CHAPTER 3

Interconnect Modeling and Extraction

Global interconnects play an important role in determining the overall performance, power consumption, signal integrity, and reliability of an integrated circuit (IC). The increasing importance of interconnects is the result of device and interconnect scaling, as described in Chapter 2. This trend has caused the emergence of interconnect driven design methodologies for high performance ICs. These methodologies affect the design process in a multitude of ways. A variety of design criteria exist for understanding the effect of the interconnects on the overall performance parameters such as latency, bandwidth, power dissipation, noise, and area. These interconnect design criteria are described in Section 3.1.

Accurate models of on-chip interconnects are critical in predicting and satisfying the performance characteristics of high performance ICs. The nature of an interconnect model depends upon certain circuit characteristics such as the process technology, operating frequency, and physical dimensions of the interconnects and devices.

An interconnect connecting two logic gates was first modeled as a short circuit, assuming metal with negligible parasitic impedances. As the device dimensions have decreased and chip dimensions have increased, the interconnect model has evolved from an ideal no impedance metal line to a simple capacitive (C) model, later to a capacitive and resistive (RC) model, and finally to a capacitive, resistive, and inductive (RLC) model. This evolution of an interconnect model is illustrated in Fig. 3.1.

The accurate model of an interconnect, i.e., short circuit, C, RC, or RLC, depends upon both global and local parameters. The global parameters that affect the model of an interconnect are the process technology and operating frequency. Alternatively, the local parameters that play a role in the interconnect model are the physical characteristics of the interconnect (such as the length, width, and thickness of the line and the type of material), and the size of the driver and driven gates.

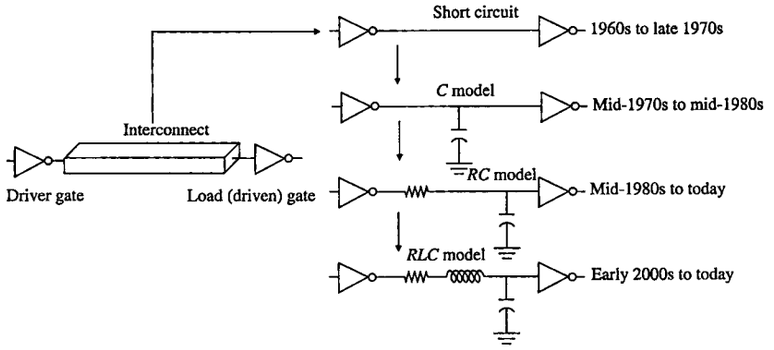

FIGURE 3.1 Evolution of an on-chip interconnect model from a short circuit to a simple capacitive line (C), capacitive and resistive line (RC), and finally a capacitive, resistive, and inductive line (RLC).

For example, a C model is required to represent an interconnect if the interconnect capacitance is comparable to the gate capacitance of the driven gate [5]. Similarly, an RC model should be adopted if the magnitude of the interconnect resistance is comparable to the gate resistance of the driver [5]. The parasitic inductance should also be considered under certain conditions, requiring an RLC model, as described in Section 3.4.3.

The capacitive characteristics of interconnects and the capacitance extraction process are described in Section 3.2. The resistive characteristics of interconnects and the resistance extraction process are described in Section 3.3. An overview of on-chip inductance, the inductive characteristics of interconnects, and the inductance extraction process are described in Section 3.4. The chapter is summarized in Section 3.5.

3.1 Interconnect Design Criteria

Interconnect optimization has started to receive considerable attention throughout the overall design process of high performance ICs. These interconnects, once modeled as a short circuit with no parasitic impedances, now have fairly complicated models to accurately analyze the effects of the interconnects on various system parameters such as circuit speed, power dissipation, signal noise, and physical area. Multiple criteria should therefore be considered during the interconnect design process, such as latency, power dissipation, bandwidth, noise, and physical area. These criteria are individually discussed in the following subsections.

3.1.1 Latency

Interconnect latency or interconnect delay is a primary design criterion since the delay of the global and semi-global interconnects play an important role in the overall system performance. Early interconnect design methodologies focused primarily on delay optimization [74], [112].

Most systems utilize a global temporal reference to manage the flow of events. This temporal reference is typically called a clock signal, which is used to synchronize the system. A simple synchronous digital circuit is shown in Fig. 3.2. Assuming the clock arrives at the registers at the same time, i.e., the difference in delay is zero (zero clock skew), the clock period T_{clock} should satisfy

$$T_{period} = \frac{1}{f_{clock}} \geq T_{C-Q} + T_{int} + T_{logic} + T_{setup} \tag{3.1}$$

where T_{C-Q} is the clock-to-Q delay of the initial register, T_{int} is the interconnect delay, T_{logic} is the delay of the combinational logic circuitry, and T_{setup} is the required set-up time of the receiving register. Note that a comprehensive discussion on synchronization is provided in Chapter 12.

Assuming the data path illustrated in Fig. 3.2 is a worst case (or critical) path, (3.1) states that the overall system speed of the circuit, i.e., the clock frequency, can be increased by reducing the interconnect delay T_{int}. Note that the effect of the interconnect delay on system speed is significant for global and semi-global interconnects where the interconnect delay T_{int} can be much greater than the gate delay T_{logic}. The signal propagation characteristics along the interconnects and the interconnect delay are further described in Chapter 4.

A reduction in interconnect delay not only increases the system speed, but also improves the computational efficiency. In advanced microprocessors, multiple computational cores are fabricated on the same die [113]. Communication among these cores and the on-chip

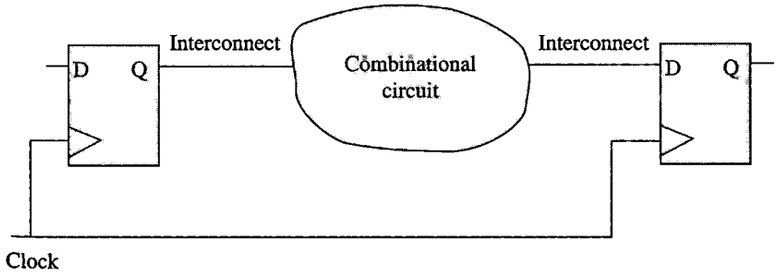

FIGURE 3.2 Simple synchronous circuit consisting of a combinational logic and two registers.

memory generally requires multiple clock cycles. The computational core sometimes enters into an idle state waiting for the required data or control signals from other regions within the IC. The computational resources of these cores, therefore, cannot be efficiently utilized due to the large amount of multicycle communication. By reducing the interconnect delay, the speed of the system, i.e., the computational efficiency of the cores, can be improved.

3.1.2 Bandwidth

The concept of bandwidth originated from the telecommunications field [114]. According to the traditional definition, bandwidth refers to the range or *band* of frequencies of a signal transmitted over a transmission medium such as a communication channel or a filter. Bandwidth is expressed in *hertz* as the difference of the maximum and minimum frequency components of the signal being transmitted.

For on-chip applications, bandwidth is used to measure the data transmission capacity of an interconnect. Specifically, bandwidth refers to the maximum number of bits that an interconnect can reliably transfer per second, i.e., the maximum bit rate [115]. Based on this definition, bandwidth B is

$$B = \frac{1}{(T_{bit})_{min}} \text{ (bit/sec)} \tag{3.2}$$

where $(T_{bit})_{min}$ is the minimum bit period that can be reliably transmitted. A bit period T_{bit} can be divided into two parts, as shown in Fig. 3.3 [116]. One part is dedicated to the transition time $t_r = t_f$, while the other part is the steady state time t_s during which the data can be latched within the receiving register. Assuming the steady state time occupies at least half of the bit period to maintain reliable transmission, the bandwidth can be expressed in terms of the

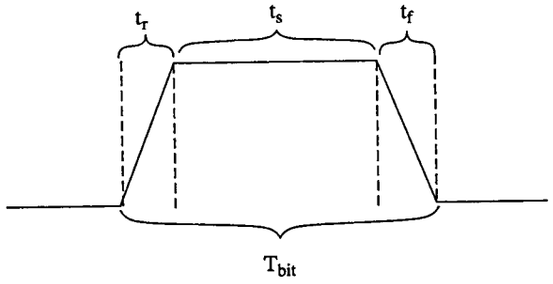

FIGURE 3.3 Bit period consisting of two parts: transition time and steady state time.

transition time $t_t = t_r = t_f$ as

$$B = \frac{1}{(T_{bit})_{min}} = \frac{1}{4t_t} \text{ (bit/sec)} \tag{3.3}$$

The bandwidth is proportional to the reciprocal of the delay, i.e., a higher bandwidth reduces the total time required to transmit a certain amount of data, thereby increasing the speed of the system. A lower delay typically implies a smaller minimum bit period, and therefore a higher bandwidth. The proportionality between the delay and the bandwidth is assumed to be inversely linear [117], [30], [118], [119]. This assumption, however, is only valid for *RC* lines, where an approximately linear relationship exists between the 50% delay and the bit period.

It is typically important to optimize the *overall* bandwidth rather than the bandwidth of a single interconnect [119]. Furthermore, the delay (or latency) and overall bandwidth should be simultaneously considered. For example, using wider wires can reduce the delay of a single interconnect, but can also reduce the overall bandwidth since the wiring density is lower [117, 118]. This situation is illustrated in Fig. 3.4.

Bandwidth is an important design criterion, particularly for global interconnects. For example, for those interconnects between the cache memory and processor or among various processors within a multicore architecture, bandwidth plays an important role in determining the overall performance [118]. Furthermore, multiplexing in time multiple bits using a single wire is an attractive option to reduce the area of on-chip data buses [115]. The bandwidth of the interconnect, however, is a limiting factor for time multiplexing.

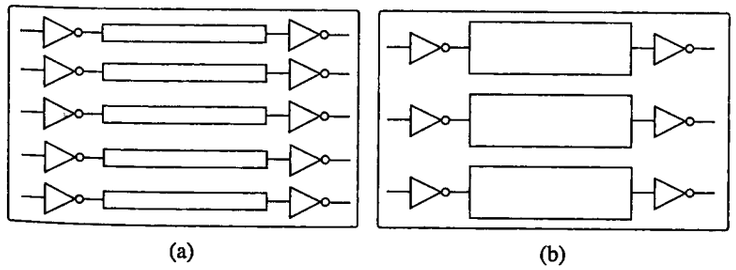

FIGURE 3.4 Optimization of overall bandwidth and delay: (a) Thinner wires exhibit a higher wire density. The overall bandwidth can be increased, but the delay per line is smaller, and (b) Wider lines have lower individual delay, but the overall bandwidth can be smaller.

3.1.3 Noise

The density of the devices within a circuit has dramatically increased, making the related issues of *noise* and *noise coupling* among various circuit elements a primary concern for nanoscale ICs. Traditionally, digital circuits consisting of static CMOS gates have been treated as inherently immune to noise due to relatively high noise margins [120,121]. Consequently, the concept of *noise* in an IC has long been associated with analog, RF, and dynamic CMOS circuits due to the higher sensitivity to noise of these types of circuits [122]. This situation, however, has changed in the last decade. Three fundamental reasons have stimulated this change:

- Noise margins have been reduced significantly due to scaling of the power supply and threshold voltages, making digital logic circuits more sensitive to noise.
- Operating frequencies have substantially increased, placing more stringent constraints on the timing requirements of the critical paths. The sensitivity of the circuit delay to noise has therefore become more significant. Furthermore, higher clock rates have enabled faster on-chip signal transitions, exacerbating the magnitude of the coupling noise.
- The delay of the interconnects has become comparable or greater than the delay of the logic gates in deep submicrometer technologies. As such, the impact of interconnect noise on the signal characteristics, i.e., *signal integrity*, as well as the overall system performance has become significant.

As a result of these three reasons, noise in digital ICs has started to receive considerable attention.

The simultaneous operation of an enormous number of devices switching at a high clock frequency with reduced physical distances among the interconnects causes two primary types of *switching noise* in a synchronous digital IC: power supply noise and interconnect noise.

