Chapter 3

TRADE-OFFS IN CMOS VLSI CIRCUITS

Andrey V. Mezhiba and Eby G. Friedman
Department of Electrical and Computer Engineering, University of Rochester

3.1. Introduction

The pace of integrated circuit (IC) technology over the past three decades is well characterized by Moore's law. It was noted in 1965 by Gordon Moore [1] that the integration density of the first commercial ICs doubled approximately every year. A prediction was made that the economically effective integration density, that is, the number of transistors on an IC leading to the minimum cost per integrated component, will continue to double every year for another decade. This prediction has held true through the early 1970s. In 1975, the prediction was revised [2] to suggest a new, slower rate of growth – the transistor count doubling every two years. This new trend of exponential growth of IC complexity has become widely known as "Moore's Law". As a result, since the start of commercial production of ICs in the early 1960s, circuit complexity has risen from a few transistors to hundreds of millions of transistors operating concurrently on a single monolithic substrate. Furthermore, Moore's law is expected to continue at a comparable pace for at least another decade [3].

The evolution of integration density of microprocessor and memory ICs is shown in Figure 3.1 along with the original prediction of [1]. As seen from the data illustrated in Figure 3.1, DRAM IC complexity has been growing at an even higher rate, quadrupling roughly every three years. The progress of microprocessor clock frequencies is shown in Figure 3.2. Associated with increasing IC complexity and clock speed is an exponential increase in overall microprocessor performance (doubling every 18–24 months). This performance trend has also been referred to as Moore's law.

Such spectacular progress could not have been sustained for three decades without multiple trade-off decisions to manage the increasing complexity of ICs at the system, circuit and physical levels. The entire field of engineering can be described as the art and science of understanding and implementing trade-offs. The topic of IC design is no exception; rather, this field is an ideal example of how trade-offs drive the design process. In fact, the progress of VLSI technology makes the topic of trade-offs in the IC design process particularly instructive.

C. Toumazou et al. (eds), Trade-Offs in Analog Circuit Design: The Designer's Companion, 75–114.

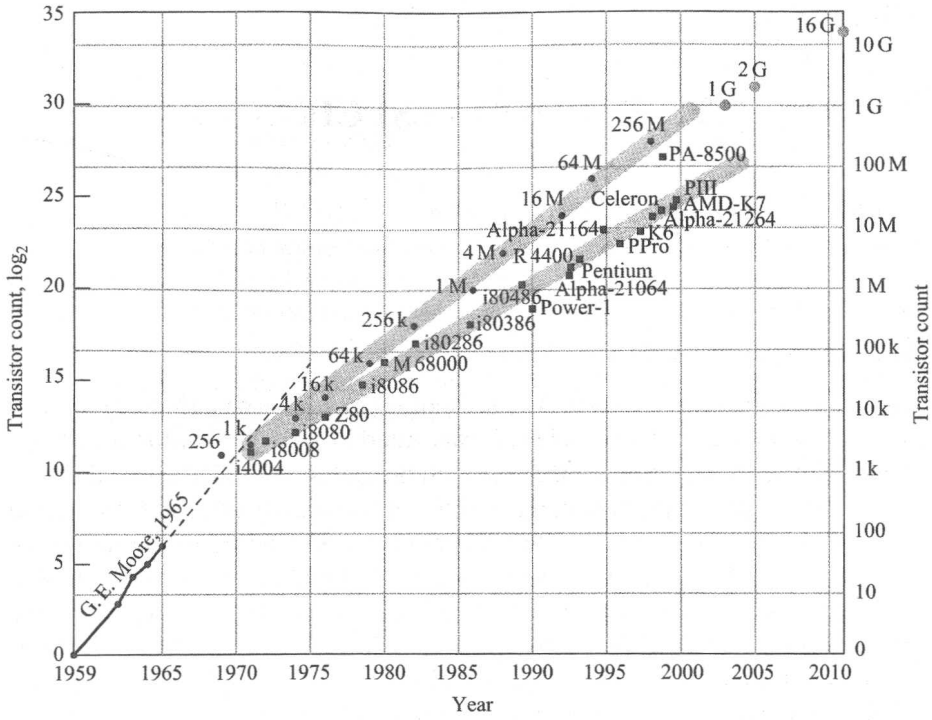

Figure 3.1. Evolution of transistor count of CPU/microprocessor and memory ICs. In the lower left corner, the original Moore's data [1] is displayed followed by the extrapolated prediction (dashed line).

The evolution of design criteria in CMOS ICs is illustrated in Figure 3.3. Design paradigm shifts shown in the figure are due to advances in the fabrication technology and the emergence of new applications. In the 1970s, yield concerns served as the primary limit to IC integration density and, as a consequence, die area was the primary issue in the IC design process. With advances in fabrication technology, yield limitations became less restricting, permitting the rise of circuit speed in the 1980s as the criterion with the highest level of priority. At the same time, new applications such as satellite electronics, digital wrist watches, portable calculators and pacemakers established a new design concept – design for ultra-low power. As device scaling progressed and a greater number of components were integrated onto a single die, on-chip power dissipation began to produce significant economic and technical difficulties. While the market for high-performance circuits could support the added cost, the design process in the 1990s has focused on optimizing speed and power, borrowing certain approaches from the ultra-low power design

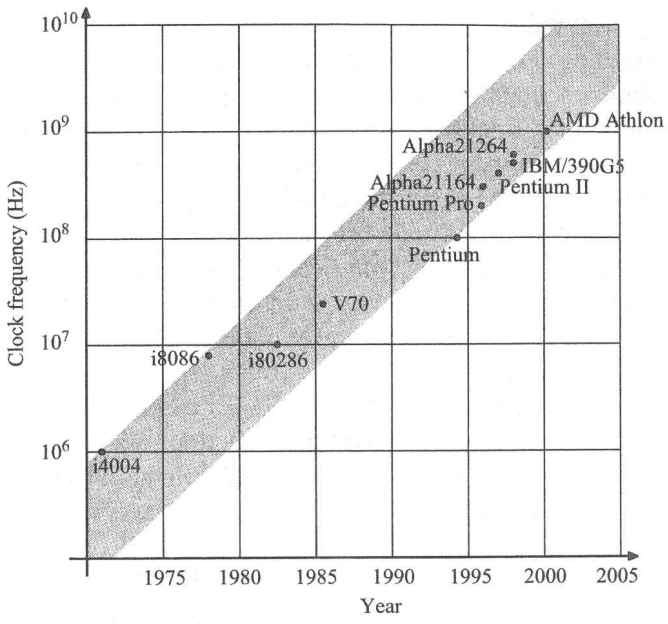

Figure 3.2. Evolution of microprocessor clock frequencies.

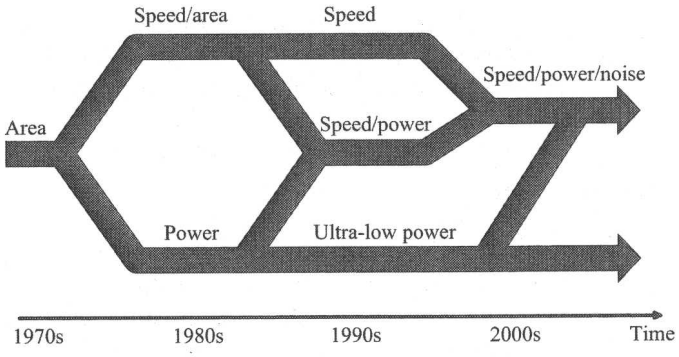

Figure 3.3. Evolution of design goals in CMOS ICs.

methodologies. Concurrently, a variety of portable electronic devices further increased the demand for power efficient and ultra-low power ICs. A continuing increase in circuit power dissipation exacerbated system price and performance, making power a primary design metric across an entire range of applications. Furthermore, aggressive device scaling and increasing circuit complexity are causing severe noise (or signal integrity) issues in VLSI circuits. Ignoring the effect of noise is no longer possible in the design of high-speed digital ICs.

These changes are reflected in the convergence of "speed" and "speed/power" trends to "speed/power/noise," as depicted in Figure 3.3.

Current semiconductor fabrication technology is able to place an entire system on a single die. Implementation of such systems-on-a-chip (SoC) has created new constraints and placed different requirements on the design process. The challenge of the VLSI design process has become the difficult problem of determining the proper set of trade-offs across high levels of complexity from system specification to the lowest physical circuit and layout details.

The material presented in this chapter on trade-offs in VLSI-based CMOS circuits is not intended to be comprehensive; rather, effort is made to summarize the primary trends and provide the reader with a general understanding of the topic of trade-offs in digital VLSI-based CMOS circuits. All types of trade-offs are available at different levels of system abstraction. Trade-offs at the higher levels of abstraction such as at the system and behavioral levels are highly application specific and difficult to systematize. These levels are not specifically treated in this chapter.

The chapter is organized as follows. Different VLSI design criteria are summarized in Section 3.2. Design trade-offs at various levels of design abstraction are considered in further sections. Architectural (register transfer) level trade-offs are treated in Section 3.3. Circuit trade-offs are considered in Section 3.4. Physical and process level trade-offs are discussed in Sections 3.5 and 3.6, respectively. The chapter closes in Section 3.7 with some conclusions and comments on future trends in trade-offs in CMOS-based VLSI systems. The terms and notations used throughout the chapter are defined in the following Glossary.

3.2. Design Criteria

Traditionally, there have been three primary figures of merit in the digital circuit design process: area, delay and power. Increasing speed, physical size, complexity and scaling of ICs have produced additional metrics to be considered as major design criteria, such as reliability, testability, noise tolerance, packaging performance and design productivity. A brief survey of these design criteria is provided in the following subsections.

3.2.1. Area

Die area is synonymous with "cost" in the VLSI field as die area has the greatest impact on die fabrication costs. Larger area reduces the number of dies that can fit onto a wafer, leading to a linear increase in processing and material costs. Much more significant, however, is the impact of die area on die yield. The yield, or the fraction of the fabricated ICs that are fully functional [4], falls sharply with die area, as shown in Figure 3.4. As a result, die manufacturing

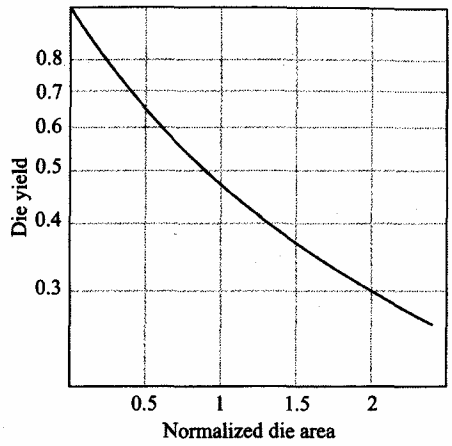

Figure 3.4. Dependence of yield on die area [6]. The area is scaled by the average defect density. The numbers on the abscissa provide the average number of defects per die.

costs quickly become prohibitive beyond some size determined by the process technology and the defect density characteristics of the manufacturing facility.

3.2.2. Speed

Although circuit performance is highly application specific, lower circuit propagation delay almost always leads to higher performance in digital systems. For this reason, VLSI performance is primarily discussed in terms of circuit speed (i.e. circuit propagation delay or the maximum clock frequency of a synchronous circuit) [4]. Therefore, the area-delay characteristics of a circuit are quite similar to the greater price-performance characteristics of that same circuit.

