On-Chip Communication Architectures

System on Chip Interconnect

Sudeep Pasricha – Nikil Dutt

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

ELSEVIER

Morgan Kaufmann is an imprint of Elsevier

MORGAN KAUFMANN PUBLISHERS

CHAPTER

11

Physical Design Trends for Interconnects

Ioannis Savidis and Eby G. Friedman

Over the past 10 years, the source of the critical signal delays has undergone a major transition. With the scaling of active device feature sizes into the deep submicrometer (DSM) regime, the on-chip interconnect has become the primary bottleneck in signal flow within high complexity, high speed integrated circuits (ICs). The smaller feature size in DSM technology nodes reduces the delay of the active devices; however, the effect on delay due to the passive interconnects has increased rapidly, as described by the 2005 International Technology Roadmap for Semiconductors (ITRS) [1]. The transition from an IC dominated by gate delays for feature sizes greater than 250μm to where the interconnects are the primary source of delay is graphically illustrated in Fig. 11.1. As noted in the figure, the disparity between the relative delay of the interconnects and the active devices is exacerbated in each successive technology node [1]. The local wire delay decreases with feature size due to a reduction in the distance among the active devices. Special attention must, however, be placed on the global lines, since the overall speed of current ICs is most often limited by the long distance global interconnects [2–6].

Low power dissipation has become a critical design criterion. With shrinking feature size and larger chip die dimensions, the sheer number of interconnects has increased exponentially [1, 7]. Interconnect capacitance often dominates the total gate load [8]; therefore, a large portion of the total transient power is dissipated by these on-chip lines. This characteristic is particularly true for those long interconnects that distribute the clock signals, where as much as 40–50% of the total power of an IC can be dissipated [1]. The gains achieved in performance, however, are often accompanied by an increase in power dissipation. As an example, additional interconnect layers enhance circuit speed at the expense of higher power consumption due to the larger interconnect capacitance. Considering both power consumption and propagation delay, interconnect design has become a dominant issue in high speed ICs.

In addition to an increase in interconnect power consumption, the design complexity of the various interconnect networks is continuing to present a significant

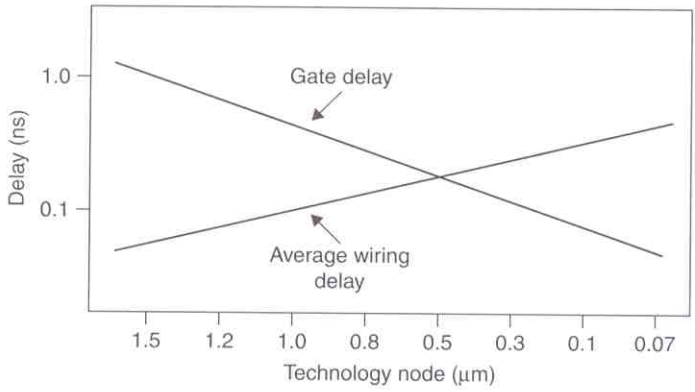

FIGURE 11.1

Comparison of interconnect (wiring) delay to gate delay

challenge. Semiconductor companies are currently building microprocessors with over 500 million transistors, and this number is increasing [9, 10]. In addition to the interconnects among the various on-chip devices, the clock and power distribution networks both require significant metal resources. Accurately modeling the clock, power, and signal nets is a difficult task; optimally allocating metal to properly design these networks presents an even greater challenge.

A significant factor that contributes to the complexity of interconnect modeling is the increase in the sheer number of lines. *RC* interconnect models are not sufficiently accurate to properly capture signal propagation in lines with fast transition times. Including inductance (*L*) in the *RC* model has become a necessary modification. *RLC* models are therefore becoming increasingly common at the expense of greater computational cost. On-chip clock rates also contribute to the complexity of the interconnect modeling process. Until recently, the semiconductor industry had been focusing primarily on faster clock rates. However, over the past couple of years, there has been a shift in this industrial paradigm toward multi-core processing. With this shift, operational frequencies have been somewhat reduced while throughput improved at the expense of increased die area. Inductance may be ignored at these lower operating frequencies under certain conditions. The length of the line, the cross-sectional area of the line, the signal waveform properties, and the available current return paths must all be considered in determining whether to include inductance in the interconnect model when operating at mid-range frequencies ranging from 1 to 3 Gigahertz [13]. For these reasons, interconnect modeling and metal allocation have become a complex design problem.

The intention of this chapter is to provide insight into the complexity of the on-chip interconnect design process, particularly in high performance applications. The effects of scaling interconnect in the DSM regime is discussed in Section 11.1. Low power, high speed circuit design techniques in support of global signaling are presented in Section 11.2. Once an understanding of current interconnect models and analysis techniques has been established, application of these models and techniques to global power and clock distribution networks is examined in

Sections 11.3 and 11.4, respectively. A brief introduction to 3-D interconnects is provided in Section 11.5 as a glimpse into future interconnect technologies. Finally, some concluding remarks are offered in Section 11.6.

11.1 DSM INTERCONNECT DESIGN

The steady decrease in the feature size of semiconductor devices has enhanced circuit complexity and performance. Scaling of the lateral dimensions in planar devices, such as MOS transistors, has produced improvements in the area, power, and speed of the devices [12]. The specific characteristics of interconnects, particularly the global lines, are degraded by current scaling trends. The power consumption and signal propagation delays of these long resistive lines have increased. Inductive effects must now be considered for possible inclusion in the interconnect models. For these reasons, accurate on-chip interconnect models are required to determine the signal characteristics and design requirements of high speed DSM interconnect.

Accurate interconnect models are used in both the design and analysis of ICs. Local interconnects can be neglected if the line capacitance is much smaller than the load capacitance. As the line capacitance becomes comparable to the load capacitance, the local line can be modeled as a single lumped capacitor, as depicted in Fig. 11.2(a) [13]. Since these lines are short, the signal propagation delay is negligible as compared to the gate delay [14]. In addition, these short interconnects have a negligible line resistance, minimally degrading the signal propagation characteristics.

Long interconnects suffer from resistive effects that impede signal propagation. The signal delay through these lines can be comparable to or exceed the gate delay. A lumped capacitor model of a global line is typically highly inaccurate with errors often exceeding 30%. A more elaborate model that includes the resistive effects of long lines is therefore better suited to represent global interconnects. As shown in Fig. 11.2(b), the simplest *RC* model is a lumped model, which does not consider the distributed nature of the impedance of a long global line. *RC* lines are often divided into sections of distributed impedances, capturing the distributed nature of the line impedance [15–17]. Each subsection is modeled as an equivalent *RC* circuit. Two common circuits often used to model long interconnect lines are T and Π circuits, depicted in Fig. 11.3(a) and 11.3(b), respectively. The accuracy of these models depends upon the number of sections used to represent the interconnect lines, with five or six sections typically being more than sufficient to accurately model a line. A T or Π equivalent circuit can model an *RC* line with less than 3% relative error in the delay, even in the case where only three ladder stages are used [15]. An *RC* model is usually adequate for low to medium operating frequencies; however, at frequencies exceeding a GHz, an *RC* model is often inadequate to accurately characterize the waveform properties along a wide interconnect line. An *RLC* model is often necessary to accurately characterize these interconnects.

Long global lines are usually much wider than local lines, exhibiting a lower resistance per unit length. With the reduction in line resistance and the higher clock frequencies, the line inductance contributes to the signal propagation

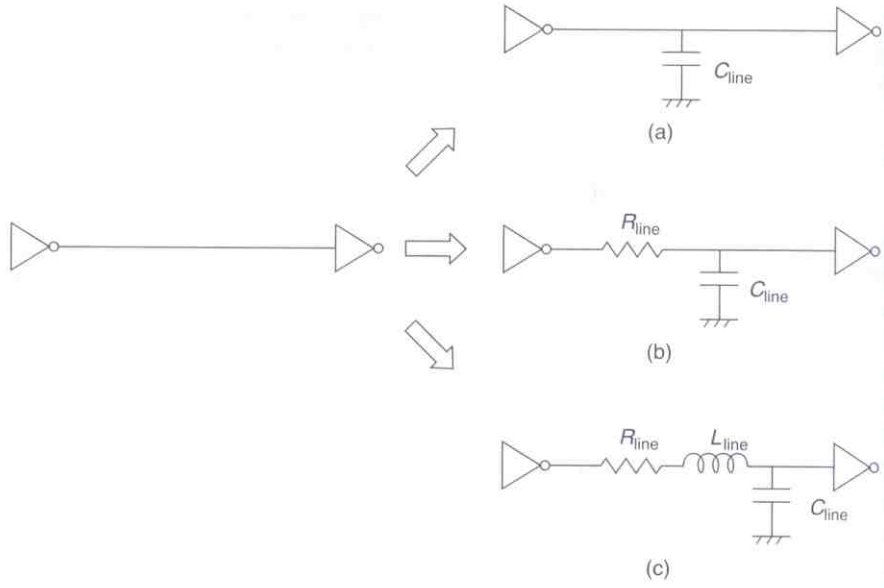

FIGURE 11.2

Lumped interconnect models: (a) *C* model, (b) *RC* model, (c) *RLC* model [13]
Figure reused with kind permission from Springer Science and Business Media

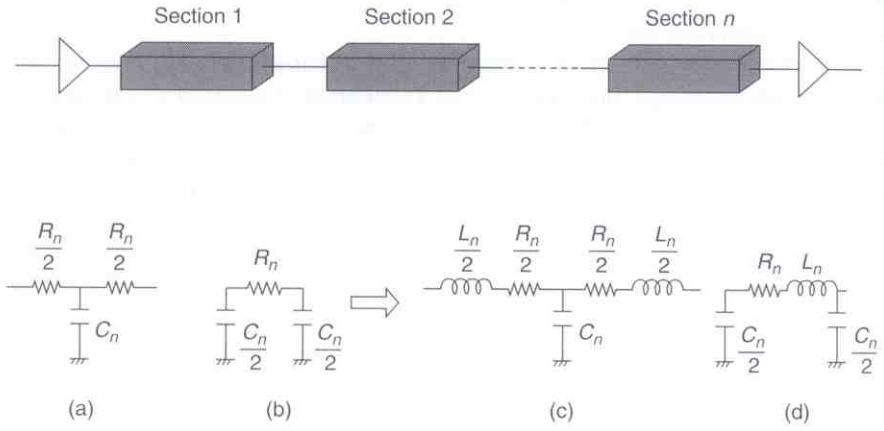

FIGURE 11.3

Distributed interconnect models: (a) *RC* T model, (b) *RC* Π model, (c) *RLC* T model, (d) *RLC* Π model [13]
Figure reused with kind permission from Springer Science and Business Media

characteristics [11]. A first order approximation of an inductive interconnect is shown in Fig. 1.2(c). Distributed T and Π equivalent models for an *RLC* line are shown in Fig. 11.3(c) and (d), respectively [18, 19].

