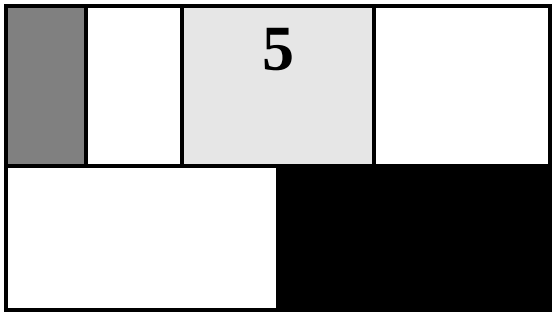
CHAPTER



THE CMOS INVERTER

Quantification of integrity, performance, and energy metrics of an inverter
Optimization of an inverter design

5.1 Introduction

The inverter is truly the nucleus of all digital designs. Once its operation and properties are clearly understood, designing more intricate structures such as NAND gates, adders, multipliers, and microprocessors is greatly simplified. The electrical behavior of these complex circuits can be almost completely derived by extrapolating the results obtained for inverters. The analysis of inverters can be extended to explain the behavior of more complex gates such as NAND, NOR, or XOR, which in turn form the building blocks for modules such as multipliers and processors.

In this chapter, we focus on one single incarnation of the inverter gate, being the static CMOS inverter — or the CMOS inverter, in short. This is certainly the most popular at present, and therefore deserves our special attention. We analyze the gate with respect to the different design metrics that were outlined in Chapter 1:

- *cost*, expressed by the complexity and area
- *integrity and robustness*, expressed by the static (or steady-state) behavior
- *performance*, determined by the dynamic (or transient) response
- *energy efficiency*, set by the energy and power consumption

From this analysis arises a model of the gate that will help us to identify the parameters of the gate and to choose their values so that the resulting design meets desired specifications. While each of these parameters can be easily quantified for a given technology, we also discuss how they are affected by *scaling of the technology*.

While this Chapter focuses uniquely on the CMOS inverter, we will see in the following Chapter that the same methodology also applies to other gate topologies.

5.2 The Static CMOS Inverter — An Intuitive Perspective

Figure 5.1 shows the circuit diagram of a static CMOS inverter. Its operation is readily understood with the aid of the simple switch model of the MOS transistor, introduced in Chapter 3 (Figure 3.25): the transistor is nothing more than a switch with an infinite off-resistance (for $|V_{GS}| < |V_T|$), and a finite on-resistance (for $|V_{GS}| > |V_T|$). This leads to the

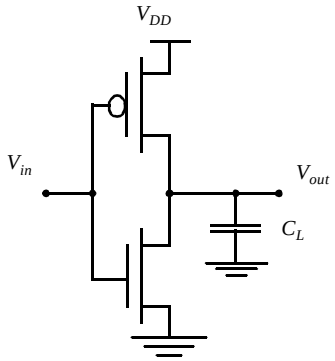
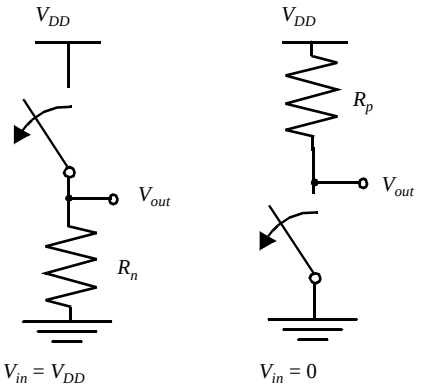


Figure 5.1 Static CMOS inverter. V_{DD} stands for the supply voltage.

following interpretation of the inverter. When V_{in} is high and equal to V_{DD}, the NMOS transistor is on, while the PMOS is off. This yields the equivalent circuit of Figure 5.2a. A direct path exists between V_{out} and the ground node, resulting in a steady-state value of 0 V. On the other hand, when the input voltage is low (0 V), NMOS and PMOS transistors are off and on, respectively. The equivalent circuit of Figure 5.2b shows that a path exists between V_{DD} and V_{out}, yielding a high output voltage. The gate clearly functions as an inverter.



(a) Model for high input (b) Model for low input

Figure 5.2 Switch models of CMOS inverter.

A number of other important properties of static CMOS can be derived from this switch-level view:

- The high and low output levels equal V_{DD} and *GND*, respectively; in other words, the voltage swing is equal to the supply voltage. This results in high noise margins.
- The logic levels are not dependent upon the relative device sizes, so that the transistors can be minimum size. Gates with this property are called *ratioless*. This is in contrast with *ratioed logic*, where logic levels are determined by the relative dimensions of the composing transistors.
- In steady state, there always exists a path with finite resistance between the output and either V_{DD} or *GND*. A well-designed CMOS inverter, therefore, has a *low output impedance*, which makes it less sensitive to noise and disturbances. Typical values of the output resistance are in kΩ range.
- The *input resistance* of the CMOS inverter is extremely high, as the gate of an MOS transistor is a virtually perfect insulator and draws no dc input current. Since the input node of the inverter only connects to transistor gates, the steady-state input current is nearly zero. A single inverter can theoretically drive an infinite number of gates (or have an infinite fan-out) and still be functionally operational; however, increasing the fan-out also increases the propagation delay, as will become clear below. So, although fan-out does not have any effect on the steady-state behavior, it degrades the transient response.

- No direct path exists between the supply and ground rails under steady-state operating conditions (this is, when the input and outputs remain constant). The absence of current flow (ignoring leakage currents) means that the gate does not consume any static power.

SIDELINE: The above observation, while seemingly obvious, is of crucial importance, and is one of the primary reasons CMOS is the digital technology of choice at present. The situation was very different in the 1970s and early 1980s. All early microprocessors, such as the Intel 4004, were implemented in a pure NMOS technology. The lack of complementary devices (such as the NMOS and PMOS transistor) in such a technology makes the realization of inverters with zero static power non-trivial. The resulting static power consumption puts a firm upper bound on the number of gates that can be integrated on a single die; hence the forced move to CMOS in the 1980s, when scaling of the technology allowed for higher integration densities.

The nature and the form of the voltage-transfer characteristic (VTC) can be graphically deduced by superimposing the current characteristics of the NMOS and the PMOS devices. Such a graphical construction is traditionally called *a load-line plot*. It requires that the *I-V* curves of the NMOS and PMOS devices are transformed onto a common coordinate set. We have selected the input voltage V_{in}, the output voltage V_{out} and the NMOS drain current I_{DN} as the variables of choice. The PMOS *I-V* relations can be translated into this variable space by the following relations (the subscripts *n* and *p* denote the NMOS and PMOS devices, respectively):

$$\begin{aligned} I_{DSp} &= -I_{DSn} \\ V_{GSn} = V_{in} \;;\; V_{GSp} &= V_{in} - V_{DD} \\ V_{DSn} = V_{out} \;;\; V_{DSp} &= V_{out} - V_{DD} \end{aligned} \tag{5.1}$$

The load-line curves of the PMOS device are obtained by a mirroring around the *x*-axis and a horizontal shift over V_{DD}. This procedure is outlined in Figure 5.3, where the subsequent steps to adjust the original PMOS *I-V* curves to the common coordinate set V_{in}, V_{out} and I_{Dn} are illustrated.

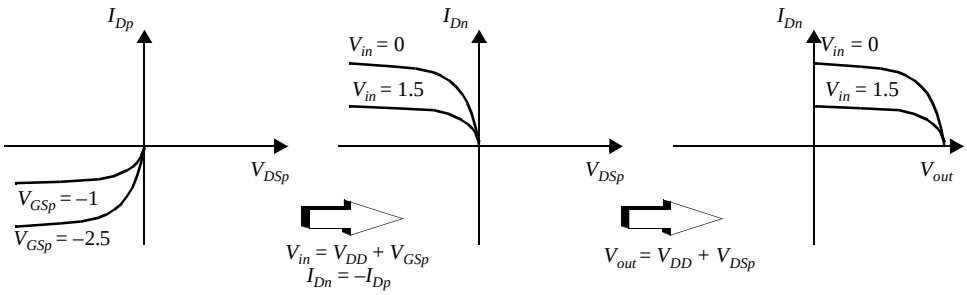


Figure 5.3 Transforming PMOS *I-V* characteristic to a common coordinate set (assuming VDD = 2.5 V).

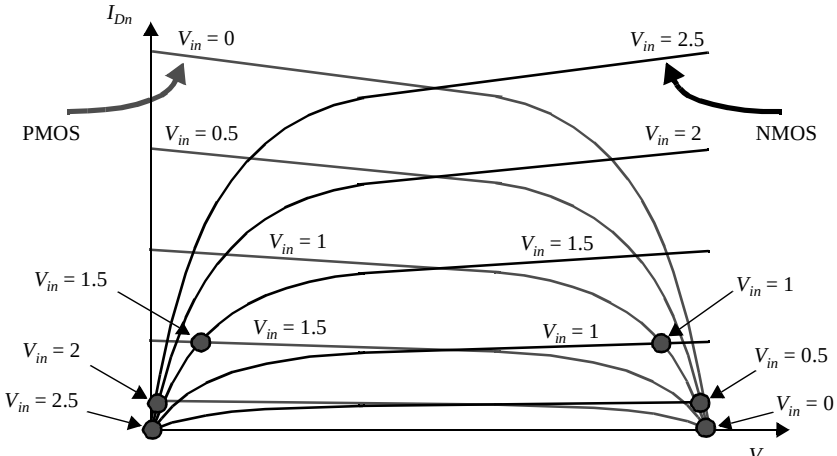


Figure 5.4 Load curves for NMOS and PMOS transistors of the static CMOS inverter (V_{DD} = 2.5 V). The dots represent the dc operation points for various input voltages.

The resulting load lines are plotted in Figure 5.4. For a dc operating points to be valid, the currents through the NMOS and PMOS devices must be equal. Graphically, this means that the dc points must be located at the intersection of corresponding load lines. A number of those points (for V_{in} = 0, 0.5, 1, 1.5, 2, and 2.5 V) are marked on the graph. As can be observed, all operating points are located either at the high or low output levels. The VTC of the inverter hence exhibits a very narrow transition zone. This results from the high gain during the switching transient, when both NMOS and PMOS are simultaneously on, and in saturation. In that operation region, a small change in the input voltage results in a large output variation. All these observations translate into the VTC of Figure 5.5.

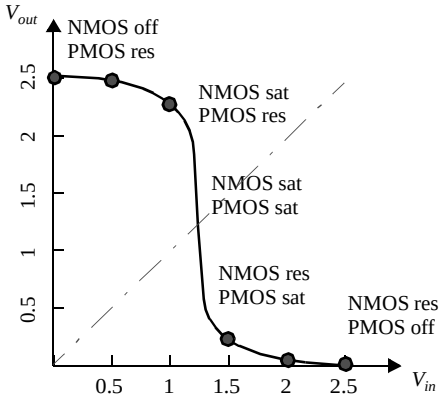


Figure 5.5 VTC of static CMOS inverter, derived from Figure 5.4 (V_{DD} = 2.5 V). For each operation region, the modes of the transistors are annotated — off, res(istive), or sat(urated).

Before going into the analytical details of the operation of the CMOS inverter, a qualitative analysis of the transient behavior of the gate is appropriate as well. This response is dominated mainly by the output capacitance of the gate, C_L, which is com-

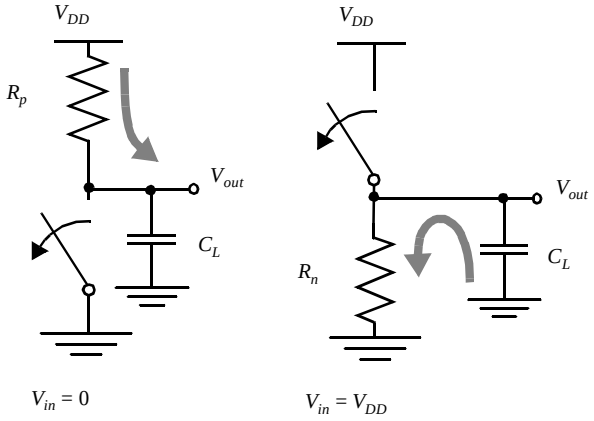


Figure 5.6 Switch model of dynamic behavior of static CMOS inverter.

posed of the drain diffusion capacitances of the NMOS and PMOS transistors, the capacitance of the connecting wires, and the input capacitance of the fan-out gates. Assuming temporarily that the transistors switch instantaneously, we can get an approximate idea of the transient response by using the simplified switch model again (Figure 5.6). Let us consider the low-to-high transition first (Figure 5.6a). The gate response time is simply determined by the time it takes to charge the capacitor C_L through the resistor R_p. In Example 4.5, we learned that the propagation delay of such a network is proportional to the its time-constant R_pC_L. **Hence, a fast gate is built either by keeping the output capacitance small or by decreasing the on-resistance of the transistor.** The latter is achieved by increasing the *W*/*L* ratio of the device. Similar considerations are valid for the high-to-low transition (Figure 5.6b), which is dominated by the R_nC_L time-constant. The reader should be aware that the on-resistance of the NMOS and PMOS transistor is not constant, but is a nonlinear function of the voltage across the transistor. This complicates the exact determination of the propagation delay. An in-depth analysis of how to analyze and optimize the performance of the static CMOS inverter is offered in Section 5.4.

5.3 Evaluating the Robustness of the CMOS Inverter: The Static Behavior

In the qualitative discussion above, the overall shape of the voltage-transfer characteristic of the static CMOS inverter was derived, as were the values of V_{OH} and V_{OL} (V_{DD} and *GND*, respectively). It remains to determine the precise values of V_M, V_{IH}, and V_{IL} as well as the noise margins.

5.3.1 Switching Threshold

The switching threshold, V_M, is defined as the point where $V_{in} = V_{out}$. Its value can be obtained graphically from the intersection of the VTC with the line given by $V_{in} = V_{out}$ (see Figure 5.5). In this region, both PMOS and NMOS are always saturated, since $V_{DS} = V_{GS}$. An analytical expression for V_M is obtained by equating the currents through the tran-

sistors. We solve the case where the supply voltage is high so that the devices can be assumed to be velocity-saturated (or $V_{DSAT} < V_M - V_T$). We furthermore ignore the channel-length modulation effects.

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right) = 0 \qquad (5.2)$$

Solving for V_M yields

$$V_M = \frac{\left(V_{Tn} + \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1 + r} \quad \text{with} \quad r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = \frac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n} \qquad (5.3)$$

assuming identical oxide thicknesses for PMOS and NMOS transistors. For large values of V_{DD} (compared to threshold and saturation voltages), Eq. (5.3) can be simplified:

$$V_M \approx \frac{r V_{DD}}{1 + r} \qquad (5.4)$$

Eq. (5.4) states that the switching threshold is set by the ratio r, which compares the relative driving strengths of the PMOS and NMOS transistors. It is generally considered to be desirable for V_M to be located around the middle of the available voltage swing (or at $V_{DD}/2$), since this results in comparable values for the low and high noise margins. This requires r to be approximately 1, which is equivalent to sizing the PMOS device so that $(W/L)_p = (W/L)_n \times (V_{DSATn}k'_n)/(V_{DSATn}k'_p)$. To move V_M upwards, a larger value of r is required, which means making the PMOS wider. Increasing the strength of the NMOS, on the other hand, moves the switching threshold closer to GND.

From Eq. (5.2), we can derive the required ratio of PMOS versus NMOS transistor sizes such that the switching threshold is set to a desired value V_M. When using this expression, please make sure that the assumption that both devices are velocity-saturated still holds for the chosen operation point.

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k'_n V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k'_p V_{DSATp}(V_{DD} - V_M + V_{Tp} + V_{DSATp}/2)} \qquad (5.5)$$

Problem 5.1 Inverter switching threshold for long-channel devices, or low supply-voltages.

