the superbuffer output is pulled high, the source terminal of the output pullup transistor, M3, will be at a low voltage. This provides a gate-to-source voltage for pullup transistor M3 that is approximately equal to a logic high voltage minus a logic low voltage, or nearly V_{DD}. For a typical NMOS process for digital circuits, it is easily shown that a gate-to-source voltage of 0 V for a depletion-mode transistor allows a current that is nearly equal to the current for a similarly sized enhancement transistor with a gate-to-source voltage of $V_{DD} - V_{TN}$. Increasing the gate-to-source voltage to V_{DD} for a depletion-mode transistor is roughly equivalent to doubling the gate-to-source voltage for an enhancement transistor. This provides improved pullup capability during the low-to-high transition of the superbuffer output.

The actively driven pullup transistor M3 improves the pullup characteristic of the superbuffer as just explained; the following analysis quantifies this improvement. For integrated digital circuitry with depletion-mode pullup transistors, the magnitude of V_{TD} is typically chosen as a large percentage of the supply voltage—for example, $-V_{TD} \approx 0.7V_{DD}$. For a standard depletion-mode pullup transistor with a gate-to-source voltage of 0 V, the transistor is in the ohmic region whenever the drain-to-source voltage is less than $0.7V_{DD}$. If the gate-to-source voltage is greater than $0.3V_{DD}$, the depletion-mode transistor is always in the ohmic region because the drain-to-source voltage is limited to V_{DD}. For a transistor in the ohmic region with constant drain-to-source voltage, its equivalent resistance is inversely proportional to the effective gate-to-source voltage, $V_{GS} - V_{TN}$ (see Eq. 3.1-8). If the effective gate-to-source voltage is doubled, the effective resistance of the pullup transistor is halved.

In the analysis of the preceding paragraph, it was shown that the equivalent resistance of the active pullup transistor of a superbuffer is initially half that of a standard inverter pullup for a low-to-high output transition. From Eq. 7.8-10, the delay t_{LH} is directly proportional to the effective resistance of the pullup transistor. Thus, the new low-to-high delay becomes

$$t'_{LH} = \frac{1}{2}t_{LH} = 2t_{HL} \tag{7.8-21}$$

The equivalent inverter-pair delay for a superbuffer output stage with the standard pullup/pulldown ratio becomes

$$t'_{ipd} = t_{HL} + t'_{LH} = 3t_{HL} \tag{7.8-22}$$

This means that a superbuffer output stage can drive a heavy capacitive load almost twice as fast as a standard inverter of the same size ($3t_{HL}$ versus $5t_{HL}$). A similar analysis applies to the inverting superbuffer of Fig. 7.8-8.

7.8.5 NMOS NAND and NOR Delays

Many times, a designer has a choice between using a NAND circuit or a NOR circuit to realize a particular logic function. For example, any combinational logic function can be realized in a sum-of-products form with a NAND-NAND logic gate combination or as a product-of-sums form with a NOR-NOR logic gate

structure.[6] Figure 7.8-9 shows the implementation of a simple logic function, the exclusive-OR, with both a NAND-NAND and a NOR-NOR realization. Because either structure can be used, it is useful to ask if one structure has advantages over the other.

In Sec. 7.4, it was determined that the typical pullup/pulldown sizing ratio for a NOR gate was about 4:1 and that a typical pullup/pulldown ratio for a two-input NAND gate was about 8:1. Assuming the input devices for the NAND and the NOR gates are sized the same as for the reference inverter, it follows from the models of Fig. 7.8-4 that the reduced circuits of Fig. 7.8-10*a* and *b* are useful to estimate the logic gate delays for two-input NAND and two-input NOR gates. The delay here refers to the worst-case delay for the NOR gate (when only one of the pulldown transistors switches on). Note that if two or more NOR gate pulldown transistors switch on simultaneously, the effective resistance, and therefore the delay, will be reduced. The reduced models of Fig. 7.8-10 show only a single pulldown path for each NOR gate and lump the series pulldown transistors of a NAND gate into a single, equivalent resistor. The circuit of Fig. 7.8-10*c* is used to compare the delay for a multi-input NOR gate with the delay for the two-input gates. The parameter values for all three cases are listed in Fig. 7.8-10*d*.

Based on the analysis methods of this section, it is easy to see that the delay for the propagation of a digital signal through two cascaded NOR gates is the same as for an inverter pair, provided that both NOR gates are loaded by a single, equivalent input. This NOR-pair delay is just t_{ipd}. The delay for two cascaded NAND gates is greater because of the high-resistance pullup device required by the 8 : 1 pullup/pulldown ratio and because of the two series transistors in the pulldown path. Because the resistances for both the pullup path and pulldown path are doubled, the NAND-pair delay is $2t_{ipd}$. Because the NAND-pair delay is double the NOR-pair delay, it is obvious that the NOR configuration is preferable to the NAND configuration from a delay viewpoint. It is further noted from Fig. 7.8-10*c* and *d* that the delay for a multi-input NOR gate is the same as that for the two-input NOR gate. Multi-input NMOS NOR gates are widely used; NMOS NAND gates with more than two inputs are rare.

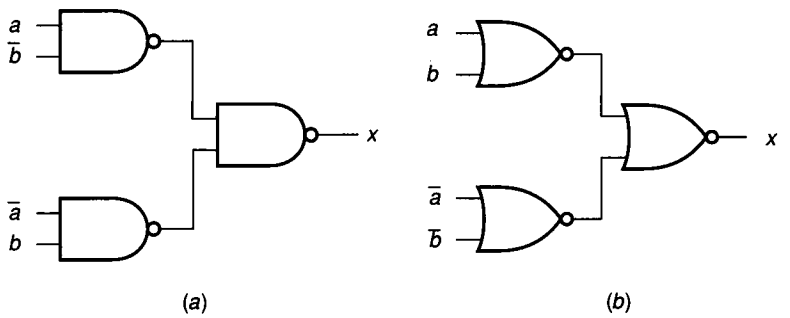
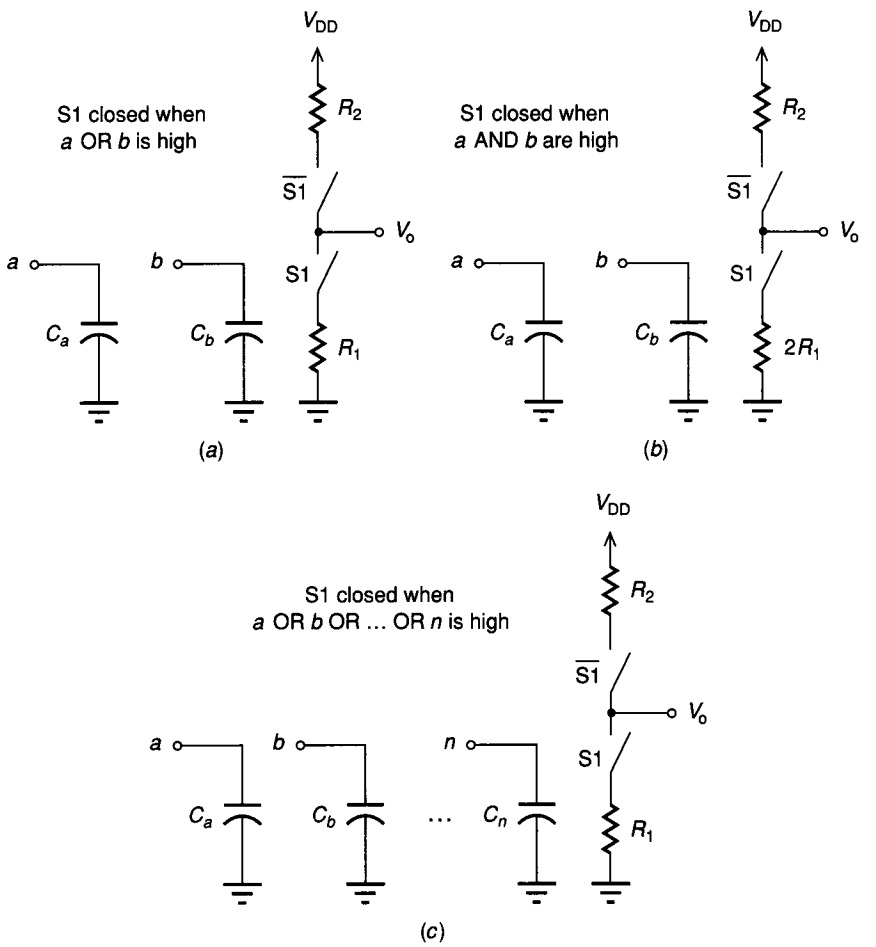