The conditions necessary for the line inductance to be included in the interconnect model have been examined in [20–22]. In high speed digital circuits,

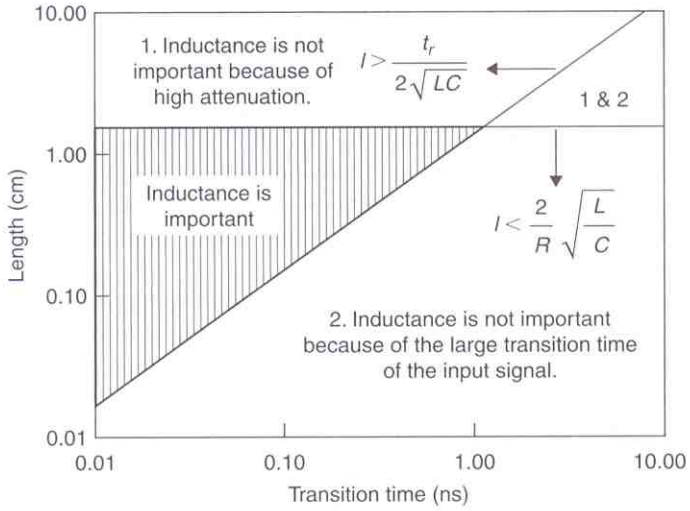

FIGURE 11.4

Transition time (t_r) versus the length of the interconnect line (l). The crosshatched area denotes the region where inductance is important ($L = 10^{-8}$H/cm, $R = 400\Omega$/cm, and $C = 10^{-12}$F/cm) [11]

limits to include the line inductance in the interconnect model based on the line impedance and transition time are described [11]. A signal propagating in an underdriven uniform lossy transmission line exhibits significant inductive effects if the line length l satisfies the following condition [11, 12],

$$\frac{t_r}{2\sqrt{LC}} < l < \frac{2}{R}\sqrt{\frac{L}{C}} \tag{11.1}$$

where R, L, and C are the per unit length resistance, inductance, and capacitance, respectively, and t_r is the rise time of the signal waveform. A graphical depiction of the length and transition time that requires the inclusion of inductance for arbitrary values of line resistance, capacitance, and inductance is depicted in Fig. 11.4 [11]. In addition, a criterion is presented in Fig. 11.4, consistent with (11.1), that graphically illustrates the conditions in which the inductance can be omitted from an interconnect model. The first condition states that as the attenuation of the line increases and the magnitude of the subsequent reflections from the load decreases, ringing on the line can be eliminated, permitting the inductance to be ignored. The second condition is that if the transition time of the output signal from the CMOS gate driving a transmission line is greater than twice the time of flight of the signal propagating across the line, the inductance can again be ignored. It is important to note that increasing signal frequencies typically require faster signal transition times. The resulting effect of a decreased signal transition

time is a lower limit on the line length, making shorter on-chip interconnects behave inductively. As a consequence, medium length lines can also behave inductively at high signal frequencies.

In addition to frequency considerations in determining whether to include inductance in an interconnect model, the length of the line must also be considered. Since the time of flight along the interconnect is dependent on the length of the line, longer lines increase the likelihood of requiring inductance in an interconnect model. The line inductance should therefore be considered in high speed, high complexity ICs [23–31].

The introduction of new materials also increases the importance of the interconnect inductance. New dielectric materials and metals have been introduced to reduce the interconnect impedance. The line capacitance can be reduced by half of the capacitance of SiO_2 with the use of low k dielectrics [13]. In addition, copper interconnect has reduced the line resistance by a factor of two to three as compared to aluminum [13]. These new materials further the need to include the line inductance in interconnect models.

Design methodologies for driving global interconnects have been proposed to reduce the propagation delay of long resistive lines; however, these techniques have also ignored the inductance of the line. Under certain conditions, ignoring the line inductance may lead to high area and power inefficient circuits. Novel techniques for designing both low power and high speed circuits to drive both *RC* and *RLC* lines are presented in the next section.

11.2 LOW POWER, HIGH SPEED CIRCUIT DESIGN TECHNIQUES

Power and noise are important characteristics when considering design techniques to optimize circuit performance in low power, high speed circuits. Noise from both inter- and intra-layer capacitive and inductive interconnect coupling, as illustrated in Fig. 11.5, affects the delay, degrades the waveform shape, and most importantly, creates the possibility of an erroneous interpretation of the digital signals [11, 32–34]. In addition, faster clock rates create higher slew rates, further increasing the on-chip noise. A variety of design techniques have been developed to mitigate deleterious on-chip noise. Lowering the power consumed by circuits, however, requires a variety of different design techniques that target a combination of static, dynamic, and short-circuit power. Power dissipation in CMOS circuits is therefore reviewed in Subsection 11.2.1. Once an understanding of the basic power dissipation principles has been established, wire sizing is introduced in Subsection 11.2.2 as a useful technique to improve circuit performance. In Subsection 11.2.3, a driver sizing optimization procedure is presented as another technique to enhance system performance. Tapered buffers are introduced in Subsection 11.2.4. Subsection 11.2.5 focuses on repeater insertion as a means to partition long interconnects into smaller segments, thereby reducing the capacitive load of each section driven by an individual repeater (as compared to the total load driven by a single large inverter). And finally, some summarizing remarks on low power, high speed circuit design techniques are presented in Subsection 11.2.6.

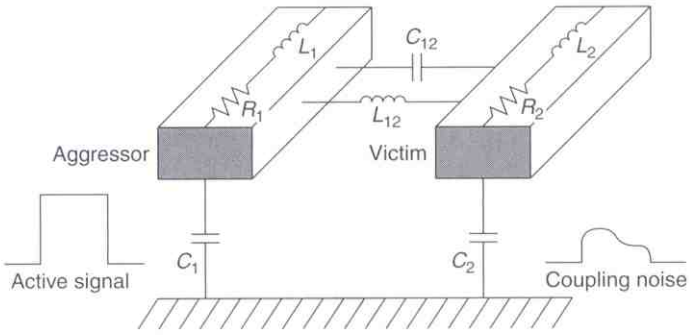

FIGURE 11.5

Cross-coupled interconnect noise in ICs. An active transition on an aggressor line can induce a coupling noise voltage on a victim line [6]

Figure reused with kind permission from Springer Science and Business Media

11.2.1 CMOS Power Dissipation

There are three primary components of power dissipation in CMOS circuits. The three components are

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{SC}} + P_{\text{leakage}} \tag{11.2}$$

where the dynamic and short-circuit power are components of the transient power, and the leakage power is a component of the static power. The dynamic, short-circuit, and leakage components can be further expanded as

$$P_{\text{total}} = CV_{\text{DD}}^2 f_{\text{switch}} + \frac{1}{2} I_{\text{peak}} t_{\text{base}} V_{\text{DD}} f_{\text{swith}} + I_{\text{leakage}} V_{\text{DD}} \tag{11.3}$$

The dynamic power P_{dynamic} accounts for the energy dissipated in charging and discharging the nodal capacitances. When a nodal capacitance C is charged, $\frac{1}{2} CV_{\text{DD}}^2$ joules of energy is stored on the capacitor, and an equal amount is dissipated in the transistors. In the discharge phase, the remaining $\frac{1}{2} CV_{\text{DD}}^2$ joules of energy stored on the capacitance is dissipated by the transistors through the discharge path, as shown in Fig. 11.6. Thus, the total energy expended in the charge and discharge cycle is CV_{DD}^2. The average dynamic power consumed is the product of CV_{DD}^2 over the frequency of the charge and discharge cycle f_{switch}, producing the well known expression for dynamic power in CMOS circuits, $CV_{\text{DD}}^2 f_{\text{switch}}$ [6].

Short-circuit current flows in a static CMOS gate when a conductive path exists from the power rail to the ground rail. A path exists when a signal transitions at the input, passing through intermediate voltage levels [35–38]. For a static CMOS inverter, this voltage range is from the n-type transistor threshold voltage V_{Tn}, the voltage at which the n-type transistor turns on, to $V_{\text{DD}} + V_{\text{Tp}}$, the voltage at which the p-type transistor turns off. Within this voltage range, both the pull-up and pull-down networks conduct DC current, producing short-circuit current, as illustrated in Fig. 11.7. The period of time when this conductive path exists is denoted as t_{base} in Eq. (11.3) [6].

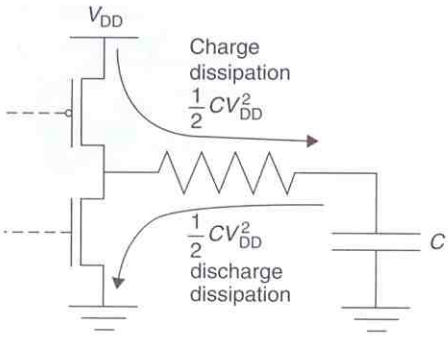

FIGURE 11.6

Energy dissipation during the charge/discharge cycle [6]
Figure reused with kind permission from Springer Science and Business Media

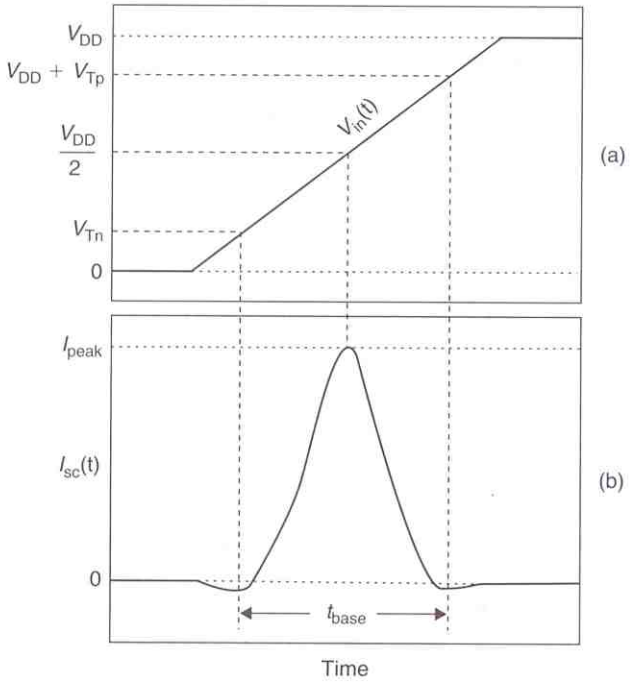

FIGURE 11.7

Short-circuit current waveform of a CMOS gate: (a) ramp-shaped input waveform, (b) short-circuit current waveform [6]
Figure reused with kind permission from Springer Science and Business Media

The transistor leakage current $I_{leakage}$ is the current that flows between the power terminals in the absence of any switching, giving rise to a leakage power component $P_{leakage}$. Typically, 50–70% of the total power dissipation is contributed by the dynamic power component [6]. Therefore, an effective strategy for reducing the total transient power consumption is to reduce the dynamic dissipation by lowering V_{DD}, operating at a lower frequency, or reducing nodal capacitances.

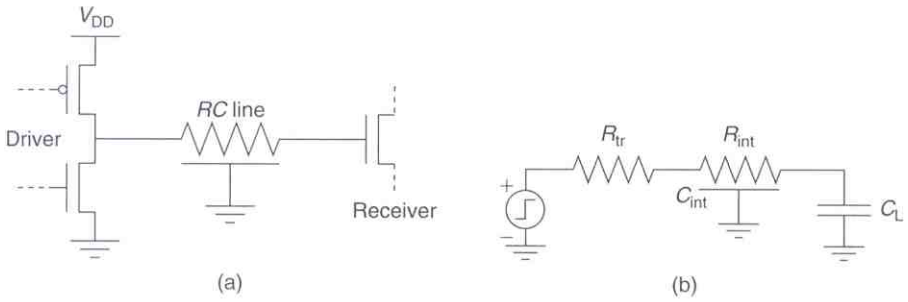

FIGURE 11.8

A CMOS circuit driving an *RC* interconnect: (a) circuit driving an *RC* line, (b) corresponding model [6]

Figure reused with kind permission from Springer Science and Business Media

Static power due to leakage current, however, is expected to grow significantly in the near future, soon exceeding 50% of the total power dissipated on-chip [39–41].

