0 V by transistor M2. Now the pulldown resistance is

$$R_{N2} \propto \frac{L_2}{W_2 K'_N} \tag{7.6-3}$$

If both inputs a and b are above $V_{DD} - |V_{TP}|$, then both transistors M3 and M4 are off and transistors M1 and M2 both conduct. This parallel connection results in a pulldown resistance that is equal to the parallel combination of R_{N1} and R_{N2}, thus ensuring that the output node is pulled to 0 V. Because the output is pulled low if input a or b or both are high, this circuit realizes the NOR logic function.

For the CMOS NOR gate, as for the CMOS inverter, the steady state output voltage is set by a ratioless connection of conducting transistors. Thus, the dc logic voltages are ideal; that is, the output voltage is pulled to either V_{DD} or 0 V. Because dc logic levels are not affected by the relative sizes of the pullup and pulldown transistors, these transistors can be sized to obtain the desired output drive characteristics, as was done previously for the CMOS inverter. Because of the presence of two inputs (and four logic input conditions) for the NOR gate, it is not possible to maintain symmetric output drive capabilities for all input conditions. A common approach is to size the devices so that the worst-case drive capability is as good as that of the reference inverter. If this strategy is adopted, then M3 and M4 can each be sized for half the effective pullup resistance of the reference inverter. If M1 and M2 are both of minimum size, then M3 and M4 would both have $W/L = 5$ to maintain this worst-case drive capability. In practice, minimum-size transistors are often used for both pullups and pulldowns to conserve area, resulting in asymmetric output drive.

A simple resistive model of the two-input NOR gate based on the device model of Fig. 7.5-4 is shown in Fig. 7.6-3. If all devices are of minimum size, then the pullup resistance is $2R_P = 5R_N$ for $K'_N = 2.5K'_P$. The pulldown resistance is R_N for a single NOR gate input high or $R_N/2$ for both inputs high. Thus, a maximum input-dependent asymmetry of 10:1 exists for this CMOS NOR structure based on minimum-size gates.

7.6.2 CMOS NAND Logic Gate

The circuit for a CMOS two-input positive logic NAND gate is also easily obtained once the reference inverter circuit has been defined. A CMOS NAND circuit is shown in Fig. 7.6-4 with its corresponding layout in Fig. 7.6-5. This circuit is called the *dual* of the CMOS NOR circuit because the two pullup transistors are connected in parallel rather than in series and the two pulldown transistors

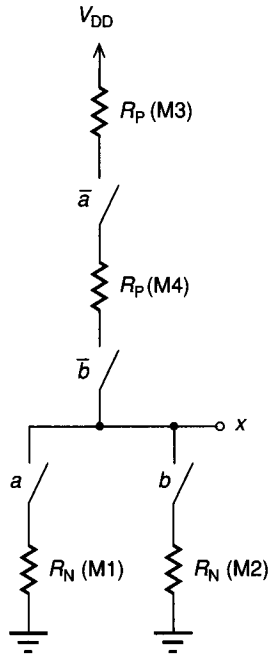

FIGURE 7.6-3
Simple resistive model of CMOS NOR gate.

are connected in series rather than in parallel. Once again, four transistors are required for a CMOS two-input NAND circuit, as compared with only three transistors for the similar NMOS circuit of Fig. 7.4-2.

The operation of the CMOS NAND circuit is similar to the CMOS NOR circuit except for the logic function realized. The interchange of the parallel and series connections for the pullup and pulldown paths changes the output voltage generated for specific input conditions, resulting in a realization of the NAND function. If both inputs a and b are below V_{TN}, the series n-channel pulldown

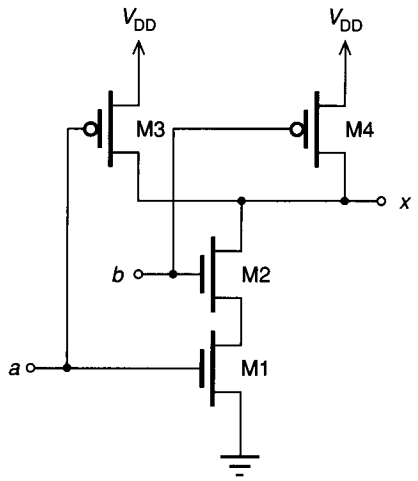

FIGURE 7.6-4
Two-input CMOS NAND gate.

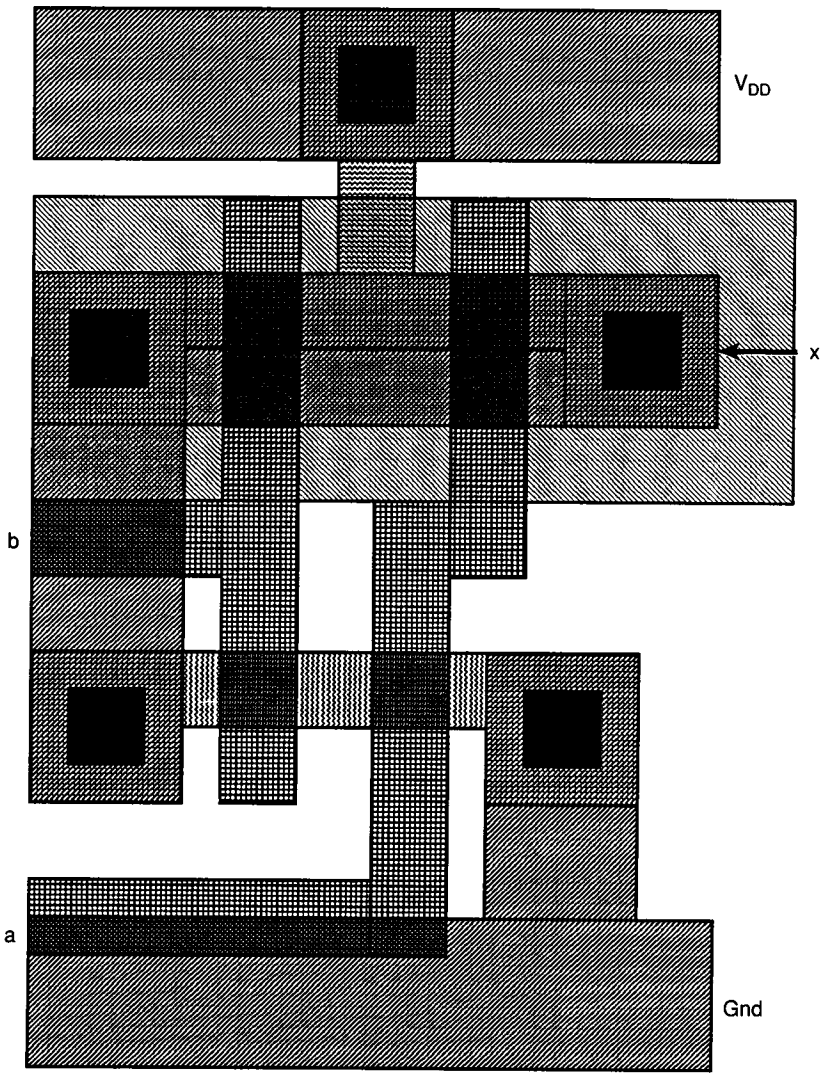

FIGURE 7.6-5
CMOS NAND gate layout with inputs a and b and output x.

transistors M1 and M2 are off while the parallel p-channel pullup transistors M3 and M4 conduct. This results in an output voltage level of V_{DD}. If only input a is below V_{TN}, then transistor M1 is off while transistor M3 conducts. Thus, the output voltage is still pulled to V_{DD}. If only input b is below V_{TN}, transistor M2 is off and transistor M4 conducts, again setting the output to V_{DD}. If both inputs a and b are above $V_{DD} - |V_{TP}|$, then transistors M3 and M4 are off and transistors M1 and M2 conduct. This condition sets the output voltage to 0 V. This input-output voltage relation realizes the positive NAND logic function.

