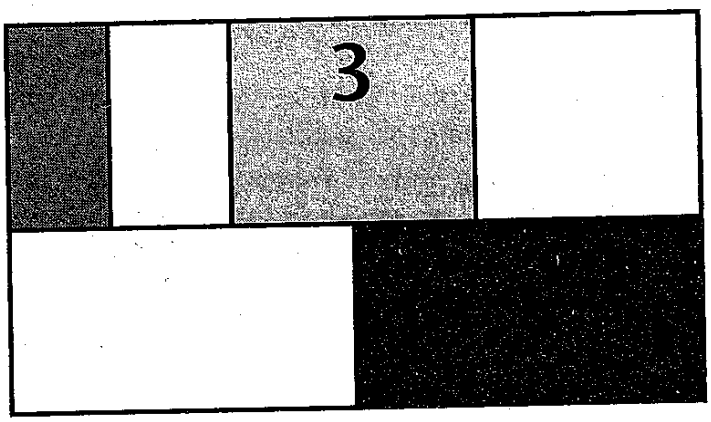
CHAPTER 3



THE INVERTER

Quality measures of a digital gate: area, robustness, speed, and energy consumption

Analyzing and optimizing an inverter design

Two contrasting approaches: static CMOS and bipolar ECL

3.1 Introduction

3.2 Definitions and Properties
- 3.2.1 Area and Complexity
- 3.2.2 Functionality and Robustness: The Static Behavior
- 3.2.3 Performance: The Dynamic Behavior
- 3.2.4 Power and Energy Consumption

3.3 The Static CMOS Inverter
- 3.3.1 A First Glance
- 3.3.2 Evaluating the Robustness of the CMOS Inverter: The Static Behavior
- 3.3.3 Performance of CMOS Inverter: The Dynamic Behavior
- 3.3.4 Power Consumption and Power-Delay Product
- 3.3.5 A Look into the Future: Effects of Technology Scaling

3.4 The Bipolar ECL Inverter
- 3.4.1 Issues in Bipolar Digital Design: A Case Study
- 3.4.2 The Emitter-Coupled Logic (ECL) Gate at a Glance
- 3.4.3 Robustness and Noise Immunity: The Steady-State Characteristics
- 3.4.4 ECL Switching Speed: The Transient Behavior
- 3.4.5 Power Consumption
- 3.4.6 Looking Ahead: Scaling the Technology

3.5 Perspective: Area, Performance, and Dissipation

3.1 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.

The choice of a technology or a design style dramatically affects the density, performance, and the power consumption of a design. To illustrate this, we discuss in detail the behavior of static complementary CMOS and bipolar ECL inverters, which are representative gates for both MOS and bipolar technologies. Although these are not the only gate topologies in use (see Chapters 4 and 5), they are certainly the most popular at present. For each gate, we analyze the following fundamental properties:

- *robustness*, expressed by the static (or steady-state) behavior
- *performance*, determined by the dynamic (or transient) response
- *heat dissipation and supply capacity requirements*, set by the power consumption

The first section provides precise definitions for each of the above properties. 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*. Finally, the properties of the presented gates are summarized, and some suggestions are provided on selecting a technology.

3.2 Definitions and Properties

This section defines a set of basic properties of a digital gate. These properties help to quantify the behavior of a gate from different perspectives: complexity, functionality, robustness, performance, and energy consumption. Although we concentrate here on the behavior of the simplest of all gates, the inverter, similar properties can also be defined for more complex components, such as NAND, NOR, and XOR gates.

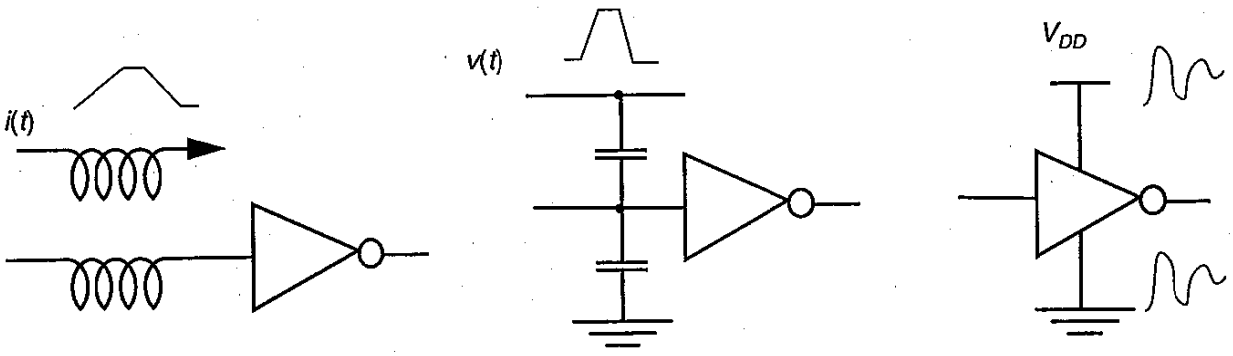
3.2.1 Area and Complexity

Having a small area is, obviously, a desirable property for a digital gate. The smaller the gate, the higher the integration density and the smaller the die size. Die size directly relates to the *fabrication cost* of a design. Smaller gates tend also to be faster, as the total gate capacitance—which is one of the dominant performance parameters—often scales with the area.

The *number of transistors* in a gate is indicative for the expected implementation area. Other parameters may have an impact, though. For instance, a complex interconnect pattern between the transistors can cause the wiring area to dominate.

3.2.2 Functionality and Robustness: The Static Behavior

A prime requirement for a digital gate is, obviously, that it perform the digital function it is designed for. The measured behavior of a manufactured gate normally deviates from the expected response. One reason for this aberration are the variations in the manufacturing process. As was explained in Chapter 2, the dimensions, threshold voltages, and currents of an MOS transistor vary between runs or even on a single wafer or die. The electrical behavior of a circuit can be profoundly affected by those variations. The presence of disturbing noise sources on or off the chip is another source of deviations in circuit response. The word *noise* in the context of digital circuits means *"unwanted variations of voltages and currents at the logic nodes."* Noise signals can enter a circuit in many ways. Some examples of digital noise sources are depicted in Figure 3.1. For instance, two wires placed side by side in an integrated circuit form a coupling capacitor and a mutual inductance. Hence, a voltage or a current change on one of the wires can influence the signals on the neighboring wire. Noise on the power and ground rails of a gate also influences the signal levels in the gate. How to cope with all these disturbances is one of the main challenges in the design of high-performance digital circuits and is a recurring topic in this book.



(a) Inductive coupling (b) Capacitive coupling (c) Power and ground noise

Figure 3.1 Noise sources in digital circuits.

The steady-state parameters of a gate measure how robust the structure is with respect to both variations in the manufacturing process and noise disturbances. The definition and derivation of these parameters requires a prior understanding of how digital signals are represented in the world of electronic circuits.

Digital circuits (DC) perform operations on *logical* (or *Boolean*) variables. A logical variable x can only assume two discrete values:

$$x \in \{0,1\}$$

As an example, the inversion (i.e., the function that an inverter performs) implements the following compositional relationship between two Boolean variables x and y:

$$y = \bar{x}: \{x = 0 \Rightarrow y = 1; x = 1 \Rightarrow y = 0\} \tag{3.1}$$

A logical variable is, however, a mathematical abstraction. In a physical implementation, such a variable is represented by an electrical quantity. This is most often a node

voltage that is not discrete but can adopt a continuous range of values. It is necessary to turn the electrical voltage into a discrete variable by associating *a nominal voltage level* with each logic state: $1 \Leftrightarrow V_{OH}$, $0 \Leftrightarrow V_{OL}$, where V_{OH} and V_{OL} represent the *high* and the *low* logic levels, respectively. Applying V_{OH} to the input of the gate yields V_{OL} at the output and vice versa. The difference between the two is called the *logic swing*.

$$\begin{aligned} V_{OH} &= \overline{(V_{OL})} \\ V_{OL} &= \overline{(V_{OH})} \end{aligned} \tag{3.2}$$

The Voltage-Transfer Characteristic

Assume now that a logical variable *in* serves as the input to an inverting gate that produces the variable *out*. The electrical function of a gate is best expressed by its *voltage-transfer characteristic* (VTC) (sometimes called the *DC transfer characteristic*), which plots the output voltage as a function of the input voltage $V_{out} = f(V_{in})$. An example of an inverter VTC is shown in Figure 3.2. The high and low nominal voltages, V_{OH} and V_{OL}, can readily be identified—$V_{OH} = f(V_{OL})$ and $V_{OL} = f(V_{OH})$. Another point of interest of the VTC is the *gate or switching threshold voltage* V_M (not to be confused with the threshold voltage of a transistor), that is defined as $V_M = f(V_M)$. V_M can also be found graphically at the intersection of the VTC curve and the line given by $V_{out} = V_{in}$. The gate threshold voltage presents the midpoint of the switching characteristics, which is obtained when the output of a gate is short-circuited to the input. This point will prove to be of particular interest when studying circuits with feedback (also called *sequential circuits*).