11.2.2 Wire Sizing

The width of an interconnect affects the power characteristics and propagation delay of an IC. As wiring becomes longer and the interconnect cross-sectional area smaller, it results in an increase in the *RC* interconnect impedances, thereby degrading the delay of the gates. Consider a CMOS inverter driving an *RC* interconnect line, as illustrated in Fig. 11.8. A simple first-order model of the delay of this circuit is [6]

$$T_{50\%} = 0.4R_{int}C_{int} + 0.7(R_{tr}C_{int} + R_{tr}C_L + R_{int}C_L) \tag{11.4}$$

If the driver load is effectively capacitive, where 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 terminating gate to model the load as a lumped capacitance, permitting the circuit delay to be characterized by an *RC* circuit delay, $0.7R_{tr}(C_{int} + C_L)$. Increasing the driver transistor width reduces R_{tr}, decreasing the circuit delay, thereby trading off circuit power and area for higher speed. This behavior, however, changes when R_{int} becomes comparable to R_{tr}. The delay cannot be reduced below $R_{int}(0.4C_{int} + 0.7C_L)$. Note that the interconnect component, $0.4R_{int}C_{int}$, increases quadratically with interconnect length since both R_{int} and C_{int} are proportional to the length of the line. Increasing the width of the interconnect to reduce R_{int} does not significantly reduce the delay caused by the *RC* interconnect impedance since this decrease in wire resistance is offset by an increase in the wire capacitance. Many algorithms have been proposed to determine the optimum wire size that minimizes a target cost function. Some of these algorithms address reliability issues by reducing clock skew [42], while most of these algorithms focus on minimizing delay [43–47]. The results described in [48–50] consider simultaneous driver and wire sizing based on the Elmore delay model [51] with capacitance, resistance, and power models.

Additionally, tradeoffs exist between the dynamic and short-circuit power characteristics as there is a dependence of the power dissipation on the interconnect width, as illustrated in Fig. 11.9. As the line inductance-to-resistance ratio increases with wider lines, the short-circuit power decreases due to a reduction in the signal transition time. For an *RC* line, the short-circuit power will remain approximately constant with increasing width as the decrease in interconnect resistance is offset by an increase in capacitance, maintaining a relatively unchanged *RC* time constant, and therefore signal transition time. If the width of the interconnect exceeds a specific limit (shown in Fig. 11.9), the short-circuit power increases for both an *RC* and *RLC* line due to the change in the matching characteristics between the driver and interconnect [52]. The dynamic power increases with line width since the line capacitance is greater. As shown in Fig. 11.9, an optimum interconnect width exists at which the total transient power is a minimum if the line exhibits inductive behavior.

11.2.3 Driver Sizing

Transistor sizing is another design approach, producing tradeoffs at the circuit level in CMOS logic families [54–63]. Wider transistors produce more current; however, the physical area and gate capacitance also increase linearly with width, increasing the circuit area and power. Thus, optimal transistor sizing is strongly dependent on the design tradeoffs among area, power, and speed.

A common objective of transistor sizing is lower delay. Consider a CMOS circuit with the output load dominated by the input capacitance of the following stage. The charge time monotonically decreases with increasing driver width. The input load of the transistor, however, also increases linearly with the driver width, loading the preceding gate. The net result is that the total delay of a data path with additional stages can be smaller. Similarly, a uniform increase in all of the transistors does not substantially change the propagation delay of a circuit where the output loads are dominated by the input capacitance of the fanout. The current

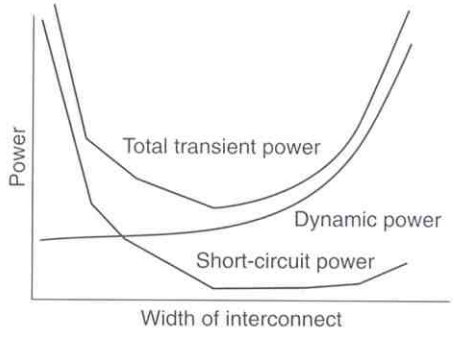

FIGURE 11.9

Dynamic, short-circuit, and total transient power as a function of interconnect line width assuming an inductive line [53]

Figure reused with kind permission from Springer Science and Business Media

drive of the gates I_{out} will increase which is offset, however, by an increase in the output capacitive load C_L. The I_{out}/C_L ratio remains essentially constant [6]. A careful balance of the current drive and output load is therefore necessary to enhance circuit performance.

Aside from improving circuit performance by increasing the current drive, transistor sizing also affects the power characteristics of a circuit. A simple approximation treats the circuit power as linearly proportional to the total active area A of a driver, that is, $P = CV^2f$, where $C = C_{ox}A$, and the gate oxide capacitance per unit area C_{ox} is constant for a given technology. An increase in the transistor width increases the area of the driver, which dissipates more power. Using the product of the power consumed and the delay of the driver as a figure of merit for optimizing the transistor size, the power-delay product is minimum when the gate output capacitance equals the sum of the interconnect and load capacitances [56, 64]. The power optimal transistor size is smaller than the power-delay optimal transistor size. An efficient tradeoff between power and delay is needed, however, with intermediate sized transistors. Non-optimal tradeoffs beyond the power-delay optimal size can be pursued in performance aggressive circuits [6].

11.2.4 Tapered Buffers

An important example of transistor sizing to drive large capacitive loads is tapered buffers [6]. The intermediate buffers are used to drive the intermediate capacitive loads. An inverter appropriately scaled for the capacitive load, as shown in Fig. 11.10(a), reduces the delay; however, the large input capacitance of the inverter loads the previous logic stage. A similar argument can be made when inserting another inverter sufficiently large to drive the inverter driving the load. This process continues until the initial input inverter of the buffer is sufficiently small to be driven by a logic gate at an acceptable speed. Thus, a tapered buffer consists of a chain of inverters of gradually increasing size, as illustrated in Fig. 11.10(b). The ratio of the size of an inverter to the size of the preceding inverter is the tapering factor β. Under the assumption that a stage load is proportional to the size of the next stage,

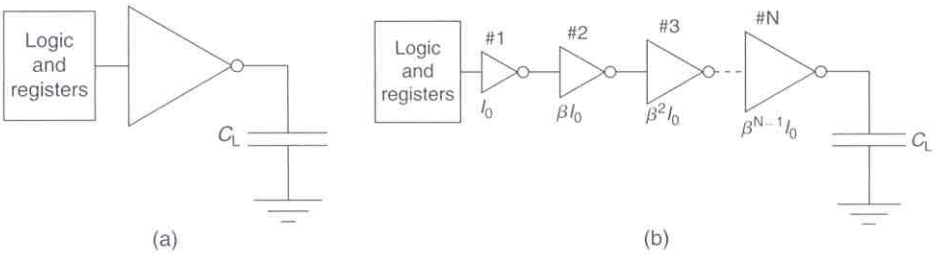

FIGURE 11.10

Circuit techniques to drive a large capacitive load: (a) a large single inverter, (b) A system of tapered buffers. The delay of the tapered buffer system is often less than the delay of a single large inverter [6]

Figure reused with kind permission from Springer Science and Business Media

thereby neglecting the interconnect and output capacitance of the gate, the delay of a tapered buffer is minimum at a constant exponential tapering factor $\beta_{opt} = e$, the base of the natural logarithm [66]. This constant tapering factor also corresponds to an optimal number of stages $N_{opt} = \ln M$, where $M = C_L/C_o$ is the ratio of the load capacitance C_L to the input capacitance C_o of the initial inverter in the chain [65, 66]. Note that since the number of buffer stages N is an integer, this condition cannot in general be satisfied precisely. Therefore, one of the two integers closest to $\ln M$ is chosen, where β is selected to satisfy $\beta^N = M$. More accurate delay models [67–69] and capacitance models [70, 71] have been employed to include the intrinsic drain and source capacitance, a ramp input, and short-circuit current. The delay optimal tapering factor increases with the ratio of the intrinsic output capacitance (which includes the diffusion and gate overlap capacitance) to the input gate capacitance [67, 68]. Further enhancements to the model, such as the effects of a finite slew rate, producing short-circuit current, have also been incorporated [70, 71].

Tradeoffs among area, power, and delay have also been considered [72–75]. For a specific load, the dependence of the buffer delay on the tapering factor is relatively flat around β_{opt}, as depicted in Fig. 11.11. The total area of the buffer is also a relatively strong function of β_{opt}. Thus, an effective tradeoff among the delay, 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. Tapered buffers with a fixed tapering factor have been compared with a geometrically increasing tapered buffer system [75]. The minimum delay of a variable-taper buffer can be reduced to within a few percent of the delay of a fixed-taper buffer by implementing the first few stages with a fixed-taper factor [6]. Optimal area-delay tradeoffs are therefore achieved in a fixed-taper buffer system with the final one to two stages utilizing a larger tapering factor.

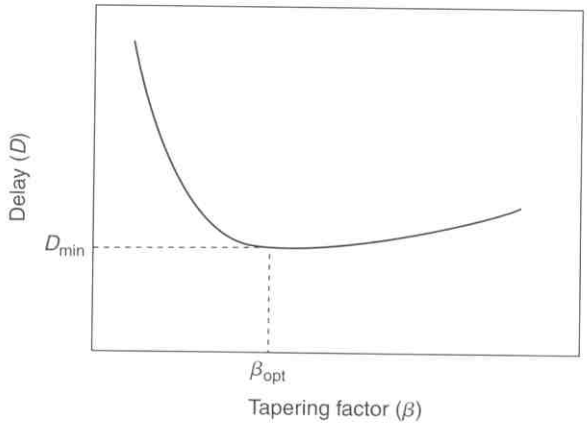

FIGURE 11.11

Dependence of the tapered buffer propagation delay on the tapering factor [6]
Figure reused with kind permission from Springer Science and Business Media

11.2.5 Repeater Insertion

Widening a uniform line has a marginal impact on the overall wire delay. A more effective strategy for reducing the delay of a long interconnect is to strategically insert buffers along a line. These buffers are typically called repeaters and the process is called repeater insertion [14]. Repeaters circumvent the quadratic increase in interconnect delay by partitioning the line into smaller and approximately equal sections, as shown in Fig. 11.12. The sum of the section delays is smaller than the delay of the original path since the delay of each section is reduced. The decreased interconnect delay is partially offset by the additional delay of the inserted repeaters. The optimal number of repeaters is determined by considering the delay of each individual repeater added to the repeater system, and determining the number of repeaters at which the increase in the repeater delay outweighs the lower interconnect delay. The optimal number of repeaters k_{opt} and the optimal size of the repeaters h_{opt} as compared to a minimum sized repeater h are

$$k_{opt} = \sqrt{\frac{a_1 R_t C_t}{a_2 R_0 C_0}} \tag{11.5}$$

$$h_{opt} = \sqrt{\frac{R_0 C_t}{R_t C_{g0}}} \tag{11.6}$$

respectively, where R_t and C_t are the total interconnect resistance and capacitance, respectively, R_0 and C_0 are the input and output repeater resistance and capacitance, respectively, and C_{g0} is the input capacitance of the repeater. The two fitting parameters, a_1 and a_2, account for the rise and fall time of the propagating signal [76].

A number of repeater insertion methods have been proposed [77–82]. Bakoglu presents a method based on characterizing the repeaters by the input capacitance and the effective output resistance of each repeater [14, 83]. The minimum delay of the resulting *RC* circuit is achieved when the delay of the repeater section equals the wire segment delay.