For the NAND gate, the series path to ground uses n-channel transistors, and the parallel path to V_{DD} uses p-channel transistors. Based on minimum-size

transistors and a 2.5:1 transconductance advantage of n-channel over p-channel transistors, the series resistance of two n-channel transistors to ground will be nearly matched by a single p-channel pullup to V_{DD}. To demonstrate this, consider the NAND gate model of Fig. 7.6-6 based on the simple resistive models of Fig. 7.5-4. Let all transistors corresponding to Fig. 7.6-6 be of minimum size. For this circuit, the effective pulldown resistance is $2R_N$ when both inputs are high. The effective pullup resistance is either R_P or $R_P/2$ depending on whether only one or both inputs are low, respectively. Noting that $R_P = 2.5R_N$ for $K'_N = 2.5K'_P$, the worst-case input-dependent asymmetry is less than 2:1, which is much better than that obtained for the corresponding NOR circuit. This near-symmetry makes the NAND gate the preferred CMOS logic form. Note that a CMOS NOR gate requires greater silicon area to achieve either a lower resistance p-channel path or a higher resistance n-channel path to reduce the worst-case input-dependent asymmetry.

7.6.3 Multi-Input CMOS Logic Gates

Classical multi-input CMOS logic gates are formed by adding an n-channel pulldown and a p-channel pullup transistor for each additional input. For the NOR gate, the pulldown transistor is added in parallel with the other n-channel transistors, and the pullup transistor is added in series with the other p-channel transistors. A three-input CMOS NOR gate is shown in Fig. 7.6-7. The three-input CMOS NAND gate of Fig. 7.6-8 is formed by inserting a pulldown transistor in series with the n-channel transistors and adding a pullup transistor in parallel with the p-channel transistors.

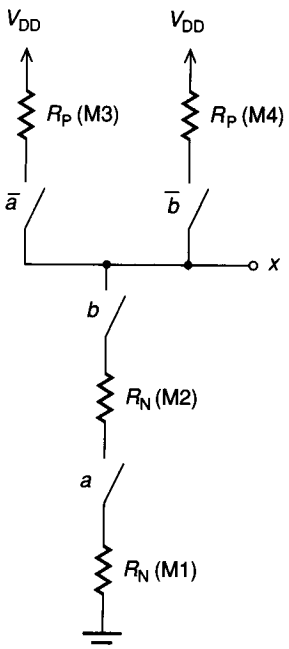

FIGURE 7.6-6
Simple resistive model of CMOS NAND gate.

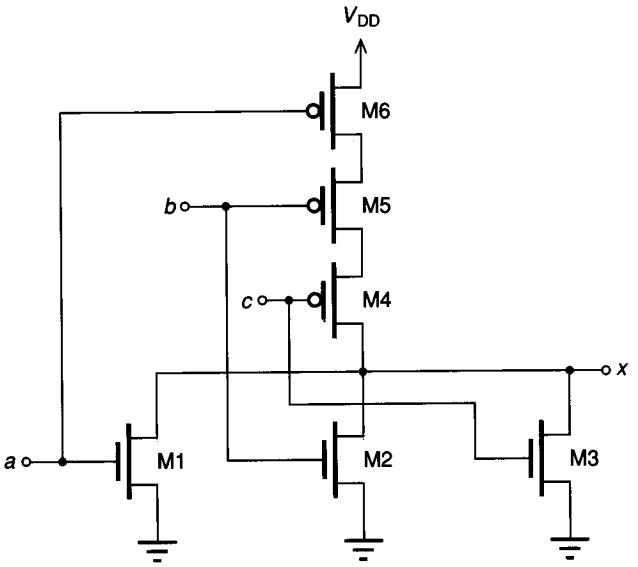

FIGURE 7.6-7
Three-input CMOS NOR gate.

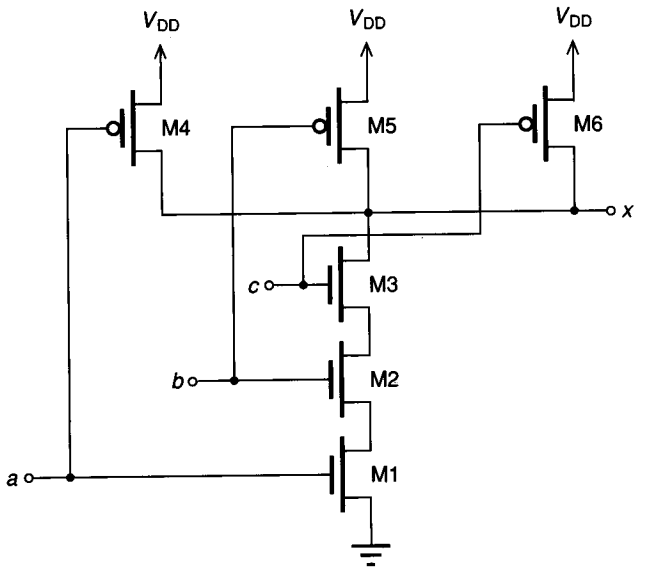

FIGURE 7.6-8
Three-input CMOS NAND gate.

Two characteristics of classical multi-input CMOS logic gates limit their use in VLSI circuits. Let N be the number of inputs to a multi-input logic gate. Let M_{NMOS} be the number of transistors required to form an N-input gate in NMOS or PMOS, and let M_{CMOS} be the number of transistors needed to form an N-input gate in CMOS. The number of transistors required to form NAND or NOR gates in either an NMOS or PMOS technology is

$$M_{NMOS} = N + 1 \tag{7.6-4}$$

whereas the number of transistors necessary to form classical NAND or NOR gates in a CMOS technology is

$$M_{CMOS} = 2N \tag{7.6-5}$$

The CMOS logic circuits introduced in this section are designated *classical* CMOS logic because of their relatively long history of use. Classical multi-input CMOS logic gates require more transistors than their NMOS counterparts for any number of inputs. As the number of inputs to a multi-input logic gate grows, a CMOS gate requires approximately twice as many transistors as an NMOS gate. To circumvent this disadvantage, clever circuit structures are commonly used for multi-input gates in CMOS. One of these techniques, domino CMOS, is discussed in a subsequent section. Alternatively, NMOS-like structures are sometimes used in CMOS, with a p-channel pullup with its gate grounded used to emulate the depletion-load device of NMOS. Like NMOS logic, this pseudo-NMOS logic has the disadvantage of static power dissipation.

A second disadvantage of classical multi-input CMOS gates arises from the requirement for a series transistor path from the output to one of the power supply nodes. For a NAND gate, this series path is the pulldown path. For a NOR gate, the series path is the pullup path. As the number of inputs to a classical CMOS gate grows, it becomes difficult to size the transistors for proper output drive through the series path. Remember that the multi-input NMOS NOR gate of Sec. 7.4 did not have this disadvantage.

Classical CMOS logic gates and their characteristics were described in this section. As with the CMOS inverter, ratioless CMOS logic gates provide ideal logic voltage levels and allows device sizing to be used for near symmetrical output drive. As a disadvantage, classical CMOS logic gates require almost twice as many transistors as their NMOS counterparts.

7.7 TRANSMISSION GATES

Because a MOS transistor is an excellent switching device, as demonstrated in Sec. 5.1, it is possible to connect this transistor in series with a logical signal to either pass or inhibit the signal. A MOS transistor connected in this way is called a *pass transistor* or *transmission gate* because it passes or transmits signals under control of its gate terminal. This connection has several advantages and some disadvantages compared to other methods of controlling a logic signal. In this section, characteristics of pass transistors are analyzed for NMOS circuits. Then the transmission gate structure for CMOS circuits is studied.