FIGURE 7.8-9
Equivalent circuits to obtain the exclusive-OR function: (*a*) NAND-NAND circuit, (*b*) NOR-NOR circuit.



	NOR	NAND	Multi-input NOR
R_1	16.7 k Ω	16.7 k Ω	16.7 k Ω
R_2	$4R_1 = 66.7$ k Ω	$8R_1 = 133.4$ k Ω	$4R_1 = 66.7$ k Ω
C_a	$C_{ox}W_a L_a$	$C_{ox}W_a L_a$	$C_{ox}W_a L_a$
C_b	$C_{ox}W_b L_b$	$C_{ox}W_b L_b$	$C_{ox}W_b L_b$
C_n	—	—	$C_{ox}W_n L_n$

(d)

FIGURE 7.8-10
Delay models for multi-input gates: (*a*) Delay model for two-input NOR, (*b*) Delay model for two-input NAND, (*c*) Delay model for multi-input NOR, (*d*) Resistance and capacitance values.

7.8.6 Enhancement versus Depletion Loads

It was shown earlier in this chapter that the depletion load is preferred to the enhancement load from a signal-swing viewpoint. Both the enhancement- and depletion-load logic gates are ratio-logic circuits. It was further stated that the same 4 : 1 sizing rule is generally used for both types of logic gates. However, within a single–power supply logic family, the dynamic performance of the two gates is substantially different.

Figure 7.8-11 shows the *I–V* characteristics for three pullup devices: an enhancement-mode transistor with its gate tied to its drain, a depletion-mode transistor with its gate tied to its source, and a linear resistor. Any one of these devices can be used with an enhancement pulldown transistor to form an inverter. All three devices can be designed to have the same equivalent large signal resistance *(V/I)* at a point *X* by suitable choice of width-to-length ratios and threshold voltages for the transistors. Let the devices be sized so that point *X* corresponds to the voltage and current for the pullup of an inverter stage when its output is low. The three devices are indistinguishable from a steady state viewpoint when used as static load devices operating at point *X*.

Now consider the dynamic operation of an inverter circuit. When the pull-down stage of the inverter is turned off, the pullup device should quickly bring the output voltage to a logic high value. This requires the lowest possible resistance for the pullup during the low-to-high transition. The *I–V* curves of Fig. 7.8-11 are particularly instructive here. The linear resistor provides a current that is directly proportional (Ohm's law) to the voltage across the load resistor, $V_{DD} - V_o$, where V_o is the output voltage of the inverter. The enhancement-mode transistor provides an equivalent resistance that tends to infinity as the output voltage rises to $V_{DD} - V_{TN}$, whereas the equivalent resistance of the depletion-mode transistor tends to a value much smaller than the linear resistor value as V_o nears V_{DD}. In the calculation of t_{LH} in Eq. 7.8-10, an equivalent resistance for the pullup

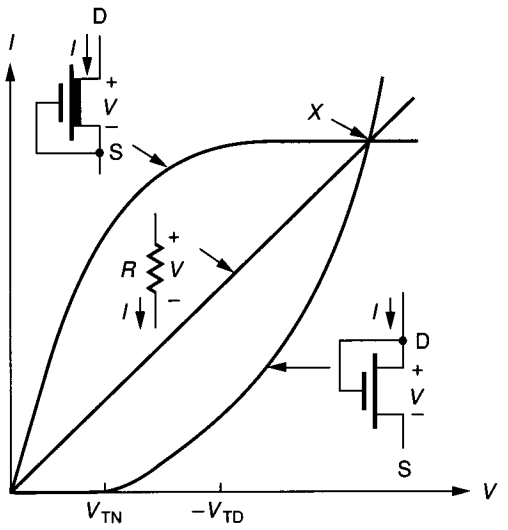


FIGURE 7.8-11
I-V curves for three load devices: depletion pullup, linear resistor, and enhancement pullup.

device determines the delay time. Thus, from the viewpoint of signal rise time, the depletion-mode transistor is a better choice than either the linear resistor or the enhancement-mode transistor. Because the output signal fall time is almost independent of the pullup device type, essentially all modern MOS ratio logic is designed with depletion pullups to reduce total gate delay times.

7.8.7 CMOS Logic Delays

So far, the delay analysis of NMOS ratio logic has been emphasized. As described in Sec 7.5, the basic CMOS inverter is not a ratio-logic device because the output is actively pulled to one logic level or the other depending on the input. Fortunately, the basic concepts of the preceding delay analysis apply equally well to CMOS logic. The unique delay characteristics of CMOS logic as compared to NMOS logic are developed in this section.