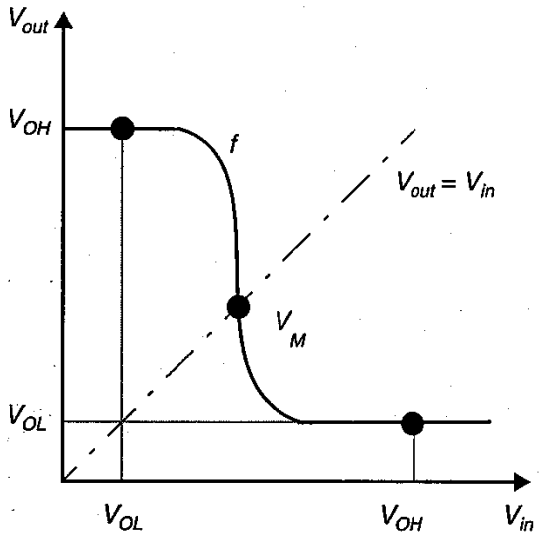
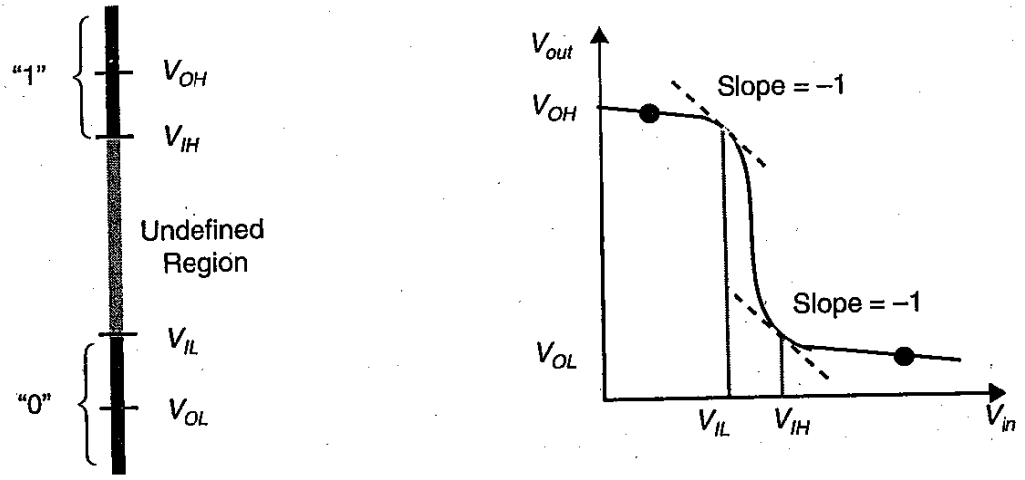


Figure 3.2 Inverter voltage-transfer characteristic.

Even if an ideal nominal value is applied at the input of a gate, the output signal often deviates from the expected nominal value. These deviations can be caused by noise or by the loading on the output of the gate (i.e., by the number of gates connected to the output signal). Figure 3.3a illustrates how a logic level is represented in reality by a range of acceptable voltages, separated by a region of uncertainty, rather than by nominal levels alone. The regions of acceptable high and low voltages are delimited by the V_{IH} and V_{IL} voltage levels, respectively. These represent by definition the points where the gain



(a) Relationship between voltage and logic levels

(b) Definition of V_{IH} and V_{IL}

Figure 3.3 Mapping logic levels to the voltage domain.

($= dV_{out} / dV_{in}$) of the VTC equals –1 as shown in Figure 3.3b. The region between V_{IH} and V_{IL} is called the *undefined region* (sometimes also referred to as *transition width*, or *TW*). Steady-state signals should avoid this region if proper circuit operation is to be ensured.

Noise Margins

For a gate to be robust and insensitive to noise disturbances, it is essential that the "0" and "1" intervals be as large as possible. A measure of the sensitivity of a gate to noise is given by the noise margins NM_L (*noise margin low*) and NM_H (*noise margin high*), which quantize the size of the legal "0" and "1", respectively:

$$NM_L = V_{IL} - V_{OL}$$
$$NM_H = V_{OH} - V_{IH} \tag{3.3}$$

The noise margins represent the levels of noise that can be sustained when gates are cascaded as illustrated in Figure 3.4. It is obvious that the margins should be larger than 0 for a digital circuit to be functional and by preference should be as large as possible.

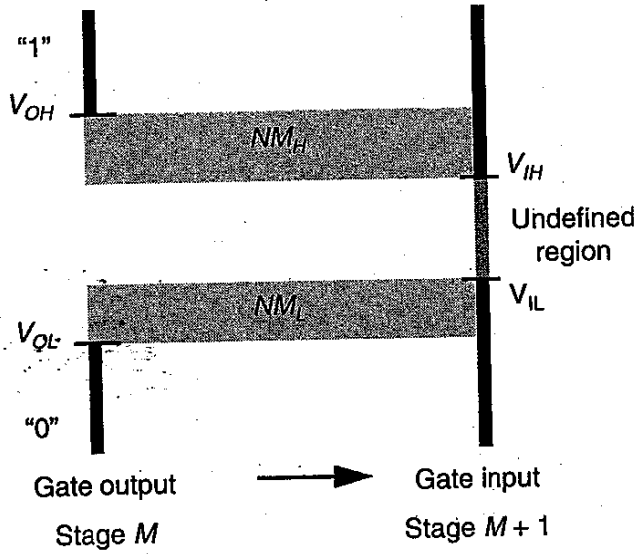
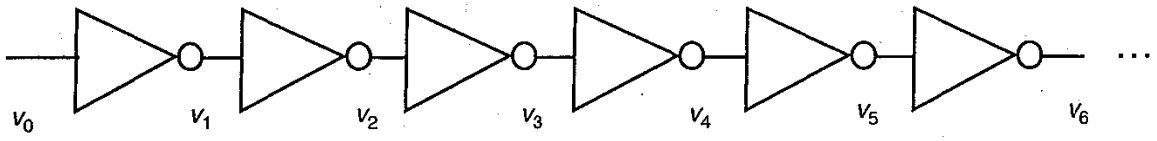


Figure 3.4 Cascaded inverter gates: definition of noise margins.

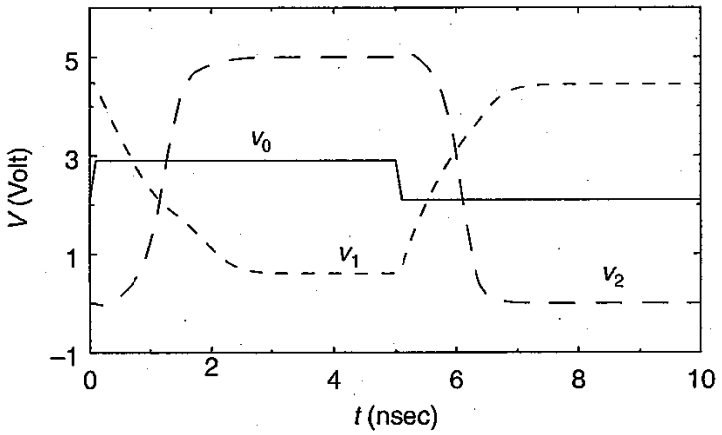
Regenerative Property

A large noise margin is a desirable, but not sufficient requirement. Assume that a signal is disturbed by noise and differs from the nominal voltage levels. As long as the signal is within the noise margins, the following gate continues to function correctly, although its output voltage varies from the nominal one. This deviation is added to the noise injected at the output node and passed to the next gate. The effect of different noise sources may accumulate and eventually force a signal level into the undefined region. This, fortunately, does not happen if the gate possesses the *regenerative property*, which ensures that a disturbed signal gradually converges back to one of the nominal voltage levels after passing through a number of logical stages. This property can be understood as follows:

An input voltage v_{in} ($v_{in} \in$ "0") is applied to a chain of N inverters (Figure 3.5a). Assuming that the number of inverters in the chain is even, the output voltage v_{out} ($N \rightarrow \infty$) will equal V_{OL} if and only if the inverter possesses the regenerative property. Similarly, when an input voltage v_{in} ($v_{in} \in$ "1") is applied to the inverter chain, the output voltage will approach the nominal value V_{OH}.



