Part One
Scaling and Challenge of Si-based CMOS

COPYRIGHTED MATERIAL

High-k Gate Dielectrics for CMOS Technology, First Edition. Edited by Gang He and Zhaoqi Sun.
© 2012 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2012 by Wiley-VCH Verlag GmbH & Co. KGaA.

1
Scaling and Limitation of Si-based CMOS

Gang He, Zhaoqi Sun, Mao Liu, and Lide Zhang

1.1
Introduction

Scaling transistor's dimensions has been the main tool to power the development of silicon integrated circuits (ICs). The more an IC is scaled, the higher its packing density and the lower its power dissipation [1]. These have been key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much reduced physical size compared to their predecessors. However, the fundamental limits of complementary metal oxide semiconductor (CMOS) technology have been discussed, reviewed, and claimed to be at hand since the first MOS processes were developed [2, 3]. The integration of semiconductor devices has gone through different stages. At each stage of evolution, limits were reached and then subsequently surpassed, and very little has changed in the basic transistor design.

Questions about the end of CMOS scaling have been discussed, but engineering ingenuity has proven the predictions wrong. The most spectacular failures in predicting the end involved the "lithography barrier," in which it was assumed that spatial resolution smaller than the wavelength used for the lithographic process is not possible [4, 5], and the "oxide scaling barrier," in which it was claimed that the gate oxide thickness cannot be reduced below ~3 nm due to gate leakage [6]. For the present and near future, it appears unlikely that lithography will limit the scaling of silicon devices. The cost of lithography tools, including that required for making masks, may, however, impede future scaling of devices. It is more likely that a fundamental limit will halt further scaling when at least one of physical dimensions of the device, be it a length, width, depth, or thickness, approaches a few silicon atoms. Manufacturing tolerance, and therefore economics, may dictate an end to the scaling of silicon devices before these fundamental limits are reached. Therefore, in this chapter on scaling of SiO_2-based gate dielectrics for MOS devices, only present perceived fundamental limits are considered.

The downscaling of an MOSFETS sets demands especially on the properties of the gate oxide, SiO_2, which nature has endowed the silicon microelectronics industry with and which has dominated as the favorite and by far most practical choice for FET

High-k Gate Dielectrics for CMOS Technology, First Edition. Edited by Gang He and Zhaoqi Sun.
© 2012 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2012 by Wiley-VCH Verlag GmbH & Co. KGaA.

gate dielectric materials since 1957. SiO_2 offers several crucial advantages and industry's acquired knowledge of its properties and processing techniques has allowed its continuous use for the past several decades. However, further scaling the FET is eventually going to be impeded by the inability to further reduce the oxide thickness without risking a breakdown of the device. Recently, however, the thickness of the gate oxide has scaled more slowly compared to its historical pace. An equivalent oxide thickness (EOT) of ~1 nm has been used for the past two to three generations because of issues such as process controllability, high leakage current, and reliability limits, signaling an end to the scaling era and the advent of a new era of material and device evolution. As we know, when it is thinned to 1 nm, which corresponds to only 4–5 atom layers, some new technology problems arose [7, 8]. A variation in thickness of only 0.1 nm could result in changes in the device operation condition, making it extremely difficult to maintain device tolerances. As one might imagine, this exponential increase in the gate dielectric leakage current has caused significant concern about the operation of CMOS devices, particularly with regard to standby power dissipation, reliability, and lifetime. This will likely be one if there is no the major contributor leading to the limited extendability of SiO_2 as the gate dielectric in the 1.5–2.0 nm regimes. Therefore, in this chapter, a detailed discussion of current dielectrics and those proposed for future generations is included. Present beliefs regarding the limitations of silicon dioxide as the gate dielectric are reviewed. Ultrathin oxides, with gate dielectrics below 40 Å, some of which are nitrided silicon oxides, are discussed. The benefits and limitations of alternative for new high-k materials for possible high-performance CMOS applications are reviewed.

1.2 Scaling and Limitation of CMOS

1.2.1 Device Scaling and Power Dissipation

The primary goal of CMOS scaling is reduction of the cost per functional power, by increasing the integration density of on-chip components. The elaboration of constant field scaling rules entails concomitant performance and power consumption improvements, which have shaped the evolution of silicon technology [1]. The concept of device scaling is illustrated in Figure 1.1. In constant field scaling, the physical dimensions of the device (gate length L_G and width W_G, oxide thickness t_{ox}, and junction depth X_j), and the supply and threshold voltages (V_{DD} and V_T, respectively), are reduced by the same factor, $\alpha > 1$, so that the two-dimensional pattern of the electric field is maintained constant, while circuit density increases by $\sim\alpha^2$. This implies that the depletion width (W_d) must also be reduced by the same amount, which is achieved by increasing the substrate doping N_B by α. Consequently, both the gate capacitance ($C = L_G W_G \varepsilon_{ox}/t_{ox}$) and the drain saturation current ($I_{D,sat}$) are scaled down by α. The saturation current determines the transistor intrinsic switching delay $\tau \sim C V_{DD}/I_{D,sat}$, which is thus reduced by α, leading to a performance

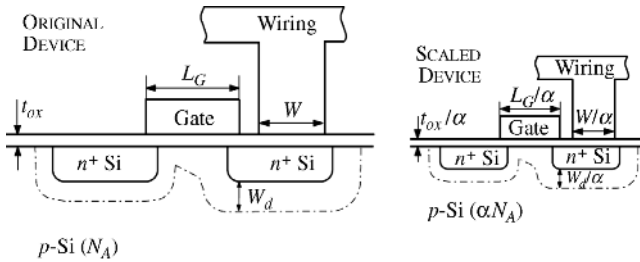

Figure 1.1 Conceptual schematic diagram of device scaling. Both device and wiring dimensions are required to scale by the same factor $1/\alpha$, in order to increase integration density by α^2. Scaling of the supply voltage by the same factor ($1/\alpha$) maintains the same 2D electric field pattern, subject to an equivalent scaling of the depletion width W_d. Reproduced from Ref. [10]. Copyright 2001, IEEE.

improvement. At the same time, the power dissipation ($P \sim I_{D,sat}V_{DD}$) is reduced by α^2, so that the power density ($P/(L_G W_G)$) remains unchanged.

Table 1.1 shows the gate length, supply voltage, and oxide thickness figures for several recent and future technology generations, as projected by the ITRS. It is obvious that present day device scaling does not adhere to the constant field scaling rules. This is because of several fundamental, nonscaling factors and practical considerations. Instead, the generalized scaling rules are followed, where the physical dimensions of the transistor are still reduced by a factor of α, providing the desired circuit density increase (α^2) and performance improvement (α), but the supply voltage is scaled by β/α, leading to an increase in the magnitude of the electrical field by $1 \leq \beta \leq \alpha$ [9]. A full list of MOSFET physical parameters, and their scaling factors, is given for constant field scaling and generalized scaling in Table 1.2. The last column in this table shows the rules for selective scaling, which relaxes one more constraint – the fixed ratio between gate length and width [10]. Such relaxation is driven mostly by the slower scaling pace of on-chip interconnect lines.

As can be seen in Table 1.2, under generalized and selective scaling scenarios, the dissipated power scales by β^2/α^2, while the power density increases by β^2. The power density is of paramount importance for chip packaging and systems design, and its increase imposes a practical limit for the exploitation of device scaling. It is worth

Table 1.1 ITRS-projected L_G, EOT and V_{DD}.

Year	L_G (nm)	EOT (nm)	V_{DD} (V)
2003	45	1.3	1.2
2005	32	1.2	1.1
2007	25	1.1	1.1
2009	20	0.9	1.0
2011	16	0.6	1.0
2013	13	0.5	0.9
2015	10	0.5	0.9

Table 1.2 Device parameters and their scaling factors. Reproduced from Ref. [10]. Copyright 2001, IEEE.

Physical parameters	Scaling rules' factors		
	Constant Field	Generalized	Selective
Gate length (L_G), oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	$1/\alpha_D$
Wiring width, channel width (W_G)	$1/\alpha$	$1/\alpha$	$1/\alpha_W$
Voltages (V_{DD}, V_T)	$1/\alpha$	β/α	β/α_D
Substrate doping (N_B)	α	$\beta\alpha$	β/α_W
Electric field	1	β	β
Gate capacitance ($C = L_G W_G \varepsilon_{ox}/t_{ox}$)	$1/\alpha$	$1/\alpha$	$1/\alpha_W$
Drive current ($I_{D,sat}$)	$1/\alpha$	β/α	β/α_W
Intrinsic delay ($\tau \sim CV_{DD}/I_{D,sat}$)	$1/\alpha$	$1/\alpha$	$1/\alpha_D$
Area ($A \propto L_G W_G$, or $\propto W_G^2$)	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_W^2$
Power dissipation ($P \sim I_{D,sat}V_{DD}$)	$1/\alpha^2$	β^2/α^2	$\beta^2/(\alpha_W\alpha_D)$
Power density (P/A)	1	β^2	$\beta^2/(\alpha_W/\alpha_D)$

noting that in the most recent technologies, and in the projections of the ITRS, supply voltage hardly scales, that is, $\beta \approx \alpha$ [11]. This already suggests that the power density grows at the same rate as the integration density. Actually, the issue is even worse than it appears considering the scaling rules alone, because of nonscaling factors, leading to the increase and dominance of the static, leakage power. This is clearly demonstrated by the crossing lines of Figure 1.2, illustrating the trends in dynamic and leakage power densities with shrinking gate length. The actual measurements for devices with gate length between 1 μm and 65 nm are shown with symbols. The rapid escalation of leakage has ultimately led to power-constrained, application-specific

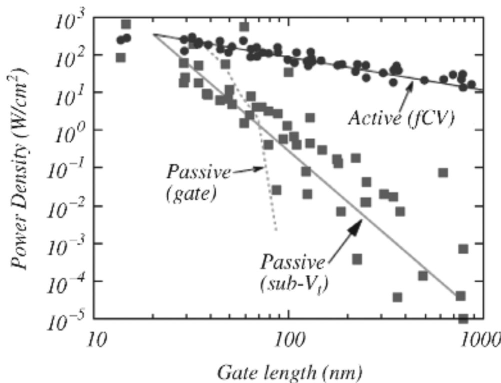

Figure 1.2 Comparison of measured active power and leakage (passive) power in devices with gate lengths ranging from 1 μm to 20 nm (symbols). Lines indicate the trend of a particular power component, as indicated. Reproduced from Ref. [11]. Copyright 2006, International Business Machines Corporation. (For a color version of this figure, please see the color plate at the beginning of this book.)