3.2.3. Power

Power dissipation in VLSI circuits also has a profound impact on both price and performance. High power dissipation penalizes the overall system since more advanced packaging and heat removal technology are necessary. Limits on power dissipation in advanced packaging can place an upper bound on economically viable integration densities before die yield limits the maximum die size. Higher power dissipation not only limits a circuit through packaging issues but also requires wider on-chip and off-chip power buses (reducing the wiring capacity available for the signal interconnect), larger on-board bypass capacitors, and often more complicated power supplies. These factors increase the system size and cost. Furthermore, portable electronic devices are limited by battery life (i.e. the time of autonomous operation); therefore, power is also a

system performance metric. In fact, the primary reason for CMOS dominatin the VLSI era has been the low power dissipation characteristics of static CMO circuits.

3.2.4. Design Productivity

Technology scaling has brought new design challenges. These challenge are caused by two primary issues. The first issue is the increasing complexit of the systems being designed. During most of the history of the semiconduc tor industry, die fabrication has been the primary constraining factor in circu complexity. The design task had been to make the most effective use of the lim ited silicon real estate. This situation has changed radically. The capabilities c the semiconductor manufacturing industry have far outpaced those of the I design industry. This "design gap" is well demonstrated by the graph show in Figure 3.5. Current multimillion transistor systems require huge amounts c highly skilled non-recurring engineering (NRE) effort. As the design produc tivity gap widens and NRE design cycles become longer, design teams hav become larger and NRE design costs have become a larger fraction of the tot cost. This trend has limited the development of high complexity SoCs to thos applications where the large NRE can be amortized over a high volume c products, such as RAMs and microprocessors. The large demand on NRE i further exacerbated when the circuits operate at high levels of performanc requiring significantly more design effort.

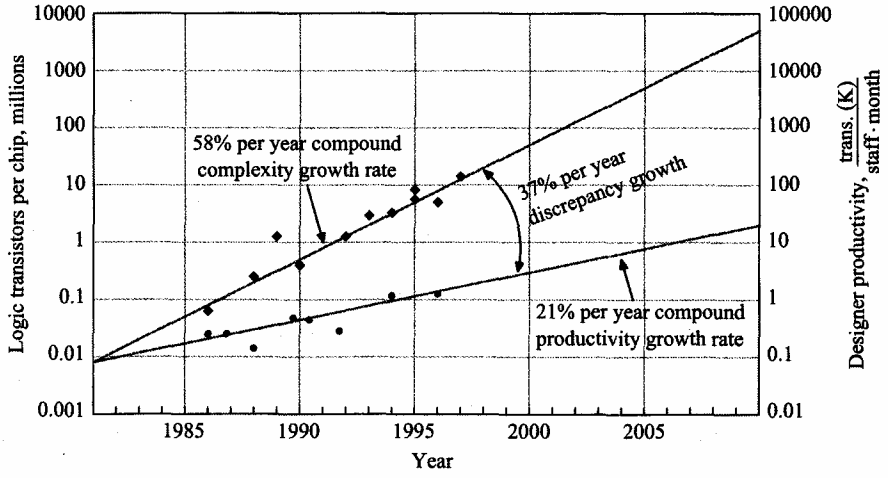

Figure 3.5. The design productivity gap [3].

In addition to the issue of cost, insufficient design productivity has made time-to-market longer and less predictable. This NRE to recurring engi-ring (RE) issue has become critical within the semiconductor industry. The rent pace of technology is shrinking the product life cycle, creating win-ws of opportunity for many products that are measured in months. Therefore, ely market introduction is paramount. Missing a product delivery deadline xtremely costly and can jeopardize the commercial success of a product. ge design teams may shorten the average development time but often do prevent design deadline slips. Trading off product capabilities and fea-es for less development effort to meet time-to-market constraints is often voidable.

2.5. Testability

Another challenge related to increasing system complexity lies in the area testing, specifically debug testing (as compared to production testing). The mber of distinct stable signal patterns a digital system can assume increases onentially with the number of inputs and the number of registers storing internal states. A state (i.e. a logic value) of a circuit node is typically t directly accessible and must be shifted to the output pins of an IC. This ocess makes the cost of exhaustive testing prohibitive for even relatively mple circuits. Limited testing of a complex system has become exorbitant less special provisions are made during the design process to ensure that the ting process is more effective. Thus, a moderate sacrifice in area and speed is tified since the increased die manufacturing cost and decreased performance compensated by a vast increase in system testability. Even with such added asures, the cost of testing per transistor has not changed significantly over years, whereas manufacturing costs have plummeted exponentially. As a ult, the portion of the test costs in terms of the total cost has grown. If current nds continue, this share will surpass all other cost components within a few ars [7]. Forecasts of the number of transistors per I/O signal pin and the cost testers for high-performance ICs are shown in Figure 3.6. The number of I/O ns increases moderately with time resulting in a large increase in the number transistors per I/O pin. The time required to test multimillion transistor logic ough dozens to hundreds of pins is not realistic, necessitating the extensive e of built-in self-test (BIST) structures. Thus, due to the high cost and limited roughput of test equipment and support personnel, a moderate reduction in st time can produce a considerable reduction in total project cost.

2.6. Reliability

Another source of circuit and physical problems is the smaller dimen-ons of the circuit elements. Changing physical dimensions and increasing

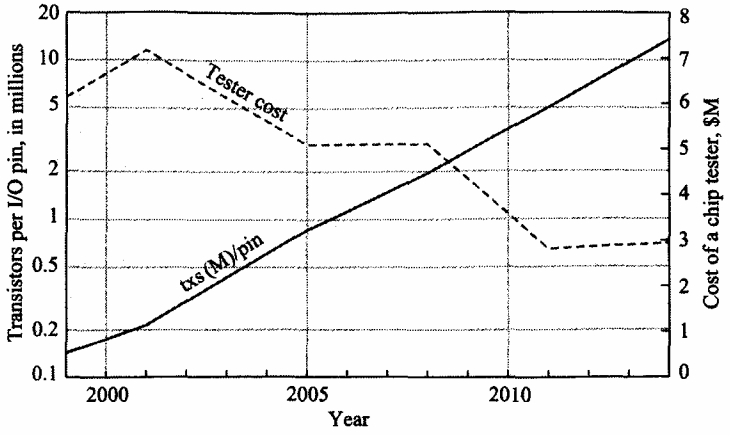

Figure 3.6. The forecast of transistor per signal I/O pin ratio and the forecast of high-performance tester cost (based on data from [3]).

speed has made reliability, packaging, and noise constraints more difficult to satisfy.

Scaling of device feature sizes without a proportional reduction in the supply voltage leads to higher electric fields, exacerbating many reliability concerns. Breakdown caused by high electric fields is one of the primary failure mechanisms. An example of a problem caused by high electric fields is that these fields give rise to hot electrons which tunnel from the channel into the gate oxide causing long-term reliability problems such as threshold voltage variations and transconductance degradation. High electric fields also produce carrier multiplication and substrate leakage current [5]. Excessive current densities in the metal lines cause electromigration problems: the metal ions are moved from the crystal lattice by colliding with the electrons propagating through the conductor [8]. The resulting voids and hillocks create open and short circuits, leading to permanent circuit failure. Electromigration can become a limitation at greater integration densities and finer feature sizes [9].

3.2.7. Noise Tolerance

Noise rejection and signal regeneration properties are two of the principal advantages of digital circuits. Nevertheless, many types of noise sources are present in VLSI systems. Inter- and intra-layer capacitive and inductive coupling of interconnect, as illustrated in Figure 3.7, results in increased delay, waveform degradation, and most importantly, the possibility of an erroneous interpretation of the digital signals [10,12,17,18]. Substrate currents result in substrate coupling which is particularly critical in dynamic

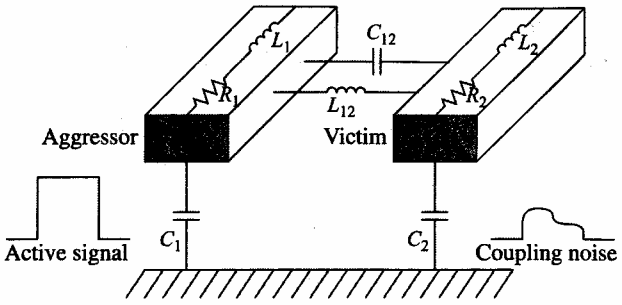

Figure 3.7. Cross-coupled interconnect noise in ICs. An active transition on an aggressor line induces a coupling noise voltage on a victim line.

circuits where a high impedance node can be easily affected [11]. As IC power consumption increases, the supply current has risen rapidly, currently reaching tens of amperes in high-performance ICs. The distribution of such high currents over increasingly larger die areas has produced challenging noise problems [5,13,14]. Due to the resistance of the power supply lines, significant *IR* voltage drops are created across the power buses, increasing the signal delay and delay uncertainty [15]. Another problem related to power distribution is simultaneous switching noise: as many amperes of current are switched on and off in subnanosecond time periods, the inductive voltage drops across the on-chip power lines and off-chip package bonding wires induce unacceptable voltage variations across the power rails [16]. Faster clock rates create higher slew rates of the signal waveforms, increasing the on-chip noise. Issues of signal integrity necessitate considering the analog nature of digital signals. These noise sources can potentially cause a circuit to both slow down and malfunction. Mitigating these noise problems has become a major VLSI challenge.

3.2.8. Packaging

Packaging is another important criterion requiring serious consideration in the design process. Packaging imposes many limits on an IC: heat dissipation, packaging price overhead, number of pins, circuit bandwidth, input cross-coupling noise, simultaneous switching noise, etc. The performance, price and power dissipation of a product are all affected (and often constrained) by the target package.

3.2.9. General Considerations

A few comments on power dissipation, technology scaling and VLSI design methodologies are offered in this section so as to better understand the trade-offs

discussed later in this chapter. These few paragraphs provide a synopsis of highly complicated topics important to the CMOS VLSI circuit design process.

Power dissipation in CMOS VLSI circuits. There are three primary components of power dissipation in CMOS circuits:

$$\begin{aligned} P_{\text{total}} &= P_{\text{dynamic}} + P_{\text{sc}} + P_{\text{leakage}} \\ &= CV_{\text{DD}}^2 f_{\text{switch}} + \tfrac{1}{2} I_{\text{peak}} t_{\text{base}} V_{\text{DD}} f_{\text{switch}} + I_{\text{leakage}} V_{\text{DD}} \end{aligned} \tag{3.1}$$

The dynamic power P_{dynamic} accounts for the energy dissipated in charging and discharging the nodal capacitances. When a capacitor C is charged, $CV_{\text{DD}}^2/2$ joules of energy is stored on the capacitor. An equivalent amount of energy is dissipated on the interconnect and transistors that are being charged. In the discharge phase, $CV_{\text{DD}}^2/2$ joules of energy that are stored on the capacitor is dissipated on the transistors and interconnect through the discharge path as shown in Figure 3.8. Thus, the total energy expended in the charge/discharge cycle is CV_{DD}^2. The average dynamic power is the CV_{DD}^2 amount of energy times the average frequency of the charge/discharge cycle f_{switch}, producing the well-known expression for dynamic power in CMOS circuits, $CV_{\text{DD}}^2 f_{\text{switch}}$.

A short-circuit current flows in a static CMOS gate when a conductive path exists from the power rail to the ground rail. It is possible for such a path to exist when a signal at one of the gate inputs transitions, passing through intermediate voltage levels [19–23]. For a static CMOS gate, this voltage range is from the n-type transistor threshold V_{Tn}, the voltage at which the n-type transistors turn on, to $V_{\text{DD}} + V_{\text{Tp}}$, the voltage at which the p-type transistors cut off. Within this voltage range, both of the pull-up and pull-down networks conduct current, producing short-circuit current, as exemplified in Figure 3.9. The period of time when this conductive path exists is denoted as t_{base} in (3.1). An analytical expression [24] that characterizes the short-circuit power P_{sc} that exhibits 15% accuracy for a wide variety of RC loads based on the Sakurai alpha-power law

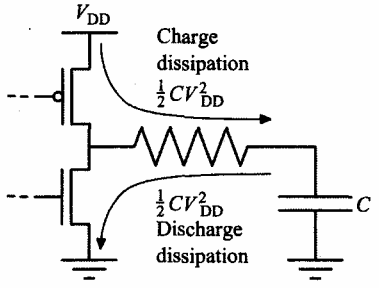

Figure 3.8. Energy dissipation during the charge/discharge cycle.

