Scaling the gate dielectric: Materials, integration, and reliability

by D. A. Buchanan

In this paper a review of the more “fundamental” concerns regarding the scaling of the gate dielectric in the ultrathin regime is presented. Material issues are discussed pertaining to the integration of silicon oxynitride and oxide/nitride stacked layers and how such films might reduce or minimize boron penetration problems and address leakage current and reliability concerns. A methodology is presented to calculate device and chip lifetimes for MOS structures on the basis of data extracted from voltage- and temperature-accelerated measurements. Some integration issues regarding higher-k materials are also discussed because of their ability to solve the scaling problem. However, difficulties are involved in integrating them into a CMOS process flow.

Introduction

The fundamental limits of metal–oxide–semiconductor (MOS) technology have been discussed, reviewed, and claimed to be at hand since the first MOS processes were developed more than 25 years ago [1–4]. The integration of semiconductor devices has gone through medium (MSI), large (LSI), very large (VLSI), and now ultralarge-scale integration (ULSI). At each stage of evolution, limits (some more fundamental than others) were reached and then subsequently surpassed.

One of the most fundamental restrictions to scaling of MOS structures has been photolithography, so much so that now each of the new generations is described by a lithographic dimension [5]. At times it was believed that optical lithography would eventually reach its limit [6, 7], yet the SIA Roadmap [5] suggests that 130-nm-deep UV optical lithography should be available to produce 0.1-μm devices. If and when the limit for optical lithography is actually surpassed, X-ray and e-beam may be introduced into CMOS manufacturing. 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 the device physical dimensions, be it a length, width, depth, or thickness, approaches the dimensions of a few silicon atoms. Manufacturing tolerances, and therefore economics, may dictate an end to the scaling of silicon devices before these fundamental limits are reached. Therefore, in this review of scaling of gate dielectrics for MOS devices, only currently perceived fundamental limits are considered.

Copyright 1999 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the Journal reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to republish any other portion of this paper must be obtained from the Editor.
Silicon dioxide has been used as the primary gate-dielectric material in field-effect devices since 1957, when the usefulness of the Si/SiO$_2$ material system was first demonstrated [1–4]. At first single devices and then integrated devices were made, and the thickness of SiO$_2$ films decreased with each generation. For currently manufactured high-performance processors, the SiO$_2$ thickness is less than 40 Å.

To demonstrate how these dimensions will continue to be reduced, in Table 1 the time line for the reduction lithography and the equivalent oxide thickness is given [5].

The phrase equivalent oxide thickness, $t_{eq}$, refers to the thickness of any dielectric scaled by the ratio of its dielectric constant to that of silicon dioxide (where $\varepsilon_{oxide} = 3.9$) such that

$$t_s = t_{eq} \frac{\varepsilon_s}{\varepsilon_{oxide}},$$

where $t_{eq}$ and $t_s$ are the equivalent oxide and physical thickness, respectively, and $\varepsilon_{oxide}$ and $\varepsilon_s$ are respectively the dielectric constant of silicon dioxide and that of the other dielectric. If the gate dielectric were to remain SiO$_2$, within 15 years (assuming a constant rate of scaling) its thickness would be only 10 Å, or about three atomic layers, whereas with an alternative dielectric $x$, with $\varepsilon_x > \varepsilon_{oxide}$, a thicker layer could be used. Figure 1 shows examples of cross-sectional high-resolution transmission electron microscope (HRTEM) images for MOS structures with (a) ~27-Å and (b) ~24-Å images. In part (a) the different orientations of the polysilicon (poly-Si) grains can be seen. Both the Si/SiO$_2$ and poly-Si/SiO$_2$ interfaces are identified in part (b) and, on this atomic scale, thickness variations of ~2–3 Å can be related to the atomic silicon steps at both interfaces. The atomic spacing of a silicon crystal lattice spacing between the $^111$ planes is 3.138 Å [8–10]. Silicon dioxide is an amorphous material, and as such its atomic dimensions are not as well defined. Average Si–O bond lengths in SiO$_2$ are of the order of 1.6 Å [11], but this is dependent upon which of the local structures is assumed (e.g., $\alpha$- or $\beta$-quartz, tridymite, cristobalite) [11–23]. Nonetheless, MOS devices have been fabricated with thermally grown dielectrics with thicknesses of the order of 15 Å [24–28]. These devices were fabricated in research facilities using limited device densities, and it is yet to be seen whether such devices can be manufactured in VLSI.