The initial delay analysis of the reference CMOS inverter is based on one inverter driving a second, identical inverter, as shown in both logic and circuit diagram form in Fig. 7.8-12. Several observations about this circuit are important here. An obvious difference from the NMOS inverter circuit of Fig. 7.8-1 is that the output of the first CMOS inverter must drive the gates of two transistors: one for the n-channel pulldown transistor and one for the p-channel pullup transistor. Both gates provide capacitive loading that slows the transition of logic signal values. Analysis of the gate capacitance for the n-channel pulldown transistor is identical to that for the NMOS inverter expressed by Eq. 7.8-8. This gate capacitance is denoted by C_{GN} and is given as

$$C_{GN} = C_G \approx C_{ox} W_N L_N \tag{7.8-23}$$

The gate capacitance of the p-channel transistor, denoted by C_{GP}, is similar to that of the gate of the n-channel transistor based on the model of Fig. 3.1-19. When the p-channel transistor is off, that is, the input voltage to the gate is a logic high, the capacitance is $C_{ox} W_P L_P$. In saturation, that is, when the input

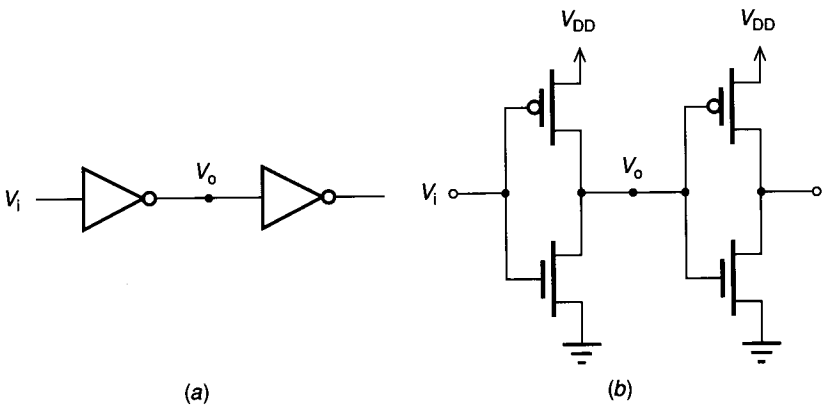


FIGURE 7.8-12
CMOS inverter driving a second, identical CMOS inverter: (*a*) Logic diagram, (*b*) Circuit diagram.

voltage to the gate is a logic low, the p-channel transistor capacitance is again $C_{ox}W_PL_P$. When the p-channel transistor is in the ohmic region, the gate capacitance is about 2/3 of its value in saturation. It is reasonable to assume that the p-channel transistor is in the ohmic region for only a short part of signal transition. For this reason, the gate capacitance of the p-channel transistor is approximated as

$$C_{GP} \approx C_{ox}W_PL_P \tag{7.8-24}$$

Because the capacitances for the two transistors of the second inverter are effectively in parallel (one to ground and the other to V_{DD}), the capacitive load seen by the first inverter is

$$C_{GC} = C_{GN} + C_{GP} \tag{7.8-25}$$

A simplified model for the CMOS inverter circuit of Fig. 7.8-12 is given in Fig. 7.8-13. This model is based on the simple resistive model of Fig. 7.8-3 for the driving transistors and the capacitance values that were just discussed. When a low-to-high step input is applied to the first inverter, the equivalent circuit of Fig. 7.8-13*b* is applicable. When a high-to-low step input is used, the equivalent circuit of Fig. 7.8-13*c* is appropriate. Approximating the 10% to 90% signal rise time by two time constants for the ideal *RC* circuit models of Fig. 7.8-13*b* and *c*, the rising and falling transition delays are given by the equations

$$t_{LH} = 2R_2C_{GC} \tag{7.8-26}$$

and

$$t_{HL} = 2R_1C_{GC} \tag{7.8-27}$$

The value of R_1 is determined using an analysis similar to that used for the NMOS inverter pulldown and is given by Eq. 7.8-4. Since the p-channel pullup

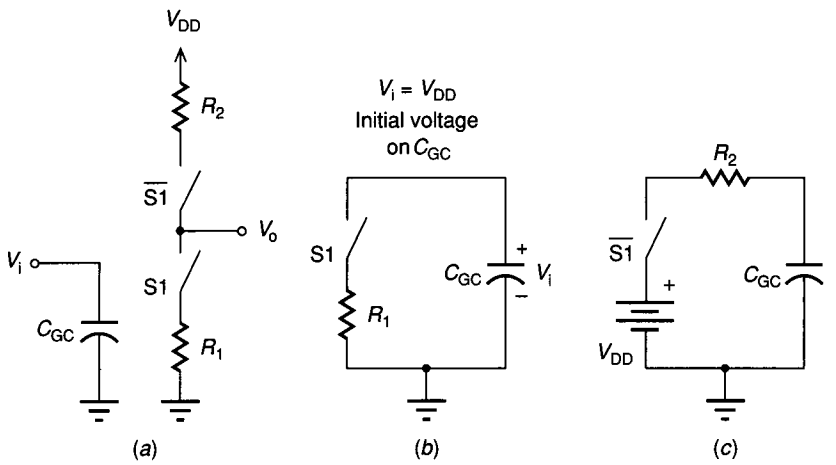


FIGURE 7.8-13
Equivalent circuits for CMOS delay analysis: (*a*) Simplified *RC* inverter model, (*b*) Equivalent circuit for high-to-low output transition, (*c*) Equivalent circuit for low-to-high output transition.

transistor is an enhancement transistor, its equivalent resistance R_2 is also given by Eq. 7.8-4 with the appropriate substitution of p-channel values for K', L, W, and V_T. If the transistors are sized for symmetrical output drive with equivalent pullup and pulldown resistances, the delay is

$$t_{LH} = t_{HL} = 2R_2C_{GC} = 2R_1C_{GC} \tag{7.8-28}$$

With the simplified RC model for a CMOS inverter as the basis, analysis of CMOS NAND and NOR gates is easy. In general, each logic gate output must drive the gate capacitance of two transistors for every connection to another CMOS logic element. The equivalent pullup and pulldown resistance for a logic gate output depends on process-dependent characterisitics for n-channel and p-channel transistors, the width-to-length ratios of the individual transistors, and the series connection required for the pulldown section of a NAND gate or the pullup section of a NOR gate.

The following example demonstrates the propagation delay analysis for CMOS inverters. Subsequently the CMOS inverter-pair delay is compared with the equivalent inverter-pair delay for NMOS inverters.

Example 7.8-1. Consider a cascade of CMOS inverters in an n-well process. The n-channel pulldown transistor is a minimum-size device in a 2 μ technology. Assume the values of K' for the n-channel and p-channel transistors are 45 μA/V^2 and 15 μA/V^2 respectively. The p-channel pullup is sized for symmetrical output drive capability. The thresholds are $V_{TN} = 1$ V and $V_{TP} = -1$ V, and the supply voltage is $V_{DD} = 5$ V. The gate-oxide capacitance for both transistor types is $C_{ox} = 1$ fF/μ^2.

Determine the inverter-pair delay for this CMOS inverter in terms of absolute time and in terms of τ_P from Eq. 7.8-14.

Solution. Because the n-channel pulldown is of minimum size, it has $W_N = L_N = 2$ μ. The p-channel pullup must have a width-to-length ratio of 3 to compensate for the relative values of K' assumed for the p- and n-channel transistors. Choose $L_P = L_N = 2$ μ and $W_P = 3L_P = 6$ μ.

