INTEGRATED CIRCUIT SIGNAL DELAY

Technologies for designing and building microelectronics-based computational equipment have been steadily advancing ever since the first commercial discrete circuits (ICs) were introduced in the late 1950s (1) (monolithic integrated circuits were introduced in the 1960s). As predicted by *Moore's law* in the 1960s (2), integrated-circuit density has been doubling approximately every 18 months, and this doubling in size has been accompanied by a similar exponential increase in circuit speed (or more precisely, clock frequency). These trends of steadily increasing circuit size and clock frequency are illustrated in Figs. 1(a) and 1(b), respectively. As a result of this amazing revolution in semiconductor technology, it is not unusual for modern integrated circuits to contain over 10 million switching elements (i.e., transistors) packed into a chip area as large as 500 mm^2 (3–5). This truly exceptional technologi-

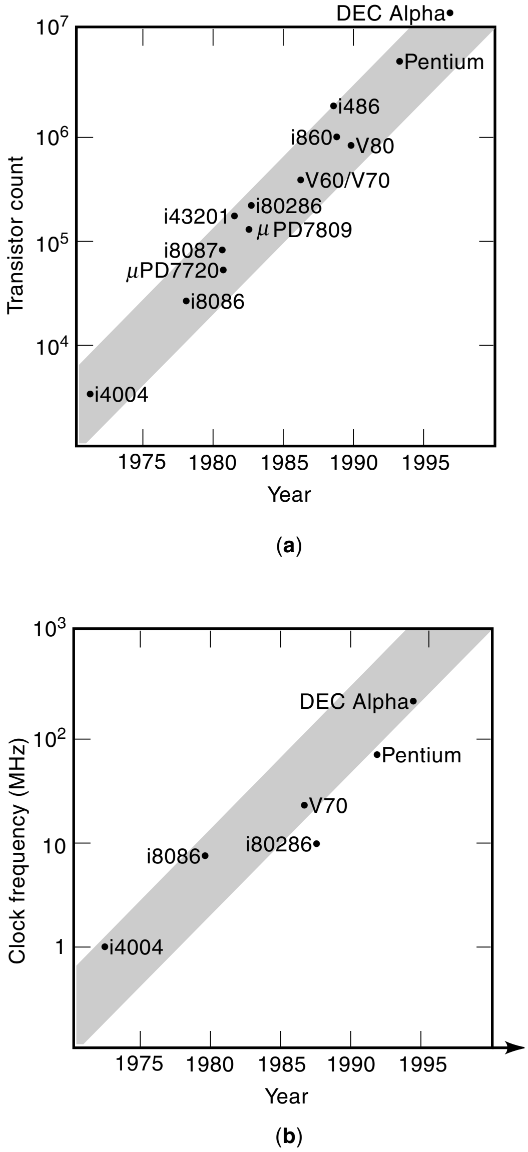

Figure 1. Moore's law—exponential increase in circuit integration and clock frequency (2). (a) Evolution of the transistor count per integrated circuit. (b) Evolution of clock frequency.

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 1999 John Wiley & Sons, Inc.

cal capability is due to advances in both design methodologies and physical manufacturing technologies. Research and experience demonstrate that this trend of exponentially increasing integrated-circuit computational power will continue into the foreseeable future.

Integrated-circuit performance is usually characterized (6) by the *speed* of operation, the available circuit *functionality,* and the *power* consumption, and there are multiple factors that directly affect these performance characteristics. While each of these factors is significant, on the technological side, increased circuit performance has been largely achieved by the following approaches:

- Reduction in feature size (technology scaling), that is, the capability of manufacturing physically smaller and faster circuit structures
- Increase in chip area, permitting a larger number of circuits and therefore greater on-chip functionality
- Advances in packaging technology, permitting the increasing volume of data traffic between an integrated circuit and its environment as well as the efficient removal of heat created during circuit operation

The most complex integrated circuits are referred to as VLSI circuits, where VLSI stands for very large scale integration. This term describes the complexity of modern integrated circuits consisting of hundreds of thousands to many millions of active transistor elements. Currently, the leading integrated-circuit manufacturers have a technological capability for the mass production of VLSI circuits with feature sizes as small as 0.12 μm (7). These sub–0.5 μm technologies are identified with the term *deep submicrometer* (DSM) since the minimum feature size is well below the 1 μm mark.

As these dramatic advances in fabricating technologies take place, integrated-circuit performance is often limited by effects closely related to the very reasons behind these advances such as small geometry interconnect structures. Circuit performance has become strongly dependent and limited by electrical issues that are particularly significant in DSM integrated circuits. *Signal delay* and related *waveform effects* are among those phenomena that have a great impact on high-performance integrated-circuit design methodologies and the resulting system implementation. In the case of fully synchronous VLSI systems, these effects have the potential to create catastrophic failures due to the limited time available for signal propagation among gates.

BACKGROUND TO SIGNAL DELAY

Data processing in the most widely available types of digital integrated circuits [complementary metal-oxide semiconductor (CMOS), bipolar junction transistor, bipolar CMOS (BiCMOS), and GaAs] is based on the transport of electrical energy from one location to another location. Typically, the information that is being processed is encoded as a physical variable that can be easily stored and transmitted to other locations while functionally manipulated along the way. Such a physical variable—also called a *signal*—is, for example, the electrical *voltage* provided by a power supply (with respect to a ground potential) and developed in circuit elements in the presence of electromagnetic fields. The voltage signal or *bit* of information (in a digital circuit) is temporarily stored in a circuit structure capable of accumulating *electric charge.* This accumulating or *storage* property is called *capacitance*—denoted by the symbol C—and, depending on the materials and the physical conditions, is created by a variety of different types of *conductor–insulator–conductor* structures commonly found in integrated circuits.

Furthermore, modern digital circuits utilize *binary logic,* in which information is encoded by two values of a signal. These logic values are typically called *false* and *true* (or *low* and *high* or *logic zero* and *logic one*) and correspond to the *minimum* and *maximum* allowable values of the signal voltage for a specific integrated-circuit implementation. Since the voltage V is proportional to the stored electric charge q ($q = CV$, where C is the storage capacitance), the logic low value corresponds to a fully *discharged* capacitance ($q = CV = 0$) while the logic high value corresponds to a capacitance storing the maximum possible charge (fully *charged* to a voltage V).

The largest and most complicated digital integrated circuits today contain many millions of circuit elements each processing these binary signals (2,6,8,9). Every element has a specific number of input terminals through which it receives data from other elements and a specific number of output terminals through which it makes the results of the processing available to other elements. For the circuit to implement a particular function, the inputs and outputs of each element must be properly connected among each other. These connections are accomplished with wires, which are collectively referred to as an *interconnect network,* while the set of binary state circuit elements is often simply called the *logic gates.* During normal circuit operation, the logic gates receive signals at their inputs, process the signals to generate new data, and transmit the resulting data signals to the corresponding logic elements through a network of interconnections. This process involves the transport of a voltage signal from one physical location to another physical location. In each case, this process takes a small yet *finite* amount of time to be completed and is often called the *propagation delay* of the signal.

The rate of data processing in a digital integrated circuit is directly related to two factors: how fast the circuit can switch between the two logic values and how precisely a circuit element can interpret a specific signal value as the intended binary logic state. Switching the state of a circuit between two logic values requires either *charging* a fully discharged capacitance or *discharging* a fully charged capacitance, depending upon the type of state transition—low to high or a high to low. This charging/discharging process is controlled by the active switching elements in the logic gates and is strongly affected by the physical properties of both the gates and the interconnections. Specifically, the signal waveform shapes change, either enhancing or degrading the signals, affecting both the ability and the time required for the logic gates to properly recognize these signals.

The concept of signal *propagation delay* between two different points A and B of a circuit is illustrated in Fig. 2. The signals at points A and B—denoted s_A and s_B, respectively—are plotted versus time for two different cases in Figs. 2(a) and 2(b), respectively. Without considering the specific electronic devices and circuits required to create these waveform shapes, it is assumed that signal s_A makes a transition from high to low and triggers a computation that causes signal s_B

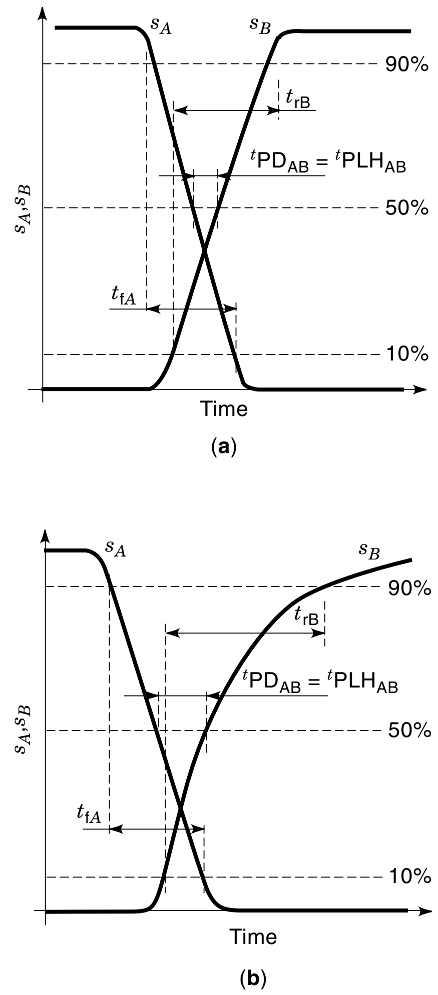

Figure 2. Signal propagation delay from point A to point B. (a) Linear ramp input and ramp output. (b) Linear ramp input and exponential output.

to make an opposite transition from low to high. Several important observations can be made from Fig. 2:

- Although s_A is the same in each case, s_B may have different shapes.
- A temporal relationship (or *causality* relationship) between s_A and s_B exists in the sense that s_A "causes" s_B, thereby preceding the switching event by an amount of time required for the physical switching process to propagate through the circuit structure.
- Regardless of shape, s_B has the same logical meaning, namely, that the state of the circuit at point B changes from low to high; this transition from low to high and the reverse state transition (signal s_A from high to low) require a positive amount of time to complete.

The temporal relationship between s_A and s_B as shown in Fig. 2 must be evaluated *quantitatively.* This information permits the speed of the signals at different points in the same circuit or in different circuits built in different semiconductor technologies to be temporally characterized. By quantifying the physical speed of the logical operations, circuit designers are provided with necessary information to design correctly functioning integrated circuits.