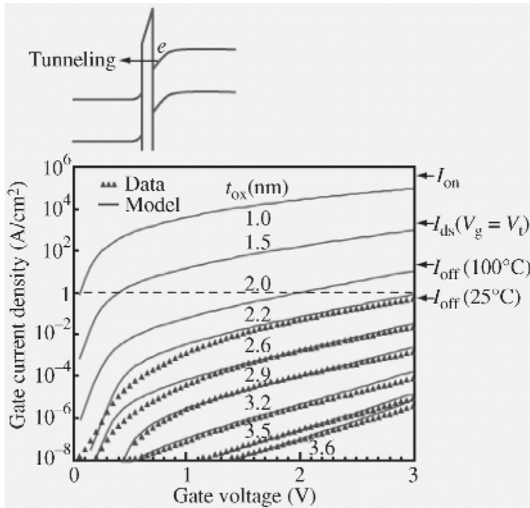

Figure 1.3 Measured and calculated oxide tunneling currents versus gate voltage for different oxide thicknesses. Reproduced from Ref. [12]. Copyright 2002, International Business Machines Corporation. (For a color version of this figure, please see the color plate at the beginning of this book.)

evaluation of scaling scenarios since different applications can tolerate different power densities.

Of the two components of consumed power, the dynamic power, dissipated during a switching between logic states, can be ameliorated by limiting the switching frequency f, to which it is proportional ($P_{dyn} \approx CV^2_{DD}f$). The other component – static, leakage power, dissipated while maintaining a logic state, is exponentially sensitive to some of the device parameters and their variability, as well as to nonscaling factors. Leakage power dissipation is time invariant and now poses the most significant scaling limit [11].

1.2.2
Gate Oxide Tunneling

When the dimensions of a MOSFET are scaled down, both the voltage level and the gate oxide thickness must also be reduced [1]. Since the electron thermal voltage, kT/q, is a constant for room-temperature electronics, the ratio between the operating voltage and the thermal voltage inevitably shrinks. This leads to higher source-to-drain leakage currents stemming from the thermal diffusion of electrons. At the same time, to keep adverse 2D electrostatic effects on threshold voltage under control, the thickness of gate oxide is reduced nearly in proportion to channel length, as shown in Figure 1.3. This is necessary in order for the gate to retain more control over the channel than the drain. For CMOS devices with channel lengths of 100 nm or less, an oxide thickness of $<$3 nm is needed. This thickness comprises only a few layers of atoms and is approaching fundamental limits.

While it is amazing that SiO_2 can take us this far without being limited by extrinsic factors such as defect density, surface roughness, or large-scale thickness, and uniformity control, oxide films this thin are subject to quantum mechanical tunneling, giving rise to a gate leakage current that increases exponentially as the oxide thickness is scaled down. Tunneling currents for oxide thicknesses ranging from 3.6 to 1.0 nm are plotted versus gate voltage in Figure 1.3 [12]. In the direct tunneling regime, the current is rather insensitive to the applied voltage or field across the oxide, so reduced voltage operation will not buy much relief. Although the gate leakage current may be at a level that is negligible compared to the on-state current of a device, it will first have an effect on the chip standby power. Note that the leakage power will be dominated by turned-on n-MOSFETs, in which electrons tunnel from the silicon inversion layer to the positively biased gate, as shown in the inset of Figure 1.3. Edge tunneling in the gate-to-drain overlap region of turned-off devices should not be a fundamental issue since one can always build up the corner oxide thickness by additional oxidation of poly-silicon after gate patterning. p-MOSFETs have a much lower leakage than n-MOSFETs because there are very few electrons in the p^+ polysilicon ("poly") gate available for tunneling to the substrate, and hole tunneling has a much lower probability. If one assumes that the total active gate area per chip is of the order of 0.1 cm^2, the maximum tolerable gate leakage current will be of the order of 10 A/cm^2. This sets a lower limit of 1.0–1.5 nm for the gate oxide thickness. Dynamic memory devices have a more stringent leakage requirement and therefore must impose a higher limit on gate oxide thickness [10].

The intolerable growth of gate tunneling leakage has triggered a radical change in the silicon technology, which aims to introduce dielectric permittivity scaling through material engineering. The goal is to scale the oxide sheet capacitance C_{ox}, which is required to maintain good electrostatic control by the gate, but avoid decreasing the physical thickness of the gate insulator. This may be achieved by increasing the gate dielectric constant of the insulator. With respect to scaling, the relevant metric is now the EOT, defined through the relation $\varepsilon_{\text{high }k}/t_{\text{high }k} = \varepsilon_{ox}/\text{EOT} = C_{ox}$, where $\varepsilon_{\text{high }k}$ and $t_{\text{high }k}$ are the permittivity and physical thickness of the high-k material, respectively, and ε_{ox} is the SiO_2 permittivity. Therefore, the introduction of high-k dielectric materials is seen as the only way to realize the projections of the ITRS for effective oxide thicknesses reaching 0.5 nm.

However, the transition to high-k gate dielectric also entails a replacement of the poly-Si gate with a metal gate, which is another formidable technological challenge, in addition to the difficulties of controlled growth of ultrathin high-k oxide films [13]. These complications have delayed the adoption of alternative gate dielectric stacks and have led to the continuous relaxation of EOT scaling requirements.

1.2.3 Gate Oxide Scaling Trends

The gate oxide has been aggressively scaled in recent generations. Figure 1.4 shows extrapolated gate oxide scaling targets based on published data from recent Intel technologies [14]. The technology node refers to the smallest poly-Si gate length that

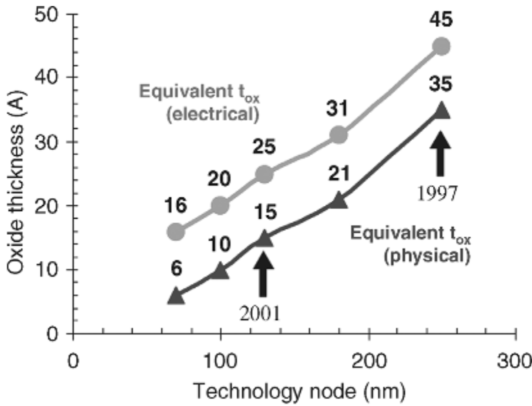

Figure 1.4 Extrapolated gate oxide scaling trend for recent CMOS technologies. Reproduced from Ref. [14]. Copyright 2000, IEEE. (For a color version of this figure, please see the color plate at the beginning of this book.)

can be defined by photolithography and roughly corresponds to the minimum channel length for a given process technology. A more complete list of projected transistor parameters is given in Table 1.3. The predictions are based on extrapolations of published state-of-the-art 180 nm technologies [15]. These projections, representative of the current targets for high-performance logic technology, aggressively outpace those compiled in the 2000 update of ITRS. The two data sets in Figure 1.4 refer to the equivalent electrical and physical thickness of the gate oxide. The EOT refers to how thin a pure SiO_2 layer would need to be in order to meet the gate capacitance requirements of a given technology. In a modern MOSFET device, the gate oxide behaves electrically as if it were 8–10 Å thicker than its physical thickness because depletion in the poly-Si gate and quantization in the inversion layer each extend the centroids of charge modulated by the gate voltage by 4–5 Å [16]. From Figure 1.4, it is clear that the physical thickness of the gate oxide is rapidly approaching atomic dimensions. The 250 nm technology, which entered volume production in 1997, used SiO_2 layer with approximately 40 Å physical t_{ox}, corresponding to approximately 20 Å monolayers of SiO_2. In contrast, the 100 and 70 nm technologies, scheduled for production in the next 5–10 years, will require gate capacitance values achievable only with SiO_2 layers as thin as 10 and 7 Å, respectively,

Table 1.3 Projected transistor parameters for future technology generations.

Generation (nm)	180	130	100	70	Scaling factor
I_{gate} (nm)	100	70	50	35	0.7×
V_{dd} (V)	1.5	1.2	1.0	0.8	0.8×
$T_{ox'}$ electrical (A)	31	25	20	16	0.8×
$T_{ox'}$ physical (A)	21	15	10	6	0.8×
I_{off} at 25 °C (nA/um)	20	40	80	160	2×

to guarantee proper device operation. A 10 Å film consists of only three–four monolayers of SiO_2.

1.2.4
Scaling and Limitation of SiO_2 Gate Dielectrics

The apparent robust nature of SiO_2 [17], coupled with industry's acquired knowledge of oxide process control, has helped the continued use of SiO_2 for over 30 years in CMOS technology. The use of amorphous, thermally grown SiO_2 as a gate dielectric provides thermodynamically and electrically stable, high-quality Si–SiO_2 interface with superior electrical isolation properties. In advanced integrated devices, the SiO_2 gate dielectrics are produced with charge densities of $10^{10}/cm^2$, mid-gap interface state densities of $10^{10}/cm^2$, and dielectric strengths of 15 MV/cm [17]. In fact, we can say that the silicon age own its existence to the superb quality of thermally grown SiO_2 as gate insulator and Si surface electrical passivator, Actually, no other known interface approaches the electrical figures quoted above the Si–SiO_2 interface. Therefore, the progress epitomized by Moore's law is best characterized as a steep but smooth development, having been achieved with no major revolution in fundamental device designs and no changes in the materials that constitute the heart of the MOSFET:Si and SiO_2.