The above expressions were derived under the assumption that the transistors are velocity-saturated. When the PMOS and NMOS are long-channel devices, or when the supply voltage is low, velocity saturation does not occur ($V_M - V_T < V_{DSAT}$). Under these circumstances, Eq. (5.6) holds for V_M. Derive.

$$V_M = \frac{V_{Tn} + r(V_{DD} + V_{Tp})}{1 + r} \quad \text{with } r = \sqrt{\frac{-k_p}{k_n}} \qquad (5.6)$$

Design Technique — Maximizing the noise margins

When designing static CMOS circuits, it is advisable to balance the driving strengths of the transistors by making the PMOS section wider than the NMOS section, if one wants to maximize the noise margins and obtain symmetrical characteristics. The required ratio is given by Eq. (5.5).

Example 5.1 Switching threshold of CMOS inverter

We derive the sizes of PMOS and NMOS transistors such that the switching threshold of a CMOS inverter, implemented in our generic 0.25 μm CMOS process, is located in the middle between the supply rails. We use the process parameters presented in Example 3.7, and assume a supply voltage of 2.5 V. The minimum size device has a width/length ratio of 1.5. With the aid of Eq. (5.5), we find

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115 \times 10^{-6}}{30 \times 10^{-6}} \times \frac{0.63}{1.0} \times \frac{(1.25 - 0.43 - 0.63/2)}{(1.25 - 0.4 - 1.0/2)} = 3.5$$

Figure 5.7 plots the values of switching threshold as a function of the PMOS/NMOS ratio, as obtained by circuit simulation. The simulated PMOS/NMOS ratio of 3.4 for a 1.25 V switching threshold confirms the value predicted by Eq. (5.5).

An analysis of the curve of Figure 5.7 produces some interesting observations:

1. V_M is relatively insensitive to variations in the device ratio. This means that small variations of the ratio (e.g., making it 3 or 2.5) do not disturb the transfer characteristic that much. It is therefore an accepted practice in industrial designs to set the width of the PMOS transistor to values smaller than those required for exact symmetry. For the above example, setting the ratio to 3, 2.5, and 2 yields switching thresholds of 1.22 V, 1.19 V, and 1.13 V, respectively.

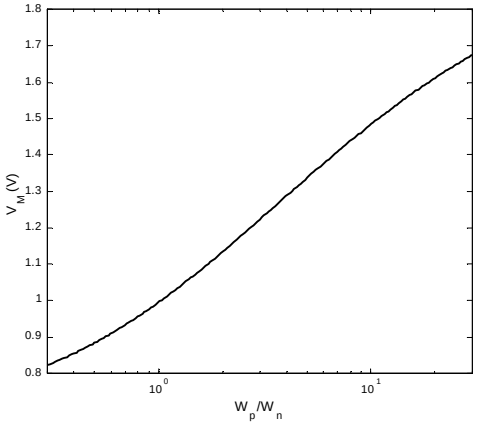
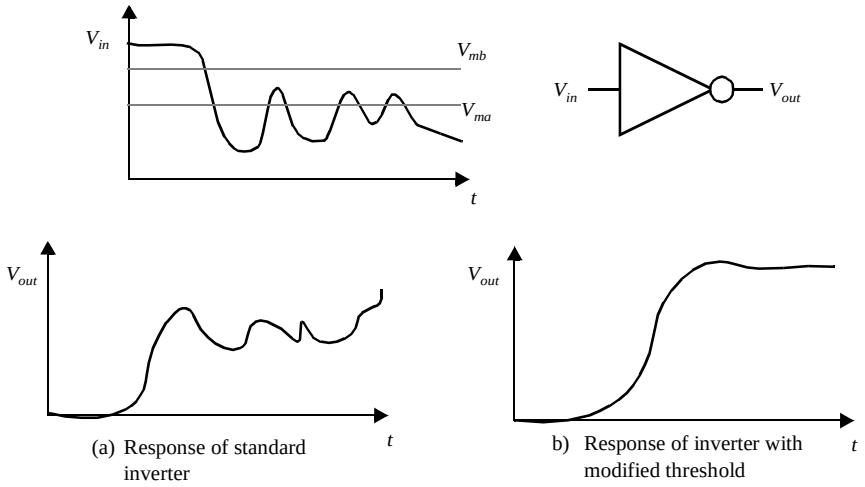


Figure 5.7 Simulated inverter switching threshold versus PMOS/NMOS ratio (0.25 μm CMOS, V_{DD} = 2.5 V)



(a) Response of standard inverter

b) Response of inverter with modified threshold

Figure 5.8 Changing the inverter threshold can improve the circuit reliability.

2. The effect of changing the W_p/W_n ratio is to shift the transient region of the VTC. Increasing the width of the PMOS or the NMOS moves V_M towards V_{DD} or *GND* respectively. This property can be very useful, as asymmetrical transfer characteristics are actually desirable in some designs. This is demonstrated by the example of Figure 5.8. The incoming signal V_{in} has a very noisy zero value. Passing this signal through a symmetrical inverter would lead to erroneous values (Figure 5.8a). This can be addressed by raising the threshold of the inverter, which results in a correct response (Figure 5.8b). Further in the text, we will see other circuit instances where inverters with asymetrical switching thresholds are desirable.
Changing the switching threshold by a considerable amount is however not easy, especially when the ratio of supply voltage to transistor threshold is relatively small (2.5/0.4 = 6 for our particular example). To move the threshold to 1.5 V requires a transistor ratio of 11, and further increases are prohibitively expensive. Observe that Figure 5.7 is plotted in a semilog format.

5.3.2 Noise Margins

By definition, V_{IH} and V_{IL} are the operational points of the inverter where $\frac{dV_{out}}{dV_{in}} = -1$. In the terminology of the analog circuit designer, these are the points where the gain g of the amplifier, formed by the inverter, is equal to −1. While it is indeed possible to derive analytical expressions for V_{IH} and V_{IL}, these tend to be unwieldy and provide little insight in what parameters are instrumental in setting the noise margins.

A simpler approach is to use a piecewise linear approximation for the VTC, as shown in Figure 5.9. The transition region is approximated by a straight line, the gain of which equals the gain g at the switching threshold V_M. The crossover with the V_{OH} and the V_{OL} lines is used to define V_{IH} and V_{IL} points. The error introduced is small and well

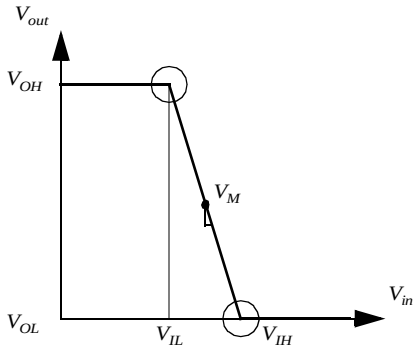


Figure 5.9 A piece-wise linear approximation of the VTC simplifies the derivation of V_{IL} and V_{IH}.

within the range of what is required for an initial design. This approach yields the following expressions for the width of the transition region V_{IH} - V_{IL}, V_{IH}, V_{IL}, and the noise margins NM_H and NM_L.

$$\begin{aligned} V_{IH} - V_{IL} &= -\frac{(V_{OH} - V_{OL})}{g} = \frac{-V_{DD}}{g} \\ V_{IH} &= V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD} - V_M}{g} \\ NM_H &= V_{DD} - V_{IH} \qquad NM_L = V_{IL} \end{aligned} \tag{5.7}$$

These expressions make it increasingly clear that a high gain in the transition region is very desirable. In the extreme case of an infinite gain, the noise margins simplify to V_{OH} - V_M and V_M - V_{OL} for NM_H and NM_L, respectively, and span the complete voltage swing.

Remains us to determine the midpoint gain of the static CMOS inverter. We assume once again that both PMOS and NMOS are velocity-saturated. It is apparent from Figure 5.4 that the gain is a strong function of the slopes of the currents in the saturation region. The channel-length modulation factor hence cannot be ignored in this analysis — doing so would lead to an infinite gain. The gain can now be derived by differentiating the current equation (5.8), valid around the switching threshold, with respect to V_{in}.

$$\begin{aligned} &k_n V_{DSATn}\left(V_{in} - V_{Tn} - \frac{V_{DSATn}}{2}\right)(1 + \lambda_n V_{out}) + \\ &k_p V_{DSATp}\left(V_{in} - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right)(1 + \lambda_p V_{out} - \lambda_p V_{DD}) = 0 \end{aligned} \tag{5.8}$$

Differentiation and solving for dV_{out}/dV_{in} yields

$$\frac{dV_{out}}{dV_{in}} = -\frac{k_n V_{DSATn}(1 + \lambda_n V_{out}) + k_p V_{DSATp}(1 + \lambda_p V_{out} - \lambda_p V_{DD})}{\lambda_n k_n V_{DSATn}(V_{in} - V_{Tn} - V_{DSATn}/2) + \lambda_p k_p V_{DSATp}(V_{in} - V_{DD} - V_{Tp} - V_{DSATp}/2)} \tag{5.9}$$

Ignoring some second-order terms, and setting $V_{in} = V_M$ results in the gain expression,

$$g = -\frac{1}{I_D(V_M)}\frac{k_n V_{DSATn} + k_p V_{DSATp}}{\lambda_n - \lambda_p} \approx \frac{1+r}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)} \tag{5.10}$$

with $I_D(V_M)$ the current flowing through the inverter for $V_{in} = V_M$. The gain is almost purely determined by technology parameters, especially the channel length modulation. It can only in a minor way be influenced by the designer through the choice of supply and switching threshold voltages.

Example 5.2 Voltage transfer characteristic and noise margins of CMOS Inverter

Assume an inverter in the generic 0.25 μm CMOS technology designed with a PMOS/NMOS ratio of 3.4 and with the NMOS transistor minimum size (W = 0.375 μm, L = 0.25 μm, W/L = 1.5). We first compute the gain at V_M (= 1.25 V),

$$I_D(V_M) = 1.5 \times 115 \times 10^{-6} \times 0.63 \times (1.25 - 0.43 - 0.63/2) \times (1 + 0.06 \times 1.25) = 59 \times 10^{-6} \text{ A}$$

$$g = -\frac{1}{59 \times 10^{-6}} \frac{1.5 \times 115 \times 10^{-6} \times 0.63 + 1.5 \times 3.4 \times 30 \times 10^{-6} \times 1.0}{0.06 + 0.1} = -27.5 \text{ (Eq. 5.10)}$$

This yields the following values for V_{IL}, V_{IH}, NM_L, NM_H:

$$V_{IL} = 1.2 \text{ V}, V_{IH} = 1.3 \text{ V}, NM_L = NM_H = 1.2.$$

Figure 5.10 plots the simulated VTC of the inverter, as well as its derivative, the gain. A close to ideal characteristic is obtained. The actual values of V_{IL} and V_{IH} are 1.03 V and 1.45 V, respectively, which leads to noise margins of 1.03 V and 1.05 V. These values are lower than those predicted for two reasons:

• Eq. (5.10) overestimates the gain. As observed in Figure 5.10b, the maximum gain (at V_M) equals only 17. This reduced gain would yield values for V_{IL} and V_{IH} of 1.17 V, and 1.33 V, respectively.

• The most important deviation is due to the piecewise linear approximation of the VTC, which is optimistic with respect to the actual noise margins.

The obtained expressions are however perfectly useful as first-order estimations as well as means of identifying the relevant parameters and their impact.

To conclude this example, we also extracted from simulations the output resistance of the inverter in the low- and high-output states. Low values of 2.4 kΩ and 3.3 kΩ were observed, respectively. The output resistance is a good measure of the sensitivity of the gate in respect to noise induced at the output, and is preferably as low as possible.

SIDELINE: Surprisingly (or not so surprisingly), the static CMOS inverter can also be used as an analog amplifier, as it has a fairly high gain in its transition region. This region is very narrow however, as is apparent in the graph of Figure 5.10b. It also receives poor marks on other amplifier properties such as supply noise rejection. Yet, this observation can be used to demonstrate one of the major differences between analog and digital design. Where the analog designer would bias the amplifier in the middle of the transient region, so that a maximum linearity is obtained, the digital designer will operate the

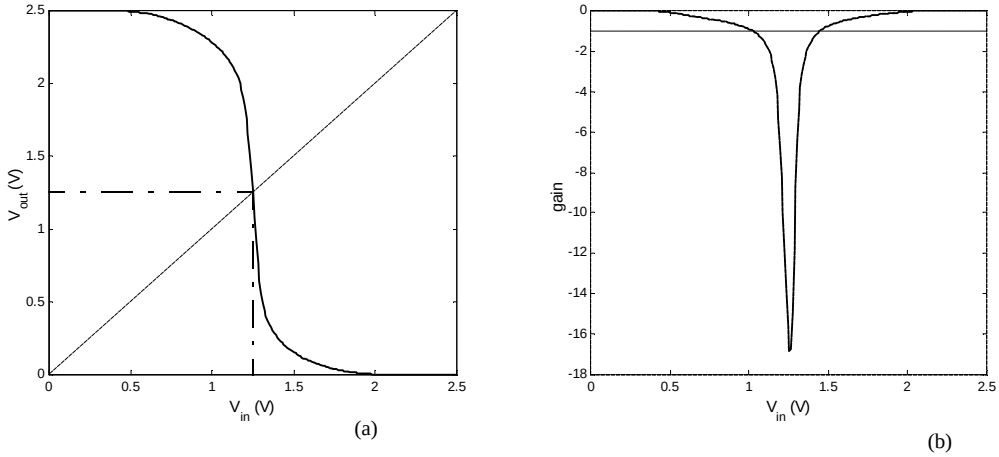


Figure 5.10 Simulated Voltage Transfer Characteristic (a) and voltage gain (b) of CMOS inverter (0.25 μm CMOS, V_{DD} = 2.5 V).

device in the regions of extreme nonlinearity, resulting in well-defined and well-separated high and low signals.

Problem 5.2 Inverter noise margins for long-channel devices

Derive expressions for the gain and noise margins assuming that PMOS and NMOS are long-channel devices (or that the supply voltage is low), so that velocity saturation does not occur.

5.3.3 Robustness Revisited

Device Variations

While we design a gate for nominal operation conditions and typical device parameters, we should always be aware that the actual operating temperature might very over a large range, and that the device parameters after fabrication probably will deviate from the nominal values we used in our design optimization process. Fortunately, the dc-characteristics of the static CMOS inverter turn out to be rather insensitive to these variations, and the gate remains functional over a wide range of operating conditions. This already became apparent in Figure 5.7, which shows that variations in the device sizes have only a minor impact on the switching threshold of the inverter. To further confirm the assumed robustness of the gate, we have re-simulated the voltage transfer characteristic by replacing the nominal devices by their worst- or best-case incarnations. Two corner-cases are plotted in Figure 5.11: a better-than-expected NMOS combined with an inferior PMOS, and the opposite scenario. Comparing the resulting curves with the nominal response shows that the variations mostly cause a shift in the switching threshold, but that the operation of the

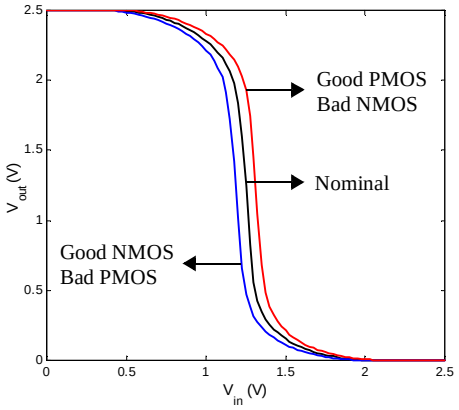


Figure 5.11 Impact of device variations on static CMOS inverter VTC. The "good" device has a smaller oxide thickness (- 3nm), a smaller length (-25 nm), a higher width (+30 nm), and a smaller threshold (-60 mV). The opposite is true for the "bad" transistor.