The design of a digital VLSI system may require a great deal of effort in order to consider a broad range of architectural and logic issues, that is, choosing the appropriate gates and interconnections among these gates to achieve the required circuit function. No design is complete, however, without considering the *dynamic* (or transient) characteristics of the signal propagation or, alternatively, the changing behavior of signals *within time.* Every computation performed by a switching circuit involves multiple signal transitions between logic states and requires a finite amount of time to complete. The voltage at every circuit node must reach a specific value for the computation to be completed. Therefore, state-of-the-art integrated-circuit design is largely centered around the difficult task of predicting and properly interpreting signal waveform shapes at various points in a circuit.

In a typical VLSI system, millions of signal transitions occur, such as those shown in Fig. 2, which determine the individual gate delays and the overall speed of the system. Some of these signal transitions can be executed *concurrently* while others must be executed in a strict *sequential* order (8). The sequential occurrence of the latter operations—or signal transition *events*—must be properly coordinated in time so that logically correct system operation is guaranteed and its results are reliable (in the sense that these results can be repeated). This coordination is known as *synchronization* and is critical to ensuring that any pair of logical operations in a circuit with a precedence relationship proceed in the proper order. In modern digital integrated circuits, synchronization is achieved at all stages of system design and system operation by a variety of techniques, known as a *timing discipline* or *timing scheme* (9–12). With few exceptions, these circuits are based on a *fully synchronous* timing scheme, specifically developed to cope with the finite speed required by the physical signals to propagate through the system.

An example of a fully synchronous system is shown in Fig. 3(a). As illustrated in Fig. 3(a), there are three easily recognizable components in this system. The first component—the logic gates, collectively referred to as the *combinational logic*—provides the range of operations that a system executes. The second component—the clocked storage elements or simply the *registers*—are elements that store the results of the logical operations. Together, the combinational logic and registers constitute the computational portion of the synchronous system and are interconnected in a way that implements the required system function. The third component of the synchronous system—known as the *clock distribution network*—is a highly specialized circuit structure that does not perform a computational process but rather provides an important control capability. The clock generation and distribution network controls the overall synchronization of the circuit by generating a time reference and properly distributing this time reference to every register.

The normal operation of a system, such as the example shown in Fig. 3(a), consists of the iterative execution of computations in the combinational logic followed by the storage of the processed results in the registers. The actual process of storing is temporally controlled by the clock signal and occurs once the signal transients in the logic gate outputs are completed and the outputs have settled to a valid state. At the beginning of each computational cycle, the inputs of the system together with the data stored in the registers initiate a new switching process. As time proceeds, the signals propagate through the logic, generating results at the logic output.

By the end of the clock period, these results are stored in the registers and are operated upon during the following clock cycle.

Therefore, the operation of a digital system can be thought of as the sequential execution of a large set of simple computations that occur concurrently in the combinational logic portion of the system. The concept of a *local data path* is a useful abstraction for each of these simple operations and is shown in Fig. 3(b). The magnitude of the delay of the combinational logic is bound by the requirement of storing data in the registers within a clock period. The initial register R_i is the storage element at the beginning of the local data path and provides some or all of the input signals for the combinational logic at the beginning of the computational cycle (defined by the beginning of the clock period). The combinational path ends with the data successfully latching within the final register R_f in which the results are stored at the end of the computational cycle. Each register acts as a source or sink for the data depending on the current phase of the system operation.

The behavior of a fully synchronous system is well defined and controllable as long as the time window provided by the clock period is sufficiently long to allow every signal in the circuit to propagate through the required logic gates and interconnect wires and successfully latch within the final register. In designing the system and choosing the proper clock period, however, two contradictory requirements must be satisfied. First, the smaller the clock period, the more computational cycles can be performed by the circuit in a given amount of time. At the same time, the time window defined by the clock period must be sufficiently long so that the slowest signals reach the destination registers before the current clock cycle is concluded and the following clock cycle is initiated.

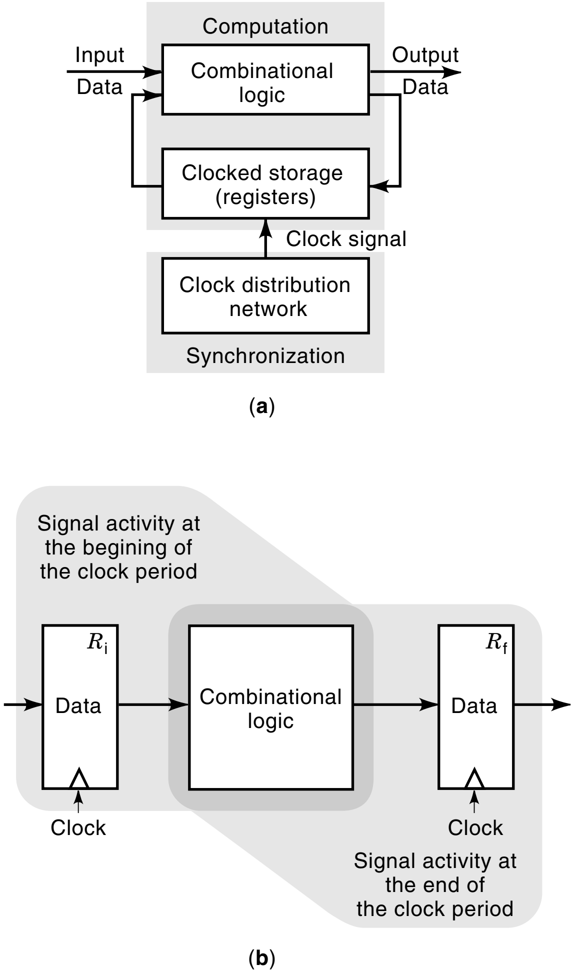
Computation
Input
Data
Combinational logic
Output
Data
Clocked storage (registers)
Clock signal
Clock distribution network
Synchronization
(a)
Signal activity at the begining of the clock period
R_i
Data
Combinational logic
R_f
Data
Clock
Clock
Signal activity at the end of the clock period
(b)

Figure 3. A synchronous system. (a) Finite-state machine model of a synchronous system. (b) A local data path.

This way of organizing computation has certain clear advantages that have made a fully synchronous timing scheme the primary choice for digital VLSI systems:

- It is easy to understand and its properties and variations are well understood.
- It eliminates the nondeterministic behavior of the propagation delay in the combinational logic (due to environmental and process fluctuations and the unknown input signal pattern) so that the system as a whole has a completely deterministic behavior corresponding to the implemented algorithm.
- The circuit design does *not* need to be concerned with glitches in the combinational logic outputs so the only relevant dynamic characteristic of the logic is the *propagation delay*.
- The state of the system is completely defined within the storage elements—this fact greatly simplifies certain aspects of the design, debug, and test phases in developing a large system.

However, the synchronous paradigm also has certain limitations that make the design of synchronous VLSI systems increasingly challenging:

- This synchronous approach has a serious drawback in that it requires the overall circuit to operate as slow as the *slowest* register-to-register path. Thus, the global speed of a fully synchronous system depends upon those paths in the combinational logic with the largest delays—these paths are also known as the *worst case* or *critical* paths. In a typical VLSI system, the propagation delays in the combinational paths are distributed unevenly so there may be many paths with delays much smaller than the clock period. Although these paths could take advantage of a lower clock period—higher clock frequency—it is the paths with the largest delays that bound the clock period, thereby imposing a limit on the overall system speed. This imbalance in propagation delays is sometimes so dramatic that the system speed is dictated by only a handful of very slow paths.
- The clock signal has to be distributed to tens of thousands of storage registers scattered throughout the system. Therefore, a significant portion of the system area and dissipated power is devoted to the clock distribution network—a circuit structure that does not perform any computational function.
- The reliable operation of the system depends upon the assumptions concerning the values of the propagation delays, which, if not satisfied, can lead to catastrophic timing violations and render the system unusable.

DELAY METRICS

The delay for a signal to propagate from one point within a circuit to another point is caused by both active electronic devices (transistors) in the logic elements and the various passive interconnect structures connecting the logic gates. While the physical principles behind the operation of transistors and interconnect are well understood at the current–voltage level, it is often computationally difficult to apply this detailed information to the densely packed multimillion transistor DSM integrated circuits of today.

A general form of a circuit with N input ($x_1, \ldots, x_N$) and M output ($y_1, \ldots, y_M$) terminals is shown in Fig. 4(a). The box labeled "Circuit" may represent a simple wire, a transistor, or a logic gate consisting of several transistors, or an arbitrarily complex combination of these elements. If the box shown in Fig. 4(a) corresponds to the portion of the logic circuit schematically outlined in Fig. 4(b), a logically possible signal activity at the circuit points X, Y, and Z is shown in Fig. 5(a). The dynamic characteristics of the signal transitions as well as their relationship in time are described and formalized in Definitions 1 to 3.

Definition 1. If X and Y are two points in a circuit and s_X and s_Y are the signals at X and Y, respectively, the signal propagation delay $t_{\text{PD}_{XY}}$ from X to Y is defined as the time interval from the 50% point of the signal transition of s_X to the 50% point of the signal transition of s_Y. (Although the delay can be defined from any point X to any other point Y, X and Y typically correspond to an input and an output of a logic gate, respectively. In such a case, the signal delay from X to Y is the propagation delay of the gate.)

This formal definition of the propagation delay is related to the concept that ideally the switching point of a logic gate is at the 50% level of the output waveform. Thus, 50% of the maximum output signal level is assumed to be the boundary point where the state of the gate switches from one binary logic state to the other binary logic state. Practically, a more physically correct definition of propagation delay is the time from the switching point of the driving circuit to the switching point of the driven circuit. Currently, however, this switching-point-based reference for signal delay is not widely used in practical computer-aided design applications because of the computational complexity of the algorithms and the increased amount of data required to estimate the delay of a path. Therefore, choosing the switching point at 50% has become a generally acceptable practice for referencing the propagation delay of a switching element.