From Eqs. 7.8-23, 7.8-24, and 7.8-25,

$$C_{GC} = C_{GN} + C_{GP} = C_{ox}(W_NL_N + W_PL_P) = C_{ox}L_N(W_N + W_P)$$

Then

$$C_{GC} = 1 \text{ fF}/\mu^2 \times 2\ \mu(2\ \mu + 6\ \mu) = 0.016 \text{ pF}$$

From Eq. 7.8-4,

$$R_N = \frac{2 \times 2\ \mu}{45\ \mu\text{A/V}^2 \times 2\ \mu(5 \text{ V} - 1 \text{ V})} = 11.1 \text{ k}\Omega$$

and

$$R_P = \frac{2 \times 2\ \mu}{15\ \mu\text{A/V}^2 \times 6\ \mu(5 \text{ V} - 1 \text{ V})} = 11.1 \text{ k}\Omega$$

From Eqs. 7.8-26 and 7.8-27,

$$t_{LH} = 2R_PC_{GC} = 2 \times 11.1 \text{ k}\Omega \times 0.016 \text{ pF} = 0.355 \text{ ns}$$

and from Eq. 7.8-28, for symmetrical output drive,

$$t_{HL} = 2R_N C_{GC} = t_{LH} = 0.355 \text{ ns}$$

The inverter-pair delay is

$$t_{ipd} = t_{LH} + t_{HL} = 0.355 + 0.355 = 0.71 \text{ ns}$$

From Eq. 7.8-13, the process characteristic time constant is

$$\tau_P = R_{ss} C_{GN} = 5.56 \text{ k}\Omega \times 0.004 \text{ pF} = 0.022 \text{ ns}$$

Thus, $t_{ipd} = 32\tau_P$.

The inverter-pair delay for CMOS inverters determined in the preceding example should be compared with the inverter-pair delay for the NMOS inverters of Eq. 7.8-18. For NMOS with a 4:1 size ratio, $t_{ipd} = 20\tau_P$. The two inverter-pair delays demonstrate a 50% decrease in raw circuit speed for CMOS as compared with NMOS. The analysis in terms of τ_P removes geometric dependencies from the comparison. This analysis does, however, depend on the relative values of K' for the n- and p-channel transistors but does not depend on the choice of symmetric drive for the CMOS inverter.

7.8.8 Interconnection Characteristics

The analysis up to this point has been concentrated on logic gates and process characteristics to analyze signal propagation delay. This section includes a brief introduction to the interconnection capacitance and resistance that results from the connection of the output of one gate to the input of another gate. This discussion is limited to on-chip interconnections. Such interconnections are possible using one of several layers such as one or more layers of metal, polysilicon, or diffusion.

The proper interconnection medium depends on the physical properties of the layers and on circuit topological constraints. The metal layer is the most flexible because it does not interact directly with either of the other two layers to create transistors. Each of the interconnection layers exhibits parasitic capacitance to substrate (ground) that slows signal propagation. The value of this capacitance per unit area is given for an NMOS process in Table 2A.4 of Appendix 2A and for a CMOS process in Table 2B.4 of Appendix 2B. The capacitance for metal and polysilicon is considered in terms of the classical case of parallel plates separated by a dielectric. As a first-order approximation, the metal and polysilicon capacitances are independent of voltage and geometrical shape. Capacitance densities from metal to substrate or polysilicon on field oxide to substrate are typically less than C_{ox} by a factor of 10 to 20. Capacitance from diffusion to substrate consists of two primary components: bottom capacitance per unit area and sidewall capacitance per unit length. The sidewall capacitance is given per unit length for convenience because the diffusion region is considered to be of constant depth. Diffusion capacitance is caused by the reverse-biased diode junction between diffusion and substrate and is voltage-dependent. As a simple approximation, the voltage dependence is eliminated by choosing a voltage that provides a nominal

capacitance value. For minimum-width geometries, diffusion capacitance per unit area is typically less than C_{ox} by a factor of 5 to 10. Even though the capacitances per unit area for metal, polysilicon, and diffusion are factors of 5 to 20 less than C_{ox}, the area associated with interconnection is usually much larger than the transistor gate area. As a consequence, these interconnection loading effects may represent the dominant capacitive loading on many nodes, particularly as device geometries scale down to the 1 μ range and below.

The interconnection layers exhibit resistance to current flow as given by the values in Table 2A.4 of Appendix 2A for an NMOS process and the values in Table 2B.4 of Appendix 2B for a CMOS process. The resistances of polysilicon and diffusion are typically greater than the resistance of metal by three orders of magnitude. For short interconnections between adjacent devices, the resistance of the interconnection may be safely ignored in comparison to the effective transistor resistances. For longer interconnections, polysilicon and diffusion present large resistances that cannot be ignored. For this reason, metal is used for long interconnections.

Logic building block characteristics and associated delays can be estimated as soon as a logic diagram or circuit diagram with device sizing is complete. Interconnection delays depend so heavily on circuit layout that their effect is often neglected until layout is available. This is an unfortunate situation because interconnection capacitance is a significant delay factor for digital logic circuits. A rough rule of thumb with which to consider interconnection delays for minimum-size digital circuits prior to circuit layout can be derived from the following premises.

1. Assume that the average interconnection capacitance per unit area is one-tenth that of C_{ox}.
2. Assume that the local interconnection area is 10 times the gate area.

With these assumptions, interconnections can be modeled by doubling the effective capacitance of each driven gate. Although this is a crude approximation, it is substantially better than ignoring interconnection delays until after layout is available. For a particular design style and technology, this approximation can be improved with measurements from previous, similar designs. For example, many gate array manufacturers provide average interconnection capacitance estimates based on anticipated die size. Even better estimates of interconnection capacitance for critical nodes can be obtained if a floor plan of the circuit is available to describe the relative placements of digital building blocks within the die area.

Many aspects of signal propagation delay have been examined in this section. The delay is a combination of gate delay and interconnection delay. Ratio logic was found to exhibit asymmetric rising and falling delay times that differ by the pullup/pulldown resistance ratio. A process charcteristic time constant was defined to allow unbiased delay comparisons of different processes. Inverter-pair delay was introduced to capture the effects of both rising and falling delays. Then a special circuit configuration called a superbuffer was introduced to minimize the asymmetric delay characteristics of ratio logic. Next, delay analysis of both

NMOS and CMOS gates was presented, followed by an example to compare the two technologies with respect to delay characteristics. Finally, the electrical characteristics of interconnections were introduced.

7.9 CAPACITIVE LOADING CONSIDERATIONS

In the preceding section, consideration was given to signal propagation delays for logic gates that were loaded by single, identical logic gates. A significant problem in large-scale integrated circuit design for digital circuits is driving the relatively large capacitive loads caused by high gate fanout, interconnections, and off-chip connections. As stated in Sec. 7.2, both signal-level degradation and propagation delay are considered when specifying the maximum fanout for a logic circuit. Because of the extremely high input resistance of MOS devices, minimal signal-level degradation occurs even in driving a large number of gates. The primary fanout consideration is the input capacitance of successive logic circuits and their interconnections. Now the analysis of the previous section is expanded to consider the effects of heavy capacitive loading and to investigate circuit techniques to minimize the associated increase in signal propagation delay.