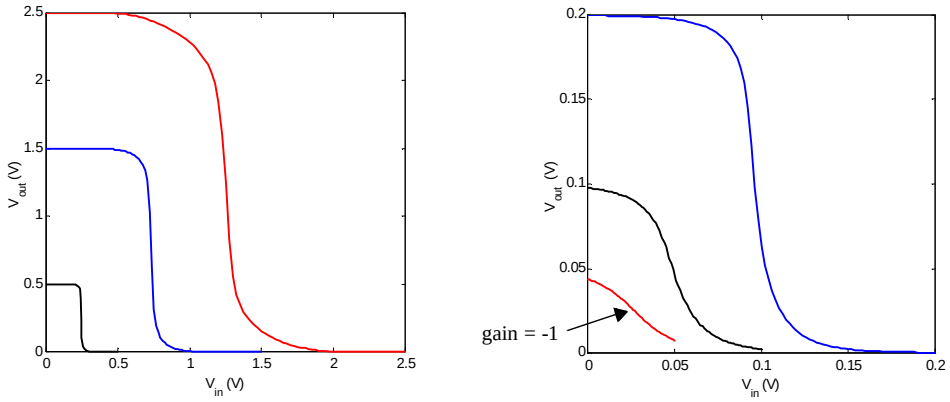
gate is by no means affected. This robust behavior that ensures functionality of the gate over a wide range of conditions has contributed in a big way to the popularity of the static CMOS gate.

Scaling the Supply Voltage

In Chapter 3, we observed that continuing technology scaling forces the supply voltages to reduce at rates similar to the device dimensions. At the same time, device threshold voltages are virtually kept constant. The reader probably wonders about the impact of this trend on the integrity parameters of the CMOS inverter. Do inverters keep on working when the voltages are scaled and are there potential limits to the supply scaling?

A first hint on what might happen was offered in Eq. (5.10), which indicates that the gain of the inverter in the transition region actually increases with a reduction of the supply voltage! Note that for a fixed transistor ratio r, V_M is approximately proportional to V_{DD}. Plotting the (normalized) VTC for different supply voltages not only confirms this conjecture, but even shows that the inverter is well and alive for supply voltages close to the threshold voltage of the composing transistors (Figure 5.12a). At a voltage of 0.5 V — which is just 100 mV above the threshold of the transistors — the width of the transition region measures only 10% of the supply voltage (for a maximum gain of 35), while it widens to 17% for 2.5 V. So, given this improvement in dc characteristics, why do we not choose to operate all our digital circuits at these low supply voltages? Three important arguments come to mind:

- In the following sections, we will learn that reducing the supply voltage indiscriminately has a positive impact on the energy dissipation, but is absolutely detrimental to the performance on the gate.
- The dc-characteristic becomes increasingly sensitive to variations in the device parameters such as the transistor threshold, once supply voltages and intrinsic voltages become comparable.
- Scaling the supply voltage means reducing the signal swing. While this typically helps to reduce the internal noise in the system (such as caused by crosstalk), it makes the design more sensitive to external noise sources that do not scale.



(a) Reducing V_{DD} improves the gain...

(b) but it detoriates for very-low supply voltages.

Figure 5.12 VTC of CMOS inverter as a function of supply voltage (0.25 µm CMOS technology).

To provide an insight into the question on potential limits to the voltage scaling, we have plotted in Figure 5.12b the voltage transfer characteristic of the same inverter for the even-lower supply voltages of 200 mV, 100 mV, and 50 mV (while keeping the transistor thresholds at the same level). Amazingly enough, we still obtain an inverter characteristic, this while the supply voltage is not even large enough to turn the transistors on! The explanation can be found in the sub-threshold operation of the transistors. The sub-threshold currents are sufficient to switch the gate between low and high levels, and provide enough gain to produce acceptable VTCs. The very low value of the switching currents ensures a very slow operation but this might be acceptable for some applications (such as watches, for example).

At around 100 mV, we start observing a major deterioration of the gate characteristic. V_{OL} and V_{OH} are no longer at the supply rails and the transition-region gain approaches 1. The latter turns out to be a fundamental show-stopper. To achieving sufficient gain for use in a digital circuit, it is necessary that the supply must be at least a couple times $\phi_T = kT/q$ (=25 mV at room temperature), the thermal voltage introduced in Chapter 3 [Swanson72]. It turns out that below this same voltage, thermal noise becomes an issue as well, potentially resulting in unreliable operation.

$$V_{DDmin} > 2\ldots4\frac{kT}{q} \tag{5.11}$$

Eq. (5.11) presents a true lower bound on supply scaling. It suggests that the only way to get CMOS inverters to operate below 100 mV is to reduce the ambient temperature, or in other words to cool the circuit.

Problem 5.3 Minimum supply voltage of CMOS inverter

Once the supply voltage drops below the threshold voltage, the transistors operate the sub-threshold region, and display an exponential current-voltage relationship (as expressed in Eq. (3.40)). Derive an expression for the gain of the inverter under these circumstances

(assume symmetrical NMOS and PMOS transistors, and a maximum gain at $V_M = V_{DD}/2$). The resulting expression demonstrates that the minimum voltage is a function of the slope factor n of the transistor.

$$g = -\left(\frac{1}{n}\right)\left(e^{V_{DD}/2\phi_T} - 1\right) \tag{5.12}$$

According to this expression, the gain drops to -1 at V_{DD} = 48 mV (for n = 1.5 and ϕ_T = 25 mV).

5.4 Performance of CMOS Inverter: The Dynamic Behavior

The qualitative analysis presented earlier concluded that the propagation delay of the CMOS inverter is determined by the time it takes to charge and discharge the load capacitor C_L through the PMOS and NMOS transistors, respectively. This observation suggests that **getting C_L as small as possible is crucial to the realization of high-performance CMOS circuits**. It is hence worthwhile to first study the major components of the load capacitance before embarking onto an in-depth analysis of the propagation delay of the gate. In addition to this detailed analysis, the section also presents a summary of techniques that a designer might use to optimize the performance of the inverter.

5.4.1 Computing the Capacitances

Manual analysis of MOS circuits where each capacitor is considered individually is virtually impossible and is exacerbated by the many nonlinear capacitances in the MOS transistor model. To make the analysis tractable, we assume that all capacitances are lumped together into one single capacitor C_L, located between V_{out} and *GND*. Be aware that this is a considerable simplification of the actual situation, even in the case of a simple inverter.

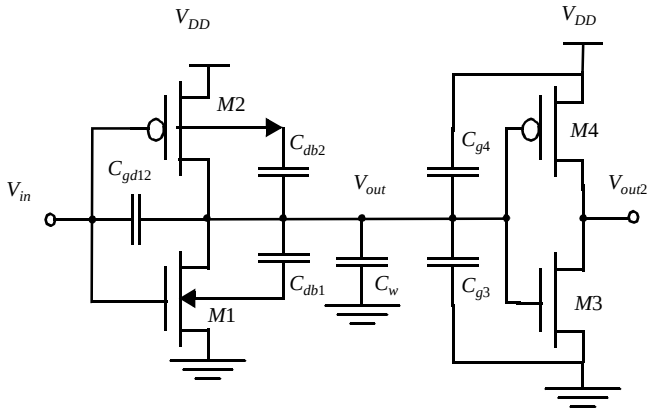


Figure 5.13 Parasitic capacitances, influencing the transient behavior of the cascaded inverter pair.

Figure 5.13 shows the schematic of a cascaded inverter pair. It includes all the capacitances influencing the transient response of node V_{out}. It is initially assumed that the input V_{in} is driven by an *ideal voltage source with zero rise and fall times*. Accounting only for capacitances connected to the output node, C_L breaks down into the following components.

Gate-Drain Capacitance C_{gd12}

M1 and M2 are either in cut-off or in the saturation mode during the first half (up to 50% point) of the output transient. Under these circumstances, the only contributions to C_{gd12} are the overlap capacitances of both M1 and M2. The channel capacitance of the MOS transistors does not play a role here, as it is located either completely between gate and bulk (cut-off) or gate and source (saturation) (see Chapter 3).

The lumped capacitor model now requires that this floating gate-drain capacitor be replaced by a capacitance-to-ground. This is accomplished by taking the so-called Miller effect into account. During a low-high or high-low transition, the terminals of the gate-drain capacitor are moving in opposite directions (Figure 5.14). The voltage change over the floating capacitor is hence twice the actual output voltage swing. To present an identical load to the output node, the capacitance-to-ground must have a value that is twice as large as the floating capacitance.

We use the following equation for the gate-drain capacitors: $C_{gd} = 2\ C_{GD0}W$ (with C_{GD0} the overlap capacitance per unit width as used in the SPICE model). For an in-depth discussion of the Miller effect, please refer to textbooks such as Sedra and Smith ([Sedra87], p. 57).[1]

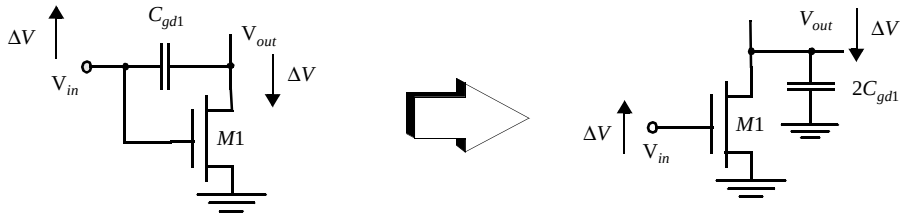


Figure 5.14 The Miller effect—A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.

Diffusion Capacitances C_{db1} and C_{db2}

The capacitance between drain and bulk is due to the reverse-biased *pn*-junction. Such a capacitor is, unfortunately, quite nonlinear and depends heavily on the applied voltage. We argued in Chapter 3 that the best approach towards simplifying the analysis is to replace the nonlinear capacitor by a linear one with the same change in charge for the voltage range of interest. A multiplication factor K_{eq} is introduced to relate the linearized capacitor to the value of the junction capacitance under zero-bias conditions.

[1] The Miller effect discussed in this context is a simplified version of the general analog case. In a digital inverter, the large scale gain between input and output always equals -1.

$$C_{eq} = K_{eq} C_{j0} \tag{5.13}$$

with C_{j0} the junction capacitance per unit area under zero-bias conditions. An expression for K_{eq} was derived in Eq. (3.11) and is repeated here for convenience

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)} [(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}] \tag{5.14}$$

with ϕ_0 the built-in junction potential and m the grading coefficient of the junction. Observe that the junction voltage is defined to be negative for reverse-biased junctions.

Example 5.3 K_{eq} for a 2.5 V CMOS Inverter

Consider the inverter of Figure 5.13 designed in the generic 0.25 μm CMOS technology. The relevant capacitance parameters for this process were summarized in Table 3.5.

Let us first analyze the NMOS transistor (C_{db1} in Figure 5.13). The propagation delay is defined by the time between the 50% transitions of the input and the output. For the CMOS inverter, this is the time-instance where V_{out} reaches 1.25 V, as the output voltage swing goes from rail to rail or equals 2.5 V. We, therefore, linearize the junction capacitance over the interval {2.5 V, 1.25 V} for the high-to-low transition, and {0, 1.25 V} for the low-to-high transition.

During the high-to-low transition at the output, V_{out} initially equals 2.5 V. Because the bulk of the NMOS device is connected to *GND*, this translates into a reverse voltage of 2.5 V over the drain junction or $V_{high} = -2.5$ V. At the 50% point, $V_{out} = 1.25$ V or $V_{low} = -1.25$ V. Evaluating Eq. (5.14) for the bottom plate and sidewall components of the diffusion capacitance yields

Bottom plate: K_{eq} ($m = 0.5$, $\phi_0 = 0.9$) = 0.57,
Sidewall: K_{eqsw} ($m = 0.44$, $\phi_0 = 0.9$) = 0.61

During the low-to-high transition, V_{low} and V_{high} equal 0 V and −1.25 V, respectively, resulting in higher values for K_{eq},

Bottom plate: K_{eq} ($m = 0.5$, $\phi_0 = 0.9$) = 0.79,
Sidewall: K_{eqsw} ($m = 0.44$, $\phi_0 = 0.9$) = 0.81

The PMOS transistor displays a reverse behavior, as its substrate is connected to 2.5 V. Hence, for the high-to-low transition ($V_{low} = 0$, $V_{high} = -1.25$ V),

Bottom plate: K_{eq} ($m = 0.48$, $\phi_0 = 0.9$) = 0.79,
Sidewall: K_{eqsw} ($m = 0.32$, $\phi_0 = 0.9$) = 0.86

and for the low-to-high transition ($V_{low} = -1.25$ V, $V_{high} = -2.5$ V)

Bottom plate: K_{eq} ($m = 0.48$, $\phi_0 = 0.9$) = 0.59,
Sidewall: K_{eqsw} ($m = 0.32$, $\phi_0 = 0.9$) = 0.7

Using this approach, the junction capacitance can be replaced by a linear component and treated as any other device capacitance. The result of the linearization is a minor distortion of the voltage waveforms. The logic delays are not significantly influenced by this simplification.

Wiring Capacitance C_w

The capacitance due to the wiring depends upon the length and width of the connecting wires, and is a function of the distance of the fanout from the driving gate and the number of fanout gates. As argued in Chapter 4, this component is growing in importance with the scaling of the technology.

Gate Capacitance of Fanout C_{g3} and C_{g4}

We assume that the fanout capacitance equals the total gate capacitance of the loading gates M3 and M4. Hence,

$$\begin{aligned} C_{fanout} &= C_{gate}(\text{NMOS}) + C_{gate}(\text{PMOS}) \\ &= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox}) \end{aligned} \tag{5.15}$$

This expression simplifies the actual situation in two ways:

- It assumes that all components of the gate capacitance are connected between V_{out} and GND (or V_{DD}), and ignores the Miller effect on the gate-drain capacitances. This has a relatively minor effect on the accuracy, since we can safely assume that the connecting gate does not switch before the 50% point is reached, and V_{out2}, therefore, remains constant in the interval of interest.
- A second approximation is that the channel capacitance of the connecting gate is constant over the interval of interest. This is not exactly the case as we discovered in Chapter 3. The total channel capacitance is a function of the operation mode of the device, and varies from approximately 1/3 of WLC_{ox} (cut-off) over 2/3 WLC_{ox} (saturation) to the full WLC_{ox} (linear). During the first half of the transient, it may be assumed that one of the load devices is always in linear mode, while the other transistor evolves from the off-mode to saturation. Ignoring the capacitance variation results in a pessimistic estimation with an error of approximately 10%, which is acceptable for a first order analysis.

Example 5.4 Capacitances of a 0.25 μm CMOS Inverter

A minimum-size, symmetrical CMOS inverter has been designed in the 0.25 μm CMOS technology. The layout is shown in Figure 5.15. The supply voltage V_{DD} is set to 2.5 V. From the layout, we derive the transistor sizes, diffusion areas, and perimeters. This data is summarized in Table 5.1. As an example, we will derive the drain area and perimeter for the NMOS transistor. The drain area is formed by the metal-diffusion contact, which has an area of $4 \times 4\ \lambda^2$, and the rectangle between contact and gate, which has an area of $3 \times 1\ \lambda^2$. This results in a total area of 19 λ^2, or 0.30 μm^2 (as $\lambda = 0.125$ μm). The perimeter of the drain area is rather involved and consists of the following components (going counterclockwise): 5 + 4 + 4 + 1 + 1 = 15 λ or PD = $15 \times 0.125 = 1.875$ μm. Notice that the gate side of the drain perimeter is not included, as this is not considered a part of the side-wall. The drain area and perimeter of the PMOS transistor are derived similarly (the rectangular shape makes the exercise considerably simpler): AD = $5 \times 9\ \lambda^2 = 45\ \lambda^2$, or 0.7 μm^2; PD = 5 + 9 + 5 = 19 λ, or 2.375 μm.