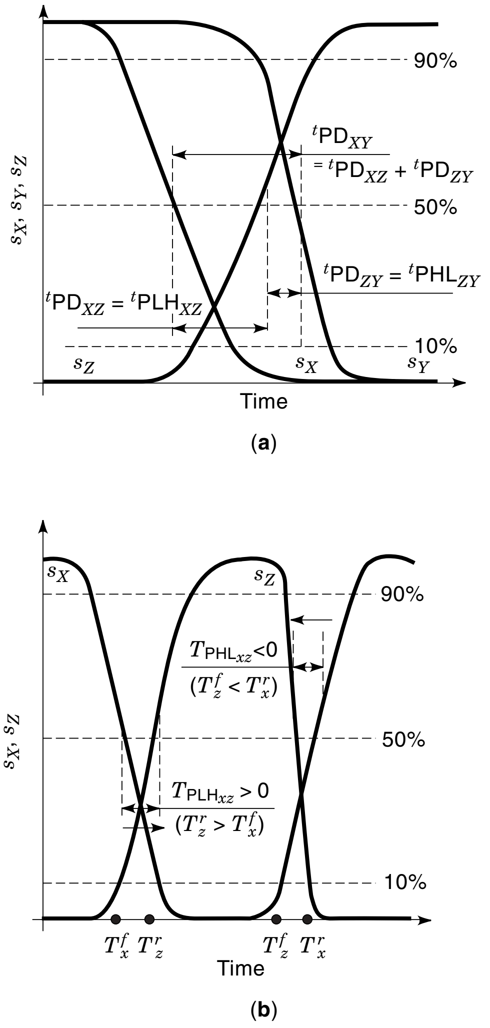

Figure 5. Switching characteristics of the circuit shown in Fig. 4(a) Signal waveforms for the circuit shown in Fig. 4(b). (b) Signal waveforms for the inverter in the circuit shown in Fig. 4(b).

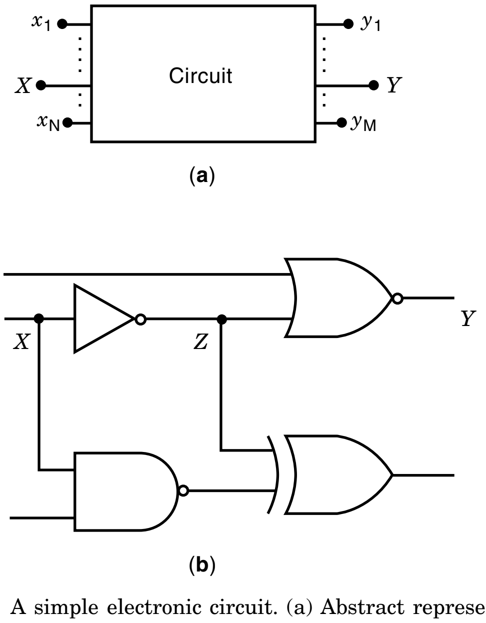

Figure 4. A simple electronic circuit. (a) Abstract representation of a circuit. (b) Logic schematic of the circuit in panel (a).

Also, note that the propagation delay t_{PD} as defined in Definition 1 is mathematically additive, thereby permitting the delay between any two points X and Y to be determined by summing the delays through the consecutive structures between X and Y. From Figs. 4(b) and 5(a), for example, $t_{\text{PD}_{XY}} = t_{\text{PD}_{XZ}} + t_{\text{PD}_{ZY}}$. However, this additivity property must be applied with caution since neither of the switching points of consecutively connected gates may occur at the 50% level. In addition, passive interconnect structures along signal paths do not exhibit switching properties although physical signals propagate through these structures with finite speed. Therefore, if the properties of a signal propagating through a series connection of logic gates and interconnections are under investigation, an analysis of the entire signal path composed of gates and wires—rather than adding 50%-to-50% delays—is necessary to avoid accumulating error.

In high-performance CMOS VLSI circuits, logic gates often switch before the input signal completes its transition. (Also, a gate may have asymmetric signal paths, whereby a gate would switch faster in one direction than in the other direction.) This difference in switching speed may be sufficiently large so that an output signal of a gate will reach its 50% point before the input signal reaches the 50% point. If this is the case, t_{PD} as defined by Definition 1 may have a negative value. Consider, for example, the inverter connected between nodes X (inverter input) and Z (inverter output) in Fig. 4(b). The specific input and output waveforms for this inverter are shown in detail in Fig. 5(b). When the input signal s_X makes a transition from high to low, the output signal s_Z makes a transition from low to high (and vice versa). In this specific example, the low-to-high transition of the signal s_Z crosses the 50% signal level after the high-to-low transition of the signal s_X. Therefore, the signal delay t_{PLH} (the signal name index is omitted for clarity) is positive as shown by the direction of the arrow in Fig. 5(b), coinciding with the positive direction of the x axis. However, when the input signal s_X makes a low-to-high transition, the output signal s_Z makes a faster high-to-low transition and crosses the 50% signal level before the input signal s_X crosses the 50% signal level. The signal delay t_{PHL} in this case is negative as shown by the direction of the arrow in Fig. 5(b), coinciding with the negative direction of the x axis.

As illustrated in Fig. 5(b), the asymmetry of the switching characteristics of a logic gate requires the ability to discriminate between the values of the propagation delay in the two different switching situations (a low-to-high or a high-to-low transition). One single value of the propagation delay t_{PD}, as defined in Definition 1, does not provide sufficient information about this possible asymmetry in the switching characteristics of a logic gate. Therefore, the concept of delay is extended further to include this missing information. Specifically, the direction of the output waveform (since the output of a gate is typically the evaluation node) is included in the definition of delay, thereby permitting the evaluation of the gate switching speed to account for the effects of the output signal transition:

Definition 2. The signal propagation delays $t_{\mathrm{PLH}_{XY}}$ and $t_{\mathrm{PHL}_{XY}}$, respectively, denote the signal delay from input X to output Y (as defined in Definition 1), where the output signal (at point Y) transitions from low to high and from high to low, respectively (the low-to-high and high-to-low transitions).

It is important to consider both t_{PLH} and t_{PHL} during circuit analysis and design. However, if only a single value of t_{PD} is specified, t_{PD} usually refers to the arithmetic average, $(t_{\mathrm{PLH}} + t_{\mathrm{PHL}})/2$.

Furthermore, Definition 2 specifies the time between switching events, but does not convey any information about the transition time of the events themselves. This transition time is finite and is characterized by the two parameters described in the following definition:

Definition 3. For a signal making a transition between two different logic states, the transition time is defined as the time interval between the 10% point and the 90% point of the signal. For a low-to-high transition, the rise transition time $t_{\mathrm{r}} = t|_{90\%} - t|_{10\%}$. For a high-to-low transition, the fall transition time $t_{\mathrm{f}} = t|_{10\%} - t|_{90\%}$.

The parameters defined in Definition 3 are illustrated in Fig. 2, where the fall time t_{f_A} and the rise time t_{r_B} for the signals s_A and s_B, respectively, are indicated.

As t_{r} and t_{f} are related to the slope of the signal transitions, the transition times also affect the values of t_{PLH} and t_{PHL}, respectively. In Fig. 5(a), for example, note that if the signal s_Y had been slower—a larger fall time t_{f_Y}—s_Y would have crossed the 50% level at a later time, effectively increasing the propagation delay $t_{\mathrm{PLH}_{XY}}$. However, as illustrated in Fig. 2, it is possible that the 50%-to-50% delay remains nearly the same, although the signal slope may change significantly [the rise time t_{r_B} in Figs. 2(a) and 2(b)].

DEVICES AND INTERCONNECTIONS

The technology of choice for most modern high-performance digital integrated circuits is based on the metal-oxide-semiconductor field-effect transistor (MOSFET) structure. The primary reasons for the wide application of MOSFETs are, among other things, high packing density and, in its complementary form, low power dissipation. In this section, the properties of both active devices and interconnections are discussed from the perspective of circuit performance.

An n-channel *enhancement* mode MOSFET transistor (NMOS) is depicted in Fig. 6(a). Note that in most digital applications, the substrate is connected to the source, i.e., $V_{\mathrm{s}} = V_{\mathrm{b}}$ and $V_{\mathrm{sb}} = 0$. Therefore, the four-terminal transistor depicted in Fig. 6(a) can be considered as a three-terminal device with the voltages V_{s}, V_{g}, and V_{d} controlling the operation of the transistor. Assuming no substrate current, $I_{\mathrm{dd}} = I_{\mathrm{ss}}$—both currents I_{dd} and I_{ss} are usually referred to by I_{ds} only. In the following discussion, the additional indices n and p are used to indicate which type of transistor is being considered, n-channel or p-channel, respectively.

To first order, the drain current I_{dsn} through the transistor is modeled by the classical Shichman–Hodges equations (13):

$$I_{\mathrm{ds}n} = \begin{cases} \beta_n[(V_{\mathrm{gs}n} - V_{\mathrm{t}n})V_{\mathrm{ds}n} - \frac{1}{2}V_{\mathrm{ds}n}^2], & \\ \qquad V_{\mathrm{gs}n} \ge V_{\mathrm{t}n} \quad \text{and} \quad V_{\mathrm{gd}n} \ge V_{\mathrm{t}n} & \text{(linear mode)} \\ \beta_n \frac{1}{2}(V_{\mathrm{gs}n} - V_{\mathrm{t}n})^2, & \\ \qquad V_{\mathrm{gs}n} \ge V_{\mathrm{t}n} \quad \text{and} \quad V_{\mathrm{gd}n} \le V_{\mathrm{t}n} & \text{(saturation mode)} \\ 0, \quad V_{\mathrm{gs}n} \le V_{\mathrm{t}n} & \text{(cutoff mode)} \end{cases} \tag{1}$$

The derivation of PMOS I-V equations is straightforward by accounting for the changes in voltages and current directions. In Eq. (1), the parameter β_n is a device parameter commonly called a *gain factor* or the *current gain* of the transistor—the dimension of β_n is [A/V^2]. The value of the current gain β_n is

$$\beta_n = K_n \frac{W_n}{L_n} \tag{2}$$

where K_n is the *process transconductance parameter* and W_n and L_n are the width and length of the transistor channel, respectively. The process transconductance K_n is found as the