7.7.1 NMOS Pass Transistor

Figure 7.7-1 shows an input signal V_i connected through an n-channel transistor to the input V_a of a standard inverter circuit. A transistor M1 connected in this manner is called a *pass transistor*. This same circuit connection is important in dynamic storage circuits, to be discussed in Sec. 9.6. If the gate voltage, V_G, of the pass transistor is held at 0 V, and if the drain and source voltages are constrained between 0 V and V_{DD}, an enhancement pass transistor can never conduct because a positive gate-to-source voltage cannot be established. Because of the high off-resistance of an enhancement transistor, the input V_i is effectively disconnected from the inverter circuit.

If, on the other hand, V_G is held at V_{DD}, the pass transistor will conduct to equalize the voltages at its source and drain. This is explained by considering the alternatives with $V_G = V_{DD}$. First, assume that V_i, the input voltage to the pass transistor, is held at a low logic level: $V_i = 0$ V. If the input terminal of the pass transistor is considered the source, the gate-to-source voltage is V_{DD} and the pass transistor conducts to bring the drain terminal voltage V_a to 0 V. Note that no current can be supplied from the inverter input, an oxide-insulated gate.

Next assume that the input is held at $V_i = V_{DD}$. The pass transistor may appear to be off because both V_G and V_i are at the same voltage. However, because an enhancement transistor is symmetrical with respect to the drain and source terminals, the V_a terminal of the pass transistor can act as the source. If the voltage at the inverter input $V_a = 0$ V, the gate-to-source voltage of the pass transistor is initially V_{DD}, the transistor will conduct, and the inverter input V_a will be pulled to a higher voltage. Note that as V_a increases, the pass transistor gate-to-source voltage is reduced. When V_a reaches $V_{DD} - V_{TN}$, the pass transistor will cease to conduct. Thus, the pass transistor can only pull the inverter input to one threshold voltage drop below the pass transistor gate voltage. This connection of the pass transistor with the V_i and V_G terminals held at V_{DD} is the same circuit configuration as the enhancement pullup load for the MOS inverter discussed in Secs. 6.1 and 7.3.

One last alternative must be considered. If both V_G and the input voltage V_i of the pass transistor are at V_{DD} and the inverter input $V_a = V_{DD}$, the pass transistor will not conduct. However, because the logic levels at the source and drain terminals of the pass transistor are already equivalent, the logical effect is the same as if the pass transistor were conducting. This analysis is summarized by the following two equations.

$$V_G = 0 \text{ V} \rightarrow \text{Pass transistor is off} \tag{7.7-1}$$

$$V_G = V_{DD} \rightarrow \text{Pass transistor conducts} \tag{7.7-2}$$

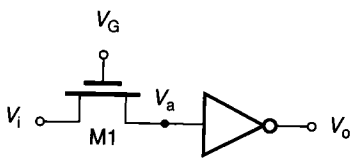

FIGURE 7.7-1
Input signal connected to an inverter through a pass transistor.

At this point the reader might infer that a pass transistor could be used to logically AND two input signals. One signal could be applied to the pass transistor input terminal, and the other could be applied to the pass transistor gate terminal. The output of the pass transistor would be pulled high only if both the gate terminal *and* the input terminal of the pass transistor were high. This is correct; however, the converse is not true. The output of the pass transistor will be pulled low if its input terminal is low, but not if its gate terminal is low. According to Eq. 7.7-1, when the gate is at 0 V, the pass transistor is off and cannot force the output to a low voltage to satisfy the AND functional requirements. Thus, this connection does not function as an AND gate. The assumption by a designer that a pass transistor can function as an AND gate would be a simple but fatal (to circuit operation) mistake.

A pass transistor used as a logic switch has important advantages in terms of integrated circuit layout constraints. First, the pass transistor consists of a single transistor and therefore requires less area than a logic gate. Even the simplest logic gate, an inverter, requires two transistors. Additionally, a pass transistor is a three-terminal device, whereas an inverter is a four-terminal device if one counts power and ground. A requirement for fewer interconnections is a major advantage when integrated circuit layout constraints are considered. For many applications, the pass transistor can be a minimum-size device, further reducing layout area. Finally, a pass transistor requires no dc power—a significant advantage.

A typical use of pass transistors to create a 4-to-1 selector circuit is shown in Fig. 7.7-2. A circuit that can connect one of N different inputs to an output is called an N-to-1 *selector circuit*. Here 8 pass transistors replace an equivalent of 21 NMOS transistors or 32 CMOS transistors if classical logic gates are used. This selector circuit can realize all 16 logic functions of the two inputs a and

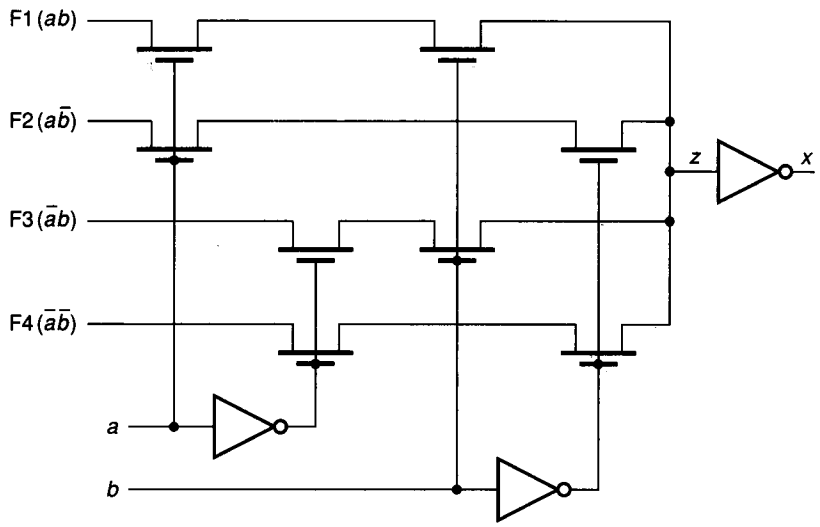

FIGURE 7.7-2
Pass transistor selector circuit.

b. For example, if F1–F4 are all set to 0, then the output z is 0. If F1 is set high and F2–F4 are set low, then the AND of a and b is given at z. If F2 and F3 are set high while F1 and F4 are set low, then the exclusive-OR function of a and b is realized at z. If F1–F3 are set high and F4 is set low, then a OR b is given at z. The 16 possible combinations of 0 or 1 at the four inputs F1–F4 yield the 16 possible logic functions of a and b. This selector circuit function block is popular within microprocessor ALUs, where both logic and arithmetic functions must be realized. Note that power dissipation, number of devices, and interconnection requirements are considerably reduced from an equivalent logic gate implementation. This is but one example of the simplicity available through the use of pass transistor logic.

A series string of pass transistors connected to control the logical path of an input signal is an appealing circuit configuration in terms of area and interconnection constraints compared to other alternatives. For example, such a string is frequently used to selectively propagate the carry in a multibit binary adder. However, at least two significant problems are encountered with a series string of pass transistors. The first problem arises because of design imposed constraints on signal propagation delays. Figure 7.7-3 shows a series string of pass transistors connecting V_i to V_o with all gates held high and an approximate equivalent circuit using a lumped circuit element model. R represents equivalent resistance between the source and drain of a pass transistor. (Bulk effects increase the equivalent resistance here because the source terminal is not grounded.) C represents the capacitance to ground for each stage; the value is determined by the gate capacitance and the drain and source diffusion capacitance of each pass transistor. It can be shown that the signal propagation delay from V_i to V_o is proportional to the square of the number of identical stages.