Nevertheless, the outstanding evolution of silicon is rapidly approaching a saturation point where device fabrication can no longer be simply scaled to progressively smaller sizes. The origin of this saturation is indicated in Figure 1.5, where not only the number of transistors in an integrated circuit but also the corresponding gate dielectric thickness are plotted as a function of time. The thinning of the gate dielectric required by scaling rules, at present between 2 and 2.5 nm in fabrication, necessary for reaching the next generations of integrated devices will give origin to

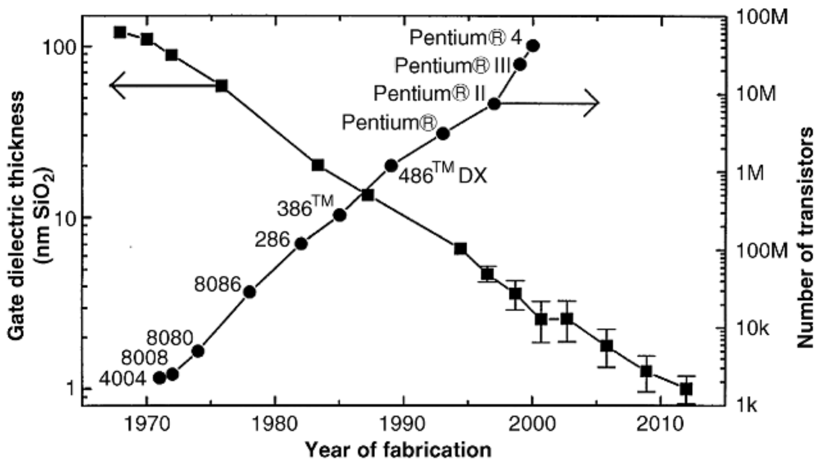

Figure 1.5 Moore's law as expressed by the number of transistors per chip. The corresponding SiO_2 gate dielectrics thickness is also shown.

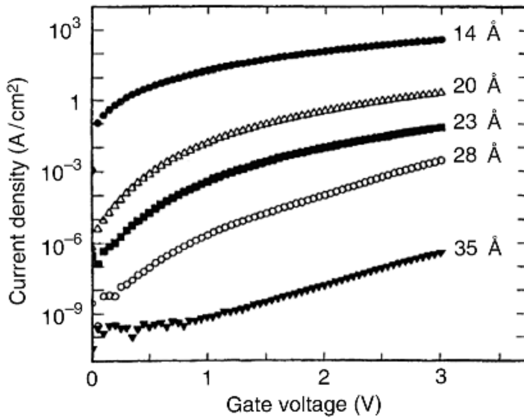

Figure 1.6 Gate current density as a function of gate voltage for MOS capacitors with different gate dielectric thickness. Reproduced from Ref. [19]. Copyright 1996, IEEE.

unacceptably high gate current arising from electron tunneling through the SiO_2 films. Gate current densities versus gate voltage are plotted in Figure 1.6 for transistor made with decreasing gate dielectric thickness below 3.5 nm, where the exponential increase in the leakage current is clearly observed. Figure 1.7 shows that progressive scaling of the gate oxide below 1.5 nm will lead to leakage current that destroys the transistor effect itself, as the ON and OFF states of the transistors might not be clearly distinguishable as necessary [18].

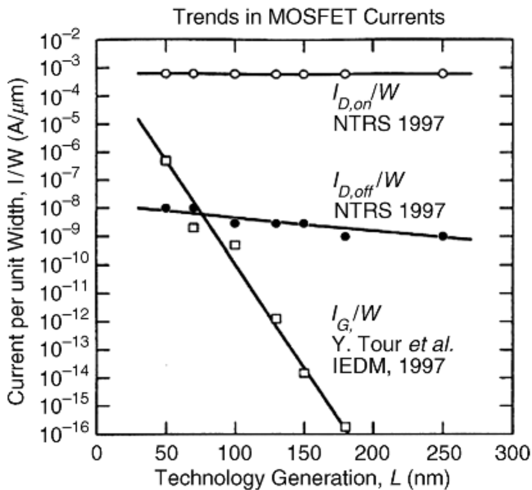

Figure 1.7 Gate current and train current per unit channel for the transistor states ON and OFF as a function of transistor channel length. Reproduced from Ref. [18]. Copyright 1997, Materials Research Society.

The inferior limit for gate oxide thickness can be extended down to 1.3 nm, owing to the characteristics of SiO_2 and the acquired knowledge of semiconductor manufacturers on oxide process control. Although high leakage current density is measured for such devices, as shown in Figure 1.6, transistors intended for high-performance microprocessor applications can sustain these currents [19]. However, further downscaling deteriorates the electrical properties rendering the devices inoperative. Several independent works showed the inconvenience of fabricated CMOSFET devices with SiO_2 gate oxides thinner than about 1.0–1.2 nm because there is no further gain in transistor drive current, setting a fundamental limit for gate oxide scaling.

The existence of a more fundamental limit to scaling SiO_2 around 1.0–1.2 nm as demonstrated at the atomic scale in a convincing experiment by Muller *et al.* from Bell Labs is reproduced in Figure 1.8 [8]. Using a scanning transmission electron microscope (STEM) probe with 2 Å resolution, they studied the chemical composition and electronic structure of oxide layers as thin as 7–12 Å using detailed electron energy loss spectroscopy (EELS) measurements. By moving the probe site-by-site through the ultrathin SiO_2 layers, they mapped the local unoccupied density of

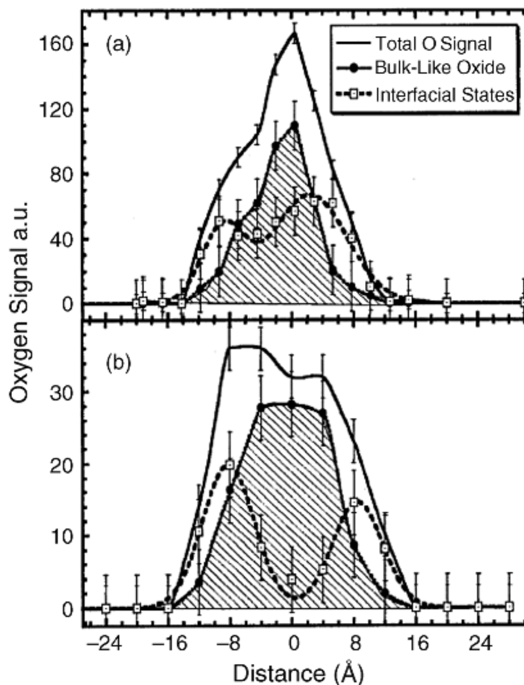

Figure 1.8 Oxygen signals in EELS performed in two different MOS capacitors made of (a) 1.3 nm and (b) 1.8 nm oxides that were analyzed in two components, one from bulk-like SiO_2 states and another from interfacial-like states. Reproduced from Ref. [8]. Copyright 1999, Nature Publishing Group.

electronic states, which provides insight into the local energy gap of the material, as a function of the probe position. In their work, the local energy gap was given by the separation between the highest occupied and lowest unoccupied states. They found that three to four monolayers of SiO_2 were needed to ensure that at least one monolayer maintained a fully bulk-like bonding environment, giving rise to the wide, insulating bandgap of SiO_2. Since the first and the last monolayers form interfaces with Si and poly-Si, respectively, they have bonding arrangements intermediate to those of bulk Si and bulk SiO_2 and hence have energy gaps smaller than that of bulk SiO_2. Based on these insights, Muller *et al.* concluded that the fundamental scaling limit of SiO_2 is likely to be in the range of 7 Å to 12 Å. Another important insight from their study was that for a 10 Å oxide, a 1 Å increase in the root mean square (RMS) interface roughness will lead to a factor of 10 increase in the gate leakage current, showing that the growth of such thin layers must be precisely controlled on atomic scales.

There has been a remarkable agreement between experiment and theory regarding the scaling limit of SiO_2. Theoretical studies by Tang *et al.* employing a Si/SiO_2 interface model based on the β-cristobalite form of SiO_2 showed that the band offset at the interface degraded substantially when the SiO_2 layer was scaled to less than three monolayers [20]. The large reduction in the band offset was attributed to a reduction in the SiO_2 bandgap and also suggested 7 Å as the scaling limit of SiO_2. A more recent study by Kaneta *et al.* using a Si/SiO_2 interface model based on quartz SiO_2 directly computed the local energy gap as a function of position through the interface [21]. While the transition from bulk Si to bulk SiO_2 in their model was structurally abrupt, it was found that the full bandgap of SiO_2 was not obtained until the second monolayer of SiO_2 was reached. Again, these calculations suggest that approximately 7 Å of SiO_2 is the minimum required for substantial band offsets to develop at the interface, indicating the formation of a large bandgap. Thus, both experiment and theory suggest that the bulk properties of SiO_2, including the wide, insulating bandgap needed to isolate the gate and channel regions, cannot be obtained for films less than 7 Å thick. Since technology roadmaps predict the need for sub-6 gate oxides in future generations, it is unlikely, from the viewpoint of both static power dissipation and fundamental materials science, which SiO_2 will scale beyond the 70 nm generation.

There are other limiting factors regarding SiO_2 gate oxide scaling. One is device reliability, which is open to debate whether the stringent 10-year reliability criterion set by industry for CMOS technology could be fulfilled by devices made of gate oxides thinner than 1.5–2 nm [22]. Indeed, dielectrics degradation due to "hot electron" irradiation of the Si–SiO_2 interface can trigger a succession of physical and chemical phenomena whose overall consequence on reliability is acceleration of dielectric breakdown of the gate oxide.

In view of the limiting factors described above, in order to achieve further scaling of integrated devices based on Si and SiO_2, the semiconductor industry has found solutions that allowed significant progress with the remarkable advantage of remaining within the Si–SiO_2 materials framework.