(a) A chain of inverters



(b) Simulated response of chain of MOS inverters

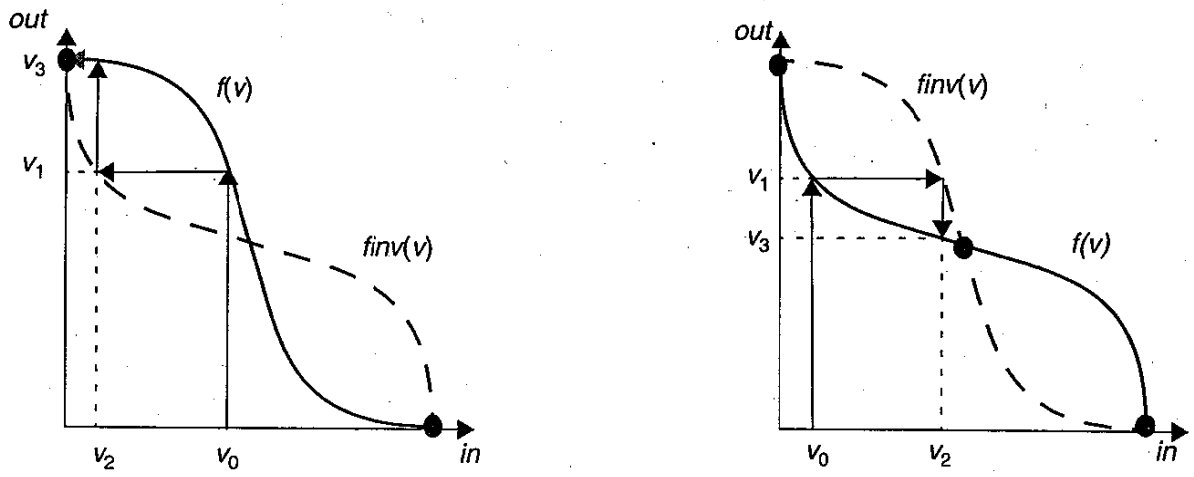
Figure 3.5 The regenerative property.

Example 3.1 Regenerative property

The concept of regeneration is illustrated in Figure 3.5b, which plots the simulated transient response of a chain of CMOS inverters. The input signal to the chain is a step-waveform with a degraded amplitude, which could be caused by noise. Instead of swinging from rail to rail, v_0 only extends between 2.1 and 2.9 V. From the simulation, it can be observed that this deviation rapidly disappears, while progressing through the chain; v_1, for instance, extends from 0.6 V to 4.45 V. Even further, v_2 already swings between the nominal V_{OL} and V_{OH}. The inverter used in this example clearly possesses the regenerative property.

The conditions under which a gate is regenerative can be intuitively derived by analyzing a simple case study. Figure 3.6(a) plots the VTC of an inverter $V_{out} = f(V_{in})$ as well as its inverse function *finv*(), which reverts the function of the x- and y-axis and is defined as follows:

$$in = f(out) \Rightarrow in = finv(out) \tag{3.4}$$



(a) Regenerative gate

(b) Nonregenerative gate

Figure 3.6 Conditions for regeneration.

Assume that a voltage v_0, deviating from the nominal voltages, is applied to the first inverter in the chain. The output voltage of this inverter equals $v_1 = f(v_0)$ and is applied to the next inverter. Graphically this corresponds to $v_1 = finv(v_2)$. The signal voltage gradually converges to the nominal signal after a number of inverter stages, as indicated by the arrows. In Figure 3.6(b) the signal does not converge to any of the nominal voltage levels but to an intermediate voltage level. Hence, the characteristic is nonregenerative. The difference between the two cases is due to the gain characteristics of the gates. To be regenerative, the VTC should have a transient region (or undefined region) with a gain *greater than* 1 in absolute value, bordered by the two legal zones, where the gain should be *smaller than* 1. Such a gate has two stable operating points. This clarifies the definition of the V_{IH} and the V_{IL} levels that form the boundaries between the legal and the transient zones.

Directivity

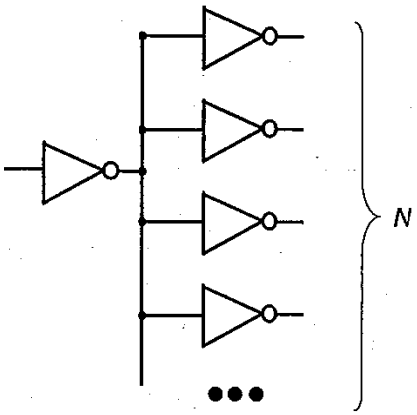
The directivity property requires a gate to be *unidirectional*, that is, changes in an output level should not appear at any unchanging input of the same circuit. If not, an output-signal transition reflects to the gate inputs as a noise signal, affecting the signal integrity.

In real gate implementations, full directivity can never be achieved. Some feedback of changes in output levels to the inputs cannot be avoided. Capacitive coupling between inputs and outputs is a typical example of such a feedback. It is important to minimize these changes so that they do not affect the logic levels of the input signals.

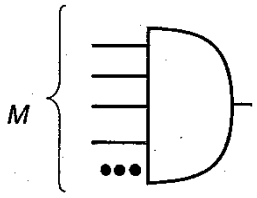
Fan-In and Fan-Out

The *fan-out* denotes *the number of load gates N that are connected to the output of the driving gate* (Figure 3.7). Increasing the fan-out of a gate can affect its logic output levels. From the world of analog amplifiers, we know that this effect is minimized by making the input resistance of the load gates as large as possible (minimizing the input currents) and by keeping the output resistance of the driving gate small (reducing the effects of load currents on the output voltage). When the fan-out is large, the added load can deteriorate the dynamic performance of the driving gate. For these reasons, many generic and library components define a *maximum fan-out* to guarantee that the static and dynamic performance of the element meet specification.

The *fan-in* of a gate is defined as the *number of inputs* to the gate (Figure 3.7b). Gates with large fan-in tend to be more complex, which often results in inferior static and dynamic properties.



(a) Fan-out *N*



(b) Fan-in *M*

Figure 3.7 Definition of fan-out and fan-in of a digital gate.

The Ideal Digital Gate

Based on the above observations, we can define the *ideal* digital gate from a static perspective. The ideal inverter model is important because it gives us a metric by which we can judge the quality of actual implementations.

Its VTC is shown in Figure 3.8 and has the following properties: infinite gain in the transition region, and gate threshold located in the middle of the logic swing, with high and low noise margins equal to half the swing. The input and output impedances of the ideal gate are infinity and zero, respectively (i.e., the gate has unlimited fan-out). While this ideal VTC is unfortunately impossible in real designs, some implementations, such as the static CMOS inverter, come close.

Example 3.2 Voltage-Transfer Characteristic

Figure 3.9 shows an example of a voltage-transfer characteristic of an actual, but outdated gate structure (as produced by SPICE in the DC analysis mode). The values of the dc-parameters are derived from inspection of the graph.

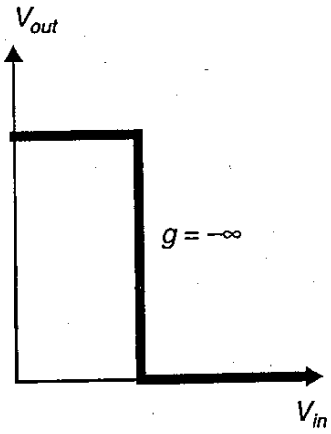


Figure 3.8 Ideal voltage-transfer characteristic.

$$V_{OH} = 3.5\text{ V}; \quad V_{OL} = 0.45\text{ V}$$
$$V_{IH} = 2.35\text{ V}; \quad V_{IL} = 0.66\text{ V}$$
$$V_M = 1.64\text{ V}$$
$$NM_H = 1.15\text{ V}; \quad NM_L = 0.21\text{ V}$$

The observed transfer characteristic, obviously, is far from ideal: it is asymmetrical, has a very low value for NM_L, and the voltage swing of 3.05 V is substantially below the maximum obtainable value of 5 V (which is the value of the supply voltage for this design).

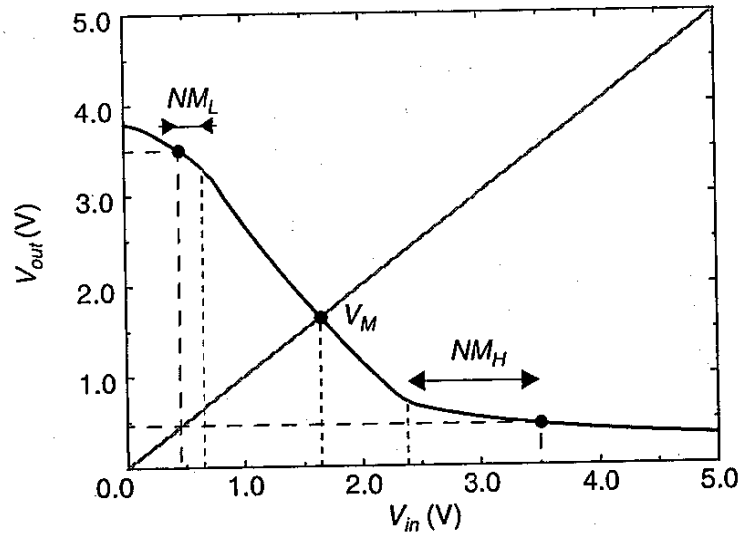


Figure 3.9 Voltage-transfer characteristic of an NMOS inverter of the 1970s.