Power and ground noise refers to voltage fluctuations, respectively, on the power and ground distribution networks. The simultaneous switching of a large number of logic gates requires a significant amount of current from the power supply. This current flows through the parasitic impedance of the power distribution network, causing both static and dynamic voltage fluctuations [123], [124]. Similarly, the current flowing from the switching logic gates to the reference ground of the power supply causes voltage fluctuations on the ground distribution network, referred to as ground noise or ground bounce [125]. A comprehensive discussion of power and ground noise is provided in Chapter 8.

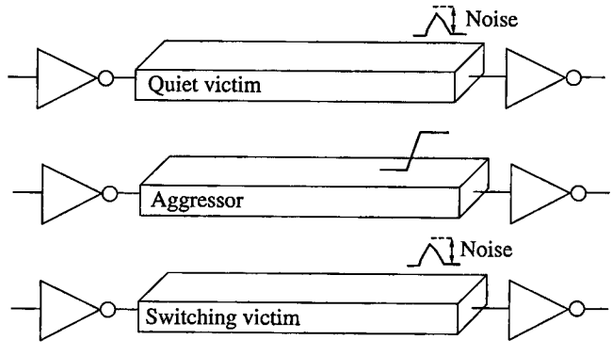

FIGURE 3.5 Interconnect noise coupling from a switching aggressor to a victim.

Interconnect noise, or *crosstalk*, refers to a voltage induced on a *victim* node due to capacitive and/or inductive coupling from a switching *aggressor* node, as depicted in Fig. 3.5. Note that a victim can either be a switching line or a quiet line.

Electromagnetic coupling has long been a primary concern for RF circuits, microwave circuits, and high speed printed circuit (PC) boards [126]. The effects of on-chip crosstalk in high speed digital ICs have started to become significant in the last decade due to asymmetric scaling between the vertical and lateral dimensions, and faster signal transitions in deep submicrometer technologies. Specifically, the lateral dimensions have been scaled to enhance performance and achieve higher density, while the vertical dimensions have not significantly changed [127]. Coupling among interconnects in a digital IC has therefore increased significantly. Similarly, the speed of on-chip signal transitions and the length of interconnects have increased, exacerbating on-chip inductive effects [128]. Noise is therefore a primary design criterion for on-chip interconnects. Note that a comprehensive discussion of interconnect coupling noise and the effects of this noise on circuit operation is provided in Chapter 5.

3.1.4 Power Dissipation

Due to increasing clock frequencies and higher levels of on-chip integration, power dissipation has significantly increased. The on-chip power dissipation of a modern microprocessor is on the order of hundreds of watts [129]. This power dissipation has various components such as dynamic, short-circuit, and leakage. Although the leakage component has been increasing, the dynamic component remains an

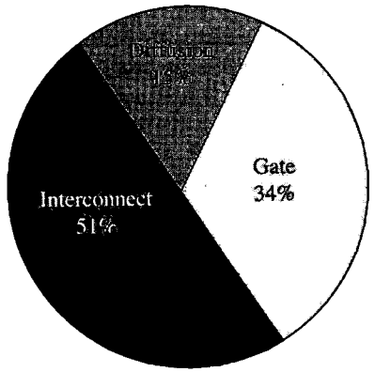

FIGURE 3.6 Components of dynamic power dissipation in a modern microprocessor due to different capacitive sources: gate capacitance, diffusion capacitance, and interconnect capacitance [130].

important fraction of the overall power dissipation. The dynamic power component can also be divided into several parts based on the type of capacitance being charged and discharged: (1) gate capacitance, (2) diffusion capacitance, and (3) interconnect capacitance. These different dynamic power components in a modern microprocessor are illustrated in Fig. 3.6 [130]. The dynamic power due to the interconnect capacitance can be greater than 50% of the total dynamic power. Furthermore, it is common practice to enhance the delay and noise characteristics of an interconnect by inserting repeaters and pipeline registers along the interconnect. These repeaters and registers consume additional power. High power dissipation increases the cost of packaging due to greater heating, and shortens the battery life in portable applications. Power dissipation is therefore a fundamental criterion in interconnect design. A comprehensive discussion of power dissipation is provided in Chapter 11.

3.1.5 Physical Area

The continuous development of CMOS technology has enabled the integration of more than two billion transistors on a single IC [131]. The number of interconnects has therefore also significantly increased. The number of metal layers is also much greater to provide sufficient metal resources for interconnect routing, reaching up to eleven layers in a modern IC technology [113]. The additional number of metal layers, however, increases the fabrication cost. Furthermore, the buffers and pipeline registers inserted along the interconnects make the constraint on silicon area more stringent. The area criterion should therefore be considered during the interconnect design process, particularly on a variety of design techniques such as wire sizing and repeater insertion.

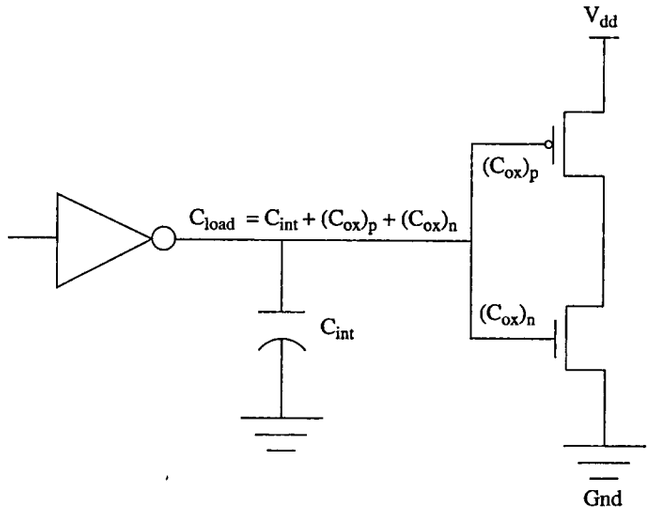

FIGURE 3.7 The overall load capacitance consists of the gate oxide capacitance of the driven gates and the capacitance of the interconnect connecting the driver and driven gates.

3.2 Interconnect Capacitance

The overall load capacitance driven by a gate consists of two components: (1) the gate oxide capacitance of the driven gates, and (2) the parasitic capacitance of the interconnect connecting the driver and driven gates, as shown in Fig. 3.7. The interconnect capacitance should be considered if the line capacitance is comparable to the gate oxide capacitance of the driven gates [5].

With decreasing device feature size and increasing die size, the interconnect capacitance has become significantly more important, particularly since the gate capacitance has been decreasing while the interconnect capacitance has increased. The different components of interconnect capacitances are described in Section 3.2.1. Techniques to extract the on-chip parasitic capacitance are discussed in Section 3.2.2.

3.2.1 Components of Interconnect Capacitance

There are three components of interconnect capacitance:

- Parallel plate capacitance, also referred to as metal-to-substrate excluding fringing effects

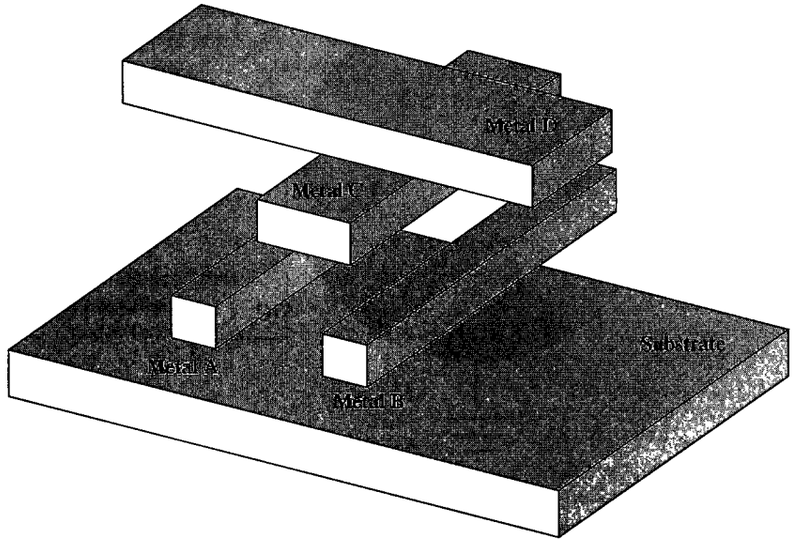

FIGURE 3.8 Three metal layer structure where metal lines A and B are located on the first layer above the substrate. Metal lines C and D are located, respectively, on layers two and three.

- Crossover capacitance, also referred to as overlap capacitance, between two nodes on different layers including fringing effects
- Lateral capacitance, also referred to as sidewall, interwire, or coupling capacitance between two nodes on the same layer

A three metal layer structure is depicted in Fig. 3.8. Metal lines A and B are located on the same layer above the substrate. Metal line C is on the second metal layer, and metal line D is located on the third metal layer. An insulator layer, typically silicon dioxide (SiO_2), is placed between each metal layer to electrically isolate the layers from each other. The three capacitive components among these interconnect lines are shown in Fig. 3.9, where the cross section of Fig. 3.8 is depicted. Note that the fringe capacitance represents the edge effects and has two components. The first component originates from the sidewall and the second component originates from the top plane of the interconnect. Also note that the ground capacitance refers to the summation of the parallel plate capacitance and fringe capacitance to ground.

Referring to Fig. 2.21, the parallel plate capacitance C_{pp} between a metal line and the substrate is

$$C_{pp} = \frac{\epsilon L_{int} W_{int}}{H} \tag{3.4}$$

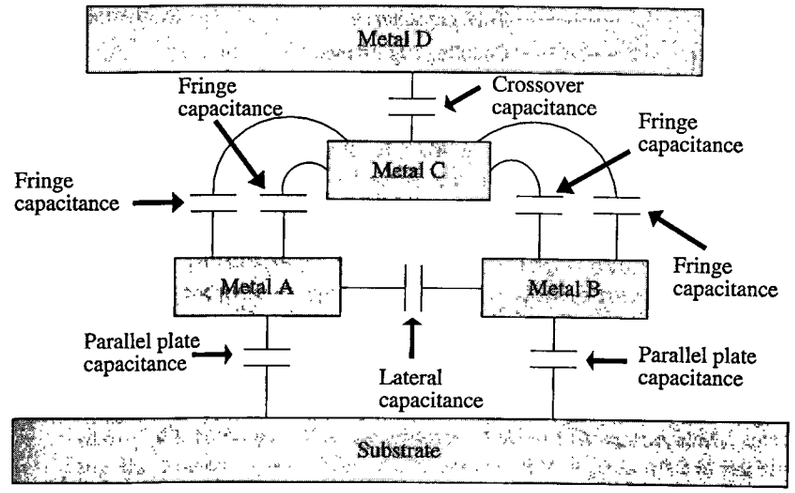

FIGURE 3.9 Components of interconnect capacitance among the three metal layers and the substrate.

where ϵ is the dielectric constant of the insulator. The lateral or coupling capacitance C_c is

$$C_c = \frac{\epsilon L_{int} T_{int}}{W_{spa}} \tag{3.5}$$

In early ICs, the parallel plate capacitance dominated the overall capacitance. However, as a result of nonideal scaling, the width W_{int} of an interconnect decreased more than the thickness T_{int} of the interconnect, causing an increase in the aspect ratio $AR = T_{int}/W_{int}$. As such, the contribution of the fringe and lateral capacitances has increased significantly [132]. For $AR = 4$, the conventional parallel plate capacitance as determined by (3.4) can underestimate the overall capacitance by as much as 70% [133]. The fringe and lateral capacitances therefore cannot be neglected when estimating the overall capacitance. This behavior is further illustrated in Fig. 3.10, where the contribution of these capacitive components is depicted as a function of line width for a constant line thickness [134]. Note that the spacing between the lines is equal to the line width. Two important conclusions from this graph are:

- The parallel plate capacitance significantly deviates from the ground capacitance due to fringing effects
- Coupling capacitance dominates as the line width and spacing between two lines are reduced

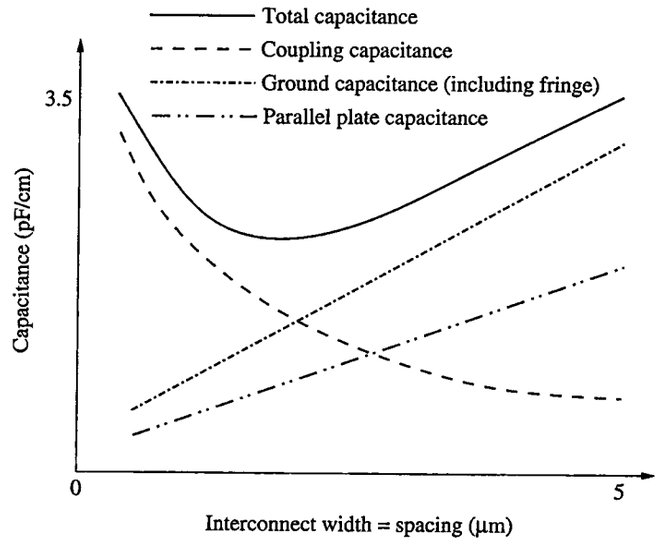

FIGURE 3.10 Contribution of coupling capacitance and ground capacitance to the overall capacitance. The parallel plate capacitance is also shown as a reference.