1.2.5
Silicon Oxynitrides

As scaling of the oxide thickness continues, it is obviously desirable to reduce the leakage current and maintain or increase the reliability of such films. The introduction of nitrogen into SiO_2 has been used to eliminate a number of concerns, although not all with equal success. The first introduction of nitrogen into SiO_2 films was for much thicker films and it was confirmed that the reliability of these films could be increased if the films were annealed in ammonia or other nitrogen-containing gas [23]. Some observations have shown that nitrogen could reduce the "defect generation rate" in these films, which can be thought to be a result of the ability of the nitrogen to getter hydrogen or reduce its diffusion [24, 25]. However, the concentration of nitrogen into the SiO_2 should be carefully controlled. Too much incorporation of nitrogen leads to large flatband or threshold-voltage shifts [26, 27]. As more experiments were performed to tailor the nitrogen profiles in oxide films, it was found that nonhydrogen nitrogen species were even more efficient at increasing device reliability and reducing defect generation rate since hydrogen was not introduced during the nitridation process. The incorporation of large amounts of nitrogen not only leads to the reduction of boron penetration but also causes threshold-voltage shifts, ΔV_t, and mobility and transconductance degradation that depend upon both the nitrogen concentration and its concentration profile, which can be due at least partially to the positive charge that results from the nitrogen incorporation into the SiO_2 matrix. The nitrogen incorporated near the Si/SiO_2 interface also reduces the mobility of the device.

The use of oxide/nitride stacks can eliminate boron penetration if the nitride film is of sufficient thickness [28–30]. With the incorporation of a nitride into the gate dielectric, the permittivity of the stack is greater than that of SiO_2. The physical thickness of the stack can be greater than that of a single SiO_2 film of equivalent thickness since the dielectric constant of silicon nitride is approximately twice that of SiO_2. Therefore, with the addition of silicon nitride to the stack, an increase in the effective thickness of approximately 30% is possible. Silicon nitride should also be able to reduce the gate leakage current [29, 30]. There are, however, many concerns involving the use of silicon nitride as a gate material, the greatest of which is the electrical stability of the material. Silicon nitride is known, at least in thicker films, to contain large amounts of positive charge; some of that charge can be unstable when a bias is applied, and this instability leads to obvious device problems. Such films are also known to contain large amounts of hydrogen, which could be a reliability concern [31, 32].

Having in hands an interim solution for reliability, the exponential increase in the leakage current with oxide thinner remains the major difficulty preventing further gate dielectric scaling. One possible solution is to consider the derive current of a MOSFET, the drain current, which in a simplified approximation can be written as

$$I_D = \frac{W}{L}\mu C\left(V_G - V_T - \frac{V_D}{2}\right)V_D$$

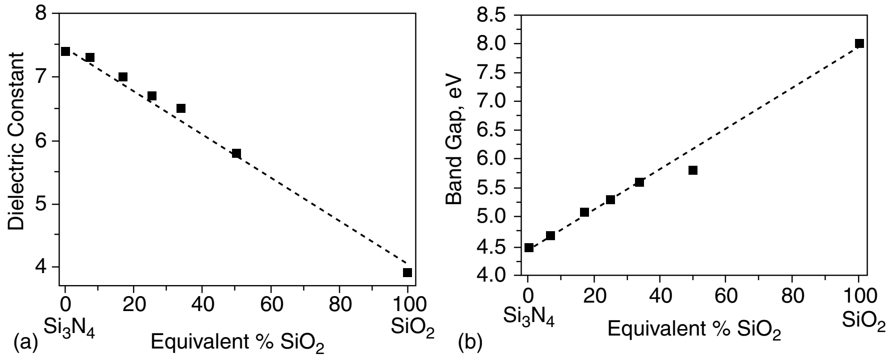

Figure 1.9 Dielectric constant (a) and bandgap (b) as a function of N content in the Si–O–N system.

where W is the width and L is the length of transistor channel, the charge carrier mobility in the channel, C the capacitance of the MOS capacitor, and $V_{\mathbf{G}}$, $V_{\mathbf{D}}$, and $V_{\mathbf{T}}$ are the gate, drain, and threshold voltages, respectively. $I_{\mathbf{D}}$ increase monotonically with $V_{\mathbf{D}}$ and then eventually saturates to a maximum when $V_{\mathrm{D.sat}} = V_{\mathbf{G}} - V_{\mathbf{T}}$ to yield

$$I_{\mathrm{D}} = \frac{W}{L}\mu C \frac{(V_{\mathrm{G}} - V_{\mathrm{T}})^2}{2}$$

Since gate voltage is limited by leakage current and reliability constraints, in order to maintain enough drain current such that the transistor can operate in safe, reliable conditions, C must be increased or at least kept constant. Now

$$C = \frac{\varepsilon A}{t}$$

where ε is the permittivity of the capacitor dielectric, and A and t are the area and the thickness, respectively. According to previous references, it can be noted that progressive scaling has been performed reducing gate oxide thickness by the same factor as the horizontal dimensions determining the area A. This scaling strategy is limited, according to Figure 1.9, by leakage current originated in electron tunneling through the ultrathin oxides required. Therefore, increasing the capacitor permittivity is the only way to prevent capacitance decrease. Figure 1.9 show that the dielectrics of silicon oxynitride increases with the N content from the pure SiO_2 value up to the pure Si_3N_4 value. As known, when materials with dielectric constant higher than that of SiO_2 are used in gate dielectric, for design purposes, the relevant magnitude is not the dielectric film thickness but rather an associated quality called equivalent oxide thickness, defined as

$$\mathrm{EOT} = \frac{k_{\mathrm{ox}}}{k_{\mathrm{high}\,k}} t_{\mathrm{high}\,k}$$

which represents the thickness of the SiO_2 layer with dielectric constant $k_{\mathbf{ox}}$ that would be required to achieve the same capacitance density as a given thickness $t_{\mathrm{high}\ k}$

of an alternative dielectric layer with dielectric constant $k_{\text{high } k}$. This means that it is possible to avoid electron tunneling that leads to unacceptable leakage current by simply increasing the physical thickness of the gate dielectric, without increasing its capacitance. Based on Figure 1.9, it is also shown that the bandgap decreases with increasing N concentration, which sets another limitation besides degradation of charge carrier mobility to the maximum N concentration in the oxynitride films. Thus, silicon oxynitride films are and will probably continue to be an interim solution for gate dielectrics for another few years. However, the limits of this solution are evident and therefore to keep pace with the historically steep progress of silicon technology it appears that the smoothness of this development will have to be abandoned in the near future.

1.3 Toward Alternative Gate Stacks Technology

1.3.1 Advances and Challenges in Dielectric Development

Fundamental challenge to the scaling of the gate dielectric is the exponential increase in tunnel current with reduction in film thickness. For films as thin as 20 Å, leakage currents can rise to 1–10A/cm^2. This incredibly high current can alter device performance, not to mention the difficulties associated with dissipation of such a large amount of power. Although higher power dissipation may be tolerable with some high-performance processors, it quickly leads to problems for portable machines. The reduction in gate leakage current is an important reason, if not the primary one, for replacing SiO_2-based dielectrics.

For a given technology, CMOS devices are designed with a specific gate capacitance, which is proportional to the dielectric constant and inversely proportional to the thickness of the gate material. To reduce the leakage current while maintaining the same gate capacitance, a thicker film with a higher dielectric constant is required. The gate leakage current, at least for direct quantum mechanical tunneling, exponentially depends upon the dielectric thickness, while the capacitance depends only linearly on the thickness. At first glance, this would seem to be a winning proposal since a substantial reduction in the current should be possible with only small increases in thickness. There is, however, another exponentially dependent term in the tunneling current – the barrier height between the cathode and the conduction band of the insulator. For a large number of dielectrics, the tunnel current exponentially depends upon the barrier height. Therefore, not only is a material with higher dielectric constant required but also this material must have a suitably large bandgap, and barrier height, to keep the gate leakage currents within reasonable limits.

Many materials have been suggested that could replace SiO_2 or SiON as a candidate for possible gate dielectric. Early, CeO_2 and Y_2O_3 were investigated to act as high-k dielectrics with an EOT of $\sim$6 nm [33, 34]. However, crystallization-

induced leakage currents and poor reliability in thick EOT regions prevent them to act as high-*k* candidates. In the 1990s, research on high-*k* gate dielectrics was updated. More attention has been focused on Ta_2O_5, TiO_2, and $SrTiO_3$, which were inherited from dynamic random access memory (DRAM) capacitor dielectric research [35–37]. Quickly, research on high-*k* dielectrics converged on the ZrO_2 and HfO_2 families with larger bandgaps [38, 39]. Owing to good thermal stability with Si, HfO_2 and ZrO_2 received most attention and many thorough reviews of research were also provided at that time [40–42]. Since then, Hf-based high-*k* gate oxides have emerged as promising candidates for high-*k* dielectrics [43–47]. Furthermore, improvement in mobility of CMOS devices using Hf-based gate dielectrics has been observed, almost matching that of nitride oxide with an EOT of $\sim$1 nm [48]. The two big breakthroughs to enable mobility enhancement are to understand transient charging behaviors in high-*k* devices and scaling of high-*k* thickness below the tunneling limit to eliminate residual tunneling carriers in high-*k* layer [49–51]. By keeping the high-*k* thickness smaller than 2 nm, mobility degradation originating from transient charging effects can be eliminated. Low standby power (LSTP) applications require low leakage current. To meet the requirement of LSTP, thickness of high-*k* layer should be increased without degrading the mobility. Meanwhile, optimization of high-*k* composition should be carried out to meet these requirements in LSTP applications.

To obtain microelectronic devices with excellent performance, aggressive EOT scaling is needed [52]. What is more, to scale the EOT of high-*k* gate stack, the growth of the interfacial layer should be controlled effectively. Figure 1.10 shows the mobility dependence on the perpendicular field at different thicknesses of the SiO_2 interfacial layer [53]. Based on this figure, it can be noted that reduction in mobility degradation has been observed due to the presence of the interfacial layer. Increasing the SiO_2 interfacial thickness leads to the increase in mobility approaching that of SiO_2. So, it can be concluded that the interfacial layer is beneficial to increase the inversion layer mobility, even if it increases EOT of the gate dielectric. The increased mobility is due to the decoupling between the motion of the electrons in the inversion layer and the phonons in the HfO_2.