Techniques to improve interconnect performance vary depending upon the electrical characteristics of the line. For an *RC* line, repeater insertion techniques outperform wire sizing [84]. Unlike an *RC* line, the minimum signal propagation delay always decreases with increasing line width for *RLC* lines if an optimum repeater system is used [85, 87]. In *RLC* lines, wire sizing outperforms repeater insertion as the minimum signal propagation delay with no repeaters is smaller

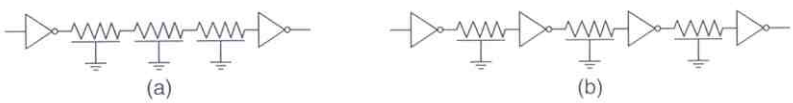

FIGURE 11.12

Repeater insertion: (a) original interconnect line, (b) interconnect line with inserted repeaters [6]
Figure reused with kind permission from Springer Science and Business Media

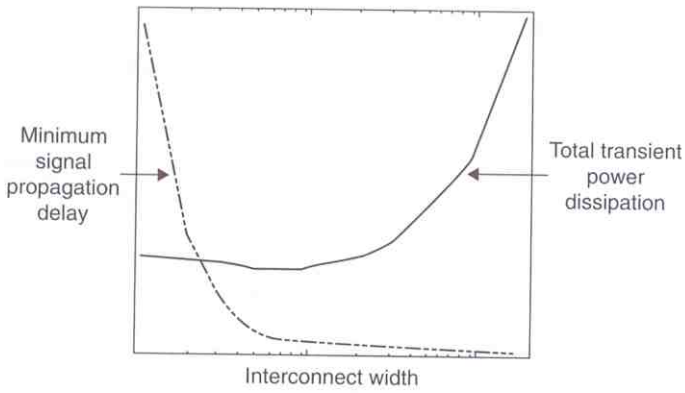

FIGURE 11.13

Minimum signal propagation delay and transient power dissipation as a function of line width for a repeater system [13]
Figure reused with kind permission from Springer Science and Business Media

than the minimum signal propagation delay using any number of repeaters. For an *RLC* line, the minimum signal propagation delay always decreases with wider lines until the number of repeaters equals zero. As shown in Fig. 11.13, the minimum propagation delay decreases while the power dissipation increases for wider interconnect, delineating the tradeoff between minimum delay and total power dissipation [13].

The interconnect resistance decreases with wider lines, increasing the ratio between the line inductance and resistance L/R, and decreasing the number of inserted repeaters to achieve the minimum propagation delay. The minimum delay produced by an optimum repeater system decreases with increasing line width as the total gate delay decreases. For an inductive interconnect line, the total signal propagation delay is [13]

$$t_{\text{pd-total}} = k_{\text{opt-}RLC} t_{\text{pd-section}} \tag{11.7}$$

where $t_{\text{pd-section}}$ is the signal delay of each *RLC* section [86] and $k_{\text{opt-}RLC}$ is the optimum number of repeaters. As shown in Fig. 11.14, for different line lengths l, the optimum number of repeaters $k_{\text{opt-}RLC}$ which minimizes the signal propagation delay decreases with increasing line width for all line lengths until the number of repeaters reaches zero, the point at which only a single driver at the beginning of the line is effective. *RC* lines, however, require repeaters, increasing $k_{\text{opt-}RC}$, as a wider line increases the line capacitance driven by each repeater. The propagation delay of an *RLC* line is therefore a decreasing function of the line width, whereas the propagation delay of an *RC* line is dependent on the delay of an increasing number of repeaters and is a function of the line width [13].

11.2.6 Summary

In this section, an introduction to several low power, high speed circuit design techniques has been presented. Wire sizing is shown to produce nominal improvements

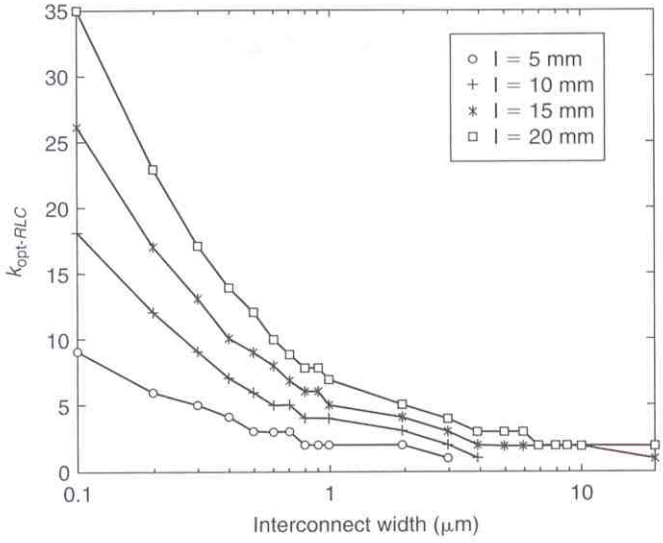

FIGURE 11.14

Optimum number of repeaters for minimum propagation delay for different line widths [13]
Figure reused with kind permission from Springer Science and Business Media

in circuit speed. Cascaded buffers have been suggested to drive large capacitive loads. Repeater insertion techniques have also been introduced to improve signal propagation delay in long resistive interconnects. The remaining sections of Chapter 11 focus on applying the basic principles presented in this chapter to power and clock distribution networks, as well as 3-D interconnect technologies.

11.3 GLOBAL POWER DISTRIBUTION NETWORKS

Distributing power in high speed, high complexity ICs has become a challenging task. This section provides insight and intuition into the behavior and design of power distribution networks. An overview of noise issues related to the power distribution network is presented. Within this noise framework, the basic model of a power grid is discussed. Decoupling capacitors are also introduced as a means to temporarily provide charge from within the power network. After a basic understanding of the models, noise issues, and use of decoupling capacitors, two important characteristics of power networks are examined in this section. The first phenomenon, multi-path current redistribution, is directly related to the frequency dependent impedance variation amongst the different levels of the on-chip metallization layers. The second topic, electromigration, is a consequence of the increasing current densities encountered in high complexity circuits.

11.3.1 Noise in Power Distribution Networks

Noise in power distribution networks is depicted by the power delivery system illustrated in Fig. 11.15. The power grid consists of a supply, load, and interconnect

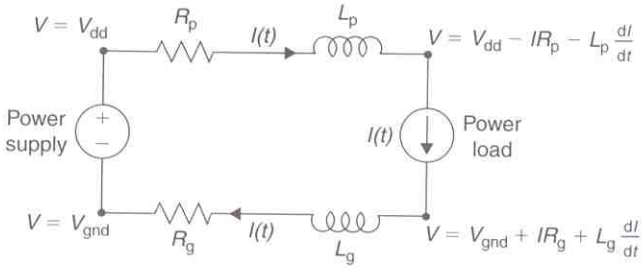

FIGURE 11.15

Power delivery system consisting of the power supply, power load, and non-ideal interconnect [12]

Figure reused with kind permission from Springer Science and Business Media

lines connecting the supply to the load. The nominal power and ground voltages, V_{dd} and V_{gnd}, are provided by an ideal power supply. A variable current source $I(t)$, which is typically a transistor or group of transistors, models the power load. The interconnect lines connecting the power supply to the load are considered non-ideal with a finite resistance and inductance, R_p, L_p and R_g, L_g, for the power and ground lines, respectively. Resistive voltage drops $\Delta V_R = IR$ and inductive voltage drops $\Delta V_L = L\,(dI/dt)$ develop across the parasitic interconnect impedances as the load draws current from the power network. The voltage levels at the load terminal change from the nominal levels provided by the supply, decreasing to $V_{dd} - IR_p - L_p\,(dI/dt)$ at the power terminal and rising to $V_{gnd} + IR_g + L_g\,(dI/dt)$ at the ground terminal. This change in supply voltages is referred to as power supply noise [12].

Power supply noise can adversely affect circuit operation. One major consequence is an increase in signal delay uncertainty. When power supply variations reduce the rail-to-rail power voltage, the gate-to-source voltage across both the NMOS and PMOS transistors also decreases, thereby lowering the output drive current of these devices. The signal delay increases as compared to the delay under a nominal power supply voltage. Conversely, a higher power voltage and a lower ground voltage shorten the propagation delay. The net effect of power noise on propagating clock and data signals is an increase in both delay and delay uncertainty within the data paths [88, 89]. Consequently, power supply noise can severely limit the maximum operating frequency of an IC [90, 91].

The power distribution network should exhibit a small impedance at the terminals of the load to ensure a small variation in the power supply voltage. Decoupling capacitors ensure correct and reliable operation of an IC in the frequency range from DC to some target operating frequency f_o by ensuring that the impedance of the power network is maintained below a specified upper bound within this target frequency range. The function of a decoupling capacitor is to provide charge when transient current demands on the power grid are high. These decoupling capacitors are distributed across a system, placed at the board, package, and on-chip levels (see Fig. 11.16). Each decoupling capacitor provides transient current to the load, effectively reducing the local transient noise. The

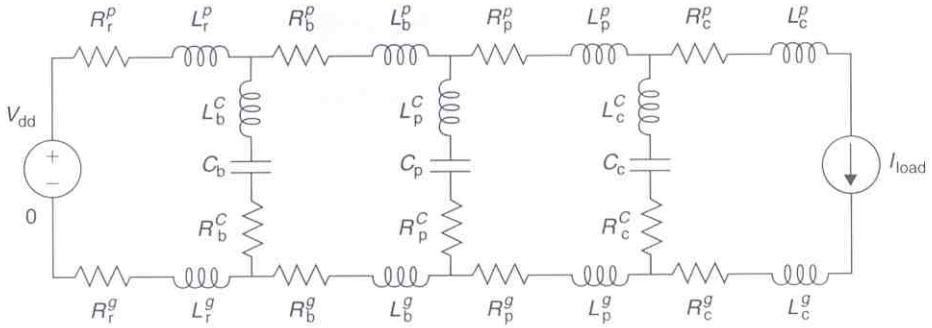

FIGURE 11.16

Power network with board, package, and on-chip decoupling capacitances [12]
Figure reused with kind permission from Springer Science and Business Media

low frequency power noise is lowered by those decoupling capacitors farthest from the load, whereas the high frequency transients require a fast injection of charge which is provided by those capacitors closest to the load.

11.3.2 Multi-Path Current Redistribution

Multi-path current redistribution is an extension of current redistribution within a single conductor to return current flowing among several parallel conductors. Adjacent signal lines, power networks, and the substrate can provide a variety of potential current return paths. Significant redistribution of the return current among these return paths can occur as signal frequencies increase. At low frequencies, the line impedance $Z(\omega) = R(\omega) + j\omega L(\omega)$ is dominated by the interconnect resistance. In this case, the path resistance determines the distribution of the return current among the available return paths, as shown in Fig. 11.17(a). At high frequencies, the line impedance $Z(\omega) = R(\omega) + j\omega L(\omega)$ is dominated by the reactive component $j\omega L(\omega)$. The minimum impedance path is primarily determined by the least inductive $L(\omega)$ path, as shown in Fig. 11.17(b). In power grids, both the forward and return currents undergo multi-path redistribution as both the forward and return paths can change with frequency since the paths consist of multiple conductors connected in parallel [12]. Current redistribution can lead to excessive current densities along certain paths. Under these conditions, a phenomenon known as electromigration must be considered.