A simplified explanation for the square-law delay for a series string of pass transistors can be given in terms of the resistance, capacitance, and associated RC time constants of the lumped circuit element model. A single pass transistor stage exhibits a delay that is a function of the series resistance R and the capacitance to ground C. This delay is proportional to the time constant $\tau = RC$. When a second

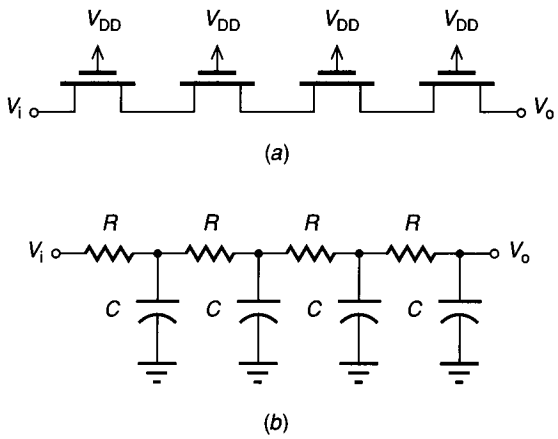

FIGURE 7.7-3
(*a*) Series string of pass transistors, (*b*) Lumped-element equivalent circuit.

pass transistor is added in series with the first and only the additional capacitance to ground is considered, the RC time constant ($\tau = R2C$), and therefore the transmission delay, is doubled. If the additional series resistance is considered but the additional capacitance to ground is neglected ($\tau = 2RC$), the delay is still doubled. The combination of twice the resistance and twice the capacitance ($\tau = 2^2RC$) causes the total delay to nearly quadruple. In general, the delay is proportional to N^2RC where N is the number of series pass transistors. For this reason, long series connections of pass transistors are usually segmented by the addition of inverters at intervals of about four pass transistors. The inverters buffer the signal and break the square-law delay effect of the series pass transistor cascade, resulting in smaller overall signal delay.

A second problem is encountered if pass transistors are cascaded with the output of one pass transistor connected to the gate of the next pass transistor as shown in Fig. 7.7-4. This problem is related to the threshold voltage drop from the gate of a pass transistor to its output terminal. As described previously for enhancement pullup transistors, the source voltage can only be pulled to a value V_{TN} less than the gate terminal voltage. If both the gate and drain terminals of a pass transistor are at a voltage V_{DD}, then the source terminal is pulled no higher than $V_{DD} - V_{TN}$. Succeeding circuits still reliably interpret this voltage as a high logic level. However, if the output (source terminal) of a pass transistor is connected to the gate of a second pass transistor, as in the series pass transistor cascade of Fig. 7.7-4, the output of the second pass transistor can only be pulled to a voltage of $V_{DD} - 2V_{TN}$. In general, for N pass transistors cascaded source to gate, the voltage at the last source terminal V_o can only be pulled to

$$V_o < V_{DD} - NV_{TN} \tag{7.7-3}$$

This voltage is too low to be recognized reliably as a high logic level if $N \geq 2$. For this reason, source-to-gate cascades of pass transistors represent an electrical rule violation and are avoided in practice.

7.7.2 CMOS Transmission Gate

The NMOS pass transistor described previously is an ideal element for performing many logic and control functions and is widely used in NMOS designs. However, is the pass transistor a useful device within CMOS designs? To help answer this question, Fig. 7.7-5 shows two CMOS inverters joined using a typical

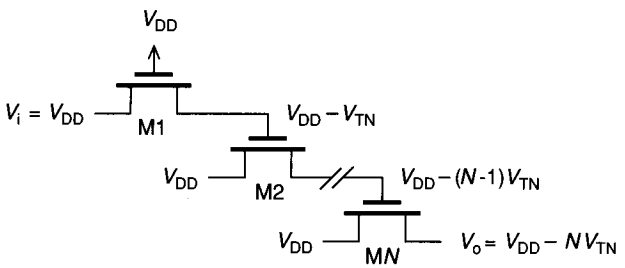

FIGURE 7.7-4
Improper cascaded connection of pass transistors.

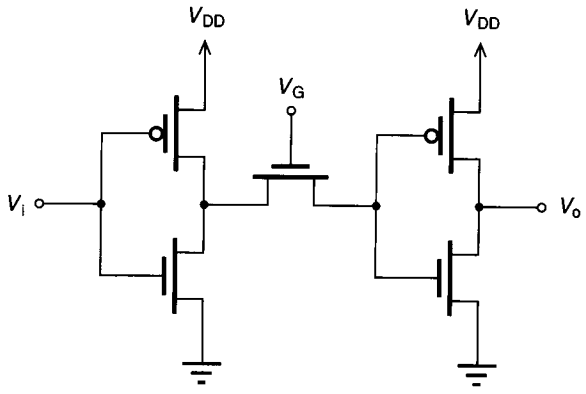

FIGURE 7.7-5
Pass transistor connecting two CMOS inverters.

n-channel pass transistor. Certainly, if the pass transistor gate voltage $V_G = 0$ V, the pass transistor isolates the two CMOS inverters, just as in the case of NMOS inverters. If the pass transistor gate voltage $V_G = V_{DD}$ and the output of the first inverter is at 0 V, the pass transistor will pull the input of the second inverter to 0 V. Once again, this is similar to the NMOS case and is satisfactory. One other condition must be considered. If both the output of the first inverter, that is, the drain terminal of the pass transistor, and the pass transistor gate are at V_{DD}, then the source terminal of the pass transistor can be pulled no higher than $V_{DD} - V_{TN}$. This voltage is sufficient to turn on the n-channel pulldown transistor of the second inverter, but it may not completely turn off the corresponding p-channel pullup transistor.

A gate terminal voltage greater than $V_{DD} - |V_{TP}|$ is required to turn off a p-channel pullup transistor. When the gate voltage drops below $V_{DD} - |V_{TP}|$, the p-channel transistor begins to conduct. If the p-channel pullup transistor conducts because of mismatched V_{TN} and V_{TP} threshold voltages, for example, static power is dissipated in the gate, and the effective noise margin is reduced. Thus, the standard n-channel pass transistor configuration is undesirable for driving a CMOS gate.

The pass transistor is such a useful circuit element in digital designs that a CMOS equivalent is used to accomplish a similar effect. The CMOS circuit is not called a pass transistor but, rather, is called a *transmission gate* because more than a single transistor is required. Figure 7.7-6 shows a CMOS transmission

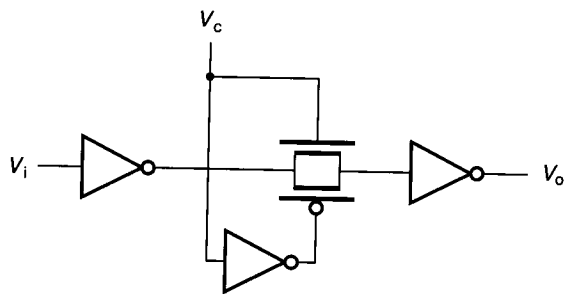

FIGURE 7.7-6
Transmission gate connecting two CMOS inverters.

gate connecting two inverters. The single n-channel pass transistor is replaced by parallel n-channel and p-channel transistors driven by opposing logic levels. Unfortunately, unless the control signal V_c for the transmission gate is available as a double-rail logic signal, an inverter is required in addition to the parallel n-channel and p-channel transistors.

Operation of a CMOS transmission gate circuit can be explained as follows. If the control signal to the transmission gate (shown in Fig. 7.7-6) is $V_c = 0$ V, the gate terminal of the n-channel transistor is also at 0 V and the gate terminal of the p-channel transistor is at V_{DD} because of the control signal inverter; thus, both transistors are off. If the control signal is $V_c = V_{DD}$, then appropriate voltages are generated to cause both the n-channel and the p-channel transistors of the transmission gate to conduct. Under this condition the two parallel transistors of the transmission gate function in a complementary manner to connect the first and second inverters of the signal path in Fig. 7.7-6. When the output of the first inverter is at 0 V, the n-channel transistor of the transmission gate pulls the input to the second inverter to 0 V. When the output of the first inverter is at V_{DD}, the p-channel transistor of the transmission gate pulls the input to the second inverter all the way to V_{DD}. This circumvents the threshold voltage drop problem encountered with a single enhancement pass transistor for CMOS logic. Thus, the CMOS transmission gate provides voltage levels that are compatible with other CMOS logic gates. Unfortunately, the cost of extra silicon layout area for transistors and interconnection is high, rendering the CMOS transmission gate a less useful device than the comparable NMOS pass transistor.