7.9.1 Capacitive Loading

Several factors contribute to capacitive loading of the output of a logic gate. These include inputs to other gates, interconnection routing or buses, bonding pads, and external loads. Regardless of the cause, each adds parasitic capacitance and contributes cumulatively and nearly linearly to the overall delay. If the total capacitive loading at the output node of a logic gate caused by these factors is found to be C_T, then the propagation delay time constant can be approximated by the expression

$$\tau_T = R_T C_T \tag{7.9-1}$$

where R_T is the equivalent charging or discharging resistance. If the capacitance C_T is driven by a reference inverter with pulldown resistance R_T and gate capacitance C_G, the average propagation delay is

$$t_{dly} = \frac{t_{apd} C_T}{C_G} \tag{7.9-2}$$

where t_{apd} is the average logic stage propagation delay of the logic family defined in terms of inverter-pair delay by

$$t_{apd} = \frac{t_{ipd}}{2} \tag{7.9-3}$$

From Eq. 7.8-18, this can be written in terms of the process characteristic time constant τ_P for an NMOS reference inverter with pullup/pulldown ratio k as

$$t_{dly} = \frac{2(1 + k)\tau_P C_T}{C_G} \tag{7.9-4}$$

7.9.2 Logic Fan-out Delays

In many situations, the output of a logic stage is required to drive more than one equivalent gate input. The response time is slowed because of the parasitic capacitance of the additional inputs. If the fan-out, that is, the number of equivalent reference inverter loads to be driven, is f, then the total capacitive load is fC_G where C_G is the capacitive load of a single reference inverter. Replacing C_T of Eq. 7.9-2 with fC_G, the average stage delay for a single stage with a fan-out of f is

$$t_{stage} = t_{apd} f \tag{7.9-5}$$

With these observations, the delay along a homogeneous signal path in a digital integrated circuit can quickly be approximated. Assume a signal passes through N levels of logic with an equivalent fan-out of f_i at the ith stage. Then the total path delay is given as

$$t_{path} = t_{apd} \sum_{i=1}^{N} f_i \tag{7.9-6}$$

For the analysis so far all stages are identical to a reference inverter, and interconnection capacitance has been neglected.

It is often convenient to decompose the total capacitive loading of Eq. 7.9-1 into that caused by MOS gate loading plus that caused by other factors, such as bus or interconnection loading. Assume a node is loaded by the equivalent of f reference inverter inputs and capacitive loading C_I from interconnections. From the interconnection area and the process-dependent capacitance per unit area (see Appendices 2A and 2B), C_I can be determined. If C_G is the input capacitance of the reference inverter, then the interconnection load C_I is equivalent to that of m reference inverter loads, where

$$m = \frac{C_I}{C_G} \tag{7.9-7}$$

It follows that the average propagation delay of this node due to both fan-out and interconnection loading is given by

$$t_{node} \approx (m + f) t_{apd} \tag{7.9-8}$$

If a signal must propagate through a sequence of N stages, where the output drive of each stage is equivalent to that of the reference inverter, it follows from Eqs. 7.9-6 and 7.9-8 that the signal propagation delay through this N-stage path can be approximated by

$$t_{path} \approx t_{apd} \sum_{i=1}^{N} (m_i + f_i) \tag{7.9-9}$$

where f_i is the equivalent number of reference inverter loads and m_i is the equivalent of interconnection loads on the ith node. Including the interconnection

loading is important in most circuits and will dominate the actual gate loading for long connections or for buses. Interconnection loading is particularly significant in submicron structures.

Equation 7.9-2 has a straightforward and useful extension. Assume that an inverter is to drive a total capacitive load C_T. The drive capability for this inverter is improved by increasing the widths of both its pullup and pulldown transistors by a factor of θ over the corresponding widths for the reference inverter. Note that this improved inverter has an input capacitance θC_G. Then it can be shown that the equivalent propagation delay for this inverter is given by

$$t_{inv} = \frac{t_{apd} C_T}{\theta C_G} \tag{7.9-10}$$

Although it could be the case that all inputs to all logic gates are equal in size to that of the reference inverter, in many situations it is advantageous to use logic circuits where the input devices have nonhomogeneous sizes. For example, this might occur where transistors are individually sized to reduce the delay in driving capacitive loads.

Equation 7.9-10 can be extended to obtain the signal propagation delay of paths that contain logic gates with varying drive capabilities. Under the assumption that the reference inverter device sizing ratio k is used for all gates in the cascade, and that the drive capability of the output of the ith gate in the cascade is θ_i times that of the reference inverter (i.e., the corresponding resistances in the model of Fig. 7.8-4 are $1/\theta_i$ times those of the reference inverter), it can be shown that the signal propagation delay through an N-stage path can be approximated by

$$t_{path} \approx t_{apd} \sum_{i=1}^{N} \frac{m_i + f_i}{\theta_i} \tag{7.9-11}$$

At the expense of some layout area, the increase in drive capability can be obtained by increasing the width of the driving transistors while keeping their length and device sizing ratio k constant.

In summary, a simple method of obtaining an approximation to the signal propagation delay along a path in digital circuits based on the average stage delay has been presented. Interconnections and other parasitic loading factors are easily included in the calculation once layout is determined. The approximation may have a significant error of $\pm 50\%$ or more. The simple analysis presented here is, however, good enough for at least two purposes. First, it allows an estimate of circuit speed for use in comparing alternative designs prior to implementation. Second, it is usually adequate to determine the critical paths that must be analyzed in detail and possibly modified to improve performance. If more accurate timing information is required, a timing analysis program or circuit simulator such as SPICE can be used. It is often prudent to identify the critical delay paths in a system and perform a detailed analysis of those paths to obtain a more precise estimate of the system delay.

7.9.3 Distributed Drivers

It can be seen from Eq. 7.9-2 that the delay associated with driving a large capacitive load from a minimum-size inverter increases linearly with the load capacitance C_T. This linear dependence is particularly troublesome because situations often arise where the total load capacitance may be as much as $100C_G$ to $10{,}000C_G$ (see Table 7.9-1). The corresponding increase in delay by a factor of 100 to 10,000 is seldom acceptable. The following question naturally arises: Is there a faster way to drive a large capacitive load? At first glance, Eq. 7.9-11 suggests that if θ_i is made larger by widening the pullup and pulldown transistors to reduce their equivalent resistances, then the delay can be reduced. (Normally the pullup/pulldown ratio k is maintained to ensure valid logic voltage levels.) However, these changes cause heavier capacitive loading on previous stages, perhaps negating the net performance gain. The following example shows that, in terms of propagation delay, it may be better to distribute the load than to increase the drive of a single stage when an output is driving a high-fanout load.

Example 7.9-1. Assume that a minimum-size inverter drives a set of 10 other minimum-size inverters, as shown in Fig. 7.9-1*a*. Estimate the propagation delay from V_i to V_c if the single inverter drives the 10 inverters directly, and if the single inverter drives two other minimum-size inverters that each drive five inverters, as shown in Fig. 7.9-1*b*. Neglect interconnection capacitance and the inversion of the logic signal.