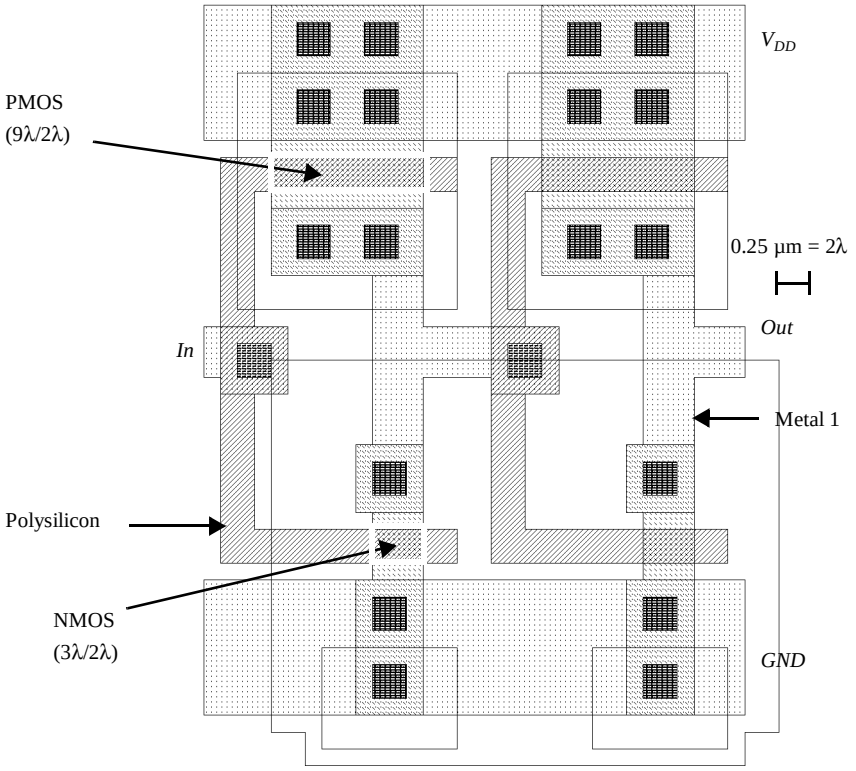


Figure 5.15 Layout of two chained, minimum-size inverters using SCMOS Design Rules (see also Color-plate 6).

Table 5.1 Inverter transistor data.

	W/L	**AD (μm^2)**	**PD (μm)**	**AS (μm^2)**	**PS (μm)**
NMOS	0.375/0.25	0.3 (19 λ^2)	1.875 (15λ)	0.3 (19 λ^2)	1.875 (15λ)
PMOS	1.125/0.25	0.7 (45 λ^2)	2.375 (19λ)	0.7 (45 λ^2)	2.375 (19λ)

This physical information can be combined with the approximations derived above to come up with an estimation of C_L. The capacitor parameters for our generic process were summarized in Table 3.5, and repeated here for convenience:

Overlap capacitance: CGD0(NMOS) = 0.31 fF/μm; CGDO(PMOS) = 0.27 fF/μm

Bottom junction capacitance: CJ(NMOS) = 2 fF/μm^2; CJ(PMOS) = 1.9 fF/μm^2

Side-wall junction capacitance: CJSW(NMOS) = 0.28 fF/μm; CJSW(PMOS) = 0.22 fF/μm

Gate capacitance: C_{ox}(NMOS) = C_{ox}(PMOS) = 6 fF/μm^2

Finally, we should also consider the capacitance contributed by the wire, connecting the gates and implemented in metal 1 and polysilicon. A layout extraction program typically

will deliver us precise values for this parasitic capacitance. Inspection of the layout helps us to form a first-order estimate and yields that the metal-1 and polsyilicon areas of the wire, that are not over active diffusion, equal 42 λ^2 and 72 λ^2, respectively. With the aid of the interconnect parameters of Table 4.2, we find the wire capacitance — observe that we ignore the fringing capacitance in this simple exercise. Due to the short length of the wire, this contribution is ignorable compared to the other parasitics.

$$C_{wire} = 42/8^2\ \mu m^2 \times 30\ aF/\mu m^2 + 72/8^2\ \mu m^2 \times 88\ aF/\mu m^2 = 0.12\ fF$$

Bringing all the components together results in Table 5.2. We use the values of K_{eq} derived in Example 5.3 for the computation of the diffusion capacitances. Notice that the load capacitance is almost evenly split between its two major components: the intrinsic capacitance, composed of diffusion and overlap capacitances, and the extrinsic load capacitance, contributed by wire and connecting gate.

Table 5.2 Components of C_L (for high-to-low and low-to-high transitions).

Capacitor	Expression	Value (fF) (H→L)	Value (fF) (L→H)
C_{gd1}	$2\ CGD0_n\ W_n$	0.23	0.23
C_{gd2}	$2\ CGD0_p\ W_p$	0.61	0.61
C_{db1}	$K_{eqn}\ AD_n\ CJ + K_{eqswn}\ PD_n\ CJSW$	0.66	0.90
C_{db2}	$K_{eqp}\ AD_p\ CJ + K_{eqswp}\ PD_p\ CJSW)$	1.5	1.15
C_{g3}	$(CGD0_n+CGSO_n)\ W_n + C_{ox}\ W_n\ L_n$	0.76	0.76
C_{g4}	$(CGD0_p+CGSO_p)\ W_p + C_{ox}\ W_p\ L_p$	2.28	2.28
C_w	From Extraction	0.12	0.12
C_L	Σ	6.1	6.0

5.4.2 Propagation Delay: First-Order Analysis

One way to compute the propagation delay of the inverter is to integrate the capacitor (dis)charge current. This results in the expression of Eq. (5.16).

$$t_p = \int_{v_1}^{v_2} \frac{C_L(v)}{i(v)} dv \tag{5.16}$$

with i the (dis)charging current, v the voltage over the capacitor, and v_1 and v_2 the initial and final voltage. An exact computation of this equation is untractable, as both $C_L(v)$ and $i(v)$ are nonlinear functions of v. We rather fall back to the simplified switch-model of the inverter introduced in Figure 5.6 to derive a reasonable approximation of the propagation delay adequate for manual analysis. The voltage-dependencies of the on-resistance and the load capacitor are addressed by replacing both by a constant linear element with a value averaged over the interval of interest. The preceding section derived precisely this value

for the load capacitance. An expression for the average on-resistance of the MOS transistor was already derived in Example 3.8, and is repeated here for convenience.

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$
$$\text{with} \quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right) \tag{5.17}$$

Deriving the propagation delay of the resulting circuit is now straightforward, and is nothing more than the analysis of a first-order linear *RC*-network, identical to the exercise of Example 4.5. There, we learned that the propagation delay of such a network for a voltage step at the input is proportional to the time-constant of the network, formed by pull-down resistor and load capacitance. Hence,

$$t_{pHL} = \ln(2)R_{eqn}C_L = 0.69R_{eqn}C_L \tag{5.18}$$

Similarly, we can obtain the propagation delay for the low-to-high transition,

$$t_{pLH} = 0.69R_{eqp}C_L \tag{5.19}$$

with R_{eqp} the equivalent on-resistance of the PMOS transistor over the interval of interest. This analysis assumes that the equivalent load-capacitance is identical for both the high-to-low and low-to-high transitions. This has been shown to be approximately the case in the example of the previous section. The overall propagation delay of the inverter is defined as the average of the two values, or

$$t_p = \frac{t_{pHL}+t_{pLH}}{2} = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right) \tag{5.20}$$

Very often, it is desirable for a gate to have identical propagation delays for both rising and falling inputs. This condition can be achieved by making the on-resistance of the NMOS and PMOS approximately equal. Remember that this condition is identical to the requirement for a symmetrical VTC.

Example 5.5 Propagation Delay of a 0.25 μm CMOS Inverter

To derive the propagation delays of the CMOS inverter of Figure 5.15, we make use of Eq. (5.18) and Eq. (5.19). The load capacitance C_L was already computed in Example 5.4, while the equivalent on-resistances of the transistors for the generic 0.25 μm CMOS process were derived in Table 3.3. For a supply voltage of 2.5 V, the normalized on-resistances of NMOS and PMOS transistors equal 13 kΩ and 31 kΩ, respectively. From the layout, we determine the (W/L) ratioes of the transistors to be 1.5 for the NMOS , and 4.5 for the PMOS. We assume that the difference between drawn and effective dimensions is small enough to be ignorable. This leads to the following values for the delays:

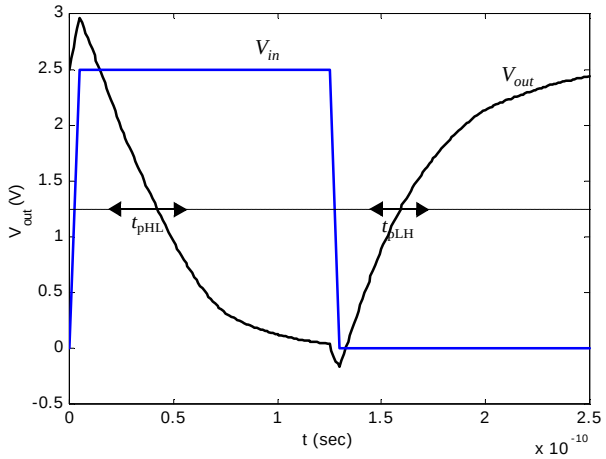


Figure 5.16 Simulated transient response of the inverter of Figure 5.15.

$$t_{pHL} = 0.69 \times \left(\frac{13\text{k}\Omega}{1.5}\right) \times 6.1\text{fF} = 36 \text{ psec}$$

$$t_{pLH} = 0.69 \times \left(\frac{31\text{k}\Omega}{4.5}\right) \times 6.0\text{fF} = 29 \text{ psec}$$

and

$$t_p = \left(\frac{36 + 29}{2}\right) = 32.5 \text{ psec}$$

The accuracy of this analysis is checked by performing a SPICE transient simulation on the circuit schematic, extracted from the layout of Figure 5.15. The computed transient response of the circuit is plotted in Figure 5.16, and determines the propagation delays to be 39.9 psec and 31.7 for the HL and LH transitions, respectively. The manual results are good considering the many simplifications made during their derivation. Notice especially the overshoots on the simulated output signals . These are caused by the gate-drain capacitances of the inverter transistors, which couple the steep voltage step at the input node directly to the output before the transistors can even start to react to the changes at the input. These overshoots clearly have a negative impact on the performance of the gate, and explain why the simulated delays are larger than the estimations.

WARNING: This example might give the impression that manual analysis always leads to close approximations of the actual response. This is not necessarily the case. Large deviations can often be observed between first- and higher-order models. The purpose of the manual analysis is to get a basic insight in the behavior of the circuit and to determine the dominant parameters. A detailed simulation is indispensable when quantitative data is required. Consider the example above a stroke of good luck.

The obvious question a designer asks herself at this point is how she can manipulate and/or optimize the delay of a gate. To provide an answer to this question, it is necessary to make the parameters governing the delay explicit by expanding R_{eq} in the delay equation. Combining Eq. (5.18) and Eq. (5.17), and assuming for the time being that the channel-length modulation factor λ is ignorable, yields the following expression for t_{pHL} (a similar analysis holds for t_{pLH})

$$t_{pHL} = 0.69\frac{3}{4}\frac{C_L V_{DD}}{I_{DSATn}} = 0.52\frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)} \tag{5.21}$$

In the majority of designs, the supply voltage is chosen high enough so that $V_{DD} >> V_{Tn} + V_{DSATn}/2$. Under these conditions, the delay becomes virtually independent of the supply voltage (Eq. (5.22)). Observe that this is a first-order approximation, and that increasing the supply voltage yields an observable, albeit small, improvement in performance due to a non-zero channel-length modulation factor.

$$t_{pHL} \approx 0.52\frac{C_L}{(W/L)_n k'_n V_{DSATn}} \tag{5.22}$$

This analysis is confirmed in Figure 5.17, which plots the propagation delay of the inverter as a function of the supply voltage. It comes as no surprise that this curve is virtually identical in shape to the one of Figure 3.27, which charts the equivalent on-resistance of the MOS transistor as a function of V_{DD}. While the delay is relative insensitive to supply variations for higher values of V_{DD}, a sharp increase can be observed starting around

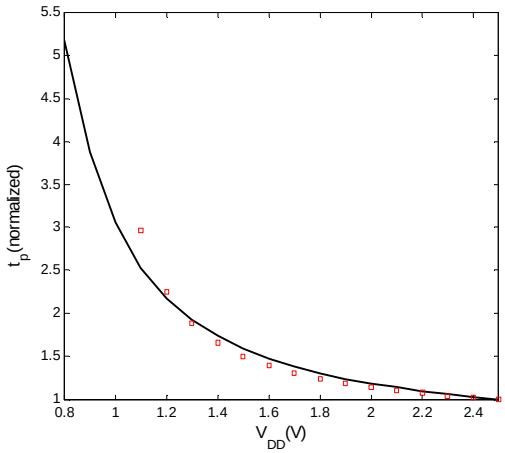


Figure 5.17 Propagation delay of CMOS inverter as a function of supply voltage (normalized with respect to the delay at 2.5 V). The dots indicate the delay values predicted by Eq. (5.21). Observe that this equation is only valid when the devices are velocity-saturated. Hence, the deviation at low supply voltages.

$\approx 2V_T$. This operation region should clearly be avoided if achieving high performance is a premier design goal.

Design Techniques

From the above, we deduce that the propagation delay of a gate can be minimized in the following ways:

- *Reduce* C_L. Remember that three major factors contribute to the load capacitance: the internal diffusion capacitance of the gate itself, the interconnect capacitance, and the fan-out. Careful layout helps to reduce the diffusion and interconnect capacitances. **Good design practice requires keeping the drain diffusion areas as small as possible.**
- *Increase the W/L ratio of the transistors*. This is the most powerful and effective performance optimization tool in the hands of the designer. Proceed however with caution when applying this approach. Increasing the transistor size also raises the diffusion capacitance and hence C_L. In fact, once the intrinsic capacitance (i.e. the diffusion capacitance) starts to dominate the extrinsic load formed by wiring and fanout, increasing the gate size does not longer help in reducing the delay, and only makes the gate larger in area. This effect is called "*self-loading*". In addition, wide transistors have a larger gate capacitance, which increases the fan-out factor of the driving gate and adversely affects its speed.
- *Increase* V_{DD}. As illustrated in Figure 5.17, the delay of a gate can be modulated by modifying the supply voltage. This flexibility allows the designer to trade-off energy dissipation for performance, as we will see in a later section. However, increasing the supply voltage above a certain level yields only very minimal improvement and hence should be avoided. Also, reliability concerns (oxide breakdown, hot-electron effects) enforce firm upper-bounds on the supply voltage in deep sub-micron processes.

Example 5.6 Device sizing for performance

Let us explore the performance improvement that can be obtained by device sizing in the design of Example 5.5. We assume the wire and fanout capacitance to be unaffected by the resizing. An insight in the potential improvement can be obtained by partitioning the load capacitance into an intrinsic (diffusion and miller) and an extrinsic (wiring and fanout) component, or

$$C_L = C_{int} + C_{ext} = C_{int}(1 + \alpha) \tag{5.23}$$

with α the ratio between extrinsic and intrinsic capacitance. Widening both NMOS and PMOS of the driving inverter with a factor S reduces their equivalent resistance by an identical factor, but also raises the intrinsic capacitance of the gate by approximately the same ratio. The propagation delay of the redesigned gate can be estimated

$$t_p = 0.69(S + \alpha)C_{int}\left(\frac{R_{eqn} + R_{eqp}}{2S}\right) = \left(1 + \frac{\alpha}{S}\right)t_{p0} \tag{5.24}$$

with t_{p0} the *intrinsic delay of the gate* (this is, no extrinsic load, or $\alpha = 0$). Making S infinitially large yields the maximum obtainable performance gain, equal to $1/(1+\alpha)$. Yet, any sizing factor S that is sufficiently larger than α will produce similar results at a substantial gain in silicon area.

For the example in question, we find from Table 5.2 that $\alpha \approx 1.05$ (C_{int} = 3.0 fF, C_{ext} = 3.15 fF). This would predict a maximum performance gain of 2.05. A scaling factor of 10 allows us to get within 10% of this optimal performance, while larger device sizes only yield ignorable performance gains.