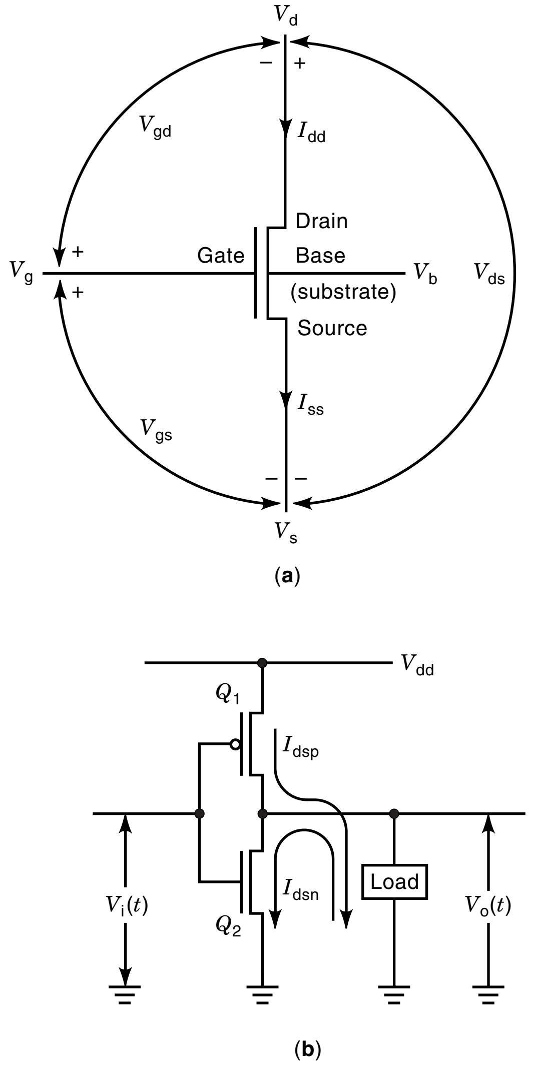

Figure 6. (a) An NMOS transistor and (b) the basic CMOS inverter gate.

product

$$K_n = \mu_n C_{ox} = \mu_n \frac{\epsilon_{ox}}{t_{ox}} \tag{3}$$

where μ_n is the *carrier mobility* and C_{ox} is the gate capacitance per unit area (ϵ_{ox} is the relative dielectric constant of the gate oxide material—3.9 for SiO_2—and t_{ox} is the gate oxide thickness). By substituting the index p for the index n in Eqs. (1) to (3), analogous expressions for β_p and K_p of a p-channel *enhancement* mode MOSFET transistor can be developed (2,6,9,14). Also note that the threshold voltage V_{tn} of an n-channel transistor is positive ($V_{tn} > 0$), while the threshold voltage V_{tp} of a p-channel transistor is negative ($V_{tp} < 0$).

Equation (1) and its counterpart for a p-channel MOS device are fundamental to both *static* and *dynamic* circuit analysis. Static, or dc, analysis refers to circuit bias conditions in which the voltages V_g, V_d, and V_s remain constant. Dynamic analysis is attractive from the signal delay perspective since it deals with voltage and current waveforms that change over time. An important goal of dynamic analysis is to determine the timing relationships connecting the transistor terminal voltages as these voltages are the signal representations of the data being processed. By performing a dynamic analysis, the signal delay from an input waveform to its corresponding output waveform can be evaluated with a certain level of accuracy.

Complementary MOS logic or CMOS logic is the most popular circuit style for most modern high-performance digital integrated circuits. An analytical analysis of a simple CMOS logic gate is presented next for one of the simplest CMOS gates—the CMOS inverter shown in Fig. 6(b). Performing such a simple analysis illustrates the process for estimating circuit performance as well as provides insight into *what* factors and *how* these factors may affect the timing characteristics of a logic gate.

Analytical Delay Analysis

Consider the CMOS inverter circuit consisting of the PMOS device Q_1 and NMOS device Q_2 shown in Fig. 6(b). For this analysis, assume that the capacitive load of the inverter—consisting of any device capacitances, interconnect capacitances, and the load capacitance of the following stage—can be lumped into a single capacitor C_L. The output voltage $V_o = V_{C_L}$ is the voltage across the capacitive load and the terminal voltages of the transistors are shown in Table 1. The regions of operation for the devices Q_1 and Q_2 are illustrated in Fig. 7 depending upon the values of V_i and V_o. Referring to Fig. 7 may be helpful in understanding the switching process in a CMOS inverter. Determining the values of the fall time t_f and the propagation delay t_{PHL} is described in the following section. Similarly, closed-form expressions for the rise time t_r and the propagation delay t_{PLH} are derived later.

The Value of t_f and t_{PHL}. The transition process used to derive t_f and t_{PHL} is illustrated in Fig. 8(a). Assume that the input signal V_i has been held at logic low ($V_i = 0$) for a sufficiently long time such that the capacitor C_L is fully charged to the value of V_{dd}. The operating point of the inverter is point ***A*** on Fig. 7. At time $t_0 = 0$ the input signal abruptly switches to a logic high. The capacitor C_L *cannot* discharge instantaneously, thereby forcing the operating point of the circuit to point ***B***—(V_i, V_o) = (V_{dd}, V_{dd}). At ***B***, the device Q_1 is cut off while Q_2 is conducting, thereby permitting C_L to begin discharging through Q_2. As this discharge process develops, the operating point moves down the line ***BD*** approaching point ***D***, where C_L is fully discharged, that is, $V_o(\boldsymbol{D}) = 0$. Observe that during the interval $0 \leq t < t_2$ the operating point is between ***B*** and ***C*** and the device Q_2 operates in saturation. At time t_2, the capacitor is discharged to $V_{dd} - V_{tn}$ and Q_2 begins to operate in the linear region. For $t \geq t_2$, the device Q_2 is in

Table 1. Terminal Voltages for the *p*-channel and *n*-channel Transistor in a CMOS Inverter Circuit

	p-Channel	n-Channel
V_g	$V_{gp} = V_{in}$	$V_{gn} = V_{in}$
V_s	$V_{sp} = V_{DD}$	$V_{sn} = 0$
V_d	$V_{dp} = V_{out}$	$V_{dn} = V_{out}$
V_{gs}	$V_{gsp} = V_{in} - V_{DD}$	$V_{gsn} = V_{in}$
V_{gd}	$V_{gdp} = V_{in} - V_{out}$	$V_{gdn} = V_{in} - V_{out}$
V_{ds}	$V_{dsp} = V_{out} - V_{DD}$	$V_{dsn} = V_{out}$

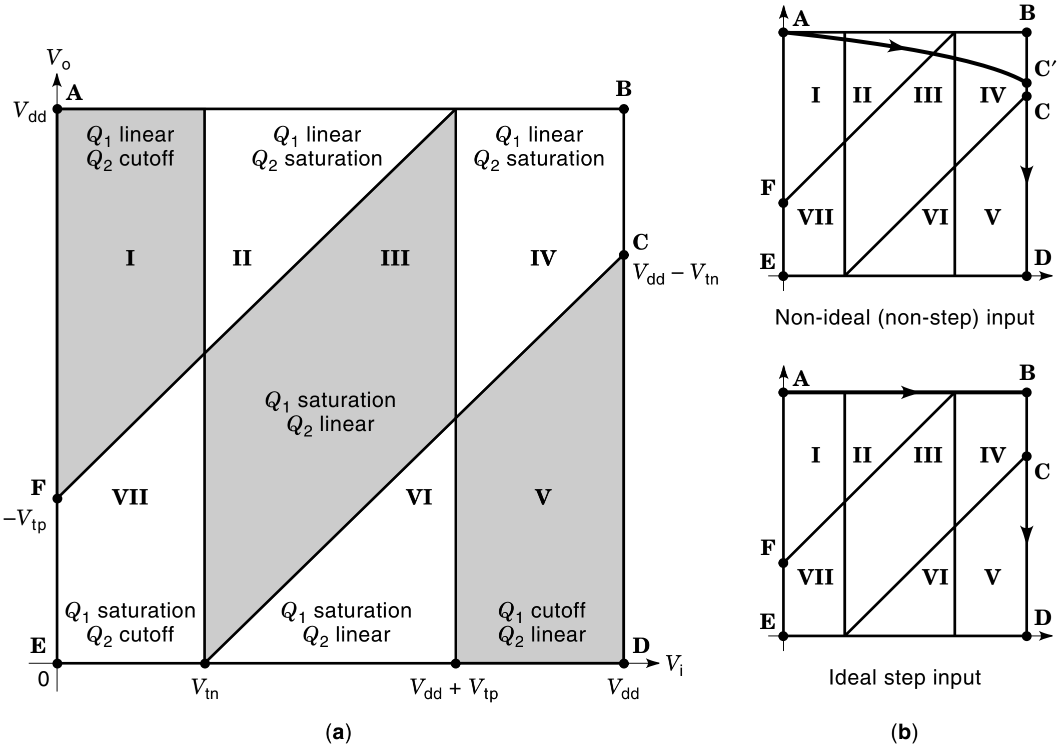

Figure 7. Modes of operation for the devices in the CMOS inverter. (a) Operating modes depending on the input voltage, V_i, and the output voltage, V_o. (b) Operating point trajectory for different input waveforms (only rising input shown).

the linear region. If (as is typical) $0.1V_{dd} < V_{tn} < 0.5V_{dd}$, then $t_1 < t_2 < t_3$ as shown in Fig. 8(a). Therefore, the fall time is $t_f = t_4 - t_1$ and the propagation time $t_{PHL} = t_3 - 0 = t_3$. To determine the values of t_f and t_{PHL}, the output waveform $V_o(t)$ must be evaluated for each of the intervals $[t_0, t_2)$ and $[t_2, \infty)$.

For $t_0 \leq t < t_2$, the current discharging the capacitor I_{dsn}, shown in Fig. 6(b), is

$$I_{dsn} = \beta_n \frac{1}{2}[(V_{dd} - V_{tn})^2] = -C_L \frac{dV_o}{dt} \tag{4}$$

Substituting

$$\eta = \frac{V_{tn}}{V_{dd}} \quad \text{and} \quad \gamma_n = \frac{\beta_n V_{dd}(1-\eta)}{C_L} \tag{5}$$

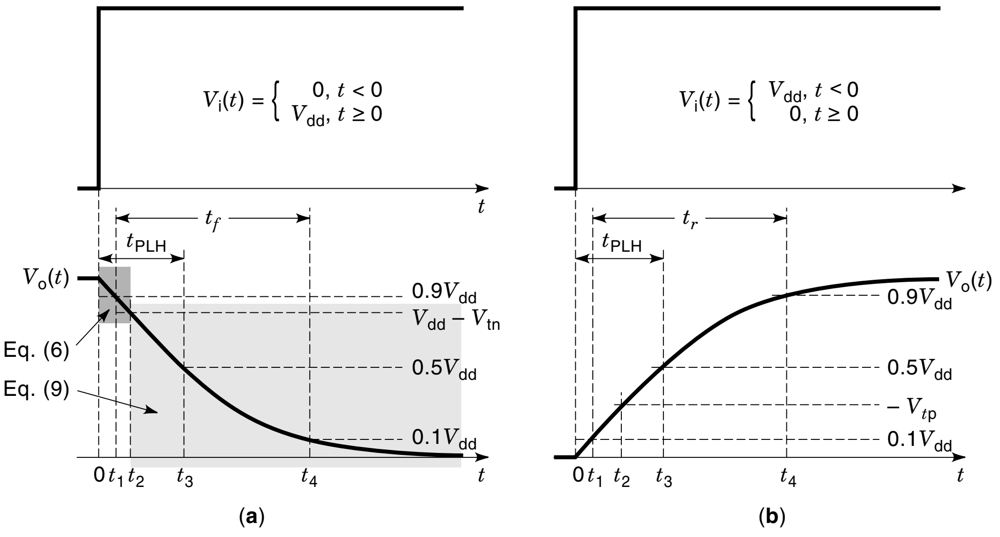

Figure 8. Switching waveforms for a step input at the CMOS inverter in Fig. 6(b). (a) High-to-low output transition. (b) Low-to-high output transition.