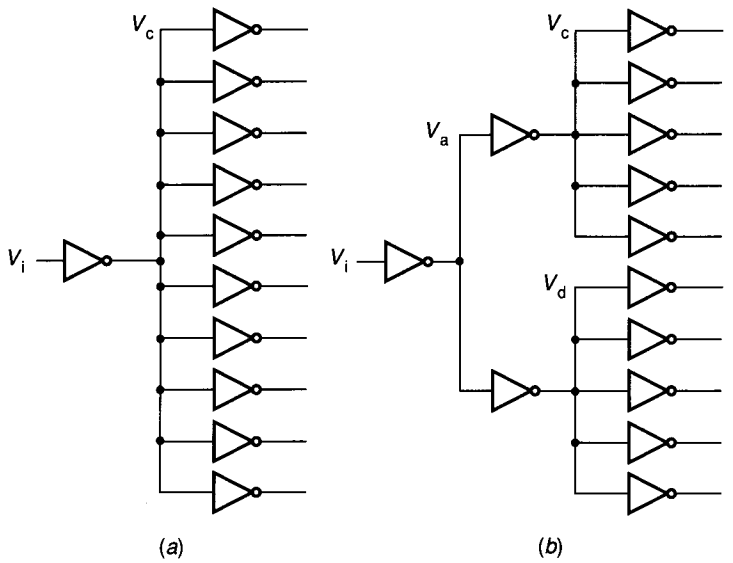


FIGURE 7.9-1
Distributed drivers versus concentrated driver: (*a*) 1:10 capacitive load, (*b*) 1:2:5 capacitive load.

Solution. Assuming that the minimum-size inverter has a pullup/pulldown ratio of $k = 4$, the average stage delay required to drive a single minimum-size inverter is t_{apd}. It follows from Eq. 7.9-2 that if the inverter drives 10 identical inverters, the total gate capacitance will be increased by a factor of 10, and the total delay becomes $10t_{apd}$.

If the minimum-size inverter drives two stages that, in turn, drive five stages each, the propagation delay will be the sum of the two delays. The average delay to drive two identical inverters will be twice the delay for a single inverter, or $2t_{apd}$. The average delay for the second stage to drive five identical stages will be five times the delay for a single inverter, or $5t_{apd}$. The combined delay from input to output with the intermediate stage will be $7t_{apd}$, which is less than the $10t_{apd}$ required for driving the 10 inputs directly.

The example shows that it may be better to include intermediate stages than to drive a heavy capacitive load from a single stage. Table 7.9-1 provides a comparison of some of the capacitive loading requirements that must be addressed for the typical process listed in Table 2A.4 of Appendix 2A. Although interconnection loading may represent capacitive loading equivalent to 100 or more reference inverters, pad loading and off-chip loading are often even larger. The divide-and-conquer strategy presented in the example will not work for a single large load. Methods of driving large, concentrated capacitive loads will be considered in the following section.

7.9.4 Driving Off-Chip Loads

Digital integrated circuits generally use the smallest possible transistors to implement logical functions. The small size is important to maximize the number of circuits per unit area and therefore minimize silicon area and cost. Ultimately, logic circuits must provide results of internal circuit operations to the outside world. These logic signals must overcome both the inherent loading effects of the bus or interconnection path from the source of the output signal to an external pin and the loading effects of some other circuit to which this output signal acts as an input. Additionally, it is frequently desirable to generate signals that are compatible, relative to logic voltage levels, with some other logic family, such as Advanced Low-Power Schottky TTL (ALSTTL), which provides current as well as capacitive loading. Other logic families are handled by increasing the

TABLE 7.9-1
Typical capacitive loading

Load	C_T	C_T/C_G
Single reference inverter (3 μ × 3 μ)	0.0063 pF	1
Ten reference inverters	0.063 pF	10
4 mm × 4.5 μ metal bus	0.450 pF	71
Standard output pad (100 μ × 100 μ)	0.250 pF	40
Oscilloscope probe	10.0 pF	1587
Memory chip address pin	5.0 pF	794

width of the driving transistors to increase current capability and by varying the pullup/pulldown ratio k to match external logic voltages. To illustrate the former case, the following example demonstrates the capacitive loading effects of driving a simple output bonding pad from a minimum-size inverter.

Example 7.9-2. Consider driving a metal output bonding pad from a minimum-size inverter in the NMOS process summarized in Table 2A.4 of Appendix 2A. Calculate the capacitance ratio between the capacitance of the bonding pad and the input gate capacitance of a minimum-size inverter. Using this ratio, estimate the delay to drive the output pad in terms of the reference delay t_{apd}.

Solution. The size of an output bonding pad is typically 100 $\mu \times 100$ μ. From the NMOS process electrical characteristics

$$C_{pad} = 0.025 \text{ fF}/\mu^2 \times 10,000 \ \mu^2 = 0.25 \text{ pF}$$

Note that this matches the value listed in Table 7.9-1. The dimensions of an input gate for a minimum-size inverter in this process are 3 $\mu \times 3$ μ.

$$C_G = 0.7 \text{ fF}/\mu^2 \times 9 \ \mu^2 = 0.0063 \text{ pF}$$

Thus, the ratio of C_{pad} to C_G is

$$C_{ratio} = \frac{C_{pad}}{C_G} = 39.7$$

From Eq. 7.9-2 the average propagation delay of the reference inverter driving the output pad is

$$t_{dly} = 39.7 t_{apd}$$

From Eqs. 7.8-14, 7.8-18, and 7.9-3, the value for the average gate delay for the NMOS process of Appendix 2A with $k = 4$ is $t_{apd} = 0.63$ ns. The average propagation delay in driving the bonding pad is thus 25 ns.

The following example demonstrates the effect of adding an external capacitive load to the bonding pad just considered.

Example 7.9-3. Consider the case of a standard oscilloscope probe connected directly to the output bonding pad of Example 7.9-2. Determine the approximate average propagation delay that will result.

Solution. The total load capacitance will be the sum of the output pad capacitance and the oscilloscope probe capacitance. Table 7.9-1 indicates that an oscilloscope probe provides 10 pF of capacitive load. Thus,

$$C_{load} = C_{pad} + C_{probe} = 0.25 \text{ pF} + 10 \text{ pF} = 10.25 \text{ pF}$$

The capacitance ratio will be

$$C_{ratio} = \frac{C_{load}}{C_G} = 1627$$

The average propagation delay is obtained from Eq. 7.9-2:

$$t_{dly} = 1627 t_{apd} = 1627 \times 0.63 \text{ ns} = 1025 \text{ ns}$$

Contrast this delay with the 0.63 ns average propagation delay of the reference inverter. Such a delay is obviously detrimental to high-speed operation of digital circuits, limiting clock frequency to less than $1/(2t_{dly}) = 488$ kHz! The oscilloscope probe capacitance is comparable in value to the total capacitance encountered in driving inputs on other integrated circuit chips (see Table 7.9-1).