To avoid the balance between EOT scaling and mobility, it is important to eliminate the charge scattering source within the high-*k* layer and increase the dielectric constant of the interfacial oxide with a minimal impact on the interface state density [55–57]. Although HfSiON can be used for 32 nm generations and below, any further extension of high-*k* dielectrics will require materials with higher dielectric constant that that of HfO_2. La_2O_3-based high-*k* oxides have been proposed as high-*k* candidates, but degraded dielectric characteristics compared to Hf oxides have been observed [58]. An alternative path to future scaling is to use other channel materials other than Si. For example, formation of interfacial oxide can be controlled when high-*k* dielectric is deposited on Ge or GaAs substrates [59, 60]. The challenges in this path are to overcome the problems associated with the degraded interface and to make use of the benefits of high carrier mobility in Ge or other channel materials. If Hf-based high-*k* gate oxides can be successfully

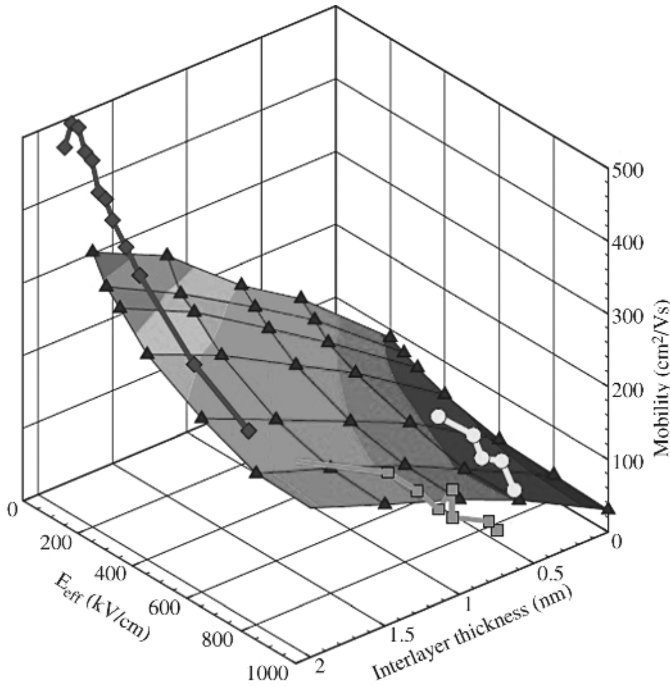

Figure 1.10 Electron mobility for HfO_2 versus effective field E_{eff} and the SiO_2 interfacial layer thickness. The mobility for pure SiO_2 and the experimental results [54] are also reported. Reproduced from Ref. [53]. Copyright 2007, Elsevier. (For a color version of this figure, please see the color plate at the beginning of this book.)

implemented with these channel materials, Hf-based oxides might then be usable beyond the 32 nm node.

Due to the thermal stability when in contact with poly-Si gate, metal electrodes have been used to integrate with high-k gate dielectrics. As we know, it is difficult to alter the effective work function (EWF) of a poly-Si gate when it is integrated with high-k dielectrics. It is attributed to the reasons that the EWF of a poly-Si/high-k dielectric stack is determined by Si-Hf bonding, not Fermi level of the poly-Si gate, which is called "Fermi-level pinning" [61]. In conclusion, PMOS devices with a poly-Si/high-k stack pose a high threshold voltage (V_{th}) that exceeds the practical range. This phenomenon has posed a serious challenge in the implementation of high-k dielectrics because to obtain the best performance in CMOS devices, the EWF of NMOS and PMOS should be close to the conduction (E_c) and valence bands (E_v), respectively [62]. To obtain high-performance applications, dual work function metal electrodes with EWFs approaching the conduction and valence bands are needed. However, metal electrode materials suffer from a limited EWF range when combined with high-k dielectrics. The following sections give more details of dielectric and electrode developments.

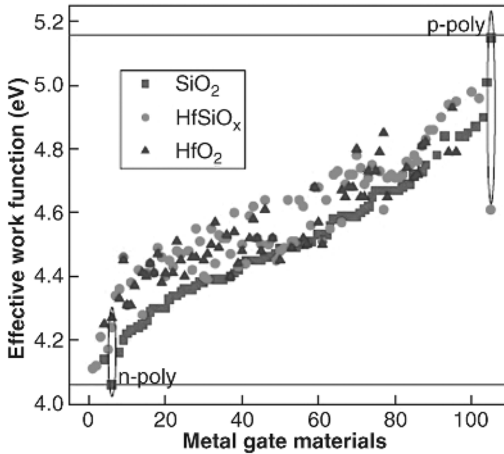

Figure 1.11 Effective work function of various metal electrode materials on SiO_2, $HfSiO_x$, and HfO_2 investigated with the terraced oxide method. Reproduced from Ref. [71]. Copyright 2008, Elsevier. (For a color version of this figure, please see the color plate at the beginning of this book.)

1.3.2
Advances and Challenges in Electrode Development

Based on the previous investigation, it can be noted that the application of poly-Si/high-k stacks are suppressed due to the V_{th} controllability, dopant penetration, and inversion oxide thickness T_{inv} scalability. Therefore, metal electrode materials with suitable EWFs close to E_c and E_v of the Si substrate should be investigated to replace poly-Si gate. The present challenge facing metal electrode is that the vacuum work function values of metals are not directly correlated with the V_{th} of devices. Thus, overall electrode processes, such as the deposition method, heat cycle, and dielectric composition, should be optimized to develop a well-defined metal electrode process. Many reported literatures suggests that experimental EWF results are consistent with the presence of a pinning effect [63, 64]; however, recently detailed studies of many of these electrode systems indicate contributions from extrinsic defect states induced by the electrode [65–67], chemical reduction of the high-k interface [68], or material characteristic changes in the electrode [69, 70] are the source of change in the EWF. Figure 1.11 shows the EWFs of vast metal systems extracted by the terraced oxide method, which were obtained using an oxide thickness series formed on a single substrate to minimize variations in interface charge [71].

The x-axis represents the material systems and three groups of data points show EWFs obtained with three different gate dielectrics. The difference in EWF between SiO_2, $HfSiO_x$, and HfO_2 is often explained by dipole formation, that is, if the EWF of the metal is smaller than the charge neutrality level (CNL) of the underlying dielectric, charge can be donated to the dielectric side and the EWF is

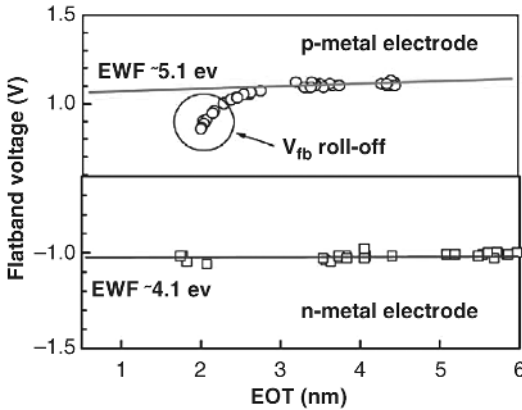

Figure 1.12 V_{fb}–EOT curve for a typical n-metal electrode and a p-metal electrode. Reproduced from Ref. [73]. Copyright 2006, Elsevier. (For a color version of this figure, please see the color plate at the beginning of this book.)

shifted toward the mid-gap value (∼4.6–4.7 eV) [72]. However, experimental data shown in Figure 1.11 indicate that the EWF shift because of dipole formation is actually unidirectional, indicating that dipole formation is more affected by metal–oxide bonding and microscopic charge transfer than the CNL of the underlying dielectric. Several metal systems show EWF values near the conduction and valence band edge of Si, even after high-temperature processing. These satisfy the requirements for gate-first CMOS integration, namely, that the EWF of the electrodes should be close to $E_{c,Si}$ for NMOS and $E_{v,Si}$ for PMOS after the CMOS heat cycle, which is typically a 1100 °C spike anneal or a 1000 °C 5 s rapid thermal anneal. The metal electrode systems are not specified in Figure 1.11 because the EWF of the metal/high-*k* stack can be controlled by many factors described above and, in fact, the electrode and high-*k* dielectric is best considered as a single material system that requires simultaneous optimization. Once thermally stable metal electrodes with band edge EWFs have been identified, the next challenge is to obtain a low V_{th} that matches the EWF values. Since the V_{th} of the actual device may be affected by additional factors, such as EOT, reactions with gate etch process gases, mixing with capping materials, and oxygen redistribution from the sidewall, it is not straightforward to predict V_{th} from the flatband voltage V_{fb}.

Figure 1.12 shows V_{fb}–EOT curve for a typical n-metal electrode and a p-metal electrode. From figure, it can be noted that high work function cannot be maintained as the device is scaled to lower EOT regimes and exhibit a “roll off”-like behavior, which can also be misinterpreted as Fermi-level pinning for p-metals. To further confirm this abnormal EWF behavior at low EOTs, a grand summary V_{fb}–EOT plot for various metal gates was presented in Figure 1.13. The V_{fb} roll-off region can be clearly found in Figure 1.14 that suggests a tradeoff in EWF in our goal of scaling the EOT. Roll off of the V_{fb} curve in the low EOT region can be contributed to several physical mechanisms including localized material interdiffusion, changes in stoichiometry,

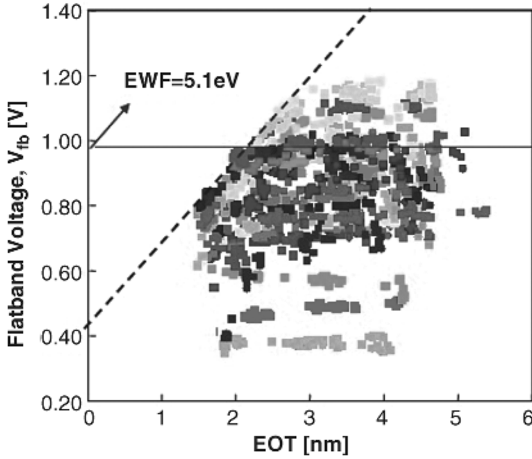

Figure 1.13 V_{fb}–EOT relationship for various materials systems investigated showing the V_{fb} roll-off for high work function metal electrodes. Reproduced from Ref. [71]. Copyright 2008, Elsevier. (For a color version of this figure, please see the color plate at the beginning of this book.)

and bulk charges of the SiO_2 bottom interface layer [65]. Based on previous analysis, it can be concluded that V_{fb} roll off should be minimized to obtain a high EWF of ~5.0 eV.