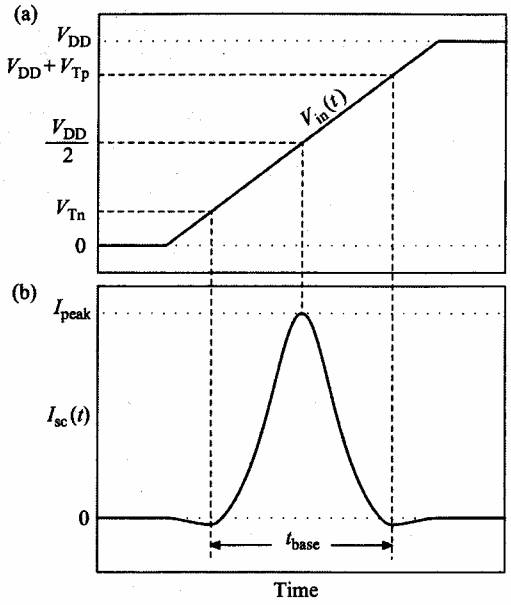

Figure 3.9. Short-circuit current waveform of a CMOS gate. (a) Ramp-shaped input waveform, (b) short-circuit current waveform.

model [25] is:

$$P_{sc} = \left| \ln \left(\frac{V_{Tn}}{V_{DD} + V_{Tp}} \right) \right| (R_{tr}C + RC)\, I_{peak} V_{DD} f_{switch} \qquad (3.2)$$

where $R_{tr} = V_{d0}/I_{d0}$ is alpha-power model effective transistor resistance in the linear region of operation.

The leakage current $I_{leakage}$ in a transistor is the current that flows between the power terminals in the absence of switching, giving rise to a leakage power component $P_{leakage}$. Typically, the dynamic power is the dominant power component, contributing 70–90% or more of the total power dissipation. Therefore, the most effective strategy for reducing the total power dissipation is to reduce the dynamic dissipation. For example, the quadratic dependence of the dynamic power on V_{DD} implies that lowering V_{DD} is an effective way to reduce both the dynamic and total power dissipation.

Technology scaling. The exponential growth of IC complexity has been largely driven by improvements in semiconductor fabrication technologies due to both technology scaling and defect density reduction. Shrinking the size of the circuit elements addresses all three of the "classical" VLSI design criteria. Capacitive loads within CMOS circuits are reduced as circuit elements become

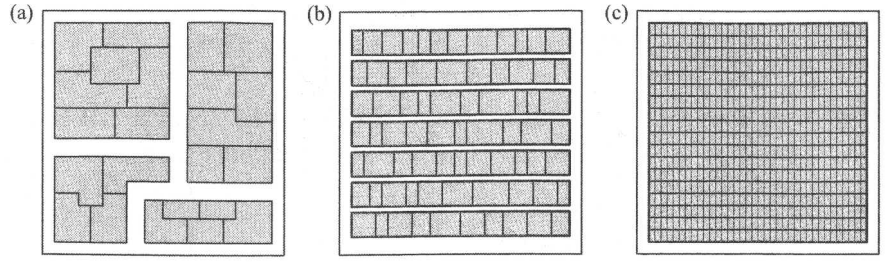

Figure 3.10. Different design approaches. (a) full custom – most laborious but most flexible, the size, shape and placement of the circuit components are freely tailored; (b) standard cell – the circuit is composed of logic gates aligned in rows of fixed height with routing channels between the rows, a library of standard logic cells is compiled prior to the design implementation and connected by metalization layers; (c) channelless gate array – wafers containing arrays of transistor cells are prefabricated and the transistors are interconnected to form a target circuit function during the metalization steps.

smaller, enhancing the delay characteristics. Circuits require less area, thereby lowering manufacturing costs, permitting the on-chip integration of larger and more complex circuits. To maintain a constant electric field, the supply voltage is often reduced. As a result, less energy is required to charge (and discharge) a capacitive load, reducing the power consumed. (For a more thorough discussion of technology scaling, see e.g. [5].)

VLSI design methodologies. The high cost and long design time of full custom circuits have prompted the creation of automated design approaches. These approaches rely on a variety of different methodologies in which circuits are automatically mapped into silicon such as automated placement and routing of standardized cells [26]. Significant geometrical constraints are imposed on the layout to make the circuit more amenable to automated place and route techniques. Circuit structures amenable to such approaches are illustrated in Figure 3.10. Automated design methods yield suboptimal designs as compared to full custom methodologies. The greater number of constraints on the design methodology makes the resulting circuit less optimal, albeit with a faster time-to-market.

3.3. Structural Level

Once a system is specified at the behavioral level, the next step in the design process is to determine which computational algorithms should be employed and the type and number of system building blocks, interfaces and connections. While the process is application specific, two types of basic trade-offs are available: parallel processing and pipelining. A simple data path, shown

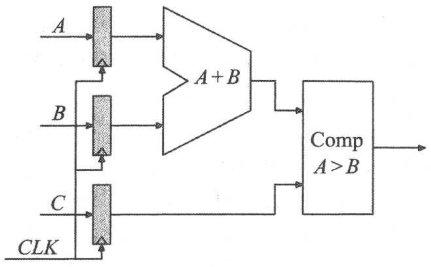

Figure 3.11. A simple data path [27].

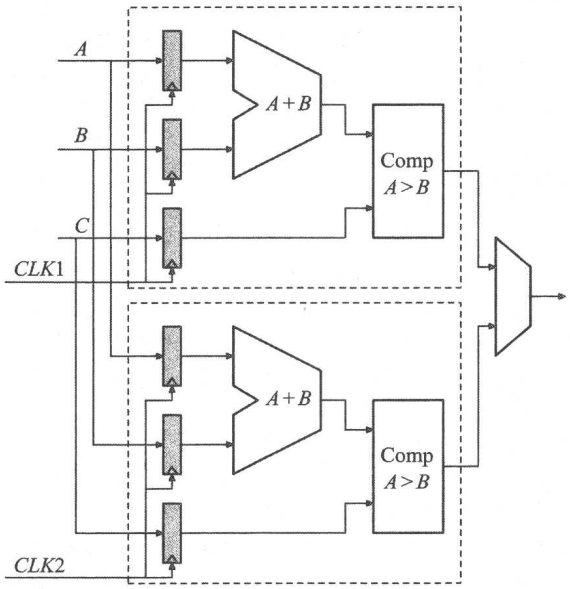

Figure 3.12. Parallel implementation of the data path [27].

in Figure 3.11, is used here to demonstrate these concepts. The speed, area and dynamic power of this circuit are compared and contrasted to a parallel implementation in Section 3.3.1 and to a pipelined implementation in Section 3.3.2.

3.3.1. Parallel Architecture

Parallel processing consists of duplicating a portion of a data path a number of times, and connecting the duplicate circuits in parallel with each other. This approach is illustrated by the circuit shown in Figure 3.12, where the parallel implementation of the data path shown in Figure 3.11 is depicted. Additional

circuitry is needed to maintain the correct data flow, such as the multiplexer and related control circuitry. Extra circuitry is also used to generate the different clocking signals for the two parallel blocks (not shown). Other conditions being equal, the computational throughput of the parallel implementation of the circuit is doubled as compared to the original serial implementation (assuming the multiplexer delay to be negligible). The circuit area is more than doubled due to the added circuitry and interconnect. As described by Chandrakasan et al. [27], the area of the parallel implementation shown in Figure 3.12 is 3.4 times larger than that of the reference circuit (a circuit implementation based on a 2 μm technology is assumed in [27]). The circuit capacitance is increased 2.15 times, leading to a proportional increase in power dissipation. For parallel processing to be effective, the algorithms should be suitable for parallelization such that high utilization of the added processing units is achieved and the overhead of the complex control circuitry is minimized.

3.3.2. Pipelining

The process of pipelining inserts new registers into a data path, breaking the path into shorter paths. This process shortens the minimum clock period from the delay of the original path to the delay of the longest of the new shorter paths [28]. Therefore, the resulting circuit can be clocked faster to achieve a higher synchronous performance. A pipelined version of the data path shown in Figure 3.11 is illustrated in Figure 3.13. If the delays of the adder and comparator are equal, the path can be clocked at almost double the original frequency, with the delay of the inserted registers preventing the system from operating at precisely double the original performance. The area penalty of pipelining is less than that of the parallel architecture approach since the processing elements are not duplicated and only the inserted registers are introduced. The area of the pipelined circuit shown in Figure 3.13 is 1.3 times larger than the area of the reference circuit shown in Figure 3.11 [27]. The capacitance (and therefore the dynamic power) is 1.15 times greater than that

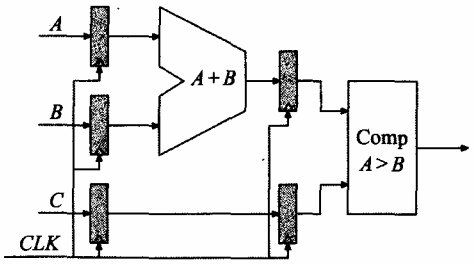

Figure 3.13. Pipelined implementation of the data path [27].

used in the serial approach. As in the parallel processing approach, pipelining is most effective in those circuits with a feed-forward nonrecursive data flow path. Pipelining also increases the latency of the system because of the added set-up and clock-to-output delays of the extra registers and any imbalance in the delays of the pipeline stages which may be detrimental to the overall system performance. As more registers are inserted into a data path, the delay of these registers becomes a larger fraction of the total path delay. Introducing more registers, therefore, has a diminishing return on performance. For a more detailed treatment of these issues, see [29,30].

The performance benefits of the parallel and pipelined approaches can improve the power characteristics, trading off area and power for speed. Instead of maintaining a fixed voltage and gaining computational throughput, another possible trade-off is to decrease the supply voltage to a level sufficient to maintain the original throughput. For the circuit shown in Figure 3.12, this strategy means decreasing the power supply until the delay doubles. The same voltage maintains the performance of the pipelined version at the original level assuming the added register delays are negligible and the delay of the new data paths are well balanced. Assuming an initial voltage of 5 V, the scaled voltage level to maintain the same effective performance is 2.9 V [27]. The power of the parallel implementation $P_{\|}$ normalized to the power of the reference path P_{ref} is:

$$P_{\|} = C_{\|} V_{\|}^2 f_{\|} = (2.15C_{\text{ref}})\,(0.58V_{\text{ref}})^2 \left(\frac{f_{\text{ref}}}{2}\right) \approx 0.36P_{\text{ref}}$$

Similarly, the normalized power of the pipelined implementation P_{pipe} is:

$$P_{\text{pipe}} = C_{\text{pipe}} V_{\text{pipe}}^2 f_{\text{pipe}} = (1.15C_{\text{ref}})\,(0.58V_{\text{ref}})^2\, f_{\text{ref}} \approx 0.39P_{\text{ref}}$$

These substantial reductions in power consumption are the result of the quadratic dependence of dynamic power on the supply voltage.