11.3.3 Electromigration

On-chip current densities can reach several hundred thousand amperes per square centimeter, making electromigration a significant issue. Electromigration is the transport of metal atoms under the force of an electron flux. The significance of electromigration has been established early in the development of ICs [93, 94]. The depletion and accumulation of metal material resulting from atomic flow

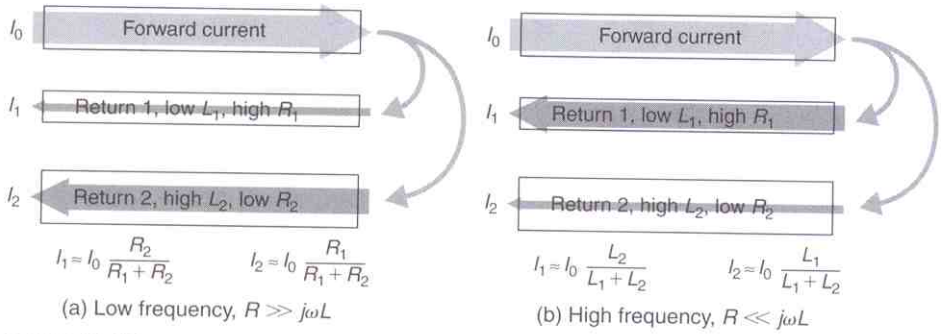

FIGURE 11.17

Current loop with two alternative current return paths. The forward current I_0 returns both through return path one with resistance R_1 and inductance L_1, and return path two with resistance R_2 and inductance L_2. In this structure, $L_1 < L_2$ and $R_1 > R_2$: (a) At low frequencies, the path impedance is dominated by the line resistance. (b) At high frequencies, the path impedance is dominated by the line inductance [12]

Figure reused with kind permission from Springer Science and Business Media

can lead to the formation of extrusions and voids in the metal structures. These extrusions and voids can lead to short circuits and open circuit faults, respectively, degrading the reliability of an IC [92]. The mass transport of metal ions through diffusion under an electrical driving force F is

$$J_a = C_a \mu F \tag{11.8}$$

where C_a is the atomic concentration and μ is the mobility of the atoms.

Two forces, an electric field force and an electron wind force, act on the metal ions. The electric field force is proportional to the electric field E and acts in the direction of the field. Conduction electrons accelerate in the direction opposite to the electric field, transferring momentum to the metal ions in the course of scattering. The force exerted by these electrons is also in the direction opposite to the field E, and is commonly referred to as the electron wind force. In metals of interest, such as aluminum and copper, the electron wind force dominates and the net force acts in the direction opposite to the electric current. The resulting atomic flux is therefore in the opposite direction of the electric current j, as shown in Fig. 11.18.

11.3.4 Summary

An overview of noise modeling is presented in this section as a means to introduce basic concepts behind the design of power distribution networks. Decoupling capacitors have been introduced as a means to reduce the noise within the power grid. Additionally, an examination of both multi-path current redistribution and electromigration provides insight into two important characteristics of interconnect, in general, and power grids, in particular. A more detailed discussion of power distribution networks can be found in [12].

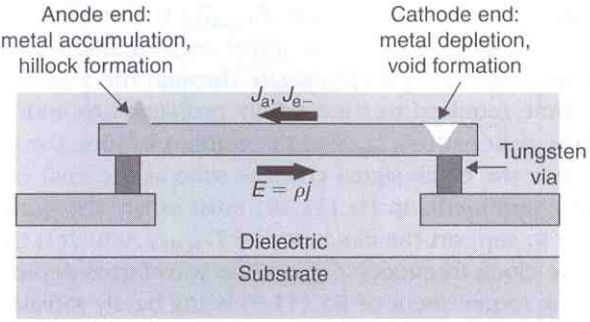

FIGURE 11.18

Electromigration mass transport in an interconnect line. An electron flux J_{e-} flowing in the opposite direction of the electric field $E = \rho j$ induces an atomic flow J_a in the direction of the electron flow [12]

Figure reused with kind permission from Springer Science and Business Media

11.4 CLOCK DISTRIBUTION NETWORKS

In a synchronous digital system, the clock signal provides a time reference for the movement of data within that system. Clock signals are typically loaded with the greatest fanout, travel over the longest distances, and operate at the highest speeds of any signal, either control or data, within the entire system. The control of any differences in the delay of the clock signals can severely limit the maximum performance of the entire system as well as create catastrophic race conditions in which an incorrect data signal may latch within a register. Therefore, understanding the basic principles governing clock distribution networks is of primary importance. An overview of the primary timing relationships is therefore provided in Subsection 11.4.1 to introduce the important concepts of local data paths and clock skew. Timing constraints and relationships are followed by a general introduction to clock topologies in Subsection 11.4.2, with more specific asymmetric and symmetric topologies presented in Subsections 11.4.3 and 11.4.4, respectively. Low power clock design is reviewed in Subsection 11.4.5, which is followed by a short summary in Subsection 11.4.6.

11.4.1 Timing Relationships

General synchronous systems are composed of the following three delay components: (i) memory storage elements; (ii) logic elements; and (iii) clocking circuitry and distribution networks [95]. The minimum allowable clock period $T_{CP(min)}$ between any two registers in a sequential data path is

$$\frac{1}{f_{clk,MAX}} = T_{CP(MIN)} = T_{PD(MAX)} + T_{Skew} \tag{11.9}$$

where

$$T_{PD(MAX)} = T_{C\text{-}Q} + T_{Logic} + T_{Int} + T_{Set\text{-}up} + T_{Hold} \tag{11.10}$$

and the total path delay of a data path $T_{PD(MAX)}$ is the sum of the maximum time required for the data to leave the initial register once the clock signal C_i arrives $T_{C\text{-}Q}$, the time necessary to propagate through the logic and interconnect $T_{Logic} + T_{Int}$, the time required to successfully propagate to and latch within the final register of the data path $T_{Set\text{-}up}$, and the amount of time the input data signal must be stable once the clock signal changes state at the final register T_{Hold}. The sum of the delay components in Eq. (11.10) must satisfy the timing constraint of Eq. (11.9) in order to support the clock period $T_{CP(MIN)}$, which is the inverse of the maximum possible clock frequency $f_{clk,MAX}$. The waveforms depicted in Fig. 11.19 illustrate the timing requirement of Eq. (11.9) being barely satisfied. Note that the clock skew T_{skewij} can be positive or negative depending on whether C_f leads or lags C_i, respectively.

Clock distribution networks are based on equipotential clocking, where the entire network is considered a temporal surface which must be maintained at a specific voltage at each half of the clock cycle. Ideally, clocking events occur simultaneously at all registers. Given this global clocking strategy, the clock signal arrival times at each register are defined with respect to a universal time reference. The difference in the clock signal arrival time between two sequentially adjacent registers is the clock skew T_{Skew}. Zero clock skew occurs if the clock signals C_i and C_f are in complete synchronism. The clock skew between two sequentially adjacent registers, R_i and R_j, and an equipotential clock distribution network is defined as

$$T_{Skewij} \equiv T_{Ci} - T_{Cj} \tag{11.11}$$

where T_{Ci} and T_{Cj} are the clock delay from the clock source to the registers R_i and R_j, respectively. Note that system-wide or chip-wide clock skew between two non-sequentially adjacent registers, from an analysis viewpoint, has no effect on the performance and reliability of a synchronous system and is essentially meaningless. System-wide global clock skew only places constraints on the permissible local clock skew. The clock skew between any two registers in a global data path which are not necessarily sequentially adjacent is the sum of the clock skew between each pair of registers along the global data path between those same two registers.

Depending upon whether C_i leads or lags C_f and upon the magnitude of T_{Skew} with respect to T_{PD}, system performance and reliability can either be degraded or enhanced. If the time of arrival of the clock signal at the final register of a data path T_{Cf} leads that of the time of arrival of the clock signal at the initial register of the same sequential data path T_{Ci}, as depicted in Fig. 11.20(a), the clock skew is referred to as positive clock skew and, under this condition, the maximum attainable operating frequency is decreased. Positive clock skew is the additional amount of time which must be added to the minimum clock period to reliably apply a new clock signal at the final register. Also note that positive clock skew only affects the maximum frequency of a system and cannot produce a race condition. If the clock signal arrives at R_i before the signal reaches R_f, as shown in Fig. 11.20(b), the clock skew is defined as being negative.

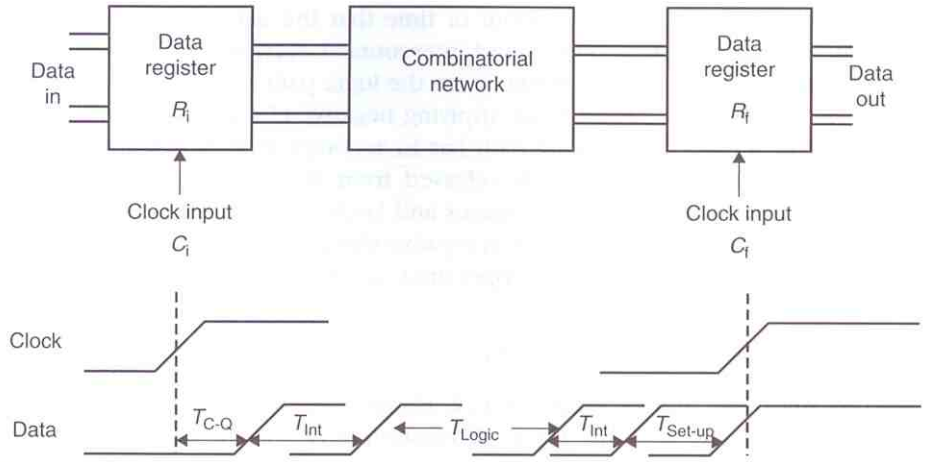

FIGURE 11.19

Timing diagram of clocked data path [95]

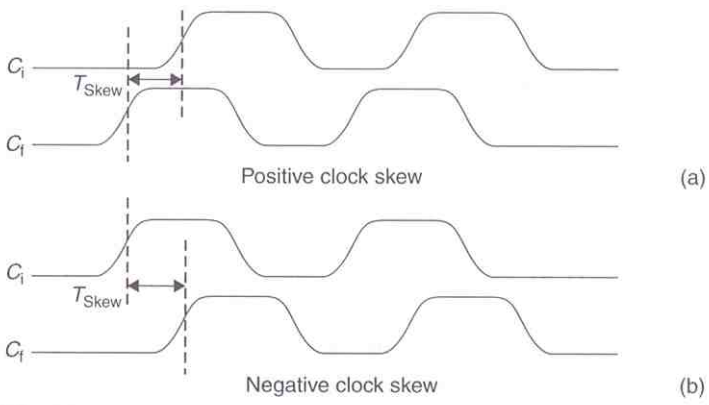

FIGURE 11.20(a, b)

Clock timing diagrams [95]

Negative clock skew can be used to improve the maximum performance of a synchronous system by decreasing the delay of a critical path; however, a potential minimum constraint can occur, creating a race condition [104–108]. In this case, when C_f lags C_i, the clock skew must be less than the time required for the data signal to leave the initial register, propagate through the interconnect and combinatorial logic, and successfully set up in the final register. If this condition is not met, the data stored in register R_f is overwritten by the data that had been stored in register R_i and has propagated through the combinatorial logic. By forcing C_i to lead C_f at each critical local data path, excess time is shifted from the neighboring less critical local data paths to the critical local data paths. This negative clock

skew represents the additional amount of time that the data signal at R_i has to propagate through the logic stages and interconnect sections and into the final register. Negative clock skew subtracts from the logic path delay, thereby decreasing the minimum clock period. Thus, applying negative clock skew increases the total time that a given critical data path has to accomplish its functional requirements by providing the data signal released from R_i extra time to propagate through the logic and interconnect stages and latch into R_f. The use of negative clock skew in a random path i results in positive clock skew in the preceding path $i - 1$, which may establish the new upper limit on the system clock frequency.

11.4.2 Clock Network Topologies

Tradeoffs that exist among system speed, physical die area, and power dissipation are greatly affected by the clock distribution network. The design methodology and topology of the clock distribution network should be considered in the development of the structure of the network for distributing the clock signals. The most common and general approach to equipotential clock distribution is the use of buffered trees. In contrast to these highly asymmetric structures, symmetric trees, such as H-trees, are also used to distribute high speed clock signals [95].