In this paper, a detailed discussion of current dielectrics and those proposed for future generations is included. Current beliefs regarding the limitations of silicon dioxide as the gate dielectric are reviewed. Ultrathin oxides, which include those gate dielectrics below 40 Å, some of which are nitrided silicon oxides (silicon oxynitrides), are discussed. The benefits and limitations of alternative or

---

**Table 1** Roadmap for technology and equivalent dielectric thickness (after [5]).

<table>
<thead>
<tr>
<th>Production year</th>
<th>Minimum feature size ($\mu$m)</th>
<th>Equivalent dielectric thickness (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>0.25</td>
<td>40–50</td>
</tr>
<tr>
<td>1999</td>
<td>0.18</td>
<td>30–40</td>
</tr>
<tr>
<td>2001</td>
<td>0.15</td>
<td>20–30</td>
</tr>
<tr>
<td>2003</td>
<td>0.13</td>
<td>20–30</td>
</tr>
<tr>
<td>2006</td>
<td>0.10</td>
<td>15–20</td>
</tr>
<tr>
<td>2009</td>
<td>0.07</td>
<td>&lt;15</td>
</tr>
<tr>
<td>2012</td>
<td>0.05</td>
<td>&lt;10</td>
</tr>
</tbody>
</table>
new high-$k$ materials for possible high-performance CMOS applications are reviewed.

Silicon dioxide has been used for more than 30 years as a gate dielectric for metal–oxide–semiconductor field-effect transistors, or MOSFETs. As the SiO$_2$ gate dielectric was made thinner, new technological problems arose. These problems have often been referred to as integration issues—specifically, issues related to the manufacture of integrated circuits. These include, but are not necessarily limited to, the dielectric thickness variation, penetration of impurities from the gate (e.g., boron) into the gate dielectric, and the reliability and lifetimes of devices made with these ultrathin films. Therefore, even when working on the detailed physics of these alternative dielectrics, it is necessary to demonstrate the manufacturability of MOS integrated circuits, where a specific device lifetime is required. Integration and reliability of both old and new materials are discussed.

**Gate-dielectric materials—old and new**

It is prudent here to begin with some information regarding the limited extendibility of SiO$_2$ as the traditional gate dielectric. Devices made with gate oxides as thin as 15 Å may not be usable, with any measure of manufacturing control and reliability, for the manufacture of high-performance CMOS [24–28]. As silicon dioxide is thinned, the gate leakage current through the film increases. For silicon dioxide, at a gate bias of $\sim$1 V, the leakage current changes from $1 \times 10^{-12}$ A/cm$^2$ at $\sim$35 Å to $1 \times 10^{-12}$ A/cm$^2$ at $\sim$15 Å: twelve orders of magnitude in current for a thickness change of little more than a factor of 2 [29–31]. As one might imagine, this exponential increase in the gate-dielectric leakage current has caused significant concern regarding the operation of CMOS devices, particularly with regard to standby power dissipation, reliability, and lifetime. This exponential increase in gate leakage current with decreasing dielectric thickness will likely be one if not the major contributor leading to the limited extendibility of SiO$_2$ as the gate dielectric in the 15–20-Å regime.

Considering for the moment the manufacturing implications of producing devices having films with thicknesses of the order of a few monolayers, the thickness variation of these films over a 200- or 300-mm silicon wafer is of substantial concern. A variation in thickness of only 1 Å could result in changes in the device operating conditions (for example, the electron or hole mobility or the device transconductance), making it extremely difficult to maintain device tolerances. The

---

1 The term “high-$k$” refers to a material with a dielectric constant significantly higher than that of SiO$_2$. The Greek letter $\kappa$ is more commonly used in the semiconductor industry to represent the relative permittivity of a material. The letter $k$ is found more often in the fields of chemistry and physics. Even though the materials are referred to as high-$k$, in this paper the symbol $\kappa$ is used for relative permittivities.