and solving Eq. (4) for V_o with the initial condition $V_o(0) = V_{dd}$ yields

$$\begin{aligned} V_o(t) &= V_{dd} - \left[\frac{\beta_n}{2C_L}(V_{dd}-V_{tn})^2\right]t \\ &= V_{dd}\left[1-\frac{\gamma_n}{2}(1-\eta)t\right] \quad \text{for} \quad t_0 \le t < t_2 \end{aligned} \tag{6}$$

From Eq. (6) it can be further shown that

$$V_o(t_2) = V_{dd} - V_{tn} \quad \text{for} \quad t_2 = \frac{2C_L}{\beta_n(V_{dd}-V_{tn})^2}V_{tn} = \frac{2\eta}{\gamma_n(1-\eta)} \tag{7}$$

The interval $t \geq t_2$ is considered next. The device Q_2 is in linear mode and I_{dsn} is given by

$$I_{dsn} = \beta_n\left[(V_{dd}-V_{tn})V_o - \frac{1}{2}V_o^2\right] = -C_L\frac{dV_o}{dt} \tag{8}$$

A closed-form expression for the output voltage $V_o(t)$ for time $t \geq t_2$ is obtained by solving Eq. (8) (a Bernoulli equation) with the initial condition $V_o(t_2) = V_{dd} - V_{tn}$:

$$V_o(t) = V_{dd}\frac{2(1-\eta)}{1+e^{\gamma_n(t-t_2)}} \quad \text{for} \quad t \ge t_2 \tag{9}$$

The values of t_1 from Eq. (6) and t_3 and t_4 from Eq. (9) are

$$\begin{aligned} t_1 &= \frac{1}{\gamma_n}\frac{0.2}{1-\eta}, \quad t_3 = \frac{1}{\gamma_n}\left[\frac{2\eta}{1-\eta} + \ln(3-4\eta)\right], \\ t_4 &= \frac{1}{\gamma_n}\left[\frac{2\eta}{1-\eta} + \ln(19-20\eta)\right] \end{aligned} \tag{10}$$

The fall time t_f is

$$t_f = t_4 - t_1 = \frac{C_L}{\beta_n}\frac{1}{V_{dd}(1-\eta)}\left(2\frac{\eta-0.1}{1-\eta} + \ln(19-20\eta)\right) \tag{11}$$

and the propagation delay t_{PHL} is

$$t_{PHL} = t_3 - 0 = t_3 = \frac{C_L}{\beta_n}\frac{1}{V_{dd}(1-\eta)}\left(\frac{2\eta}{1-\eta} + \ln(3-4\eta)\right) \tag{12}$$

The Value of t_r and t_{PLH}. The rise time t_r and the propagation delay t_{PLH} are determined from the switching process illustrated in Fig. 8(b) (similarly to t_f and t_{PHL} earlier). Assume that the input signal V_i has been held at logic high ($V_i = V_{dd}$) for a sufficiently long time such that the capacitor C_L is fully discharged to $V_o = 0$. The operating point of the inverter is point $\boldsymbol{D}$ shown in Fig. 7. At time $t_0 = 0$, the input signal abruptly switches to a logic low. Since the voltage on C_L *cannot* change instantaneously, the operating point is forced at point $\boldsymbol{E}$. At $\boldsymbol{E}$, the device Q_2 is cut off while Q_1 is conducting, thereby permitting C_L to begin charging through Q_1. As this charging process develops, the operating point moves up the line $\boldsymbol{EA}$ toward point $\boldsymbol{A}$ at which C_L is fully charged, that is, $V_o(\boldsymbol{A}) = V_{dd}$. Note that during the interval $0 \leq t < t_2$, the operating point is between $\boldsymbol{E}$ and $\boldsymbol{F}$ and the device Q_1 operates in the saturation region. At time t_2, the capacitor is charged to $-V_{tp}$ (recall that $V_{tp} < 0$) and Q_1 starts operating in the linear region. For $t \geq t_2$, the device Q_1 is in the linear region. If (as is typical) $0.1V_{dd} < |V_{tp}| < 0.5V_{dd}$, then $t_1 < t_2 < t_3$ as shown in Fig. 8(b). Therefore, the rise time is $t_r = t_4 - t_1$ and the propagation delay is $t_{PLH} = t_3 - 0 = t_3$. To determine the values of t_r and t_{PLH}, the output waveform $V_o(t)$ must be evaluated for each of the intervals $[t_0, t_2)$ and $[t_2, \infty)$.

An analysis similar to that described earlier can be performed to derive expressions for t_1, t_3, and t_4 in Fig. 8(b). Substituting

$$\pi = -\frac{V_{tp}}{V_{dd}}, \qquad \gamma_p = \frac{\beta_p V_{dd}(1-\pi)}{C_L} \tag{13}$$

t_1, t_3, and t_4 are

$$\begin{aligned} t_1 &= \frac{1}{\gamma_p}\frac{0.2}{1-\pi}, \qquad t_3 = \frac{1}{\gamma_p}\left[\frac{2\pi}{1-\pi} + \ln(3-4\pi)\right], \\ t_4 &= \frac{1}{\gamma_p}\left[\frac{2\pi}{1-\pi} + \ln(19-20\pi)\right] \end{aligned} \tag{14}$$

Therefore, the value of the rise time t_r is

$$t_r = t_4 - t_1 = \frac{C_L}{\beta_p}\frac{1}{V_{dd}(1-\pi)}\left(2\frac{\pi-0.1}{1-\pi} + \ln(19-20\pi)\right) \tag{15}$$

and the value of the propagation delay t_{PLH} is

$$t_{PLH} = t_3 - 0 = t_3 = \frac{C_L}{\beta_p}\frac{1}{V_{dd}(1-\pi)}\left(\frac{2\pi}{1-\pi} + \ln(3-4\pi)\right) \tag{16}$$

Several observations can be made by analyzing the expressions derived earlier for t_r, t_f, t_{PHL}, and t_{PLH}. First, the factors that affect the inverter delays are analyzed. Following this analysis, the related waveform effects are considered and *short-channel* effects of submicrometer devices are then described.

Controlling the Delay

Note that in Eqs. (11) and (15), the fall and rise times, respectively, are the product of a term of the form C_L/β and another *process-dependent* term (a function solely of V_{dd} and V_t). These relationships imply that for a given manufacturing process, improvements in individual gate delays are possible by reducing the load impedance C_L or by increasing the current gain of the transistors. Reducing the load impedance is possible by controlling physical aspects of the design (the specific gate layout). Alternatively, increasing β of the devices (recall that $\beta \propto W/L$) is typically accomplished by controlling the value of W—a process known as transistor or gate sizing. (Typically, the device channel length is chosen to be the minimum permitted by the technology and therefore cannot be decreased to further increase β.) Transistor sizing, however, has limits—area requirements may limit the maximum channel width W, and increasing W will also increase the input load capacitance of the gates.

Waveform Effects

The ideal step input waveform used to derive the delay expressions presented earlier is a physical abstraction. Such an ideal waveform does not exist naturally, although it is used to simplify the analysis presented before. Note that despite ideally fast input waveforms, the output signal of a CMOS logic gate has a finite slope, thereby contributing to a certain gate delay. In a practical VLSI integrated circuit, both the input and output signals have a nonzero rise and fall time due to the impedances along any signal path. Fast input waveforms can be effectively considered as step inputs, and the delay expressions derived in Eqs. (11) and (15) model the delays for such cases with reasonable accuracy. Slow input waveforms, however, contribute significantly to the delay of the charge–discharge path in a gate output (6,9,14,15), making the aforementioned delay expression inaccurate.

Furthermore, it is considerably more difficult to derive closed-form delay expressions for nonstep input waveforms. Consider, for example, the derivation of the fall time of the inverter shown in Fig. 6(b) assuming a nonideal input, such as the linear ramp signal s_A in Fig. 2(a). Referring to Fig. 7(b), the trajectory of the operating point relating V_i and V_o for a nonideal (nonstep) input is as shown in the upper diagram. This trajectory is a curve passing through regions **I, II, III,** and **IV** (through regions **I, II, III, IV,** and **V** for slower input signals), and down the line $\boldsymbol{C'} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{D}$ rather than the two straight-line segments $\boldsymbol{A} \rightarrow \boldsymbol{B}$ and $\boldsymbol{B} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{D}$ (as shown in the lower diagram). Therefore, calculating an exact expression for t_f in this case would require separately evaluating the delay for all five portions of the output V_o—one for each region.

Analysis of the CMOS inverter shown in Fig. 6(b) with other than an ideal step input, as well as the respective delay expressions, can be found in Ref. 15. Consider, for example, the linear ramp input described by

$$V_i(t) = \begin{cases} 0, & t < 0 \\ \dfrac{t}{t_{r_i}} V_{dd}, & 0 \le t < t_{r_i} \\ V_{dd}, & t \ge t_{r_i} \end{cases} \tag{17}$$

where t_{r_i} is the rise time of the input voltage signal $V_i(t)$. In the case depicted in the upper diagram shown in Fig. 7(b), the total propagation delay $t_{PHLramp}$ at the 50% level (15) is given by

$$t_{PHLramp} = \tfrac{1}{6}(1 + 2\eta)t_{r_i} + t_{PHLstep} \tag{18}$$

where $t_{PHLstep}$ is the propagation delay time for a step input given by Eq. (12).

Note that the ramp input described by Eq. (17) is also an idealization intended to simplify analysis. In a practical integrated circuit, the input waveform to the inverter is not a linear ramp, but rather the output waveform of another gate within the circuit. For such an input—also known as a *characteristic* input—it is practical to regard the propagation delay through the inverter gate shown in Fig. 6(b) as a function of the C_L/β ratio of the preceding gate or, equivalently, as a function of the step response delay of the preceding stage (15). This kind of direct analytical solution—by breaking the output waveform in regions depending upon the trajectory of the operating point—becomes even more complicated for a gate with more than one input arriving at an arbitrary time and with arbitrary waveforms. Because of the growing complexity of such an analytical solution, it is imperative that alternative methods for the delay calculation be developed and used in practice.