Note that the effect of a coupling capacitance on the overall signal integrity of the circuit is described in Chapter 5. Techniques to extract the capacitance considering fringing effects are discussed in the following subsection.

3.2.2 Interconnect Capacitance Extraction

As described in the previous section, a conventional parallel plate capacitance model is insufficient to accurately estimate the overall capacitance. Several techniques exist to determine the capacitance of a specific geometry representing an interconnect structure such as shown in Fig. 3.8. These techniques exhibit different advantages and disadvantages in terms of accuracy and computational complexity. For example, using electromagnetic theory to extract the capacitance produces highly accurate results at the expense of increased computational complexity. Alternatively, compact models can be developed to estimate the capacitance with moderate accuracy and low computational complexity.

The process of determining the capacitance should fundamentally treat the interconnect as a three-dimensional (3-D) physical structure. Poisson's equation is solved within this 3-D structure. In electrostatics, Poisson's equation relates the electric potential field ϕ to the charge

density ρ,

$$\Delta^2\phi = \frac{-\rho}{\epsilon} \tag{3.6}$$

An exact analytic solution of (3.6) is possible only for sufficiently simple geometries. Closed-form solutions of (3.6) for practical circuits, however, do not exist [135]. Alternatively, (3.6) can be numerically solved using finite difference [136] or finite element methods [137] to determine the capacitance. Finite difference or finite element methods discretize the volume to determine the electric potential at each point within the grid. These potentials are used to extract the capacitance by applying Gauss law, as described by Dierking and Bastian [136]. Another approach is to use the boundary element method with an appropriate Green's function to discretize (3.6) [138]. Note that the capacitance obtained by numerically solving the Poisson's equation considers fringing effects and is highly accurate [139].

The primary disadvantage of applying these numerical techniques to solve Poisson's equation is high computational complexity. For a typical high complexity IC, the process of extracting the parasitic capacitances using finite difference, finite element, or boundary element methods is computationally prohibitive.

An alternative approach to determine the parasitic capacitance is to develop compact models. These models can be developed either by a simplified analytic solution or by *fitting* the data obtained from numerical simulations or measurements. The models developed by fitting are also called empirical models. These simplified analytic and empirical models provide fast estimation of the parasitic capacitances with moderate accuracy, while providing physical intuition. These models, however, are typically valid for a specific range of width and thickness of the interconnect, and insulator thickness. It is important that these models are sufficiently accurate over a reasonably large range of on-chip interconnect characteristics.

Analytic formulas have been proposed by Chang [140] in 1976 using conformal transformation techniques. The accuracy of these models is within 1%, provided that the width of the interconnect is greater than the dielectric thickness.

In 1982, Elmasry [133] developed a significantly simpler analytic expression to estimate the parasitic capacitance, exhibiting less than 10% error as compared to a two-dimensional numerical solution. Elmasry partitioned the conductor into four equipotentials, as illustrated in Fig. 3.11, and estimated the overall capacitance C as the parallel connection of three capacitive components, C_1, C_2, and C_3,

$$\frac{C}{C_1} = 1 + 2\frac{H}{W_{int}} ln\left(1 + \frac{T_{int}}{H}\right) + 2\frac{T_{int}}{W_{int}} ln\left(1 + \frac{W_{int}/2}{H + T_{int}}\right) \tag{3.7}$$

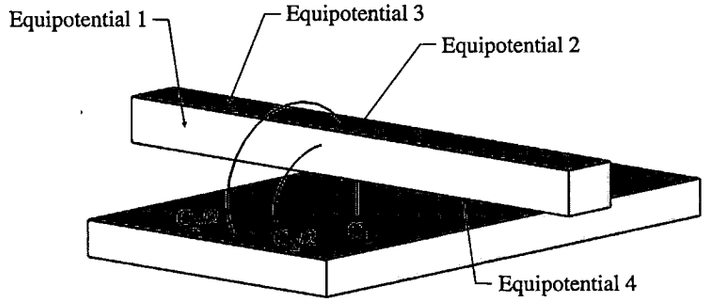

FIGURE 3.11 Four equipotentials based on the Elmasry model. Two sidewalls are represented as equipotentials 1 and 2. The top plate and bottom plate represent, respectively, equipotentials 3 and 4. The overall capacitance is the parallel connection of the three capacitive components, C_1, C_2, and C_3.

where C_1 is the conventional parallel plate capacitance as determined by $\epsilon L_{int} W_{int}/H$, and the second and third terms represent, respectively, the sidewall fringe capacitance C_2 and the topwall fringe capacitance C_3. Note that the interactions among the equipotentials are neglected, resulting in this simple expression.

Sakurai and Tamaru [141] developed an empirical formula rather than provide an enhanced analytic approach. From Sakurai and Tamaru, the capacitance per unit length of a single line is

$$C = \epsilon\left[1.15\left(\frac{W_{int}}{H}\right) + 2.80\left(\frac{T_{int}}{H}\right)^{0.222}\right] \tag{3.8}$$

where the first term represents the bottom and top plates of the interconnect and the second term represents the sidewalls. The accuracy is within 6% for $0.3 < W_{int}/H < 30$ and $0.3 < T_{int}/H < 30$. This expression was obtained by fitting the results from a two-dimensional numerical analysis. Note that the authors also proposed expressions for the lateral (coupling) capacitive component, which dominates in high aspect ratio interconnect with small wire spacing.

Another empirical formula to estimate interconnect capacitance was developed by Chern et al. [142]. Simpler structures were exploited to determine the capacitance of *multilevel metal capacitances*, exhibiting an accuracy within 8% of simulations and measurements. According to Chern et al., the ground capacitance per unit length of a line is

$$C = \epsilon\left[\frac{W_{int}}{H} + 3.28\left(\frac{T_{int}}{T_{int}+2H}\right)^{0.23} + \left(\frac{W_{spa}}{W_{spa}+2H}\right)^{1.16}\right] \tag{3.9}$$

where W_{spa} is the spacing between two parallel lines on the same layer. Note the dependence of the ground capacitance on W_{spa}. As the

spacing between the lines increases, the ground capacitance slowly increases since additional electric field lines couple to ground from the sidewall of the interconnect.

Although efficient and sufficiently accurate, these compact models have been developed for a specific process and geometry, and therefore exhibit greater error when the process or geometry changes. A methodology for automatically generating compact models for interconnect capacitances is described by Umakanta [132] to overcome this issue.

A different and practical technique to determine the interconnect capacitance is to provide look-up tables where the numerical value of the capacitance (typically obtained from 3-D numerical simulations or measurement results) is stored in a table as a function of several interconnect parameters such as the metal width, metal thickness, and insulator thickness [132]. The dimensions of these look-up tables and memory requirements grow rapidly as the number of parameters increases. Furthermore, the number and specific value of these instances for each parameter are critical to accurately interpolate the capacitance for those values not included in the look-up tables. Although look-up tables provide a practical solution to the issue of capacitance extraction, another disadvantage of these look-up tables is the lack of physical intuition and insight during the circuit design process.

A typical flow for a modern 3-D capacitance extraction process is shown in Fig. 3.12 [143]. The first step is to generate the fundamental test patterns where each pattern corresponds to a specific geometry or building block of a more complicated interconnect structure. These test patterns are measured or simulated (using a 3-D field solver) to obtain realistic data. The generated data are used for two reasons. The first reason is to generate look-up tables and the second reason is to derive compact models by fitting the data to analytic expressions. In either case, the next step is to extract the geometric parameters of the interconnects for the circuit under examination. The last step is to match these geometric parameters to the test patterns, and use look-up tables or compact models to determine the capacitance values, as illustrated in Fig. 3.12. This technique is called *pattern matching* [143]. Note that only the nearest neighbors are considered during the capacitance extraction process since the electrostatic interaction of the capacitance exhibits a short-range behavior.

3.3 Interconnect Resistance

The resistance characterizes the ability of an interconnect to pass electrical charge. The parasitic resistance of an interconnect becomes important if this resistance is comparable to the channel on-resistance of the driving transistor [5], as depicted in Fig. 3.13. Assuming the

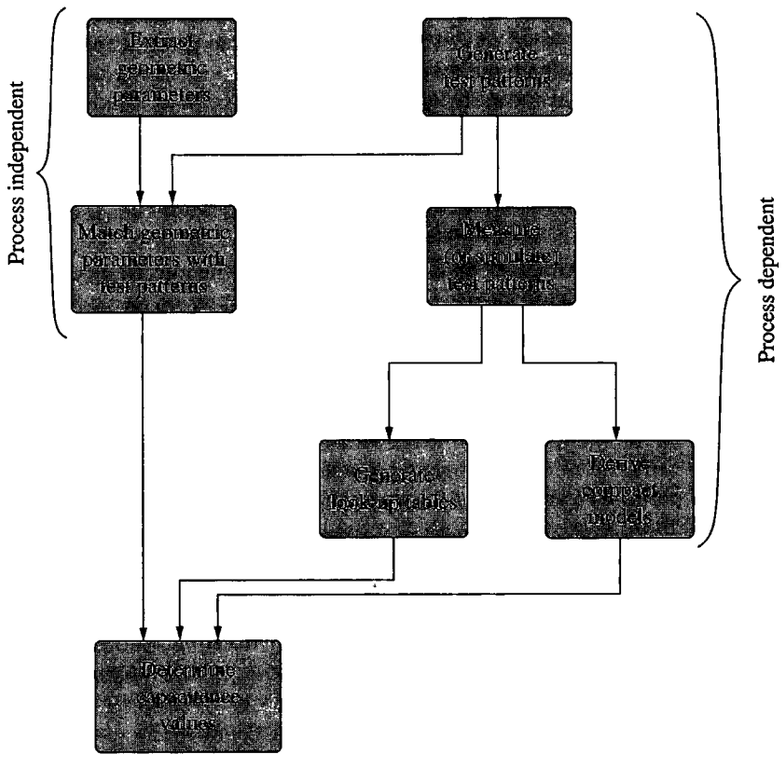

FIGURE 3.12 Flow of 3-D capacitance extraction process.

driving transistor operates in the triode region, a first order approximation for the on-resistance R_{tr} of a transistor is

$$R_{tr} \approx \frac{1}{\mu C_{ox} \frac{W}{L}(V_{GS} - V_{TH})} \quad . \tag{3.10}$$

where μ is the mobility of the charge carriers, C_{ox} is the gate capacitance, and W and L are, respectively, the channel width and length of the transistor. The transistor on-resistance R_{tr} remains approximately constant with scaling since the increase in C_{ox} (due to the reduction in oxide thickness) is partly compensated by the decrease in $V_{GS} - V_{TH}$ (note that for technologies below 0.18 μm, V_{GS} scales more than V_{TH}, resulting in an effective decrease in $V_{GS} - V_{TH}$) [5]. Alternatively, the interconnect resistance R_{int} of the global lines significantly increases with scaling, as described in Chapter 2. Note that the transistor operates in both the triode region and the saturation region during the signal transition. When the transistor is operating in saturation, the

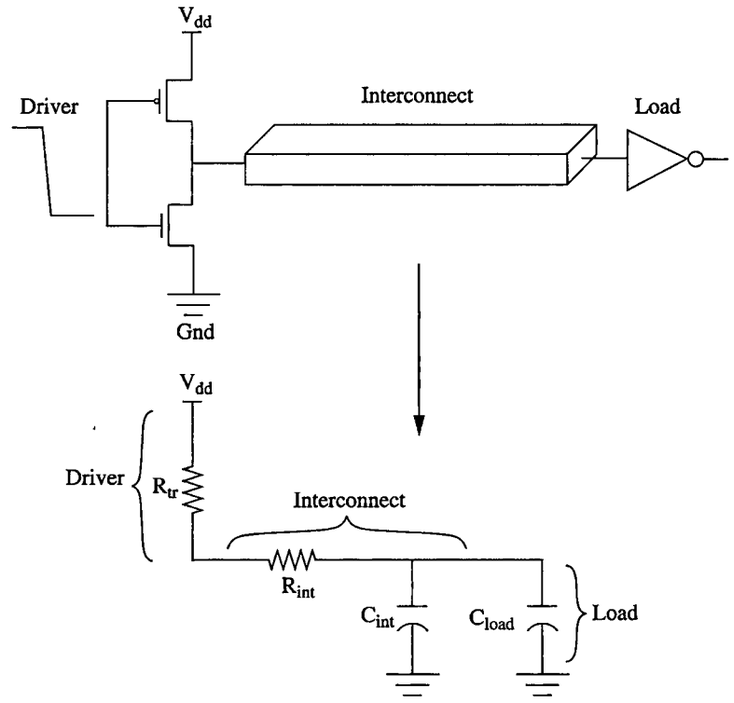

FIGURE 3.13 Interconnect resistance R_{int} should be considered if R_{int} is comparable to the channel on-resistance R_{tr} of the transistor.

channel resistance is significantly higher since the drain current is relatively constant.