Both pass transistors and transmission gates were described in this section. These building blocks are most advantageous where they can select or control the flow of a logic signal. Care must be exercised in cascading these devices to prevent large signal propagation delays.

7.8 SIGNAL PROPAGATION DELAYS

Earlier sections of this chapter dealt with steady state analysis of MOS circuits that perform digital logic functions. Now an additional parameter—time—is injected into the analysis. Real logic circuits require nonzero but finite time for signals to propagate from input to output. Models to analyze and predict the propagation delay are crucial for practical logic design.

Delays encountered in digital circuitry are composed of two principle components: *gate delay* and *interconnection delay*. Logic gate delay, the time required for a signal to propagate from the input of a logic gate to the output of the same gate, is an important parameter in determining the capabilities of a logic family such as TTL or NMOS logic. Historically, logic gate delay has been the major limiting factor in setting clock rates, and therefore the computational speed, of single-chip digital integrated circuits. Digital systems, comprising many integrated circuit chips, require analysis of interconnection delay in addition to the logic gate delay of the circuits themselves. Digital system interconnection delays are those arising from integrated circuit package connections, printed circuit board connections, and chassis back-plane connections. As integrated circuits

have been manufactured with reduced device sizes, internal gate-to-gate interconnection delays have increased in importance relative to logic gate delays. As device sizes reach submicron dimensions, internal interconnection delays dominate the gate delays. The characterization of signal propagation delays for logic gates loaded by a single, identical logic gate with minimal interconnection is addressed in this section. In a subsequent section delays caused by nonhomogeneous logic gate characteristics, logic fanout, interconnection capacitance, and off-chip loads are analyzed.

7.8.1 Ratio-Logic Model

To address the problem of signal propagation delay, the simple case of Fig. 7.8-1*a*, where the output of one inverter drives the input to a second, identical inverter, will be considered initially. For this analysis, depletion-load inverters such as the one analyzed in Fig. 7.3-1*a*, and redrawn in Fig. 7.8-1*b*, will be used. Figure 7.8-2*a* shows an ideal voltage step of amplitude V_{DD}. Assume that this voltage step is applied to the input of the first inverter of Fig. 7.8-1*a* at $t = t_0$. For the output of the first inverter, the response of Fig. 7.8-2*b* is ideal, but the actual response will resemble that of Fig. 7.8-2*c*. In this case, the signal propagation delay is the difference between the time of the input transition, t_0, and the time that the output is recognized as a valid logic low voltage, t_1. This delay is caused by parasitic capacitances in the MOSFETs, as discussed in Sec. 3.1, and interconnection capacitance. The slight initial overshoot in the actual response is caused by the gate-drain overlap capacitance of M1. The overall delay time, however, is dominated by the effects of the capacitive load caused by the second inverter including the interconnection capacitance. A closed-form mathematical expression for the output waveform of Fig. 7.8-2*c* is unwieldy because of the

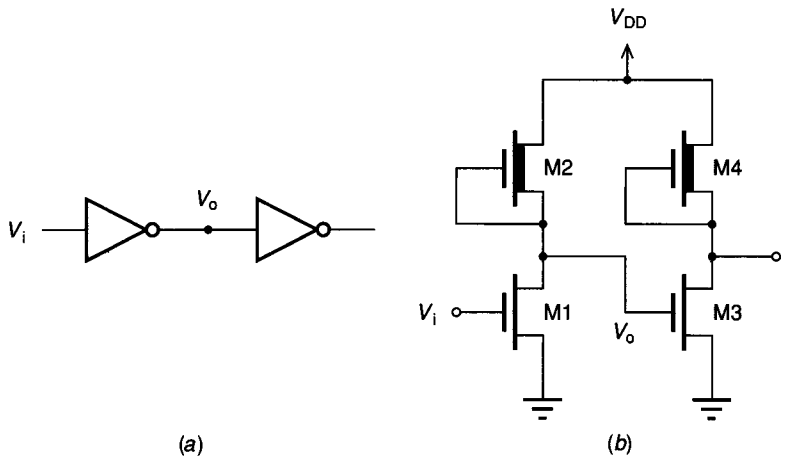

FIGURE 7.8-1
Single inverter driving a second, identical inverter: (*a*) Logic diagram, (*b*) Circuit diagram.

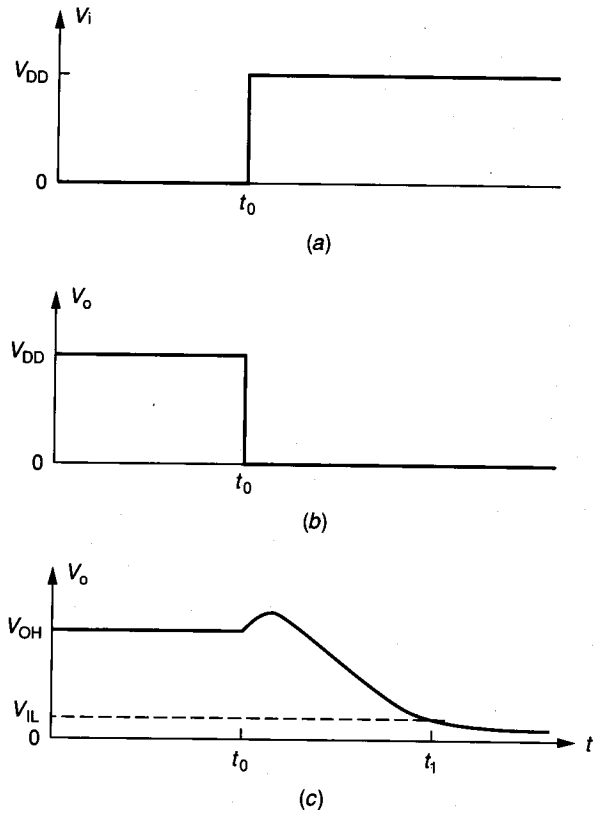

FIGURE 7.8-2
Voltage step input and inverter response: (*a*) Ideal voltage step, (*b*) Ideal response, (*c*) Typical response.

nonlinear nature of the device *I–V* characteristics, the voltage dependence of all parasitic capacitors, and the changes in operating region for the MOSFETs during the transition. A closed-form solution requires the simultaneous solution of a set of nonlinear partial differential equations.