Pipelining and parallelism can also be combined to further improve performance and power. To maintain a critical delay of the original data path, the supply voltage can be lowered to 2 V, reducing the power by a factor of five,

$$P_{\|\text{pipe}} = C_{\|\text{pipe}} V_{\|\text{pipe}}^2 f_{\|\text{pipe}} = 2.5C_{\text{ref}} \cdot (0.4V_{\text{ref}})^2 \cdot \left(\frac{f_{\text{ref}}}{2}\right) \approx 0.2P_{\text{ref}}$$

The data and the implicit trade-offs are summarized in Table 3.1.

3.4. Circuit Level

Most CMOS-specific trade-offs are made at the circuit abstraction level. Trade-offs involved in selecting dynamic or static implementations are discussed in Section 3.4.1. Transistor sizing, a central issue in CMOS circuit

Table 3.1. Summary of trade-off data (from [27]).

Data path implementation	*Voltage (V)*	*Area*	*Power*
Original data path	5.0	1	1
Parallel data path	2.9	3.4	0.36
Pipelined data path	2.9	1.3	0.39
Parallel and pipelined	2.0	3.7	0.2

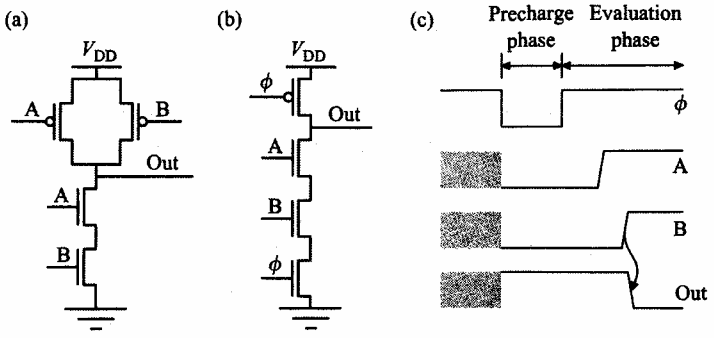

Figure 3.14. CMOS implementations of a two-input NAND gate. (a) Static implementation of the NAND gate, (b) dynamic implementation of the NAND gate, (c) timing diagram of the dynamic implementation.

design, is considered in Section 3.4.2. Trade-offs in tapered buffers are reviewed in Section 3.4.3.

3.4.1. Static versus Dynamic

The use of dynamic or static CMOS structures to implement a circuit function is an important decision that is made at the circuit level. The concepts of static and dynamic styles are illustrated in Figure 3.14. Both choices have virtues. Static CMOS is relatively simple to design, and is both robust and noise tolerant. Alternatively, dynamic CMOS uses fewer transistors to implement a given logic function, requires less area, has smaller parasitic capacitances, and is able to operate at higher speeds. Dynamic circuits also do not dissipate power due to spurious transitions (or glitches). However, dynamic circuits have higher switching activities. All of the nodes are charged during a precharge phase; many of these nodes are charged and immediately discharged during the evaluation phase. In contrast, in static circuits, except for spurious transitions, the output nodes are charged or discharged only when the logic values change. Static circuits can also be easily powered down by gating the clock signal; dynamic circuits, alternatively, require a small amount of additional circuitry

to preserve a state in the absence of the clock signal, increasing the parasitic capacitance and decreasing the circuit speed. The choice of circuit type is, however, not mutually exclusive. Static and dynamic circuits can both be used within the same IC. More complex design and verification of dynamic circuits is required in order to avoid potential hazards. It is, therefore, common to implement performance critical parts of a circuit design in dynamic CMOS in order to meet stringent performance goals and to implement the remaining circuitry in static CMOS in order to save design time while improving overall circuit robustness.

3.4.2. Transistor Sizing

Transistor sizing is another fundamental trade-off at the circuit level in CMOS logic families [31–45]. As transistors become wider, the current drive increases (the output resistance decreases) linearly with the transistor width, decreasing the propagation delay. The physical area and gate capacitance also increase linearly with width, increasing the circuit area and power. Thus, the optimal transistor size is strongly dependent on the trade-off of area and power for speed. Furthermore, the same type of optimization process may produce different approaches satisfying different design goals. Consider, for example, a static CMOS inverter in which the NMOS to PMOS transistor width ratio is chosen to minimize the propagation delay. The ratio

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p} \tag{3.3}$$

balances the output rise and fall transition times. An alternative option is the ratio

$$\frac{W_p}{W_n} = \sqrt{\frac{\mu_n}{\mu_p}} \tag{3.4}$$

which minimizes the average of the rise and fall delays [48]. Note, however, that either of these choices can produce the worst case signal delay depending upon the input rise and fall transition times. Transistor sizing depends, therefore, on both the optimization criteria and the circuit context. The primary transistor sizing trade-offs are considered below.

A common objective of transistor sizing is delay minimization. Consider a CMOS circuit with the output loads dominated by the input capacitances of the following stages. The typical dependence of the capacitor charging delay on the transistor width is shown in Figure 3.15. The charge time monotonically decreases with increasing transistor width. However, a caveat is that the input load of the transistor increases linearly with the transistor width, delaying the preceding gate. The net result is that the total delay of a data path with more stages can be smaller; an example circuit is illustrated in Figure 3.16. Similarly,

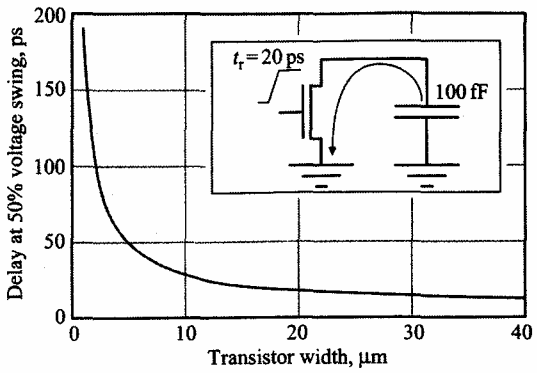

Figure 3.15. SPICE simulation of a capacitor discharge time as a function of the transistor width. TSMC 0.25 μm CMOS technology (MOSIS), 100 fF external load and $V_{DD} = 2.5$ V.

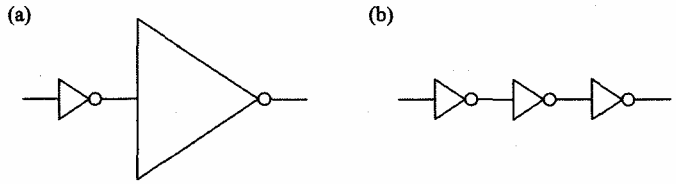

Figure 3.16. The effect of the gate input load on the propagation delay of a data path. Due to the large input capacitance it is possible for the path shown in (a) to have a longer delay than the path shown in (b) although there are less gates in the path shown in (a).

a uniform increase of all of the transistor sizes does not substantially change the propagation delay of a circuit in which the output loads are dominated by the input capacitances of the fanout while a linear increase in power and area will occur. The current drive of the gates I_{out} will increase which is offset by an increase in the output capacitive loads C_L. The I_{out}/C_L ratio remains essentially constant. A careful balance of the current drive and input load is therefore necessary to enhance circuit performance.

Two iterative algorithms, one algorithm for minimum delay and the other for minimum active area under a delay constraint, are described by Lee and Soukup in [31] for combinational circuits driving large capacitive loads. An important conclusion of these algorithms is the rapid rise of silicon area as the minimum area is approached. For example, the area of a tristate output buffer designed in a 2.5 μm CMOS technology to drive a 25 pF load more than triples when the delay is reduced from 28 ns to the minimum of 22 ns. This behavior is further aggravated in deep submicrometer (DSM) technologies as the interconnect impedances increase the area penalty while degrading any

device delay advantages. Design for minimum delay is therefore seldom a practical solution.

The area and speed trade-offs achieved by transistor sizing are also dependent on the design style. Full-custom design is the most flexible; semi-custom design strategies impose certain geometrical constraints such as a fixed cell height in standard cell methodologies (see Figure 3.10). The size of the transistors within the cells can be either fixed (a typical cell library contains several cells implementing the same function with different output current capabilities) or adjusted at the time of cell invocation to satisfy a target current drive; gate array circuits are the most restrictive, with transistor sizes being multiples of the width of the prefabricated transistors, producing inefficient area utilization. A comparison of area-delay trade-offs in these design styles is presented in [32]. A full custom style offers the most efficient and flexible area-delay trade-off. Area-delay trade-offs among different implementations of a combinational path driving a large capacitive load are compared in [32]. One implementation uses a unit-sized cell with a tapered output driver between the last logic stage and the capacitive load. The second implementation uses tapered logic gates [32], where the final stage is sufficient to effectively drive the capacitive load. A circuit consisting of a chain of three inverters and a capacitive load one hundred times larger than the input load of a unit size inverter is considered for comparison. The first approach yields a circuit area and delay of 29 minimum inverter delays and 22 minimum inverter areas, while the second approach produces a delay and area of 16.8 minimum inverter delays and 32 minimum inverter areas, respectively.

The transistor size also affects the circuit power dissipation characteristics. A simple approximation is to consider the circuit power as linearly proportional to the total active area of a circuit A, that is, $P = CV^2 f$, where $C = C_{ox}A$, and the gate oxide capacitance per unit area C_{ox} is constant for a given technology. Under this assumption (i.e. no interconnect capacitance), power optimization and circuit area optimization are the same since the circuit area is assumed to be proportional to the active area. Therefore, a power optimal design should use only minimum size transistors as long as correct circuit operation is not affected.

Yuan and Svensson [33] discuss transistor sizing with respect to power-delay optimization. For a one-stage pass transistor circuit or inverter with a symmetric voltage transfer characteristic (VTC), the power-delay reaches a minimum when the gate output parasitic capacitance equals the load capacitance. Power optimal loading ratios are also calculated for more complex structures. Transistor size optimization for an energy-delay performance metric is considered in [34]. However, [33,34] neglect the short-circuit current contribution to the power dissipation, a significant power component in circuits with large fanout and, therefore, long transition times. An analytical power dissipation model

characterizing short-circuit power is described in [35]. In this case, the power optimal size of the transistors is dependent on the input slew rate, which in turn is a function of the input driver size and output load. The power optimal size for inputs with high slew rates are smaller than for inputs with low slew rates, as the short-circuit power is inversely proportional to the slew rate s [20]. If driving large capacitive loads, the power savings is substantial for optimally sized gates as compared to minimally sized gates. For an inverter driving ten minimal inverters, the power savings is 35%. If the load is 20 inverters, the power savings is 58%. However, the fraction of such high fanout gates in practical circuits is typically small.

The power optimal transistor size is smaller than the power-delay optimal transistor size. An efficient trade-off of power for delay occurs at intermediate sizes. Trade-offs beyond the power-delay optimum can be pursued in aggressive circuit designs. Two algorithms based on a power model are also developed in [35]. The first algorithm searches for the power optimal transistor size. Benchmark circuits optimized with this algorithm have average power savings of approximately 5%, average area increases of approximately 5%, and, typically, a lower delay as compared to those circuits with minimum active area. The second algorithm performs power optimization under a delay constraint. In benchmark circuits, this algorithm achieves a power savings of about 1–5% over a similar algorithm with minimum active area as the only design criterion.

The power supply has been assumed fixed in the discussion of transistor sizing. Releasing this restriction provides an added degree of freedom for power-area trade-offs. As described in [36], power savings through transistor sizing in order to lower the supply voltage is not effective for long channel devices; only a marginal savings can be achieved under a limited set of load conditions. However, for short channel devices, when the device current is linearly proportional to $V_{GS} - V_T$, a wide opportunity for such optimization exists. Beyond this power optimal size, any power saving through lower voltages is lost due to the larger amount of capacitance being switched. The increase in interconnect capacitance due to an increase in circuit area is neglected in the analytical model presented in [36].