The increase in interconnect resistance has deleterious effects on the global signal lines (such as increased delay and degradation in the signal waveform characteristics), global clock lines (inducing greater clock skew), and global power distribution lines (producing voltage variations caused by IR drops). These effects are discussed in the following parts of this book.

The resistivity of copper interconnects and the parameters that affect this resistivity are explained in Section 3.3.1. Techniques to extract the interconnect resistance are discussed in Section 3.3.2.

3.3.1 Copper Resistivity

The resistance of a conductor with a rectangular cross section, as depicted in Fig. 3.14, is

$$R = \rho \frac{l}{WH} \tag{3.11}$$

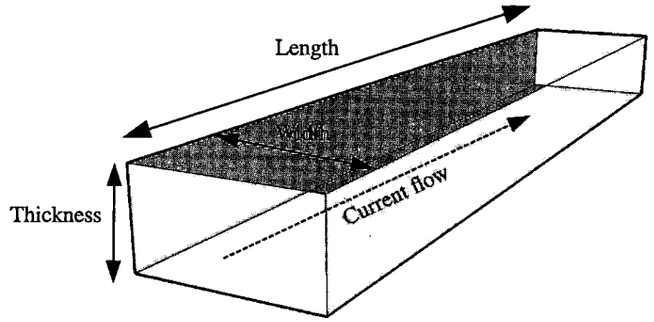

FIGURE 3.14 Conductor with a rectangular cross section representing an on-chip interconnect.

where ρ is the resistivity of the material, and l, W, and H are, respectively, the length, width, and thickness of the interconnect. In current CMOS technologies, aluminum interconnect has been replaced with copper interconnect to relieve the interconnect bottleneck.

At 22°C, pure copper has a lower resistivity of approximately 1.7 $\mu\Omega$-cm as compared to the resistivity of pure aluminum, 2.7 $\mu\Omega$-cm [79]. Furthermore, copper has enhanced electromigration behavior as compared to aluminum, permitting higher current densities [79], [144]. Yet another advantage of copper interconnect is the relatively easier procedure of obtaining a clean interface, reducing the contact resistance among different metal layers [79].

The *effective resistivity*, however, deviates from the ideal or pure resistivity due to several reasons caused by certain processing steps, operating temperature, and frequency. These effects are discussed in the following subsections.

Diffusion Barrier

A thin barrier layer is built on three sides of an on-chip copper interconnect to prevent copper from diffusing into the surrounding dielectric [144], as shown in Fig. 3.15. Although this diffusion process is negligible in aluminum wires [30], this barrier layer increases the effective resistivity of copper since a portion of the cross sectional area is consumed by this highly resistive barrier. The increase in resistivity due to the barrier layer is dependent upon the minimum required barrier thickness and the cross sectional barrier profile as determined by the deposition technology [144]. Referring to Fig. 3.15 and assuming the diffusion barrier does not conduct, the effective resistivity ρ_b of copper is [144]

$$\rho_b = \rho_0 \frac{1}{1 - \frac{A_b}{WH}} \tag{3.12}$$

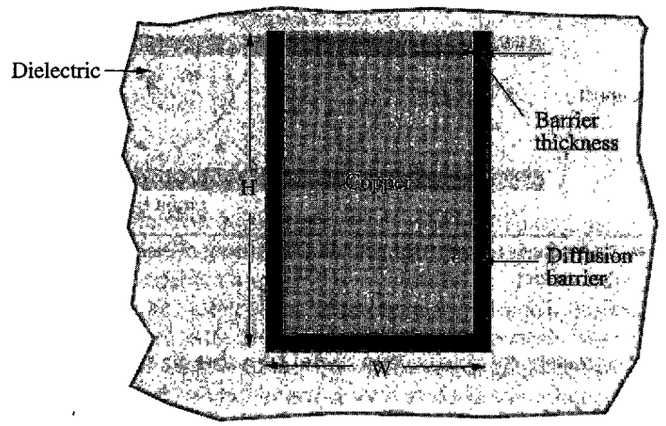

FIGURE 3.15 Cross section of an on-chip copper interconnect illustrating the barrier layer built on three sides to prevent copper from diffusing into the surrounding dielectric.

where ρ_0 is the bulk resistivity at a specific temperature, A_b is the cross sectional area of the barrier, and $W \times H$ is the cross sectional area of the interconnect.

Note that the effect of the diffusion barrier becomes more dominant as the copper interconnect dimensions are scaled since the barrier does not scale at the same rate as the interconnect due to reliability constraints [144]. Consequently, the cross sectional area occupied by the diffusion barrier is a larger fraction of the overall area of the interconnect, increasing the effective resistivity of the copper.

Surface and Grain Boundary Scattering

The second factor affecting the resistivity of copper interconnect is a mechanism called surface and grain boundary scattering [145]. This mechanism is also referred to as the size effect of interconnect lines after the Fuchs' size effect theory [146]. Surface and grain boundary scattering is typically ignored in older technologies where the dimensions of the interconnect are relatively greater. As the dimensions of the on-chip interconnect are reduced in newer technologies, the mean free path of the electrons (the average distance an electron travels between collisions with other electrons) becomes comparable to the interconnect dimensions, increasing the effect of surface and grain boundary scattering [79], [144], [145], as illustrated in Fig. 3.16. When the physical dimensions (thickness and/or width) of a wire shrink, the electrons experience additional collisions and reflections at the surface, increasing the effective resistivity. An approximate expression for the

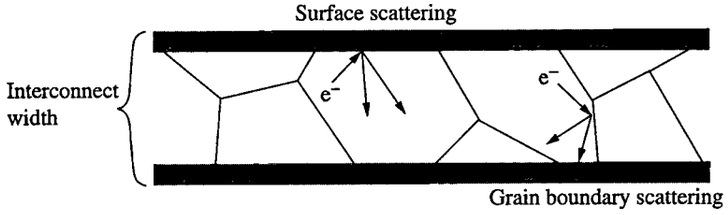

FIGURE 3.16 Surface and grain boundary scattering of electrons within a metal interconnect.

effective resistivity ρ_s of a thin metal film due to surface scattering is [145]

$$\rho_s = \frac{\rho_0}{1 - \frac{3(1-p)}{8k}} \quad (k >> 1) \tag{3.13}$$

$$\rho_s = \frac{\rho_0}{\frac{3k}{4}(1+2p)(\ln\frac{1}{k} + 0.423)} \quad (k << 1) \tag{3.14}$$

where $k = d/\lambda$ is the ratio of the metal film thickness to the electron mean free path, p is the fraction of the electrons that is elastically scattered at the surface and varies between 0 and 1, and ρ_0 is the bulk resistivity at a specific temperature. Note that elastically scattered electrons do not increase the resistivity since these electrons conserve the direction of the current flow [144]. ρ_s is therefore equal to ρ_0 if p is 1. Also note that (3.13) and (3.14) are valid for a thin metal film, which implicitly assumes that the electric field along the metal is uniform, neglecting the skin effect. The consequences of skin effect are described in the following section.

The effective resistivity of copper has been shown to be more sensitive to the line dimensions as compared to aluminum if the p value of copper is larger than the p value of aluminum [145]. Also note that (3.13) and (3.14) exhibit an implicit temperature dependence since the electron mean free path decreases with increasing temperature [144]. The effect of surface scattering is therefore lower at higher temperatures.

Grain boundaries in a polycrystalline interconnect (such as aluminum and copper) are the interface between grains (also referred to as crystallites) within the structure [147]. These boundaries are considered defects in a crystal structure, decreasing the electrical and thermal conductivity of the material by acting similar to "partially reflecting planes located perpendicular to the electric field" [145]. This behavior, referred to as grain boundary scattering, is illustrated in Fig. 3.16.

The grain size of an interconnect typically scales linearly with the wire dimensions [148]. If the grain size becomes comparable to the mean free electron path, the effect of grain boundary scattering on

the effective resistivity increases. This effect is characterized as

$$\rho_g = \frac{\rho_0}{3}\left[\frac{1}{3} - \frac{1}{2}\alpha + \alpha^2 - \alpha^3 \ln(1 + 1/\alpha)\right]^{-1} \tag{3.15}$$

$$\alpha = \frac{\lambda}{d}\frac{k}{1-k}, \tag{3.16}$$

where λ is the electron mean free path, d is the grain parameter, k is the grain boundary reflection coefficient ranging between 0 and 1, and ρ_0 is the bulk resistivity at a specific temperature.

Note that an alternative interconnect material, *carbon nanotubes* [149], offers attractive surface and grain boundary scattering characteristics as compared to copper. The use of graphene-based carbon nanotubes for on-chip interconnects, and the advantages and limitations of these advanced interconnect structures are summarized in Chapter 2.

Skin Effect

The study of surface and grain boundary scattering has typically focused on the resistivity of *thin films* [144], [145]. A thin film implicitly assumes that the electric field along the film is uniform. This assumption, however, is not true at sufficiently high frequencies due to the skin effect. At these high frequencies, electrons tend to flow near the interconnect surface and the current density exponentially decreases with the depth of the interconnect [150], as illustrated in Fig. 3.17.

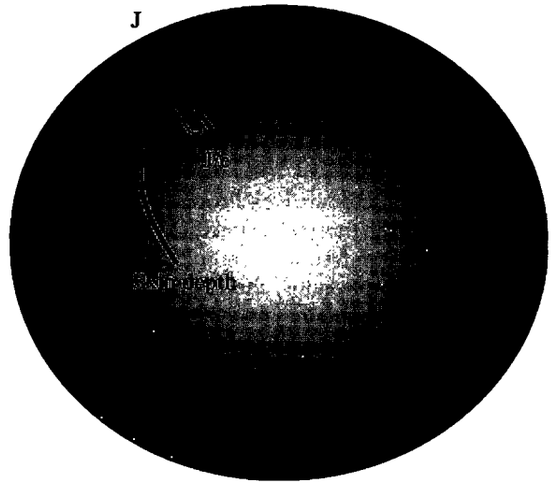

FIGURE 3.17 Cross section of a cylindrical conductor illustrating current distribution at high frequencies. The darker color indicates a higher current density.

Consequently, the effective cross sectional area of the conductor is reduced, thereby increasing the effective resistance of the interconnect. This phenomenon is called the skin effect. Note that the resistance modeled by (3.11) does not consider the skin effect and is called the DC resistance. The AC resistance of a conductor is therefore greater than the DC resistance due to this skin effect.