3.2.3 Performance: The Dynamic Behavior

The *propagation delay* t_p of a gate defines how quickly it responds to a change at its input and relates directly to the speed and performance metrics. The propagation delay expresses *the delay experienced by a signal when passing through a gate*. It is measured

between the 50% transition points of the input and output waveforms, as shown in Figure 3.10 for an inverting gate.[1] Because a gate displays different response times for rising or falling input waveforms, two definitions of the propagation delay are necessary. The t_{pLH} defines the response time of the gate for a *low to high* (or positive) output transition, while t_{pHL} refers to a *high to low* (or negative) transition. The overall propagation delay t_p is defined as the average of the two,

$$t_p = \frac{t_{pLH} + t_{pHL}}{2} \tag{3.5}$$

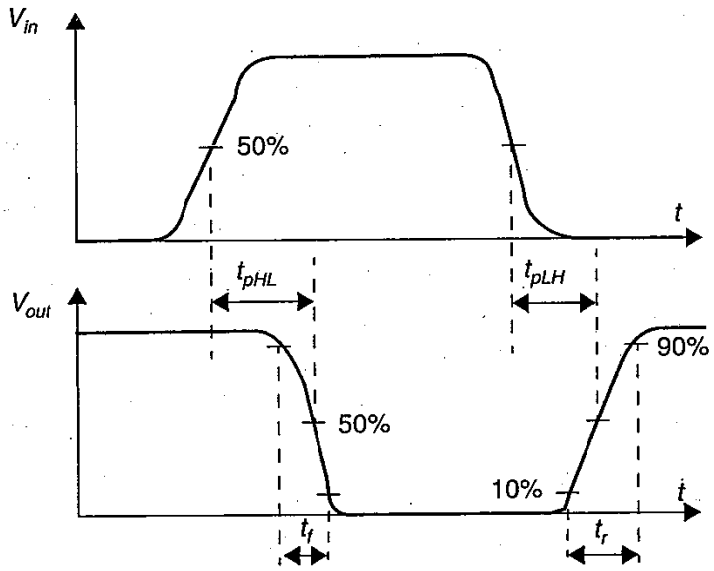


Figure 3.10 Definition of propagation delays and rise and fall times.

Knowledge of t_p is, however, not sufficient to completely characterize circuit performance. The power consumption, noise behavior, and, indirectly, the speed of a gate are also strong functions of the *signal slopes* (as will become clear later in this chapter). This can be quantified with the *rise and fall time* measures t_r and t_f, which are defined between the 10% and 90% points of the waveforms (Figure 3.10).

The propagation delay of a gate is a function of its fan-in and fan-out. Fan-out gates present an increased load (mostly capacitive) to the driving gate and slow its performance. The increased complexity of a gate due to a large fan-in also has a negative influence on the performance. When comparing the performance of gates in different technologies, it is important not to confuse the picture by including second-order parameters such as fan-in and fan-out. It is therefore useful to find a uniform way of measuring the t_p of a gate, so that technologies can be judged on an equal footing. The de-facto standard circuit for delay measurement is the *ring oscillator*, which consists of an odd number of inverters connected in a circular chain (Figure 3.11). Due to the odd number of inversions, this circuit does not have a stable operating point and oscillates. The period T of the oscillation is determined by the propagation time of a signal transition through the complete chain, or

[1] The 50% definition is inspired the assumption that the switching threshold V_M is typically located in the middle of the logic swing.

$T = 2 \times t_p \times N$ with N the number of inverters in the chain. The factor 2 results from the observation that a full cycle requires both a low-to-high and a high-to-low transition. Note that this equation is only valid for $2Nt_p >> t_f + t_r$. If this condition is not met, the circuit might not oscillate—one "wave" of signals propagating through the ring will overlap with a successor and eventually dampen the oscillation. Typically, a ring oscillator needs a least five stages to be operational.

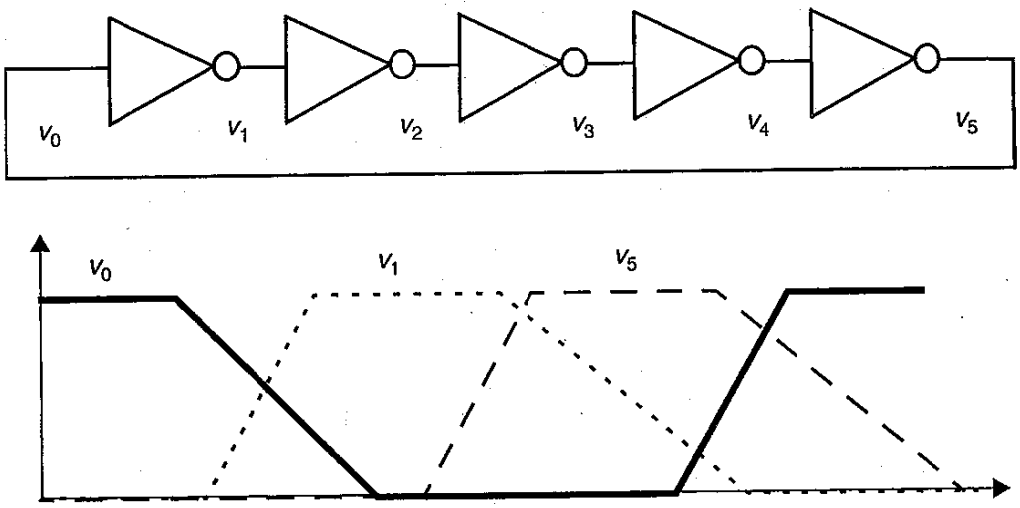


Figure 3.11 Ring oscillator circuit for propagation-delay measurement.

We must be extremely careful with results obtained from ring oscillator measurements. A t_p of 100 psec by no means implies that a circuit built with those gates will operate at 10 GHz. The oscillator results are primarily useful for quantifying the differences between various manufacturing technologies and gate topologies. The oscillator is an idealized circuit where each gate has a fan-in and fan-out of exactly one and parasitic loads are minimal. In more realistic digital circuits, fan-ins and fan-outs are higher, and interconnect delays are non-negligible. The gate functionality is also substantialy more complex than a simple invert operation. As a result, the achievable clock frequency on average is 50 to a 100 times slower than the frequency predicted from ring oscillator measurements. This is an average observation; carefully optimized designs might approach the ideal frequency more closely.

Example 3.3 Propagation Delay of First-Order *RC* Network

Digital circuits are often modeled as first-order *RC* networks of the type shown in Figure 3.12. The propagation delay of such a network is thus of considerable interest.

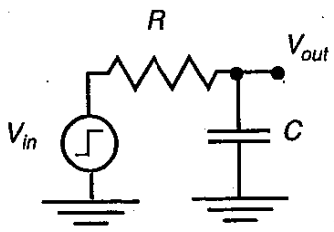


Figure 3.12 First-order *RC* network.

When applying a step input (with V_{in} going from 0 to V), the transient response of this circuit is known to be an exponential function, and is given by the following expression (where $\tau = RC$, the time constant of the network):

$$V_{out}(t) = (1 - e^{-t/\tau})\, V \tag{3.6}$$

The time to reach the 50% point is easily computed as $t = \ln(2)\tau = 0.69\tau$. Similarly, it takes $t = \ln(9)\tau = 2.2\tau$ to get to the 90% point. It is worth memorizing these numbers, as they are extensively used in the rest of the text.

3.2.4 Power and Energy Consumption

The power consumption of a gate determines how much heat the circuit dissipates and how much energy is consumed per operation. These factors influence a great number of critical design decisions, such as the packaging and cooling requirements, supply-line sizing, power-supply capacity, and, most important, the number of circuits that can be integrated onto a single chip. For instance, it is mainly power considerations that prevent the development of very large bipolar digital integrated circuits. Therefore, power dissipation is an important property of a gate that affects feasibility, cost, and reliability. Depending upon the design problem at hand, different dissipation measures have to be considered. For instance, the peak power P_{peak} is important when studying supply-line sizing. When addressing cooling or battery requirements, one is predominantly interested in the average power dissipation P_{av} Both measures are defined in equation Eq. (3.7):

$$\begin{aligned} P_{peak} &= i_{peak} V_{supply} = max[p(t)] \\ P_{av} &= \frac{1}{T}\int_0^T p(t)dt = \frac{V_{supply}}{T}\int_0^T i_{supply}(t)dt \end{aligned} \tag{3.7}$$

where i_{supply} is the current being drawn from the supply voltage V_{supply} over the interval $t \in [0,T]$, and i_{peak} is the maximum value of i_{supply} over that interval. The dissipation can further be decomposed into static and dynamic components. The latter occurs only during transients, when the gate is switching. It is due to the charging of capacitors and temporary current paths between the supply rails, and is, therefore, proportional to the switching frequency: *the higher the number of switching events, the higher the power consumption.* The static component on the other hand is present even when no switching occurs and is caused by static conductive paths between the supply rails or by leakage currents. It is always present, even when the circuit is in stand-by. Minimization of this consumption source is a worthwhile goal.