A prescaler, consisting of four identical toggle flip flops, is investigated through SPICE simulations. An optimal size of four times the minimum width for uniformly scaled transistors is determined. At 300 MHz, the optimally scaled circuit consumes 50% less power (in a 1 μm CMOS technology). Two versions of the prescaler have also been manufactured in a 1 μm CMOS technology, one version based on minimum sized transistors and another version based on large, individually optimized transistors. To operate at 300 MHz, the first circuit requires a 5 V supply and consumes 1.740 mW, whereas the optimized circuit requires only a 1.5 V power supply and dissipates only 0.575 mW.

A variety of tools for automated transistor sizing has also been reported [37–39]. In [38], an average 50% reduction in power is reported for optimized circuits as compared to standard cell implementations operating at the same clock frequency. Alternatively, an average 25% gain in clock frequency is achieved dissipating the same amount of power. Techniques can also be applied to perform transistor size optimization under noise margin and charge sharing constraints [37]. Research on transistor size and input reordering optimization with respect to hot carrier reliability has been described in [43]. It is shown that optimization for hot carrier reliability and for power dissipation are quite different.

3.4.3. Tapered Buffers

An important special case of transistor sizing is tapered buffers. Consider the problem of driving a large capacitive load. Driving board traces and on-chip buses, where capacitances are typically two to four orders of magnitude larger than on-chip logic levels, is an example of such a task. To drive such large capacitive loads at an acceptable speed, an intermediate buffer is often used. Using an inverter appropriately scaled for the capacitive load (as shown in Figure 3.17(a)) reduces the delay; however, the large input capacitance of the inverter loads the previous logic with too large a capacitive load. A similar argument can be made when inserting another inverter, large enough to drive the inverter driving the load, and so on until the initial input inverter of the buffer is sufficiently small to be driven by the previous logic gate at an acceptable speed. Thus, a tapered buffer consists of a chain of inverters of gradually increasing size as illustrated in Figure 3.17(b). The ratio of an inverter size to the size of the preceding inverter is called the tapering factor β.

The idea of tapering was first introduced by Lin and Linholm [46]; these authors investigated trade-offs based on a weighted product of a *per-stage* delay and used the total buffer area as a figure of merit. Following Lin and Linholm, Jaeger [47] considered minimization of the *total* buffer delay as the primary optimization objective. Jaeger showed that, under the assumption that a stage

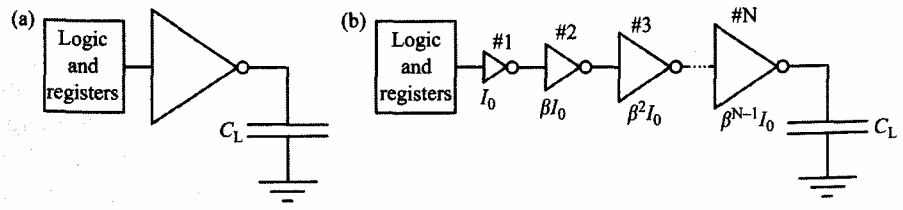

Figure 3.17. Circuit approaches to driving a large capacitive load. (a) A large inverter, (b) a tapered buffer. The delay of the tapered buffer is less than the delay of a single large inverter.

load is proportional to the next stage size (i.e. neglecting the intrinsic load of the gate), the delay of a tapered buffer reaches a minimum at a constant tapering factor $\beta_{opt} = e$ (the base of the natural logarithm) with a corresponding number of stages $N_{opt} = \ln M$, where $M = C_L/C_0$ is the ratio of the load capacitance C_L to the input capacitance C_0 of the initial inverter in the chain (usually considered to be minimum size). Note that because the number of buffer stages N is an integer, the aforementioned condition cannot in general be satisfied precisely. Therefore, one of the two integers closest to $\ln M$ is chosen, and β is calculated to satisfy $\beta^N = M$.

The approach of Lin and Linholm followed by Jaeger has been improved in several directions. More accurate delay models [48]–[50,54] and capacitance models [48,50,51,53] have been employed to allow for the intrinsic load capacitance, a ramp input signal, and short-circuit current. Initially, the effect of the intrinsic load capacitance was investigated by Kanuma [48] and Nemes [49]. These authors determined that the delay optimal tapering factor increases with the ratio of the intrinsic output capacitance (diffusion and gate overlap) to the input gate capacitance. Further improvements to account for the effects of the finite input slew rate and resulting short-circuit current were developed in [50,51,53]. In [51], the intrinsic output capacitance is increased by an analytically calculated value to account for the slower charging of the nodal capacitance. In [53], empirical data from circuit simulations are used to calculate the increased equivalent capacitance. A model considering both a finite slew rate and intrinsic loading is described in [50]. Further discussion of this topic can be found in [55,56]. To summarize these results, the delay optimal tapering factor varies from three to five, depending upon the target technology (i.e. the ratio of the input capacitance to the intrinsic output capacitance). The delay optimal transistor ratio of the inverter stages is $\sqrt{\mu_n/\mu_p}$, which minimizes the average output delay [50], although less than a 10% gain in delay is achieved as compared to equally sized transistors. A possible exception from this rule is the final stage where equal rise and fall times are often preferred over average delay minimization.

Area-power-delay trade-offs have also been considered [20,50,52,57,58]. It has been observed that for a given load the buffer delay versus tapering factor dependence is relatively flat around β_{opt}, as illustrated in Figure 3.18. Also, the total area of the buffer is a relatively strong function of β, falling with β. Thus, an effective trade-off of delay for area and power is possible. For example, if a buffer with an optimum number of stages is implemented with both four stages and three stages, the buffer delay rises by 3% and 22% but the area shrinks by 35% and 54%, respectively. Similar results are obtained by Vemuru in the investigation of tapered buffers with a geometrically increasing tapering factor [52]. While producing higher minimum delays (less than 15% greater than the smallest delays in a fixed-taper (FT) buffer), such buffers have

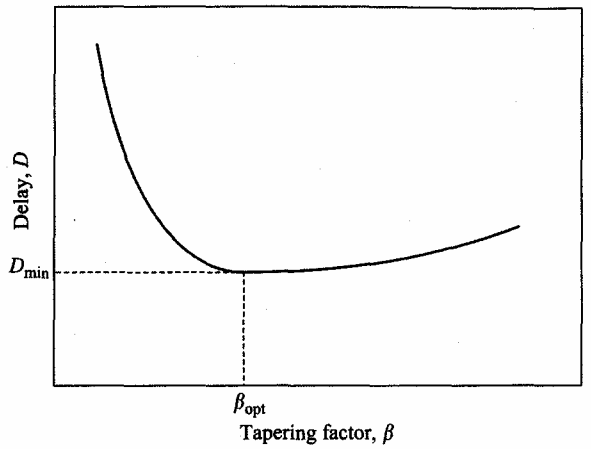

Figure 3.18. Dependence of the tapered buffer propagation delay on the tapering factor.

lower area and power at comparable suboptimal delays. The minimum delay of variable-taper buffers can be reduced and brought to within a few percent of the delay of a FT buffer by implementing the first few stages with a FT factor. Therefore, the optimal area-delay trade-offs are achieved in a FT buffer with the final one to two stages utilizing a larger tapering factor. This strategy is consistent with the observation in [50] that the buffer delay is reduced when the tapering factor of the final stage is increased. Power-delay product optimization of the tapered buffers is considered in [57]. Power-delay optimized buffers require fewer stages (and, consequently, a higher β) where the power-delay product improves by 15–35% as compared to delay optimal buffers.

Cherkauer and Friedman integrated these disparate approaches to CMOS tapered buffers into a unified design methodology, considering speed, area, power and reliability together [58]. Enhanced short channel expressions are presented for tapered buffer delay and power dissipation based on the alpha-power law short channel transistor model [25]. Analytic expressions of similar form are produced for the four performance metrics, permitting the combination of these metrics into different weighted optimization criteria. An important result is that short channel effects do not change the form of the propagation delay through a tapered buffer chain. The I–V model affects the absolute value of the delay, but does not change the process of delay optimization. Consequently, delay optimization schemes developed under long channel assumptions are also applicable to short channel devices.

A design methodology is presented in [59] for the optimal tapering of cascaded buffers in the presence of interconnect capacitance. Though interconnect capacitance is typically small in a full custom circuit, in those circuits based on

channel routing, physical proximity of the stages is not necessary and the capacitive interconnect load can often be substantial. Also, as shown in the paper, neglecting interconnect capacitance may result in suboptimal circuits even in those cases where the interconnect capacitance is small. A method, called constant capacitance-to-current ratio tapering (C^3RT), is based on maintaining the capacitive load to current drive ratio constant, such that the delay of each buffer stage also remains constant. Hence, in the presence of high interconnect loads, it is possible for the C^3RT methodology to produce a buffer in which a particular stage is smaller than the preceding stage, that is, with a tapering factor of less than unity between the stages. The importance of interconnect capacitance can vary from small to significant, depending upon the ratio of the interconnect capacitance to the total load capacitance at a node. The larger the interconnect load and the closer the load to the input of the buffer, the greater the impact on the circuit, as the input and output capacitances of the stages close to the load are larger and, therefore, the interconnect load is typically a proportionally smaller fraction of the total load. To demonstrate this methodology, a case study is conducted on a five-stage buffer driving a 5 pF load at 5 MHz. Implementation in a 2 μm technology is assumed; the interconnect capacitance is varied from 10 fF (the best case scenario in practice where the stages are physically abutted) to 500 fF (a severe case, possibly a gate array or standard cell circuit). C^3RT optimized buffers exhibit delay, area, and power advantages over FT buffers, as listed in Table 3.2. Note the steady *absolute* decrease in power of the C^3RT buffer as the interconnect capacitance increases. Although it may appear counterintuitive, this absolute decrease is accounted for by the reduction in the active area capacitance which offsets the increase in the interconnect capacitance such that the *total* capacitance is decreased. In general, the omission of interconnect capacitance leads to suboptimal designs in DSM CMOS circuits.

Table 3.2. C^3RT buffers as compared to FT buffers in the presence of interconnect capacitance [59].

C_{int} (fF)	*Propagation delay*			*Power dissipation*			*Active area*		
	FT (ns)	C^3RT (ns)	*Improvement* (%)	*FT* (mW)	C^3RT (mW)	*Improvement* (%)	*FT* (μm^2)	C^3RT (μm^2)	*Improvement* (%)
10	2.29	2.28	0.6	1.42	1.38	2.8	1900	1795	5.5
100	2.56	2.54	0.9	1.44	1.32	8.3	1900	1591	16.3
250	2.87	2.85	0.7	1.48	1.25	15.6	1900	1313	30.9
500	3.38	3.32	2.0	1.56	1.21	22.4	1900	1024	46.1

3.5. Physical Level

Coping with interconnect is a major problem in VLSI circuits. Interconnect affects system performance, power consumption and circuit area. The increasing importance of interconnect [5,67,68] is due to the classical scaling trend that while device feature size is *decreasing*, interconnect feature sizes are shrinking, and the die size is *increasing*, doubling every eight to ten years [3]. Thus, the wiring tends to become longer and the interconnect cross-section area smaller.

