CHAPTER 2

Technology Scaling

Over the past 15 years, the high performance integrated circuit (IC) design process has undergone a major transition. With scaling of active device feature sizes into the nanoscale regime, the effects of interconnects on system delay, power consumption, noise, and reliability have become fundamentally important [28–31]. This transition from a logic centric IC design process into an interconnect driven design process, as shown in Fig. 2.1, has caused the development of new design methodologies applied at multiple abstraction levels [29].

A smaller feature size reduces the delay of the active devices; however, the effect on delay due to the passive interconnects has increased rapidly. The International Technology Roadmap for Semiconductors (ITRS) identifies interconnect delay as "the most critical phenomenon affecting high performance products" [32]. In a 1 μm CMOS technology, a typical transistor delay was in the range of 20 ps [32]. This delay is expected to decrease to approximately 1.45 ps for a 15 nm CMOS technology [33]. Alternatively, the minimum achievable interconnect delay remains effectively fixed at approximately 20 ps/mm [34]. Note that this estimate assumes an optimum width and number of repeaters that achieve the minimum possible delay. Thus, the overall speed of current ICs is most often limited by the long distance global interconnects. This characteristic was also described by ITRS in 2005 with a relatively famous graph, as shown in Fig. 2.2 [35]. The starting technology node for this figure is 0.25 μm, where the interconnect delay is approximately equal to the gate delay. According to this figure, the gate and local interconnect delay are both reduced with smaller feature sizes. The delay of the global interconnects, however, has substantially increased. Use of repeaters somewhat alleviates this issue, but cannot prevent the global interconnects from becoming the primary bottleneck for system performance. Furthermore, each additional buffer increases overall power consumption and physical area.

Global interconnects directly increase not only the delay, but also the power consumption. With shrinking feature size and larger chip die dimensions, the shear number of interconnects has increased

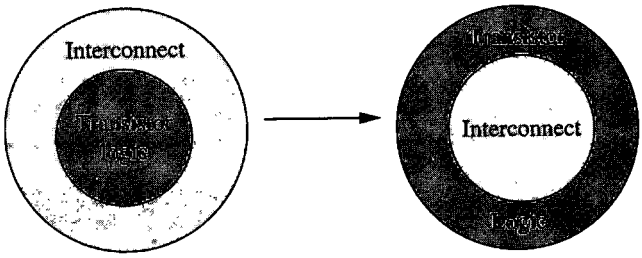


FIGURE 2.1 Transition from a logic-centric IC design process to an interconnect driven design process [29].

exponentially [28, 30, 36]. Interconnect capacitance often dominates the total gate load [37]; therefore, a large portion of the total transient power is dissipated by these on-chip capacitive lines. In DSM technologies, the coupling capacitance between wires is typically the dominant capacitive component in interconnects. The introduction of novel dielectric materials with smaller permittivity (low-K dielectric) does not sufficiently alleviate this issue due to the high-K materials required during other manufacturing processes [32]. The contribution of the interconnects to the overall power consumption is particularly significant in those long interconnects that distribute the clock signals, where as much as 40% to 50% of the total power of an IC can be dissipated [35], as further discussed in Chapter 15.

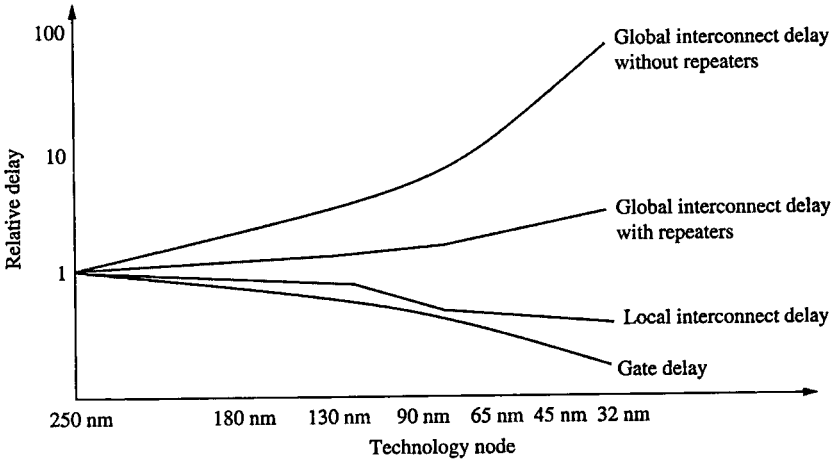


FIGURE 2.2 Dominance of interconnect delay over gate delay in deeply scaled CMOS technologies [35].

In addition to reducing system performance and increasing power consumption, higher interconnect coupling capacitances also degrade on-chip signal integrity, as described in Chapter 5. Thus, the circuit is not only more susceptible to logical failure, but also exhibits greater delay uncertainty, thereby causing synchronization faults in time critical data signals. Furthermore, on-chip interconnects dedicated to the power and ground lines suffer from power supply noise due to high parasitic impedances and fast signal transitions, as further described in Chapter 8. Variations in the power supply voltage affect circuit performance and degrade power integrity. Degradation in both signal and power integrity produces a less robust circuit with poor predictability characteristics.

Considering these primary design objectives (propagation delay, power consumption, and noise), interconnect has become a dominant issue in high performance ICs. The scaling characteristics of both the active devices and passive interconnects are reviewed in this chapter. Specifically, device scaling characteristics are described in Section 2.1. Small geometry effects due to device scaling are discussed in Section 2.2. Several techniques to alleviate small geometry effects are summarized in Section 2.3. Interconnect scaling characteristics are described in Section 2.4. The significance of the global interconnects is also demonstrated in this section by considering the scaling factor for each primary design objective. Several evolving techniques to reduce the effects of the interconnect parasitic impedances are described in Section 2.5. Finally, the chapter is summarized in Section 2.6.

2.1 Device Scaling

Continuous miniaturization has been the primary driving force in the dominance of CMOS technology within the electronics industry. The computational capability of the ICs has significantly increased, while the overall cost per IC has decreased. The motivation (from a device perspective) behind technology scaling, also referred to as Moore's law [24,25], is described by the following benefits:

- Higher packing density, and therefore, higher yield since the defect density is a function of physical area
- Higher circuit speed due to enhanced transistor performance
- Higher system speed due to smaller chip-to-chip delays
- Lower system power consumption due to smaller chip-to-chip capactive loads
- Enhanced system reliability due to fewer leads and off-chip connections
- Heterogeneous systems integration

Two technology scaling schemes (constant electric field and constant voltage) have been utilized throughout the history of IC scaling. Starting from the 2 μm technology node, a coarse chronology of these scaling schemes is illustrated in Fig. 2.3:

- 1: Between the 2 μm and 0.8 μm technology nodes: constant voltage scaling
- 2: Between the 0.8 μm and 130/90 nm technology nodes: constant electric field scaling
- 3: Below the 130/90 nm technology node: constant voltage scaling
- 4: Below the 45 nm technology node: structural enhancements

Between the 2 μm and 0.8 μm technology nodes, the power supply voltage had been maintained constant, primarily due to two reasons: (1) to avoid multiple power supply voltages since the interface and peripheral circuits typically require constant input/output voltages, and (2) to reduce the delay. Since the voltage is constant, the vertical and lateral electric fields increase as the technology is scaled. The increased electric fields cause hot carrier related issues [38]. The overall power consumption also increases. Due to these two detrimental effects, below the 0.8 μm technology node, a constant electric field scaling scheme was utilized where the power supply voltage and threshold voltage both scale with technology. Note that the threshold voltage should ideally be decreased by the same ratio as the power supply voltage to maintain performance. This trend continued until approximately the 130/90 nm technology node, where the power supply voltage has been scaled to approximately 1 volt.

Below the 130/90 nm technology node, however, further scaling of the power supply voltage is limited since the threshold voltage does

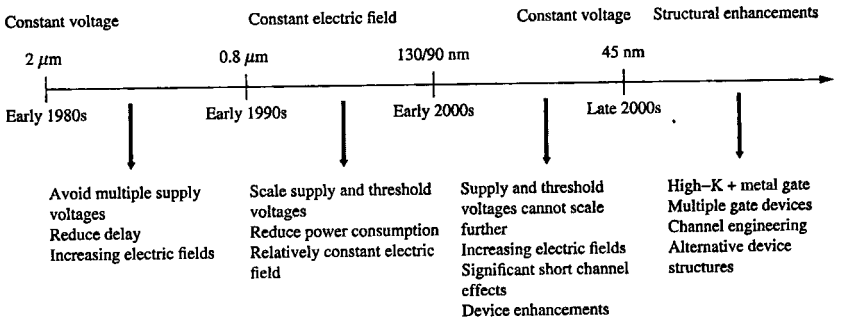


FIGURE 2.3 A coarse chronology of CMOS technology scaling schemes starting from the 2 μm node.

not scale well due to two primary reasons: (1) a significant increase in subthreshold leakage current, and (2) difficulty in controlling the manufacturing process steps that determine the threshold voltage during fabrication. Note that transitioning from a constant electric field to constant voltage scaling did not occur at a specific technology node. As the feature size was reduced, the aforementioned reasons limited the scaling of the threshold voltage, which, in turn, slowed scaling of the power supply voltage. The electric fields therefore began to increase before the 130/90 nm node. Ultimately, around the 130/90 nm technology node, the power supply reached approximately one volt. Any further scaling in the voltage has been minimal [39]. Thus, the electric fields started to rapidly increase below the 130/90 nm node, causing detrimental small geometry effects, as further described in Section 2.2.

An important detrimental effect of the high vertical electric field has been an increase in gate leakage current, significantly limiting technology scaling below the 65 nm node. To alleviate this issue, certain structural enhancements such as alternative gate insulator materials (other than silicon dioxide) with a higher dielectric permittivity (high-K) have been utilized, starting at the 45 nm technology node [40]. The use of high-K materials decreased the vertical electric fields since the current drive capability is increased by exploiting the higher permittivity of the insulator rather than reducing the insulator thickness. Note that both subthreshold and gate leakage mechanisms are reviewed in Chapter 11.

A brief overview of MOS device operation is provided in Section 2.1.1. The two scaling scenarios, constant electric field and constant voltage scaling, are described, respectively, in Sections 2.1.2 and 2.1.3. Finally, these two scenarios are compared in Section 2.1.4.

2.1.1 MOS Device Operation

A metal oxide semiconductor field effect transistor (MOSFET) consists of four terminals: gate, source, drain, and bulk. As depicted in Fig. 2.4, a thin dielectric layer underneath the gate electrode isolates the gate terminal from the source and drain terminals. The source and drain terminals are doped with opposite polarity to the substrate. For example, in an NMOS transistor with a P-type substrate, the source and drain are doped with boron impurities, whereas in a PMOS transistor with an N-type substrate, the source and drain are doped with phosphorus or arsenic.

If the voltage at the gate terminal is smaller than the threshold voltage of the device, the majority carriers within the channel are pushed deep into the substrate, producing a depletion region within the silicon surface (or transistor channel). Note that this depletion region extends to the source and drain terminals since each of the two terminals and the substrate produce PN diodes. A conducting layer

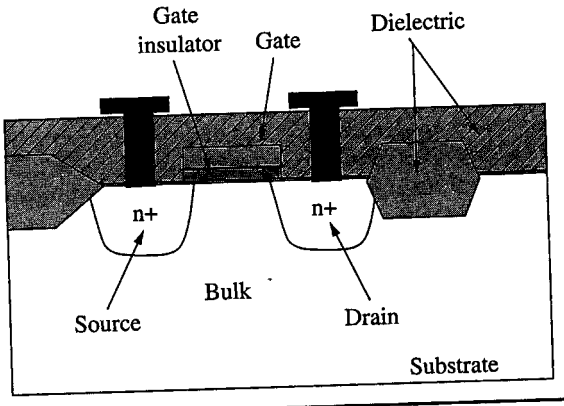


FIGURE 2.4 Cross section of an N-type metal oxide field effect transistor (MOSFET) that consists of four terminals: gate, source, drain, and bulk.

is formed between the source and drain terminals when a sufficiently large voltage (greater than the threshold voltage) is applied to the gate terminal. This conducting layer is formed since a large number of minority carriers are attracted to the silicon surface due to the vertical electric field. Once a conducting layer is formed, the transistor channel or the silicon surface is referred to as inverted since minority carriers within the substrate dominate the channel. After the channel is inverted, current flows between the drain and source terminals, assuming that there is a voltage difference (and therefore a lateral electric field) between these two terminals. Thus, an essential aspect of an MOS device is the ability to control the current flow by the vertical electric field generated by the gate electrode.

The quality of a MOSFET in a digital IC is determined by two factors [41,42]:

- The ratio of the on-current to off-current I_{on}/I_{off}, where I_{on} is the transistor current when $V_{GS} = V_{DS} = V_{DD}$ and I_{off} is the transistor current when $V_{GS} = 0$ and $V_{DS} = V_{DD}$. A higher I_{on}/I_{off} ratio is desirable to increase speed and lower standby power.
- The time required for a transistor to switch from the on-state to the off-state. This delay is determined by two primary factors: (1) the distance between the source and drain terminals, i.e., channel length, and (2) the velocity of the charge carriers that flow from the source to the drain. A smaller switching time is desirable to enhance circuit speed.

The two scaling schemes and the effect of these schemes on transistor performance are described in the following sections.

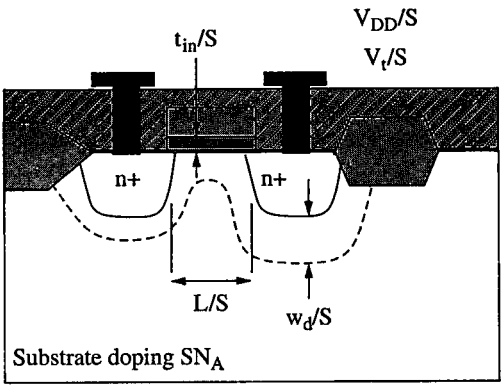


FIGURE 2.5 Constant electric field scaling framework developed by Dennard et al. in the early 1970s [43].

2.1.2 Constant Electric Field Scaling

In the early 1970s, Dennard et al. developed a theoretical framework for scaling MOS transistors [43]. These principles, also referred to as constant electric field scaling, have been largely followed by the semiconductor industry. Referring to Fig. 2.5, Dennard's framework relies on the following steps, which are based on simultaneously scaling three variables: (1) physical dimensions, (2) voltages, and (3) doping concentrations [43]:

- Increase the current drive capability by scaling the gate insulator thickness and channel length
- Control the high vertical electric field and power dissipation by scaling the power supply voltage
- Maintain the current drive capability by scaling the threshold voltage
- Decrease the depletion layer width by increasing the substrate doping concentration

Despite the various device level challenges that were faced during the scaling process, the trend has continued and the fundamental structure of the MOS device has remained basically the same. One of the undesirable effects of reducing the distance between the drain and source, the channel length, occurs when the depletion region surrounding the drain and source occupy a greater portion of the substrate under the gate electrode, thereby modifying the surface potential [43]. The capability of the gate electrode to control the channel is therefore degraded, particularly at high drain voltages. This short channel effect reduces the threshold voltage of the device, degrading the transistor quality. To prevent this issue, the vertical dimensions, such as the gate insulator thickness and junction depths, have been scaled along with the horizontal dimensions such as the

channel length and width. Note that the power supply voltage has also been reduced and the substrate doping concentration has been increased [43].

Assuming a unitless scaling factor S ($S > 1$) and constant electric field scaling, the transistor width W, channel length L, gate insulator thickness t_{in}, and power supply voltage V_{DD} each scale by $1/S$. Note that the electric field ideally remains the same. The substrate doping concentration N_a scales by S. Under these conditions, the scaling factor for multiple parameters is described in the following subsections.

Scaling Factor for Depletion Layer Width

The relationship between the original depletion layer width w_d and the scaled depletion layer width w'_d is [43]

$$w'_d = \sqrt{\frac{2\epsilon_{si}(\psi_b + V_{SB}/S)}{qSN_a}} \approx \frac{w_d}{S}, \tag{2.1}$$

where ϵ_{si}, ψ_b, N_a, and V_{SB} represent, respectively, the silicon permittivity, built-in junction potential, substrate doping concentration, and source-to-bulk voltage of an MOS transistor. Note that ψ_b is assumed to be approximately constant with scaling due to the weak logarithmic dependence of ψ_b on N_a [43]. Thus, the depletion layer width scales approximately by the same scaling factor as the channel length. Note that this reduction in the source/drain depletion layer width prevents the depletion layer from penetrating into the transistor channel since the channel length is shorter.

Scaling Factor for Threshold Voltage

A primary parameter that affects the constant electric field scaling characteristics is the threshold voltage. The threshold voltage of a long channel MOS device is the sum of four voltage components [44, 45]:

- The voltage across the insulator due to the fixed charges Q_{eff} at the nonideal insulator-silicon interface
- The voltage across the insulator due to the depletion layer charge
- The work function difference between the gate and substrate ΔW_f, also referred to as the flatband voltage
- The surface potential of silicon ψ_s

Based on these components, the scaling factor for the threshold voltage V_t is [43]

$$V'_t = \frac{t_{in}}{S\epsilon_{in}}\left[-Q_{eff} + \sqrt{2\epsilon_{si}qSN_a(\psi_s + V_{SB}/S)}\right] + (\Delta W_f + \psi_s) \approx \frac{V_t}{S}, \tag{2.2}$$

where ϵ_{in} and t_{in} are, respectively, the insulator dielectric permittivity and thickness. According to (2.2), the change in threshold voltage due to the doping concentration is mostly compensated by the change in voltage. Thus, the threshold voltage scales primarily due to a reduction in the insulator thickness. Note that similar to (2.1), ψ_s is assumed to remain approximately constant due to the weak logarithmic dependence of ψ_s on N_a.