A parameter called the skin depth describes the intensity of the skin effect. Specifically, skin depth represents the distance below the conductor surface where the current density drops to $1/e$ of that at the surface, as depicted in Fig. 3.17 for a cylindrical conductor. A higher skin depth therefore represents a more uniform current flow within the conductor. Alternatively, a lower skin depth indicates a more significant skin effect. The skin depth δ of a *good conductor*, i.e., $1/\rho >> \omega\epsilon$ where ρ is the resistivity, ϵ is the dielectric constant, and ω is the angular frequency, is [150]

$$\delta(f) = \sqrt{\frac{\rho}{\pi \mu f}} \tag{3.17}$$

where μ represents the permeability of the conductor. At a certain frequency, a comparison of the skin depth to the wire dimensions is an important criterion to determine whether the skin effect should be considered. If the skin depth is sufficiently small and comparable to the wire dimensions, the skin effect should be considered when determining the resistance of the interconnect. This decision also depends upon the accuracy required from the models. The skin depth of copper, gold, and aluminum is shown in Fig. 3.18 as a function of frequency. Note that the skin depth is comparable to the wire dimensions in deep submicrometer technologies at frequencies above several tens of GHz.

Temperature

The resistivity of a conductor such as a copper interconnect is proportional to the electron mean free path. Specifically, a higher temperature reduces the electron mean free path by increasing the probability of electrons colliding with phonons, thereby increasing the resistivity. This relationship is approximately linear for copper at relatively high temperatures and can be characterized as [151], [79]

$$\rho_t = \rho_0(1 + \beta \Delta T) \tag{3.18}$$

where β is the temperature coefficient of resistivity (TCR) and ΔT is the difference in temperature. The typical TCR of a thin film copper wire is approximately 0.36%/°C [79]. A 10°C to 20°C increase in temperature therefore increases the resistivity by approximately 5%.

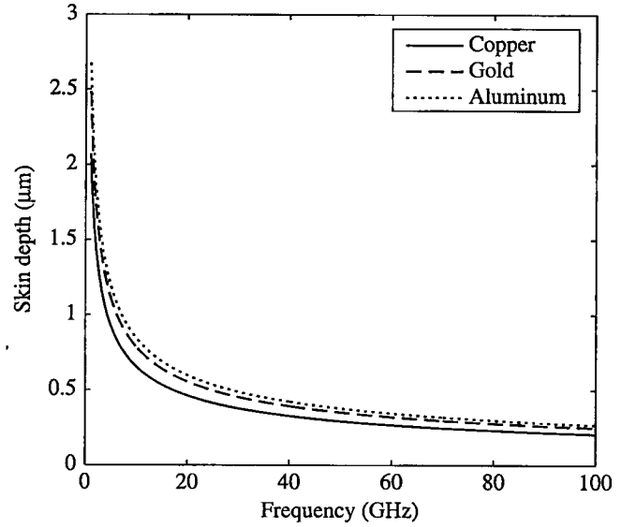

FIGURE 3.18 Variation of skin depth as a function of frequency for copper, gold, and aluminum.

3.3.2 Interconnect Resistance Extraction

Similar to the parasitic capacitance extraction process described in Section 3.2.2, the extraction of the resistance of an arbitrary shaped conductor requires solving the Poisson's equation described by (3.6) to determine the scalar electric potential ϕ. Since the closed-form solution of (3.6) is highly complex for even simple geometries, numerical solution techniques such as finite element [152], [153], finite difference [154], and boundary element method [155], [156] are often utilized. Once the electric potential is determined using any of these numerical techniques, the electric field vector E is obtained as the gradient of the electric potential,

$$E = -\Delta\phi \tag{3.19}$$

The resistance between the two ports is extracted from Ohm's law,

$$R = \frac{V}{I} = \frac{-\int_l E\,\partial l}{\oint_S \sigma E\,\partial s} \tag{3.20}$$

where the voltage across the resistor is equal to the line integral of the electric field between the ports, and the current is equal to the surface integral of the current density through the surface of the conductor.

In a typical modern IC where the interconnects occupy a significant portion of the overall circuit, solving Poisson's equation for

each interconnect is computationally prohibitive. The geometric constraints of the interconnect structures can be exploited to significantly reduce the computational complexity [157]. Specifically, interconnects are composed of a finite number of straight perpendicular lines, i.e., polygons, including orthogonal and some non-90° diagonal lines. More complicated shapes such as arcs are typically not used. These polygons can be divided into simpler regions to determine the overall resistance of the interconnect [157].

For a rectangular interconnect, the resistance is extracted from the process dependent sheet resistance of the material and the number of squares, i.e., the length divided by the width of the rectangle [158]. The length and width of the line are design related parameters characterizing an interconnect line. The sheet resistance $R_\square$ is described with the remaining parameters, i.e., resistivity and thickness,

$$R_\square = \frac{\rho}{H} \tag{3.21}$$

permitting the interconnect resistance to be described in terms of the sheet resistance per square times the effective number of squares,

$$R_{tot} = R_\square \frac{l}{W} = R_\square N \tag{3.22}$$

where $N = l/W$ is the number of squares along the rectangle. For example, for the interconnect shown in Fig. 3.19, the total parasitic resistance R_{tot} is

$$R_{tot} = R_\square \left(\frac{L_1}{W_1} + \frac{L_2}{W_2} \right) \tag{3.23}$$

Once a polygon is divided into simpler regions, the overall resistance is determined by adding the resistance of the individual segments. An important consideration when decomposing a polygon is to evaluate the flow of current [157]. Horowitz and Dutton describe heuristics for determining the resistance of a polygon that considers *bends* along an interconnect [157]. Typically, a bend is equivalent to 0.4 to 0.5 of a square.

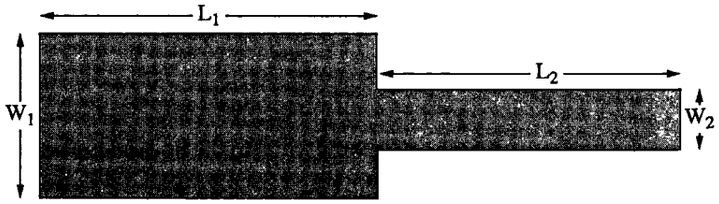

FIGURE 3.19 Interconnect to illustrate the calculation of resistance through the sheet resistance and number of squares.

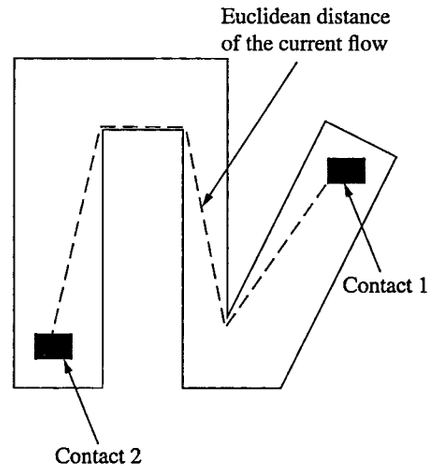

FIGURE 3.20 Approximation of the polygon length through the Euclidean distance assuming that most of the charge carriers flow along the shortest path.

An alternative resistance extraction methodology is described by Ladage and Leupers [159]. Rather than decomposing a polygon into multiple rectangles, a routing algorithm is proposed to estimate an approximate number of squares. The length is approximated by assuming that most of the charge carriers flow along the shortest path, i.e., the Euclidean distance between two contacts, as illustrated in Fig. 3.20 [159]. The width is approximated by several heuristics, where the total resistance is the sheet resistance times the number of squares [159].

A heterogeneous extraction methodology is also proposed where a polygon is decomposed into two types of segments: *regular* (or linear) and *irregular* (or nonlinear) [160], [161]. The irregular segments are composed of turns, contacts, and abrupt changes in the width of the line. The remaining segments are regular segments composed of a simple rectangle. These segments are illustrated in Fig. 3.21. The resistance of the regular segments is determined by the sheet resistance to reduce the computational complexity. Alternatively, the resistance of the irregular sections is determined by the boundary element method [160], analytic equations, look-up tables, or the finite element method to achieve reasonable accuracy [161].

Note that the individual contacts along the interconnect should be carefully examined when estimating the resistance due to two reasons: (1) the resistance of a contact can be significant, and (2) the current density through a contact is typically not homogeneous, making the extraction of the contact resistance difficult. This nonhomogeneous distribution of current within a contact is called current crowding [158].

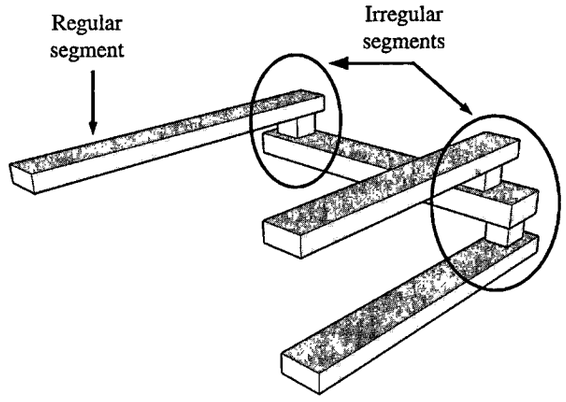

FIGURE 3.21 Regular (or linear) and irregular (or nonlinear) segments of an interconnect. The resistance of regular segments is determined by the sheet resistance and the resistance of the irregular segments is typically determined by the finite element, finite difference, or boundary element method.

3.4 Interconnect Inductance

Modeling an interconnect with an *RC* impedance does not consider the effects of on-chip inductance. Over the past decade, on-chip inductance has emerged as an important design parameter due to several factors [162]. These factors include longer and wider global interconnects; the use of copper, which exhibits a relatively low resistivity as compared to aluminum; and higher clock frequencies, i.e., faster signal transition times [163]. Consequently, on-chip inductance should be considered within interconnect impedance models under certain conditions to accurately determine the signal delay, evaluate signal integrity (inductive crosstalk), and satisfy noise constraints due to $L\,di/dt$ voltage drops.

On-chip parasitic inductance is reviewed in this section. Several definitions of inductance and the applicability of these definitions to the circuit analysis process are described in Section 3.4.1. The variation of inductance with frequency is described in Section 3.4.2. Figures of merit to determine those conditions where on-chip inductance should be considered during the circuit analysis process are explained in Section 3.4.3. Finally, the on-chip inductance extraction process is discussed in Section 3.4.4.

3.4.1 Definitions of Inductance

Inductance represents the ability of a circuit to store magnetic energy. The inductive properties of a circuit can be analyzed in several ways, as described in the following subsections.

Field Energy Formulation

The most general characterization of inductance is based on the *field energy formulation*. According to this formulation, inductance relates the electrical current to the magnetic field energy and is determined by current distribution functions within the media [164]. While the field energy formulation is sufficiently general, this approach requires numerical field analysis to determine the inductance. Numerical field analysis is computationally prohibitive for large scale circuits.