11.4.3 Asymmetric Topologies

The most common strategy for distributing on-chip clock signals is to insert buffers at the clock source and along the clock path, forming a tree structure [96, 97]. The clock source is frequently described as the root of the tree, the initial portion of the tree as the trunk, individual paths as the branches, and the registers being driven as the leaves. This metaphor describing a clock distribution network is illustrated in Fig. 11.21. Occasionally, a mesh version of the clock tree structure is used in which shunt paths further down the clock network are used to minimize the interconnect resistance within the clock tree. This mesh structure, an extension of the standard clock tree depicted in Fig. 11.21, effectively places branch resistances in parallel, minimizing the clock skew but at the cost of greater power dissipation. If the interconnect resistance of the buffer at the clock source is small as compared to the buffer output resistance, a single buffer is often used to drive the entire clock distribution network. The primary requirement of a single buffer system is that the buffer should provide sufficient current to drive the network capacitance, which includes both interconnect and fanout, while maintaining high quality waveform shapes. Additionally, a single buffer can be used if the output resistance of the buffer is much greater than the resistance of the interconnect section being driven. Alternatively, buffers, acting as repeaters, distributed throughout the clock network may be used in place of a single buffer. This approach requires additional area but greatly improves the precision and control of the clock signal waveforms and is necessary if the resistance of the interconnect lines is non-negligible. The distributed buffers serve the double function of amplifying the clock signals degraded by the distributed interconnect impedances while isolating the local clock nets from the upstream load impedances. Note that

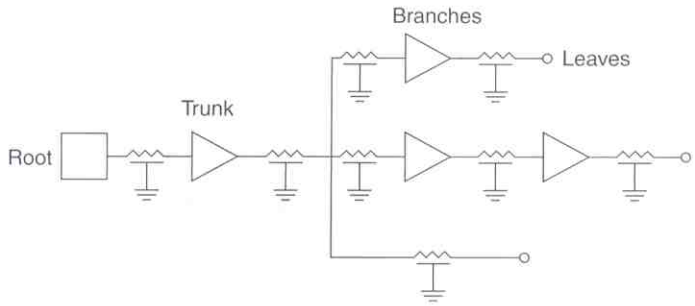

FIGURE 11.21

Tree structure of clock distribution network [95]

the buffers are a primary source of clock skew within a well-balanced clock distribution network since the active device characteristics vary much more greatly than the passive interconnect characteristics.

11.4.4 Symmetric Topologies

Another approach for distributing clock signals, an extension of the distributed buffer approach depicted in Fig. 11.21, utilizes a hierarchy of planar symmetric H-tree or X-tree structures [14, 98, 99] to ensure zero clock skew by maintaining identical distributed interconnect and buffer paths from the clock signal source to the clocked registers. This approach ensures that each clock path from the source to each register has practically the same delay. The primary delay difference among the clock signal paths is due to variations in process parameters that affect the interconnect impedance and, in particular, any active distributed amplifying buffers. The clock skew within an H-tree clock network is dependent upon the physical size, the semiconductor process, and the number of active buffers and clocked latches distributed within the H-tree structure. The conductor widths in H-tree structures are designed to progressively decrease as the signal propagates to lower levels of the hierarchy, ensuring that reflections are minimized at the branch points. Specifically, the impedance of the conductor leaving each branch point Z_{K+1} must be twice the impedance of the conductor providing the signal at that branch point Z_K for an H-tree structure and four times the impedance for an X-tree structure. Therefore, for the tapered H-tree structure illustrated in Fig. 11.22, $Z_K = Z_{K+1}/2$. A drawback to H-tree structures as compared to standard clock trees is that the interconnect capacitance and therefore the power dissipation is much greater since the total wire length tends to be much longer. Symmetric clock structures such as H-trees are used to minimize clock skew; however, an increase in the clock signal delay is incurred. The increase in clock delay must be considered when choosing between buffered asymmetric and symmetric clock networks. Additionally, H- and X-tree distribution networks are difficult to implement in high complexity integrated systems which are typically irregular in nature [95].

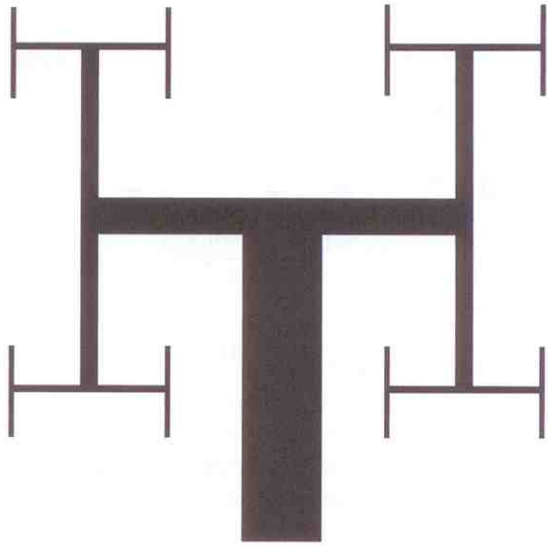

FIGURE 11.22

Tapered H-tree clock distribution network

11.4.5 Power Considerations

In modern integrated systems, the clock signal may drive many tens of thousands of registers, placing a large capacitive load on the network. The combination of a large capacitive load and a demand for higher clock frequencies has led to an increasingly larger proportion of the total power of a system dissipated within the clock network, in some applications much greater than 40% of the total power [100–102]. The primary component of power dissipation in most CMOS-based digital circuits is dynamic power. It is possible to reduce CV^2f dynamic power by lowering the clock frequency, the power supply, and/or the capacitive load of the clock distribution network. Multi-core processors target dynamic power reduction by maintaining the same logical throughput but at a reduced frequency. Each core can operate at a lower frequency while maintaining the same total workload as a single processor operating at a much higher frequency. Additionally, decreasing the total effective capacitance required to implement a clock tree can also reduce power consumption. Reductions of 10–25% in power dissipated within the clock tree have been reported with no degradation in clock frequency. Targeting the quadratic voltage term of the dynamic power has the greatest potential for power savings. A technique has been described for designing clock buffers and pipeline registers such that the clock distribution network operates at half the power supply swing, reducing the power dissipated in the clock tree by 60% without compromising the clock frequency [103]. The degradation in system speed is minimal since the data signals operate over the full power supply rail. The voltage is therefore only reduced in the clocking circuitry, resulting in significantly lower power with a minimal degradation in system speed [95].

11.4.6 Summary

In this section, clock distribution timing constraints are introduced and an overview of clock skew in terms of the local data path delay is provided. Once an understanding of the basic timing relationships is established, several asymmetric and symmetric clock topologies are presented, and the benefits and drawbacks of these networks are discussed. Finally, methods to minimize the power consumed within the clock distribution network are addressed as the clock network can consume more than 40% of the total power dissipated on-chip.

11.5 3-D INTERCONNECTS

As previously mentioned, two major effects of device scaling on interconnects are the increase in the number of metal layers necessary to achieve a higher integration density and an increase in the number and length of the global lines. These effects, caused by device scaling, increase both the line delay and capacitive crosstalk. 3-D interconnects have been proposed as a possible solution to address these issues. The introduction of a third dimension significantly alters the distribution of the interconnect length in ICs. As the number of planes is increased, the length and number of the global interconnects decrease as depicted in Fig. 11.23. Since the total number of IC interconnects are the same, the number of short interconnects increases [109]. Various characteristics, including the power dissipation and area allocated for metallization, can be improved with 3-D interconnects.

The corner-to-corner interconnect length is one such characteristic that benefits directly from 3-D integration. Since it is feasible to partition a 2-D IC into multiple subsections and stack these sections in the vertical dimension, the corner-to-corner interconnect length significantly decreases. As a result, for a constant clock frequency, several global interconnects in the upper metallization levels can be transferred to local, smaller aspect ratio metal layers. This strategy, in turn, implies a reduction in the total number of metal levels within a 3-D circuit. Alternatively, assuming a constant number of registers along a sequential data path and number of metal layers, an increase in the clock frequency is possible as the worst case data path delay can be reduced (decreasing T_{int} in Eq. (11.10)). For a constant clock frequency, the wiring pitch can be reduced, lowering the interconnect area. In addition, 3-D ICs with smaller interconnect lengths consume less power as compared to 2-D ICs as a consequence of the reduced capacitive load of the global interconnect lines. A graphical depiction of the variation in gate pitch, interconnect length, and power consumption as a function of the number of planes for two different values of Rent's exponent is illustrated in Fig. 11.24 [110]. Rent's exponent is a component of Rent's rule, a rule that correlates the number of I/O terminals with the number of circuit elements. Rent's rule describes the increase in the number of interconnects due to scaling similar to how Moore's Law describes the increase in device density due to scaling. Note that Rent's exponent increases as the parallelism of a system increases [111, 112].

Increasing the number of planes that can be integrated into a single 3-D system requires interplane interconnects that connect signals between vertically stacked

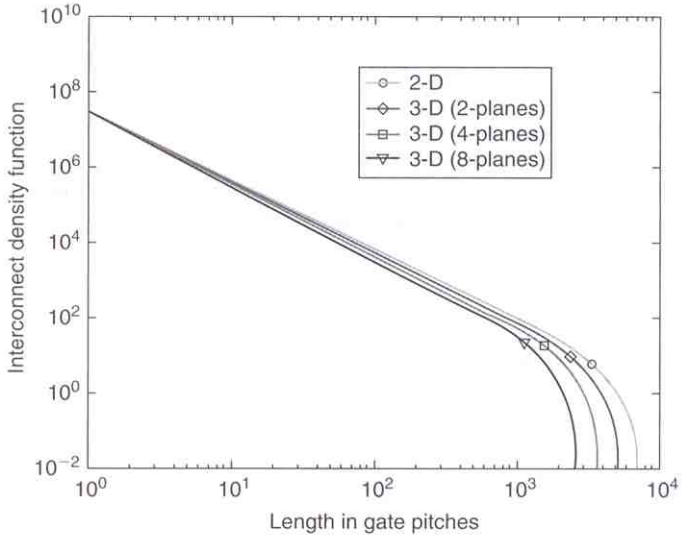

FIGURE 11.23

Interconnect length distribution for 2-D and 3-D ICs

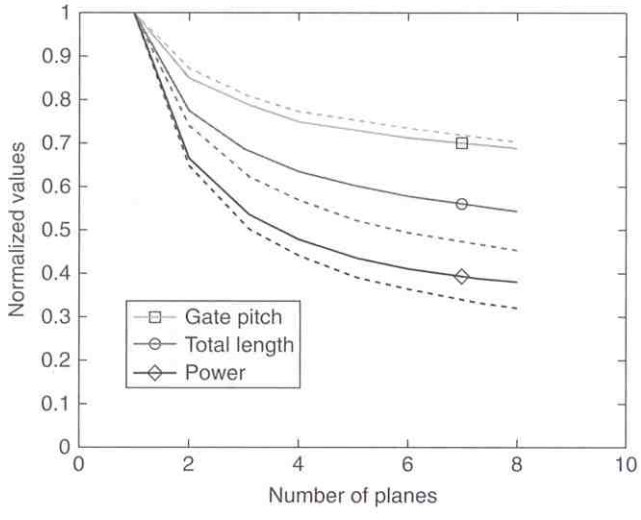

FIGURE 11.24

Variation of gate pitch, total interconnect length, and interconnect power as a function of the number of planes

devices. Interplane interconnects implemented as through silicon vias (TSV) or 3-D vias can produce the shortest path within a 3-D system, as compared to wire bonding, peripheral vertical interconnects, and solder ball arrays. Fabrication processes utilizing 3-D vias should be reliable and inexpensive, exhibit low