This is confirmed by simulation results, which predict a maximum obtainable perfor-

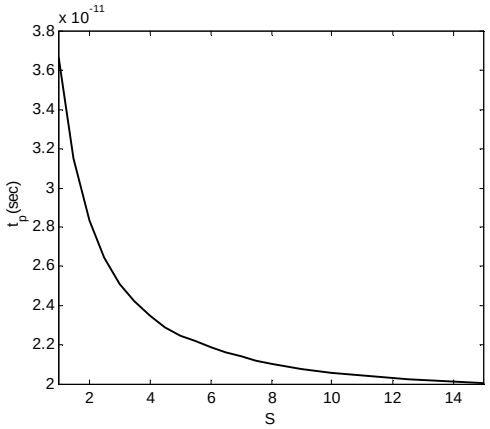


Figure 5.18 Increasing inverter performance by sizing the NMOS and PMOS transistor with an identical factor S for a fixed fanout (inverter of Figure 5.15).

mance improvement of 1.9 (t_{p0} = 19.3 psec). From the graph of Figure 5.18, we observe that the bulk of the improvement is already obtained for S = 5, and that sizing factors larger than 10 barely yield any extra gain.

Problem 5.4 Propagation Delay as a Function of (dis)charge Current

So far, we have expressed the propagation delay as a function of the equivalent resistance of the transistors. Another approach would be replace the transistor by a current source with value equal to the average (dis)charge current over the interval of interest. Derive an expression of the propagation delay using this alternative approach.

5.4.3 Propagation Delay Revisited

A detailed analysis of the transient response of the complementary MOS inverter yields some extra observations and design trade-off's, worth analyzing.

Impact of Fanout

Eq. (5.23) states that the load capacitance of the inverter can be divided into an intrinsic and an extrinsic component. The latter factor is an obvious function of the fanout of the gate: the larger the fanout, the larger the external load. Assuming that each fanout gate presents an identical load, and that the wiring capacitance is proportional to the fanout,[2] we can rewrite the delay equation as a function of the fanout N.

$$t_p(N) = t_{p0}(1 + \alpha N) \tag{5.25}$$

[2] The linear relationship between fanout and wiring capacitance has been confirmed by a number of heuristic studies [REF].

A linear dependence can be observed. Large fanout factors should hence be avoided if performance is an issue. From the preceding discussions, it is furthermore apparent that increasing the sizing factor S of the driving inverter is appropriate and recommendable in the presence of fanout.

NMOS/PMOS Ratio

So far, we have consistently widened the PMOS transistor so that its resistance matches that of the pull-down NMOS device. This typically requires a ratio of 3 to 3.5 between PMOS and NMOS width. The motivation behind this approach is to create an inverter with a symmetrical VTC, and to equate the high-to-low and low-to-high propagation delays. However, this does not imply that this ratio also yields the minimum overall propagation delay. If symmetry and reduced noise margins are not of prime concern, it is actually possible to speed up the inverter by reducing the width of the PMOS device!

The reasoning behind this statement is that, while widening the PMOS improves the t_{pLH} of the inverter by increasing the charging current, it also degrades the t_{pHL} by cause of a larger parasitic capacitance. When two contradictory effects are present, there must exist a transistor ratio that optimizes the propagation delay of the inverter.

This optimum ratio can be derived through the following simple analysis. Consider two identical, cascaded CMOS inverters. The load capacitance of the first gate equals approximately

$$C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_W \tag{5.26}$$

where C_{dp1} and C_{dn1} are the equivalent drain diffusion capacitances of PMOS and NMOS transistors of the first inverter, while C_{gp2} and C_{gn2} are the gate capacitances of the second gate. C_W represents the wiring capacitance.

When the PMOS devices are made β times larger than the NMOS ones ($\beta = (W/L)_p / (W/L)_n$), all transistor capacitances will scale in approximately the same way, or $C_{dp1} \approx \beta\, C_{dn1}$, and $C_{gp2} \approx \beta\, C_{gn2}$. Eq. (5.26) can then be rewritten:

$$C_L = (1 + \beta)(C_{dn1} + C_{gn2}) + C_W \tag{5.27}$$

An expression for the propagation delay can be derived, based on Eq. (5.20).

$$\begin{aligned} t_p &= \frac{0.69}{2}((1 + \beta)(C_{dn1} + C_{gn2}) + C_W)\left(R_{eqn} + \frac{R_{eqp}}{\beta}\right) \\ &= 0.345((1 + \beta)(C_{dn1} + C_{gn2}) + C_W)R_{eqn}\left(1 + \frac{r}{\beta}\right) \end{aligned} \tag{5.28}$$

r ($= R_{eqp}/R_{eqn}$) represents the resistance ratio of identically-sized PMOS and NMOS transistors. The optimal value of β can be found by setting $\frac{\partial t_p}{\partial \beta}$ to 0, which yields

$$\beta_{opt} = \sqrt{r\left(1 + \frac{C_w}{C_{dn1} + C_{gn2}}\right)} \tag{5.29}$$

This means that when the wiring capacitance is negligible ($C_{dn1}+C_{gn2} >> C_W$), β_{opt} equals $\sqrt{r}$, in contrast to the factor r normally used in the noncascaded case. If the wiring capacitance dominates, larger values of β should be used. The surprising result of this analysis is that smaller device sizes (and hence smaller design area) yield a faster design at the expense of symmetry and noise margin.

Example 5.7 Sizing of CMOS Inverter Loaded by an Identical Gate

Consider again our standard design example. From the values of the equivalent resitances (Table 3.3), we find that a ratio β of 2.4 (= 31 kΩ / 13 kΩ) would yield a symmetrical transient response. Eq. (5.29) now predicts that the device ratio for an optimal performance should equal 1.6. These results are verified in Figure 5.19, which plots the simulated propagation delay as a function of the transistor ratio β. The graph clearly illustrates how a changing β trades off between t_{pLH} and t_{pHL}. The optimum point occurs around β = 1.9, which is somewhat higher than predicted. Observe also that the rising and falling delays are identical at the predicted point of β equal to 2.4.

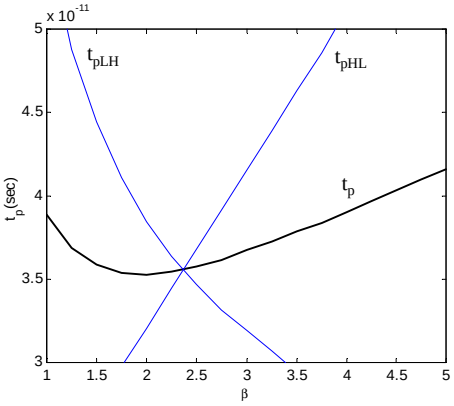


Figure 5.19 Propagation delay of CMOS inverter as a function of the PMOS/NMOS transistor ratio β.

.

The rise/fall time of the input signal

All the above expressions were derived under the assumption that the input signal to the inverter abruptly changed from 0 to V_{DD} or vice-versa. Only one of the devices is assumed to be on during the (dis)charging process. In reality, the input signal changes gradually and, temporarily, PMOS and NMOS transistors conduct simultaneously. This affects the total current available for (dis)charging and impacts the propagation delay. Figure 5.20 plots the propagation delay of a minimum-size inverter as a function of the input signal slope—as obtained from SPICE. It can be observed that t_p increases (approximately) linearly with increasing input slope, once $t_s > t_p(t_s=0)$.

While it is possible to derive an analytical expression describing the relationship between input signal slope and propagation delay, the result tends to be complex and of limited value. From a design perspective, it is more valuable to relate the impact of the finite slope on the performance directly to its cause, which is the limited driving capability of the preceding gate. If the latter would be infinitely strong, its output slope would be zero, and the performance of the gate under examination would be unaffected. The major

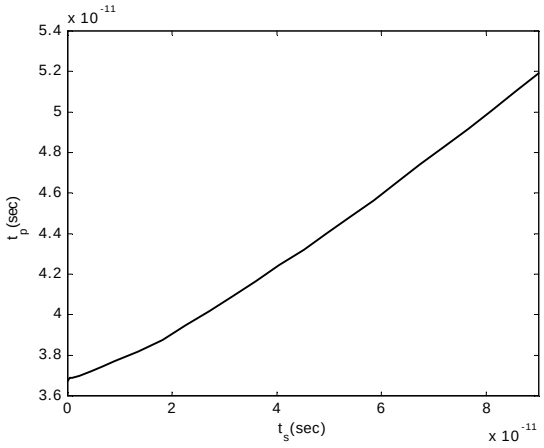


Figure 5.20 t_p as a function of the input signal slope (10-90% rise or fall time) for minimum-size inverter with fan-out of a single gate.

advantage of this approach is that it realizes that a gate is never designed in isolation, and that its performance is both affected by the fanout, and the driving strength of the gate(s) feeding into its inputs. This leads to a revised expression for the propagation delay of an inverter *i* in a chain of inverters [Hedenstierna87]:

$$t_p^i = t_{step}^i + \eta t_{step}^{i-1} \tag{5.30}$$

Eq. (5.30) states that the propagation delay of inverter *i* equals the sum of the delay of the same gate for a step input (t_{step}^i) (i.e. zero input slope) augmented with a fraction of the step-input delay of the preceding gate (*i*-1). The fraction η is an empirical constant. This expression has the advantage of being very simple, yet it exposes all relationships necessary for global delay computations of complex circuits.

Example 5.8 Delay of Inverter embedded in Network

Consider for instance the circuit of . All inverters in this example are assumed to be identical, and to have an intrinsic propagation delay t_{p0}. With the aid of Eq. (5.30) and Eq. (5.25), we can derive an expression for the delay of inverter *i*.

$$\begin{aligned} t_p^i &= t_{p0}(1 + \alpha N) + \eta t_{p0}(1 + \alpha M) \\ &= t_{p0}(1 + \eta + \alpha(N + \eta M)) \end{aligned} \tag{5.31}$$

with *N* and *M* the fanout factors of inverters *i* and *i*-1, respectively. Typical values for the parameters α and η are around 1 and 0.25, respectively. Experiments have demonstrated that the model of Eq. (5.31) forms a good approximation of the actual dependencies, although some important deviations can be observed for small values of *N* and *M*.

Design Challenge

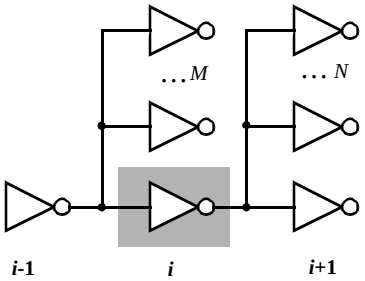


Figure 5.21 Inverter (in shaded box) embedded in network of identical inverters. *M* and *N* denote the fanout factors of inverter *i*-1 and *i*, respectively.

It is advantageous to keep the signal rise times smaller than or equal to the gate propagation delays. This proves to be true not only for performance, but also for power consumption considerations as will be discussed later. Keeping the rise and fall times of the signals small and of approximately equal values is one of the major challenges in high-performance design, and is often called '*slope engineering*'.

Problem 5.5 Impact of input slope

Determine if reducing the supply voltage increases or decreases the influence of the input signal slope on the propagation delay. Explain your answer.

Delay in the Presence of (Long) Interconnect Wires

The interconnect wire has played a minimal role in our analysis so far. When gates get farther apart, the wire capacitance and resistance can no longer be ignored, and may even dominate the transient response. Earlier delay expressions can be adjusted to accomodate these extra contributions by employing the wire modeling techniques introduced in the previous Chapter. The analysis detailed in Example 4.9 is directly applicable to the porblem at hand. Consider the circuit of Figure 5.22, where an inverter drives a single fanout through a wire of length L. The driver is represented by a single resistance R_{dr}, wich is the average between R_{eqn} and R_{eqp}. C_{int} and C_{fan} account for the intrinsic capacitance of the driver, and the input capacitance of the fanout gate, respectively.

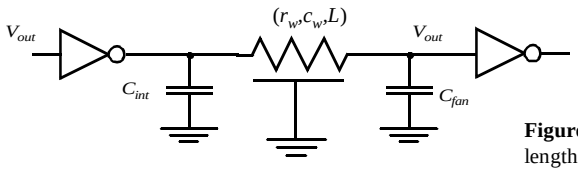


Figure 5.22 Inverter driving single fanout through wire of length *L*.

The propagation delay of the circuit can be obtained by applying the Ellmore delay expression.

$$\begin{aligned} t_p &= 0.69R_{dr}C_{int} + (0.69R_{dr} + 0.38R_w)C_w + 0.69(R_{dr} + R_w)C_{fan} \\ &= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_wC_{fan})L + 0.38r_wc_wL^2 \end{aligned} \quad (5.32)$$

The 0.38 factor accounts for the fact that the wire represents a distributed delay. Cw and Rw stand for the total capacitance and resistance of the wire, respectively. The delay expressions contains a component that is linear with the wire length, as well a quadratic one. It is the latter that causes the wire delay to rapidly become the dominant factor in the delay budget for longer wires.

Example 5.9 Inverter delay in presence of interconnect

Consider the circuit of Figure 5.22, and assume the device parameters of Example 5.5: C_{int} = 3 fF, C_{fan} = 3 fF, and R_{dr} = 0.5(13/1.5 + 31/4.5) = 7.8 kΩ. The wire is implemented in metal1 and has a width of 0.4 μm—the minimum allowed. This yields the following parameters: c_w = 92 aF/μm, and r_w = 0.19 Ω/μm (Example 4.4). With the aid of Eq. (5.32), we can compute at what wire length the delay of the interconnect becomes equal to the intrinsic delay caused purely by device parasitics. Solving the following quadratic equation yields a single useful solution.

$$6.6 \times 10^{-18}L^2 + 0.5 \times 10^{-12}L = 32.29 \times 10^{-12}$$

or

$$L = 65\ \mu\text{m}$$

Observe that the extra delay is solely due to the linear factor in the equation, and more specifically due to the extra capacitance introduced by the wire. The quadratic factor (this is, the distributed wire delay) only becomes dominant at much larger wire lengths (> 7 cm). This is due to the high resistance of the (minimum-size) driver transistors. A different balance emerges when wider transistors are used. Analyze, for instance, the same problem with the driver transistors 100 times wider, as is typical in high-speed, large fan-out drivers.

5.5 Power, Energy, and Energy-Delay

So far, we have seen that the static CMOS inverter with its almost ideal VTC—symmetrical shape, full logic swing, and high noise margins—offers a superior robustness, which simplifies the design process considerably and opens the door for design automation. Another major attractor for static CMOS is the almost complete absence of power consumption in steady-state operation mode. It is this combination of robustness and low static power that has made static CMOS the technology of choice of most contemporary digital designs. The power dissipation of a CMOS circuit is instead dominated by the dynamic dissipation resulting from charging and discharging capacitances.

5.5.1 Dynamic Power Consumption

Dynamic Dissipation due to Charging and Discharging Capacitances

Each time the capacitor C_L gets charged through the PMOS transistor, its voltage rises from 0 to V_{DD}, and a certain amount of energy is drawn from the power supply. Part of this energy is dissipated in the PMOS device, while the remainder is stored on the load capacitor. During the high-to-low transition, this capacitor is discharged, and the stored energy is dissipated in the NMOS transistor.[3]

A precise measure for this energy consumption can be derived. Let us first consider the low-to-high transition. We assume, initially, that the input waveform has zero rise and fall times, or, in other words, that the NMOS and PMOS devices are never on simultaneously. Therefore, the equivalent circuit of Figure 5.23 is valid. The values of the energy E_{VDD}, taken from the supply during the transition, as well as the energy E_C, stored on the capacitor at the end of the transition, can be derived by integrating the instantaneous power over the period of interest. The corresponding waveforms of $v_{out}(t)$ and $i_{VDD}(t)$ are pictured in Figure 5.24.