Loop Flux Definition

An alternative formulation of inductance is the *loop flux definition*. According to this definition, inductance is defined as a constant relating the magnetic flux induced in a loop to a current in another loop [164], [165], as illustrated in Fig. 3.22. The mutual loop inductance L_{ij} is

$$L_{ij} = \frac{\phi_{ij}}{I_j} \tag{3.24}$$

where ϕ_{ij} is the magnetic flux in loop i induced by the current I_j flowing through loop j. Assuming a uniform current distribution within a conductor, the mutual loop inductance L_{ij} can be expressed as

$$L_{ij} = \frac{\mu}{4\pi}\frac{1}{a_i a_j}\oint_{l_i}\oint_{l_j}\int_{a_i}\int_{a_j}\frac{\partial a_i \partial a_j\,\partial l_i \partial l_j}{|d|} \tag{3.25}$$

where a_i and a_j represent, respectively, the cross sectional area of the conductors i, j, l_i and l_j represent, respectively, the circumference of the loops i and j, and $|d|$ is the magnitude of the vector distance

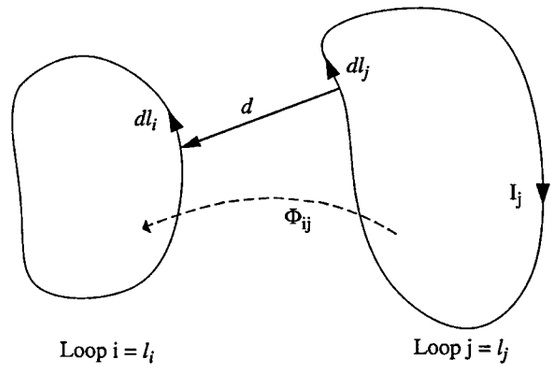

FIGURE 3.22 According to the loop flux formulation, the inductance is a constant that relates the magnetic flux in loop i induced by a current in loop j.

between the differential elements of the cross sections ∂a_i and ∂a_j. If the loops are the same ($i = j$), L_{ii} is the loop self-inductance. Note that the differential elements ∂l_i and ∂l_j are vectors whose signs are determined according to the direction of the current flow, respectively, in loops i and j. The mutual loop inductance can therefore be either positive or negative depending upon the relative direction of the current flow within the two loops. Also note that the mutual loop inductance decreases as the distance between the loops is increased. Alternatively, both the mutual loop and loop self-inductance increase as the circumference of the loop(s) is increased.

The loop flux definition of inductance is a special case of the field energy formulation where the current is assumed to flow in well formed loops. While the loop flux definition is more efficient since numerical field analysis is not required, a significant disadvantage remains: a closed loop for the current needs to be identified to determine the loop inductance. Traditional circuit analysis tools, however, are based on two terminal circuit elements rather than a loop. The concept of partial inductance has therefore been introduced to overcome this difficulty [166], [167].

Partial Inductance

According to the definition of *partial inductance*, a loop can be divided into multiple segments where the loop inductance is the sum of each self- and mutual inductance among the segments, i.e., the partial self- and partial mutual inductances [167]. The partial inductance formulation is utilized to describe the loop self-inductance in terms of the partial inductances. Assuming a current loop i consisting of N segments, the loop self-inductance is

$$L_{loop} = L_{ii} = \sum_{k=1}^{N} \sum_{l=1}^{N} L_{kl} \tag{3.26}$$

where L_{kl} is the partial inductance. Specifically, L_{kl} represents the partial self-inductance if $k = l$ and the partial mutual inductance if $k \neq l$. Consider, for example, the loop illustrated in Fig. 3.23 consisting of four discrete segments. The loop self-inductance in terms of the partial inductances is

$$\begin{aligned} L_{loop} = L_{ii} = {} & [L_1 + L_{12} - L_{13} - L_{14}] + [L_2 + L_{21} - L_{23} - L_{24}] \\ & + [L_3 + L_{34} - L_{32} - L_{31}] + [L_4 + L_{43} - L_{41} - L_{42}] \end{aligned} \tag{3.27}$$

Note that the current flows in the opposite direction for segments 1 and 3, 1 and 4, 2 and 3, and 2 and 4. The partial mutual inductance among these segments is therefore negative, as indicated by (3.27).

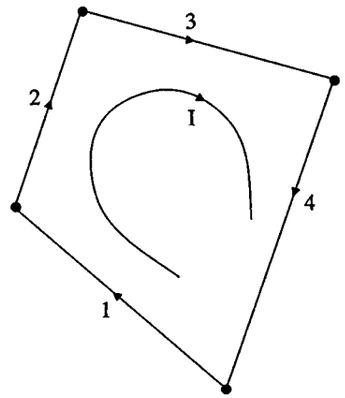

FIGURE 3.23 Loop consisting of four piecewise linear segments to illustrate partial inductances.

Net Inductance

The concept of *net inductance* is used to determine the contribution of a segment current to the overall magnetic flux [164]. Specifically, the net inductance of a segment is the summation of the partial self-inductance of the segment and the partial mutual inductances of the segment with all of the remaining segments. Equation (3.26) can be rearranged as

$$L_{loop} = L_{ii} = \sum_{k=1}^{N} L_k^{net} \tag{3.28}$$

where the net inductance of segment k, L_k^{net}, is

$$L_k^{net} = \sum_{k,l=1}^{N} L_{kl} \tag{3.29}$$

For example, referring to Fig. 3.23, the net inductance L_{net}^1 of segment 1 is

$$L_{net}^1 = L_1 + L_{12} - L_{13} - L_{14} \tag{3.30}$$

The self-inductance of loop i is therefore the sum of the net inductance of each segment,

$$L_{loop} = L_{ii} = L_1^{net} + L_2^{net} + L_3^{net} + L_4^{net} \tag{3.31}$$

Application of Inductance Formulations to the Circuit Analysis Process

An example is provided in this subsection to demonstrate how the concepts of loop, partial, and net inductance can be used to analyze

inductive effects within a circuit. Two interconnects are depicted in Fig. 3.24(*a*), where the first interconnect represents a current path of a signal and the second interconnect represents the corresponding current return path. This return path can be an adjacent signal line, power/ground line, or the substrate. Assuming the amount of current flowing through the signal path and the current return path is the same, this structure can be treated as a complete current loop.

The inductance of the signal and return paths can be separately determined by using partial inductances. Furthermore, the voltage drop along these paths caused by the inductance can be described by the net inductance. As shown in Fig. 3.24(*b*), the signal path has a partial self-inductance L^p_{sig}, a current return path has a partial self-inductance L^p_{ret}, and a partial mutual inductance M^p exists between the two paths. The voltage drop V_{sig} along the signal path is

$$V_{sig} = L^{net}_{sig}\frac{\partial i}{\partial t} \tag{3.32}$$

where $L^{net}_{sig} = L^p_{sig} - M^p$ is the net inductance of the signal path. Similarly, the voltage drop V_{ret} along the return path is

$$V_{ret} = L^{net}_{ret}\frac{\partial i}{\partial t} \tag{3.33}$$

where $L^{net}_{ret} = L^p_{ret} - M^p$ is the net inductance of the return path. The self-inductance of the loop created by these signal and return paths is expressed in terms of the partial self- and partial mutual inductances as

$$L_{Loop} = L^{net}_{sig} + L^{net}_{ret} = L^p_{sig} + L^p_{ret} - 2M^p \tag{3.34}$$

As illustrated in this example, the current return path of a signal should be known before the concept of a loop inductance can be applied. A significant problem in analyzing the effect of on-chip inductance is to accurately identify the current return paths. Accurate and efficient identification of the current return paths, however, is quite difficult due to the complex and high density interconnects in modern ICs.

3.4.2 Frequency Dependence of Inductance

Similar to resistance, inductance is also a function of frequency due to the variation of the current distribution within a conductor, as described in Section 3.3.1. In (3.25), current distribution within a conductor is assumed to be uniform. This assumption is only valid if the

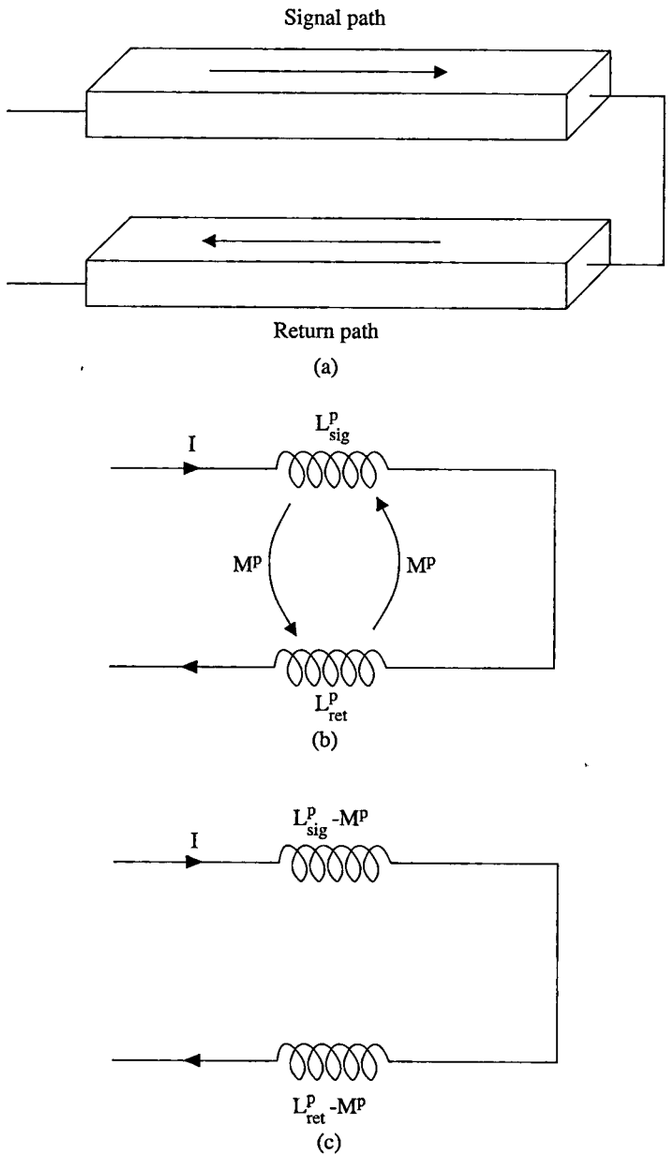

FIGURE 3.24 Current loop with a signal current path and the corresponding current return path: (a) Physical representation of the loop, (b) Equivalent partial inductance model of the loop, and (3) equivalent net inductance model.

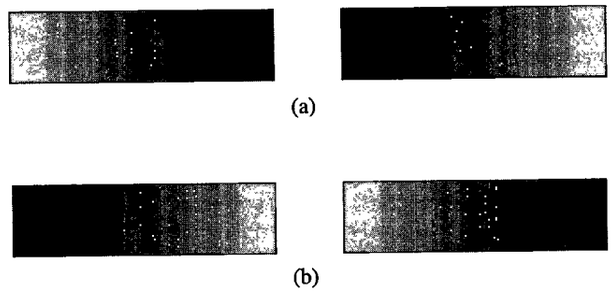

FIGURE 3.25 Cross section of two parallel interconnects illustrating the proximity effect. The darker color indicates a higher current density: (a) Current flows in opposite directions, and (b) current flows in the same direction.

magnetic field does not significantly affect the current flow within the conductor. There are two conditions that satisfy this assumption [164]:

- The resistive impedance R is much greater than the magnetic impedance $j\omega L$
- The separation between the conductors is much greater than the cross sectional dimensions

Note that either of these conditions is sufficient to assume the current distribution is uniform within a conductor.

The first mechanism that affects the current distribution is the skin effect, as described in Section 3.3.1 (see Fig. 3.17). At sufficiently high frequencies, the current tends to flow near the surface, decreasing the magnetic field within the conductor, reducing the inductance. In fact, the circuit inductance consists of two types of inductance [150]:

- The *external inductance* due to the magnetic fields outside the conductor (this component of the inductance has been described in the previous section)
- The *internal inductance* due to the magnetic fields within the conductor

The internal inductance is reduced due to the skin effect while the external inductance is unaffected. The individual contribution of the internal and external inductance to the overall inductance is therefore important to quantify the significance of the skin effect.

The second mechanism that affects the distribution of the current is the proximity effect. The proximity effect in two parallel interconnects is illustrated in Fig. 3.25. If the current in these two wires flows in opposite directions, the currents concentrate toward each other, as shown in Fig. 3.25(*a*); if the currents flow in the same direction, the two currents shift away from each other, as depicted in Fig. 3.25(*b*).

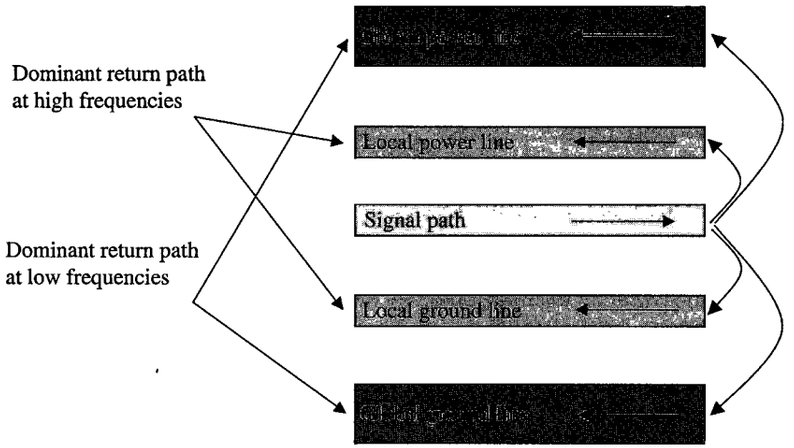

FIGURE 3.26 Return current distribution so as to minimize the total impedance at a specific frequency.