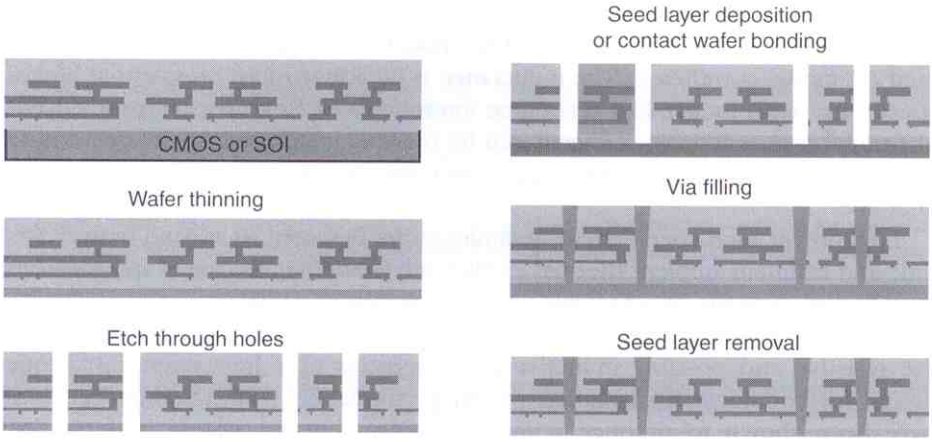

FIGURE 11.25

The via last approach to TSV fabrication and filling after wafer thinning

impedance characteristics, and have a negligible effect on the performance and reliability of the nearby active devices. There are two industrial fabrication approaches to produce TSVs. In the via first approach, the 3-D vias are formed after fabricating the active devices on each plane and prior to wafer thinning. A major disadvantage of this approach is the degradation in the reliability of the TSV due to wafer thinning and bonding. The advantages of this technique include simpler wafer handling, and compatibility with existing process flows [113]. In the via last approach, fabrication of the 3-D vias is performed after wafer thinning, as depicted in Fig. 11.25 [114–117]. The reliability of the TSVs is much improved as compared to the via first approach; however, the via last approach requires handling the thin wafers for several processing steps. Currently, the most popular method of TSV fabrication is via first.

3-D integration is a novel technology of growing importance that has the potential to offer significant performance and functional benefits as compared to 2-D ICs. Much work, however, is needed to properly characterize and model the interplane TSV, which is the primary technological innovation required to exploit the benefits of 3-D integration.

11.6 SUMMARY AND CONCLUDING REMARKS

The complexity of properly designing interconnects in the DSM regime increases with each successive technology generation. Modeling, following the same trend as design, has also increased in complexity. Many factors contribute to the complexity of properly modeling the interconnect. Deciding between modeling the interconnect as a distributed *RC* or *RLC* is dependent on such factors as the signal waveform characteristics and the length of the line. A relevant criterion

notes that if the transition time of the signal is shorter than twice the time of flight and the line is not too resistive, an *RLC* model is required; otherwise, a distributed *RC* model is sufficient. The inductance is therefore more prevalent at higher speeds and with longer, low resistance interconnects. However, with the advent of parallel processing, clock speeds can be reduced, complicating the decision as to whether including the inductance is necessary to properly model an interconnect line.

Low power, high speed circuit techniques are essential to expand battery lifetime and maintain ambient thermal levels. Understanding the causes and methods to minimize dynamic, short-circuit, and leakage power consumption as applied to interconnect design has become critical. Long, narrow interconnect lines increase the resistive and possibly inductive characteristics of a line, requiring unique design methodologies to enhance system performance. Wide interconnect lines have been shown to produce nominal improvements in system performance; however, the insertion of repeaters to partition a load into smaller, more easily driven segments decreases the signal propagation delay. Proper sizing of tapered buffers, while considering tradeoffs among area, power, and speed, is an effective technique to drive nodes with large capacitive loads.

Clock and power distribution are important applications of the general interconnect design problem. Power distribution networks are greatly affected by *IR* and $L(\mathrm{d}i/\mathrm{d}t)$ noise. Special treatment of power lines is required to assure that these noise sources do not adversely affect both the delay and the delay uncertainty of the clock and data signals. Decoupling capacitors are added between the source and sink nodes of a power network to provide charge at the terminals of the load. The large currents that are often present on the power grid can cause electromigration, the mass transport of metal ions through diffusion under an electrical driving force. These extrusions and voids can cause circuit failures in the form of short or open circuits. Clock distribution networks affect circuit behavior in other ways. Negative clock skew along a sequential data path can create race conditions although negative skew can also be used to improve the performance of a path, whereas positive clock skew degrades the maximum operating frequency. Topologies that distribute clock signals can be either asymmetric or symmetric, with buffers to maintain proper waveform characteristics. Symmetric clock distribution schemes, such as an H-tree structure, have an additional benefit of minimizing clock skew.

Novel techniques in interconnect design can also help alleviate deleterious trends that include longer line lengths, greater line impedances, and increased propagation delays. 3-D interconnects is one such technique that can help reduce the line length, and therefore the line impedance. Additional research is required to fully exploit the benefits of 3-D integration, including design, analysis, modeling, and process manufacturing technologies. As interconnect continues to become more complex, requiring greater resources, an understanding of the basic principles, as presented in this chapter, is essential to the development of new process technologies and design techniques while ensuring that electrical interconnect remains a viable means of signal transmission within the foreseeable future.

REFERENCES

[1] International Technology Roadmap for Semiconductors: Semiconductor Industry Association, 2005.

[2] K. Lee, "On-chip interconnects—Gigahertz and beyond," in *Proceedings of the IEEE International Interconnect Technology Conference*, December 1998, pp. 15–17.

[3] C. S. Chang, "Interconnection challenges and the national technology roadmap for semiconductors," in *Proceedings of the IEEE International Interconnect Technology Conference*, December 1998, pp. 3–6.

[4] M.T. Bohr, "Interconnect scaling—The real limiter to high performance ULSI," in *Proceedings of the IEEE International Electron Devices Meeting*, December 1995, pp. 241–244.

[5] F. Caignet, S. Delmas-Bendhia and E. Sicard, "The challenge of signal integrity in deep-submicrometer CMOS technology," in *Proceedings of the IEEE*, Vol. 89, No. 4, April 2001, pp. 556–573.

[6] A. V. Mezhiba and E. G. Friedman, "Trade-offs in CMOS VLSI circuits," *Trade-offs in Analog Circuit Design: The Designer's Companion*, C. Toumazou, G. Moschytz and B. Gilbert (Eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 75–114.

[7] R. Ho, K. W. Mai and M. A. Horowitz, "The future of wires," in *Proceedings of the IEEE*, Vol. 89, No. 4, April 2001, pp. 490–504.

[8] Y. Shin and T. Sakurai, "Power distribution analysis of VLSI interconnects using model order reduction," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 6, June 2002, pp. 739–745.

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

[10] S. Borkar, "Obeying Moore's law beyond 0.18 micron," in *Proceedings of the ASIC/SOC Conference*, September 2000, pp. 26–31.

[11] Y. I. Ismail, E. G. Friedman and J. L. Neves, "Figures of merit to characterize the importance of on-chip inductance," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 7, No. 4, December 1999, pp. 442–449.

[12] M. Popovich, A. V. Mezhiba and E. G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors*, Springer Science+ Business Media, New York, 2008.

[13] M. A. El-Moursy and E. G. Friedman, "Design methodologies for on-chip inductive interconnect," *Interconnect-Centric Design for Advanced SoC and NoC*, J. Nurmi, H. Tenhunen, J. Isoaho and A. Jantsch (Eds.), Kluwer Academic Publishers, Boston, 2004.

[14] H. B. Bakoglu, *Circuits, Interconnects, and Packaging for VLSI, Reading*, Addison-Wesley Publishing Company, MA, 1990.

[15] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE Journal of Solid-State Circuits*, Vol. 18, No. 4, August 1983, pp. 418–426.

[16] R. Antinone and G. W. Brown, "The modeling of resistive interconnects for integrated circuits," *IEEE Journal of Solid-State Circuit*, Vol. 18, No. 2, April 1983, pp. 200–203.

[17] G. Y. Yacoub, H. Pham, M. Ma and E. G. Friedman, "A system for critical path analysis based on back annotation and distributed interconnect impedance models," *Microelectronics Journal*, Vol. 19, No. 3, May/June 1988, pp. 21–30.

[18] N. Gopal, E. Tuncer, D. P. Neikirk and L. T. Pillage, "Non-uniform lumped models for transmission line analysis," in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, April 1992, pp. 119–121.

[19] T. Dhaene and D. D. Zutter, "Selection of lumper element models for coupled lossy transmission lines," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 11, No. 7, July 1992, pp. 805–815.

[20] L. Chang, K. Chang and R. Mathews, "When should on-chip inductance modeling become necessary for VLSI timing analysis?," in *Proceedings of the IEEE International Technology Conference*, 2000, pp. 170–172.

[21] A. Deutsch et al., "When are transmission-line effects important for on-chip interconnections?," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 45, No. 10, October 1997, pp. 1836–1846.

[22] B. Krauter, S. Mehrotra and V. Chandramouli, "Including inductive effects in interconnect timing analysis," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, May 1999, pp. 445–452.

[23] A. Deutsch et al., "Electrical characteristics of interconnections for high-performance systems," in *Proceedings of the IEEE*, Vol. 86, No. 2, February 1998, pp. 313–355.

[24] P. J. Restle, A. E. Ruehli, S. G. Walker and G. Papadopoulos, "Full-wave PEEC time-domain method for the modeling of on-chip interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 20, No. 7, July 2001, pp. 877–887.

[25] J. A. Davis and J. D. Meindl, "Compact distributed RLC interconnect models—Part I: Single line transient, time delay, and overshoot expressions," *IEEE Transactions on Electron Devices*, Vol. 47, No. 11, November 2000, pp. 2068–2077.

[26] J.A. Davis and J. D. Meindl, "Compact distributed RLC interconnect models—Part II: Coupled line transient expressions and peak crosstalk in multilevel networks," *IEEE Transactions on Electron Devices*, Vol. 47, No. 11, November 2000, pp. 2078–2087.

[27] Y. I. Ismail and E. G. Friedman, *On-chip Inductance in High Speed Integrated Circuits*, Kluwer Academic Publishers, Boston, 2001.

[28] A. Deutsch et al., "Functional high-speed characterization and modeling of a six-layer copper wiring structure and performance comparison with aluminum on-chip interconnections, in *Proceedings of the IEEE International Electron Devices Meeting*, December 1998, pp. 295–298.

[29] A. Deutsch et al., "Design guidelines for short, medium, and long on-chip interconnections," in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, October 1996, pp. 30–32.

[30] S. V. Morton, "On-chip inductance issues in multiconductor systems," in *Proceedings of the IEEE/ACM Design Automation Conference*, October 1999, pp. 921, 926.

[31] A. Deutsch et al., "The importance of inductance and inductive coupling for on-chip design guidelines for short, medium, and long on-chip interconnections," in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, October 1996, pp. 30–32.

[32] I. Catt, "Crosstalk (noise) in digital systems," *IEEE Transactions on Electronic Computers*, Vol. 16, No. 6, December 1967, pp. 743–763.