Silicides are the promising candidates for metal electrode applications. Fully silicided (FUSI) gates have been paid more attention because the poly-Si gate can be silicided with a relatively low heat cycle and minimal damage to the gate dielectric after device fabrication. $CoSi_2$ was used in the initial demonstration of a silicided gate, but the focus was shifted quickly to NiSi and $NiSi_3$ because the work function of Ni silicides can be modified with dopants in the poly-Si gate or a phase of Ni silicides [74, 75]. However, the EWF range with a silicide gate is still limited, even though the FUSI process uses only low-temperature steps. Various silicides have been studied to find a gate showing a band edge EWF. The most promising silicide gate materials are HfSi (~4.2 eV) and ErSi (~4.2 eV) for NMOS and PtSi (~4.9 eV) for PMOS [76, 77]. Even though the range of EWF is not wide enough for high-performance applications, these silicide gates can still be used in low-power applications or applications requiring quarter gap electrodes (~4.3 eV for NMOS and ~4.8 eV for PMOS), such as FinFETs or fully depleted silicon-on-insulator (FDSOI) devices. There are several potential challenges in the FUSI approach, such as pattern-dependent silicidation [78], strain control arising from volume expansion, and process complexity for a dual-silicide approach. Once electrode materials have been chosen for each application, the impact of reliability should be studied in detail. Metal electrodes can affect the reliability of gate dielectrics in many ways: diffusion and intermixing with the dielectric, oxygen, and nitrogen redistribution, and impurity contamination. None of these topics has been well investigated, primarily because the material systems have not been finalized yet.

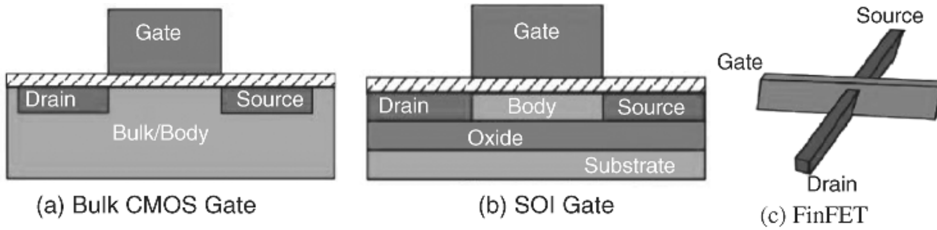

Figure 1.14 (a) Bulk, (b) SOI transistor structures, and (c) FinFET structure. (For a color version of this figure, please see the color plate at the beginning of this book.)

1.4
Improvements and Alternative to CMOS Technologies

1.4.1
Improvement to CMOS

1.4.1.1 New Materials

As analyzed in the past, the high-mobility III–V materials do not confer an advantage at the end of the scaling path since the isotropic, low-mass C valley does not provide a strong electron confinement and, furthermore, the light electron mass promotes source–drain tunneling. In addition, the smooth heterobarriers, responsible for the spectacular transport, are of low height and do not scale well. The single-transport valley also does not provide the large charge densities needed for these small devices. Materials such as Ge can provide larger carrier densities, but the low bandgap is a considerable disadvantage unless quantum confinement can be used to increase the bandgap in structures of practical dimensions.

High-k dielectrics are designed to address one particular aspect of off-state power consumptions: gate tunneling currents. It is likely that the gate dielectric thickness will be the first parameter to reach atomic dimensions. This is because the dielectric thickness indirectly controls the gate length. In general, the effective gate length needs to be 40 times the dielectric thickness to properly control short channel effects (SCE). Thus, the scaling that reduces gate length must also reduce the dielectric thickness. However, as the dielectric thickness decreases, electron tunneling through the dielectric becomes a significant issue. The proposed solution to this problem is to find a material that has a higher k value than the SiO_2 used at present as the gate dielectric. This would allow the actual thickness of the gate dielectric to be increased while still maintaining the same electric field in the channel.

While this sounds good in theory, implementation has proved to be challenging. To do this, a material must be found that meets many criteria. The material must be compatible with the surrounding silicon and the fabrication processes used. Also, it must have a breakdown time at least as long as silica. While there are many other requirements, suffice it to say that many materials have been proposed, but no good substitute has yet been found.

1.4.1.2 New Structures

It has been proposed to try and change the structure of the transistor itself. Here, two most prominent structural changes, silicon on insulator (SOI) and double-gate CMOS (DGCMOS) have been discussed. As we know, bulk CMOS structure is based on the concept that the transistor is connected to the substrate, as demonstrated in Figure 1.14a. However, for silicon on insulator, an insulating oxide is first deposited on the substrate and then the transistor is fabricated on top of that (Figure 1.14b). By doing this structure, the body is electrically isolated from its surroundings. This design leads to the improvement in performance. What is more, this new structure also lends itself to some new uses, such as using the insulating layer for a high-resistance element. However, there are also some drawbacks facing this new structure. Some techniques have been developed to address some of these issues. Based on SOI structure, IBM Company redesigned some PC line chips and performance improvement has been observed compared to bulk CMOS structure.

The second structure is DGCMOS, which is more experimental, but it may demonstrate promising application in future. In this structure, an additional gate has been added to increase coupling between the gate and the channel, also called "deal structure for scalability" [79]. However, it is difficult to design and realize this structure. Based on traditional method, second gate can be added to the body. However, the alignment issues of such a gate are troublesome. So, FinFET structure has been proposed, as shown in Figure 1.14c. This is a daunting challenge because the gate length is usually the smallest dimension that can be fabricated. There are some technologies that may address this, but more work needs to be carried out in this area.

1.5 Potential Technologies Beyond CMOS

So far, scaling has tremendous benefits that no alternative technologies can compete with the power of mainstream CMOS devices. In addition, more attention and knowledge have been focused on the investigation of MOS technology. Even if much efforts made, scaling CMOS is unquestionably approaching its limits. Although many issues have been resolved, scaling still cannot progress past the size of the molecule. The questions of what technology might surpass CMOS have come out. By now, there are many alternative devices that show promising replacement to CMOS for the future mentioned as follows.

1) Electrical-dependent nanodevices [80–81].

   Based either on ballistic transport and tunneling or on electrostatic phenomenon, these nanodevices are being investigated, such as carbon nanotubes field-effect transistors (CNTFETs), semiconductor nanowire field-effect transistors (NWFETs), resonant tunneling diodes (RTDs), and electrical quantum dot cellular automata (EQCA).

2) Magnetic-dependent nanodevices.
   Magnetic quantum dot cellular automata (MQCA) and spin field-effect transistors (spinFETs) are included in this class. Magnetostatic and spin transport are the phenomena for the operation of the devices.
3) Mechanical-dependent nanodevices.
   Compared to traditional CMOS devices, these nanodevices demonstrate some advantages. For example, CNTFETs have an extraordinary mechanical strength, low power consumption, better thermal stability, and higher resistance to electromigration [80]. The advantages of NWFETs over CMOS are similar to CNTFETs [80], plus the ability to operate at high speed produces saturated current at low bias voltage, and the potential to behave as either active or passive devices in single nanowires [81]. EQCA and MQCA exhibit a higher low-power dissipation, nonvolatility, and reconfigurability [81]. SpinFETs possess many advantages, such as small off-current, high operating speed, high-power gain, and low-power consumption compared to traditional CMOS devices. Except for the innovations already mentioned, there still exist some other proposed technologies, which may develop in parallel with CMOS allowing developers to choose the technology to satisfy their demands.

1.6 Conclusions

With the end of CMOS roadmap looming, there has been tremendous research in order to identify promising technologies to continue the historical trend of performance scaling. This chapter mainly explored the present status and challenges associated with alternative gate stack technology for future generations. Present beliefs regarding the limitations and main challenges faced by CMOS technology are reviewed. The benefits and limitations of alternative new high-k materials as candidate of SiO_2 and nitrided SiO_2 for possible high-performance CMOS applications are reviewed. Considering the progress of technology development, an alternative gate stack with a metal gate/high-k dielectric will be implemented for Si-based technology in the near future. Finally, we have also presented alternative devices that are potentially able to overcome the limitation in CMOS technology. These technologies will probably develop in parallel with CMOS allowing developers to choose the technology that best fits their needs.

Acknowledgments

The authors acknowledge the support from the National Natural Science Foundation of China (Grant Nos. 10804109 and 11104269) and Outstanding Young Scientific Foundation of Anhui University (Grant No. KJJQ1103). The authors are indebted to publishers/authors concerned for their kind permissions to reproduce their works, especially figures, used in this chapter.

References

1 Dennard, R.H., Gaensslen, F.H., Yu, H.N., Rideout, V.L., Bassous, E., and LeBlanc, A.R. (1974) Design of ion-implanted MOSFETs with very small physical dimensions. *IEEE J. Solid St. Circ.*, **SC-9**, 256.

2 Ligenza, J.R. and Spitzer, W.G. (1960) The mechanisms for silicon oxidation in steam and oxygen. *J. Phys. Chem. Solids*, **14**, 131.

3 Ligenza, J.R. and Spitzer, W.G. (1961) Effects of crystal orientation on oxidation rates of silicon in high pressure steam. *J. Phys. Chem.*, **65**, 2011.

4 Hoeneisen, B. and Mead, C.A. (1972) Fundamental limitations in microelectronics-I. MOS technology. *Solid State Electron.*, **15**, 819.

5 Wallmark, J.T. (1975) Fundamental physical limitations in integrated electronic circuits. *Inst. Phys. Conf. Ser.*, **25**, 133.

6 Stathis, J.H. and Dimaria, D.J. (1998) Reliability projection for ultra-thin oxides at low voltage. IEDM Technical Digest, p. 167.