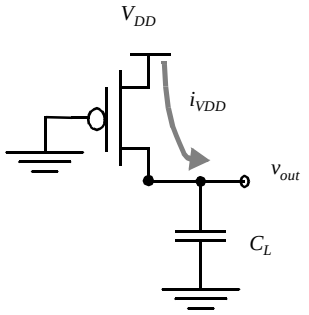


Figure 5.23 Equivalent circuit during the low-to-high transition.

$$E_{VDD} = \int_0^\infty i_{VDD}(t)V_{DD}dt = V_{DD}\int_0^\infty C_L\frac{dv_{out}}{dt}dt = C_L V_{DD}\int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2 \qquad (5.33)$$

$$E_C = \int_0^\infty i_{VDD}(t)v_{out}dt = \int_0^\infty C_L\frac{dv_{out}}{dt}v_{out}dt = C_L\int_0^{V_{DD}} v_{out}dv_{out} = \frac{C_L V_{DD}^2}{2} \qquad (5.34)$$

These results can also be derived by observing that during the low-to-high transition, C_L is loaded with a charge $C_L V_{DD}$. Providing this charge requires an energy from the supply equal to $C_L V_{DD}^2$ $(= Q \times V_{DD})$. The energy stored on the capacitor equals $C_L V_{DD}^2/2$. This means that only half of the energy supplied by the power source is stored on C_L. The other half has been dissipated by the PMOS transistor. Notice that this energy dissipation is independent of the size (and hence the resistance) of the PMOS device! During the discharge phase, the charge is removed from the capacitor, and its energy is dissipated in the NMOS device. Once again, there is no dependence on the size of the device. In summary, each switching cycle (consisting of an L→H and an H→L transition) takes a fixed amount of energy, equal to $C_L V_{DD}^2$. In order to compute the power consumption, we have to take

[3] Observe that this model is a simplification of the actual circuit. In reality, the load capacitance consists of multiple components some of which are located between the output node and GND, others between output node and V_{DD}. The latter experience a charge-discharge cycle that is out of phase with the capacitances to GND, i.e. they get charged when V_{out} goes low and discharged when V_{out} rises. While this distributes the energy delivery by the supply over the two phases, it does not impact the overall dissipation, and the results presented in this section are still valid.

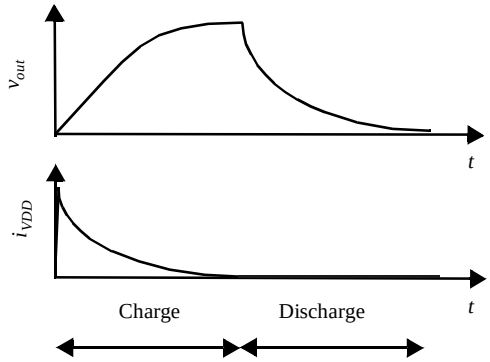


Figure 5.24 Output voltages and supply current during (dis)charge of C_L.

into account how often the device is switched. If the gate is switched **on and off** $f_{0\to1}$ times per second, the power consumption equals

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} \tag{5.35}$$

$f_{0\to1}$ represents the frequency of energy-consuming transitions, this is $0 \to 1$ transitions for static CMOS.

Advances in technology result in ever-higher of values of $f_{0\to1}$ (as t_p decreases). At the same time, the total capacitance on the chip (C_L) increases as more and more gates are placed on a single die. Consider for instance a 0.25 μm CMOS chip with a clock rate of 500 Mhz and an average load capacitance of 15 fF/gate, assuming a fanout of 4. The power consumption per gate for a 2.5 V supply then equals approximately 50 μW. For a design with 1 million gates and assuming that a transition occurs at every clock edge, this would result in a power consumption of 50 W! This evaluation presents, fortunately, a pessimistic perspective. In reality, not all gates in the complete IC switch at the full rate of 500 Mhz. The actual activity in the circuit is substantially lower.

Example 5.10 Capacitive power dissipation of inverter

The capacitive dissipation of the CMOS inverter of Example 5.4 is now easily computed. In Table 5.2, the value of the load capacaitance was determined to equal 6 fF. For a supply voltage of 2.5 V, the amount of energy needed to charge and discharge that capacitance equals

$$E_{dyn} = C_L V_{DD}^2 = 37.5 \text{ fJ}$$

Assume that the inverter is switched at the maximum possible rate ($T = 1/f = t_{pLH} + t_{pHL} = 2\,t_p$). For a t_p of 32.5 psec (Example 5.5), we find that the dynamic power dissipation of the circuit is

$$P_{dyn} = E_{dyn}/(2t_p) = 580\ \mu\text{W}$$

Of course, an inverter in an actual circuit is rarely switched at this maximum rate, and even if done so, the output does not swing from rail-to-rail. The power dissipation will hence be substantially lower. For a rate of 4 GHz (T = 250 psec), the dissipation reduces to 150 μW. This is confirmed by simulations, which yield a power consumption of 155 μW.

Computing the dissipation of a complex circuit is complicated by the $f_{0\to1}$ factor, also called the *switching activity*. While the switching activity is easily computed for an inverter, it turns out to be far more complex in the case of higher-order gates and circuits. One concern is that the switching activity of a network is a function of the nature and the statistics of the input signals: If the input signals remain unchanged, no switching happens, and the dynamic power consumption is zero! On the other hand, rapidly changing signals provoke plenty of switching and hence dissipation. Other factors influencing the activity are the overall network topology and the function to be implemented. We can accomodate this by another rewrite of the equation, or

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} = C_L V_{DD}^2 P_{0\to1} f = C_{EFF} V_{DD}^2 f \tag{5.36}$$

where f now presents the maximum possible event rate of the inputs (which is often the clock rate) and $P_{0\to1}$ the probability that a clock event results in a $0 \to 1$ (or power-consuming) event at the output of the gate. $C_{EFF} = P_{0\to1} C_L$ is called the *effective capacitance* and represents the average capacitance switched every clock cycle. For our example, an activity factor of 10% ($P_{0\to1} = 0.1$) reduces the average consumption to 5 W.

Example 5.11 Switching activity

Consider the waveforms on the right where the upper waveform represents the idealized clock signal , and the botttom one shows the signal at the output of the gate. Power consuming transitions occur 2 out of 8 times, which is equaivalent to a transition probability of 0.25 (or 25%).

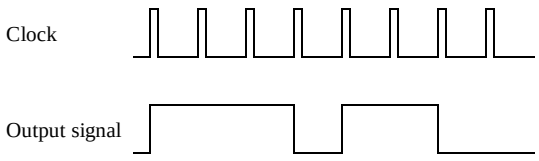


Figure 5.25 Clock and signal waverforms

Low Energy/Power Design Techniques

With the increasing complexity of the digital integrated circuits, it is anticipated that the power problem will only worsen in future technologies. This is one of the reasons that lower supply voltages are becoming more and more attractive. **Reducing V_{DD} has a quadratic effect on P_{dyn}.** For instance, reducing V_{DD} from 2.5 V to 1.25 V for our example drops the power dissipation from 5 W to 1.25 W. This assumes that the same clock rate can be sustained. Figure 5.17 demonstrates that this assumption is not that unrealistic as long as the supply voltage is substantially higher than the threshold voltage. An important performance penalty occurs once V_{DD} approaches 2 V_T.

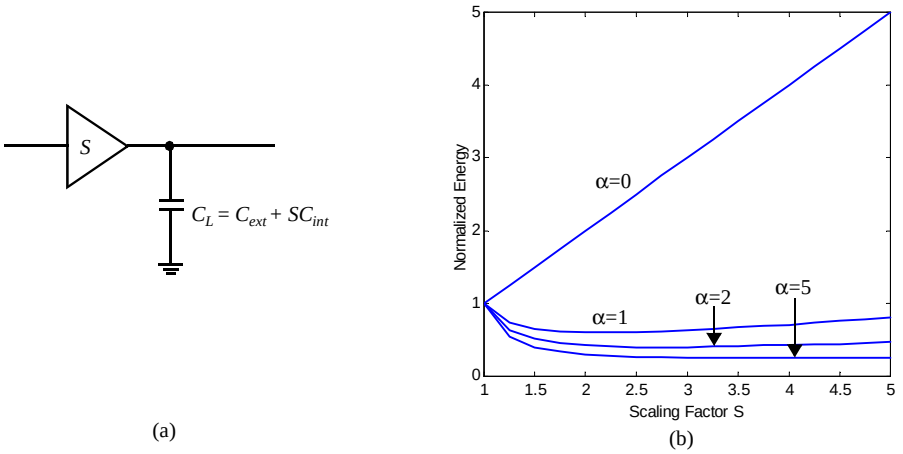
When a lower bound on the supply voltage is set by external constraints (as often happens in real-world designs), or when the performance degradation due to lowering the supply voltage is intolerable, the only means of reducing the dissipation is by lowering the effective capacitance. This can be achieved by addressing both of its components: the physical capacitance and the switching activity.

A *reduction in the switching activity* can only be accomplished at the logic and architectural abstraction levels, and will be discussed in more detail in later Chapters. *Lowering the*

physical capacitance is an overall worthwhile goal, which also helps to improve the performance of the circuit. As most of the capacitance in a combinational logic circuit is due to transistor capacitances (gate and diffusion), it makes sense to keep those contributions to a minimum when designing for low power. This means that transistors should be kept to *minimal size* whenever possible or reasonable. This definitely affects the performance of the circuit, but the effect can be offset by using logic or architectural speed-up techniques. The only instances where transistors should be sized up is when the load capacitance is dominated by extrinsic capacitances (such as fan-out or wiring capacitance). This is contrary to common design practices used in cell libraries, where transistors are generally made large to accommodate a range of loading and performance requirements.

The above observations lead to an interesting design challenge. Assume we have to minimize the energy dissipation of a circuit with a specified lower-bound on the performance. The obvious approach is to lower the supply voltage as much as possible, and to compensate the loss in performance by increasing the transistor sizes. Yet, the latter causes the capacitance to increase. It may be foreseen that at a low enough supply voltage, the latter factor may start to dominate and cause energy to increase with a further drop in the supply voltage.

To analyze the transistor-sizing for minimum energy problem, let us analyze the simple case of a static CMOS inverter driving a load capacitance consisting of an intrinsic (C_{int}) and an extrinsic component (C_{ext}) (Figure 5.26a). While the former represents the diffusion capacitances, the latter stands for wiring capacitance and fan-out. It is assumed that the ratio between PMOS and NMOS transistors is constant. The factor S stands for the inverter sizing factor, where S is equal to 1 for an inverter constructed of minimum-size devices. We can see that the intrinsic capacitance of the scaled device is proportional to S (or $C_{int}(scaled) = SC_{int}$). Figure 5.26b plots the normalized energy (per transition) as a function of the scaling factor S with the ratio between the extrinsic and intrinsic capacitance as a parameter: $\alpha = C_{ext}/C_{int}$. The speed of all implementations is kept constant by appropriately adjusting the supply voltage: larger values of S normally mean lower values of the supply voltage.



(a) (b)

Figure 5.26 Normalized energy of a MOS inverter with load capacitance C_L, as a function of the inverter size S and the ratio between the extrinsic and intrinsic capacitance α (= C_{ext}/C_{int}). (assuming a reference supply voltage of 2.5 V).

When $\alpha = 0$ (or the load capacitance is zero), the lowest energy consumption is obtained when using minimum-size devices. Only when the extrinsic capacitances dominate ($\alpha \geq 1$) does it make sense to widen the devices. This result should come as no surprise: transistor sizing to increase performance—and reduce the energy by lowering the supply voltage—only makes sense as long as performance is dominated by the extrinsic capacitance. Once the intrinsic capacitance becomes the primary factor, further increases in the device sizes only raise the energy consumption while no longer lowering the propagation delay. For example, a sizing factor S of 3.75 minimizes the energy for a load of $\alpha = 5$. The energy-reduction—with a factor of 4 with respect to the circuit instance with minimum-size devices—requires that the supply voltage be reduced to 1.03 V.

Example 5.12 Transistor Sizing for Inverter

We derive a simplified expression for the normalized energy of the inverter of Figure 5.26 as a function of S and α. The energy is normalized with respect to the case for $S = 1$, which is called the *reference*.

An expression for the propagation delay of the gate was already derived in Eq. (5.24), and is repeated here for convenience.

$$t_p = 0.69(S + \alpha)C_{int}\left(\frac{R_{eqn} + R_{eqp}}{2S}\right) = \left(1 + \frac{\alpha}{S}\right)t_{p0}$$

t_{po} stands for the intrinsic delay of the gate at the reference voltage V_{DD}. Its dependence upon V_{DD} is approximated by the following expression, derived from Eq. (5.21).

$$t_{p0} \sim \frac{V_{DD}}{V_{DD} - V_{TE}} = \frac{1}{1 - V_{TE}/V_{DD}} \tag{5.37}$$

with $V_{TE} = V_T + V_{DSAT}/2$ (assume a value averaged over NMOS and PMOS).

Keeping the propagation delay of the scaled inverter constant with respect to the reference case means lowering the supply voltage:

$$\frac{V'_{DD}}{V_{TE}} = \frac{S(1 + \alpha)}{\alpha(S - 1) + (S + \alpha)(V_{TE}/V_{DD})}$$

where V'_{DD} and V_{DD} are the supply voltages of the scaled and reference inverters, respectively. The (normalized) dissipated energy of the scaled inverter is now derived:

$$\frac{E'}{E_{ref}} = \frac{(S + \alpha)C_{int}(V'_{DD})^2}{(1 + \alpha)C_{int}(V_{DD})^2} = \frac{(S + \alpha)S^2(1 + \alpha)}{(\alpha(S - 1)(V_{DD}/V_{TE}) + (S + \alpha))^2}$$

The charts for $V_{DD} = 2.5$ V and $V_{TE} = 0.75$ V are plotted in Figure 5.26. Observe that for $S >> \alpha$, the dissipation increases linearly with S. The reader should further be aware that the presented model is somewhat optimistic, as it ignores the extra energy dissipation related to the increased gate capacitance of the driving transistors.

Dissipation Due to Direct-Path Currents

In actual designs, the assumption of the zero rise and fall times of the input wave forms is not correct. The finite slope of the input signal causes a direct current path between V_{DD} and *GND* for a short period of time during switching, while the NMOS and the PMOS transistors are conducting simultaneously. This is illustrated in Figure 5.27. Under the (reasonable) assumption that the resulting current spikes can be approximated as triangles and that the inverter is symmetrical in its rising and falling responses, we can compute the energy consumed per switching period,

$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak} \tag{5.38}$$

as well as the average power consumption

$$P_{dp} = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f \tag{5.39}$$

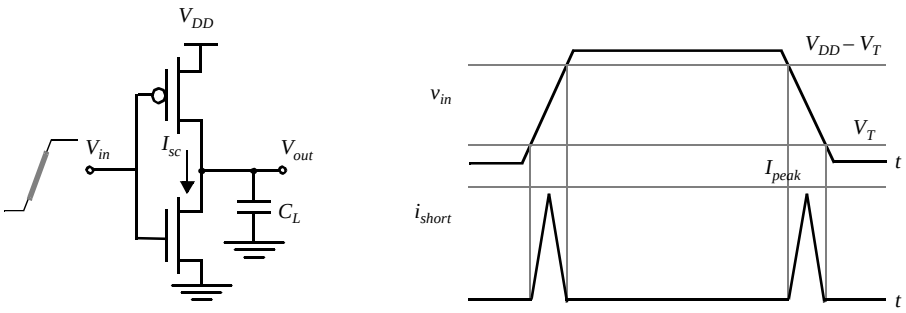
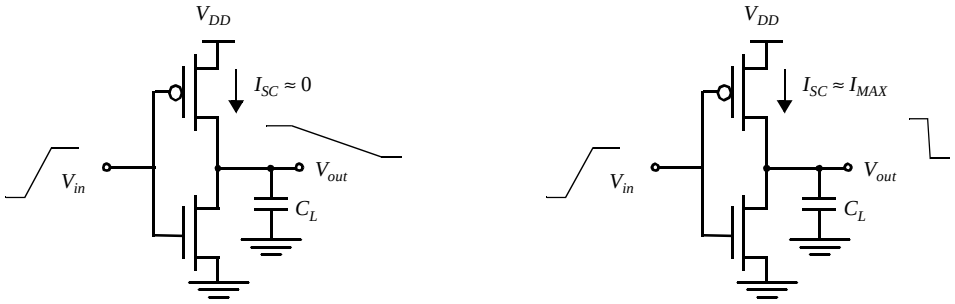


Figure 5.27 Short-circuit currents during transients.

The direct-path power dissipation is proportional to the switching activity, similar to the capacitive power dissipation. t_{sc} represents the time both devices are conducting. For a linear input slope, this time is reasonably well approximated by Eq. (5.40) where t_s represents the 0-100% transition time.