Nonideal input waveforms also have implication on the power dissipation of individual logic gates and therefore of the entire circuit. Observe that in regions II, IV, and VI, shown in Fig. 7, both devices simultaneously conduct, creating a temporary direct path for the current from V_{dd} to ground. The *short-circuit* current in this direct current path is only slightly related to the output voltage of the gate and adds to the total power dissipation. This added component is known as *short-circuit* power. The short-circuit power can be a substantial fraction of the total power dissipation of a circuit and can become an obstacle to meeting a specific design goal. Faster waveforms throughout the circuit generally mean less time spent switching within regions II, IV, and VI and therefore decreased short-circuit current and short-circuit power.

Short-Channel Effects

The active device model used in the analyses described earlier, Eq. (1), is accurate for long-channel devices. As technology is scaled down into the submicrometer range, a variety of physical phenomena develop that requires improved device models in order to preserve accuracy. In this section, certain key effects, known as *short-channel* effects, are described, as related to the discussion of propagation delay.

Channel-Length Modulation. A MOSFET device modeled by Eq. (1) has an infinite output resistance in saturation and acts as a voltage-controlled current source. Recall the linear portion of the falling or rising output waveforms from the analysis presented earlier. The device acts as a current source because of the complete independence in saturation of the drain current I_{dsn} from the voltage V_{dsn} assumed in Eq. (1). This independence, however, is an idealization that does not take into account the effect of the voltage V_{dsn} on the shape of the channel. In practice, as V_{dsn} is increased beyond the value required for saturation (such that $V_{gdn} < V_{tn}$ or $V_{gdp} > V_{tp}$ for a PMOS device), the channel pinch-off point moves towards the source. Therefore, the effective channel length is reduced, an effect known as *channel-length modulation.*

To account for channel-length modulation analytically, the expression for the current in saturation in Eq. (1) is modified as follows:

$$I_{dsn} = \beta_n \tfrac{1}{2}(V_{gsn} - V_{tn})^2(1 + \lambda_n V_{dsn}) \tag{19}$$

The additional factor $(1 + \lambda_n V_{dsn})$ in Eq. (19) is the cause of a finite device output resistance $\partial V_{dsn}/\partial I_{dsn} = 2(V_{gsn} - V_{tn})^{-2}/(\lambda_n \beta_n)$ in saturation. The output waveform is degraded due to the degradation of the transfer characteristic of the inverter.

Velocity Saturation. In a long-channel transistor, the drift velocity of the carriers in the channel is proportional to both the carrier mobility and the lateral electric field in the channel (parallel to the source–drain path). In short-channel devices, however, the velocity of the carriers eventually saturates for some value of the voltage V_{ds} within the operating range of the circuit. This velocity saturation phenomenon is due to the fact that the power supply voltage is not scaled down as quickly as the device dimensions due to system constraints.

The saturation in carrier velocity for high electric field strengths—caused by the high voltage $V_{\rm ds}$ applied over a short channel—causes a reduction in both the process transconductance [see Eq. (3)] and the current gain of a saturated device. This reduction in the current gain β has a direct effect on the ability of the devices to drive a specific load, resulting in increased delay times. Recall that the propagation delays described in Eqs. (11), (12), (15), and (16) are inversely proportional to β.

A more realistic device model for DSM devices—known as the α-power model—has been proposed in Ref. 16 to include the carrier velocity saturation effect in submicrometer devices (short-channel devices in general):

$$I_{{\rm ds}n} = \begin{cases} I'_{{\rm D0}_n}, & V_{{\rm gs}n} \geq V_{{\rm t}n}, V_{{\rm ds}n} \geq V'_{{\rm D0}_n} \\ & \text{(pentode or saturation region)} \\ \dfrac{I'_{{\rm D0}_n}}{V'_{{\rm D0}_n}} V_{{\rm ds}n}, & V_{{\rm gs}n} \geq V_{{\rm t}n}, V_{{\rm ds}n} < V'_{{\rm D0}_n} \\ & \text{(triode or linear region)} \\ 0, & V_{{\rm gs}n} \leq V_{{\rm t}n} \ \text{(cutoff mode)} \end{cases} \qquad (20)$$

where

$$I'_{{\rm D0}_n} = I_{\rm D0}\left(\frac{V_{{\rm gs}n} - V_{{\rm t}n}}{V_{\rm dd} - V_{{\rm t}n}}\right)^{\alpha}, \qquad V'_{{\rm D0}_n} = V_{\rm D0}\left(\frac{V_{{\rm gs}n} - V_{{\rm t}n}}{V_{\rm dd} - V_{{\rm t}n}}\right)^{\alpha/2} \qquad (21)$$

In Eqs. (20) and (21), α is the velocity saturation index, $V_{\rm D0}$ is the drain saturation voltage for $V_{{\rm gs}n} = V_{\rm dd}$, and $I_{\rm D0}$ is the drain saturation current for $V_{{\rm gs}n} = V_{{\rm ds}n} = V_{\rm dd}$. A typical value for the velocity saturation index of short-channel devices is $1 \leq \alpha \leq 2$, where Eq. (20) becomes Eq. (1) for $\alpha = 2$.

Analytical solutions for the output voltage of a CMOS inverter with a purely capacitive load $C_{\rm L}$ for a step, linear ramp, and exponential input waveforms can be found in Ref. 17. Closed-form expressions for the delay of the CMOS inverter shown in Fig. 6(b) under the α-power model are given in Ref. 16 and are repeated here:

$$t_{\rm PHL} = t_{\rm PLH} = \left(\frac{1}{2} - \frac{1-\eta}{1+\alpha}\right)t_{\rm T} + \frac{C_{\rm L}V_{\rm dd}}{2I_{\rm D0}}, \ \eta = \frac{V_{{\rm t}n}}{V_{{\rm d}n}} \qquad (22)$$

The propagation delay described by Eq. (22) can be applied to nonideal input waveforms and consists of two terms. The first term reflects the effect on the gate delay of the input waveform shape and is proportional to the input waveform transition time $t_{\rm T}$. The second term reflects the dependency of the delay on the gate load, similarly to the $C_{\rm L}/\beta$ term in Eqs. (12) and (16).

The Importance of Interconnections

The analysis of the CMOS gate delay as described earlier is based on the assumption that the load of the inverter shown in Fig. 6(b) is a purely capacitive load (C). This assumption is generally true for logic gates placed close to each other in the physical layout of an integrated circuit. In a multimillion transistor VLSI circuit, however, certain connected logic gates may be relatively distant from each other. In this situation, the impedance of the interconnect wires cannot be considered as being purely capacitive, but rather as being *resistive–capacitive* (*RC*). An important type of global circuit structure where the gates can be very far apart is the clock distribution network (18).

The interconnect has become a major concern due to the high resistance that can limit overall circuit performance. These interconnect impedances have become significant as the minimum line dimensions have been scaled down into the deep-submicrometer region while the overall chip dimensions have increased. Perhaps the most important consequence of these trends of scaling transistor and interconnect dimensions and increasing chip sizes is that the primary source of signal propagation delay has shifted from the active transistors to the passive interconnect lines. Therefore, the nature of the load impedance has shifted from a lumped capacitance to a distributed resistance–capacitance, thereby requiring new qualitative and quantitative interpretations of the signal switching processes.

To illustrate the effects of scaling, consider ideal scaling (6) where devices are scaled down by a factor of S ($S > 1$) and chip sizes are scaled up by a factor of $S_{\rm c}$ ($S_{\rm c} > 1$). The delay of the logic gates decreases by $1/S$ while the delay due to the interconnect increases by $S^2S_{\rm c}^2$ (6,19). Therefore, the ratio of interconnect delay to gate delay after ideal scaling increases by a factor of $S^3S_{\rm c}^2$. For example, if $S = 4$ (corresponding to scaling down from a 2 μm CMOS technology to a 0.5 μm CMOS technology) and $S_{\rm c} = 1.225$ (corresponding to the chip area increasing by 50%), the ratio of interconnect delay to gate delay will increase by a factor of $4^3 \times 1.225 = 78.4$ times.

Delay Estimation in *RC* Interconnect. Interconnect delay can be analyzed by considering the CMOS inverter shown in Fig. 6(b) with the capacitive load $C_{\rm L}$ representing the accumulated capacitance of the fanout of the inverter. The interconnect connecting the drains of the devices Q_1 and Q_2 to the upper terminal of the load is replaced by a distributed RC line with a resistance and capacitance $R_{\rm int}$ and $C_{\rm int}$, respectively (19).

Closed-form expressions for the signal delay with an RC load are described by Wilnai (20). The delay values for both a *distributed* and *lumped RC* load are summarized in Table 2. These delay values are obtained assuming a step input signal.

The results listed in Table 2 are illustrated graphically in Fig. 9 (20). Two waveforms of the signal output response making a low-to-high transition are shown in Fig. 9. These two waveforms are based on the RC load being distributed and lumped, respectively.

Assuming an on-resistance $R_{\rm tr}$ of the driving transistor (19), the interconnect delay $T_{\rm int}$ can be characterized by the following expression:

Table 2. Closed-Form Expressions for the Signal Delay Response Driving a Distributed and Lumped *RC* Load—An Ideal Step Input Is Assumed

	Signal Delay	
Output Voltage Range	Distributed *RC*	Lumped *RC*
0% to 90%	1.0*RC*	2.3*RC*
10% to 90% (rise time $t_{\rm r}$)	0.9*RC*	2.2*RC*
0% to 63%	0.5*RC*	1.0*RC*
0% to 50% (delay $t_{\rm PLH}$)	0.4*RC*	0.7*RC*
0% to 10%	0.1*RC*	0.1*RC*

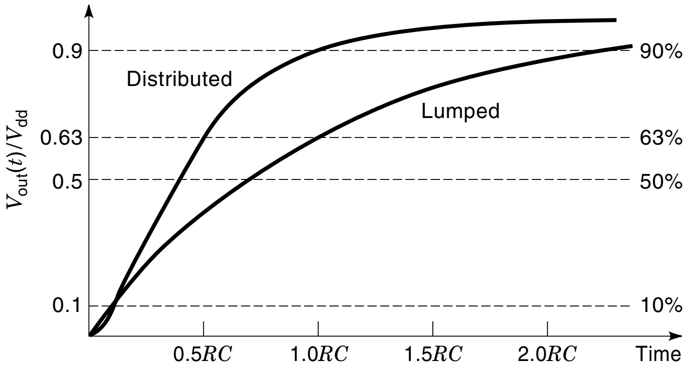

Figure 9. Illustration of the *RC* signal delay expressions in Table 2 (20). Waveforms are shown for both a distributed and lumped *RC* load.

$$T_{\text{int}} = R_{\text{int}}C_{\text{int}} + 2.3(R_{\text{tr}}C_{\text{int}} + R_{\text{tr}}C_{\text{L}} + R_{\text{int}}C_{\text{L}}) \tag{23}$$

$$\approx (2.3R_{\text{tr}} + R_{\text{int}})C_{\text{int}} \tag{24}$$

The on-resistance of the driving transistor R_{tr} in Eqs. (23) and (24) can be approximated (19) by

$$R_{\text{tr}} \approx \frac{1}{\beta V_{\text{dd}}} \tag{25}$$

where the term β in Eq. (25) is the current gain of the driving transistor [see Eq. (2)].