Scaling Factor for Current

The scaling factor for the transistor current is based on two different current-voltage equations: (1) the Shockley square-law MOS model that is valid for long channel transistors [18], where the channel length is greater than 1 μm, and (2) the Sakurai alpha-power law model that is valid for short channel transistors, where the channel length is around or smaller than 1 μm [46]. According to the square-law model, the scaled linear $I'^{\square}_{lin}$ and saturation currents $I'^{\square}_{sat}$ are, respectively,

$$I'^{\square}_{lin} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}\left(\frac{V_{GS} - V_t - V_{DS}/2}{S}\right)\frac{V_{DS}}{S} = \frac{I^{\square}_{lin}}{S}, \tag{2.3}$$

$$I'^{\square}_{sat} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}\left(\frac{V_{GS} - V_t}{S}\right)^2 = \frac{I^{\square}_{sat}}{S}, \tag{2.4}$$

where μ_{eff} is the effective surface mobility of the charge carriers. As determined by both (2.3) and (2.4), according to the square-law model, the transistor current scales by $1/S$. Note that while the overall current is reduced, the current density (current per unit channel width) remains the same. Also note that the mobility is assumed constant with technology, and channel length modulation is neglected.

According to the alpha-power law model, the scaled linear I'^{α}_{lin} and saturation currents I'^{α}_{sat} are, respectively,

$$I'^{\alpha}_{lin} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}\frac{[(V_{GS} - V_t)/S]^{\alpha/2}[(V_{DD} - V_t)/S]^{\alpha/2}}{(V_{DD} - V_t)/S}V_{DS}/S$$
$$= I^{\alpha}_{lin}S^{1-\alpha}. \tag{2.5}$$

$$I'^{\alpha}_{sat} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}\left(\frac{V_{GS} - V_t}{S}\right)^{\alpha} = I^{\alpha}_{sat}S^{1-\alpha}. \tag{2.6}$$

Note that in (2.3), (2.4), (2.5), and (2.6) the threshold voltage is assumed to scale by the same factor ($1/S$) as the power supply voltage. Also note that in (2.5) and (2.6), α refers to the velocity saturation index which varies between one and two depending upon the technology [46]. For a short channel MOS device, α is close to one. In this case, according to (2.5) and (2.6), the current scaling factor is one. Thus, as opposed to long channel transistors where the device current scales by $1/S$, in short channel transistors, the device current remains the same with

technology, assuming constant electric field scaling. The effects of a constant device current on the delay and power are described in the following subsections. Also note that when $\alpha = 2$, the alpha-power law model based scaling factor for the current is equal to the square-law model based scaling factor.

Scaling Factor for Subthreshold Slope

The current of an MOS device in the subthreshold or weak inversion region is another important characteristic since an important figure-of-merit characterizing a transistor is the ratio of the on-current to the off-current. When the gate-to-source voltage is smaller than the threshold voltage (for an NMOS transistor), the current is exponentially dependent on the gate voltage. The slope S_{sub} describing this dependence is referred to as the "subthreshold slope." A lower subthreshold slope is desirable to achieve sharper turn-on characteristics. Alternatively, a higher subthreshold slope causes higher subthreshold leakage current. The scaling characteristic of the subthreshold slope is [43]

$$S'_{sub} = \frac{\partial V'_{GS}}{\partial \log_{10} I'_{DS}} = \ln 10 \frac{kT}{q}\left[1 + \frac{(\epsilon_{si}S)/w_d}{(\epsilon_{in}S)/t_{in}}\right] = S_{sub} \text{ volts/decade.} \tag{2.7}$$

Unfortunately, the subthreshold slope does not scale well with technology. This issue is an undesirable result of technology scaling since nanoscale MOS devices suffer from significant subthreshold leakage current.

Scaling Factor for Capacitance and On-Resistance

Assuming the device capacitance is dominated by the gate capacitance, the scaling characteristic of the gate capacitance C_g is

$$C'_g = \frac{\epsilon_{in}}{t_{in}/S}\frac{W}{S}\frac{L}{S} = \frac{C_g}{S}. \tag{2.8}$$

Despite the increase in the unit capacitance due to a reduction in the insulator thickness, the overall gate capacitance decreases since both the width and channel length are reduced.

Assuming that the transistor operates in the linear region, the scaling characteristic of the on-resistance is

$$R'^{\square}_{on} = \frac{1}{\mu_{eff}\frac{\epsilon_{in}}{t_{in}/S}\frac{W/S}{L/S}\frac{V_{GS}-V_t}{S}} = R^{\square}_{on}. \tag{2.9}$$

Thus, the on-resistance of a transistor does not scale with technology, assuming a square-law model for the current. Note that a similar result can be obtained by dividing the voltage by the current since both quantities scale with $1/S$.

If the alpha-power law model is utilized for the current, the scaling characteristic of the on-resistance changes to

$$R'^{\alpha}_{on} = \frac{1}{\mu_{eff}\frac{\epsilon_{in}}{t_{in}/S}\frac{W/S}{L/S}\frac{[(V_{GS}-V_t)/S]^{\alpha/2}[(V_{DD}-V_t)/S]^{\alpha/2}}{(V_{DD}-V_t)/S}} = R^{\alpha}_{on}S^{\alpha-2}. \tag{2.10}$$

According to (2.10), the on-resistance starts to decrease with technology as α approaches one. A similar result is obtained by dividing the voltage scaling factor ($1/S$) by the alpha-power law model scaling factor for the current ($S^{1-\alpha}$). The scaling characteristics of the gate capacitance and on-resistance play an important role in the scaling factor for delay, as described in the following subsection.

Scaling Factor for Delay

Assuming a first order approximation, the scaling factor for the delay is determined by the overall load capacitance, the difference between the final and initial voltages, and average current. The scaled propagation delay based on the square-law model $t'^{\square}_d$ and alpha-power law model t'^{α}_d are, respectively,

$$t'^{\square}_d = \frac{(C_g/S)(\Delta V/S)}{I^{\square}_{DS}/S} = \frac{t_d}{S}, \tag{2.11}$$

$$t'^{\alpha}_d = \frac{(C_g/S)(\Delta V/S)}{I^{\alpha}_{DS}S^{1-\alpha}} = t_d S^{\alpha-3}, \tag{2.12}$$

where ΔV is the difference between the final and initial voltage across the capacitance C_g. The delay therefore decreases linearly ($1/S$) for a long channel transistor, as determined by (2.11). If a short channel transistor is assumed, the scaling factor varies between $1/S$ and $1/S^2$, as described by (2.12). Thus, the reduction in delay with scaling is higher for a short channel transistor assuming that small geometry effects are neglected. From a delay perspective, constant electric field scaling is desirable for both long and short channel transistors, as described, respectively, by (2.11) and (2.12).

Scaling Factor for Power, Power-Delay Product, and Power Density

The overall power consumption of a device is determined by multiplying the voltage and average channel current. Thus, the power consumption after scaling based on the square-law model $P'^{\square}$ and alpha-power law model P'^{α} is, respectively,

$$P'^{\square} = \frac{I^{\square}_{DS}}{S}\frac{V_{DD}}{S} = \frac{P^{\square}}{S^2}, \tag{2.13}$$

$$P'^{\alpha} = I^{\alpha}_{DS}S^{1-\alpha}\frac{V_{DD}}{S} = P^{\alpha}S^{-\alpha}. \tag{2.14}$$

According to (2.13) and (2.14), constant electric field scaling is desirable not only for delay, but also power consumption. Note, however, that the reduction in power is higher for long channel transistors since the scaling factor is $1/S^2$. Alternatively, for short channel transistors, the power scaling factor varies between $1/S^2$ and $1/S$. Thus, in the worst case, the power consumption decreases linearly with technology, assuming constant electric field scaling.

A useful design criterion that considers both delay and power consumption with equal weight is the power-delay product. Utilizing (2.11), (2.12), (2.13), and (2.14), the scaled power-delay product $PDP'^{\square}$ for the square-law model and alpha-power law model PDP'^{α} is, respectively,

$$PDP'^{\square} = \frac{P^{\square}}{S^2}\frac{t_d^{\square}}{S} = \frac{PDP^{\square}}{S^3}, \tag{2.15}$$

$$PDP'^{\alpha} = P^{\alpha}S^{-\alpha}t_d^{\alpha}S^{\alpha-3} = \frac{PDP^{\alpha}}{S^3}. \tag{2.16}$$

According to (2.15) and (2.16), the power-delay product scales with $1/S^3$ for both short and long channel transistors. This significant reduction in power-delay product with technology had been the primary motivation and one of the most significant advantages of constant electric field scaling. Note that for long channel transistors, the reduction in the power-delay product is achieved mostly due to a reduction in power, whereas for short channel transistors, the reduction in the power-delay product is primarily due to a smaller delay. In both cases, however, the power-delay product scales by $1/S^3$.

Finally, the scaling characteristics of the power density are considered. In a typical IC, the maximum power density is usually limited by the maximum junction temperature and cooling capability [47]. The scaled power density for the square-law model $P'^{\square}_{den}$ and alpha-power law model P'^{α}_{den} is, respectively,

$$P'^{\square}_{den} = \frac{P^{\square}/S^2}{(WL)/S^2} = P^{\square}_{den}, \tag{2.17}$$

$$P'^{\alpha}_{den} = \frac{P^{\alpha}S^{-\alpha}}{(WL)/S^2} = P^{\alpha}_{den}S^{2-\alpha}. \tag{2.18}$$

As determined by (2.17) and (2.18), the scaling behavior of the power density is significantly different in short and long channel transistors. For a long channel transistor, the power density does not change with technology, whereas for a short channel transistor, the power density scaling factor varies between one and S. Thus, in short channel devices, the power density increases with technology, causing significant thermal integrity issues.

2.1.3 Constant Voltage Scaling

As mentioned previously, in constant electric field scaling, the threshold voltage is scaled by $1/S$ to maintain performance, as predicted by (2.11) and (2.12). Challenges in controlling the subthreshold leakage current and related process parameters that determine the threshold voltage, however, have limited the ability to scale the threshold voltage. Consequently, scaling of the power supply voltage has also slowed down. Starting at the 130/90 nm technology node, the power supply voltage has been maintained constant at approximately one volt [42]. It is therefore important to understand constant voltage scaling characteristics for technologies below the 130/90 nm nodes.

In the constant voltage scaling scheme, only the physical dimensions and doping profiles are scaled, whereas the voltages remain the same with technology scaling. Again, assuming a unitless scaling factor S ($S > 1$), the transistor width W, channel length L, and gate insulator thickness t_{in} scale by $1/S$. Note that the electric fields increase since the voltage does not scale. The substrate doping concentration N_a increases by S^2 due to two primary reasons: (1) to maintain a constant threshold voltage, and (2) to scale the depletion layer width in proportion to the channel length. Under these conditions, the scaling factor for multiple parameters is described in the following subsections.

Scaling Factor for Depletion Layer Width

Utilizing (2.1), the depletion layer width after constant voltage scaling is

$$w'_d = \sqrt{\frac{2\epsilon_{si}(\psi_b + V_{SB})}{q\,S^2N_a}} = \frac{w_d}{S}. \tag{2.19}$$

Note that as opposed to constant field scaling where the doping concentration scales by S, in constant voltage scaling, the doping concentration scales by S^2 to compensate for the constant voltage. According to (2.19), the depletion layer width therefore scales by $1/S$, as desired.

Scaling Factor for Threshold Voltage

Utilizing (2.2), the threshold voltage after constant voltage scaling is

$$V'_t = \frac{t_{in}}{S\epsilon_{in}}\left[-Q_{eff} + \sqrt{2\epsilon_{si}q\,S^2N_a(\psi_s + V_{SB})}\right] + (\Delta W_f + \psi_s) \approx V_t. \tag{2.20}$$

According to (2.2), the threshold voltage remains approximately the same since the increase in the threshold voltage due to a higher doping concentration is compensated by the decrease in the threshold voltage due to a smaller gate insulator thickness. This result is consistent with constant voltage scaling.

Scaling Factor for Current

Unlike constant electric field scaling, in a constant voltage scaling scenario, the scaling factor for the transistor current is not dependent upon the current model used in the analysis. For both the Shockley square-law MOS model and Sakurai alpha-power law model, the scaling factor for the current is the same since the voltage remains the same. Therefore, the exponent of the voltage in the current equations does not affect the current scaling factor. Thus, either current expression can be used in the analysis.

Assuming the alpha-power law model, the scaled linear and saturation currents are, respectively,

$$I'^{\alpha}_{lin} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}\frac{(V_{GS}-V_t)^{\alpha/2}(V_{DD}-V_t)^{\alpha/2}}{V_{DD}-V_t}V_{DS} = SI^{\alpha}_{lin}. \quad (2.21)$$

$$I'^{\alpha}_{sat} = \mu_{eff}\frac{\epsilon_{in}}{t_{in/S}}\frac{W/S}{L/S}(V_{GS}-V_t)^{\alpha} = SI^{\alpha}_{sat}. \quad (2.22)$$

The transistor current therefore increases by S with technology under a constant voltage scenario.

Scaling Factor for Subthreshold Slope

The scaling characteristic of the subthreshold slope for the constant voltage scaling scheme is

$$S'_{sub} = \frac{\partial V'_{GS}}{\partial \log_{10} I'_{DS}} = \ln 10\frac{kT}{q}\left[1+\frac{(\epsilon_{si}S)/w_d}{(\epsilon_{in}S)/t_{in}}\right] = S_{sub}\ \text{volts/decade}. \quad (2.23)$$

Based on constant voltage scaling, the subthreshold slope remains the same with technology. This result is similar to the constant electric field scaling scheme. Thus, according to both scaling schemes, MOS devices suffer from subthreshold leakage current as the device dimensions are reduced.

Scaling Factor for Capacitance and On-Resistance

According to the constant voltage scaling scheme, the scaling characteristic of the gate capacitance C_g is

$$C'_g = \frac{\epsilon_{in}}{t_{in}/S}\frac{W}{S}\frac{L}{S} = \frac{C_g}{S}. \quad (2.24)$$

Similar to constant electric field scaling, the overall gate capacitance scales by $1/S$ since both the width and channel length are reduced.

Assuming that the transistor operates in the linear region, the scaling characteristic of the on-resistance is

$$R'^{\alpha}_{on} = \frac{1}{\mu_{eff}\frac{\epsilon_{in}}{t_{in}/S}\frac{W/S}{L/S}\frac{(V_{GS}-V_t)^{\alpha/2}(V_{DD}-V_t)^{\alpha/2}}{V_{DD}-V_t}} = \frac{R^{\alpha}_{on}}{S}. \quad (2.25)$$

Note that since the power supply voltage is maintained constant, the transistor on-resistance decreases with technology. A similar result is obtained by dividing the voltage scaling factor (1) by the transistor current scaling factor, as determined by (2.21) and (2.22).

Scaling Factor for Delay

The scaled propagation delay under a constant voltage scaling scenario is

$$t'^{\alpha}_d = \frac{(C_g/S)\Delta V}{SI^{\alpha}_{DS}} = \frac{t_d}{S^2}. \quad (2.26)$$

Thus, the propagation delay decreases quadratically with technology due to a reduction in the load capacitance and an increase in the transistor current. Note that in the constant electric field scaling scenario, the scaling factor for delay varies between $1/S$ and $1/S^2$ depending upon whether the square-power or alpha-power law is used for the transistor current. Thus, considering only the transistor delay, constant voltage scaling is more beneficial than constant electric field scaling.

Scaling Factor for Power, Power-Delay Product, and Power Density

Despite the quadratic reduction in delay, the constant voltage scaling scenario has deleterious effects on power consumption and power density. Specifically, the scaled power consumption is

$$P'^{\alpha} = I^{\alpha}_{DS}SV_{DD} = SP^{\alpha}. \quad (2.27)$$

According to (2.27), the power consumption increases by S since the voltage does not scale and the current increases by S. Note that in the constant electric field scaling scenario, the scaling factor for the power varies between $1/S$ and $1/S^2$. Thus, power is linearly reduced, even in the worst case where the alpha-power law model is assumed. Alternatively, a linear increase in power under a constant voltage scaling scenario places a practical limit on technology scaling due to the degraded thermal stability and cooling requirements that exceed existing capabilities.

The scaled power-delay product for the constant voltage scaling scenario is

$$PDP'^{\alpha} = SP^{\alpha}\frac{t^{\alpha}_d}{S^2} = \frac{PDP^{\alpha}}{S}. \quad (2.28)$$

Thus, the power-delay product scales only by $1/S$ since the reduction in delay is partly compensated by the increase in power consumption.

Finally, the scaled power density is

$$P'^{\alpha}_{den} = \frac{SP^{\alpha}}{(WL)/S^2} = S^3 P^{\alpha}_{den}. \tag{2.29}$$

According to (2.29), under a constant voltage scaling scenario, the power density increases significantly, degrading system reliability and requiring expensive cooling techniques.