For hand analysis and insight into design, a quick and simple method of approximating the response of this circuit is needed for both high-to-low and low-to-high output transitions. This initial analysis neglects interconnection capacitance and is based on the simple switch resistor model of Fig. 7.5-4, with the addition of gate capacitance. Figure 7.8-3*a* shows a symbolic convention that emphasizes the resistive models for transistors M1 and M2. This can be partitioned further to the resistance-capacitance inverter model of Fig. 7.8-3*b*. Note that M2 is modeled by a resistor without a switch since the depletion transistor with $V_{GS} = 0$ V is always on. M1 is modeled by C_G, S1, and R_1. Switch S1 is open for V_i low and closed for V_i high. The resistance R_2 models the pullup transistor resistance, and the resistance R_1 models the pulldown transistor resistance. The capacitance C_G is the input capacitance to the inverter. It will be demonstrated later that C_G is approximately equal to $C_{ox}WL$ where W and L are the width and length of the pulldown transistor (M1 or M3 in Fig. 7.8-1*b*). For this analysis, the equivalent input capacitance to the reference inverter in a logic family is termed C_G. From

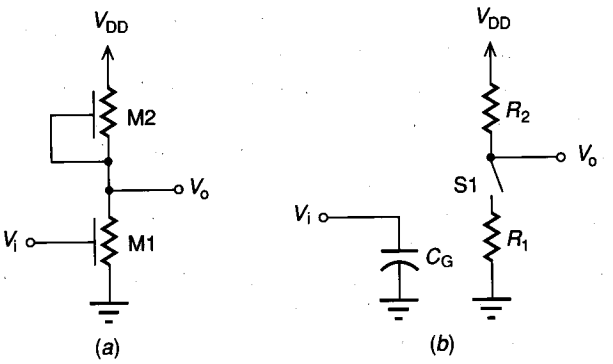

FIGURE 7.8-3
Inverter models for delay calculations: (*a*) Symbolic representation, (*b*) Resistance-capacitance inverter model.

a dc viewpoint, it follows from the model of Fig. 7.8-3*b* that the high output voltage level is V_{DD} and the low output voltage level is

$$V_L = \frac{V_{DD}R_1}{R_1 + R_2} \tag{7.8-1}$$

Logic circuits where the dc output voltage is determined by the ratio of two series resistors are termed *ratio logic* circuits. As was stated in Eqs. 7.5-2 and 7.5-3, the equivalent resistance of a MOS transistor in the model of Fig. 7.8-3*b* is proportional to its length L divided by its width W. Thus, the relationship between resistance R_2 and resistance R_1 can be approximated as

$$R_2 = \frac{L_2W_1}{W_2L_1}R_1 = kR_1 \tag{7.8-2}$$

With the 4 : 1 sizing rule of Sec. 7.3, $k = 4$ and $R_2 = 4R_1$. With this sizing rule and the model of Fig. 7.8-3*b*, it follows from Eq. 7.8-1 that the high and low logic levels for the output are V_{DD} and $V_{DD}/5$, respectively. For $V_{DD} = 5$ V, this gives a high logic level of 5 V and a low logic level of 1 V. Although the high logic level is the same as that obtained by the more exact analysis of Sec. 7.3, the low logic level differs from the more exact value of 9/32 V obtained previously with $V_{DD} = 5$ V. This difference in dc voltage levels is not significant for an approximate delay analysis.

To provide a simplified model for delay calculations, approximate values for resistances R_1 and R_2 of Fig. 7.8-3*b* must be found. Observe that when V_i is high, S1 is closed and M1 is in the ohmic region for much of the high-to-low output transition. From Eq. 3.1-8 with $V_{GS} = V_i = V_{DD}$, a good approximation for the resistance R_1 near $V_o = V_{DS} = 0$ V is the small signal equivalent resistance

$$R_{ss} = \frac{L_1}{K'W_1(V_{DD} - V_{TN})} \tag{7.8-3}$$

(Note that Eq. 7.8-3 demonstrates the proportionality factor described for Eqs. 7.5-2 and 7.5-3.) This resistance is a good reference point but does not consider the effects of the pullup resistance or the nonlinear characteristics of the enhancement transistor over the actual output voltage range.

A better model for delay estimation is found by considering the large signal equivalent resistance. The equivalent resistance at the start of the output transition is found by dividing the initial voltage across M1 by the initial current through M1. The pullup transistor M2 contributes no current at this point because it has $V_{DS} = 0$ V. Assuming $V_i = V_{DD}$, the equivalent resistance is given by

$$R_1 = \frac{2V_{DD}L_1}{K'W_1(V_{DD} - V_{TN})^2} \approx 2R_{ss} \tag{7.8-4}$$

using the approximation that $V_{DD} - V_{TN} \approx V_{DD}$. As the output voltage falls, the resistance of the enhancement transistor approaches R_{ss}, but the depletion pullup transistor begins to supply current, thereby increasing the effective resistance of the inverter output. Analysis of the large signal equivalent resistance at an output voltage near V_{TN} with the effects of pullup transistor M2 considered gives

$$R_1 = \frac{V_o L_1}{K'W_1\left[(V_{DD} - V_{TN} - V_o/2)V_o - V_{TD}^2/2k\right]} \tag{7.8-5}$$

It can be shown that this expression also reduces to $R_1 \approx 2R_{ss}$ with typical parameters for the reference inverter. For the approximate propagation delay analysis in this section, it will be assumed that R_1 is given by Eq. 7.8-4.

For a low-to-high output transition, transistor M2 starts in the saturated region and finishes in the ohmic region. With typical parameter values for the 4:1 reference inverter, the large signal equivalent resistance of M2 for a low-to-high output transition varies from about $13R_{ss}$ to $5R_{ss}$ as V_o increases from 0 V to V_{DD}. An average equivalent resistance of $8R_{ss}$ is a reasonable compromise. Note that considering the 4:1 resistance ratio for the reference inverter, the equivalent pullup resistance is

$$R_2 \approx 4R_1 = 2kR_{ss} \tag{7.8-6}$$

It should be noted that for this approximate analysis, significant errors of $\pm 50\%$ or more can be anticipated using the simple model of Fig. 7.8-3*b*. These errors are justifiable for a quick approximate analysis. The alternative when better accuracy is required is to use circuit simulation programs such as SPICE to account for the nonlinear characteristics of MOS transistors and the effect of nonzero rise time for the input voltage. If $L_1 = W_1$, $L_2 = 4W_2$, $K' = 30\ \mu A/V^2$, $V_{TN} = 1$ V, $V_{TD} = -3.5$ V, and $V_{DD} = 5$ V, then

$$R_1 \approx 16.6\ k\Omega \quad \text{and} \quad R_2 \approx 66.4\ k\Omega \tag{7.8-7}$$

With approximate models for the transistor resistances in hand, the analysis now turns to the input capacitance. Because M1 of Fig. 7.8-1*b* is ohmic for $V_i = V_{DD}$ and cutoff for $V_i = 0$ V, it follows from Fig. 3.1-19*b* that the parasitic input capacitance at either logic level is approximately $C_{ox}W_1L_1$. During transitions, M1 enters the saturation region momentarily, causing the input capacitance to drop to approximately $(2/3)C_{ox}W_1L_1$, as indicated in Fig. 3.1-19*b*. Because the capacitance change is relatively small and M1 is normally in the saturation region for only part of the output transition time, the change is neglected in the model of Fig. 7.8-3*b*. Thus, C_G is approximated as

$$C_G \approx C_{ox}WL \tag{7.8-8}$$

To estimate propagation delays, the equivalent circuit model of Fig. 7.8-3*b* is simplified to the model of Fig. 7.8-4*a* by adding switch $\overline{S1}$. This model isolates the effects of R_1 and R_2 to simplify the analysis further. Remember that the effects of the pullup transistor during a high-to-low transition were included when R_1 was approximated earlier.