Approximating a distributed *RC* line by a combination of lumped resistances (*R*) and capacitances (*C*) is commonly used in circuit simulation programs. Three typical ladder circuits are illustrated in Fig. 10. The names of the ladder circuits shown in Fig. 10 are derived based on the similarities between the shape of the circuit and a known structure such as a letter. The *RC* interconnect is replaced in circuit simulation programs with circuit ladder structures such as those shown in Fig. 10. To increase the accuracy of simulation, more detailed Π_n and T_n ladder models can be used (21).

A lumped Π and T ladder circuit model better approximates a distributed *RC* model than a lumped L ladder circuit (21) by up to 30%. As described in Ref. 21, the strategy to model a distributed *RC* line depends upon two circuit parameters:

1. The ratio $C_{\text{T}} = C_{\text{L}}/C$ of the load capacitance C_{L} of the fanout to the capacitance C of the interconnect line
2. The ratio $R_{\text{T}} = R_{\text{tr}}/R$ of the output resistance of the driving MOSFET device R_{tr} to the resistance R of the interconnect line

The appropriate ladder circuit (from Ref. 21) to model a distributed *RC* interconnect line properly within 3% error as a function of R_{T} and C_{T} is shown in Table 3. By using the appropriate *RC* model (21), the computational time of the simulation can be more efficiently reduced while preserving the accuracy of the circuit simulation (22).

Table 3. Circuit Network to Model Distributed *RC* Line with a Maximum Error of 3%[a]

	R_{T}											
C_{T}	0	0.01	0.1	0.2	0.5	1	2	5	10	20	50	100
0	Π3	Π3	Π2	Π2	Π1	Π1	Π1	Π1	Π1	C	C	C
0.01	Π3	Π3	Π2	Π2	Π1	Π1	Π1	Π1	Π1	C	C	C
0.1	T2	T2	Π2	Π2	Π1	Π1	Π1	Π1	Π1	C	C	C
0.2	T2	T2	Π2	Π2	Π1	Π1	Π1	Π1	Π1	C	C	C
0.5	T1	T1	T1	T1	Π1	Π1	Π1	Π1	Π1	C	C	C
1	T1	T1	T1	T1	Π1	Π1	Π1	Π1	Π1	C	C	C
2	T1	T1	T1	T1	Π1	Π1	Π1	Π1	L1	L1	C	C
5	Π1	Π1	Π1	Π1	Π1	Π1	Π1	L1	L1	L1	C	C
10	Π1	Π1	Π1	Π1	Π1	Π1	L1	L1	L1	L1	C	C
20	R	R	R	R	R	R	L1	L1	L1	L1	C	C
50	R	R	R	R	R	R	R	R	R	R	C	N
100	R	R	R	R	R	R	R	R	R	R	N	N

[a] From Ref. 21. The notations Π, T, and L correspond to a Π, T, and L model, respectively. The notations R and C correspond to a single lumped resistance and capacitance, respectively. The notation N means that the interconnect impedance can be ignored. The number after certain models (e.g., Π3) correspond to a multiple model structure.

Delay Mitigation

A variety of different techniques have been developed to improve the signal delay characteristics depending upon the type of load and other circuit parameters. Among the most important techniques are as follows:

- Gate sizing to increase the output current drive capability of the transistors along the logic chain (23–25). Gate sizing must be applied with caution, however, because of the resulting increase in area and power dissipation.
- Tapered buffer circuit structures are often used to drive large capacitive loads (such as at the output pad of a chip) (8,26–31). A series of CMOS inverters such as the circuit shown in Fig. 6(b) are cascaded where the output drive of each buffer is increased by a constant tapering factor.
- The use of *repeater* circuit structures to drive resistive–capacitive (*RC*) loads. Unlike tapered buffers, repeaters are typically CMOS inverters of uniform size (drive capability) that are inserted at uniform intervals along an interconnect line (6,32–37).
- A different timing discipline such as *asynchronous timing* (2,8,38). Unlike fully synchronous circuits, the order of execution of logic operations in an asynchronous circuit

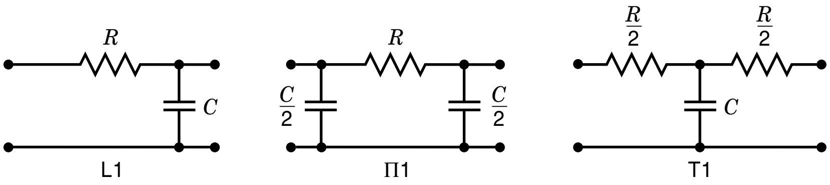

Figure 10. L, Π, and T ladder circuits to approximate an *RC* interconnect impedance.

is *not* controlled by a global clock signal. Therefore, asynchronous circuits are essentially independent of the signal delays. The logical order of operations in an asynchronous circuit is enforced by requiring the generation of special *handshaking* signals that communicate the status of the computation.

Among other useful techniques to improve the signal delay characteristics are the use of *dynamic* CMOS logic circuits including *Domino* logic (9) and differential circuit logic styles, such as *cascade voltage switch logic* or CVSL (9).

IMPACT OF DSM ON DESIGN METHODOLOGIES

The capability of applying and analyzing timing relationships and delays in deep-submicrometer integrated circuits requires a great amount of knowledge describing the physical phenomena of these circuits. As described earlier, the development of purely analytic equations is practically impossible to carry out even for very simple circuits. Furthermore, from a design perspective, it is important to be able to apply intuitive knowledge that incorporates circuit physics and operation when both analyzing existing circuits and synthesizing, or designing, new circuits based on their topological, functional, and timing characteristics.

The use of powerful computers coupled with efficient algorithms is absolutely fundamental to the successful analysis and synthesis of multi-million transistors integrated circuits. In fact, the majority of these algorithms are specifically developed with circuit complexity in mind and the related issues of accuracy, run time, and memory requirements. Therefore, CAD software tools play a vital role in the circuit design and manufacturing process. As noted earlier, however, improvements in technology and the demand for greater functionality and performance are changing the physical models of devices and interconnects in DSM circuits. A serious consequence of these changes is that the traditional design flow (the sequence of steps involved in the design and analysis of circuits) is no longer able to handle the required circuit complexity in an efficient manner.

In the traditional design flow, a great amount of effort and time is devoted to the architectural and logical aspects of the circuit. In this *front-end* portion of the design process, the circuit is partitioned into smaller subsystems and the individual logical networks. At the front end, the emphasis is on the behavioral, register transfer level (RTL), and logic levels of abstraction, concentrating on satisfying the functional design goals. Approximate timing information is used at the front end to estimate the delay of the logic gates and to determine the correct architectural (rather than physical) placement of the registers within a circuit.

Actual circuit and physical design are at the *back end* of the design process and consist of determining the circuit description of the specific physical transistor and interconnect patterns corresponding to the previously developed networks of logic gates. During this phase, the locations of the logic gates on the chip area are determined and wires are routed among the terminals of these gates as required by the logic network specifications. Besides being a time-consuming process targeted to satisfy these many geometrical and connectivity constraints, the physical design process must also preserve the dynamic specifications of the circuit assumed during the front-end design process. Alternatively, the gates must be placed and the wires routed among them to guarantee that the circuit will function correctly given the system input signals.

The primary difficulty with this approach is that the front-end methodologies and CAD tools largely ignore the details and problems of the physical domain. Such an approach cannot be tolerated in the design of DSM circuits for multiple reasons, among which the following are related to the signal delay and waveform shapes in an important way:

1. With advances in technology, transistor devices and gates become smaller and faster, while the size of the integrated circuit increases. These trends lead to the appearance of many global interconnect wires the length of which increases proportionally with increasing die size. Not only are the devices smaller but these transistors often have to drive relatively larger loads due to the long global interconnections.
2. As the average length of a wire increases, the electrical model of an interconnect wire changes from a purely capacitive (C) model to a resistive–capacitive (RC) model and finally to an inductive (RLC) model. The wire geometry also changes in order to satisfy performance, density, and yield objectives. Therefore, fringing capacitances between lines and cross-wire signal coupling begin to play an increasingly important role in signal integrity and circuit speed.
3. Multiple wire planes are often used, thereby increasing the complexity of the routing tools and making it significantly more difficult to account for any coupling and noise effects during analysis and synthesis portions of the design process.
4. Fast turnaround times and increased market pressure often require the reuse of large circuit subsystems (known as ASIC cores or megacells) surrounded by customized glue logic. (Application-specific integrated circuits—or ASIC—are specialized circuits developed to satisfy a specific manufacturer's need rather than be distributed as off-the-shelf parts.) The reusable portions and the glue logic are naturally separated from each other on the surface of the integrated circuit, requiring multiple long interconnect wires.

CAD software tool developers and circuit designers have become increasingly concerned with new approaches to the integrated-circuit design flow in order to cope with these aforementioned effects. A paradigm shift towards merging the capabilities of front-end and back-end tools is currently emerging as an alternative to the traditional methodology of separating these design efforts. Thus, in order to relieve constraints on the back-end tools and increase the likelihood of a successful design, the front-end tools must account for the lower-level DSM-related physical effects at a much earlier stage of the system design process.

An important approach in circuit extraction and simulation is applying advanced mathematical methods to extract parasitic wire impedances and to reduce the complexity of the extracted data. This reduction is needed so that the analysis and simulation of the critical wires in a circuit can be performed in a reasonable amount of time while not sacrificing precision.