The propagation delay and the power consumption of a gate are related—the propagation delay is mostly determined by the speed at which a given amount of energy can be stored on the gate capacitors. The faster the energy transfer (or the higher the power consumption), the faster the gate. For a given technology and gate topology, the product of power consumption and propagation delay is generally a constant. This product is called the *power-delay product* (or PDP) and can be considered as a quality measure for a switching device. The PDP is simply the *energy* consumed by the gate *per switching event.* The ring oscillator is again the circuit of choice for measuring the PDP of a logic family.

3 The Static CMOS Inverter

In the following sections, we will derive the static and dynamic parameters for two popular gates, i.e., the CMOS and the ECL inverter. For each structure, we will initiate the discussion with an intuitive analysis of the gate operation.

3.3.1 A First Glance

Figure 3.13 shows the circuit diagram of a static CMOS inverter. Its operation is readily understood from a simplified circuit model. Qualitatively, an MOS transistor can be modeled as *a switch with a finite on-resistance* R_{on}. When $|V_{GS}| < |V_T|$, the switch is open; when $V_{GS} > V_T$, the transistor behaves as a finite resistance (Figure 3.14a). This leads to the following interpretation of the inverter. When V_{in} is high (or equal to V_{DD}), the NMOS transistor is on, while the PMOS is off. This yields the equivalent circuit of Figure 3.14b. 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 3.14c shows that a path exists between V_{DD} and V_{out}, yielding a high output voltage. Notice that no path exists between the supply and ground in steady-state operation. Consequently, the inverter does not consume any static power (ignoring leakage).

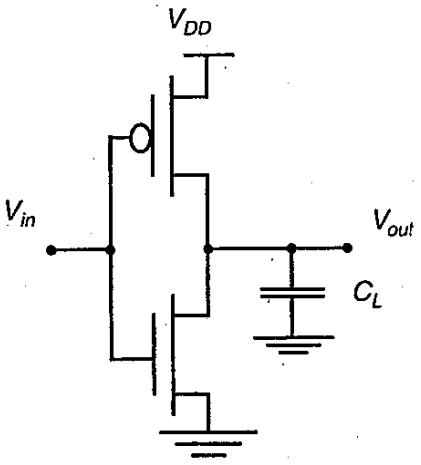


Figure 3.13 Static CMOS inverter. V_{DD} stands for the supply voltage.

SIDELINE: The above observation is one of the prime reasons CMOS superceded NMOS as the digital technology of choice. NMOS was very popular in the 1970s and early 1980s. All early microprocessors, such as the Intel 4004, were pure NMOS designs. The lack of complementary devices (such as the NMOS and PMOS transistor) in a pure NMOS technology makes the realization of inverters with zero static power nontrivial. This static power consumption of the basic NMOS gate puts an upper bound on the number of gates that can be integrated on a single chip; hence the move to CMOS in the 1980s.

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

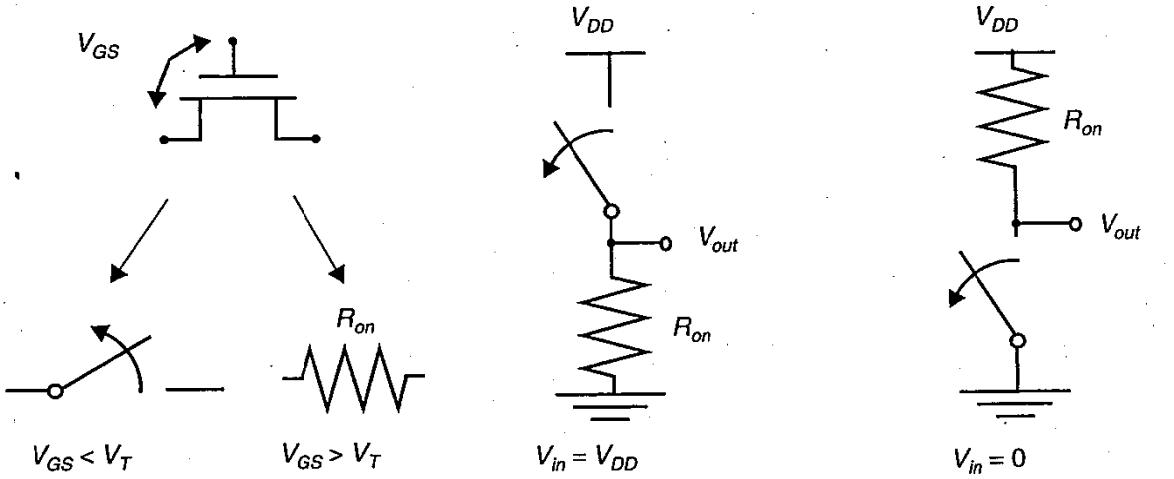


Figure 3.14 Switch models of CMOS inverter.

- 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 the range of 10 kΩ (for the technology under consideration).
- 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.

The nature and the form of the voltage-transfer characteristic (VTC) can be graphically deduced from the load-line plots, which superimpose the current characteristics of the NMOS and the PMOS devices. Creating such a graph requires that the *I-V* curves of the NMOS and PMOS devices are transformed onto a common coordinate set. We have selected the output voltage and the NMOS drain current I_{DN} as the independent variables. 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).

$$I_{DSp} = -I_{DSn}$$
$$V_{GSn} = V_{in} \; ; \; V_{GSp} = V_{in} - V_{DD} \tag{3.8}$$
$$V_{DSn} = V_{out} \; ; \; V_{DSp} = V_{out} - V_{DD}$$

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 illustrated in Figure 3.15, 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 enumerated.

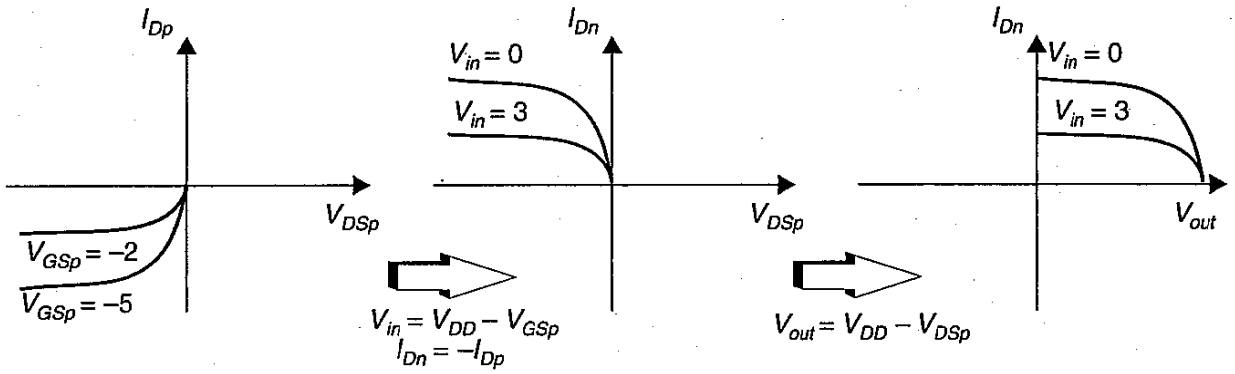


Figure 3.15 Transforming PMOS I-V characteristic to a common coordinate set.

The resulting load lines are plotted in Figure 3.16.[2] For valid dc operating points, 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, 1, 2, 3, 4, and 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

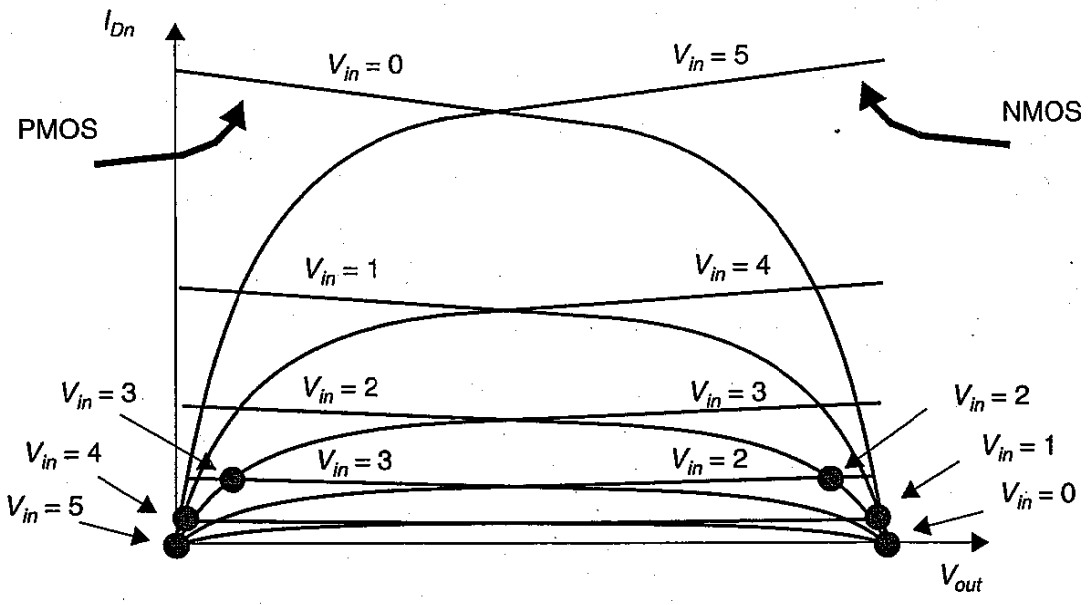


Figure 3.16 Load curves for NMOS and PMOS transistors of the static CMOS inverter (V_{DD} = 5 V). The dots represent the dc operation points for various input voltages.