2.1.4 Comparison between Device Scaling Scenarios

The device scaling characteristics for both constant electric field and constant voltage scaling scenarios are summarized in Table 2.1, where the scaling factor for each device characteristic is listed. For a constant electric field, the square-power and alpha-power laws are considered separately. According to the scaling factors listed in this table, technology scaling is largely favorable for active devices. An important motivation for technology scaling has been a reduction in transistor delay and increase in transistor density, which are valid for each scaling scenario. Furthermore, the power-delay product is also reduced in each case, justifying the positive effect of technology scaling on performance. Note, however, that the decrease in the power-delay product has slowed down with constant voltage scaling. Also note that if

Performance parameter	Constant electric field		Constant voltage
	Square-power law	Alpha-power law	
Threshold voltage	$1/S$	$1/S$	1
Current	$1/S$	$S^{1-\alpha}$	S
Subthreshold slope	1	1	1
Capacitance	$1/S$	$1/S$	$1/S$
On-resistance	1	$S^{\alpha-2}$	$1/S$
Delay	$1/S$	$S^{\alpha-3}$	$1/S^2$
Power	$1/S^2$	$S^{-\alpha}$	S
Power-delay product	$1/S^3$	$1/S^3$	$1/S$
Power density	1	$S^{2-\alpha}$	S^3

Note: For constant electric field scaling, two transistor current models, square-power and alpha-power law, are considered separately.

TABLE 2.1 Scaling Characteristics of a Transistor for Both Constant Electric Field and Constant Voltage Scaling Scenarios

only power consumption is considered, there are several important drawbacks of technology scaling:

- A significant increase in power density and power consumption under the constant voltage scaling scenario
- The constant subthreshold slope in each scaling scenario

As predicted by these two drawbacks, reducing the power consumption and power density has become one of the most significant challenges of technology scaling. This issue is further exacerbated by the constant voltage scaling scheme, which replaced the constant electric field scheme since any further reduction in the power supply voltage was not feasible below the 130/90 nm technology nodes. Furthermore, the transistor off-current (I_{off}) does not scale due to a constant subthreshold slope. Thus, the transistor quality is degraded since the I_{on}/I_{off} ratio has been gradually decreased. Due to these reasons, nanoscale CMOS technology is typically assumed to be a power-constrained era.

The scaling analysis performed in this section does not consider nonideal effects due to small geometries in MOS devices. Also referred to as short- and narrow-channel effects, these mechanisms significantly limit the benefits of technology scaling, requiring further advances in material structures and fabrication processes. These effects are discussed in the following section.

2.2 Small Geometry Effects

As device dimensions have been reduced, the Shockley model of the current-voltage characteristics of a device has become inaccurate. Specifically, according to the one-dimensional gradual channel approximation (GCA), the lateral and vertical electric field components do not interact within an MOS device [45]. Thus, the voltage within the channel from the source to the drain varies independently of the perpendicular voltage from the gate to the substrate. This assumption enables a device to be modeled with simple one-dimensional equations and is relatively accurate for long channel transistors. With smaller dimensions, however, the accuracy of the GCA degrades due to several small geometry effects. Understanding and accurately characterizing these effects is crucial since MOS device operation is significantly influenced by these effects.

Gaensslen investigated small geometry effects as early as 1979 [44]. He used the *electric device size* as opposed to the geometric size of a device, where the electric size refers to the geometric size divided by the depletion layer width. According to this definition, a short channel transistor refers to a device with a channel length of the same order of magnitude as the depletion layer width. Similarly, a narrow channel

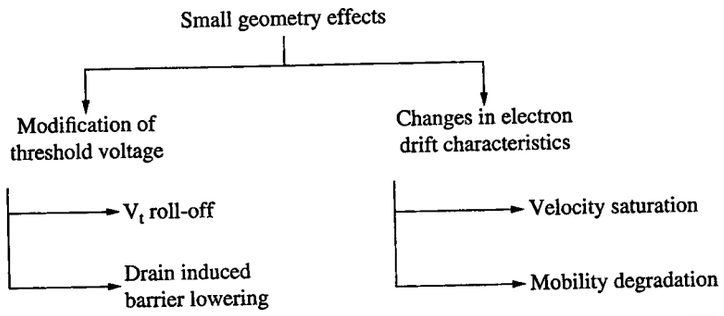


FIGURE 2.6 Classification of small geometry effects in MOS transistors.

transistor refers to a device with a channel width that is of the same order of magnitude as the depletion layer width. Note that the depletion layer width is a function of the reverse bias voltage, as described by (2.1). A conservative approach is to assume that the reverse bias voltage is equal to the power supply voltage, producing the widest depletion layer. If the channel length and width are comparable to this depletion layer width, the device is assumed to exhibit, respectively, short and narrow channel effects [44].

As illustrated in Fig. 2.6, small geometry effects can be classified into two categories [45]:

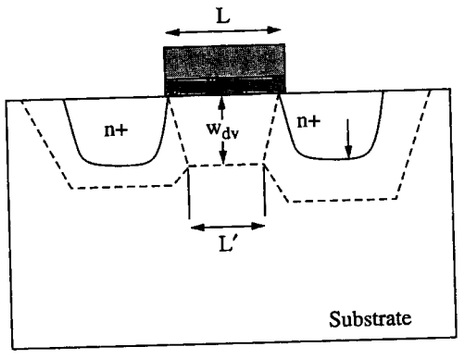
- Modification of the threshold voltage
- Changes in the electron drift characteristics within the channel

Small geometry effects that modify the threshold voltage V_{th} are included in the first category. These effects are V_t roll-off and drain induced barrier lowering (DIBL), as described respectively, in Sections 2.2.1 and 2.2.2. Alternatively, small geometry effects due to varying drift characteristics are included in the second category. These effects are velocity saturation and mobility degradation as described, respectively, in Sections 2.2.3 and 2.2.4.

2.2.1 Threshold Voltage Roll-Off

As mentioned previously, the threshold voltage of an MOS device consists of four voltage components. The work function difference (flatband voltage) and the surface potential do not change with device geometry [44]. The voltage across the insulator due to the depletion layer charge, as described by the second term in (2.2), however, varies with the size of the MOS device. In long channel MOS devices, this depletion charge is primarily due to the vertical electric field generated by the gate voltage. This charge is typically referred to as the bulk depletion charge [45]. Alternatively, in short channel MOS devices, the

FIGURE 2.7 Contribution of the source and drain regions to the overall depletion charge within the channel [48].



source and drain regions induce a nonnegligible amount of depletion charge within the channel. A significant portion of the overall depletion charge is therefore provided by the source and drain regions, as depicted in Fig. 2.7 [48]. In a long channel transistor, the overall depletion charge is proportional to $w_{dv} \times L$, where w_{dv} is the depletion width in the vertical direction and L is the channel length. Alternatively, in a short channel transistor, the overall depletion charge is proportional to the area of the trapezoid $(w_{dv} \times (L + L')/2)$, as shown in Fig. 2.7. The bulk depletion charge induced by the gate voltage is therefore lower. Thus, the second term in (2.2) overestimates the gate induced depletion charge, thereby producing a threshold voltage higher than the actual threshold voltage of a short channel MOS device. This effect is referred to as threshold voltage (V_t) roll-off and is one of the more significant short channel effects.

Considering this effect, the zero bias threshold voltage V_{t0}^{sc} of a short channel MOS device is

$$V_{t0}^{sc} = V_{t0} - \Delta V_t, \tag{2.30}$$

where V_{t0} and ΔV_t refer, respectively, to the zero bias threshold voltage of a long channel MOS device and a change in the threshold voltage due to short channel effects. Approximating the gate induced bulk depletion region with a trapezoid and the overall depletion region (due to both the gate voltage and source/drain regions) with a rectangle, ΔV_t can be described as the charge difference between a rectangular and trapezoidal shaped depletion region [45]. Utilizing this approximation, ΔV_t is

$$\Delta V_t = \frac{t_{in}}{\epsilon_{in}} \sqrt{2\epsilon_{si} q N_a \psi_s} \frac{x_j}{2L} \left[\left(\sqrt{1 + \frac{2w_{dS}}{x_j}} - 1 \right) + \left(\sqrt{1 + \frac{2w_{dD}}{x_j}} - 1 \right) \right], \tag{2.31}$$

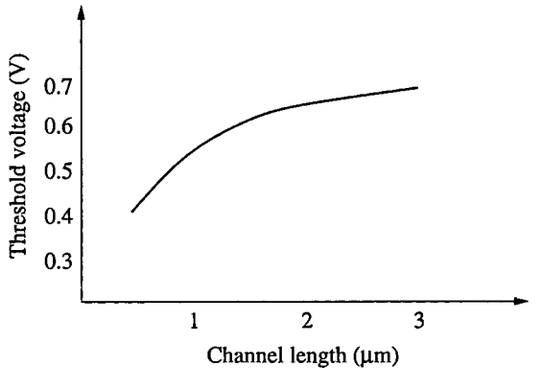


FIGURE 2.8 Variation of the threshold voltage as a function of channel length in an N-type MOS transistor [45].

where x_j is the source/drain junction depth, and w_{dS} and w_{dD} refer, respectively, to the source and drain depletion widths. Note that ΔV_t is proportional to the channel length L. For shorter channel lengths, ΔV_t increases significantly. Alternatively, for longer channel lengths where $L >> x_j$, ΔV_t approaches zero. The variation of the threshold voltage as a function of channel length is illustrated in Fig. 2.8, assuming the other parameters remain the same [45].

In addition to the channel length, several other parameters affect ΔV_t. According to (2.31), ΔV_t increases with a larger gate insulator thickness and junction depth. Thus, V_t roll-off is alleviated by two methods

- Decreasing the gate insulator thickness t_{in}
- Decreasing the junction depth, i.e., producing a shallow junction

A thinner insulator layer enhances the ability of the gate to control the channel. Similarly, a shallow junction reduces the depletion charge induced by the source/drain regions. The drawback of the first method is a significant increase in gate leakage current. Furthermore, according to (2.20), as the oxide thickness is reduced, the substrate doping concentration should be increased to maintain a constant threshold voltage. A higher doping concentration, however, degrades the carrier mobility due to channel impurity scattering [39]. A primary drawback of the second technique is an increase in the junction resistance due to a reduction in the cross sectional area.

2.2.2 Drain-Induced Barrier Lowering

In small geometry MOS devices, the channel depletion is controlled not only by the gate voltage, but also the drain voltage. At high drain

voltages, the channel potential (potential barrier) is affected by the drain-induced depletion region since the depletion layer is wider due to a positive bias voltage, as determined by

$$w_{dD} = \sqrt{\frac{2\epsilon_{si}(\psi_b + V_{DB})}{q N_a}}. \tag{2.32}$$

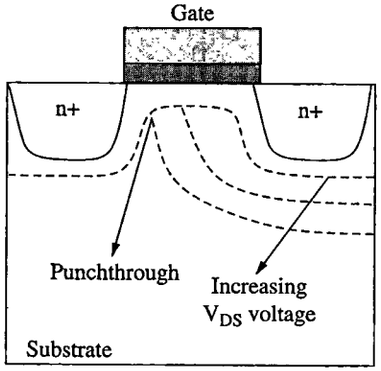
Thus, the transistor channel is inverted at a lower gate voltage than the threshold voltage since the drain of the transistor contributes significantly to the channel depletion region. This mechanism is referred to as drain-induced barrier lowering (DIBL). Note that DIBL is similar to V_t roll-off since both mechanisms are due to junction-induced depletion regions. DIBL is more significant since the drain-induced depletion region is larger due to a higher drain voltage. The effect of DIBL is quantified by analyzing the change in the threshold voltage. At a specific drain-to-bulk voltage V_{DB}, (2.32) is replaced in (2.31) to describe the change in the threshold voltage. The magnitude of DIBL is

$$DIBL = V_t|_{V_D \approx 100\ \text{mV}} - V_t|_{V_D = V_{DD}}, \tag{2.33}$$

where $V_t|_{V_D \approx 100\ \text{mV}}$ is the threshold voltage at a drain voltage of approximately 100 mV, and $V_t|_{V_D = V_{DD}}$ refers to the threshold voltage when the drain voltage is equal to the power supply voltage.

If the drain-induced depletion region is sufficiently large, the drain- and source-induced depletion regions can merge, producing an inverted channel that is independent of the gate voltage, as depicted in Fig. 2.9. This condition, referred to as punchthrough [48], should be avoided since the gate voltage cannot control the channel under this condition.

FIGURE 2.9 Illustration of punchthrough when drain- and source-induced depletion regions merge, producing an inverted channel that is independent of the gate voltage.



2.2.3 Velocity Saturation

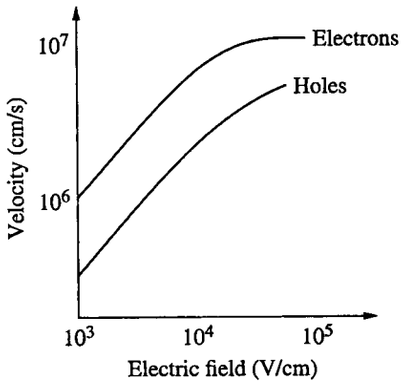
Velocity saturation is related to a change in the drift characteristics of the charge carriers under a high electric field. Specifically, at low lateral (drain-to-source) electric fields, as described by V_{DS}/L, the electron drift velocity is proportional to the electric field, where the proportionality constant is referred to as the carrier mobility. Alternatively, at high electric fields, the interaction of the charge carriers with the silicon lattice changes due to a higher number of collisions with the optical phonons [49]. Thus, the drift velocity cannot increase any further. This saturation is reached at a lateral electric field of approximately 10^5 V/cm [45]. At this electric field, the saturated drift velocities for electrons and holes are, respectively, 10^7 cm/s and 8×10^6 cm/s [50]. The relationship between the drift velocity and lateral electric field is depicted in Fig. 2.10 [51].

According to the constant voltage scaling scheme, the drain voltage remains constant whereas the channel length is reduced, thereby increasing the lateral electric field. Velocity saturation is therefore an important effect in short channel MOS devices operating in the saturation region since the lateral electric field is higher due to a greater drain voltage.

According to the Shockley square-law current model [18,19], the transistor current is quadratically dependent on the gate-to-source voltage. For a short channel MOS device, however, the dependence of the current on voltage is less than quadratic due to velocity saturation. As previously noted, the Sakurai alpha-power law model considers this effect, where the coefficient alpha varies between one and two. Furthermore, the current is independent of the effective channel length for a velocity saturated device. Assuming that velocity saturation has been reached, the maximum transconductance g_m^{max} of an MOS device is [52]

$$g_m^{max} = W v_{sat} C_{in}, \tag{2.34}$$

FIGURE 2.10
Relationship between drift velocity and lateral electric field for both electrons and holes in an MOS transistor [51].



where v_{sat} is the drift velocity under saturation and $C_{in} = \epsilon_{in}/t_{in}$ is the gate insulator capacitance per unit area.

2.2.4 Mobility Degradation

The drift characteristics of the charge carriers within the channel of an MOS device are affected not only by the lateral electrical field but also the vertical electric field. Specifically, the vertical electric field from the gate to the bulk node modifies the scattering of the characteristics of the carriers. A high gate voltage increases the number of collisions at the silicon-insulator interface. Thus, at high vertical electric fields, the effective surface mobility is reduced below the bulk mobility. The relationship between the vertical electric field and mobility is illustrated in Fig. 2.11 [53].

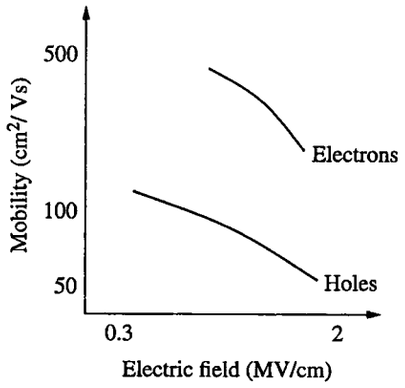
An empirical model has been developed to estimate the carrier mobility under high electric fields [54]. According to this model, developed by Chen et al. [54], the high field effective mobility is determined by scaling the low field surface mobility. The electron and hole carrier mobilities μ_n and μ_p as functions of the gate-to-source voltage, threshold voltage, and insulator thickness are, respectively [54],

$$\mu_n(V_{GS}, V_t, t_{in}) = \frac{540\ cm^2/Vs}{1+(\frac{E_{eff}}{0.9})^{1.85}} = \frac{540}{1+(\frac{V_{GS}+V_t}{5.4t_{in}})^{1.85}}, \tag{2.35}$$

$$\mu_p(V_{GS}, V_t, t_{in}) = \frac{185\ cm^2/Vs}{1+\frac{E_{eff}}{0.45}} = \frac{185}{1+\frac{|V_{GS}|+|1.5V_t|}{3.38t_{in}}}, \tag{2.36}$$

where E_{eff} (MV/cm) is the effective vertical electric field and 540 and 185 are, respectively, the electron and hole surface mobility under a low electric field. The units of the voltage and insulator thickness are, respectively, MV (million volts) and cm.

FIGURE 2.11
Relationship between vertical electric field and mobility for both electrons and holes in an MOS transistor [53].



2.3 Device Enhancements

The small geometry effects described in the previous section degrade MOS device performance and increase undesirable leakage currents. For CMOS technologies below 90 nm, these effects negate the speed gains due to device scaling [42]. Thus, structural enhancements to produce higher carrier mobilities while suppressing small geometry effects play an important role in improving performance in sub-90 nm technologies.