7.9.5 Cascaded Drivers

It is obvious from the two examples of the preceding section that the signal delay encountered in driving off-chip loads directly from a minimum-size inverter is unacceptable. Fortunately, there are circuit configurations that reduce the effective delay in driving large capacitive loads. One good circuit configuration employs a cascade of inverters with increasing current-drive capability to minimize this delay.

Assume a signal is available at the output of a minimum-size inverter (reference inverter) and that it is to drive a load C_L. From Eq. 7.9-2 the average propagation delay associated with driving this load directly is

$$t_{dir} = \frac{t_{apd} C_L}{C_G} \tag{7.9-12}$$

where t_{apd} is the average logic stage delay and C_G is the input capacitance of the reference inverter. For any integer $n \geq 1$, define α by the expression

$$\alpha = \left(\frac{C_L}{C_G}\right)^{1/n} \tag{7.9-13}$$

Alternatively, n can be represented in terms of α as

$$n = \frac{\ln(C_L/C_G)}{\ln \alpha} \tag{7.9-14}$$

Consider now the alternative structure of Fig. 7.9-2 for driving a load C_L. This structure is composed of a cascade of n inverters (including the initial reference inverter) each sized by the 4 : 1 sizing rule and each with a drive capability that is α times as large as the previous stage. The width and length of the kth stage can be characterized by the equations

$$\begin{aligned} W_{dk} &= \alpha^{k-1} W_{d1} \\ L_{dk} &= L_{d1} \\ W_{uk} &= W_{dk} \\ L_{uk} &= 4L_{dk} \end{aligned} \tag{7.9-15}$$

where the device dimensions W_{dk} and L_{dk} correspond to the pulldown transistor and W_{uk} and L_{uk} correspond to the pullup transistor of the kth inverter structure in the cascade, as indicated in Fig. 7.9-2. It can be observed that the load on the kth stage C_{Lk} is related to the reference inverter input capacitance C_G by the expression

$$C_{Lk} = \alpha^k C_G \tag{7.9-16}$$

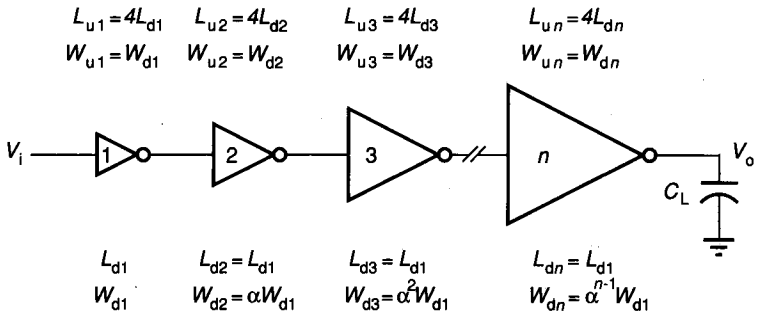


FIGURE 7.9-2
Cascaded drivers for a concentrated load.

From Eq. 7.9-2 it follows that the average propagation delay of the first inverter is αt_{apd}. It can be shown (Prob. 7.41) that the average propagation delay for each inverter in this geometric cascade is αt_{apd}. Hence, it follows from Eq. 7.9-11 with $f_i/\theta_i = \alpha$ and $m_i = 0$, that the total delay for the cascade is

$$t_{cas} = n\alpha t_{apd} \tag{7.9-17}$$

Let r be the ratio between the propagation delays of the direct-drive circuit and of the geometric cascade approach. From Eqs. 7.9-12 and 7.9-17 it follows that

$$r = \frac{t_{cas}}{t_{dir}} = \frac{n\alpha t_{apd}}{t_{apd}C_L/C_G} = \frac{n\alpha C_G}{C_L} \tag{7.9-18}$$

It is our goal to determine n and α to minimize r and thus minimize the propagation delay in driving the load. From Eq. 7.9-14 it follows that n can be eliminated from the expression for r to obtain the expression

$$r = \frac{\ln(C_L/C_G)}{C_L/C_G}\frac{\alpha}{\ln\alpha} \tag{7.9-19}$$

The terms involving capacitance are fixed by the load requirements. The goal is thus to determine α in Eq. 7.9-19 to minimize r. The second term on the right-hand side of Eq. 7.9-19 is plotted in Fig. 7.9-3. It is easy to see that $\alpha/\ln\alpha$ has a wide local minimum at $\alpha = e$ with value e. A plot of the number of stages n versus C_L/C_G for minimizing the delay with $\alpha = e$, as obtained from Eq. 7.9-14, is shown in Fig. 7.9-4. Plots of n versus C_L/C_G for $\alpha = 3$ and $\alpha = 5$ are also shown in Fig. 7.9-4.

It should be noted that because n is the number of cascade stages, n must assume an integer value greater than or equal to 1. Quantization of device geometries during layout precludes setting α to an exact ratio of e. In fact, α is usually set to a value greater than e to reduce the number of cascade stages required while still reducing the propagation delay significantly. As can be seen from Fig. 7.9-3, as long as α is between 2 and 4, the deviation from minimum delay is less than about 5%. For conservative design, the values for n and α can be selected to drive a load a little larger than C_L within the allotted delay time.

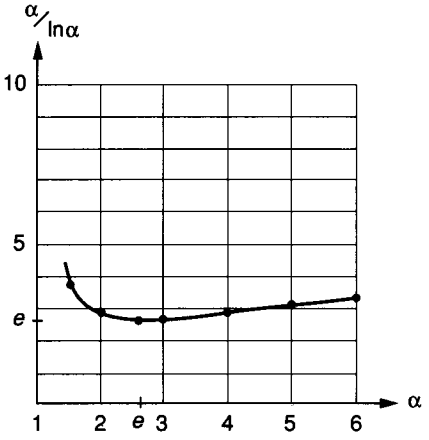


FIGURE 7.9-3
Plot of $\alpha/\ln\alpha$ versus α.

It is instructive to obtain an appreciation for how much benefit is practically derived from the cascaded driver approach. It can be shown from Eq. 7.9-17 that for small loading ratios, the speed improvements are small and the area overhead associated with the cascade may not be justified. For large capacitive load ratios, the speed improvement offered by the cascade is significant. For example, from Eq. 7.9-18 a seven-stage optimally sized cascade would drive a capacitive load that is approximately $1100C_G$ in 1.7% of the time required with direct drive!

Two final points deserve mention. First, if the number of inverter stages in the cascade is odd, the output signal is inverted. If inversion is unacceptable, a minimal delay increase occurs if the circuit is preceded by a reference inverter. Second, even though the speed improvements are significant for large n, the silicon area penalty is also quite high. The active silicon area grows as a geometric function of the number of stages. For example, the final stage in a seven-stage optimally weighted cascade requires $e^6 = 403$ times as much active area as a reference inverter. This is almost twice the area of all the preceding cascaded stages combined.

The following example compares the delay for three stages and the optimal number of stages driving the oscilloscope probe of Example 7.9-3.