[2] If not familiar with the concept of load lines, please refer to [Sedra87].

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. The gain of the circuit is determined by the transconductances of the transistors and their output resistances. The latter are large in the transition region as both devices are in saturation. This translates into the VTC of Figure 3.17.

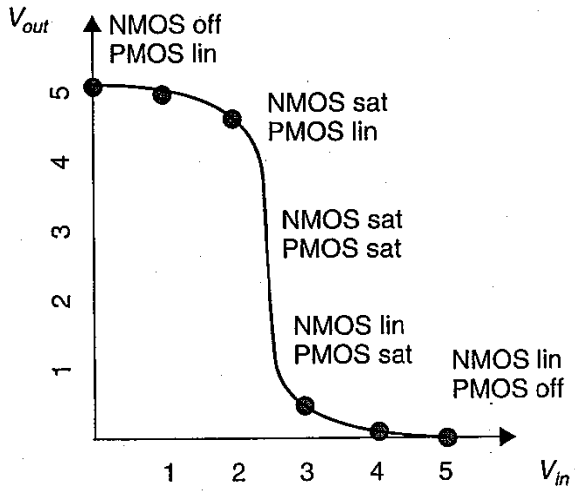


Figure 3.17 VTC of static CMOS inverter, derived from Figure 3.16 (V_{DD} = 5 V). For each operation region, the modes of the transistors are annotated.

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. This response is dominated mainly by the output capacitance of the gate, C_L, which is composed of the diffusion capacitances of the NMOS and PMOS transistors, the capacitance of the interconnect wires, and the input capacitance of the fan-out. Assuming temporarily that the transistors switch instantaneously, the transient response can be approximated again using the simplified switch model (Figure 3.18). Let us consider the low-to-high transition (Figure 3.18a). The gate response time is determined by the time it takes to charge the capacitor C_L through the resistor R_{on} (i.e., by the time constant $C_L R_{on}$). A fast gate is built either

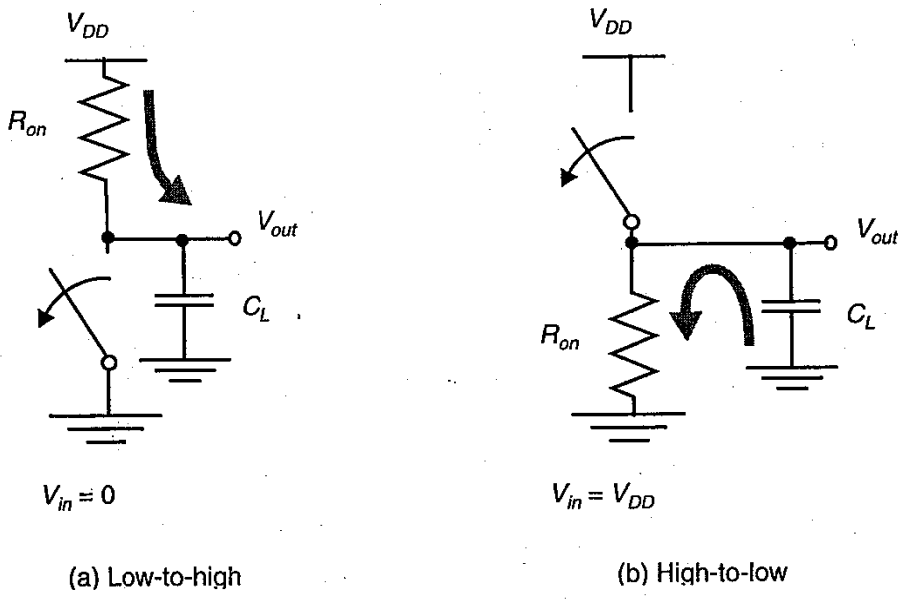


Figure 3.18 Switch model of dynamic behavior of static CMOS inverter.

by keeping the output capacitance small or by decreasing the on-resistance of the PMOS transistor. The latter is achieved by increasing the *W/L* ratio of the device, as is apparent from an inspection of the current equations of an MOS device. A similar consideration is true for the high-to-low transition Figure 3.18(b). One 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.

3.3.2 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. By definition, the values of V_{IH} and V_{IL} are the points at which

$$\frac{\partial V_{out}}{\partial V_{in}} = -1 .$$

In the terminology of the analog circuit designer, these are the points where the small-signal gain g of the amplifier, formed by the inverter, is equal to -1.

SIDELINE: Surprisingly (or not so surprisingly), an inverter can also be used as an analog amplifier. The static CMOS inverter/amplifier actually has a very high gain in its transition region (although the rest of its amplifier properties such as supply noise rejection are rather poor). 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 device in the regions of extreme nonlinearity, resulting in well-defined and well-separated high and low signals.

The small-signal model of the MOS transistor and expressions for its parameters were introduced in Appendix B. By inserting the model for the NMOS and PMOS transistor and by shorting all dc voltage sources, the small-signal model of the inverter is obtained as shown in Figure 3.19. The gain of this circuit is determined by inspection of the circuit and is set to -1 for $V_{in} = V_{IH}$ and V_{IL}

$$g = \frac{v_{out}}{v_{in}} = -(g_{mn} + g_{mp}) \times (r_{on} \parallel r_{op}) = -1 \qquad (3.9)$$

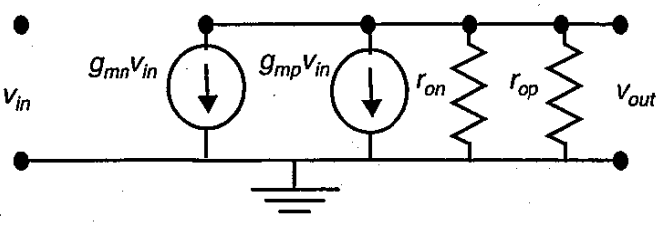


Figure 3.19 Small-signal model of a CMOS inverter.

The parameters of the model depend upon the operating modes of the transistors. Consider for instance V_{IH}. For $V_{in} = V_{IH}$, the PMOS and NMOS transistors can be assumed to be in the *saturation* and *linear* regions, respectively (Figure 3.17). Using the expressions presented in Table B.1, this results in the following values for the small-signal parameters. To simplify the analysis, the channel-length modulation is ignored, or $\lambda_p = 0$.

$$\begin{aligned} g_{mn} &= k_n V_{out} \\ g_{mp} &= k_p(V_{DD} - V_{IH} - |V_{Tp}|) \\ r_{on} &= \frac{1}{k_n(V_{IH} - V_{out} - V_{Tn})} \\ r_{op} &= \infty \end{aligned} \tag{3.10}$$

Inserting those expressions into the gain formula Eq. (3.9) results in a first relation between V_{IH} and V_{out}. A second relation is obtained by noting that the static currents through the NMOS and PMOS transistors must be identical.

$$g = -(g_{mn} + g_{mp}) \times (r_{on} \| r_{op}) = -\frac{k_n V_{out} + k_p(V_{DD} - V_{IH} - |V_{Tp}|)}{k_n(V_{IH} - V_{out} - V_{Tn})} = -1 \tag{3.11}$$

and

$$k_n\left[(V_{IH} - V_{Tn})V_{out} - \frac{V_{out}^2}{2}\right] = \frac{k_p}{2}(V_{DD} - V_{IH} - |V_{Tp}|)^2 \tag{3.12}$$

Solving V_{out} from Eq. (3.11) and substituting the result into Eq. (3.12) produces a second-order equation with one root between 0 and V_{DD}. An analytic expression for V_{IH} is complex and, therefore, not included.