Both the skin effect and the proximity effect can be explained due to the same mechanism: at sufficiently high frequencies, the current attempts to minimize the inductive impedance by concentrating closer to the current return path. The inductance is decreased due to the reduced circumference of the current loop. This behavior of current can be generalized to better understand the concept of current redistribution as a function of frequency, as illustrated in Fig. 3.26 [164].

At low frequencies, the impedance is dominated by the interconnect resistance. In this case, the return current flows through the low resistive global power and ground lines. Alternatively, at high frequencies, the return current flows through the local power and ground lines since the inductive component dominates the overall impedance. In this case, the return current flows through nearby paths to minimize the inductance by reducing the circumference of the current loop. The current return path is distributed among various return paths to minimize the total impedance at a specific frequency.

3.4.3 When is On-Chip Inductance Important?

As described in Section 3.4.1, the current return paths should be identified to determine the inductance. Identification of the current return paths, however, is a difficult and computationally expensive process. It is therefore of practical importance to determine those conditions under which *inductive effects* are dominant within a circuit. Only if these conditions are satisfied, should the on-chip inductance be thoroughly analyzed.

Note that a high value of inductance does not necessarily result in inductive behavior. Whether a circuit exhibits inductive behavior

depends not only upon the absolute magnitude of the inductance, but also the magnitude of the resistance, capacitance, signal transition time, and length of the interconnect [164].

Two criteria are important to characterize the importance of on-chip inductance [168]:

- The ratio of the transition time of the input signal to the time of flight of the signal across the line
- The damping factor of a lumped RLC circuit

Specifically, a circuit exhibits inductive behavior if the transition time t_r of the input signal is smaller than the round trip time of flight T_o,

$$t_r < 2T_o = \frac{2l}{\upsilon} = 2l\sqrt{LC} \tag{3.35}$$

where l is the length of the interconnect, $\upsilon = 1/\sqrt{LC}$ is the velocity of the electromagnetic signal propagation along the line, and L and C are, respectively, the inductance and capacitance per unit length. This condition specifies that the circuit exhibits inductive behavior if the length of the interconnect is sufficiently longer than the shortest signal wavelength observed *within the spectral content of the signal*. Note that the ratio of the length of the interconnect to the signal wavelength l/λ is referred to as the *electrical size* [169], which is used to determine whether lumped or distributed elements are required to accurately characterize an interconnect. This behavior is further discussed in Chapter 4.

The second criterion to characterize the importance of on-chip inductance is the damping factor ζ of the interconnect. Specifically, if the damping factor is smaller than one, the circuit is underdamped and oscillations will occur, indicating an inductive response. The second condition is therefore

$$\zeta = \frac{Rl}{2}\sqrt{\frac{C}{L}} < 1 \tag{3.36}$$

where l is the length of the interconnect, and R, L, and C are, respectively, the resistance, inductance, and capacitance per unit length.

Rearranging (3.35) and (3.36) enables a *range of interconnect length* to be determined where a circuit exhibits inductive behavior [168],

$$\frac{t_r}{2\sqrt{LC}} < l < \frac{2}{R}\sqrt{\frac{L}{C}} \tag{3.37}$$

The lower bound of the interconnect length is determined by the rise time constraint since for sufficiently short lines, the transition time is greater than the time of flight across a line. The upper bound of

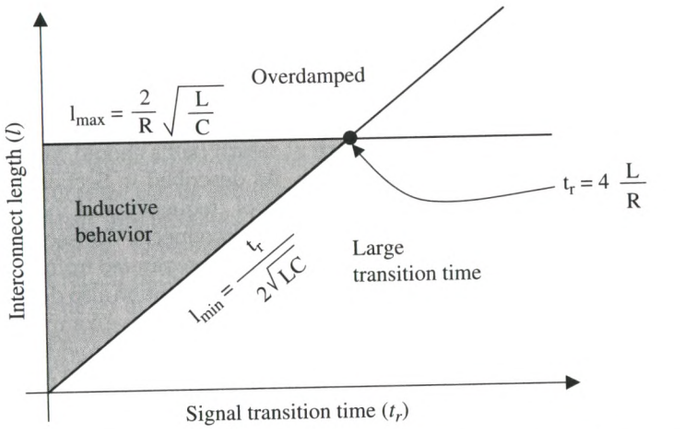

FIGURE 3.27 Range of length where an interconnect exhibits inductive behavior, as determined by (3.37) [168].

the interconnect length is determined by the damping factor constraint since for sufficiently long lines, the resistance of the interconnect dominates. A larger attenuation caused by the increased resistance makes the inductive effects negligible. The range of interconnect length where the circuit exhibits inductive behavior is depicted in Fig. 3.27. Note that the damping characteristics of a circuit are not dependent upon the rise time, and therefore a unique line length is determined beyond which the circuit is overdamped for any signal transition time, as illustrated by l_{max} in Fig. 3.27. Also note that l_{max} and l_{min} intersect at $t_r = 4L/R$. For transition times beyond $4L/R$, the circuit does not exhibit inductive behavior for any interconnect length.

3.4.4 Interconnect Inductance Extraction Process

As compared to resistance and capacitance, the on-chip interconnect inductance is significantly more difficult to extract due to several reasons:

- It is difficult to accurately determine the current return paths
- Inductance is a long range phenomenon, making the mutual inductance of nonadjacent interconnects nonnegligible
- Inductance changes as a function of frequency due to skin and proximity effects
- The density and complexity of on-chip interconnects, distribution of the return currents into multiple paths, and the long

range effect of inductance result in extremely large and dense inductance matrices, increasing the computational complexity of the simulation process

The inductance of a wire not forming a closed loop does not have a physical meaning [167]. The current return paths should therefore be identified to obtain a closed loop. As described in Section 3.4.2, these return paths change as a function of frequency and the resistive characteristics of the surrounding interconnects. Consequently, the inductance extraction process should not be separated from the resistance extraction process since the current distribution also depends upon the interconnect resistance. Furthermore, capacitive components such as coupling, ground, and decoupling capacitances provide alternative current loops, as depicted in Fig. 3.28. Assuming all of the current flows uniformly toward the end of the interconnect is therefore a pessimistic approach since the local capacitances form shorter return paths, as depicted by L_2 and L_3 in Fig. 3.28 [170].

Long range inductive effects is another difficulty for the inductance extraction process. Artificially restricting the inductance extraction process to nearby geometries not only introduces inaccuracies, but may also result in instability. The pattern matching method used for capacitance extraction, therefore, cannot be used for inductance extraction due to the complex geometries common in modern

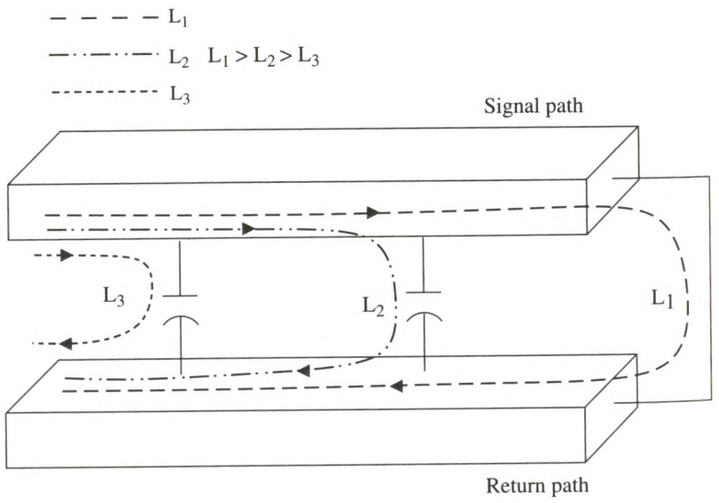

FIGURE 3.28 Alternative current return paths formed by the interconnect capacitances. Assuming all of the current flows uniformly toward the end of the interconnect results in a pessimistic estimate of the loop inductance.

ICs [143]. Several methods such as numerical techniques and closed-form solutions for inductance extraction are discussed in the following subsections.

Numerical Techniques

Assuming the current return paths forming a closed loop are known, the inductance can be extracted using Maxwell's equations based on field energy and loop flux formulations, as described in Section 3.4.1. This solution, however, is not feasible due to two reasons: (1) the current return paths are difficult to determine, and (2) discretization (or meshing) of the conductors is computationally prohibitive due to the structural complexity of the on-chip interconnects. This loop-based technique is effective for well designed interconnect structures, such as shielded buses and clock distribution networks, rather than a full circuit analysis composed of a variety of arbitrary and complex physical structures.

The concept of partial inductance, as described in Section 3.4.1, is one approach to avoid determining the current return paths. The current return paths are assumed to be at infinity or at a specific reference common to each signal path [30]. An application of partial inductance is the *partial element equivalent circuit* (PEEC) model [171] where each conductor is divided or discretized into multiple *filaments*. Note that these filaments are essential to accurately capture skin and proximity effects [172]. Based on the PEEC model, the loop inductance is uniquely determined as the sum of the partial self-inductance of each segment and the partial mutual inductances between any segments. Also note that when applying the concept of partial inductance in circuit models, such as a PEEC model, all of the wires that form the current loops should be included to obtain accurate results. The primary issue of the PEEC model is the extremely large and dense inductance matrices, increasing computational complexity. Various techniques have been proposed to sparsify the inductance matrices [163] such as the shell technique [173], the halo technique [174], and the K matrix technique [175].

FastHenry [170] is a commonly used numerical tool for determining the loop inductance of simple interconnect structures. Several other loop based inductance extraction techniques have also been developed by estimating the distribution of the return current [176–179]

Another solution for the current return path is to shield the critical signal lines, such as a clock signal [163]. Based on this technique, the critical interconnect is designed to make the inductance less difficult to extract. Specifically, shield lines are placed on both sides of the critical signal line, as depicted in Fig. 3.29, which constrains the current return path to these shield lines. Hence, the loop inductance is not only minimized, but also accurately determined.

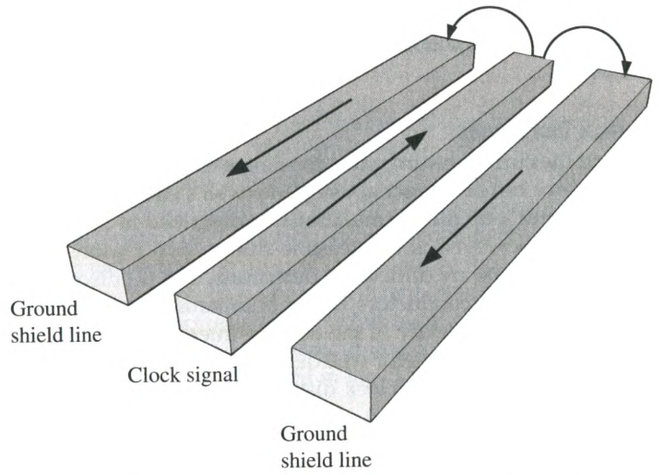

FIGURE 3.29 Shield lines placed on both sides of a clock signal to force the return current to flow through these shield lines.

Closed-Form Solutions

Closed-form solutions, also referred to as rule-based methods [176], offer significant improvement in computational complexity over numerical techniques. Furthermore, closed-form solutions provide intuition during the design process. These closed-form solutions, however, can only be obtained for simple structures that are a small subset of the structures required in a full circuit inductance analysis.

For example, the partial self-inductance L_{part} of a rectangular line, as shown in Fig. 3.30(*a*), is approximated as [180]

$$L^{p}_{self} = 0.2l\left(\ln\frac{2l}{H+W} + \frac{1}{2} - \ln\gamma\right) \quad \mu\text{H} \tag{3.38}$$

where l represents the length of the interconnect in meters. H and W represent, respectively, the thickness and width of the interconnect.