Three scaling scenarios are compared in Fig. 2.12 [42]. In the first case, scaling an MOS device assumes constant voltage scaling without any structural enhancements. In the second case, electron and hole mobilities are increased by 1.5 times per each technology generation. Finally, in the last case, in addition to an increase in the mobility, the effect of V_t roll-off and DIBL are maintained constant. As depicted in this figure, for the first case, the performance initially remains constant until approximately the 45 nm node, and degrades with more advanced technology nodes since small geometry effects outperform any performance gains achieved from scaling. If the mobilities are higher, as in the second case, the performance improvement is initially greater. As the technology advances, however, the small geometry effects become dominant and the performance starts to degrade at approximately the 35 nm node. A continuous performance improvement is achieved only in the last case. Note that even in this case, the improvement slows as the technology advances.

As emphasized in Fig. 2.12, both mobility enhancement and suppression of small geometry effects are critical to maintaining

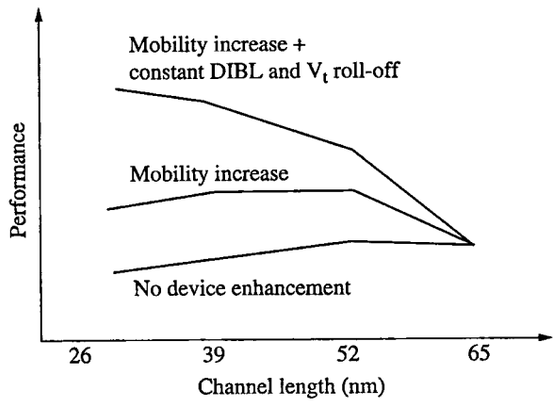


FIGURE 2.12 Comparison of three scaling scenarios demonstrating that the suppression of short channel effects is critical to achieving increased performance [42].

the benefits provided by technology scaling. Several techniques to achieve these goals are described in the following sections. Specifically, nonuniform channel doping is discussed in Section 2.3.1 to better control an MOS device. Strain engineering is described in Section 2.3.2 to enhance device mobility. Utilizing a higher gate dielectric permittivity is reviewed in Section 2.3.3. Finally, multiple gate MOS devices are discussed in Section 2.3.4.

2.3.1 Nonuniform Channel Doping

Nonuniform channel doping is a standard method for suppressing certain small geometry effects (specifically, V_t roll-off and DIBL) below 0.25 μm CMOS technologies. As described in Sections 2.2.1 and 2.2.2, a wider depletion layer worsens short channel effects since the ability of the gate node to control the channel is degraded. According to (2.32), a higher substrate doping concentration reduces the depletion layer width. The two important drawbacks of a higher substrate doping concentration are:

- An increase in the threshold voltage, as described by (2.2)
- A decrease in the carrier mobility

For these two reasons, a uniform increase in the doping concentration is typically avoided.

The primary purpose of a nonuniform channel doping [48], also referred to as halo doping, is to alleviate the effects of V_{th} roll-off and DIBL by reducing the source/drain depletion regions, while maintaining a constant threshold voltage and carrier mobility. As illustrated in Fig. 2.13, near the source/drain and channel boundaries, the substrate doping concentration is greater. Both the source- and drain-induced depletion layer widths are therefore reduced since charge sharing

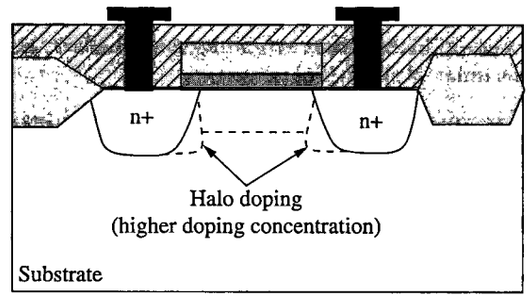


FIGURE 2.13 Higher doping concentration, also referred to as halo doping, near the source/drain and channel boundaries to alleviate the effects of V_{th} roll-off and DIBL [48].

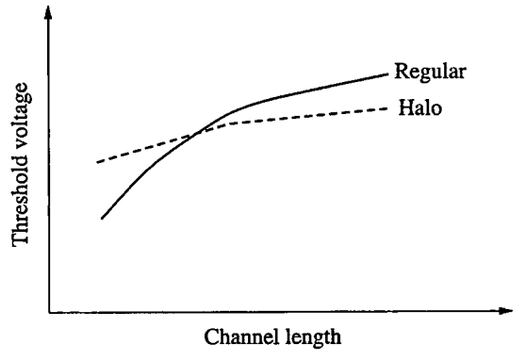


FIGURE 2.14 Weaker dependence of the threshold voltage on channel length for halo doping [48].

between the source/drain and bulk is alleviated due to the higher doping concentration within the substrate. Thus, as depicted in Fig. 2.14, the dependence of the threshold voltage on the channel length is weaker with halo doping. Furthermore, the device is less likely to exhibit punchthrough since the depletion layer is more narrow.

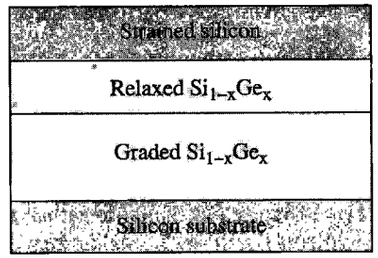
2.3.2 Strain Engineering

Strain engineering plays an important role in enhancing the carrier mobility for sub-90 nm CMOS technologies [55]. If a molecule is strained, the energy band structure of the molecule changes due to stress, thereby affecting the scattering characteristics [55]. Thus, the mobility of the device changes. There are two primary ways of introducing strain within the channel of an MOS device to enhance mobility:

- Substrate-induced strain producing a somewhat global effect, also referred to as bi-axial strain [56]
- Process-induced strain exhibiting a local effect, also referred to as uniaxial strain [56]

An effective technique to introduce strain is to grow strained silicon (Si) on a silicon-germanium (SiGe) layer, as shown in Fig. 2.15. Due to the lattice mismatch between these two layers, the lattice of the silicon layer (the channel of the device) experiences bi-axial tensile strain. The energy band diagram changes, resulting in reduced scattering and a smaller effective mass. Thus, the mobility is enhanced [57]. A 15% to 25% increase in the drive current has been demonstrated in bi-axial strained MOS devices [58,59]. However, substrate-induced strain primarily increases the electron mobility. The performance of the weaker PMOS transistors does not significantly increase. Another important

FIGURE 2.15 Strained silicon grown on a silicon germanium (SiGe) layer [55].



drawback of a strained Si/SiGe device is degraded thermal performance since the thermal conductivity of the SiGe layer is relatively low [59–61].

As opposed to substrate-induced strain, process-induced strain can be utilized to independently change the mobility characteristics of NMOS and PMOS devices. This difference is an important advantage since electrons and holes react differently to induced stress. Specifically, electron mobility is enhanced by tensile stress, whereas compressive stress enhances hole mobility. In process-induced strain, stress is introduced by utilizing additional processing steps such as stress liners, stress memorization, and filling source/drain regions with SiGe [57]. In these techniques, an additional layer such as nitride is deposited over the device to induce mechanical stress. Note that both tensile and compressive stress can be simultaneously induced by utilizing a dual stress liner technology [62]. The stress level depends upon the thickness of the nitride layer. In addition to nitride, SiGe can also be used to induce strain into the transistor channel. Source/drain regions are filled with SiGe, thereby producing compressive uni-axial strain, enhancing hole mobility. Tensile uni-axial strain can also be generated in NMOS transistors by replacing SiGe with SiC. Note that both electron and hole mobilities are also dependent upon the orientation of the substrate crystal and the direction of the channel [62]. Different substrate orientations and channel directions can therefore be utilized to further increase carrier mobilities [62].

2.3.3 Combining High-K and Metal Gate Structures

As previously mentioned, the use of a high permittivity (high-K) dielectric material rather than silicon dioxide has enabled the continuation of technology scaling below the 65 nm node by reducing the gate leakage current and alleviating short channel effects related to high vertical electric fields. A high-K material such as hafnium permits an increase in the insulator thickness, thereby reducing the leakage current. Electrically, a high-K insulator behaves as a thin oxide with an equivalent oxide thickness (EOT). Since an EOT of 0.9 nm has been demonstrated in a 32 nm CMOS technology [63], a high-K dielectric

material can be used to improve performance and alleviate short channel effects.

As further discussed in Chapter 11, several challenges exist to achieving a reliable MOS device with high-K dielectric materials. These challenges include threshold voltage instability, mobility degradation due to soft optical phonons, negative bias temperature instability, and time dependent dielectric breakdown. Some of these issues are alleviated by utilizing a metal gate rather than a polysilicon gate due to the enhanced stability characteristics of the high-K material with metal. For sub-22 nm CMOS technologies, an EOT smaller than 0.8 nm is required to control small channel effects and maintain any gains in performance. Thus, next generation devices will require dielectric materials with a relative permittivity greater than 30 [64].

2.3.4 Multiple Gate Devices

As mentioned previously, nonuniform channel doping has been utilized below 0.25 μm CMOS technologies to control V_t roll-off and DIBL, thereby decreasing the subthreshold current. For more advanced sub-32 nm technologies, however, nonuniform channel doping alone is not a sufficiently effective mechanism to reduce the subthreshold current.

According to the ITRS [32], enhanced electrostatic control of the transistor channel is a critical requirement for sub-25 nm technologies since the threshold voltage should be decreased to produce a sufficient gate overdrive voltage ($V_{GS} - V_t$) without increasing the subthreshold current [26]. Multigate devices have been proposed to achieve this requirement [65, 66]. There are two primary types of multigate devices:

- Planar multigate devices
- Vertical multigate devices

As shown in Fig. 2.16, planar double gate transistors consist of both a front gate and a back gate fabricated on a buried oxide, as in

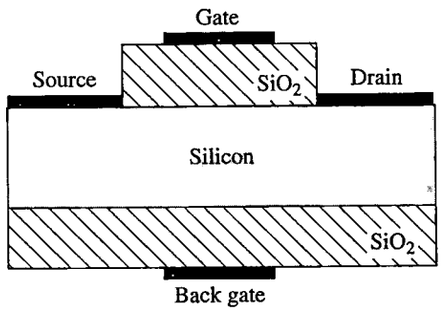


FIGURE 2.16
Planar double gate MOS transistor consisting of both a front gate and a back gate fabricated on a buried oxide [67].

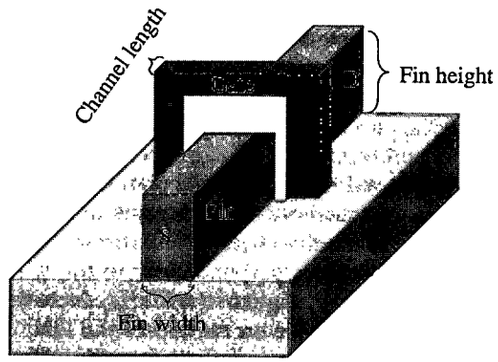


FIGURE 2.17
FinFET as a vertical multigate device [69].

silicon-on-insulator technology. Since the channel is controlled by the two gates, the energy barrier between the source and drain terminals is primarily determined by the gate terminal, thereby alleviating short channel effects and reducing subthreshold leakage current. Alignment of the two gates, however, is a difficult fabrication challenge in planar double gate devices [64]. This issue is exacerbated as the channel length is reduced [68]. Misalignment produces additional gate-to-source and gate-to-drain overlap capacitances, thereby degrading performance. Another challenge is to obtain a sufficiently thin film for the back gate since the back gate is not on top of the wafer [68].

Two emerging vertical multigate devices are FinFETs and triple gate FETs, as depicted, respectively, in Figs. 2.17 and 2.18 [69]. In both devices, the gate terminal achieves enhanced control of the transistor channel as compared to planar bulk CMOS technologies since the gate surrounds a larger area of the channel. Referring to Fig. 2.17, in a FinFET, the fin width W_{fin} should be sufficiently small as compared to the fin height H_{fin} to minimize subthreshold leakage current and decrease the source/drain series resistance [70]. The primary difference

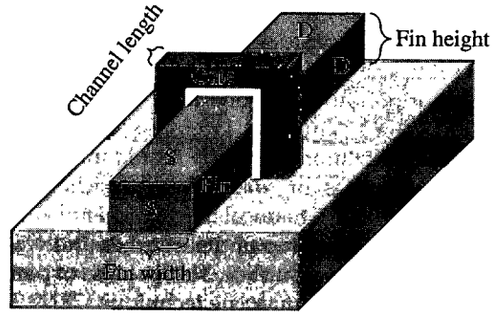


FIGURE 2.18
Triple gate FET consisting of an additional gate at the top of the FIN [69].

between a FinFET and a triple gate FET is the additional third gate that exists at the top of the fin in triple gate FETs. This additional gate further enhances the ability of the gate terminal to control the channel. Thus, process requirements are relatively more relaxed in triple gate FETs, allowing relatively larger fin widths while maintaining a constant subthreshold leakage current [69].

As opposed to conventional planar devices, in FinFETs and triple gate FETs the transistor channel forms along the vertical surface of the silicon fin. Thus, the current flows from the drain to source along the vertical surfaces perpendicular to the wafer. This dimension is determined by the height of the fin H_{fin}. Note that in a triple gate FET, the current also flows on the top of the FIN, as shown in Fig. 2.18. Thus, the effective channel width for a FinFET and triple gate FET are, respectively, $2 \times H_{fin}$ and $2 \times H_{fin} + W_{fin}$. Since the effective width of a triple gate device is higher, the current drive is enhanced. Note that the fin height H_{fin} is a process dependent parameter. Thus, the effective width of the vertical multigate transistors can only be increased by utilizing multifinger transistors, where the gate terminal is common. Only discrete effective widths can therefore be achieved. This drawback is important in static random access memory (SRAM) circuits since the transistor widths are carefully chosen to simultaneously satisfy both the read and write operations [69].

Vertical multigate devices achieve a significantly higher I_{on}/I_{off} ratio as compared to planar technologies. A triple gate MOS device is typically superior to a FinFET due to the additional gate. The clock frequency of a circuit composed of triple gate devices with a TiN metal gate has been shown to be 65% higher than a FinFET based circuit with a polysilicon gate [69].

2.4 Interconnect Scaling

System performance depends upon not only the scaling characteristics of the transistors, but also the interconnects. To increase overall integration density, the size of the interconnects is decreased. Alternatively, to maintain a steady improvement in system performance, the parasitic interconnect impedances should scale by the same scaling factor as the transistors. Unfortunately, these two requirements are difficult to simultaneously satisfy, as described in this section.

An important milestone in interconnect technology, first introduced by Andy Grove in 1959, was the introduction of planar multilevel metallization [71]. The insulating layer is first patterned to produce trenches or vias. These trenches and vias are filled by depositing metal. Chemical mechanical polishing (CMP) removes the additional metal on top of the trenches and vias, thereby achieving a highly planar metal surface. This process of patterning metal layers is referred to as the damascene technique [72].

Producing a highly planar surface is a critical step since scaling interconnects has been greatly facilitated by this capability. For example, the number of metal layers has steadily increased. A modern 32 nm CMOS technology consists of nine metallization layers [63]. The metal pitch has been reduced from 1.8 μm in 1994 to several hundred nanometers in 2011 [63,71].

The interconnect density has therefore increased due to these improvements, following a similar trend as for the transistors. The reduction in the cross sectional area increases the parasitic resistance, whereas a smaller pitch increases the lateral parasitic capacitance. A partial solution to this issue has been the development of a hierarchical wiring scheme that exploits these multiple metal layers, as described in Section 2.4.1. Despite this hierarchical scheme, interconnects have a significant effect on multiple performance parameters such as the delay, power consumption, noise, and reliability. The scaling characteristics of on-chip interconnects are described in this section. Specifically, interconnect characteristics under the ideal scaling scheme are described in Section 2.4.2. The effects of more realistic scaling scenarios are discussed in Section 2.4.3. Finally, several scaling scenarios are compared in Section 2.4.4.

2.4.1 Global versus Local Interconnects

As depicted in Fig. 2.19, in a hierarchical wiring scheme, the lower metal layers with smaller cross sectional dimensions achieve a high

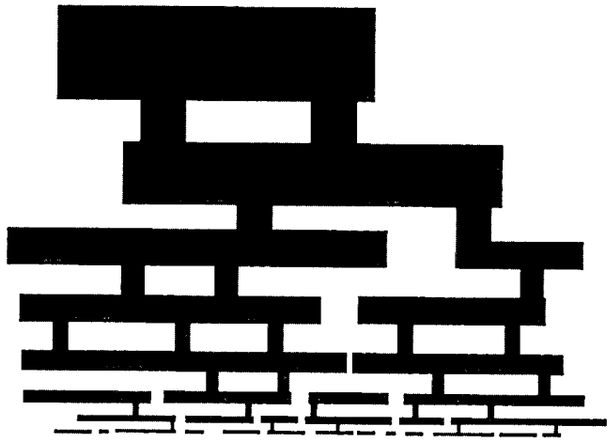


FIGURE 2.19 Cross section of a hierarchical wiring scheme where the lower metal layers with smaller cross sectional dimensions are utilized to achieve high density interconnections, whereas the upper metal layers with greater cross sectional dimensions provide a relatively low parasitic resistance and capacitance for global signaling.