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}}t_s \approx \frac{V_{DD} - 2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8} \tag{5.40}$$

I_{peak} is determined by the saturation current of the devices and is hence directly proportional to the sizes of the transistors. The peak current is also **a strong function of the ratio between input and output slopes**. This relationship is best illustrated by the following simple analyis: Consider a static CMOS inverter with a $0 \rightarrow 1$ transition at the input. Assume first that the load capacitance is very large, so that the output fall time is significantly larger than the input rise time (Figure 5.28a). Under those circumstances, the input moves through the transient region before the output starts to change. As the source-drain voltage of the PMOS device is approximately 0 during that period, the device shuts off without ever delivering any current. The short-circuit current is close to zero in this case.



(a) Large capacitive load (b) Small capacitive load

Figure 5.28 Impact of load capacitance on short-circuit current.

Consider now the reverse case, where the output capacitance is very small, and the output fall time is substantially smaller than the input rise time (Figure 5.28b). The drain-source voltage of the PMOS device equals V_{DD} for most of the transition period, guaranteeing the maximal short-circuit current (equal to the saturation current of the PMOS). This clearly represents the worst-case condition. The conclusions of the above analysis are confirmed in Figure 5.29, which plots the short-circuit current through the NMOS transistor during a low-to-high transition as a function of the load capacitance.

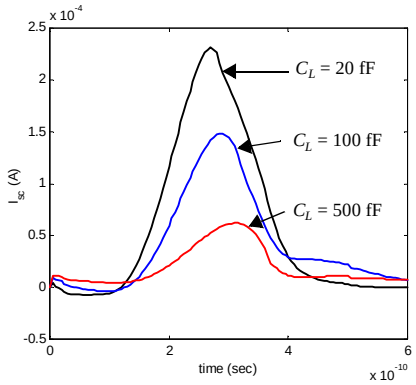


Figure 5.29 CMOS inverter short-circuit current through NMOS transistor as a function of the load capacitance (for a fixed input slope of 500 psec).

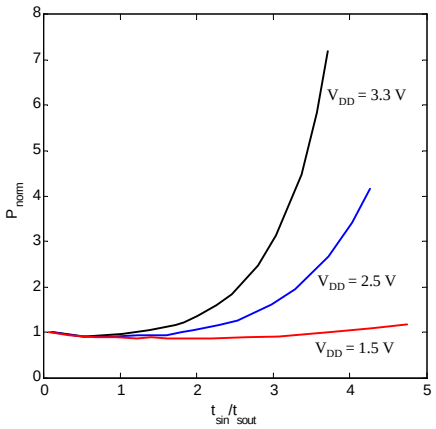
This analysis leads to the conclusion that the short-circuit dissipation is minimized by making the output rise/fall time larger than the input rise/fall time. On the other hand, making the output rise/fall time too large slows down the circuit and can cause short-circuit currents in the fan-out gates. This presents a perfect example of how local optimization and forgetting the global picture can lead to an inferior solution.

Design Techniques

A more practical rule, which optimizes the power consumption in a global way, can be formulated (Veendrick84]):

The power dissipation due to short-circuit currents is minimized by matching the rise/fall times of the input and output signals. At the overall circuit level, this means that rise/fall times of all signals should be kept constant within a range.

Making the input and output rise times of a gate identical is not the optimum solution for that particular gate on its own, but keeps the overall short-circuit current within bounds. This is shown in Figure 5.30, which plots the short-circuit energy dissipation of an inverter (normalized with respect to the zero-input rise time dissipation) as a function of the ratio r between input and output rise/fall times. When the load capacitance is too small for a given inverter size (r > 2…3 for V_{DD} = 5 V), the power is dominated by the short-circuit current. For very large capacitance values, all power dissipation is devoted to charging and discharging the load capacitance. When the rise/fall times of inputs and outputs are equalized, most power dissipation is associated with the dynamic power and only a minor fraction (< 10%) is devoted to short-circuit currents.



$W/L|_P$ = 1.125 µm/0.25 µm
$W/L|_N$ = 0.375 µm/0.25 µm
C_L = 30 fF

Figure 5.30 Power dissipation of a static CMOS inverter as a function of the ratio between input and output rise/fall times. The power is normalized with respect to zero input rise-time dissipation. At low values of the slope ratio, input-output coupling leads to some extra dissipation.

Observe also that the impact of **short-circuit current is reduced when we lower the supply voltage**, as is apparent from Eq. (5.40). In the extreme case, when $V_{DD} < V_{Tn} + |V_{Tp}|$, short-circuit dissipation is completely eliminated, because both devices are never on simultaneously. With threshold voltages scaling at a slower rate than the supply voltage, short-circuit power dissipation is becoming of a lesser importance in deep-submicron technologies. At a supply voltage of 2.5 V and thresholds around 0.5 V, an input/output slope ratio of 2 is needed to cause a 10% degradation in dissipation.

Finally, its is worth observing that the short-circuit power dissipation can be modeled by adding a load capacitance $C_{sc} = t_{sc}I_{peak}/V_{DD}$ in parallel with C_L, as is apparent in Eq. (5.39). The value of this short-circuit capacitance is a function of V_{DD}, the transistor sizes, and the input-output slope ratio.

5.5.2 Static Consumption

The static (or steady-state) power dissipation of a circuit is is expressed by Eq. (5.41), where I_{stat} is the current that flows between the supply rails in the absence of switching activity

$$P_{stat} = I_{stat} V_{DD} \tag{5.41}$$

Ideally, the static current of the CMOS inverter is equal to zero, as the PMOS and NMOS devices are never on simultaneously in steady-state operation. There is, unfortunately, a leakage current flowing through the reverse-biased diode junctions of the transistors, located between the source or drain and the substrate as shown in Figure 5.31. This contribution is, in general, very small and can be ignored. For the device sizes under consideration, the leakage current per unit drain area typically ranges between 10-100 pA/μm^2 at room temperature. For a die with 1 million gates, each with a drain area of 0.5 μm^2 and operated at a supply voltage of 2.5 V, the worst-case power consumption due to diode leakage equals 0.125 mW, which is clearly not much of an issue.

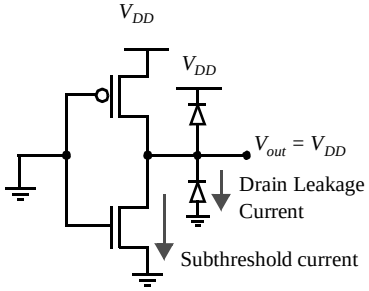


Figure 5.31 Sources of leakage currents in CMOS inverter (for V_{in} = 0 V).

However, be aware that the junction leakage currents are caused by thermally generated carriers. Their value increases with increasing junction temperature, and this occurs in an exponential fashion. At 85°C (a common junction temperature limit for commercial hardware), the leakage currents increase by a factor of 60 over their room-temperature values. Keeping the overall operation temperature of a circuit low is consequently a desirable goal. As the temperature is a strong function of the dissipated heat and its removal mechanisms, this can only be accomplished by limiting the power dissipation of the circuit and/or by using chip packages that support efficient heat removal.

An emerging source of leakage current is the subthreshold current of the transistors. As discussed in Chapter 3, an MOS transistor can experience a drain-source current, even when V_{GS} is smaller than the threshold voltage (Figure 5.32). The closer the threshold voltage is to zero volts, the larger the leakage current at V_{GS} = 0 V and the larger the static power consumption. To offset this effect, the threshold voltage of the device has generally been kept high enough. Standard processes feature V_T values that are never smaller than 0.5-0.6V and that in some cases are even substantially higher (~ 0.75V).

This approach is being challenged by the reduction in supply voltages that typically goes with deep-submicron technology scaling as became apparent in Figure 3.40. We con-

cluded earlier (Figure 5.17) that scaling the supply voltages while keeping the threshold voltage constant results in an important loss in performance, especially when V_{DD} approaches 2 V_T . One approach to address this performance issue is to scale the device thresholds down as well. This moves the curve of Figure 5.17 to the left, which means that the performance penalty for lowering the supply voltage is reduced. Unfortunately, the threshold voltages are lower-bounded by the amount of allowable subthreshold leakage current, as demonstrated in Figure 5.32. The choice of the threshold voltage hence represents a trade-off between performance and static power dissipation. The continued scaling of the supply voltage predicted for the next generations of CMOS technologies will however force the threshold voltages ever downwards, and will make subthreshold conduction a dominant source of power dissipation. Process technologies that contain devices with

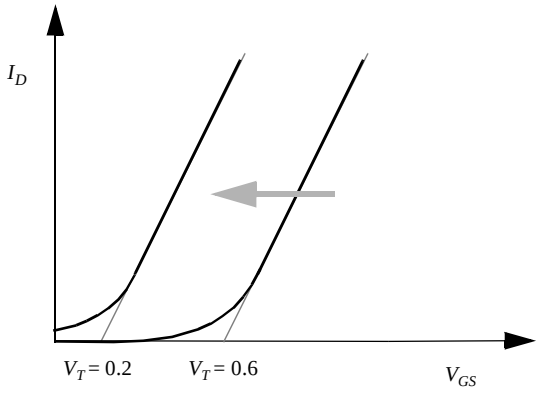


Figure 5.32 Decreasing the threshold increases the subthreshold current at V_{GS} = 0.

sharper turn-off characteristic will therefore become more attractive. An example of the latter is the SOI (Silicon-on-Insulator) technology whose MOS transistors have slope-factors that are close to the ideal 60 mV/decade.

Example 5.13 Impact of threshold reduction on performance and static power dissipation

Consider a minimum size NMOS transistor in the 0.25 μm CMOS technology. In Chapter 3, we derived that the slope factor S for this device equals 90 mV/decade. The off-current (at V_{GS} = 0) of the transistor for a V_T of approximately 0.5V equals 10^{-11}A (Figure 3.22). Reducing the threshold with 200 mV to 0.3 V multiplies the off-current of the transistors with a factor of 170! Assuming a million gate design with a supply voltage of 1.5 V, this translates into a static power dissipation of $10^6 \times 170 \times 10^{-11} \times 1.5$ = 2.6 mW. A further reduction of the threshold to 100 mV results in an unacceptable dissipation of almost 0.5 W! At that supply voltage, the threshold reductions correspond to a performance improvement of 25% and 40%, respectively.

This lower bound on the thresholds is in some sense artificial. The idea that the leakage current in a static CMOS circuit has to be zero is a preconception. Certainly, the presence of leakage currents degrades the noise margins, because the logic levels are no longer equal to the supply rails. As long as the noise margins are within range, this is not a compelling issue. The leakage currents, of course, cause an increase in static power dissipa-

tion. This is offset by the drop in supply voltage, that is enabled by the reduced thresholds at no cost in performance, and results in a quadratic reduction in dynamic power. For a 0.25 μm CMOS process, the following circuit configurations obtain the same performance: 3 V supply–0.7 V V_T; and 0.45 V supply–0.1 V V_T. The dynamic power consumption of the latter is, however, 45 times smaller [Liu93]! Choosing the correct values of supply and threshold voltages once again requires a trade-off. The optimal operation point depends upon the activity of the circuit. In the presence of a sizable static power dissipation, it is essential that non-active modules are *powered down*, lest static power dissipation would become dominant. Power-down (also called *standby*) can be accomplished by disconnecting the unit from the supply rails, or by lowering the supply voltage.

5.5.3 Putting It All Together

The total power consumption of the CMOS inverter is now expressed as the sum of its three components:

$$P_{tot} = P_{dyn} + P_{dp} + P_{stat} = (C_L V_{DD}^2 + V_{DD} I_{peak} t_s) f_{0\rightarrow 1} + V_{DD} I_{leak} \tag{5.42}$$

In typical CMOS circuits, the capacitive dissipation is by far the dominant factor. The direct-path consumption can be kept within bounds by careful design, and should hence not be an issue. Leakage is ignorable at present, but this might change in the not too distant future.

The Power-Delay Product, or Energy per Operation

In Chapter 1, we introduced the *power-delay product* (PDP) as a quality measure for a logic gate.

$$PDP = P_{av} t_p \tag{5.43}$$

The PDP presents a measure of energy, as is apparent from the units (Wsec = Joule). Assuming that the gate is switched at its maximum possible rate of $f_{max} = 1/(2t_p)$, and ignoring the contributions of the static and direct-path currents to the power consumption, we find

$$PDP = C_L V_{DD}^2 f_{max} t_p = \frac{C_L V_{DD}^2}{2} \tag{5.44}$$

The PDP stands for the **average energy consumed per switching event** (this is, for a 0→1, or a 1→0 transition). Remember that earlier we had defined E_{av} as the average energy per switching cycle (or per energy-consuming event). As each inverter cycle contains a 0→1, and a 1→0 transition, E_{av} hence is twice the PDP.

Energy-Delay Product

The validity of the PDP as a quality metric for a process technology or gate topology is questionable. It measures the energy needed to switch the gate, which is an important

property for sure. Yet for a given structure, this number can be made arbitrarily low by reducing the supply voltage. From this perspective, the optimum voltage to run the circuit at would be the lowest possible value that still ensures functionality. This comes at the major expense in performance, at discussed earlier. A more relevant metric should combine a measure of performance and energy. The energy-delay product (EDP) does exactly that.

$$EDP = PDP \times t_p = P_{av} t_p^2 = \frac{C_L V_{DD}^2}{2} t_p \tag{5.45}$$

It is worth analyzing the voltage dependence of the EDP. Higher supply voltages reduce delay, but harm the energy, and the opposite is true for low voltages. An optimum operation point should hence exist. Assuming that NMOS and PMOS transistors have comparible threshold and saturation voltages, we can simplify the propagation delay expression Eq. (5.21).

$$t_p \approx \frac{\alpha C_L V_{DD}}{V_{DD} - V_{Te}} \tag{5.46}$$

where $V_{Te} = V_T + V_{DSAT}/2$, and α technology parameter. Combining Eq. (5.45) and Eq. (5.46), [4]

$$EDP = \frac{\alpha C_L^2 V_{DD}^3}{2(V_{DD} - V_{TE})} \tag{5.47}$$

The optimum supply voltage can be obtained by taking the derivative of Eq. (5.47) with respect to V_{DD}, and equating the result to 0.

$$V_{DDopt} = \frac{3}{2} V_{TE} \tag{5.48}$$

The remarkable outcome from this analysis is the low value of the supply voltage that simultaneously optimizes performance and energy. For sub-micron technologies with thresholds in the range of 0.5 V, the optimum supply is situated around 1 V.

Example 5.14 Optimum supply voltage for 0.25 μm CMOS inverter

From the technology parameters for our generic CMOS process presented in Chapter 3, the value of V_{TE} can be derived.

$$V_{Tn} = 0.43 \text{ V}, V_{Dsatn} = 0.63 \text{ V}, V_{TEn} = 0.74 \text{ V}.$$
$$V_{Tp} = -0.4 \text{ V}, V_{Dsatp} = -1 \text{ V}, V_{TEp} = -0.9 \text{ V}.$$
$$V_{TE} \approx (V_{TEn} + |V_{TEp}|)/2 = 0.8 \text{ V}$$

Hence, $V_{DDopt} = (3/2) \times 0.8 \text{ V} = 1.2 \text{ V}$. The simulated graphs of Figure 5.33, plotting normalized delay, energy, and energy-delay product, confirm this result. The optimum supply volt-

[4] This equation is only accurate as long as the devices remain in velocity saturation, which is probably not the case for the lower supply voltages. This introduces some inaccurarcy in the analysis, but will not distort the overall result.

age is predicted to equal 1.1 V. The charts clearly illustrate the trade-off between delay and energy.

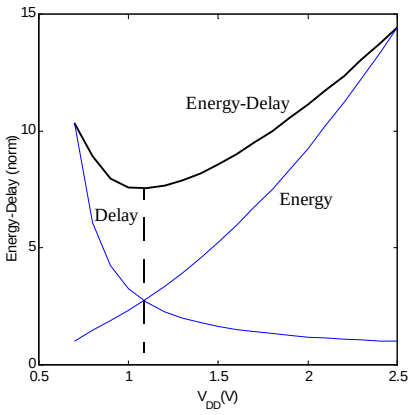


Figure 5.33 Normalized delay, energy, and energy-delay plots for CMOS inverter in 0.25 μm CMOS technology.