A problem resulting from this trend is increased *RC* interconnect impedances, degrading the delay of the gates. Consider a CMOS inverter driving an *RC* interconnect line as illustrated in Figure 3.19 (the driver output capacitance is omitted for the sake of simplicity). A first-order model of the delay of the circuit is [5]

$$T_{50\%} = 0.4R_{\text{int}}C_{\text{int}} + 0.7(R_{\text{tr}}C_{\text{int}} + R_{\text{tr}}C_{\text{L}} + R_{\text{int}}C_{\text{L}}) \qquad (3.5)$$

If the driver load is effectively capacitive, that is, the interconnect resistance R_{int} is much less than the effective driver resistance R_{tr}, the interconnect capacitance can be combined with the input capacitance of the gate to form a lumped load capacitance, permitting the circuit delay to be characterized by a lumped *RC* circuit delay, $0.7R_{\text{tr}}(C_{\text{int}} + C_{\text{L}})$. The signal propagation delay is due to the capacitive load being charged by the driver. Increasing the driver transistor width and consequently reducing R_{tr} decrease the circuit delay, trading off circuit power and area for higher speed. However, this behavior changes when R_{int} becomes comparable to R_{tr}. The delay cannot be reduced below $R_{\text{int}}(0.4C_{\text{int}} + 0.7C_{\text{L}})$. Note that the purely interconnect-related delay component, $0.7R_{\text{int}}C_{\text{int}}$, increases quadratically with interconnect length as both R_{int} and C_{int} are proportional to the length of the interconnect. This component of the total delay quickly becomes dominant in long interconnect. This interconnect delay component cannot be reduced significantly by making the interconnect wider as a decrease in wire resistance is offset by an increase in the

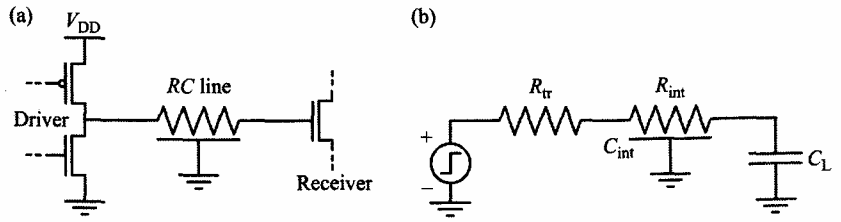

Figure 3.19. A model of a CMOS circuit driving an *RC* interconnect. (a) A circuit driving an *RC* line, (b) a corresponding simple model [5].

Table 3.3. Typical interconnect physical characteristics of a 0.18 μm CMOS process.

Layer characteristic	*Impedance*
Sheet resistance	$0.08\frac{\Omega}{\square}$
Metal pitch	0.5 μm
Metal thickness	0.5 μm
Interlayer spacing	0.5 μm
Metal width	0.25 μm
Line spacing	0.25 μm
Minimum width wire capacitance	0.2 fF/μm

wire capacitance. The increasing importance of interconnect delay is demonstrated by a 0.18 μm CMOS technology with aluminum interconnect. The physical characteristics of the first level metal interconnect typical for this technology are listed in Table 3.3. As an example, consider a local interconnect spanning several gates which is minimum width and 100 wire widths long, that is, 25 μm. The total resistance and capacitance are

$$R_{\text{int}} = 100\square \cdot 0.08\frac{\Omega}{\square} = 8\,\Omega \quad \text{and} \quad C_{\text{int}} = 0.2\frac{\text{fF}}{\mu\text{m}} \cdot 25\,\mu\text{m} = 5\,\text{fF}$$

respectively, yielding an interconnect delay of $t_{\text{int}} = 0.4 \cdot 8\,\Omega \cdot 5\,\text{fF} = 16\,\text{fs}$. Thus, interconnect delay is not important for local interconnect. As the dimensions scale with feature size, the local interconnect delay is expected to remain relatively insignificant with technology scaling [3]. The interconnect becomes significant, however, at the level of intermediate interconnect, where the length is approximately a half perimeter of a functional block, typically 3–4 mm. At such a length,

$$R_{\text{int}} = \frac{4\,\text{mm}}{0.25\,\mu\text{m}} \cdot 0.08\frac{\Omega}{\square} = 1280\,\Omega \quad \text{and} \quad C_{\text{int}} = 0.2\frac{\text{fF}}{\mu\text{m}} \cdot 4000\,\mu\text{m} = 0.8\,\text{pF}$$

and the interconnect delay is $t_{\text{int}} = 0.4 \cdot 1280\,\Omega \cdot 0.8\,\text{pF} = 0.4\,\text{ns}$. This delay exceeds typical gate delays in a 0.18 μm CMOS technology and is a significant fraction of the minimum clock period of a high-performance circuit (1–3 ns). Global interconnections can be as long as half a perimeter of the die (and longer for bus structures). Assuming a die with $1 \times 1\,\text{cm}^2$ dimensions, a moderate size for current fabrication capabilities, a half perimeter line would have the following parameters,

$$R_{\text{int}} = \frac{20\,\text{mm}}{0.25\,\mu\text{m}} \cdot 0.08\frac{\Omega}{\square} = 6400\,\Omega \quad \text{and} \quad C_{\text{int}} = 0.2\frac{\text{fF}}{\mu\text{m}} \cdot 20,000\,\mu\text{m} = 4\,\text{pF}$$

and the interconnect delay would equal $t_{int} = 0.4 \cdot 6400\,\Omega \cdot 4\,pF \simeq 10\,ns$. A 10 ns path delay (equal to a 100 MHz clock frequency) exceeds the clock period of many circuits, dwarfing the delay of the logic elements. The delay of global interconnect is, therefore, a central topic of concern in high-performance VLSI circuits.

Widening a uniform line has a marginal impact on the overall wire delay. These delay estimations are based on the thin first layer metal. The thickness of the upper metal layers is typically increased to provide less resistive interconnections. The pitch and interlayer spacing of the top layers are also wider, therefore, the line capacitance does not significantly change as compared to the first metal layer. The impedance characteristics of the metal lines in the upper layers are about an order of magnitude lower than the impedance characteristics of the lower metal levels. While mitigating the problem, the thick upper metal layers do not solve the overall problem as global line impedances severely limit circuit performance.

An effective strategy for reducing long interconnect delay is inserting intermediate buffers, typically called repeaters [5]. Repeaters circumvent the quadratic increase in interconnect delay by partitioning the interconnect line into smaller and approximately equal sections, as shown in Figure 3.20. The sum of the section delays is smaller than the delay of the original path since the delay of each section is quadratically reduced. The decreased interconnect delay is partially offset by the added delays of the inserted repeaters.

A number of repeater insertion methods has been proposed [69–74]. Bakoglu presents a method based on characterizing the repeaters by the input capacitance and the effective output resistance deduced from the repeater size [5,67]. The minimum delay of the resulting *RC* circuit is achieved when the repeater section delay equals the wire segment delay. Another method has been described by Wu and Shiau [69]; in this method a linearized form of the Shichman–Hodges

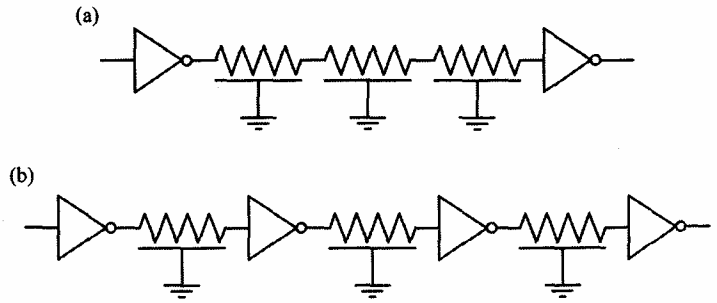

Figure 3.20. Repeater insertion. (a) The original interconnect line, (b) the interconnect line with inserted repeaters.

equations is used to determine the points of repeater insertion. Nekili and Savaria have introduced the concept of parallel regeneration in which precharge circuitry is added to the repeaters to decrease the evaluation time [70,71]. This technique reduces the number of repeaters, but requires extra area and a precharge signal to maintain correct operation. A mathematical treatment of repeater optimization with and without area constraints is described by Dhar and Franklin [72]. Elegant solutions are obtained; however, a simple resistor–capacitor model is used to characterize the repeaters and no closed form solutions are described.

A repeater design methodology is presented by Adler and Friedman [74]. A timing model of a CMOS inverter driving an *RC* load based on the alpha-power law transistor model is used to account for short channel velocity saturation effects. The closed form expression for the overall signal delay of a uniform repeater chain driving a large distributed *RC* load is described. The analytical delay estimates are within 16% of SPICE simulations of representative long interconnect loads. A comparison of uniform and tapered-buffer repeaters is also described. Uniform repeater are found to outperform tapered-buffer repeaters when driving even relatively low resistive *RC* loads. Power issues in the repeater design process are also considered. An analytic expression for the short-circuit power in a repeater chain is described which exhibits a maximum error of 15% as compared to SPICE simulations within the primary regions of interest. It is shown that short-circuit power can represent up to 20% of the total dynamic power dissipation. It also shown that a 4% increase in delay over the minimum delay of a repeater chain can be traded off for a 40% savings in area and 15% savings in power.

3.6. Process Level

Changing technology is typically not a design or trade-off option, however, it is sometimes feasible to choose different semiconductor manufacturers or specialized technologies. Two technologies, both described as "0.25 μm CMOS technologies," can be substantially different. The notion of "0.25 μm" refers to the smallest resolvable feature size in a process, typically the transistor channel length L. While L is the primary parameter controlling the transistor current drive, the channel length is just one of the many dozens of design rules that characterize a process. As the interconnect system (with related contacts and vias) occupy an increasing portion of the total die area, these design rules are of great significance in the overall circuit performance and area characteristics. The effects of technology scaling are discussed in Section 3.6.1. The trade-offs involved in the choice of threshold voltage and power supply voltage are discussed in Sections 3.6.2 and 3.6.3, respectively. The impact of improved materials on design trade-offs is considered in Section 3.6.4.

3.6.1. Scaling

Shrinking dimensions (i.e. length, height and width) directly improve circuit area and power. The circuit area decreases rapidly (quadratically, assuming linear scaling). The parasitic capacitance of the transistors and interconnect is reduced; therefore, the power consumed by the circuit is also reduced. These gains in area and power can be traded for increased speed. The effects of changes in the vertical dimensions differ depending upon the circuit component. Thinner gate oxides translate to increased transistor transconductance and therefore higher speed, which can be effectively traded for lower power by lowering the power supply voltage. Thicker intermetal oxide reduces the parasitic wiring capacitance, leading to shorter *RC* delays (i.e. higher speed) and lower cross-coupling noise. Increased metal thickness lowers the sheet resistance of the metal layers. The wiring is denser and the total die area is reduced; however, there is also an increase in interwire capacitive coupling and noise.

3.6.2. Threshold Voltage

The control of the threshold voltage V_T is one of the primary issues at the process level. A higher V_T means higher noise margins and lower leakage currents when the transistors are cut off. However, the leakage current contribution to the total power dissipated in most low power systems is typically small and the coupling noise can be proportionally lowered as the supply voltage is decreased. The relative magnitude of the capacitive cross-coupling noise to the signal level is determined by the circuit geometries. The magnitude of the switched current is decreased as V_{DD} is lowered; therefore, the *IR*, inductive and simultaneous switching components of noise also scale. A lower V_T, however, enhances the transistor current drive, permitting the circuit to operate faster or, alternatively, providing a substantial power saving by lowering the supply voltage without a significant increase in the logic delay. Threshold voltage process variations set a limit on the maximum V_T reduction. If a statistical deviation of V_T in just one transistor is above some critical value, an entire multimillion transistor IC can be lost. As more and more transistors are integrated onto one die, tens of millions transistor ICs have become commonplace, making tight control of the threshold voltage ever more challenging.