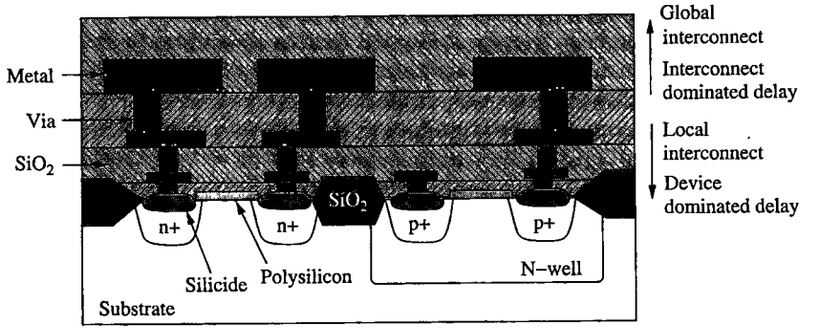


FIGURE 2.20 Local versus global interconnects. The global interconnects are typically located on the upper metal layers, whereas the local interconnects are closer to the devices.

interconnection density, whereas the less dense upper metal layers with greater cross sectional dimensions produce a relatively low parasitic resistance and capacitance to achieve global signaling. Interconnects at the lower metal layers are typically used as local interconnects, whereas the higher metal layers are often used as global interconnects. Those metal layers closer to the upper levels are referred to as semi-global interconnects.

As illustrated in Fig. 2.20, the local interconnects typically transmit signals at the device level, within a gate or circuit block. The gate delays typically dominate at this level. Alternatively, the global interconnects typically transmit signals across the integrated circuit. Important examples of global interconnects include global signal lines (see Chapter 6), power networks (see Chapter 8), and clock distribution networks (see Chapter 15). The interconnect delay dominates the gate delay in these global interconnects. Due to an increase in overall die area, the length of the global interconnects typically grows by a factor S_c ($S_c > 1$), referred to as the chip scaling factor [73,74].

The scaling characteristics of the local and global interconnect dimensions are listed in Table 2.2 for four cases:

- Ideal scaling
- Quasi-ideal scaling
- Scaling based on constant resistance
- Scaling based on constant thickness

The interconnect scaling characteristics for each of these cases are described in the following sections.

Interconnect parameters	Ideal scaling		Quasi-ideal scaling		Constant resistance		Constant thickness	
	Local	Global	Local	Global	Local	Global	Local	Global
Length (L_{int})	1/S	S_c	1/S	S_c	1/S	S_c	1/S	S_c
Width (W_{int})	1/S	1/S	1/S	1/S	$1/\sqrt{S}$	$1/\sqrt{S}$	1/S	1/S
Thickness (T_{int})	1/S	1/S	$1/\sqrt{S}$	$1/\sqrt{S}$	$1/\sqrt{S}$	$S_c\sqrt{S}$	1	1
Height (H)	1/S	1/S	$1/\sqrt{S}$	$1/\sqrt{S}$	$1/\sqrt{S}$	$1/\sqrt{S}$	1/S	1/S
Spacing (W_{spa})	1/S	1/S	1/S	1/S	$1/\sqrt{S}$	$1/\sqrt{S}$	1/S	1/S
Aspect ratio (AR)	1	1	$S/\sqrt{S}$	$S/\sqrt{S}$	1	S_cS	S	S
Resistance	S	S_cS^2	$\sqrt{S}$	$S_cS\sqrt{S}$	1	1	1	SS_c
Coupling capacitance (C^c)	1/S	S_c	$1/\sqrt{S}$	$S_c\sqrt{S}$	1/S	S_c^2S	1	S_cS
Ground capacitance (C^g)	1/S	S_c	$1/(S\sqrt{S})$	$S_c/\sqrt{S}$	1/S	S_c	1/S	S_c
C^c/C^g	1	1	S	S	1	S_cS	S	S
RC delay	1	$S_c^2S^2$	1/S to 1	$S_c^2S^2$ to $S_c^2S^2$	1/S	S_c to S_c^2S	1/S to S	S_c^2S to $S_c^2S^2$

Note: S is the scaling factor for the local interconnects and S_c is the scaling factor of the die size.

TABLE 2.2 Scaling Characteristics of Local and Global Interconnects for Four Scaling Scenarios: Ideal, Quasi-Ideal, Constant Resistance, and Constant Thickness

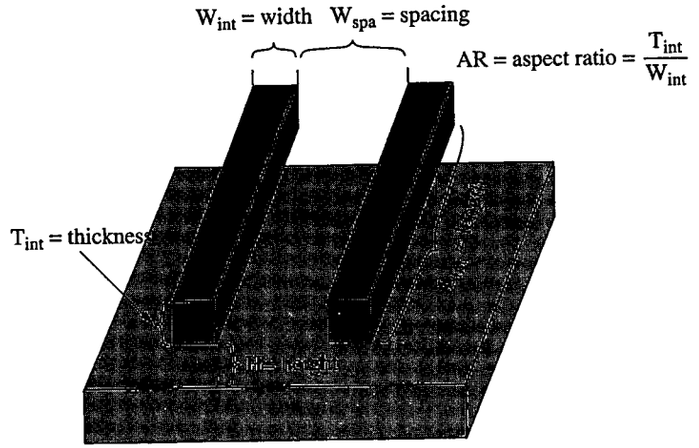


FIGURE 2.21 Physical dimensions of two parallel conductors located on the same metalization layer.

2.4.2 Ideal Scaling

The dimensions of two parallel conductors are illustrated in Fig. 2.21. L_{int}, W_{int}, and T_{int} refer, respectively, to the length, width, and thickness of a conductor. W_{sp} refers to the spacing between two parallel conductors located on the same metalization layer, and H is the distance between the conductor and the substrate or, equivalently, the thickness of the dielectric between two adjacent metal layers. The aspect ratio of an interconnect is

$$AR = \frac{T_{int}}{W_{int}}. \tag{2.37}$$

Under the ideal scaling scheme, each dimension of an interconnect scales by the same scaling factor S ($S > 1$) except for the length of the global interconnects [73,74]. Specifically, the length scaling factor is dependent upon the type of interconnect, i.e., local versus global. A local interconnect length decreases by S, whereas in a global interconnect, the length increases by the chip scaling factor S_c.

According to ideal scaling, the resistance of a local interconnect increases by S,

$$R'_L = \frac{\rho L_{int}/S}{(W_{int}T_{int})/S^2} = SR_L, \tag{2.38}$$

where R'_L is the scaled resistance and ρ is the resistivity of the material. Note that the cross sectional area A of a conductor is $A = W_{int}T_{int}$.

Alternatively, for a global interconnect, the scaled resistance R'_G is

$$R'_G = \frac{\rho S_c L_{int}}{(W_{int}T_{int})/S^2} = S_c S^2 R_G. \tag{2.39}$$

Thus, for global interconnects, the parasitic resistance increases significantly under the ideal scaling scheme. The high interconnect resistance increases the power loss and delay, while degrading signal quality. Due to these critical drawbacks, practical scaling differs from ideal scaling, as described in Section 2.4.3.

The scaling characteristics of the interconnect capacitance are considered in terms of two capacitive components: (1) coupling capacitance, and (2) ground capacitance. Note that a more detailed description of these capacitive components is provided in Chapter 3. Only the scaling characteristics of these capacitances are discussed in this section. As depicted in Fig. 2.22, the coupling capacitance refers to the lateral sidewall capacitance between two interconnects located on the same metal layer. Alternatively, the ground capacitance refers to the parallel plate and fringing capacitance between the interconnect and the substrate.

Assuming ideal scaling and a local interconnect, the scaled coupling C'^c_L and ground capacitances C'^g_L are, respectively,

$$C'^c_L = \frac{\epsilon(L_{int}T_{int})/S^2}{W_{spa}/S} = C^c_L\frac{1}{S}, \tag{2.40}$$

$$C'^g_L = \frac{\epsilon(L_{int}W_{int})/S^2}{H/S} = C^g_L\frac{1}{S}, \tag{2.41}$$

where ϵ is the permittivity of the interlayer dielectric. Thus, both coupling and ground capacitances scale with $1/S$ for local interconnects. Alternatively, for a global interconnect, the scaled coupling C'^c_G and

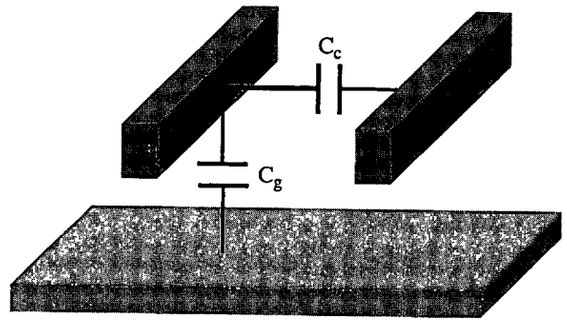


FIGURE 2.22 Sidewall coupling capacitance C_c and ground capacitance C_g between the interconnect and substrate.

ground capacitances $C_G^{'g}$ are, respectively,

$$C_G^{'c} = \frac{\epsilon S_c L_{int} T_{int}/S}{W_{spa}/S} = S_c C_G^c, \tag{2.42}$$

$$C_G^{'g} = \frac{\epsilon S_c L_{int} W_{int}/S}{H/S} = S_c C_G^g. \tag{2.43}$$

As opposed to the local interconnect, the capacitance of a global interconnect increases by the chip scaling factor S_c. For both local and global interconnects, the ratio of the coupling capacitance to the ground capacitance C_c/C_g remains the same with technology scaling.

Finally, the scaled RC delay for local and global interconnects are, respectively,

$$t'_{dL} = SR_L \frac{1}{S} C_L = t_{dL}, \tag{2.44}$$

$$t'_{dG} = S_c S_G^R 2 S_c C_L = S_c S^2 R_G S_c C_L. \tag{2.45}$$

Despite the constant delay of the local interconnects with technology, the situation is significantly different for the global interconnects, where the delay increases rapidly (with a scaling factor of $S_c^2 S^2$). This issue is the primary reason why a practical scaling scheme differs from ideal scaling, as described in the following section.

2.4.3 More Realistic Scaling Scenarios

Three more realistic scaling scenarios are considered in this section. These scenarios represent a more practical interconnect scaling framework since under ideal scaling, the resistance of the global interconnects rapidly increases [74]. These three more realistic scaling scenarios are depicted in Fig. 2.23, and the scaling characteristics for each case are described in the following subsections. These results are also listed in Table 2.2.

Quasi-Ideal Scaling

According to the first scenario, referred to as quasi-ideal scaling, the vertical dimensions do not scale by the same scaling factor as the lateral dimensions. Specifically, H and T_{int} scale by $1/\sqrt{S}$ rather than $1/S$. Consequently, the scaled resistance of the local and global interconnects is, respectively,

$$R'_L = \frac{\rho L_{int}/S}{(W_{int} T_{int})/(S\sqrt{S})} = \sqrt{S} R_L, \tag{2.46}$$

$$R'_G = \frac{\rho S_c L_{int}}{(W_{int} T_{int})/(S\sqrt{S})} = S_c S \sqrt{S} R_G. \tag{2.47}$$

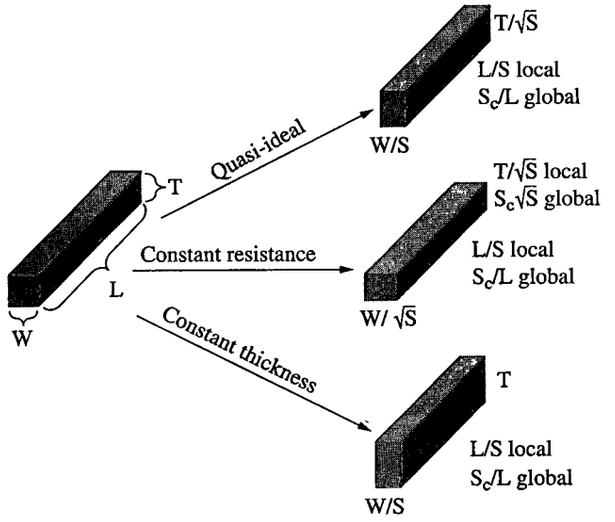


FIGURE 2.23 More realistic scaling scenarios.

As compared to ideal scaling, the pace of increase in the interconnect resistance is slowed by $\sqrt{S}$. Despite the decrease in the resistive scaling factor, the resistance of the global interconnects rapidly increases.

The scaled coupling capacitance $C_L^{'c}$ and ground capacitances $C_L^{'g}$ of a local interconnect are, respectively,

$$C_L^{'c} = \frac{\epsilon (L_{int} T_{int})/(S\sqrt{S})}{W_{spa}/S} = \frac{1}{\sqrt{S}} C_L^c, \tag{2.48}$$

$$C_L^{'g} = \frac{\epsilon (L_{int} W_{int})/S^2}{H/\sqrt{S}} = \frac{1}{S\sqrt{S}} C_L^g. \tag{2.49}$$

Alternatively, for a global interconnect, the scaled coupling capacitance $C_G^{'c}$ and ground capacitances $C_G^{'g}$ are, respectively,

$$C_G^{'c} = \frac{\epsilon S_c L_{int} T_{int}/\sqrt{S}}{W_{spa}/S} = S_c \sqrt{S} C_G^c, \tag{2.50}$$

$$C_G^{'g} = \frac{\epsilon S_c L_{int} W_{int}/S}{H/\sqrt{S}} = \frac{S_c}{\sqrt{S}} C_G^g. \tag{2.51}$$

Note that quasi-ideal scaling lowers the ground capacitance since the vertical distance between the interconnect and substrate is scaled by $\sqrt{S}$ rather than S. The coupling capacitance is, however, degraded since the thickness of the interconnect scales more slowly. Thus, the ratio of the coupling capacitance to the ground capacitance increases

by S. The improvement in the interconnect resistance is therefore achieved at the expense of a degradation in the coupling capacitance, which directly affects signal integrity and delay uncertainty (see Chapter 5).

Finally, the scaled RC delay for a local and global interconnect is, respectively,

$$t'_{dL} = \sqrt{S}R_L\frac{1}{S\sqrt{S}}C_L \text{ to } \sqrt{S}R_L\frac{1}{\sqrt{S}}C_L = \frac{1}{S}t_{dL} \text{ to } t_{dL}, \tag{2.52}$$

$$\begin{aligned} t'_{dG} &= S_c S\sqrt{S}R_G(S_c/\sqrt{S})C_L \text{ to } S_c S\sqrt{S}R_G S_c\sqrt{S}C_L \\ &= S_c^2 S t_{dG} \text{ to } S_c^2 S^2 t_{dG}. \end{aligned} \tag{2.53}$$

Note that the lower and upper bounds of the scaled RC delay are determined, respectively, by the ground and coupling capacitance. This issue is further discussed in Section 2.4.4.

Scaling Based on Constant Resistance

In the second scheme, referred to as constant resistance scaling, the interconnect dimensions are scaled to ensure that the local and global interconnect resistances remain the same. To achieve this condition for the local interconnects, not only the vertical dimensions such as H and T_{int} but also the width scales by $1/\sqrt{S}$. Alternatively, for the global interconnects, the thickness should be *increased* by a factor of $S_c\sqrt{S}$ to maintain a constant resistance.

Under these conditions, the scaled coupling capacitance C'^c_L and ground capacitances C'^g_L of a local interconnect are, respectively,

$$C'^c_L = \frac{\epsilon(L_{int}T_{int})/(S\sqrt{S})}{W_{spa}/\sqrt{S}} = C^c_L\frac{1}{S}, \tag{2.54}$$

$$C'^g_L = \frac{\epsilon(L_{int}W_{int})/(S\sqrt{S})}{H/\sqrt{S}} = C^g_L\frac{1}{S}. \tag{2.55}$$

Alternatively, for a global interconnect, the scaled coupling capacitance C'^c_G and ground capacitances C'^g_G are, respectively,

$$C'^c_G = \frac{\epsilon S_c L_{int} S_c\sqrt{S}T_{int}}{W_{spa}/\sqrt{S}} = S_c^2 S C^c_G, \tag{2.56}$$

$$C'^g_G = \frac{\epsilon S_c L_{int} W_{int}/\sqrt{S}}{H/\sqrt{S}} = S_c C^g_G. \tag{2.57}$$

Thus, the ratio of the coupling capacitance to the ground capacitance increases by $S_c S$.

Under the constant resistance scheme, the scaled RC delay for a local and global interconnect is, respectively,

$$t'_{dL} = R_L\frac{1}{S}C_L = \frac{1}{S}t_{dL}, \tag{2.58}$$

$$t'_{dG} = R_G S_c C_L \text{ to } R_G S_c^2 S C_L = S_c t_{dG} \text{ to } S_c^2 S t_{dG}. \tag{2.59}$$

According to (2.58), this scaling scheme works well for local interconnects. The detrimental effects of scaling on the global interconnects, however, continue since the coupling capacitance of a global interconnect increases significantly when the resistance is maintained constant.