Wire impedance and signal coupling effects are quite important in DSM circuits and must not be overlooked during the design process. These signal-integrity-related effects are also extremely difficult to deal with during the post-layout verification phase. Circuit design methodologies are emerging that are targeted to identifying possible wire delay bottlenecks and dealing with these effects before the actual physical layout is completed. Among the most promising techniques are the automatic repeater insertion to reduce the *RC* delays of long wires and circuit and architectural techniques that effectively "balance" wire delay distribution to ease the physical design process. Clock skew scheduling, retiming, wave pipelining, and a combination of these methods are potentially feasible and useful techniques for balancing the distribution of the delays within a circuit.

CONCLUSIONS

With the incessant advances of integrated-circuit design and manufacturing technologies, the performance of CMOS integrated circuits has become strongly dependent on low-level physical effects. Lower supply voltages and effects such as velocity saturation and channel-length modulation are contributing to the degradation of signal waveforms as signals propagate through the logic gates. Furthermore, the changing nature of the interconnect from a lumped capacitive to a distributed resistance-capacitance has increased the signal propagation delay through passive interconnect structures.

Therefore, it is becoming increasingly difficult to design correctly functioning circuits while satisfying performance criteria such as higher clock frequencies (lower clock periods) and low power. The smaller device sizes and larger integrated-circuit chip dimensions create conditions for the existence of long interconnect structures. These structures account for a growing portion of the combined logic and interconnect delay and shift the primary cause of the delay from the logic elements to the interconnect. In addition, the variations of the values of the physical parameters introduced during circuit manufacturing can substantially change the overall behavior and timing characteristics of the physical structures. It is therefore imperative that the signal properties be well understood and properly applied in the design of high-performance VLSI integrated circuits.

BIBLIOGRAPHY

1. J. S. Kilby, Invention of the integrated circuit, *IEEE Trans. Electron Devices,* **ED-23**: 648–654, 1976.
2. J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective,* Upper Saddle River, NJ: Prentice-Hall, 1996.
3. N. Gaddis and J. Lotz, A 64-b quad-issue CMOS RISC microprocessor, *IEEE J. Solid-State Circuits,* **SC-31**: 1697–1702, 1996.
4. P. E. Gronowski et al., A 433-MHz 64-bit quad-issue RISC microprocessor, *IEEE J. Solid-State Circuits,* **SC-31**: 1687–1696, 1996.
5. N. Vasseghi et al., 200-MHz superscalar RISC microprocessor, *IEEE J. Solid-State Circuits,* **SC-31**: 1675–1686, 1996.
6. H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI,* Reading, MA: Addison-Wesley, 1990.
7. S. Bothra et al., Analysis of the effects of scaling on interconnect delay in ULSI circuits, *IEEE Trans. Electron Devices,* **ED-40**: 591–597, 1993.
8. C. Mead and L. Conway, *Introduction to VLSI Systems,* Reading, MA: Addison-Wesley, 1980.
9. N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective,* 2nd ed., Reading, MA: Addison-Wesley, 1993.
10. F. Anceau, A synchronous approach for clocking VLSI systems, *IEEE J. Solid-State Circuits,* **SC-17**: 51–56, 1982.
11. M. Afghani and C. Svensson, A unified clocking scheme for VLSI systems, *IEEE J. Solid-State Circuits,* **SC-25**: 225–233, 1990.
12. S. H. Unger and C.-J. Tan, Clocking schemes for high-speed digital systems, *IEEE Trans. Comput.,* **C-35**: 880–895, 1986.
13. H. Shichman and D. A. Hodges, Modeling and simulation of insulated-gate field-effect transistor switching circuits, *IEEE J. Solid-State Circuits,* **SC-3**: 285–289, 1968.
14. S.-M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design,* New York: McGraw-Hill, 1996.
15. N. Hedenstierna and K. O. Jeppson, CMOS circuit speed and buffer optimization, *IEEE Trans. Comput.-Aided Design,* **CAD-6**: 270–281, 1987.
16. T. Sakurai and A. R. Newton, Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas, *IEEE J. Solid-State Circuits,* **SC-25**: 584–594, 1990.
17. A. I. Kayssi, K. A. Sakallah, and T. M. Burks, Analytical transient response of CMOS inverters, *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.,* **CAS I-39**: 42–45, 1992.
18. E. G. Friedman, *Clock Distribution Networks in VLSI Circuits and Systems.* IEEE Press, 1995.
19. H. B. Bakoglu and J. D. Meindl, Optimal interconnection circuits for VLSI, *IEEE Trans. Electron Devices,* **ED-32**: 903–909, 1985.
20. A. Wilnai, Open-ended RC line model predicts MOSFET IC response, *Electron. Design News,* 53–54, December 1971.
21. T. Sakurai, Approximation of wiring delay in MOSFET LSI, *IEEE J. Solid-State Circuits,* **SC-18**: 418–426, 1983.
22. G. Y. Yacoub et al., A system for critical path analysis based on back annotation and distributed interconnect impedance models, *Microelectron. J.,* **19**: 21–30, 1988.
23. M. R. C. M. Berkelaar and J. A. G. Jess, Gate sizing in MOS digital circuits with linear programming, *EDAC: Proc. Eur. Design Automat. Conf.,* 1990, pp. 217–221.
24. O. Coudert, Gate sizing for constrained delay/power/area optimization, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* **VLSI-5**: 465–472, 1997.
25. U. Ko and P. T. Balsara, Short-circuit power driven gate sizing technique for reducing power dissipation, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* **VLSI-3**: 450–455, 1995.
26. S. R. Vemuru and A. R. Thorbjornsen, Variable-taper CMOS buffer, *IEEE J. Solid-State Circuits,* **SC-26**: 1265–1269, 1991.
27. C. Prunty and L. Gal, Optimum tapered buffer, *IEEE J. Solid-State Circuits,* **SC-27**: 118–119, 1992.
28. N. Hedenstierna and K. O. Jeppson, Comments on the optimum CMOS tapered buffer problem, *IEEE J. Solid-State Circuits,* **SC-29**: 155–158, 1994.
29. B. S. Cherkauer and E. G. Friedman, Channel width tapering of serially connected MOSFET's with emphasis on power dissipation, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* **VLSI-2**: 100–114, 1994.
30. B. S. Cherkauer and E. G. Friedman, Design of tapered buffers with local interconnect capacitance, *IEEE J. Solid-State Circuits,* **SC-30**: 151–155, 1995.
31. B. S. Cherkauer and E. G. Friedman, A unified design methodology for CMOS tapered buffers, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* **VLSI-3**: 99–111, 1995.
32. V. Adler and E. G. Friedman, Repeater insertion to reduce delay and power in *RC* tree structures, *Proc. Asilomar Conf. Sign., Syst., Comput.,* 1997.

33. V. S. Adler and E. G. Friedman, Delay and power expressions for a CMOS inverter driving a resistive-capacitive load, *1996 IEEE Int. Symp. Circuits Syst.,* **4**: 1996, pp. 101–104.

34. V. Adler and E. G. Friedman, Repeater design to reduce delay and power in resistive interconnect, *Proc. IEEE Int. Symp. Circuits Syst.,* 1997, pp. 2148–2151.

35. V. Adler and E. G. Friedman, Timing and power models for CMOS repeaters driving resistive interconnect, *Proc. IEEE ASIC Conf.,* 1996, pp. 201–204.

36. C. J. Alpert, Wire segmenting for improved buffer insertion, *Proc. IEEE/ACM Design Automat. Conf.,* 1997.

37. V. E. Adler and E. G. Friedman, Repeater design to reduce delay and power in resistive interconnect, *IEEE Trans. Circuits Syst. II, Analog Digital Sign. Process.,* **CAS II-45**: 607–616, 1998.

38. I. E. Sutherland, Micropipelines, *Commun. ACM,* **32**: 720–738, 1989.

Reading List

R. J. Antinone and G. W. Brown, The modeling of resistive interconnects for integrated circuits, *IEEE J. Solid-State Circuits,* **SC-18**: 200–203, 1983.

E. Barke, Line-to-ground capacitance calculation for VLSI: A comparison, *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.,* **CAD-7**: 295–298, 1988.

J. R. Black, Electromigration—A brief survey and some recent results, *IEEE Trans. Electron Devices,* **ED-16**: 338–347, 1969.

J.-H. Chern et al., Multilevel metal capacitance models for CAD design synthesis systems, *IEEE Electron Device Lett.,* **13**: 32–34, 1992.

Z. Chen et al., CMOS technology scaling for low voltage low power applications, *1994 IEEE Symp. Low Power Electron., Dig. Tech. Papers,* 1994, pp. 56–57.

S. Dutta, S. S. M. Shetti, and S. L. Lusky, A Comprehensive Delay Model for CMOS Inverters, *IEEE J. Solid-State Circuits,* **SC-30**: 864–871, 1995.

M. Gilligan and S. Gupta, A methodology for estimating interconnect capacitance for signal propagation delay in VLSIs, *Microelectron. J.,* **26**: 327–336, 1995.

F. S. Lai, A generalized algorithm for CMOS circuit delay, power, and area optimization, *Solid-State Electron.,* **31**: 1619–1627, 1988.

J. Qian, S. Pullela, and L. Pillage, Modeling the "effective capacitance" for the *RC* interconnect of CMOS gates, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.,* **CAD-13**: 1526–1535, 1994.

J. Rubinstein, P. Penfield, and M. A. Horowitz, Signal delay in *RC* tree networks, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.,* **CAD-2**: 202–211, 1983.

T. Sakurai, Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSIs, *IEEE Trans. Electron Devices,* **ED-40**: 118–124, 1993.

T. Sakurai and K. Tamaru, Simple formulas for two- and three-dimensional capacitances, *IEEE Trans. Electron Devices,* **ED-30**: 183–185, 1983.

S. R. Vemuru and N. Scheinberg, Short-circuit power dissipation estimation for CMOS logic gates, *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.,* **CAS I-41**: 762–765, 1994.

IVAN S. KOURTEV
EBY G. FRIEDMAN
University of Rochester

INTEGRATED CIRCUITS, MICROWAVE. See MICROWAVE INTEGRATED CIRCUITS.

INTEGRATED CIRCUITS, NEURAL. See NEURAL CHIPS.

INTEGRATED CIRCUITS, OPTOELECTRONIC. See OPTOELECTRONICS IN VLSI TECHNOLOGY.

INTEGRATED CIRCUITS, POWER. See POWER INTEGRATED CIRCUITS.

INTEGRATED CIRCUIT TESTING. See AUTOMATIC TESTING.