3.6.3. Power Supply

Power supply voltage V_{DD}, strictly speaking, is not a process parameter; however, the power supply voltage is effectively defined by the process technology. The delay rises dramatically as V_T approaches V_{DD}, see Figure 3.21. Increasing V_{DD} above several V_T is often not practical due to a small increase in speed at the expense of a quadratic increase in the power consumption [61–63]. Due to carrier velocity saturation effects, the transistor current increases

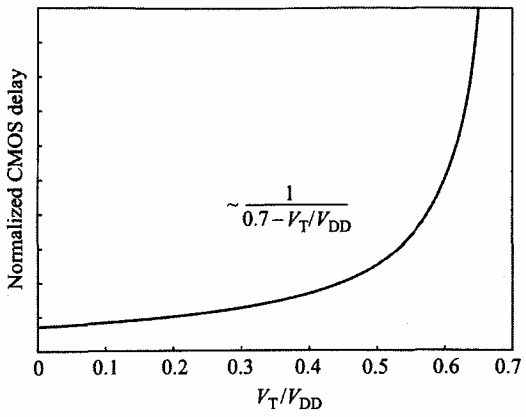

Figure 3.21. Dependence of transistor switching delay on threshold voltage V_T [75].

almost linearly with voltage and no significant speed benefits are attained with further voltage increases. Furthermore, reliability issues such as gate oxide breakdown, hot electron injection, carrier multiplication and electromigration place an upper limit on the magnitude of the power supply and current density.

3.6.4. Improved Interconnect and Dielectric Materials

The introduction of copper as a low resistance interconnect material and low dielectric constant materials as interlayer isolators is a relatively recent phenomenon. The immediate effect on existing circuits is higher operating speeds due to reduced interconnect impedances. Though copper-based CMOS processes cost more than conventional aluminum-based CMOS processes at the present time, once matured, the cost should drop below the cost of aluminum interconnect. By some estimates, a layer of copper interconnect costs 20% less than a comparable layer of aluminum [76]. Much greater speed improvements are expected for those ICs originally designed for copper interconnect processes. The higher wiring capacity of a copper metal layer as compared to an aluminum layer will also result in a substantial decrease in die area and/or the use of less interconnect layers, further decreasing overall fabrication costs.

3.7. Future Trends

Semiconductor fabrication technology has reached the point where integrating an entire large system on a single chip is possible. The increased level of integration will, however, exacerbate design productivity issues, greatly affecting design time and cost. Designing an SoC at the transistor level is considered impractical from both a cost and a design time point of view. A large fraction

of an SoC consists of functional cores either reused from previous circuits or automatically synthesized from a high level description (such as RTL). Another reason for design reuse is that the design of certain functional units of a system may not be within a particular company's areas of expertise. The circuit design information, therefore, must be purchased from other IC design houses, raising complicated intellectual property (IP) issues. The necessary business and legal framework is required to support the use of expertise accumulated from extensive IP outsourcing. The current CMOS circuit design approach of choosing a design style for a specific circuit is likely to continue: noncritical regular circuit structures are likely to be automatically synthesized from high level descriptions while performance critical parts of the circuits are likely to be customized or reused from previous high-performance circuits. Extensive reuse of high-speed functional blocks will likely become a common practice even among the more aggressive IC design companies.

Furthermore, a move to system-scale integration has produced qualitatively new issues. SoCs are heterogeneous in nature. These systems integrate a combination of circuit functionality: digital logic, signal processing and conditioning, memory, communications, analog signal processing. Such diverse functionality necessitates a number of heterogeneous circuits being designed into an integral system and fabricated within a single semiconductor technological process: digital circuitry for control and computation; SRAM, FLASH or embedded DRAM memory for code and data storage; RF for communications; sensors, analog and mixed-signal for interfacing to physical signals, high-speed buses for communication among functional units. This diversity is presenting the circuit design industry with formidable challenges and design trade-offs.

Since the reused cores have not been specifically designed for a target SoC, these circuits can place different constraints on system-wide signals such as the clock and power distribution; additionally, protocols for reset, test and data exchange interfaces can be quite different. Multiple clock domains and asynchronous intercore communication may emerge as viable solutions for multiple core integration. Significant design effort will be required to properly integrate the reused cores into a cohesive SoC without drastically affecting performance. Detailed specifications of the cores are required: circuit delay versus power supply voltage dependence, power consumed versus power supply level dependence, power supply tolerance, peak current, maximum inductance of the power supply bus, clock signal load, clock duty cycle, period, and rise/fall time constraints are just a few of the many system level trade-offs which will need to be integrated into core-based design methodologies. Multiple cores share many system-wide signals; thus, system level trade-offs will have to be integrated into core-based SoC design methodologies. These compatibility and specification issues will require an entirely new set of standards for circuit reuse. Furthermore, the reuse of analog and mixed-signal circuits is far more

difficult than that of digital circuits due to the higher sensitivity to input and output load and parasitic impedances within the linear circuits. New problems affecting the proper operation of the analog circuits have developed such as core-to-core substrate coupling. Substrate coupling remains an open design problem which must be surmounted.

The technical and economic necessity of design reuse will likely lead to a new design paradigm – design for reusability. With a goal of ease of design effort, reuse being a principal design merit, design trade-offs will need to be made at every level of design to render the circuit more easily reusable to a wider range of applications, circuit environments and fabrication processes at the expense of performance, area and power. Multiple versions of the same functional core may need to be individually optimized with each version tailored to a specific application.

While system functionality has been growing at an exponential rate according to Moore's law, the cost of fabricating state-of-the-art ICs has remained relatively flat [3] and the cost of fabricating the same IC drops with time as it is implemented in the newer scaled processes with finer feature sizes. The market has proven to be highly elastic, that is, the cost reduction of semiconductor products has greatly expanded the consumption of ICs. The history of the fast growth of the semiconductor market and projections for the next few years are shown in Figure 3.22. To a great part, this growth has been primarily due to the boom in personal computers. The next major opportunities for high sustained growth in the semiconductor market is internet infrastructure

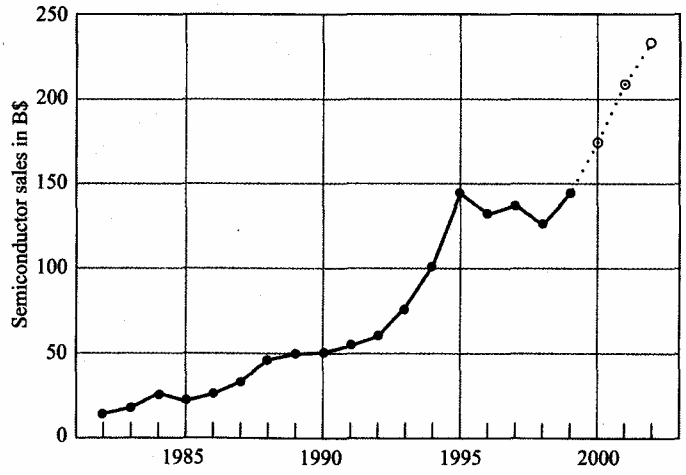

Figure 3.22. History and forecast of semiconductor market sales (with the forecast represented by the dotted line). Data from Semiconductor Industry Association [77].

products, and personal information and wireless communication appliances. While potentially lucrative, personal appliances are a consumer market with inherently tough competition, thin margins, and tight and unforgiving product windows. Low risk design strategies that consider multiple trade-offs at all levels of design abstraction will be required to produce commercial success in a market of commodities.

Summarizing, the following trends will shape the immediate future of CMOS VLSI circuits. As CMOS fabrication technologies are continually scaled at a breathtaking pace and as SoC integration emerges, the process of developing VLSI circuits will become increasingly design productivity constrained rather than technology constrained. High level design capture and design reuse will become important solutions to increase design productivity. Incremental design approaches and design standardization will also be instrumental for effectively reusing existing circuits. Design cost and time will likely dominate decision making in the development of next generation products [78]. Cost, as always, will be crucial in making design trade-offs in semiconductor products that target commodity markets.

Glossary

The following notations and abbreviations are used in this chapter.

Acronyms used in terminology pertaining to VLSI circuits:

IC	integrated circuit
CMOS	complementary metal oxide semiconductor
DSM	deep submicrometer
CPU	central processing unit
RAM	random access memory
DRAM	dynamic random access memory
SRAM	static random access memory
RF	radio frequency
IP	intellectual property
VTC	voltage transfer characteristic
RTL	register transfer level
BIST	built-in-self test
SoC	system on a chip
RE	recurring engineering
NRE	non-recurring engineering

Circuit-specific parameters:

P_{dynamic}	dynamic power
P_{sc}	short-circuit power
P_{leakage}	power dissipated due to the leakage current

$I_{leakage}$	transistor leakage current when operating in the cut-off mode
I_{out}	gate drive current
I_{peak}	peak magnitude of short-circuit current
I_{D0}	drain current at $V_{GS} = V_{DS} = V_{DD}$
V_{DD}	power supply voltage
V_{Tn}	N-channel transistor threshold voltage
V_{Tp}	P-channel transistor threshold voltage (negative for an enhancement mode device)
V_{D0}	drain saturation voltage at $V_{GS} = V_{DD}$
R_{int}	interconnect resistance
R_{tr}	effective transistor "on" resistance
C_{int}	interconnect capacitance
C_{ox}	gate oxide capacitance per unit area
C_0	input capacitance of a minimum size inverter
C_{gate}	transistor gate capacitance
C_L	load capacitance
μ_n	electron mobility in n-type transistor
μ_p	hole mobility in p-type transistor
W_n	N-channel transistor width
W_p	P-channel transistor width
f	circuit clock frequency
f_{switch}	average charge/discharge cycle frequency
t_{base}	duration of short-circuit current
s	slew rate of a ramp-shaped signal
β	multistage buffer tapering factor
β_{opt}	delay optimal tapering factor
N	number of stages in a tapered buffer
N_{opt}	delay optimal number of stages in a tapered buffer
K	transistor gain factor
A	total active area of a CMOS circuit

References

[1] G. E. Moore, "Cramming more components onto integrated circuits", *Electronics*, pp. 114–117, 19 April 1965.

[2] G. E. Moore, "Progress in Digital Integrated Electronics", *Proceedings of the IEEE International Electron Devices Meeting*, pp. 11–13, December 1975.

[3] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors, 1998 Update.*

[4] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[5] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.

[6] C. H. Stapper, "The effects of wafer to wafer defect density variations on integrated circuit defect and fault distributions", *IBM Journal of Research and Development*, vol. 29, no. 1, pp. 87–97, January 1985.

[7] E. A. Bretz, "Test & measurement", *IEEE Spectrum*, pp. 75–79, January 2000.

[8] J. R. Black, "Electromigration – a brief survey and some recent results", *IEEE Transactions on Electron Devices*, vol. ED-16, no. 4, pp. 338–347, April 1969.

[9] Y.-W. Yi, K. Ihara, M. Saitoh and N. Mikoshiba, "Electromigration-induced integration limits on the future ULSI's and the beneficial effects of lower operation temperatures", *IEEE Transactions on Electron Devices*, vol. 42, no. 4, pp. 683–688, April 1995.

[10] I. Catt, "Crosstalk (noise) in digital systems", *IEEE Transactions on Electronic Computers*, vol. EC-16, no. 6, pp. 743–763, December 1967.

[11] M. Shoji, *Theory of CMOS Digital Circuits and Circuit Failures*, Princeton University Press, 1992.

[12] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's", *IEEE Transactions on Electron Devices*, vol. ED-40, no. 1, pp. 118–124, January 1993.

[13] M. Shoji, *High-Speed Digital Circuits*, Addison-Wesley, 1996.

[14] W. S. Song and L. A. Glasser, "Power distribution techniques for VLSI circuits", *IEEE Journal of Solid-State Circuits*, vol. SC-21, no. 1, pp. 150–156, February 1986.