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

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

[35] L. Bisduonis, S. Nikolaidis, O. Koufopavlou and C. E. Goutis, "Modeling the CMOS short-circuit power dissipation," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol. 4, May 1996, pp. 469–472.

[36] 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. 19, No. 4, August 1984, pp. 468–473.

[37] 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 Application*, Vol. 41, No. 11, November 1994, pp. 762–766.

[38] A. M. Hill and S.-M. Kang, "Statistical estimation of short-circuit power in VLSI design," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol. 4, May 1996, pp. 105–108.

[39] A. Chanrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," in *Proceedings of the IEEE*, Vol. 83, No. 4, April 1995.

[40] J. Kao, S. Narendra and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, November 2002, pp. 141–148.

[41] R. Rao, A. Srivastava, D. Blaauw and D. Sylvester, "Statistical estimation of leakage current considering inter- and intra-die process variation," in *Proceedings of the International Symposium on Low Power Electronics and Design*, August 2003, pp. 84–89.

[42] S. Pullela, N. Menezes and L. T. Pillage, "Reliable non-zero skew clock trees using wire width optimization," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 1998, pp. 165–170.

[43] S. Pullela, N. Menezes and L. T. Pillage, "Moment-sensitivity-based wire sizing for skew reduction in on-chip clock nets," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 16, No. 2, February 1997, pp. 210–215.

[44] T. D. Hodes, B. A. McCoy and G. Robins, Dynamically wire-sized elmore-based routing constructions," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol. 1, May 1994, pp. 463–466.

[45] M. Edahiro, "Delay minimization for zero-skew routing," in *Proceedings of the IEEE International Conference on Computer-Aided Design*, November 1993, pp. 563, 566.

[46] S. S. Sapatnekar, "RC interconnect optimization under the elmore delay model," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 1994, pp. 387, 391.

[47] J. J. Cong and K. Leung, "Optimal wiresizing under elmore delay model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 14, No. 3, March 1995, pp. 321–336.

[48] S. Pullela, N. Menezes, L. T. Pillage, "Simultaneous gate and interconnect sizing for circuit-level delay optimization," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 1995, pp. 690–695.

[49] J. J. Cong and C.-K. Koh, "Simultaneous driver and wire sizing for performance and power optimization," *IEEE Transactions of Very Large Scale Integration (VLSI) Systems*, Vol. 2, No. 4, December 1994, pp. 408–425.

[50] C. P. Chen and N. Menezes, "Spec-based repeater insertion and wire sizing for on-chip interconnect," in *Proceedings of the IEEE International Conference on VLSI Design*, January 1999, pp. 476–483.

[51] W. C. Elmore, "The transient response of damped linear networks with particular regard to wideband amplifiers," *Journal of Applied Physics*, Vol. 19, No. 1, January 1948, pp. 55–63.

[52] M. A. El-Moursy and E. G. Friedman, "Optimizing inductive interconnect for low power," *System-on-Chip for Real-Time Applications*, W. Badawy and G. A. Jullien (Eds.), Kluwer Academic Publishers, 2003, pp. 380–391.

[53] M. A. El-Moursy and E. G. Friedman, "Power characteristics of inductive interconnect," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 12, December 2004, pp. 1295–1306.

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

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

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

[57] M. Borah, R. M. Owens and M. J. Irwin, "Transistor sizing for low power CMOS circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 15, No. 6, June 1996, pp. 665–671.

[58] 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, July 1997, pp. 1142–1145.

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

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

[61] T. Xiao and M. Marek-Sadowska, "Crosstalk reduction by transistor sizing," in *Proceedings of the Asia and Pacific Design Automation Conference*, January 1999, pp. 137–150.

[62] A. Vittal, L. H. Chen, M. Marek-Sadowska, K.-P. Wang and S. Yang, "Crosstalk in VLSI interconnection," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 12, December 1999, pp. 1817–1824.

[63] 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, November 1996, pp. 1–94.

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

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

[66] R. C. Jaeger, "Comments on 'an optimized output stage for mos integrated circuits'," *IEEE Journal of Solid-State Circuits*, Vol. 10, No. 3, June 1975, pp. 185–186.

[67] A. Kanuma, "CMOS circuit optimization," *Solid-State Electronics*, Vol. 26, No. 1, January 1983, pp. 47–58.

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

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

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

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

[72] N. Hedenstiera and K. O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 6, No. 2, March 1987, pp. 270–281.

[73] B. S. Cherkauer and E. G. Friedman, "A unified design methodology for CMOS tapered buffers," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 3, No. 1, March 1995, pp. 99–111.

[74] 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, September 1994, pp. 1142–1145.

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

[76] G. Chen and E. G. Friedman, "Low-power repeaters driving RC and RLC interconnects with delay and bandwidth constraints," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 14, No. 2, February 2006, pp. 161–172.

[77] 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, October 1990, pp. 1247–1256.

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

[79] M. Nekili and Y. Savaria, "Parallel regeneration of interconnections in VLSI & ULSI circuits," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 1992, pp. 2023–2026.

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

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

[82] 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, May 1998, pp. 607–616.

[83] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Transactions on Electron Devices*, Vol. 32, No. 5, May 1985, pp. 903–909.

[84] C. J. Alpert, A. Devgan, J. P. Fishburn and S. T. Quay, "Interconnect synthesis without wire tapering," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 20, No. 1, January 2001, pp. 90–104.

[85] M. A. El-Moursy and E. G. Friedman, "Optimum wire sizing of RLC interconnect with repeaters," *Integration, the VLSI Journal*, Vol. 38, 2004, pp. 205–225.

[86] Y. I. Ismail and E. G. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 8, No. 2, April 2000, pp. 195–206.

[87] M. A. El-Moursy and E. G. Friedman, "Optimum wire sizing of RLC interconnect with repeaters," in *Proceedings of the IEEE Great Lakes Symposium on VLSI*, April 2003, pp. 27–32.

[88] K. T. Tang and E. G. Friedman, "Delay uncertainty due to on-chip simultaneous switching noise in high performance CMOS integrated circuits," in *Proceedings of the IEEE Workshop on Signal Processing Systems*, October 2000, pp. 633–642.

[89] K. T. Tang and E. G. Friedman, "Incorporating voltage fluctuations of the power distribution network into the transient analysis of CMOS logic gates," *Analog Integrated Circuits and Signal Processing*, Vol. 31, No. 3, June 2002, pp. 249–259.

[90] M. Saint-Laurent and M. Swaminathan, "Impact of power supply noise on timing in high-frequency microprocessors," in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, October 2002, pp. 261–264.

[91] A. Waizman and C.-Y. Chung, "Package capacitor impact on microprocessor maximum operating frequency," in *Proceedings of the IEEE Electronic Components and Technology Conference*, June 2001, pp. 118–122.

[92] J. J. Clement, "Electromigration reliability," *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. Bowhill and F. Fox (Eds.), IEEE Press, New York, 2001, pp. 429–448, Chapter 20.

[93] I. A. Blech and H. Sello, *Mass Transport of Aluminum by Moment Exchange with Conducting Electrons*, Vol. 5, 1966, pp. 496–505. USAF-RADC Series

[94] J. R. Black, "Mass transport of aluminum by moment exchange with conducting electrons," in *Proceedings of the IEEE International Reliability Physics Symposium*, April 1967, pp. 148–159.

[95] E. G. Friedman, "Clock distribution networks in synchronous digital integrated circuits," in *Proceedings of the IEEE*, Vol. 89, No. 5, May 2001, pp. 665–690.

[96] 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, April 1991, pp. 930–934.

[97] S.Y. Kung, *VLSI Array Processors*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[98] M. Nekili, Y. Savaria, G. Bois and M. Bennani, "Logic-based H-trees for large VLSI processor arrays: A novel skew modeling and high-speed clocking method," in *Proceedings of the International Conference on Microelectronics*, December 1993, pp. 1–4.

[99] H. B. Bakoglu, J. T. Walker and J. D. Meindl, "A symmetric clock-distribution tree and optimized high-speed interconnections for reduced clock skew in ULSI and WSI circuits," in *Proceedings of the IEEE International Conference on Computer Design*, October 1996, pp. 118–122.

[100] H. Kojima, S. Tanaka and K. Sasaki, "Half-swing clocking scheme for 75% power saving in clocking circuitry," in *Proceedings of the IEEE Symposium on VLSI Circuits*, June 1994, pp. 23–24.

[101] D. W. Dobberpuhl et al., "A 200-MHz 65-b dual-issue CMOS RISC microprocessor," *IEEE Journal on Solid-State Circuits*, Vol. 27, November 1992, pp. 1555–1565.

[102] J. L. Neves and E. G. Friedman, "Minimizing power dissipation in nonzero skew-based clock distribution networks," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 1995, pp. 1576–1579.

[103] E. De Man and M. Schobinger, "Power dissipation in the clock system of highly pipelined ULSI CMOS circuits," in *Proceedings of the International Workshop on Low Power Design*, April 1994, pp. 133–138.

[104] M. Hatamian and G. L. Cash, "Parallel bit-level pipelined VLSI designs for high-speed signal processing," in *Proceedings of the IEEE*, Vol. 75, September 1987, pp. 1192–1202.

[105] M. Hatamian, L. A. Hornak, T. E. Little, S. T. Tewksbury and P. Franzon, "Fundamental interconnection issues," *AT&T Technical Journal*, Vol. 66, July/August 1987, pp. 13–30.

[106] J. P. Fishburn, "Clock skew optimization," *IEEE Transactions on Computation*, Vol. 39, July 1990, pp. 945–951.

[107] E. G. Friedman, "Performance limitations in synchronous digital systems," Ph.D. dissertation, University of California, Irvine, June 1989.

[108] M. Hatamian, "Understanding clock skew in synchronous systems," *Concurrent Computations (Algorithms, Architectures and Technology)*, S. K. Tewksbury, B. W. Dickinson and S. C. Schwartz (Eds.), Plenum, New York, 1988, pp. 87–96.

[109] J. W. Joyner et al., "Impact of three-dimensional architectures on interconnects in gigascale integration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 9, No. 6, December 2001, pp. 922–928.

[110] J. W. Joyner and J. D. Meindl, "Opportunities for reduced power distribution using three-dimensional integration," in *Proceedings of the IEEE International Interconnect Technology Conference*, June 2002, pp. 148–150.

[111] W. E. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Transactions on Circuits and Systems*, Vol. 26, No. 4, April 1999, pp. 272–277.

[112] D. Stroobandt, *A Priori Wire Length Estimates for Digital Design*, Kluwer Academic Publishers, Netherlands, 2001.

[113] B. Kim et al., "Factors affecting copper filling process within high aspect ratio deep vias for 3D chip stacking," in *Proceedings of the IEEE International Electronic Components and Technology Conference*, June 2006, pp. 838–843.

[114] M. W. Newman et al., "Fabrication and electrical characterization of 3D vertical interconnects," in *Proceedings of the IEEE International Electronic Components and Technology Conference*, June 2006, pp. 394–398.

[115] N. T. Nguyen et al., "Through-wafer copper electroplating for three-dimensional interconnects," *Journal of Micromechanics and Microengineering*, Vol. 12, No. 4, July 2002, pp. 395–399.

[116] C. S. Premachandran et al., "A vertical wafer level packaging using through hole filled via interconnect by lift-off polymer method for MEMS and 3D stacking applications," in *Proceedings of the IEEE International Electronic Components and Technology Conference*, June 2005, pp. 1094–1098.

[117] V. F. Pavlidis, *Interconnect-Based Design Methodologies for Three-Dimensional Integrated Circuits*, Ph.D. Dissertation, University of Rochester, Rochester, New York, June 2008.