The high-to-low and low-to-high transition times for an inverter loaded by an identical inverter will now be determined. Because the input to the load inverter looks like a capacitor of value C_G, the equivalent circuit for a low-to-high step input voltage is as shown in Fig. 7.8-4*b*. Note that this input causes a high-to-low output transition. Because this model is a simple *RC* circuit with an initial voltage on C_G, it follows that a major portion of a high-to-low output transition, designated t_{HL}, occurs in two *RC* time constants, where $R = R_1$ and $C = C_G$ (for an ideal *RC* circuit, the 90% to 10% transition requires 2.2 time constants). Thus, the high-to-low output transition is approximated here by

$$t_{HL} \approx 2R_1C_G \tag{7.8-9}$$

Through a similar analysis of the circuit in Fig. 7.8-4*c*, the low-to-high output transition time is approximated by

$$t_{LH} \approx 2R_2C_G \tag{7.8-10}$$

For ratio-logic circuits, R_1 and R_2 of Fig. 7.8-4*a* are related by the device-sizing parameter, k. From Eqs. 7.8-2, 7.8-9, and 7.8-10, it follows that

$$t_{LH} = kt_{HL} \tag{7.8-11}$$

For the 4:1 sizing rule,

$$t_{LH} = 4t_{HL} \tag{7.8-12}$$

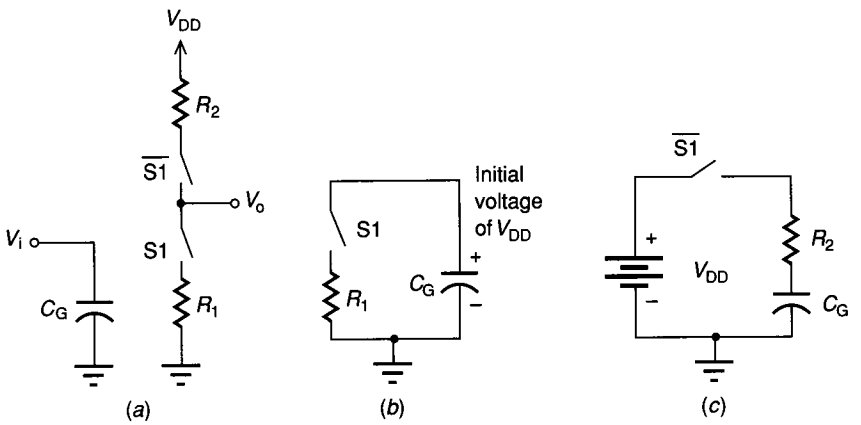

FIGURE 7.8-4
Equivalent circuits for inverter delay analysis: (*a*) Simplified *RC* inverter model, (*b*) Equivalent circuit for high-to-low output transition, (*c*) Equivalent circuit for low-to-high output transition.

7.8.2 Process Characteristic Time Constant

Each process will have geometrical design rules that limit the minimum size for a transistor. The process will also have electrical design rules that specify the desired supply voltage V_{DD}, the transconductance parameter K', the threshold voltage V_T, and the gate capacitance per unit area C_{ox}. From these parameters, a characteristic time constant for the process can be determined. This time constant, designated as τ_P, and defined as $R_{ss}C_G$ is useful for comparing delay characteristics of different processes. For both NMOS and CMOS technologies the values of K' and V_T for the n-channel enchancement transistor are used; and for PMOS technologies the values of K' and V_T for the p-channel enchancement transistor are used. From Eqs. 7.8-3 and 7.8-8,

$$\tau_P = R_{ss}C_G = \frac{L_1}{K'W_1(V_{DD} - V_{TN})} C_{ox}W_1L_1 \tag{7.8-13}$$

This reduces to

$$\tau_P = \frac{L_1^2 C_{ox}}{K'(V_{DD} - V_{TN})} \tag{7.8-14}$$

Assume the typical minimum transistor dimension is 2 μ, $K' = 45$ $\mu A/V^2$, $C_{ox} = 1$ fF/μ^2, $V_{TN} = 1$ V, and $V_{DD} = 5$ V. Using these values in Eq. 7.8-14, the process characteristic time constant is

$$\tau_P = 0.02 \text{ ns} \tag{7.8-15}$$

It must be noted that τ_P is not a measure of expected circuit delay. The value of τ_P depends only on process geometrical and electrical parameters and is thus independent of a particular circuit implementation. Therein lies the usefulness of τ_P. From Eqs. 7.8-4, 7.8-6, 7.8-9, and 7.8-10, it can be observed that t_{HL} and t_{LH} can be expressed in terms of τ_P as $t_{HL} = 4\tau_P$ and $t_{LH} = 16\tau_P$ for a minimum-size inverter.

7.8.3 Inverter-Pair Delay

As was the case in the analysis of logic levels in Sec. 7.2, the inverter pair plays an important part in logic gate delay analysis. Because the high-to-low and low-to-high transition times are asymmetric, neither transition time is adequate to characterize a logic family. If the signal propagation delay from V_i to V_o for a pair of identical cascaded inverters shown in Fig. 7.8-5 is considered, then both

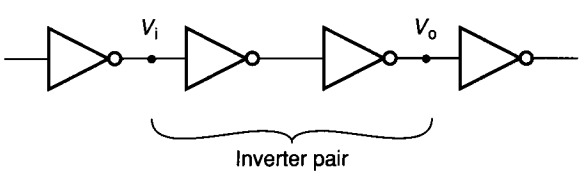

FIGURE 7.8-5
Cascade of identical inverters.

high-to-low and low-to-high transitions contribute to the delay. Thus, the inverter-pair delay, designated t_{ipd}, is defined as the sum of a high-to-low transition and a low-to-high transition time:

$$t_{ipd} = t_{HL} + t_{LH} \tag{7.8-16}$$

This is often expressed in terms of the device sizing ratio as

$$t_{ipd} = (1 + k)t_{HL} \tag{7.8-17}$$

or from Eqs. 7.8-4, 7.8-9, and 7.8-13 in terms of τ_P:

$$t_{ipd} = 4(1 + k)\tau_P \tag{7.8-18}$$

The inverter-pair delay is a key parameter in characterizing the speed of operation of a digital logic family. It represents a fundamental lower bound on the clock period of a synchronous system because a signal must be able to experience at least one high-to-low and one low-to-high transition during a system clock cycle. It will be seen later that the system clock period is typically much longer than the inverter-pair delay because of the contribution of interconnection delays and multiple levels of logic.

A physical interpretation of the inverter-pair delay deserves consideration. If either a high-to-low or a low-to-high input transition is applied to a cascade of inverters with identical loads, as in Fig. 7.8-5, then the delay associated with propagation of the signal through any two inverters (i.e., an inverter pair) is the inverter-pair delay. Assume the inverter cascade of Fig. 7.8-5 is designed in the typical NMOS process used for Eq. 7.8-15. Further assume that the gate of the pulldown device for each inverter is the minimum size of 2 $\mu \times 2$ μ, it follows from Eqs. 7.8-15 and 7.8-18 that the inverter-pair delay is

$$t_{ipd} = 4(1 + 4)0.02 \text{ ns} = 0.4 \text{ ns} \tag{7.8-19}$$

Note that based on the simplified device model used in these calculations, the response time for the reference inverter pair is very fast.

Three important observations must be made at this point. First, the present analysis gives unrealistically small delay times because it neglects interconnection capacitance. Second, the low-to-high transition delay for ratio logic is about k times as long as the high-to-low transition delay, as indicated by Eq. 7.8-11. This introduces significant asymmetry in the two transitions and is of concern in many applications, particularly if an output is loaded by a large capacitance. A typical response of the reference inverter to a fast square-wave excitation in a ratio-logic circuit is given in Fig. 7.8-6. This figure clearly shows the asymmetric rise and fall times for a ratio inverter circuit. The asymmetry occurs because the active pulldown device, M1, has much lower resistance than the passive pullup device, M2. One might be tempted to improve the pullup characteristics of M2 to decrease t_{LH} by resizing M2 for lower equivalent resistance. Unfortunately, this solution is not feasible because it would cause a degradation in logic signal levels, as can be seen from Eq. 7.8-1. The asymmetry in transition times is thus inherent in ratio-logic systems, as indicated by Eq. 7.8-11.