Scaling Based on Constant Thickness

Finally, the constant thickness scaling scheme assumes that the thickness of the interconnects remains the same with future technologies. The length of the global interconnects scales by S_c, whereas the remaining interconnect parameters scale by $1/S$. The resistance of the local and global interconnects scales, respectively, by one and $S_c S$. Thus, the scaled resistance of the local and global interconnects is, respectively,

$$R'_L = \frac{\rho L_{int}/S}{(W_{int}T_{int})/S} = R_L, \tag{2.60}$$

$$R'_G = \frac{\rho S_c L_{int}}{(W_{int}T_{int})/S} = S_c S R_G. \tag{2.61}$$

Under the constant thickness scheme, the scaled coupling capacitance C'^c_L and ground capacitances C'^g_L of a local interconnect are, respectively,

$$C'^c_L = \frac{\epsilon(L_{int}T_{int})/S}{W_{spa}/S} = C^c_L, \tag{2.62}$$

$$C'^g_L = \frac{\epsilon(L_{int}W_{int})/S^2}{H/S} = \frac{1}{S}C^g_L. \tag{2.63}$$

Alternatively, for a global interconnect, the scaled coupling capacitance C'^c_G and ground capacitances C'^g_G are, respectively,

$$C'^c_G = \frac{\epsilon S_c L_{int}T_{int}}{W_{spa}/S} = S_c S C^c_G, \tag{2.64}$$

$$C'^g_G = \frac{\epsilon S_c L_{int}W_{int}/S}{H/S} = S_c C^g_G. \tag{2.65}$$

Thus, the ratio of the coupling capacitance to the ground capacitance increases by S for both the local and global interconnects.

Finally, the scaled RC delay for a local and global interconnect is, respectively,

$$t'_{dL} = R_L \frac{1}{S} C_L \text{ to } R_L C_L = \frac{1}{S} t_{dL} \text{ to } t_{dL}, \tag{2.66}$$

$$t'_{dG} = S S_c R_G S_c C_L \text{ to } S S_c R_G S_c S C_L = S_c^2 S t_{dG} \text{ to } S_c^2 S^2 t_{dG}. \tag{2.67}$$

Despite maintaining a constant thickness, the RC delay grows rapidly for global interconnects due to the growing IC dimensions and reduction in the interconnect width, spacing, and height.

2.4.4 Comparison between Interconnect Scaling Scenarios

The primary purpose of these more realistic scaling scenarios is to alleviate the rapid increase in the global interconnect delay. The motivation is to slow the scaling of the vertical dimensions. As summarized in Table 2.2, however, the scaling factor for the RC delay is only marginally decreased as compared to the ideal scaling scheme. This behavior is due to the increasing coupling capacitance when the interconnect thickness does not scale in proportion to the interconnect length and width. Thus, for both the quasi-ideal and constant thickness scaling scenarios, the upper bound on the global interconnect delay scaling factor is the same as for ideal scaling, as listed in Table 2.2. This upper bound is valid when the coupling capacitance dominates the ground capacitance. If this condition is satisfied, neither the quasi-ideal nor constant thickness scaling scenario will reduce the global interconnect delay.

As listed in Table 2.2, the only scaling scenario where the upper bound on the global interconnect delay scaling factor is reduced is the constant resistance scheme. In this case, the scaling factor is reduced from $S_c^2 S^2$ to $S_c^2 S$. This reduction is achieved when the interconnect thickness is *increased* to maintain a constant resistance. To control the coupling capacitance and ground capacitance, the spacing and height are only scaled by $1/\sqrt{S}$. Although the increase in global interconnect delay is somewhat alleviated in this scaling scheme, the overall integration density suffers since the interconnect thickness is increased and the width and height are only marginally decreased. Note that increasing the cross sectional dimensions of an interconnect is referred to as reverse scaling and is typically applied to the semi-global and global lines [75]. Also note that the aspect ratio of the global interconnects increases in all cases other than ideal scaling, as listed in Table 2.2. The implications of this increase on the interconnect capacitance and signal integrity are discussed in Chapter 5.

2.5 Interconnect Enhancements

Until the late 1990s, aluminum had been the primary interconnect material due to the low diffusion coefficient of aluminum in silicon. A low diffusion coefficient provides enhanced adhesion characteristics to native silicon insulators such as silicon nitride. Despite this important advantage, two primary drawbacks of the aluminum interconnect prevented the use of aluminum in CMOS technologies below the 0.25 μm technology node [76]:

- Higher resistivity
- Degraded electromigration characteristics

The first limitation is the relatively high resistivity of pure aluminum. In practice, the effective resistivity is higher than pure resistivity since an alloy of copper and silicon, and thin titanium layers are utilized in aluminum interconnections to enhance reliability [71,76]. The second limitation is the degraded electromigration characteristics. High current densities cause momentum transfer between electrons and metal atoms. Due to the influence of high electric fields, the metal atoms of the interconnect are transported, causing open or short circuit failures. Electromigration is one of the most significant reliability issues for on-chip interconnects [77]. Due to the low activation energy, aluminum based interconnects are more sensitive to electromigration [78].

Starting in the late 1990s, copper interconnect was adopted to alleviate these challenges related to the high resistance of aluminum based interconnect. At room temperature, the resistivity of pure copper is approximately 1.7 times lower than the resistivity of pure aluminum [79]. Despite the one time reduction in resistance, copper is only a temporary solution since the resistance of the interconnect continues to increase with scaling, as described in the previous section. Furthermore, with decreasing interconnect dimensions, mechanisms such as surface and grain boundary scattering have become more important. These nonideal material effects increase the effective resistivity of the interconnect, as described in Chapter 3.

As discussed in the previous section, different scaling scenarios have been proposed to slow the increase in the interconnect resistance. These scaling schemes typically rely on adopting a smaller scaling factor for the vertical dimensions or maintaining constant vertical dimensions constant with technology. Unfortunately, these schemes increase the lateral fringing capacitance. The overall performance of a global interconnect therefore degrades with technology. Several techniques at different levels of abstraction have been proposed to alleviate challenges related to the interconnect. Utilizing an ultra-low dielectric permittivity material to achieve interlayer insulation is described in Section 2.5.1. Three-dimensional integration to achieve a shorter and

smaller number of global interconnects is discussed in Section 2.5.2. Utilizing on-chip optical interconnects for global communication is summarized in Section 2.5.3. Finally, carbon-based on-chip interconnects are described in Section 2.5.4.

2.5.1 Ultra-Low-K Dielectric Material

Isolation among adjacent interconnect lines is achieved with a dielectric material. As described in Chapter 3, this dielectric material produces several capacitive components among the metal lines. These parasitic interconnect capacitances not only increase signal delay, but also degrade signal integrity due to capacitive crosstalk. An insulator with a low dielectric constant (low-K) is desirable since the capacitance is directly proportional to this constant. A low-K dielectric material has therefore three important advantages:

- Reduction in interconnect parasitic capacitance
- Reduction in dynamic power consumption since the interconnect portion of the overall switching capacitance is lower
- Indirect reduction in interconnect parasitic resistance since a lower dielectric constant enables a greater interconnect thickness at a constant lateral capacitance

These benefits are illustrated in Fig. 2.24, where aluminum and silicon dioxide are compared with copper and low-K dielectric material in terms of the number of required metal layers, relative power consumption, and relative delay [76]. Despite these advantages, other reliability issues exist to integrate low-K dielectric materials into the interconnect fabrication process. These issues include mechanical stability,

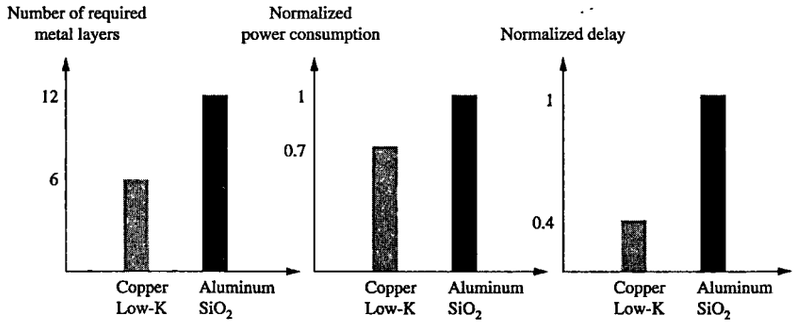


FIGURE 2.24 Aluminum and silicon dioxide are compared with copper and low-K dielectric material in terms of the number of required metal layers, relative power consumption, and relative delay [76].

the ability to withstand polishing (planarization capability) and thermal cycling processes, and the ability to adhere to the deposited metal [71].

Carbon Doped Silicon

Prior to the acceptance of copper as the primary interconnect material, oxide/nitride was utilized as the interconnect insulator, where the relative dielectric permittivity is approximately four [80]. A low-K dielectric refers to insulators with a relative permittivity in the range of three. These low-K dielectric materials such as carbon and hydrogen doped silicon, also referred to as organic silicon or organosilicate glass (SiCOH), were first used in a 90 nm technology node using conventional plasma enhanced chemical vapor deposition (PE-CVD) techniques [64]. The silicon dioxide is doped with methyl ($-CH_3$) to reduce the dielectric permittivity.

Porous Dielectric Film

Ultra-low-K dielectric material typically refers to those insulators with a relative dielectric permittivity in the range of 2.5 and smaller. An example is a porous dielectric film that achieves a relative permittivity of approximately two. A porous film relies on producing voids (empty spaces) within the dielectric material to reduce permittivity, which is dependent upon the pore size and distribution [81]. These parameters also affect the mechanical stability of the dielectric material. An ultra-low-K porous film typically suffers from poor mechanical characteristics. This issue is partly alleviated by utilizing a hybrid integration where the dense lower metal layers utilize an ultra-low-K porous material, whereas a more conventional low-K material is used as an insulator between the upper metal layers [64].

Spin-On Dielectric

Another technique to form dielectric material is spin-on deposition as opposed to PE-CVD [82]. A spin-on process typically exhibits enhanced planarization characteristics [83]. Polymer based organic materials are typically used as the dielectric [64]. Porosity can be introduced to a spin-on dielectric, further lowering the permittivity. The achieved permittivity level is typically in the same range as those for porous films that utilize PE-CVD. Unfortunately, spin-on–based dielectric materials also suffer from mechanical and thermal stability.

Air-Gap–Based Dielectric

Finally, an interesting approach to obtain an ultra-low-K material is air gaps. Note that air has a relative dielectric permittivity of approximately one. Thus, an ultimate low-K insulator can be produced by replacing porous material with air. An air gap based dielectric was first used as an intra-metal insulator to isolate metals located at

different layers [84]. A temporary, also referred to as sacrificial, dielectric material is first deposited. In the second step, this dielectric is removed either by thermal decomposition or wet etching, while permitting air to diffuse to form an air cavity [84]. Another technique is to utilize the nonideal step coverage characteristics of PE-CVD to introduce and control air filled voids [64,84]. Utilizing an air-gap–based dielectric as an inter-metal insulator has also been proposed, achieving further reductions in the parasitic interconnect capacitance [84].

2.5.2 Three-Dimensional Integration

In the past decade, three-dimensional (3-D) integration has emerged as a promising technology that maintains the benefits of miniaturization by utilizing the vertical dimension rather than decreasing the size of the devices in two dimensions [85–89]. The first demonstration of vertically integrated devices dates back to the early 1980s, where the NMOS and PMOS transistors with a CMOS inverter shared a common gate, as depicted in Fig. 2.25 [90]. With the increasing dominance of interconnects in determining system performance, 3-D integration started to receive considerable attention in the early 2000s. The advantages of 3-D integration technology, several existing fabrication techniques, and related challenges are discussed in the following subsections.

Advantages

In a 3-D integrated system, multiple active silicon and/or nonsilicon layers are vertically stacked, thereby enabling heterogeneity and higher integration density, as shown in Fig. 2.26 [85]. Utilizing the vertical dimension in a monolithic fashion not only increases the integration density, but also reduces the length and number of global interconnects. Assuming a first order approximation, the reduction in

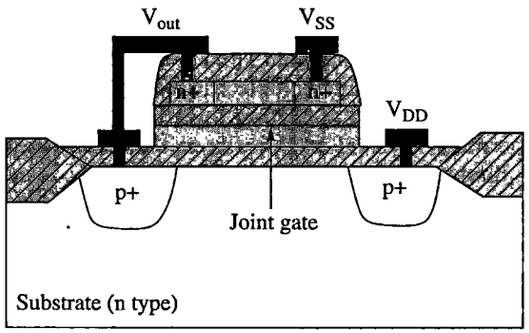


FIGURE 2.25 Vertically integrated inverter, also referred to as joint metal oxide semiconductor (JMOS), where the NMOS and PMOS transistors share a common gate [90].

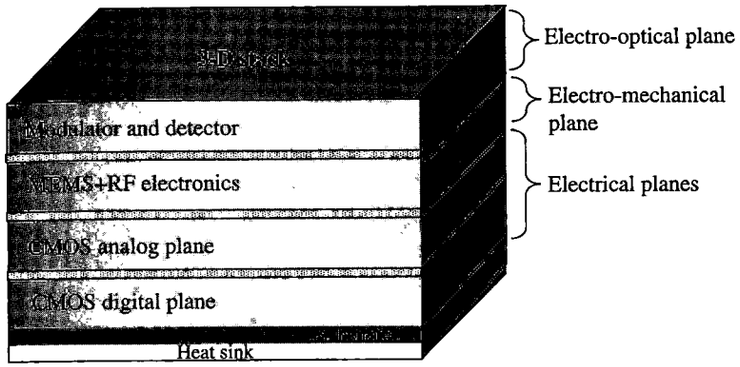


FIGURE 2.26 3-D integrated system where multiple active silicon and nonsilicon layers are vertically stacked, thereby enabling heterogeneity and a higher integration density.

the length of the longest interconnect is proportional to $\sqrt{n}$, where n is the number of planes within a 3-D stack [85]. The reduction in the length and number of global interconnects enhances system performance (due to reduced RC delay) while lowering the power consumption (due to reduced interconnect capacitance). The interconnect bottleneck of conventional planar technologies is thereby alleviated [85].

In addition to advantages in speed and power consumption, 3-D integration provides unique opportunities for heterogeneous integration. The ability to merge disparate materials and technologies is essential for multifunctional systems since each layer can be optimized according to the requirements of that particular function. This advantage expands the application domain of 3-D integrated circuits from high performance computing to embedded systems consisting of, for example, sensors, analog interface circuit, digital processing blocks, and RF wireless transmission circuitry. Additional advantages of 3-D integration include a smaller form factor and potentially higher yield and lower cost. Note that these advantages largely apply to a wafer-level 3-D integration methodology. Different approaches to fabricate 3-D circuits are summarized in the following subsection.

Approaches to 3-D Integration

Existing efforts to produce 3-D integrated circuits can be broadly classified under three primary categories [91]:

- Transistor-level 3-D integration
- System-in-package (SiP) and system-on-package (SoP) 3-D integration
- Wafer-level through silicon via (TSV) 3-D integration

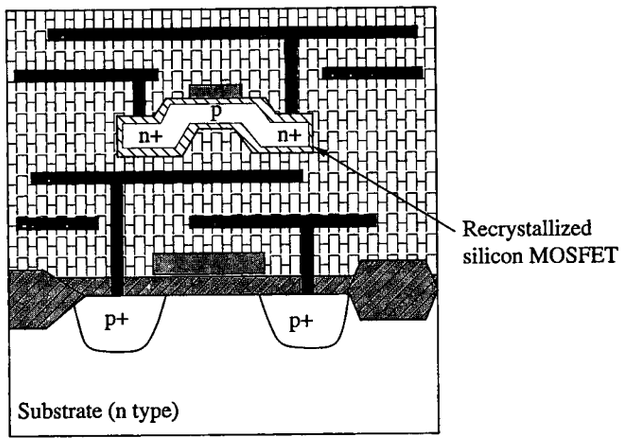


FIGURE 2.27 Transistor-level 3-D integration where the active devices within a single logic gate are fabricated on different layers.

Transistor-Level 3-D Integration

In transistor-level 3-D integration, the active devices within a single logic gate are fabricated on different layers, as illustrated in Fig. 2.27. The transistors located within the first device layer are fabricated by conventional processes. Alternatively, the formation of the devices located within the upper layers is the primary challenge for transistor-level 3-D integration. Different fabrication techniques are required to form the upper layer devices. These techniques are typically based on recrystallization of the silicon [85].

In each of these techniques, after the formation of the first layer transistors using conventional approaches, an isolated amorphous silicon film is deposited. A catalyst (such as germanium or nickel) and laser heating or rapid thermal annealing are used to recrystallize the silicon, on which the upper layer devices are fabricated. Thus, in this technique, the upper layer transistors are formed within the on-chip interconnect. Depending upon the order of fabrication, the interplane interconnects can either be a doped semiconductor resistant to high processing temperatures or a lower resistivity metal if the interconnects are fabricated after the formation of the upper layer devices. In another approach, the upper layer transistors are formed on a polysilicon film produced by laser heating or rapid thermal annealing of the amorphous silicon. In this case, doped polysilicon or tungsten is used as the vertical interconnect since the processing temperature is relatively high.

System-In-Package 3-D Integration

In SiP 3-D integration, multiple bare or packaged dies are assembled, where the interconnections among the dies are achieved by one of the following methods:

- Wire bonding
- Low density vertical interconnects and solder balls typically located along the periphery of the package
- Area array based C4 (controlled collapse chip connect) bumps located on the side surface of the stacked dies

Wire bonding, illustrated in Fig. 2.28, is a commonly used technique in SiP 3-D integration due to low cost and shorter time to manufacture. The reliability and packaging efficiency are determined by the thickness of the dies (that depends on the wafer thinning capability), spacers between the dies, adhesive materials, and size of the bumps that connect the bottom most die to the package substrate [85]. The length of the bonding wires, also referred to as loop overhang, is typically in the range of millimeters and plays an important role in the overall SiP performance. The parasitic impedances of the package are largely determined by these bonding wires.