WARNING: While the above example demonstrates that there exists a supply voltage that minimizes the energy-delay product of a gate, this voltage does not necessarily represent the optimum voltage for a given design problem. For instance, some designs require a minimum performance, which requires a higher voltage at the expense of energy. Similarly, a lower-energy design is possible by operating by circuit at a lower voltage and by obtaining the overall system performance through the use of architectural techniques such as pipelining or concurrency.

5.5.4 Analyzing Power Consumption Using SPICE

A definition of the average power consumption of a circuit was provided in Chapter 1, and is repeated here for the sake of convenience.

$$P_{av} = \frac{1}{T}\int_0^T p(t)dt = \frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt \tag{5.49}$$

with T the period of interest, and V_{DD} and i_{DD} the supply voltage and current, respectively. Some implementations of SPICE provide built-in functions to measure the average value of a circuit signal. For instance, the HSPICE *.MEASURE TRAN I(VDD) AVG* command computes the area under a computed transient response (*I*(*VDD*)) and divides it by the period of interest. This is identical to the definition given in Eq. (5.49). Other implementations of SPICE are, unfortunately, not as extensive. This is not as bad as it seems, as long as one realizes that SPICE is actually a differential equation solver. A small circuit can easily be conceived that acts as an integrator and whose output signal is nothing but the average power.

Consider, for instance, the circuit of Figure 5.34. The current delivered by the power supply is measured by the current-controlled current source and integrated on the capacitor C. The resistance R is only provided for DC-convergence reasons and should be chosen as high as possible to minimize leakage. A clever choice of the element parameter ensures that the output voltage P_{av} equals the average power consumption. The operation of the circuit is summarized in Eq. (5.50) under the assumption that the initial voltage on the capacitor C is zero.

$$C\frac{dP_{av}}{dt} = ki_{DD}$$
$$or \tag{5.50}$$
$$P_{av} = \frac{k}{C}\int_0^T i_{DD}dt$$

Equating Eq. (5.49) and Eq. (5.50) yields the necessary conditions for the equivalent circuit parameters: $k/C = V_{DD}/T$. Under these circumstances, the equivalent circuit shown presents a convenient means of tracking the average power in a digital circuit.

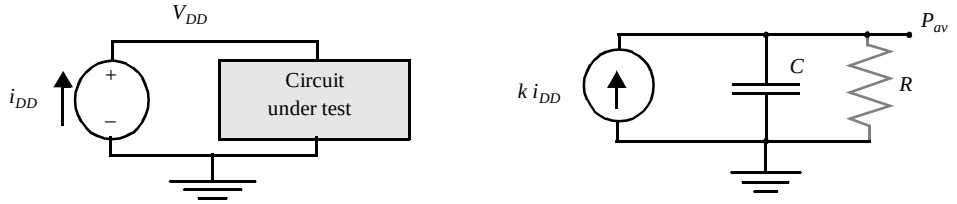


Figure 5.34 Equivalent circuit to measure average power in SPICE.

Example 5.15 Average Power of Inverter

The average power consumption of the inverter of Example 5.4 is analyzed using the above technique for a toggle period of 250 psec (T = 250 psec, k = 1, V_{DD} = 2.5 V, hence C = 100 pF). The resulting power consumption is plotted in Figure 5.35, showing an average power consumption of approximately 157.3 μW. The .MEAS AVG command yields a value of 160.32 μW, which demonstrates the approximate equivalence of both methods. These numberes are equivalent to an energy of 39 fJ (which is close to the 37.5 fJ derived in Example 5.10). Observe the slightly negative dip during the high-to-low transition. This is due to the injection of current into the supply, when the output briefly overshoots V_{DD} as a result of the capacitive coupling between input and output (as is apparent from in the transient response of Figure 5.16).

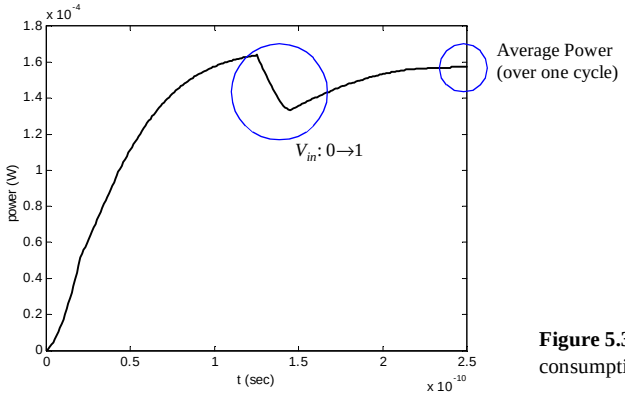


Figure 5.35 Deriving the power consumption using SPICE.

5.6 Perspective: Technology Scaling and its Impact on the Inverter Metrics

In section 3.5, we have explored the impact of the scaling of technology on the some of the important design parameters such as area, delay, and power. For the sake of clarity, we repeat here some of the most important entries in the resulting scaling table (Table 3.8).

Table 5.3 Scaling scenarios for short-channel devices (*S* and *U* represent the technology and voltage scaling parameters, respectively).

Parameter	Relation	Full Scaling	General Scaling	Fixed-Voltage Scaling
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
Intrinsic Delay	$R_{on}C_{gate}$	$1/S$	$1/S$	$1/S$
Intrinsic Energy	$C_{gate}V^2$	$1/S^3$	$1/SU^2$	$1/S$
Intrinsic Power	*Energy/Delay*	$1/S^2$	$1/U^2$	1
Power Density	*P/Area*	1	S^2/U^2	S^2

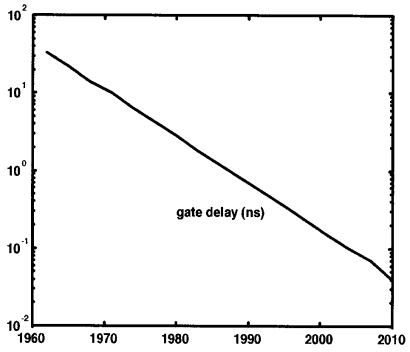


Figure 5.36 Scaling of the gate delay (from [Dally98]).

To validity of these theoretical projections can be verified by looking back and observing the trends during the past decades. From , we can derive that the gate delay indeed decreases exponentially at a rate of 13%/year, or halving every five years. This rate is on course with the prediction of Table 5.3, since *S* averages approximately 1.15 as we had already observed in Figure 3.39. The delay of a 2-input NAND gate with a fanout of four has gone from tens of nanoseconds in the 1960s to a tenth of a nanosecond in the year 2000, and is projected to be a few tens of picoseconds by 2010.

Reducing power dissipation has only been a second-order priority until recently. Hence, statistics on dissipation-per-gate or design are only marginally available. An interesting chart is shown in Figure 5.37, which plots the power density measured over a large number of designs produced between 1980 and 1995. Although the variation is large—even for a fixed technology—it shows the power density to increase approximately with S^2. This is in correspondence with the fixed-voltage scaling scenario presented in Table 5.3. For more recent years, we expect a scenario more in line with the the full-scaling model—which predicts a constant power density—due to the accelerated supply-voltage scaling and the increased attention to power-reducing design techniques. Even under these circumstances, power dissipation-per-chip will continue to increase due to the ever-larger die sizes.

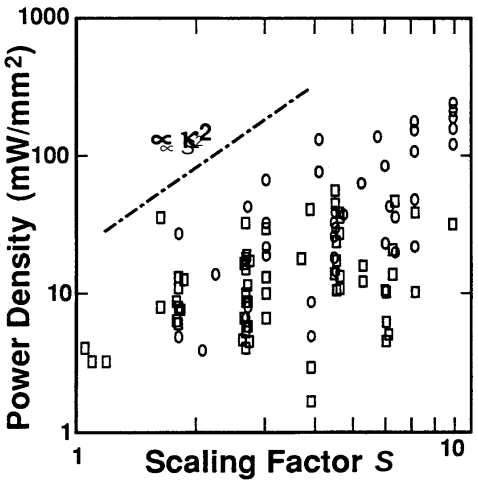


Figure 5.37 Evolution of power-density in micro- and DSP processors, as a function of the scaling factor S ([Sakurai]). *S* is normalized to 1 for a 4 µm process.

The presented scaling model has one fatal flaw however: the performance and power predictions produce purely "intrinsic" numbers that take only device parameters into account. In Chapter 4, it was concluded that the interconnect wires exhibit a different scaling behavior, and that wire parasitics may come to dominate the overall performance. Similarly, charging and discharging the wire capacitances may dominate the energy bud-

get. To get a crisper perspective, one has to construct a combined model that considers device and wire scaling models simultaneously. The impact of the wire capacitance and its scaling behavior is summarized in Table 5.4. We adopt the fixed-resistance model introduced in Chapter 4. We furthermore assume that the resistance of the driver dominates the wire resistance, which is definitely the case for short to medium-long wires.

Table 5.4 Scaling scenarios for wire capacitance. *S* and *U* represent the technology and voltage scaling parameters, respectively, while S_L stands for the wire-length scaling factor. ε_c represents the impact of fringing and interwire capacitances.

Parameter	Relation	General Scaling
Wire Capacitance	WL/t	ε_c/S_L
Wire Delay	$R_{on}C_{int}$	ε_c/S_L
Wire Energy	$C_{int}V^2$	$\varepsilon_c/S_L U^2$
Wire Delay / Intrinsic Delay		$\varepsilon_c S/S_L$
Wire Energy / Intrinsic Energy		$\varepsilon_c S/S_L$

The model predicts that the interconnect-caused delay (and energy) gain in importance with the scaling of technology. This impact is limited to an increase with ε_c for short wires ($S = S_L$), but it becomes increasingly more outspoken for medium-range and long wires ($S_L < S$). These conclusions have been confirmed by a number of studies, an example of which is shown in Figure 5.38. How the ratio of wire over intrinsic contributions

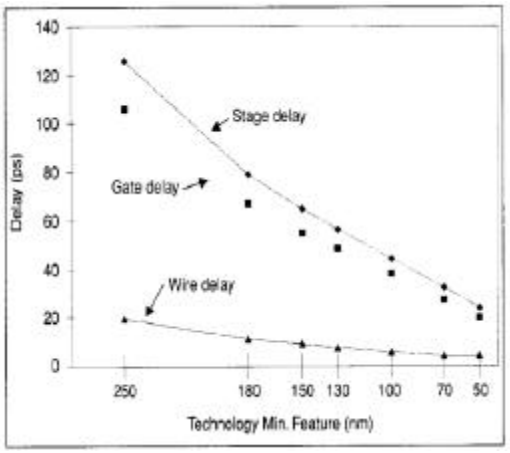


Figure 5.38 Evolution of wire delay / gate delay ratio with respect to technology (from [Fisher98]).

will actually evolve is debatable, as it depends upon a wide range of independent parameters such as system architecture, design methodology, transistor sizing, and interconnect materials. The doomday scenario that interconnect may cause CMOS performance to saturate in the very near future hence may be exagerated. Yet, it is clear to that increased attention to interconnect is an absolute necessity, and may change the way the next-generation circuits are designed and optimized (e.g. Sylvester99]).

5.7 Summary

This chapter presented a rigourous and in-depth analysis of the static CMOS inverter. The key characteristics of the gate are summarized:

- The static CMOS inverter combines a pull-up PMOS section with a pull-down NMOS device. The PMOS is normally made wider than the NMOS due to its inferior current-driving capabilities.
- The gate has an almost ideal voltage-transfer characteristic. The logic swing is equal to the supply voltage and is not a function of the transistor sizes. The noise margins of a symmetrical inverter (where PMOS and NMOS transistor have equal current-driving strength) approach $V_{DD}/2$. The steady-state response is not affected by fan-out.
- Its propagation delay is dominated by the time it takes to charge or discharge the load capacitor C_L. To a first order, it can be approximated as

$$t_p = 0.69 C_L \left(\frac{R_{eqn} + R_{eqp}}{2} \right)$$

  Keeping the load capacitance small is the most effective means of implementing high-performance circuits. Transistor sizing may help to improve performance as long as the delay is dominated by the extrinsic (or load) capacitance of fanout and wiring.
- The power dissipation is dominated by the dynamic power consumed in charging and discharging the load capacitor. It is given by $P_{0\rightarrow1} C_L V_{DD}^2 f$. The dissipation is proportional to the activity in the network. The dissipation due to the direct-path currents occuring during switching can be limited by careful tailoring of the signal slopes. The static dissipation can usually be ignored but might become a major factor in the future as a result of subthreshold currents.
- Scaling the technology is an effective means of reducing the area, propagation delay and power consumption of a gate. The impact is even more striking if the supply voltage is scaled simultaneously.
- The interconnect component is gradually taking a larger fraction of the delay and performance budget.

5.8 To Probe Further

The operation of the CMOS inverter has been the topic of numerous publications and textbooks. Virtually every book on digital design devotes a substantial number of pages to the analysis of the basic inverter gate. An extensive list of references was presented in Chapter 1. Some references of particular interest that were explicitly quoted in this chapter are given below.

REFERENCES

[Baccarani84] G. Baccarani, M. Wordeman, and R. Dennard, "Generalized Scaling Theory and Its Application to 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Devices*, ED-31(4): p. 452, 1984.

[Brews89] J. Brews et al., "The Submicrometer Silicon MOSFET," in [Watts89].

[De Man87] De Man H., *Computer Aided Design of Digital Integrated Circuits*, Lecture Notes, Katholieke Universiteit Leuven, Belgium.

[Dennard74] R. Dennard et al., "Design of Ion-Implanted MOSFETS with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, SC-9, pp. 256–258, 1974.

[Embabi93] S. Embabi, A. Bellaouar, M. Elmasry, *Digital BiCMOS Integrated Circuit Design*, Kluwer Academic Publishers, 1993.

[Hodges88] D. Hodges and H. Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill, 1988.

[Jouppi93] N. Jouppi et al., "A 300 MHz 115W 32b Bipolar ECL Microprocessor with On-Chip Caches," *Proc. IEEE ISSCC Conf*., pp. 84–85, February 1993.

[Kakumu90] M. Kakumu and M. Kinugawa, "Power-Supply Voltage Impact on Circuit Performance for Half and Lower Submicrometer CMOS LSI," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 8, pp. 1900–1908, August 1990.

[Lohstroh81] J. Lohstroh, "Devices and Circuits for Bipolar (V)LSI," *Proceedings of the IEEE*, vol. 69, pp. 812–826, July 1981.

[Masaki92] A. Masaki, "Deep-Submicron CMOS Warms Up to High-Speed Logic," *IEEE Circuits and Devices Magazine*, pp. 18–24, November 1992.

[Schutz94] J. Schutz, "A 3.3V 0.6 mm BiCMOS Superscaler Microprocessor," *ISSCC Digest of Technical Papers*, pp. 202–203, February 1994.

[Sedra87] Sedra and Smith, *MicroElectronic Circuits*, Holt, Rinehart and Winston, 1987.

[Shoji88] Shoji, M., *CMOS Digital Circuit Technology*, Prentice Hall, 1988.

[Tang89] D. Tang, "Scaling the Silicon Bipolar Transistor," in [Watts89].

[Veendrick84] H. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 4, pp. 468–473, 1984.

[Watts89] Watts R., ed., *SubMicron Integrated Circuits*, Wiley, 1989.

5.9 Exercises and Design Problems

For all problems, use the device parameters provided in Chapter 3 (as well as the inside back cover), unless otherwise mentioned.

1. [M, SPICE, 3.3.2] The layout of a static CMOS inverter is given in Figure 5.39. ($1\lambda = 0.6$ μm).

a. Determine the sizes of the NMOS and PMOS transistor.

b. Derive the VTC and its parameters (V_{OH}, V_{OL}, V_M, V_{IH}, and V_{IL}).

c. Is the VTC affected when the output of the gates is connected to the inputs of 4 similar gates?