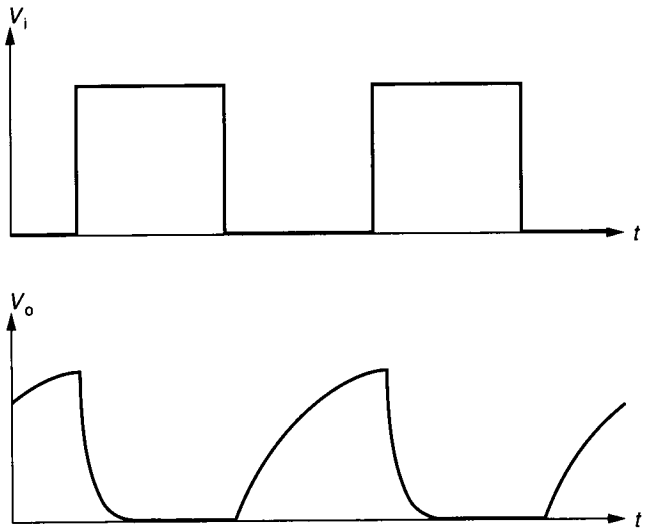

FIGURE 7.8-6
Asymmetric response for a ratio-logic inverter driven with a square wave.

A comparison of results obtained from the approximate delay analysis with those obtained from a detailed circuit simulation shows some differences.[7] These are attributable primarily to the terms R_1 and R_2 in Eqs. 7.8-9 and 7.8-10. These differences are due, in part, to the fact that the pullup and pulldown devices are not linear throughout their entire output transitions. Also, the previous analysis estimated a stage delay based on an ideal voltage step input. For a cascade of MOS inverters, the ideal step input applies only to the first stage; subsequent stages typically have an input rise time almost as long as the delay of the preceding stage. This partially invalidates the earlier assumption that the inverter input voltage was at V_{DD} during the entire high-to-low output transition. Thus, if the delay for a cascade of inverters is computed as the sum of single inverter-stage responses with ideal step inputs, further error is introduced.

Two good ways exist to improve the accuracy of a simplified inverter-delay analysis without increasing its complexity. First, the inverter-pair delay t_{ipd} (or even t_{HL} and t_{LH}) for a reference inverter can be experimentally measured for a specific process or determined from an accurate computer simulation. These delay parameters can then be thought of as design parameters. Overall delays can be expressed directly in terms of these delay parameters of the reference inverter. Alternatively, the resistance values R_1 in Eq. 7.8-4 and R_2 in Eq. 7.8-6 can be modified to compensate for errors in t_{HL} and t_{LH}. It should be remembered that the approximate analysis is useful primarily as a quick estimate of circuit performance. Attempts to create a simple, accurate delay analysis are undermined by many factors in practical circuits.[8]

The third observation is that R_1 and R_2 from Eqs. 7.8-4 and 7.8-6 are dependent on device length-to-width ratios but not on the actual lengths and widths of a transistor. As device sizes shrink—for example, from 2 μ to 1 μ—the

transistor resistances remain relatively constant. The gate area of M3, the input device of the second stage in Fig. 7.8-1*b*, provides the parasitic capacitance load to the first stage. In a linearly scaled process, the gate area of M3 decreases with the square of the feature size, while the gate oxide thickness decreases directly with device size (see constant field scaling, Sec. 1.1). For these approximations, the total capacitive loading decreases linearly with feature size. It thus follows that the inverter-pair delay, which depends on effective transistor resistance and parasitic capacitive loading, decreases approximately linearly with feature size. This decrease offers the potential for significant system speed improvements as devices become smaller. The process characteristic time constant τ_P for the scaled-down process shows this speed increase.

The previous analysis was for a single inverter loaded by a single, identical inverter stage. If an input signal transition must ripple through a cascade of identical logic circuits, a first-order approximation of the total delay for the cascade assumes that the start of the transitions on successive stages is delayed until the transition of the preceding stage is complete. This allows the total delay to be simply computed as the sum of the individual stage delays. If the cascade delay is denoted as t_{cas}, the number of identical stages is N, and the average stage delay is one-half the inverter-pair delay t_{ipd}, then an approximation to the total delay is

$$t_{cas} = \frac{Nt_{ipd}}{2} \tag{7.8-20}$$

It will be shown in a subsequent section that capacitive loading from interconnections and fan-out causes delays larger than that predicted by Eq. 7.8-20.

7.8.4 Superbuffers

The asymmetric output delay of a ratio logic circuit is particularly undesirable when a highly capacitive bus or a large number of secondary device inputs must be driven. The slow pullup capability can seriously limit clock speeds for a system. One partial solution to this problem is a special circuit configuration called a *superbuffer*. Figures 7.8-7 and 7.8-8 show circuits for a noninverting super-

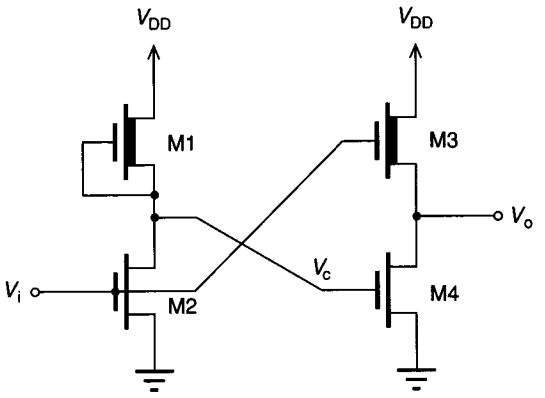

FIGURE 7.8-7
Noninverting superbuffer circuit.

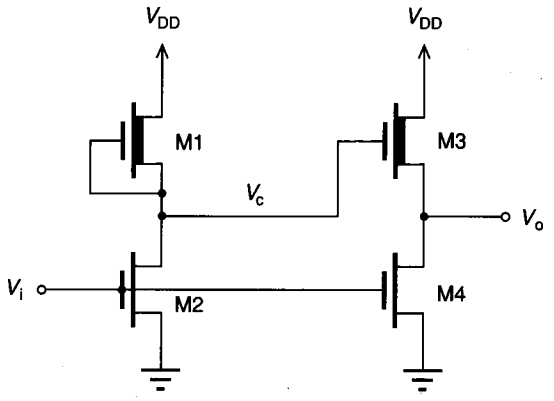

FIGURE 7.8-8
Inverting superbuffer circuit.

buffer and an inverting superbuffer, respectively. The output stage of each of these circuits is modified from the standard ratio-logic inverter in that the pullup transistor has its gate connected to an active logic signal. This connection increases the current sourcing ability of the pullup device, allowing faster drive for capacitive loads.

As can be seen from Figs. 7.8-7 and 7.8-8, a superbuffer consists of a standard inverter stage connected to a second inverter stage with an active pullup. The first inverter stage provides the input signal and its logical complement to drive the transistors of the output inverter stage. Both the input signal and its complement are necessary to drive the pulldown transistor and pullup transistor of the output stage at complementary times.

The logical operation of the noninverting superbuffer of Fig. 7.8-7 will now be examined. Assume that the output stage with transistors M3 and M4 has the 4:1 pullup/pulldown ratio that was used for the reference inverter. In this case, both transistors may remain unchanged sizewise from the reference inverter circuit. If the input to the noninverting superbuffer is low, then the output should also be low. Because the M1-M2 circuit is an inverter, the intermediate voltage V_c will be high when the superbuffer input is a logic low. Because this intermediate voltage drives the gate of pulldown transistor M4 to a logic high voltage, the output of the superbuffer is pulled low. The gate of the pullup transistor M3 is tied directly to the superbuffer input; thus, its gate is connected to a low voltage. If the superbuffer output is low, pullup transistor M3 has its gate and source each connected to a low voltage. This provides a gate-to-source voltage of 0 V, and the circuit is therefore equivalent to the standard depletion-load inverter when the superbuffer input is low. Based on this analysis, the superbuffer output stage functions as a standard depletion-load inverter stage when the superbuffer output is low.

The logical operation of the noninverting superbuffer will now be examined in response to a high input. This high input at the first inverter stage will drive its output, V_c, to a low level. The signal V_c is connected to the gate of pulldown transistor M4, causing it to turn off. The pullup transistor M3 has its gate connected directly to the superbuffer input, which is high. Initially, before