[15] S. R. Vemuru, "Accurate simultaneous switching noise estimation including velocity-saturation effects", *IEEE Transactions on Components, Packaging, and Manufacturing Technology – Part B*, vol. 19, no. 2, pp. 344–349, May 1996.

[16] P. Larsson, "di/dt noise in CMOS integrated circuits", *Analog Integrated Circuits and Signal Processing*, vol. 14, no. 1/2, pp. 113–129, September 1997.

[17] Y. I. Ismail, E. G. Friedman and J. L. Neves, "Figures of merit to characterize the importance of on-chip Inductance", *IEEE Transactions on VLSI Systems*, vol. 7, no. 4, pp. 83–97, December 1999.

[18] K. T. Tang and E. G. Friedman, "Interconnect coupling noise in CMOS VLSI circuits", *Proceedings of the ACM/IEEE International Symposium on Physical Design*, pp. 48–53, April 1999.

[19] L. Bisduonis, S. Nikolaidis, O. Koufopavlou and C. E. Goutis, "Modeling the CMOS short-circuit power dissipation", *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 4.469–4.472, May 1966.

[20] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits", *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 4, pp. 468–473, August 1984.

[21] S. R. Vemuru and N. Scheinberg, "Short-circuit power dissipation estimation for CMOS logic gates", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 41, no. 11, pp. 762–766, November 1994.

[22] A. M. Hill and S.-M. Kang, "Statistical estimation of short-circuit power in VLSI design", *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 4.105–4.108, May 1996.

[23] A. Hirata, H. Onodera and K. Tamaru, "Estimation of short-circuit power dissipation for static CMOS gates", *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science*, vol. E79-A, no. 3, pp. 304–311, March 1996.

[24] V. Adler and E. G. Friedman, "Delay and power expressions for a CMOS inverter driving a resistive–capacitive load", *Analog Integrated Circuits and Signal Processing*, vol. 14, no. 1/2, pp. 29–39, September 1997.

[25] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas", *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, April 1990.

[26] M. J. S. Smith, *Application-Specific Integrated Circuits*, Addison-Wesley, 1997.

[27] A. P. Chandrakasan, S. Sheng and R. W. Brodersen, "Low power CMOS digital design", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–483, April 1992.

[28] E. G. Friedman and J. H. Mulligan, Jr., "Clock frequency and latency in synchronous digital systems", *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 930–934, April 1991.

[29] E. G. Friedman and J. H. Mulligan, Jr., "Pipelining of high performance synchronous digital systems", *International Journal of Electronics*, vol. 70, no. 5, pp. 917–935, May 1991.

[30] E. G. Friedman and J. H. Mulligan, Jr., "Pipelining and clocking of high performance synchronous digital systems", in: M. A. Bayoumi and E. E. Swartzlander, Jr. (eds), *VLSI Signal Processing Technology*, Kluwer Academic Publishers, ch. 4, pp. 97–133, 1994.

[31] C. M. Lee and H. Soukup, "An algorithm for CMOS timing and area optimization", *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 5, pp. 781–787, October 1984.

[32] E. T. Lewis, "Optimization of device area and overall delay for CMOS VLSI designs", *Proceedings of the IEEE*, vol. 72, no. 5, pp. 670–689, June 1984.

[33] J. Yuan and C. Svensson, "Principle of CMOS circuit power-delay optimization with transistor sizing", *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 637–640, May 1996.

[34] C. Tretz and C. Zukowski, "CMOS transistor sizing minimization of energy-delay product", *Proceedings of the IEEE Great Lakes Symposium on VLSI*, pp. 168–173, March 1996.

[35] M. Borah, R. M. Owens and M. J. Irwin, "Transistor sizing for low power CMOS circuits", *IEEE Transactions on Computer-Aided Design*, vol. 15, no. 6, pp. 665–671, June 1996.

[36] R. Rogenmoser and H. Kaeslin, "The impact of transistor sizing on power efficiency in submicron CMOS circuits", *IEEE Journal of Solid-State Circuits*, vol. 32, no. 7, pp. 1142–1145, July 1997.

[37] H. Y. Chen and S. M. Kang, "A new circuit optimization technique for high performance CMOS circuits", *IEEE Transactions on Computer-Aided Design*, vol. 10, no. 5, pp. 670–676, May 1991.

[38] J. P. Fishburn and S. Taneja, "Transistor sizing for high performance and low power", *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 591–594, May 1997.

[39] A. R. Conn, P. K. Coulman, R. A. Haring, et al., "Optimization of custom MOS circuits by transistor sizing", *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 174–180, November 1996.

[40] M. Tachibana, S. Kurosawa, R. Nojima, N. Kojima, M. Yamada, T. Mitsuhashi and N. Goto, "Power and area minimization by reorganizing CMOS complex gates", *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences*, vol. E79-A, no. 3, pp. 312–319, March 1996.

[41] T. Xiao and M. Marek-Sadowska, "Crosstalk reduction by transistor sizing", *Proceedings of the Asia and Pacific Design Automation Conference*, pp. 137–140, January 1999.

[42] A. Vittal, L. H. Chen, M. Marek-Sadowska, K.-P. Wang, S. Yang, "Crosstalk in VLSI interconnection", *IEEE Transactions on Computer-Aided Design*, vol. 18, no. 12, pp. 1817–1824, December 1999.

[43] A. Dasgupta and R. Karri, "Hot-carrier reliability enhancement via input reordering and transistor sizing", *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 819–824, June 1996.

[44] J. Cong, L. He, C.-K. Koh and P. H. Madden "Performance optimization of VLSI interconnect layout", *Integration, The VLSI Journal*, vol. 21, no. 1/2, pp. 1–94, November 1996.

[45] L. S. Heusler and W. Fichtner, "Transistor sizing for large combinational digital CMOS circuits", *Integration, The VLSI Journal*, vol. 10, no. 2, pp. 155–168, January 1991.

[46] H. C. Lin and L. W. Linholm, "An optimized output stage for MOS integrated circuits", *IEEE Journal of Solid-State Circuits*, vol. SC-10, no. 2, pp. 106–109, April 1975.

[47] R. C. Jaeger, "Comments on 'An Optimized Output Stage for MOS Integrated Circuits' ", *IEEE Journal of Solid-State Circuits*, vol. SC-10, no. 3, pp. 185–186, June 1975.

[48] A. Kanuma, "CMOS circuit optimization", *Solid-State Electronics*, vol. 26, no. 1, pp. 47–58, January 1983.

[49] M. Nemes, "Driving large capacitances in MOS LSI systems", *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 1, pp. 159–161, February 1984.

[50] N. Hedenstierna and K. O. Jeppson, "CMOS circuit speed and buffer optimization", *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, no. 2, pp. 270–281, March 1987.

[51] N. C. Li, G. L. Haviland and A. A. Tuszynski, "CMOS tapered buffer", *IEEE Journal of Solid-State Circuits*, vol. 25, no. 4, pp. 1005–1008, August 1990.

[52] S. R. Vemuru and A. R. Thorbjornsen, "Variable-taper CMOS buffer", *IEEE Journal of Solid-State Circuits*, vol. 26, no. 9, pp. 1265–1269, September 1991.

[53] C. Prunty and L. Gal, "Optimum tapered buffer", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 1, pp. 118–119, January 1992.

[54] T. Sakurai, "A unified theory for mixed CMOS/BiCMOS buffer optimization", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 7, pp. 1014–1019, July 1992.

[55] N. Hedenstierna and K. O. Jeppson, "Comments on the optimum CMOS tapered buffer problem", *IEEE Journal of Solid-State Circuits*, vol. 29, no. 2, pp. 155–158, February 1994.

[56] L. Gal, "Reply to comments on the optimum CMOS tapered buffer problem", *IEEE Journal of Solid-State Circuits*, vol. 29, no. 2, pp. 158–159, February 1994.

[57] J.-S. Choi and K. Lee, "Design of CMOS tapered buffer for minimum power-delay product", *IEEE Journal of Solid-State Circuits*, vol. 29, no. 9, pp. 1142–1145, September 1994.

[58] B. S. Cherkauer and E. G. Friedman, "A unified design methodology for CMOS tapered buffers", *IEEE Transactions on VLSI Systems*, vol. 3, no. 1, pp. 99–111, March 1995.

[59] B. S. Cherkauer and E. G. Friedman, "Design of tapered buffers with local interconnect capacitance", *IEEE Journal of Solid-State Circuits*, vol. 30, no. 2, pp. 151–155, February 1995.

[60] B. S. Carlson and S.-J. Lee, "Delay optimization of digital CMOS VLSI circuits by transistor reordering", *IEEE Transactions on Computer-Aided Design*, vol. 14, no. 10, pp. 1183–1192, October 1995.

[61] M. Kakumu and M. Kinugawa, "Power supply voltage impact on circuit performance for half and lower submicrometer CMOS LSI", *IEEE Transactions on Electron Devices*, vol. 37, no. 8, pp. 1902–1908, August 1990.

[62] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages", *IEEE Journal of Solid-State Circuits*, vol. 28, no. 1, pp. 10–17, January 1993.

[63] K. Chen and C. Hu, "Performance and V_{DD} scaling in deep submicrometer CMOS", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 10, pp. 1586–1589, October 1998.

[64] F. Mu and C. Svensson, "Analysis and optimization of a uniform long wire and driver", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 46, no. 9, pp. 1086–1100, September 1999.

[65] C. Nagendra, M. J. Irwin and R. M. Owens, "Area-time-power tradeoffs in parallel adders", *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 43, no. 10, pp. 689–702, October 1996.

[66] C. Nagendra, R. M. Owens and M. J. Irwin, "Power-delay characteristics of CMOS adders," *IEEE Transactions on VLSI Systems*, vol. 2, no. 3, pp. 377–381, September 1994.

[67] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI", *IEEE Transactions on Electron Devices*, vol. ED-32, no. 5, pp. 903–909, May 1985.

[68] S. Bothra, B. Rogers, M. Kellam and C. M. Osburn, "Analysis of the effects of scaling on interconnect delay in ULSI circuits", *IEEE Transactions on Electron Devices*, vol. 40, no. 3, pp. 591–597, March 1993.

[69] C. Y. Wu and M. Shiau, "Delay models and speed improvement techniques for *RC* tree interconnections among small-geometry CMOS inverters", *IEEE Journal of Solid-State Circuits*, vol. 25, no. 10, pp. 1247–1256, October 1990.

[70] M. Nekili and Y. Savaria, "Optimal methods of driving interconnections in VLSI circuits", *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 21–23, May 1992.

[71] M. Nekili and Y. Savaria, "Parallel regeneration of interconnections in VLSI & ULSI circuits", *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 2023–2026, May 1993.

[72] S. Dhar and M. A. Franklin, "Optimum buffer circuits for driving long uniform lines", *IEEE Journal of Solid-State Circuits*, vol. 26, no. 1, pp. 32–40, January 1991.

[73] C. J. Alpert, "Wire segmenting for improved buffer insertion", *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 588–593, June 1997.

[74] V. Adler and E. G. Friedman, "Repeater design to reduce delay and power in resistive interconnect", *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 5, pp. 607–616, May 1998.

[75] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998.

[76] A. E. Braun, "Aluminum persists as copper age dawns", *Semiconductor International*, pp. 58–66, August 1999.

[77] Semiconductor Industry Association, *http://www.semichips.org/stats*

[78] H. Chang, L. Cooke, M. Hunt, G. Martin, A. McNelly and L. Todd, *Surviving the SOC Revolution – A Guide to Platform-Based Design*, Kluwer Academic Publishers, 1999.