7 Schulz, M. (1999) The end of the road for silicon. *Nature*, **399**, 729.

8 Muller, D.A., Sorsch, T., Moccio, S., Baumann, F., Evans-Lutterodt, K., and Timp, G. (1999) The electronic structure at the atomic scale of ultrathin gate oxide. *Nature*, **399**, 758.

9 Baccarani, G., Wordeman, M., and Dennard, R. (1984) Generalised scaling theory and its application to a 1/4 micrometer MOSFET design. *IEEE Trans. Electron. Dev.*, **31**, 452.

10 Frank, D., Dennard, R., Nowak, E., Solomon, P., Taur, Y., and Wong, H.-S. (2001) Device scaling limits of Si MOSFETs and their application dependencies. *Proc. IEEE*, **89**, 259.

11 Haensch, W., Nowak, E., Dennard, R., Solomon, P., Bryant, A., Dokumaci, O., Kumar, A., Wang, X., Johnson, J., and Fischetti, M. (2006) Silicon CMOS devices beyond scaling. *IBM J. Res. Dev.*, **50**, 339.

12 Taur, Y. (2002) CMOS design near the limit of scaling. *IBM J. Res. Dev.*, **46**, 213.

13 Gusev, E., Narayanan, V., and Frank, M. (2006) Advanced high-*k* dielectric stacks with poly-Si and metal gates. *IBM J. Res. Dev.*, **50**, 387.

14 Ghani, T., Mistry, K., Packan, P., Thompson, S., Stettler, M., Tyagi, S., and Bohr, M. (2000) Scaling challenges and device design requirements for high performance sub-50nm gate length planar CMOS transistors. Symposium on VLSI Technology Digest, p. 174.

15 Hargrove, M., Crowder, S., Nowak, E., Logan, R., Han, L., Ng, H., Ray, A., Sinitsky, D., Smeys, P., Guarin, F., Oberschmidt, J., Crabbe, E., Yee, D., and Su, L. (1998). High-performance sub-0.08 μm CMOS with dual gate oxide and 9.7ps inverter delay. IEDM Technical Digest, p. 627.

16 Lo, S.-H., Buchanan, D., and Taur, Y. (1999) Modeling and characterization of quantization, polysilicon depletion, and direct tunelling effects in MOSFETs with ultrathin oxides. *IBM J. Res. Dev.*, **43**, 327.

17 Green, M.L., Gusev, E.P., Degraeve, R., and Garfunkel, E.L. (2001) Ultrathin (<4 nm) SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics: understanding the processing, structure, and physical and electrical limits. *J. Appl. Phys.*, **90**, 2057.

18 Massoud, H.Z., Shiely, J.P., and Shanware, A. (1999) Self-consistent MOSFET tunneling simulations? Trends in the gate and substrate currents and the drain-current turnaround effect with oxide scaling. *Mater. Res. Soc. Symp.*, **567**, 227.

19 Momose, H.S., Ono, M., Yoshitomi, T., Ohguro, T., Nakamura, S.I., Saito, M., and Iwai, H. (1996) 1.5-nm direct-tunneling gate oxide Si MOSFET's. *IEEE Trans. Electron. Dev.*, **43**, 1233.

20 Tang, S., Wallace, R., Seabaugh, A., and King-Smith, D. (1998) Evaluating the minimum thickness of gate oxide on silicon using first-principles method. *Appl. Surf. Sci.*, **135**, 137.

21 Kaneta, C., Yamasaki, T., Uchiyama, T., Uda, T., and Terakura, K. (1999) Structure and electronic property of Si

(100)/SiO_2 interface. *Microelectron. Eng.*, **48**, 117.

22 Cao, M., Voorde, P.V., Cox, M., and Greene, W. (1998) Boron diffusion and penetration in ultrathin oxide with poly-Si gate. *IEEE Electron. Device Lett.*, **19**, 291.

23 Krisch, K.S. and Sodini, C.G. (1994) Suppression of interface state generation in reoxidized nitrided oxide gate dielectrics. *J. Appl. Phys.*, **76**, 2284.

24 Cartier, E., Buchanan, D.A., and Dunn, G.J. (1994) Atomic hydrogen-induced interface degradation of reoxidized-nitrided silicon dioxide on silicon. *Appl. Phys. Lett.*, **64**, 901.

25 Buchanan, D.A., Marwick, A.D., DiMaria, D.J., and Dori, L. (1994) Hot-electron induced redistribution and defect generation in metal-oxide-semiconductor capacitors. *J. Appl. Phys.*, **76**, 3595.

26 Joshi, A.B., Ahn, J., and Kwong, D.L. (1993) Oxynitride gate dielectrics for p1 polysilicon gate MOS devices. *IEEE Electron. Device Lett.*, **14**, 560.

27 Ma, Z.L., Chen, J.C., Liu, H., Krick, J.T., Cheng, Y.C., Hu, C., and Ko, P.K. (1994) Suppression of boron penetration in p1 polysilicon gate *p*-MOSFETs using low-temperature gate oxide N_2O anneal. *IEEE Electron. Device Lett.*, **15**, 109.

28 Taur, Y., Mii, Y.-J., Frank, D.J., Wong, H.-S., Buchanan, D.A., Wind, S.J., Rishton, S.A., Sai-Halasz, G.A., and Nowak, E.J. (1995) CMOS scaling into the 21st century: 0.1mm and beyond. *IBM J. Res. Dev.*, **39**, 245.

29 Ma, T.P. (1998) Making silicon nitride film a viable gate dielectric. *IEEE Trans. Electron. Dev.*, **45**, 680.

30 Ma, T.P. (1997) Gate dielectric properties of silicon nitride films formed by jet vapor deposition. *Appl. Surf. Sci.*, **117/118**, 259.

31 Buchanan, D.A., DiMaria, D.J., Chang, C.-A., and Taur, Y. (1994) Defect generation in 3.5nm silicon dioxide films. *Appl. Phys. Lett.*, **65**, 1820.

32 Stathis, J.H. and Cartier, E. (1994) Atomic hydrogen reactions with Pb centers at the (100)Si/SiO_2 interface. *Phys. Rev. Lett.*, **72**, 2745.

33 Fukumoto, H., Imura, T., and Osaka, Y. (1989) Heteroepitaxial growth of Y_2O_3 films on silicon. *Appl. Phys. Lett.*, **55**, 360.

34 Inoue, T., Yamamoto, Y., Koyama, S., Suzuki, S., and Ueda, Y. (1990) Epitaxial growth of CeO_2 layers on silicon. *Appl. Phys. Lett.*, **56**, 1332.

35 Alers, G.B., Werder, D.J., Chabal, Y., Lu, H.C., Gusev, E.P., Garfunkel, E., Gustafsson, T., and Urgahl, R.S. (1998) Intermixing at the tantalum oxide/silicon interface in gate dielectric structures. *Appl. Phys. Lett.*, **73**, 1517.

36 Ha, H.-K., Yoshimoto, M., Koinuma, H., Moon, B.-K., and Ishiwara, H. (1996) Open air plasma chemical vapor deposition of highly dielectric amorphous TiO_2 films. *Appl. Phys. Lett.*, **68**, 2965.

37 Pallecchi, I., Grassano, G., Marre, D., Pellegrino, L., Putti, M., and Siri, A.S. (2001) $SrTiO_3$-based metal-insulator–semiconductor heterostructures. *Appl. Phys. Lett.*, **78**, 2244.

38 Wilk, G.D. and Wallace, R.M. (2000) Stable zirconium silicate gate dielectrics deposited directly on silicon. *Appl. Phys. Lett.*, **76**, 112.

39 Lee, B.H., Kang, L., Nieh, R., Qi, W.J., and Lee, J.C. (2000) Thermal stability and electrical characteristics of ultrathin hafnium oxide gate dielectric reoxidized with rapid thermal annealing. *Appl. Phys. Lett.*, **76**, 1926.

40 Wilk, G.D. and Wallace, R.M. (1999) Electrical properties of hafnium silicate gate dielectrics deposited directly on silicon. *Appl. Phys. Lett.*, **74**, 2854.

41 Qi, W.J., Nieh, R., Dharmarajan, E., Lee, B.H., Jeon, Y., Kang, L., Onishi, K., and Lee, J.C. (2000) Ultrathin zirconium silicate film with good thermal stability for alternative gate dielectric application. *Appl. Phys. Lett.*, **77**, 1704.

42 Wilk, G.D., Wallace, R.M., and Anthony, J.M. (2001) High-*k* gate dielectrics: current status and materials properties considerations. *J. Appl. Phys.*, **89**, 5243.

43 He, G., Liu, M., Zhu, L.Q., Chang, M., Fang, Q., and Zhang, L.D. (2005) Effect of postdeposition annealing on the thermal

stability and structural characteristics of sputtered HfO_2 films on Si (100). *Surf. Sci.*, **576**, 67.

44 He, G., Fang, Q., Liu, M., Zhu, L.Q., and Zhang, L.D. (2007) Structural and optical properties of nitrogen-incorporated HfO_2 gate dielectrics deposited by reactive sputtering. *Appl. Surf. Sci.*, **253**, 8483.

45 He, G., Zhang, L.D., Meng, G.W., Li, G.H., Fang, Q., and Zhang, J.P. (2007) Temperature-dependent thermal stability and optical properties of ultrathin $HfAlO_x$ films on Si (100) grown by reactive sputtering. *J. Appl. Phys.*, **102**, 094103.

46 He, G., Zhang, L.D., Meng, G.W., Li, G.H., Fei, G.T., Wang, X.J., Zhang, J.P., Liu, M., Fang, Q., and Boyd, I.W. (2008) Composition dependence of electronic structure and optical properties of $Hf_{1-x}Si_xO_y$ gate dielectrics. *J. Appl. Phys.*, **104**, 104116.

47 He, G., Fang, Q., and Zhang, L.D. (2006) High-k $HfSi_xO_y$ gate dielectrics grown by solid phase reaction between sputtered Hf layer and SiO_2/Si. *J. Appl. Phys.*, **100**, 083517.