Similar equations can be derived for the $V_{in} = V_{IL}$. In this case however, the NMOS operates in *saturation*, while the PMOS device is in the *linear* mode.

$$\begin{aligned} g_{mn} &= k_n(V_{IL} - V_{Tn}) \\ g_{mp} &= k_p(V_{DD} - V_{out}) \\ r_{on} &= \infty \\ r_{op} &= \frac{1}{k_p(v_{out} - V_{IL} - |V_{Tp}|)} \end{aligned} \tag{3.13}$$

so that

$$g = -(g_{mn} + g_{mp}) \times (r_{on} \| r_{op}) = -\frac{k_n(V_{IL} - V_{Tn}) + k_p(V_{DD} - V_{out})}{k_p(V_{out} - V_{IL} - |V_{Tp}|)} = -1 \tag{3.14}$$

and

$$k_p\left[(V_{DD} - V_{IL} - |V_{Tp}|)(V_{DD} - V_{out}) - \frac{(V_{DD} - V_{out})^2}{2}\right] = \frac{k_n}{2}(V_{IL} - V_{Tn})^2 \tag{3.15}$$

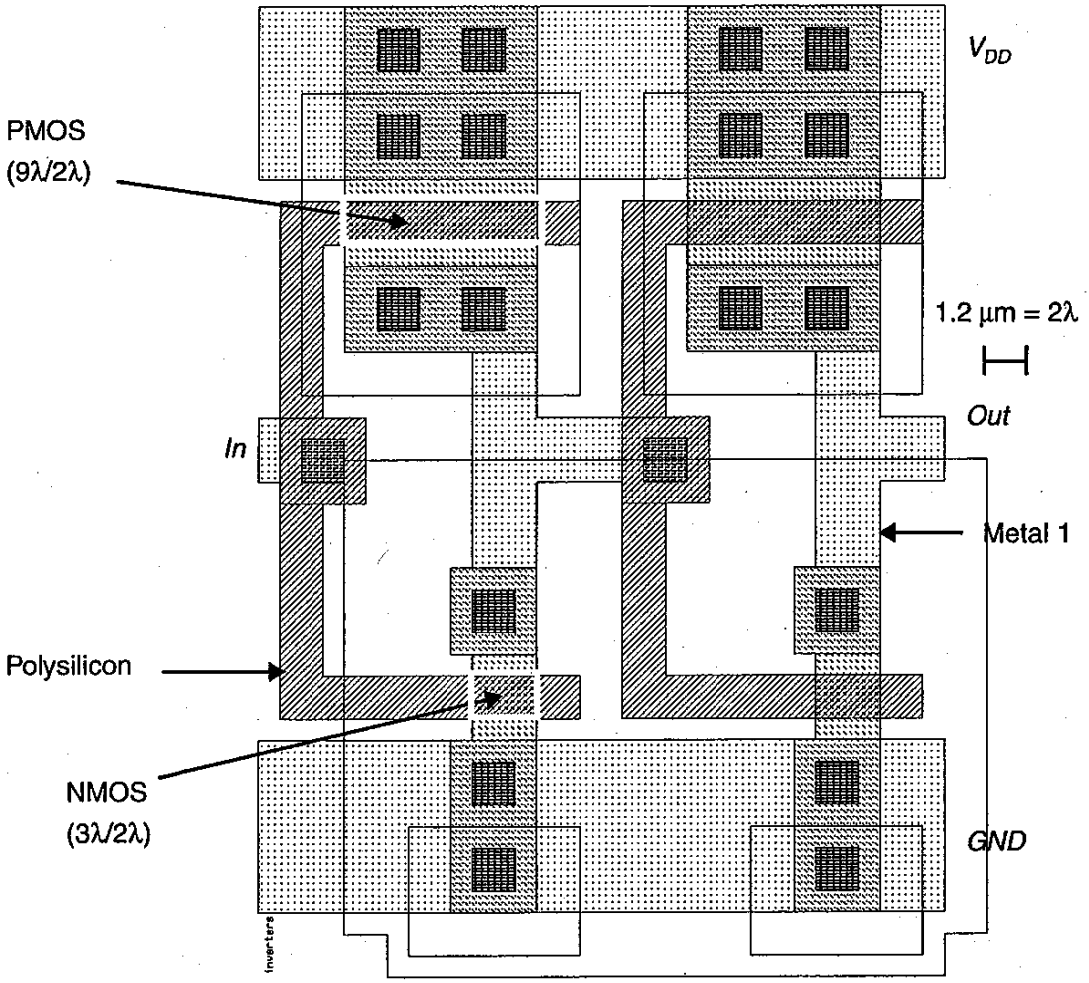


Figure 3.25 Layout of two chained, minimum-size inverters (1.2 μm CMOS technology) (see also Color-plate 6).

Table 3.1 Inverter transistor data.

	W/L	AD (μm²)	PD (μm)	AS (μm²)	PS (μm)
NMOS	1.8/1.2	6.84 (19 λ²)	9.0 (15λ)	6.8 (19 λ²)	9.0 (15λ)
PMOS	5.4/1.2	3*5.4 = 16.2(45 λ²)	3*2 + 5.4 = 11.4 (19λ)	16.2 (45 λ²)	11.4 (19λ)

This physical information can be combined with the approximations derived above to come up with an estimation of C_L. From the SPICE model, the following capacitor parameters are obtained:

Overlap capacitance: CGD0(NMOS) = 0.43 fF/μm; CGDO(PMOS) = 0.43 fF/μm

Bottom junction capacitance: CJ(NMOS) = 0.3 fF/μm²; CJ(PMOS) = 0.5 fF/μm²

Side-wall junction capacitance: CJSW(NMOS) = 0.8 fF/μm; CJSW(PMOS) = 0.135 fF/μm

Gate capacitance: C_{ox}(NMOS) = C_{ox}(PMOS) = ε_{ox}/t_{ox} = 1.76 fF/μm²

Bringing all the components together results in Table 3.2. Notice that the load capacitance is almost evenly split between its two major components: the internal capacitance (diffusion and overlap capacitances) and the gate capacitance of the connecting gate. Technically speaking, we should use different K_{eq} values for the bottom-plate and side-wall capacitances (as the latter is of the linear-gradient type). To simplify the derivation, we consider both of them to be abrupt and use the values of Example 3.5.

Table 3.2 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 W_n	1.55	1.55
C_{gd2}	2 CGD0 W_p	4.65	4.65
C_{db1}	K_{eqn} (AD_n CJ + PD_n CJSW)	3.45	5.6
C_{db2}	K_{eqp} (AD_p CJ + PD_p CJSW)	5.9	3.6
C_{g3}	C_{ox} W_n L_n	3.8	3.8
C_{g4}	C_{ox} W_p L_p	11.4	11.4
C_w	From Extraction	2	2
C_L	Σ	32.75	32.6

Propagation Delay: First-Order Analysis

The propagation delay can be computed by integrating the capacitor (dis)charge current, which results in Eq. (3.20).

$$t_p = C_L \int_{v_1}^{v_2} \frac{dv}{i(v)} \tag{3.20}$$

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 complex, as $i(v)$ is a nonlinear function of v. A reasonable approximation of the propagation delay, adequate for manual analysis, can be obtained by replacing the time-varying charging current by a fixed current I_{av} which is the average of the currents at the end points of the voltage transition. This simplification transforms Eq. (3.20) into a more tractable expression.

$$t_p = \frac{C_L(v_2 - v_1)}{I_{av}} \tag{3.21}$$

Remember that the propagation delay is defined as the time it takes for the output to reach the 50% point. For the low-to-high transition, $v_1 = V_{OL}$ and $v_2 = (V_{OH}+V_{OL})/2$. For the high-to-low transition, $v_1 = V_{OH}$, and $v_2 = (V_{OH} + V_{OL})/2$. As a result, the following holds for both t_{pLH} and t_{pHL}

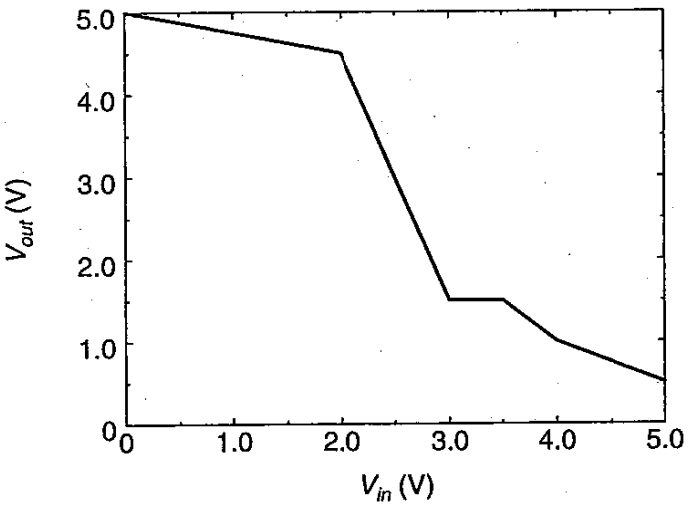


Figure 3.61 Voltage transfer characteristic of fictitious inverter.

c. Is the gate regenerative? Determine the input and output resistance for high and low inputs (assume $V_{in,low} = -3$ V and $V_{in,high} = 3$ V).

d. Determine the power consumption of the gate for $V_{in} = 0$ V and $V_{in} = 3$ V. Assume that all internal and external capacitances can be ignored.

e. Derive the VTC for a fan-out of 1 identical gate.