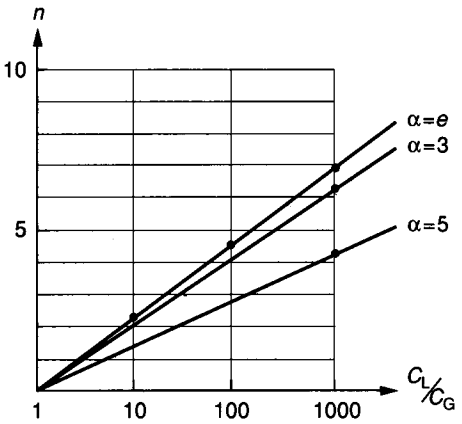


FIGURE 7.9-4
Plot of n versus C_L/C_G for $\alpha = 3, e$, and 5.

Example 7.9-4. Determine the number of stages required to drive the output pad of Example 7.9-3 with minimum delay. Calculate the delay for this minimum delay case and for the case where three stages sized according to Eq. 7.9-13 are used.

Solution. From Eq. 7.9-14 the optimal number of stages n is found with $\alpha = e$. From Example 7.9-3, $C_L/C_G = 1627$, so $n = \ln(1627) = 7.39$. This is rounded to seven stages, including the reference inverter at the signal source. The propagation delay for the seven stages is given by Eqs. 7.9-17 and 7.9-13 as

$$t_7 = 7 \times 2.88 \times t_{apd} = 20.16 t_{apd}$$

Note from Eq. 7.9-17 that each stage contributes equally to the delay.

If $n = 3$ is chosen, then from Eq. 7.9-13, $\alpha = 11.76$ for $C_L/C_G = 1627$. From Eq. 7.9-17, this gives a delay of

$$t_3 = 3 \times 11.76 t_{apd} = 35.3 t_{apd}$$

Note that the delay for three cascaded driver stages is only 75% more than the delay for seven cascaded driver stages. If this delay is acceptable, the area savings is significant. Even with just three cascaded driver stages, the delay is reduced to only 2% of the delay for the load driven directly from a reference inverter.

Standard output driver stages are needed to drive most output nodes for high-speed digital circuits. These circuits are called *pad drivers* and are usually available in libraries of standard circuits. Pad drivers are generally just a cascade of inverters with a geometrically increasing drive capability sized to give reduced delay. Figure 7.9-5 shows a plot of a Low-Power Schottky TTL-compatible output pad driver circuit. Figure 7.9-6 gives the circuit schematic showing the number of stages and the device sizing for each stage. Note the use of two superbuffers and of enhancement transistors for both the pullup and pulldown in the final stage. The use of an active enhancement pullup prevents static power dissipation in the unloaded pad driver output stage.

The analysis in this section has covered the various aspects of large capacitive loads. First, delays from capacitive loading and from logic fanout were demonstrated. Then the delays caused by driving large off-chip capacitances were analyzed. Two approaches to driving these loads were presented: distributed drivers and optimally cascaded drivers.

7.10 POWER DISSIPATION

One major limitation of MOS ICs is internal power dissipation. Although the power dissipation of MOS circuits is generally much less than that of bipolar integrated circuits performing the same function, it becomes a major factor limiting the size of VLSI MOS circuits. Electrical power dissipation in an integrated circuit is converted to heat that must be removed through the circuit's packaging. Integrated circuit packaging offers a resistance to heat removal. The heat flow through this resistance generates a temperature difference across the package analogous to the voltage difference caused by current flow through an electrical resistance. For a given integrated circuit package, a specified maximum temperature of the integrated circuit, and a specified ambient temperature, a maximum

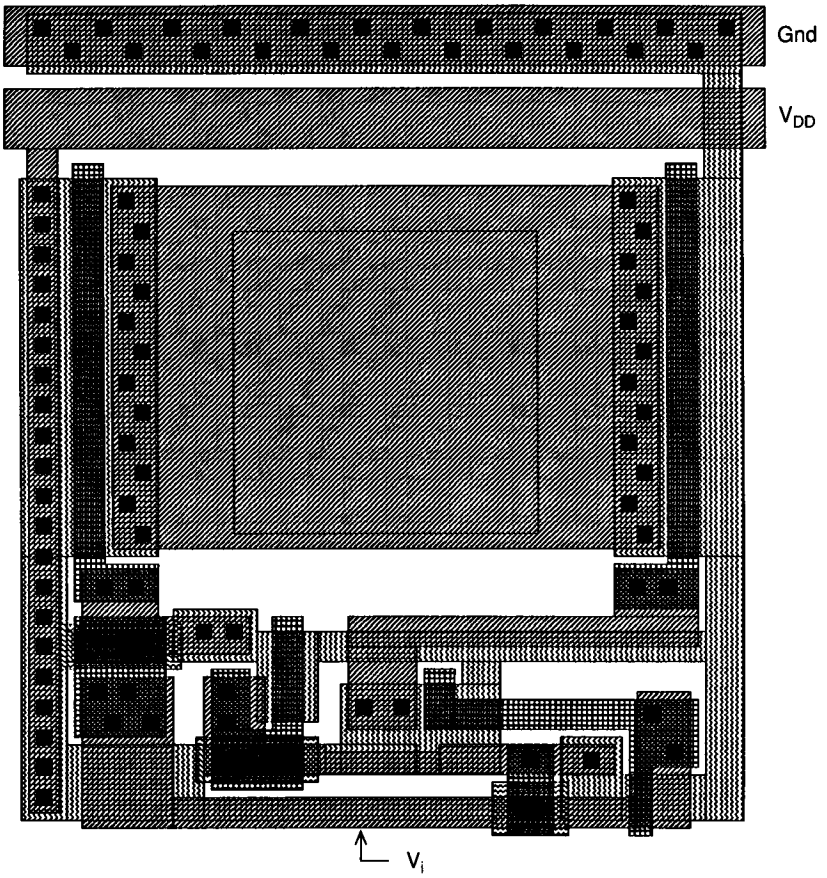


FIGURE 7.9-5
LSTTL output pad driver layout.

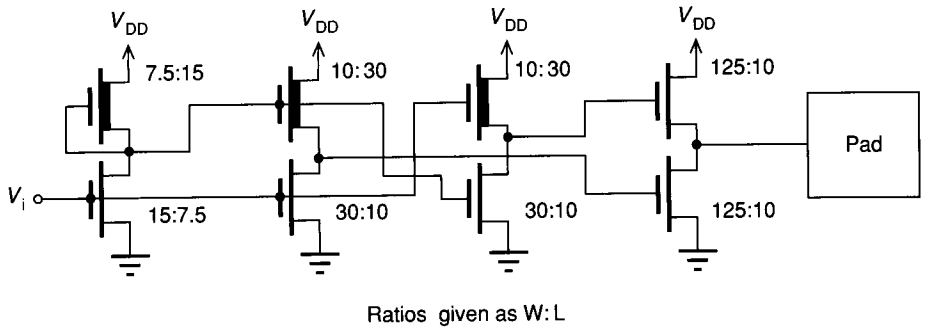


FIGURE 7.9-6
Circuit diagram for LSTTL output pad driver.