48 Quevedo-Lopez, M.A., Krishnan, S.A., Kirsch, P.D., Li, H.J., Sim, J.H., and Huffman, C. (2005) High performance gate first HfSiON dielectric satisfying 45nm node requirements. IEDM Technical Digest, 2005, p. 438.

49 Lee, B.H., Young, C., Choi, R., Sim, J.H., and Bersuker, G. (2005) Transient charging and relaxation in high-k gate dielectrics and their implications. *Jpn. J. Appl. Phys.*, **44**, 2415.

50 Choi, R., Song, S.C., Young, C.D., and Bersuker, G., and Lee, B.H. (2005) Charge trapping and detrapping characteristics in hafnium silicate gate dielectric using an inversion pulse measurement technique. *Appl. Phys. Lett.*, **87**, 122901.

51 Sim, J.H., Song, S.C., Kirsch, P.D., Young, C.D., Choi, R., Kwong, D.L., Lee, B.H., and Bersuker, G. (2005) Effects of ALD HfO_2 thickness on charge trapping and mobility. *Microelectron. Eng.*, **80**, 218.

52 Kirsch, P.D., Quevedo-Lopez, M.A., Li, H.J., Senzaki, Y., Peterson, J.J., Song, S.C., Wang, Q., Gay, D., and Ekerdt, J.G. (2006) Nucleation and growth study of atomic layer deposited HfO_2 gate dielectrics resulting in improved scaling and electron mobility. *J. Appl. Phys.*, **99**, 023508.

53 Ferrari, G., Watling, J.R., Roy, S., Barker, J.R., and Asenov, A. (2007) Beyond SiO_2 technology: simulation of the impact of high-k dielectrics on mobility. *J. Non-Cryst. Solids*, **353**, 630.

54 Ragnarsson, L.A., Severi, S., Trojman, L., Brunco, D.P., Johnson, K.D., Del abie, A., Schram, T., Tsai, W., Groeseneken, G., De Meyer, K., De Gendt, S., and Heyns, M. (2005). High performing 8 angstrom EOT HfO_2/TaN low thermal-budget n-channel FETs with solid-phase epitaxially regrown (SPER) junctions. Symposium on VLSI Technology Digest, p. 234.

55 Lai, C.S., Wu, W.C., Wang, J.C., and Chao, T.S. (2005) Characterization of CF_4-plasma fluorinated HfO_2 gate dielectrics with TaN metal gate. *Appl. Phys. Lett.*, **86**, 222905.

56 Kita, K., Kyuno, K., and Toriumi, A. (2005) Permittivity increase of yttrium-doped HfO_2 through structural phase transformation. *Appl. Phys. Lett.*, **86**, 102906.

57 Osten, H.J., Liu, P., Gaworzewski, P., Bugiel, E., and Zaumseil, P. (2000) High-k gate dielectrics with ultra-low leakage current based on praseodymium oxide. IEDM Technical Digest, p. 653.

58 Lu, X.B., Liu, Z.G., Wang, Y.P., Wang, X.P., Zhou, H.W., and Nguyen, B.Y. (2003) Structure and dielectric properties of amorphous $LaAlO_3$ and $LaAlO_xN_y$ films as alternative gate dielectric materials. *J. Appl. Phys.*, **94**, 1229.

59 Chen, J.H. Jr., Bojarczuk, N.A., Shang, H., Copel, M., and Hannon, J.B. (2004) Ultrathin Al and HfO_2 gate dielectrics on surface-nitrided Ge. *IEEE Trans. Electron. Dev.*, **51**, 1441.

60 Frank, M.M., Wilk, G.D., Starodub, D., Gustafsson, T., Garfunkel, E., Chabal, Y.J., Grazul, J., and Muller, D.A. (2005) HfO_2 and Al_2O_3 gate dielectrics on GaAs grown by atomic layer deposition. *Appl. Phys. Lett.*, **86**, 152904.

61 Hobbs, C.C., Fonseca, L.R.C., Knizhnik, A., Dhandapani, V.,

Anderson, S.G.H., and Tobin, P.J. (2004) Fermi-level pinning at the polysilicon/metal oxide interface. *IEEE Trans. Electron. Dev.*, **51**, 971.

62 De, I., Jonri, D., Srivastava, A., and Osburn, C.M. (2000) Impact of gate work function on device performance at the 50nm technology node. *Solid State Electron.*, **44**, 1077.

63 Samavedam, S.B., La, L.B., Tobin, P.J., White, B., Hobbs, C., Fonseca, L.R.C., Demkov, A.A., Schaeffer, J., Luckowski, E., Martinez, A., Raymond, M., Triyoso, D., Roan, D., Dhandapani, V., Garcia, R., Anderson, S.G.H., Moore, K., Tseng, H.H., Capasso, C., Adetutu, O., Gilmer, D.C., Taylor, W.J., Hegde, R., and Grant, J. (2003). Fermi-level pinning with sub-monolayer MeO_x and metal gates. IEDM Technical Digest, p. 307.

64 Yu, H.Y., Ren, C., Yeo, Y.C., Kang, J.F., Wang, X.P., Ma, H.H.H., Li, M.F., Chan, D.S.H., and Kwong, A.L. (2004) Fermi pinning-induced thermal instability of metal-gate work functions. *IEEE Electron. Device Lett.*, **25**, 337.

65 Cartier, E., McFeely, F.R., Narayanan, V., Jamison, P., Linder, B.P., Copel, M., Paruchuri, V.K., Basker, V.S., Haight, R., Lim, D., Carruthers, R., Shaw, T., Steen, M., Sleight, J., Rubino, J., Deligianni, H., Guha, S., Jammy, R., and Shahidi, G. (2005). Role of oxygen vacancies in V-FB/V-t stability of pFET metals on HfO_2. Symposium on VLSI Technology Digest, p. 230.

66 Schaeffer, J.K., Fonseca, L.R.C., Samavedam, S.B., Liang, Y., Tobin, P.J., and White, B.E. (2004) Contributions to the effective work function of platinum on hafnium dioxide. *Appl. Phys. Lett.*, **85**, 1826.

67 Liang, Y., Curless, J., Tracy, C.J., Gilmer, D.C., Schaeffer, J.K., Triyoso, D.H., and Tobin, P.J. (2006) Interface dipole and effective work function of Re in Re/HfO_2/SiO_x/n-Si gate stack. *Appl. Phys. Lett.*, **88**, 072907.

68 Copel, M., Pezzi, R.P., Neumayer, D., and Jamison, P. (2006) Reduction of hafnium oxide and hafnium silicate by rhenium and platinum. *Appl. Phys. Lett.*, **88**, 072914.

69 Pantisano, L., Schram, T., Li, Z., Lisoni, J.G., Pourtois, G., Gendt, S.D., Brunco, D.P., Akheyar, A., Afanas'ev, V.V., Shamuilia, S., and Stesmans, A. (2006) Ruthenium gate electrodes on SiO_2 and HfO_2: sensitivity to hydrogen and oxygen ambients. *Appl. Phys. Lett.*, **88**, 243514.

70 Alshareef, H.N., Wen, H.C., Luan, H., Choi, K., Harris, H.R., Senzaki, Y., Majhi, P., Lee, B.H., Foran, B., and Lian, G. (2006) Temperature dependence of the work function of ruthenium-based gate electrodes. *Thin Solid Films*, **515**, 1294.

71 Wen, H.C., Majhi, P., Choi, K., Park, C.S., Alshareef, H.N., Harris, H.R., Luan, H., Niimi, H., Park, H.B., Bersuker, G., Lysaght, P.S., Kwong, D.L., Song, S.C., Lee, B.H., and Jammy, R. (2008). Decoupling the Fermi-level pinning effect and intrinsic limitations on p-type effective work function metal electrodes. *Microelectron. Eng.*, **85**, 2.

72 Yeo, Y.-C., Ranade, P., King, T.-J., and Hu, C. (2002) Effects of high-*k* gate dielectric materials on metal and silicon gate workfunctions. *IEEE Electron. Device Lett.*, **23**, 342.

73 Lee, B.H., Oh, J., Tseng, H.H., Jammy, R., and Huff, H. (2006) Gate stack technology for nanoscale devices. *Mater. Today*, **9**, 32.

74 Tavel, B., Skotnicki, T., Pares, G., Carriere, N., Rivoire, M., Leverd, F., Julien, C., Torres, J., and Pantel, R. (2001) Totally silicided ($CoSi_2$) polysilicon: a novel approach to very low resistive gate without metal CMP or etching. IEDM Technical Digest, p. 825.

75 Kang, C.Y., Lysaght, P., Choi, R., Lee, B.H., Rhee, S.J., Choi, C.H., Akbar, M.S., and Lee, J.C. (2005) Nickel-silicide phase effects on flatband voltage shift and equivalent oxide thickness decrease of hafnium silicon oxynitride metal-silicon-oxide capacitors. *Appl. Phys. Lett.*, **86**, 222906.

76 Park, C.S., Cho, B.J., and Kwong, D.L. (2004) Thermal stable fully

silicided Hf-silicide metal-gate electrode. *IEEE Electron. Device Lett.*, **25**, 372.

77 Nabatame, T., Kadoshima, M., Iwamoto, K., Mise, N., Migita, S., and Ohno, M. (2004) Partial silicides technology for tunable work function electrodes on high-*k* gate dielectrics: Fermi level pinning controlled $PtSi_x$ for HfO_x (N) pMOSFET. IEDM Technical Digest, p. 83.

78 Kedzierski, J. (2003) Issues in NiSi-gated FDSOI device integration. IEDM Technical Digest, p. 441.

79 Nowak, E.J. (2002) Maintaining the benefits of CMOS scaling when scaling bogs down. *IBM J. Res. Dev.*, **46**, 169.

80 Goser, K. (2004) *Nanoelectronics and Nanosystems: From Transistor to Molecular and Quantum Devices*, Springer.

81 Zhimov, V.V. (2005) Emerging research logic devices. *IEEE Circuits Devices Mag.*, **21**, 37.