An alternative to wire bonding is vertical interconnects and solder balls along the periphery of the package. In this technique, solder balls and through hole vias provide communication between the stacked dies. An illustration is shown in Fig. 2.29. Since the parasitic impedance of the vertical interconnects is significantly reduced as compared to wire bonding, more dies can be stacked. Despite the reduction in parasitic impedances, this type of vertical interconnect cannot satisfy the interconnection density required by high performance systems.

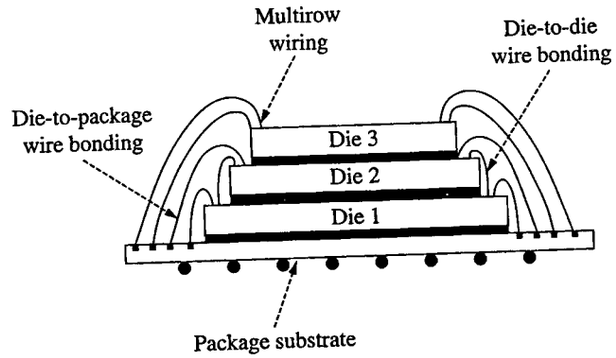


FIGURE 2.28 System-in-package (SiP) 3-D integration utilizing wire bonding.

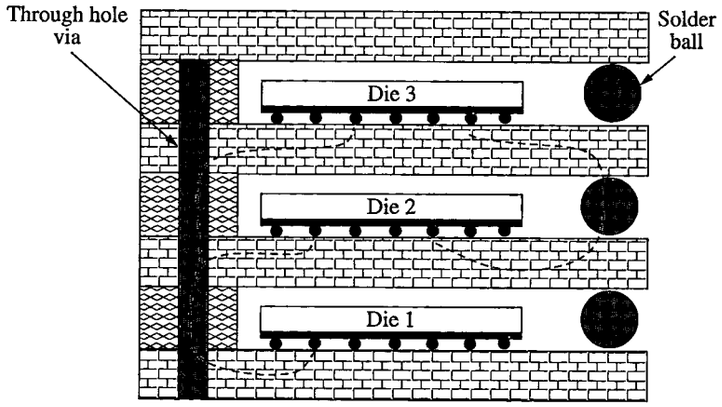


FIGURE 2.29 System-in-package (SiP) 3-D integration utilizing low density vertical interconnects and solder balls located along the periphery of the package.

The last example of an SiP 3-D system utilizes area array based C4 bumps. The I/O pads of a die are connected to a metal that laterally carries the signal to a side surface, where another metal vertically carries the signal to other dies, as depicted in Fig. 2.30 [92]. The vertical metal is also connected to a solder bump. Thus, each die within a 3-D system shares a common area bump array and the vertical interconnects

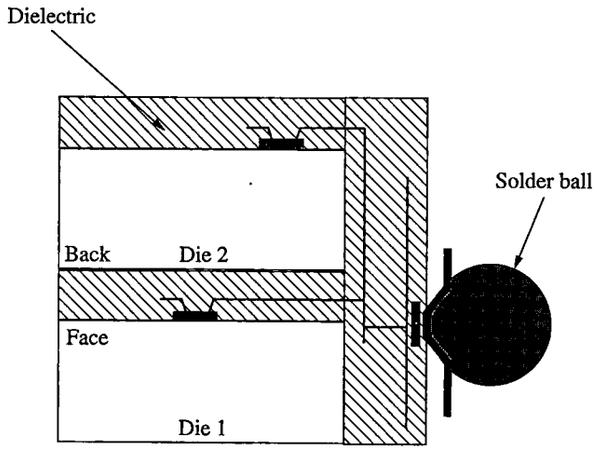


FIGURE 2.30 System-in-package (SiP) 3-D integration utilizing area array based C4 (controlled collapse chip connect) bumps located on the side surface of stacked dies [92].

do not pass through the substrate of the planes. These bumps are attached to a ceramic substrate. Although the interconnect density of an area array based SiP is higher than a peripheral vertical interconnect based SiP, the relatively large size of the solder bumps constitutes a limitation in achieving ultra-high interconnect density.

Wafer-Level TSV 3-D Integration

Finally, in wafer-level 3-D integration technology, multiple wafers are thinned, aligned, and bonded. As depicted in Fig. 2.31, communication among the dies is achieved by high density through silicon vias (TSVs) that pass through the substrate of the planes [85]. A three plane TSV 3-D circuit has recently been demonstrated to operate well over a gigahertz [86].

Multiple TSV fabrication techniques exist. If the TSVs are formed before bonding, these TSVs are referred to as via-first TSVs. Alternatively, via-last TSVs are formed after the bonding is complete. Note that, as opposed to transistor-level 3-D integration, each wafer is individually and separately processed, decreasing the overall manufacturing time and potentially increasing yield. Since the TSV dimensions have decreased, the communication bandwidth among the dies is significantly higher than other 3-D integration technologies. The greatest reduction in wire length is therefore achieved by TSV wafer-level 3-D integration technology.

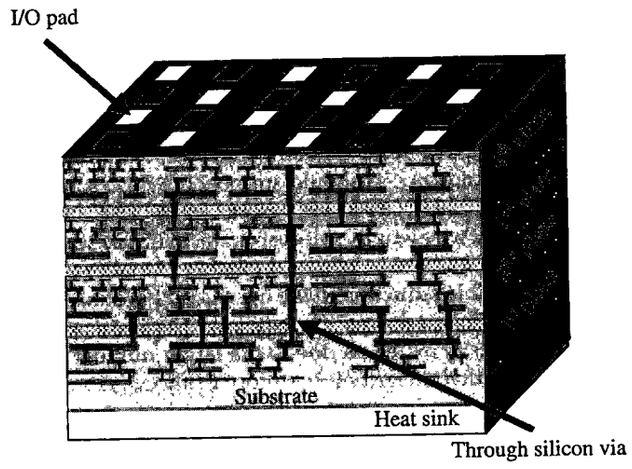


FIGURE 2.31 Wafer-level 3-D integration technology where communication among the dies is achieved by the high density through silicon vias (TSVs) that pass through the substrate of the planes [85].

Challenges

As mentioned in the previous subsection, TSV wafer-level 3-D integration technology is the most promising option to alleviate the interconnect bottleneck. Several challenges, however, exist for TSV 3-D ICs to evolve into a mainstream technology. These challenges are classified into the following categories [85,91]:

- Manufacturing challenges
- Testing challenges
- Thermal challenges
- Circuit and computer-aided design challenges

Wafer bonding is a critical manufacturing step that determines the integrity of a 3-D system. The bonding process should not affect the performance of the individual planes. Furthermore, the planes should remain bonded during the lifetime of the system. Other manufacturing steps such as wafer alignment, wafer thinning, and TSV formation determine the length and density of the TSVs, thereby affecting the reduction in wire length. Overall system performance is therefore directly related to these manufacturing steps.

Testing a 3-D system with multiple stacked dies is a challenging task. A typical 3-D test flow consists of a pre-bond test, also referred to as a known good die (KGD) test; post-bond test, also referred to as known good stack (KGS) test; and the final test after assembly and packaging is complete [93]. To achieve a pre-bond test, dedicated input/output (I/O) pads are utilized to achieving wafer probing. Methods should be developed to reduce the number of these test pads since additional area is consumed. The TSVs are also tested in the pre-bond phase to enhance yield. The post-bond test is a critical step to achieving high yield. Once a fabrication fault is detected during the stacking process, the remaining dies are not stacked, thereby saving the working dies for a different stack. Also note that design-for-testability techniques [94] should support the post-bond test. For example, if a functional circuit block is partitioned into multiple planes to reduce the global wire length, self-test circuity such as scan registers should detect this characteristic to evaluate functionality.

Despite a possible reduction in the overall power consumption of a 3-D system as compared to a 2-D system with equivalent functionality, the *power density* increases due to the larger number of devices per unit area. Thermal integrity is therefore a primary concern in 3-D ICs. The temperature of those planes located farther from the heat sink increases due to the higher thermal resistance, thereby degrading performance and reliability. Efficient temperature-aware 3-D floorplanning techniques, packaging solutions, and more effective heat sinks should therefore be developed.

Finally, design methodologies and CAD algorithms are required to efficiently verify a hybrid 3-D system. These design methodologies should support the integration of heterogeneous technologies and functionalities. Furthermore, these tools should effectively exploit the benefits provided by 3-D integration technology, while producing a robust system with sufficient signal, power, and thermal integrity. An important challenge is the computational complexity of these tools since significantly more computational resources are needed to design and analyze these 3-D systems as compared to a 2-D system.

2.5.3 On-Chip Optical Interconnect

The concept of on-chip optical interconnect was first proposed in 1984 [95]. Critical electrical interconnects are replaced with optical links, thereby significantly improving the speed and power characteristics. An optical link, however, requires electrical-to-optical and optical-to-electrical conversion that consumes power and introduces latency. On-chip optical interconnects are therefore considered only for global nets such as data buses and clock distribution networks.

Introducing optical interconnects into high complexity integrated systems requires compatibility with CMOS technology. This requirement limits the available materials and processes to fabricate optical links. An important issue in the development of on-chip optical interconnects is the absence of a monolithic, silicon based laser. Thus, on-chip optical interconnect technologies typically rely on an external laser.

As illustrated in Fig. 2.32, a typical on-chip optical interconnect data path consists of the following elements [34]:

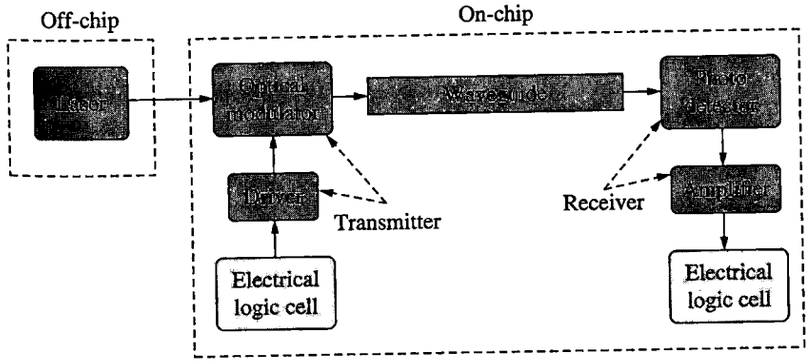


FIGURE 2.32 Optical interconnect data path consisting of an off-chip laser, on-chip transmitter, receiver, and waveguide [34].

- A transmitter that includes an electro-optical modulator and driver circuit
- An optical waveguide
- A receiver consisting of a photodetector and amplifier

The design of a fast, silicon compatible electro-optical modulator is a challenging task. The primary principle behind electro-optical modulation is the change in certain optical characteristics of the medium, such as the refractive index, by electrical signals. Utilizing this change, the optical signals are modulated in amplitude or phase. Pure silicon, however, does not exhibit suitable mechanisms to vary the refractive index. One of the candidate mechanisms is the free carrier plasma dispersion effect. An example device based on a Mach-Zehnder interferometer has been demonstrated [96]. This device is, however, 10 mm long, and requires a large capacitance and therefore higher delay and power consumption for the driver circuitry. A possible structure that can significantly reduce the size is an optical microcavity that produces less than 10 pF of capacitance for the driver. The driver consists of a series of tapered inverters, where the optimum number of stages is primarily determined by the modulator capacitance.

Optical waveguides have the advantage of high signal propagation speed as compared to electrical propagation since an optical waveguide does not exhibit *RLC* impedances. The superior delay characteristic of an optical waveguide is valid regardless of the waveguide material [97]. A low signal dispersion characteristic, however, is desirable to further reduce the delay of an optical waveguide. An important limitation of optical waveguides is a relatively high area requirement. For example, the minimum pitch of a practical integrated waveguide varies between 10 and 20 μm [98]. Multiplexing the optical signals, also referred to as wavelength division multiplexing (WDM), has been proposed to alleviate this limitation [99]. In WDM, an optical link replaces multiple electrical wires, thereby reducing the overall area. Significant challenges, however, exist to implement on-chip WDM since each WDM requires fast transmitters and receivers in addition to low area [97].

On the receiver side, a photodetector converts light back into the electrical domain. The output current of the photodetector is first amplified by a transimpedance amplifier (TIA). Additional amplifying stages strengthen the analog signal to a digital voltage level. Note that a photodetector requires high optical absorption. Alternatively, the modulator of the transmitter requires negligible optical absorption. Since the modulator and detector of the same optical link have to operate at the same wavelength, light with a 1.5 μm wavelength and a silicon germanium or germanium based detector are typically used [34, 97]. Similar to a transmitter, the size, power consumption, and response time are important characteristics for a receiver within an on-chip optical link.

Another approach to developing on-chip optical communication relies on free space propagation of optical signals [100, 101]. Free space propagation eliminates the need for bulky waveguides by utilizing micro-lenses and mirrors. A dedicated optical interposer has also been proposed [100]. In this hybrid system, a separate interposer IC consisting of highly dense vertical-cavity surface-emitting lasers (VCSELs), electro-optical modulators, and optical detectors is bonded to a silicon IC consisting of CMOS circuitry [100]. Note that 3-D integration technology is highly compatible with this approach where an optical plane optimized for the lasers, modulators, and detectors can be stacked near an electrical plane optimized for CMOS technology [101].

2.5.4 Carbon Based On-Chip Interconnect

A carbon based on-chip interconnect is a relatively radical approach to alleviating the communication bottleneck. The use of lower resistivity copper and low-K dielectrics for reduced capacitance are one time solutions that can only temporarily alleviate the interconnect issue. As the interconnect dimensions continue to decrease, the parasitic impedances will once again start to dominate, despite the initial reduction in resistivity and permittivity. The use of carbon nanotubes and graphene nanoribbons represent long term solutions to the interconnect bottleneck and are discussed in the following subsections.

Carbon Nanotubes

Single-walled carbon nanotubes (SWCNTs) have been proposed to replace metallic interconnects such as copper [102]. SWCNTs have two primary advantages:

- Enhanced reliability characteristics
- Higher conductivity when fabricated in bundles

SWCNTs are rolled graphite sheets forming a cylinder where the rolling angle, i.e., chirality, and diameter determine whether the nanotube exhibits metallic or semiconducting characteristics [103]. An SWCNT is shown in Fig. 2.33(*a*). Carbon nanotubes are covalently bonded, and are therefore highly resistant to physical breakdown mechanisms. Furthermore, SWCNTs can carry significantly larger current densities as compared to copper without exhibiting electromigration-related issues, which have been experimentally demonstrated [104]. For example, no degradation in device characteristics has been observed at 250°C when a current density of 10^9 A/cm^2

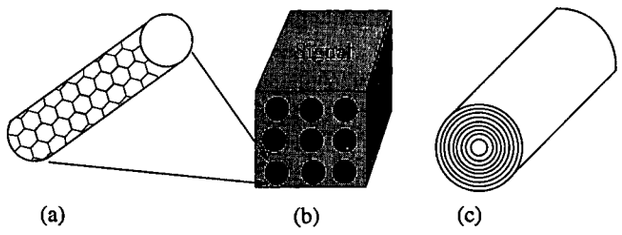


FIGURE 2.33 Carbon nanotube as on-chip interconnect: (a) single-walled carbon nanotube (SWCNT), (b) SWCNT bundle, and (c) multiwalled carbon nanotube (MWCNT).

is applied to carbon nanotubes for approximately two weeks [104]. Note that in a metal interconnect, typical current densities are on the order of 10^5 A/cm^2, which is four orders of magnitude smaller than the current density of an SWCNT. Existing metal interconnects cannot satisfy future current density requirements, as predicted by the ITRS [32].

The second important advantage of carbon nanotubes is a potential increase in electrical conductivity as compared to copper. Despite the high intrinsic resistance (approximately 6.5 kΩ) of an individual nanotube [102], *bundles* (also referred to as *ropes*) of SWCNTs in parallel are utilized to reduce this resistance, as illustrated in Fig. 2.33(*b*). Furthermore, multiple SWCNTs with different diameters are nested to form multiwalled carbon nanotubes (MWCNTs), where the co-centric interior SWCNTs are interconnected with low contact resistance [105], as shown in Fig. 2.33(*c*).

Certain nanotubes within a bundle are metallic, whereas the remaining nanotubes are semiconducting depending upon the chirality and diameter. If no special techniques are utilized, approximately one third of the SWCNTs in a bundle is expected to be metallic. Note that several techniques exist to increase this ratio. Also note that the thermal conductivity of carbon nanotubes is approximately fifteen times higher than copper, thereby exhibiting enhanced heat removal characteristics [106].

Design-centric research in carbon nanotubes has focused on developing *RLC* models for individual SWCNTs, bundled SWCNTs, and MWCNTs [102,103]. These models are utilized to compare the power and delay characteristics of carbon nanotubes with metallic interconnects such as copper optimized for advanced CMOS technologies. This research has demonstrated that for short, local interconnects, SWCNTs achieve up to a 50% reduction in capacitance and therefore dynamic energy per switching operation [107]. Thus, to reduce power consumption, SWCNTs are preferable for short interconnects. Similarly, delay is also reduced with SWCNT bundles for local interconnects.