The $\ln\gamma$ term is a function of the ratio of the thickness to the width H/W, and is typically small as compared to the other terms. Note that (3.38) is valid only for a conductor with a length that is much greater than the summation of the thickness and width ($l >> H + W$). Also note that the partial self-inductance is a superlinear function of line length. The partial mutual inductance L^{p}_{mutual} between two parallel straight lines, as illustrated in Fig. 3.30(*b*), is approximated as [180]

$$L^{p}_{mutual} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right) \quad \mu\text{H} \tag{3.39}$$

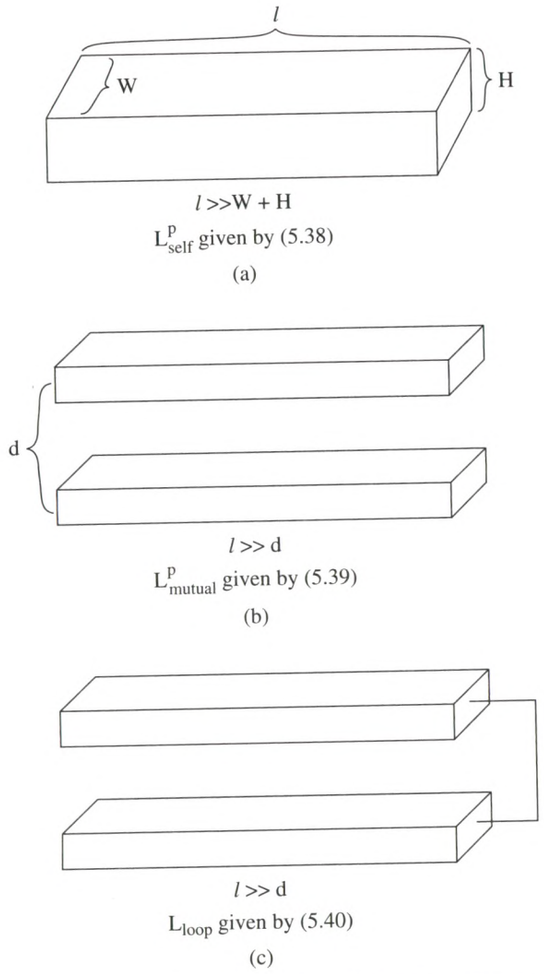

FIGURE 3.30 Closed-form solutions to determine the inductance of simple geometric structures: (a) Rectangular line for the partial self-inductance, (b) two parallel rectangular lines for the partial mutual inductance, and (c) current loop formed by two rectangular lines for the loop inductance.

where d is the pitch, i.e., the distance between the line centers. Note that (3.39) is valid if the length of the parallel lines is equal and much greater than the pitch, $l >> d$. The loop inductance L_{loop} of a complete current loop formed by two rectangular lines, as depicted in

Fig. 3.30(*c*), can be approximated as [180]

$$L_{loop} = 0.4l\left(\ln\frac{d}{H+W} + \frac{3}{2} - \ln\gamma + \ln k\right) \ \mu\text{H} \qquad (3.40)$$

Similar to (3.38) and (3.39), (3.40) is also accurate for long lines where $l >> d$. Also note that the loop inductance is a linear function of line length.

As illustrated in Fig. 3.30, these closed-form expressions are only effective for simple geometries with specific constraints on the length and pitch. These expressions can be useful for certain interconnects that satisfy these constraints, such as global power and clock distribution networks that consist of highly regular and long interconnects.

Also note that the sensitivity of the signal waveform characteristics such as the propagation delay and rise time to an inaccuracy in the inductance value is low [181]. This relatively weak dependence of delay on inductance can be exploited to improve the computational efficiency of the inductance extraction process. For example, a specific amount of error in the inductance can be tolerated since the computational efficiency is greatly improved while the effect of this error on the signal waveform characteristics is less significant [181].

3.5 Chapter Summary

The physical and electrical characteristics of on-chip interconnects are presented in this chapter, as summarized below:

- Global interconnects play an important role in affecting the overall performance characteristics of an IC such as the speed, power, noise, and area.
- Multiple criteria such as latency, bandwidth, power dissipation, noise, and area should be simultaneously considered during the interconnect design process.
- The system performance and computational efficiency of a processor can be improved by reducing the latency of those global interconnects that provide communication among various blocks within a multicore architecture.
- Bandwidth refers to the maximum bit rate that can be reliably transmitted across the interconnect per second.
- The bandwidth of an interconnect is proportional to the reciprocal of the delay of the interconnect. This proportionality, however, is not necessarily linear.

- Interconnect noise in digital circuits has become a primary concern due to several reasons such as reduced noise margins, faster signal transitions, and the increasing effect of interconnects on system performance, e.g., signal integrity.
- Power dissipation and physical area are two additional important design criteria for on-chip interconnects.
- Accurately modeling these interconnects is crucial to predict and satisfy target performance parameters.
- The electrical model of an on-chip interconnect has evolved from a short circuit to a simple capacitive model, and later to a capacitive and resistive model, and finally to a capacitive, resistive, and inductive model.
- The parasitic capacitance of an interconnect should be considered if the line capacitance is comparable or greater than the overall gate oxide capacitance of the driven gate(s).
- Three components of the interconnect capacitance are the parallel plate capacitance, fringe capacitance, and lateral (or coupling) capacitance.
- In deeply scaled technologies, the contribution of the fringe and lateral capacitance has significantly increased due to the greater aspect ratio of the interconnects.
- A conventional parallel plate capacitance model significantly underestimates the overall capacitance in deeply scaled technologies.
- Two primary approaches exist to extract the parasitic capacitance: (1) numerically solving Poisson's equation using finite difference, finite element, or boundary element methods, and (2) developing simplified empirical or analytic models.
- The primary disadvantage of solving Poisson's equation is high computational complexity, making the capacitance extraction process of modern on-chip interconnects increasingly infeasible.
- Empirical models provide fast estimation of capacitance with moderate accuracy.
- The range of physical interconnect parameters such as width, thickness, and length for which these empirical models are sufficiently accurate is an important consideration.
- Pattern matching is a technique used in modern capacitance extraction processes where the geometric parameters of the circuit are matched to a previously extracted set of test structures to efficiently obtain an estimate of the capacitance.

- The parasitic resistance of an interconnect should be considered if the line resistance is comparable or greater than the channel on-resistance of the driving transistor.
- The channel on-resistance of a transistor remains approximately constant with technology scaling while the parasitic resistance of the global interconnects increases with scaling.
- The parasitic resistance has deleterious effects on the signal characteristics such as increased delay, signal degradation, clock jitter, and interconnect noise.
- Copper has replaced aluminum in modern ICs due to lower resistivity, ability to carry higher current densities, and reduced contact resistance.
- The effective resistivity of copper deviates from the pure resistivity of copper due to operating temperature, signal frequency, and certain necessary processing steps.
- A barrier layer (built on the sides of the interconnect to prevent copper from diffusing into the surrounding dielectric) increases the resistivity of copper since the effective cross sectional area of the interconnect is reduced.
- Surface scattering is a mechanism that increases copper resistivity since the electrons experience more collisions and reflections at the surface due to scaled dimensions.
- Grain boundary scattering is another mechanism that increases the resistivity of copper due to the partially reflecting behavior of the crystallites within the interconnect.
- The resistance of an interconnect increases at high frequencies due to the skin effect where the electrons tend to flow near the surface, reducing the effective cross sectional area.
- At a certain frequency, if the skin depth is comparable to the wire dimensions, the skin effect should be considered when estimating the resistance.
- A higher temperature increases the resistivity of an interconnect by increasing the electron collision probability with phonons.
- Accurately extracting the resistance of an arbitrarily shaped conductor requires the solution of Poisson's equation, which is computationally prohibitive in most practical circuits.
- Typical shapes of on-chip interconnects (polygons where orthogonal and diagonal lines are allowed) can be exploited to simplify the extraction process.
- A polygon can be decomposed into simpler regions, permitting the overall resistance to be determined through the sheet resistance.

- Those regions composed of turns, contacts, and abrupt changes in the line width should be carefully examined since the current density through these regions is typically non-homogeneous.
- The on-chip inductance has become an important design parameter due to several reasons such as longer and wider global interconnects, the use of lower resistivity copper, and faster signal transition times.
- The inductive properties of a circuit can be analyzed in several ways: (1) field energy formulation, (2) loop flux definition, and (3) partial inductance.
- The field energy formulation is sufficiently general, but requires numerical field analysis, which is computationally prohibitive in large scale circuits.
- Although the loop flux definition is more convenient since numerical field analysis is not required, identification of a closed current loop is difficult due to the geometric complexity of on-chip interconnects.
- The concept of a partial inductance can be used to determine the loop inductance by dividing the loop into multiple segments, where the return path for each segment is assumed to be at infinity or at a specific reference common to each signal path.
- The net inductance of a segment is the summation of the partial self-inductance of the segment and the partial mutual inductances of the segment with all of the remaining segments.
- The inductive voltage drop along a segment is the net inductance of the segment times the rate of change in the current.
- Current distribution within a conductor is not uniform at sufficiently high frequencies due to skin and proximity effects.
- The inductance decreases at sufficiently high frequencies due to the reduced circumference of the current loop since the current concentrates closer to the current return path.
- A current return path is frequency dependent since the current flow will minimize the overall impedance along a path.
- A high value of inductance does not necessarily result in inductive behavior within a circuit.
- A circuit exhibits inductive behavior in two situations: (1) the transition time of the input signal is smaller than the round trip time of flight, or (2) the damping factor of the interconnect is smaller than one, i.e., the inductive time constant should be greater than half the resistive time constant.

- The on-chip inductance extraction process is challenging for several reasons: (1) inductance is a long range phenomenon, (2) the current return paths are difficult to identify, and (3) inductance is frequency dependent.
- On-chip inductance can be extracted using Maxwell's equations if the current return paths are known.
- Shielding is a design technique used to constrain the return current to flow through a nearby shield line, thereby minimizing the inductance and enhancing the estimation accuracy.
- The partial element equivalent circuit model is a numerical technique that applies the concept of a partial inductance to extract the on-chip parasitic inductance.
- The inductance of simple geometries can be determined from closed-form solutions or rule-based methods that are computationally efficient, but are relatively inaccurate when evaluating the inductance of complicated geometries and structures.
- The signal delay characteristics are a weak function of the magnitude of the inductance.

CHAPTER 4

Signal Propagation Analysis

The physical characteristics of the interconnect and the extraction of the parasitic impedances, i.e., capacitance, resistance, and inductance, have been described in the previous chapter. Models for characterizing *signal propagation* along an interconnect and techniques to evaluate the *delay of an interconnect* are discussed in this chapter.

As emphasized in Chapter 2, the global interconnects play a primary role in determining the critical performance characteristics of an integrated circuit (IC). One such characteristic is the speed of a circuit. The effects of the interconnects on the signal propagation characteristics should be considered to accurately estimate the delay [5,182,183]. To properly design complex circuits, accurate characterization and simulation of the interconnect behavior and signal transients are required. This high accuracy is necessary, particularly for analyzing performance critical blocks to accurately anticipate possible timing violations and hazards during switching activity. Furthermore, increasing clock frequencies require safety margins to be reduced. Note that safety margins ensure that the IC operates correctly under process and environmental variations that are difficult to predict. The safety margins can be reduced if more accurate interconnect delay information is available. Thus, the process of characterizing signal waveforms in on-chip interconnects is of primary importance.

Accurate and computationally efficient circuit models are required to analyze the delay of an interconnect carrying a data or clock signal. Achieving both accurate and computationally efficient delay analysis, however, is challenging. For example, utilizing a dynamic simulator such as Simulation Program with Integrated Circuit Emphasis (SPICE) to analyze the delay characteristics produces sufficiently accurate results. The computational complexity and memory required by SPICE, however, is prohibitive in analyzing large scale circuits.

Several approaches are described in this chapter that exhibit different tradeoffs between accuracy and computational complexity.