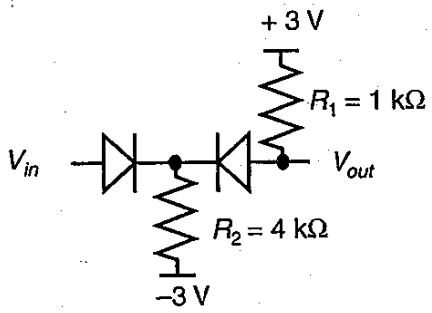


Figure 3.62 A diode-based digital gate.

3. [E, None, 3.3.2] The gate of Figure 3.63a uses a fictitious device as load. The device is characterized by the *I-V* curve of Figure 3.58b. Determine V_{OH} and V_{OL} of the gate. (Hint: use a graphical solution approach). Assume $V_T = 1$ V, $V_{DD} = 5$ V, $\lambda_n = 0$ V^{-1}, $k_n = 0.2286$ mA/V^2 for the NMOS transistor.

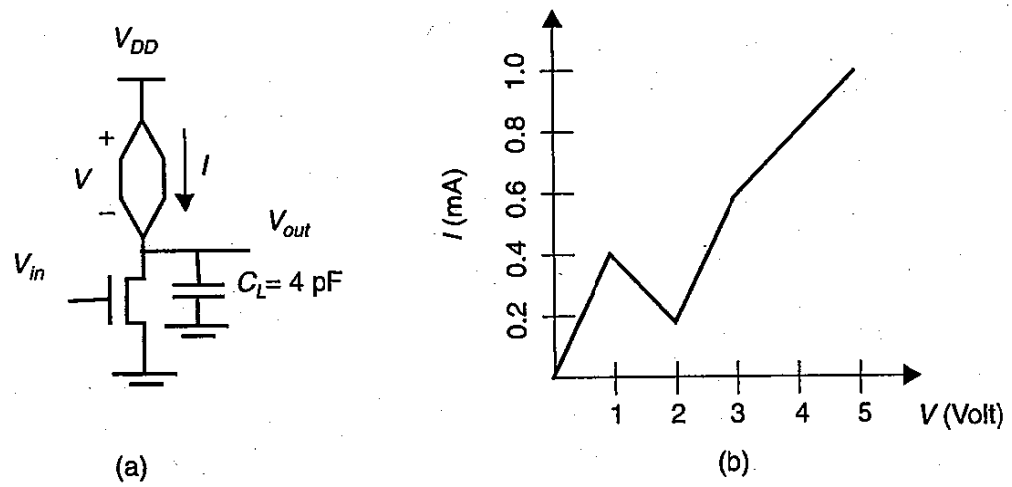


Figure 3.63 An inverter with a fictitious load, whose *I-V* characteristic is shown in (b).

4. [M, SPICE, 3.3.2] The layout of a static CMOS inverter is given in Figure 3.64. ($1\lambda = 0.6\ \mu m$).
   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?
   d. Compare your results with the VTC obtained by SPICE (using the LEVEL-2 model provided in Chapter 2).

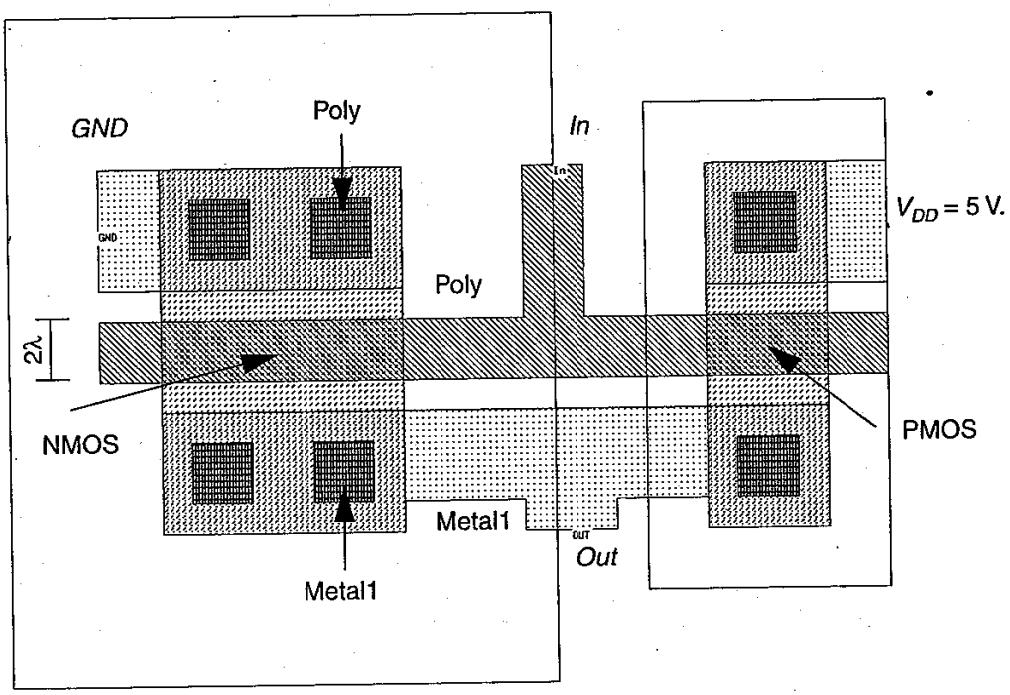


Figure 3.64 CMOS inverter layout.

5. [E, None, 3.3.2] Redesign the inverter of Figure 3.64 to achieve a switching threshold of approximately 1.5 V (only transistor sizes are needed). How are the noise margins affected by this modification?
6. The text defines the V_{IH} and V_{IL} points of the VTC as the points where the gain equals -1. This leads to complex expressions. A simpler approach is to use a piecewise linear approximation for the VTC, as shown in Figure 3.65. The transient region is approximated by a straight line, whose gain equals the gain in the mid-point V_M. The crossover with the V_{OH} and the V_{OL} lines is used to define V_{IH} and V_{IL} points.
   a. Using this approach, derive expressions for the noise margins of the CMOS inverter. Assume that $k_n = k_p$, $V_M = V_{DD}/2$, and $V_{Tn} = |V_{Tp}|$. To solve this problem, you have to take into account the effect of channel-length modulation. This can be accounted for by using the following expression for the small-signal resistance in the saturation region: $r_{o,sat} = 1/(\lambda I_D)$ (instead of the value of ∞, used in the text).
   b. Compare the obtained results with the values derived using the gain $=-1$ approach. For the comparison, assume $V_{DD} = 5$V, $V_{Tn} = |V_{Tp}| = 0.75$ V, $k_n = k_p = 40\ \mu A/V^2$, $\lambda_n = 0.06$, $\lambda_p = 0.2$.
7. [M, SPICE, 3.3.2] The noise margins of a CMOS inverter are highly dependent on the sizing ratio, $r = k_p/k_n$, of the NMOS and PMOS transistors. Use SPICE with $V_{Tn} = |V_{Tp}|$ to answer the following:

b. Determine t_{pLH}. You may assume that all internal capacitances can be ignored and that C_L is the only capacitance of interest. Assume the following values: $V_{swing} = 1$ V, $V_{in(low)} = 1$ V, $V_{in(high)} = 2$ V, $R_E = 800\ \Omega$, $R_C = 1$ kΩ

c. Determine t_{pHL}. Use the parameter values of part *b*.

d. Determine the dynamic energy dissipation during a low-to-high transition at the output.

e. Will any of the two bipolar transistors ever saturate (for V_{in} between 0 and 3 V)? Explain your answer.

DESIGN PROBLEM

Using the 1.2 μm CMOS introduced in Chapter 2, design a static CMOS inverter that meets the following requirements:

1. Matched pull-up and pull-down times (i.e., $t_{pHL} = t_{pLH}$).
2. $t_p = 5$ nsec (± 0.1 nsec).

The load capacitance connected to the output is equal to 4 pF. Notice that this capacitance is substantially larger than the internal capacitances of the gate.

Determine the W and L of the transistors. To reduce the parasitics, use minimal lengths ($L = 1.2$ μm) for all transistors. Verify and optimize the design using SPICE after proposing a first design using manual computations. Compute also the energy consumed per transition. If you have a layout editor (such as MAGIC) available, perform the physical design, extract the real circuit parameters, and compare the simulated results with the ones obtained earlier.