Alternatively, for longer interconnect lengths (semi-global and global), it is more advantageous to use densely packed SWCNT bundles or MWCNTs with large diameters to reduce the resistance and propagation delay. Another possibility is to utilize small aspect ratios, thereby reducing power dissipation and crosstalk while maintaining constant delay. Note that a critical length exists for MWCNTs beyond which a larger diameter produces higher conductivity. Wider MWCNTs are therefore desirable for longer interconnects with lengths greater than 100 μm [105]. If the length of a MWCNT is shorter than the critical length, a higher diameter produces a lower conductivity [105]. Thus, for interconnects shorter than 10 μm, dense SWCNT bundles achieve a higher conductivity than MWCNTs. For vertical vias shorter than a few micrometers, SWCNT bundles or MWCNTs with small diameters are preferable [105].

Graphene Nanoribbons

Despite the aforementioned potential advantages, carbon nanotubes suffer from limited manufacturing maturity. Only small scale fabrication of SWNTs is currently available [108]. Furthermore, available synthesis techniques typically produce a relatively high impurity concentration. These impurities cause the conductivities to deviate from theoretical conductivities.

Graphene nanoribbon (GNR) interconnects have been proposed to alleviate some of these fabrication related issues. GNRs are obtained by patterning a sheet of graphene [109]. Thus, a GNR is an unrolled SWCNT. The fabrication process of a GNR is more controllable as compared to a carbon nanotube due to the planar characteristic of graphene. Conventional lithography techniques are utilized to pattern the graphene. This characteristic makes GNR a potential candidate for the horizontal interconnects [110]. The edge structure of a GNR plays an important role in determining the conductivity. Among the two possible edge structures, armchair GNRs can be either metallic or semiconducting, whereas zigzag GNRs are always metallic [110].

Despite the advantage of a relatively controlled fabrication process, GNRs have several important drawbacks. For example, GNRs exhibit edge scattering, which reduces the electron mean free path. Furthermore, the conductivity of a multilayer GNR is reduced due to intersheet electron hopping [111]. Fabrication procedures require further investigation since the substrate should satisfy certain conditions before the GNR can be deposited. Design-centric research has also been performed for GNRs to determine the conditions under which GNRs are superior to copper and carbon nanotubes [109,111]. According to this research, specific technology requirements (such as doping with arsenic pentaflouride and generating specular edges) should be achieved for the GNRs to outperform copper and carbon nanotubes when used as on-chip interconnects [109,111].

2.6 Chapter Summary

Transistor and on-chip interconnect scaling characteristics are described in this chapter. Several emerging technologies for both transistors and interconnects are also discussed. A summary of the chapter is provided below:

- With continuous scaling of the active device feature sizes, the effect of interconnects on system delay, power consumption, noise, and reliability has become significant, thereby causing a transition from a logic-centric to an interconnect-driven design process.
- Transistor delay is expected to decrease to approximately 1.45 ps for a 15 nm CMOS technology, whereas the interconnect delay remains effectively fixed at approximately 20 ps/mm.
- The overall speed of current ICs is most often limited by long distance global interconnects even if these on-chip lines are delay-optimized for the optimum width and number of repeaters.
- A large portion of the total transient power is dissipated by the on-chip lines since the interconnect capacitance often dominates the total gate load.
- Increasing the interconnect coupling capacitance degrades signal integrity, producing greater delay uncertainty.
- The on-chip interconnects dedicated to the power and ground lines suffer from power supply noise due to high parasitic resistances and inductances.
- The expected benefits of technology scaling (from a device perspective) are: (1) higher packing density and therefore higher yield, (2) higher circuit speed, (3) higher system speed, (4) lower system power consumption, (5) enhanced system reliability, and (6) heterogeneous systems integration.
- Constant voltage and constant electric field scaling are the two primary scaling scenarios.
- Constant voltage scaling was adopted until approximately the 0.8 μm CMOS technology node to reduce delay and avoid multiple supply voltages.
- Due to increasing power consumption and electric fields within a transistor, constant electric field scaling was utilized below the 0.8 μm CMOS technology node.
- Constant electric field scaling continued until approximately the 130/90 nm technology node where any further reduction in the power supply and threshold voltages was limited by the significant increase in leakage current and process variations.
- Since the power supply and threshold voltages have been maintained approximately constant below the 130/90 nm technology nodes, both lateral and vertical electric fields have increased significantly, producing deleterious short channel effects.
- The quality of an MOS transistor in digital ICs is determined by two factors: (1) the ratio of the on-current to the off-current, and (2) the time required for the transistor to switch from the on-state to the off-state.
- Dennard's scaling framework (constant electric field scaling) relies on the following steps: (1) increased current drive capability due to a reduced gate insulator thickness, channel length, and threshold voltage, (2) controlled electric fields and power dissipation due to a reduced power supply voltage, and (3) reduced depletion layer width due to a higher substrate doping concentration.
- The reduction in the depletion layer width is necessary to prevent the depletion layer from penetrating into the transistor channel since the channel length is shorter.
- According to constant electric field scaling, the transistor delay scales approximately by $1/S$ for a long channel transistor and by $1/S^2$ for a short channel transistor, whereas the transistor power consumption scales approximately by $1/S^2$ for a long channel transistor and by $1/S$ for a short channel transistor.
- For both short and long channel transistors, the power-delay product scales by $1/S^3$, demonstrating the fundamental advantage of constant electric field scaling.
- The scaling behavior of the power density is significantly different in short and long channel transistors where, for a long channel transistor, the power density remains constant, whereas in a short channel transistor, the power density increases by S, causing thermal integrity issues.
- In constant voltage scaling, the substrate doping concentration increases by S^2 to reduce the depletion layer width in proportion to the channel length to maintain a constant threshold voltage.
- For both the constant electric field and constant voltage scaling schemes, the subthreshold slope remains constant with technology, causing subthreshold leakage current to become dominant in nanoscale technologies.

- Considering only the transistor delay, constant voltage scaling is more beneficial since the transistor delay decreases by $1/S^2$.
- According to constant voltage scaling, the scaling factor for the power, power-delay product, and power density is, respectively, S, $1/S$, and S^3, demonstrating the fundamental limitations of this scheme.
- As predicted by the power consumption and power density scaling factors, nanoscale CMOS technology is typically power constrained.
- The accuracy of the one-dimensional gradual channel approximation and therefore Shockley's transistor model degrade with smaller dimensions due to short channel effects.
- A transistor is referred to as a short channel device if the channel length is of the same order of magnitude as the depletion layer width.
- Small geometry effects are classified into two categories: (1) modification of the threshold voltage, and (2) changes in the electron drift characteristics.
- Small geometry effects that modify the threshold voltage of a device are (1) V_t roll-off and (2) DIBL, whereas small geometry effects due to varying drift characteristics are (1) velocity saturation, and (2) mobility degradation.
- In short channel devices, V_t roll-off refers to a mechanism where the source and drain regions induce a nonnegligible amount of depletion charge within the channel, thereby reducing the threshold voltage.
- V_t roll-off is alleviated by decreasing the gate insulator thickness and junction depth, thereby producing a more shallow junction.
- Decreasing the gate insulator thickness significantly increases the gate leakage current, whereas decreasing the junction depth increases the junction resistance.
- At high drain voltages, the potential barrier of the channel is affected by the drain-induced depletion region due to a greater depletion layer width, thereby lowering the threshold voltage of a device.
- DIBL is quantified as the difference between the threshold voltage when the drain voltage is 100 mV and the threshold voltage when the drain voltage is equal to the power supply voltage.
- Punchthrough occurs when the drain-induced depletion region is sufficiently large and therefore the drain- and source-induced depletion regions merge, producing an inverted channel independent of the gate voltage.

- Velocity saturation occurs at high lateral electric fields of approximately 10^5 V/cm, where the electron drift velocity saturates due to a greater number of collisions with the optical phonons.
- In a velocity saturated device, the dependence of the current on the gate-to-source voltage is weaker than a quadratic function, and the current is independent of the effective channel length.
- Mobility degradation occurs at high vertical electric fields where the carriers collide at the silicon-insulator interface, thereby reducing the effective surface mobility below the bulk mobility.
- For CMOS technologies smaller than 90 nm, suppression of small geometry effects and structural device enhancements are highly critical since short channel effects negate any performance gains from device scaling.
- Nonuniform channel doping is a typical practice to alleviate the effects of V_t roll-off and DIBL by reducing the source/drain depletion region, while maintaining a constant threshold voltage and carrier mobility.
- In nonuniform channel doping, the substrate doping concentration is increased near the source/drain and channel boundaries, thereby reducing the source- and drain-induced depletion layer widths.
- Strain engineering (changing the energy-band structure and scattering characteristics of a molecule by introducing stress) alleviates mobility degradation in short channel devices.
- Substrate-induced strain (e.g., growing strained silicon on a silicon-germanium layer) exhibits a more global effect, but only enhances the electron mobility, while the performance of the PMOS devices does not improve.
- Process-induced strain is utilized to independently change the mobility of the NMOS and PMOS devices at the expense of additional processing steps such as stress liners, stress memorization, and filling source/drain regions with silicon-germanium.
- The use of a high-permittivity (high-K) dielectric and metal gate has enabled the continuation of technology scaling below the 65 nm node by reducing the gate leakage current and alleviating short channel effects related to high vertical electric fields.

- Electrically, a high-K gate insulator behaves as a thin oxide, where an EOT of 0.9 nm has been demonstrated for the 32 nm CMOS technology node.
- Nonuniform channel doping is not an effective mechanism in reducing the subthreshold current for sub-32 nm CMOS technologies, requiring more radical changes in the device structure.
- Planar and vertical multigate devices have been proposed to enhance the capability of the gate terminal in controlling the transistor channel.
- Planar double gate transistors consist of a front and back gate, both of which control the transistor channel, thereby alleviating short channel effects.
- The primary challenge in building planar double gate transistors is to achieve sufficient alignment of the two gates, an issue that is exacerbated with smaller channel lengths.
- Vertical multigate devices such as FinFETs, which consist of two gates and triple gate FETs, have been proposed where the gate terminal surrounds a larger area of the channel to suppress short channel effects.
- Vertical multigate devices achieve a significantly higher I_{on}/I_{off} ratio as compared to planar technologies.
- Triple gate FETs consist of three gates as opposed to FinFETs where the additional gate further suppresses the short channel effects and relaxes the process requirements, allowing relatively larger fin widths under a constant subthreshold current.
- In vertical multigate devices, only discrete effective widths can be achieved since the fin height is a process-constrained parameter.
- To maintain a continuous improvement in system performance, the interconnect parasitic impedances should decrease in proportion to the transistor resistance and capacitance.
- Smaller on-chip interconnect dimensions and reduced interconnect parasitic impedances are two conflicting requirements.
- The damascene technique has been utilized to achieve a sufficiently planar metal surface, enabling a higher number of metal layers and a smaller metal pitch.
- The continuous reduction in the interconnect cross sectional area and pitch increases the parasitic resistance and capacitance.

- A planar multilevel and hierarchical metallization architecture has been proposed to achieve high density local interconnect while maintaining low parasitic impedances for the global interconnect.
- The chip scaling factor S_c ($S_c > 1$) has been introduced to consider the growing die area since the length of the global interconnects increases with this scaling factor.
- In ideal scaling, each interconnect dimension scales by S ($S > 1$) except for the length of the global interconnects, which increases by S_c.
- Under ideal scaling, the resistance and capacitance of a global interconnect significantly increase, degrading the delay, power consumption, and signal integrity of a circuit.
- Practical scaling scenarios differ from ideal scaling since under ideal scaling, the resistance of a global interconnect rapidly increases.
- In the quasi-ideal scaling scenario, the vertical dimensions scale by $1/\sqrt{S}$ rather than S, thereby decreasing the pace of the increase in interconnect resistance.
- The quasi-ideal scaling scenario lowers the ground capacitance (since the distance between the interconnect and substrate is reduced with a smaller scaling factor), whereas the coupling capacitance increases since the thickness of an interconnect scales more slowly.
- In the constant resistance scaling scenario, the interconnect dimensions are scaled to ensure that the local and global interconnect resistances remain the same.
- Under the constant resistance scaling scenario, the coupling capacitance among the global interconnects is increased since the thickness increases to maintain a constant resistance.
- The constant resistance scaling scenario operates well for local interconnects, but the delay increases rapidly for global interconnects due to the increasing coupling capacitance.
- The constant thickness scaling scenario assumes that the thickness of the interconnects remains the same, whereas the remaining parameters scale by S except the length of the global interconnects, which scale by S_c.
- Despite maintaining a constant thickness, the delay of the global interconnects grows rapidly due to increasing IC dimensions and the reduction in the interconnect width, spacing, and height.

- A tradeoff exists between the parasitic resistance and the coupling capacitance since a larger cross sectional area reduces the resistance at the expense of an increase in coupling capacitance.
- If the coupling capacitance dominates over the ground capacitance, neither quasi-ideal nor constant thickness scaling scenarios reduce the global interconnect delay.
- The constant resistance scaling scheme partly alleviates the rapid increase in global interconnect delay at the expense of a lower integration density.
- Several techniques have been proposed at different abstraction levels to alleviate the adverse effects of scaling on the global interconnects.
- Despite a relatively higher resistivity and degraded electromigration characteristics, aluminum had been the primary material for on-chip interconnects until the late 1990s due to the low diffusion coefficient in silicon.
- Starting in the late 1990s, copper replaced aluminum due to the lower resistivity and enhanced electromigration characteristics of copper.
- An interlayer insulator with a low relative dielectric constant (low-K) is desirable since a smaller interconnect capacitance reduces the delay and power consumption while also enhancing signal integrity due to less capacitive crosstalk.
- Low-K interlayer materials ($K \approx 3$) such as carbon and hydrogen doped silicon were first introduced for the 90 nm technology node.
- Porous dielectric film, spin-on dielectrics, and air-gap based dielectrics are examples of ultra-low-K insulators where $K < 2.5$.
- 3-D integration has emerged as a promising technology to alleviate the interconnect bottleneck by reducing the length and number of global interconnects.
- 3-D integration also provides unique opportunities for heterogeneous integration by merging disparate materials and technologies within a 3-D stack.
- Several different 3-D integration approaches exist, such as: (1) transistor level 3-D integration, (2) SiP 3-D integration, and (3) wafer-level TSV 3-D integration.
- In transistor-level 3-D integration, the active devices within a logic gate are fabricated on different layers, where the formation of the active devices within the upper layers is a primary challenge.

- In SiP 3-D integration, multiple bare or unpackaged dies are assembled, where the interconnections among the dies are achieved by one of three methods: (1) wire bonding, (2) low density vertical interconnects and solder balls located along the periphery of the package, and (3) area array based C4 bumps located on the side surface of the package.
- A primary disadvantage of SiP 3-D integration is the low density of the vertical interconnections, which limits the reduction in the number and length of the global interconnects.
- In wafer-level TSV 3-D integration, multiple wafers are thinned, aligned, and bonded where the high density TSVs pass signals through the substrate of the planes to communicate between dies.
- The communication bandwidth between dies is highest in TSV wafer-level 3-D integration due to the small TSV dimensions.
- Challenges related to TSV wafer-level 3-D integration are classified into the following categories: (1) manufacturing challenges, (2) testing challenges, (3) thermal challenges, and (4) circuit and computer-aided design challenges.
- Replacing the global nets such as the data buses and clock networks with on-chip optical interconnects has been proposed to increase speed and lower power consumption.
- An optical link requires electrical-to-optical and optical-to-electrical conversion that consumes power and introduces latency.
- An on-chip optical interconnect data path consists of a transmitter including a modulator and driver, an optical waveguide, and a receiver including a photodetector and amplifier.
- The elements of an on-chip optical interconnect link should be compatible with CMOS technology, limiting the available materials and processes.
- Despite exhibiting superior delay characteristics, an important limitation of on-chip optical waveguides is a relatively high area requirement where the minimum pitch of the waveguide varies between 10 μm and 20 μm.
- Another approach for on-chip optical communication relies on free space propagation of optical signals rather than bulky waveguides.
- In free space propagation of optical signals, a separate die consisting of micro-lenses, mirrors, and vertical cavity surface emitting lasers is bonded or 3-D integrated to a silicon IC consisting of CMOS circuits.

- SWCNTs have been proposed to replace metallic interconnects because of two primary advantages: (1) enhanced reliability characteristics due to the ability to carry significantly higher current densities, and (2) greater electrical and thermal conductivity when fabricated in bundles.
- Bundles of SWCNTs and multiwalled MWCNTs consisting of nested and co-centric SWCNTs reduce the resistance of carbon nanotubes.
- For global interconnects, densely packed SWCNT bundles or MWCNTs with large diameters achieve a lower propagation delay as compared to copper.
- Despite a potential reduction in delay, carbon nanotubes suffer from limited manufacturing techniques that produce relatively high impurity concentrations, thereby causing the conductivity of carbon nanotubes to significantly deviate from theoretical expectations.
- Graphene nanoribbons have been proposed to alleviate fabrication challenges in carbon nanotubes since GNRs are planar.
- Despite certain advantages during the fabrication process, GNRs suffer from several issues such as edge scattering (and therefore a reduced electron mean free path) and intersheet electron hopping.
- Specific technology requirements need to be satisfied for GNRs to outperform copper and carbon nanotubes when used as on-chip interconnects.

PART II

Interconnect Networks