---

control of a 23-Å SiO$_2$ thickness can be maintained such that variations across a 200-mm wafer can be limited to $<0.5$ Å and even $<0.1$ Å for some nitried processes (which typically have a slower dielectric growth rate) [32–39]. In Figures 2(a) and 2(b) the capacitance–voltage ($C$–$V$) and current–voltage ($I$–$V$) curves are shown for a series of $p^+$ poly-Si gate capacitors with a gate-dielectric film $\sim$23 Å thick and having a narrow thickness variation such that the standard deviation across a 200-mm wafer is $\sigma = 0.09$ Å ($\sim$0.4%). These films were grown using nitrous oxide (N$_2$O) at 850°C in a standard thermal diffusion furnace. The area of the capacitors is $10^{-4}$ cm$^2$.

---

![Figure 2](image-url)
a 200-mm wafer is \( \sigma = 0.09 \, \text{Å} \) (~0.4%). These films were grown using nitrous oxide (N\(_2\)O) in a standard thermal diffusion furnace. The area of these capacitors is \( 10^{-4} \, \text{cm}^2 \). Plotted in the figure are data from a \( 5 \times 5 \) matrix of devices distributed over a 200-mm silicon wafer. There are 25 curves in each graph! These data demonstrate quite well that with reasonable control and care of the growth conditions, it is possible to surpass most concerns regarding thickness variations of ultrathin SiO\(_2\)-based films. Therefore, some of the required manufacturing tolerances, at least at 23 Å, are attainable. The results of Figure 2 show only within-wafer uniformity. The variation from lot to lot is also an extremely important manufacturing concern that is not addressed here.

Although some encouraging results at 23 Å have been obtained, this suggests that perhaps thinner films could be grown with at least the same uniformity.

As scaling of the oxide thickness continues, it is obviously desirable to reduce the leakage current. It is also desirable to maintain or increase the reliability of such films. The introduction of nitrogen into silicon dioxide has been used to eliminate a number of concerns, although not all with equal success. The first introduction of nitrogen into silicon dioxide films was for much thicker films in the range of hundreds of angstroms. It was shown that the reliability of these films could be increased if the films were annealed in ammonia (NH\(_3\)) or other nitrogen-containing gas. Some studies have shown [43] that the nitrogen could reduce the “defect-generation rate” (\( P_{\text{gen}} \)) in these films. (The defect-generation rate is the density of defects produced per electron of a given energy that traverses the film, where the electron energy is determined by the applied potential or electric field.) The details of these processes of electron transport and \( P_{\text{gen}} \) are discussed in detail in another section of this paper. The incorporation of nitrogen reduced the generation of defects, and this was thought to be a result of the ability of the nitrogen to getter hydrogen or reduce its diffusion, which was believed to be a cause of defect generation [43–53]. The incorporation of too much nitrogen leads to large flat-band or threshold-voltage shifts [24, 54–58]. For SiO\(_2\), a nitrogen concentration of <1 atomic percent (at.%) close to the Si/SiO\(_2\) was found to be useful without degrading electrical properties [24, 54–58].

As more experiments were performed to tailor the nitrogen profiles in oxide films, it was found that nonhydrogen nitrogen species (e.g., N\(_2\)O and NO) were even more efficient at increasing device reliability and reducing \( P_{\text{gen}} \), since hydrogen was not introduced during the nitridation process [59, 60]. The introduction of nitrous oxide (N\(_2\)O) and later nitric oxide (NO) was found to reduce the defect-generation rate to a greater extent than ammonia (NH\(_3\)). Subsequently, it was observed that postoxidation annealing with N\(_2\)O not only incorporated nitrogen within the film, but also increased the film thickness. In general, this was not viewed as a desirable effect. Independent control of the oxide film and the nitrogen concentration and its profile was desired. The study of the gas-phase and surface chemistry of N\(_2\)O and NO oxidation and annealing was conducted primarily by groups at Motorola [61–65] and at the universities of Cornell [39, 66–70], Rutgers [37, 71–79], and Paris [80, 81]. It was found [39, 66–70] that the primary source for the incorporation of nitrogen was NO. The N\(_2\)O is first reduced to NO and oxygen radicals during the growth or anneal steps. The films produced using NO also had higher total nitrogen concentrations. Oxide films could be grown to contain 0.1–1 at.% nitrogen with N\(_2\)O growth or annealing, whereas substantially higher nitrogen concentrations are obtained with NO using similar thermal processing [37, 39, 61–73, 75–79].

Oxide films that were either grown or annealed in N\(_2\)O typically have total integrated nitrogen concentrations of less than 1 at.%. Such small concentrations of nitrogen are desirable for controlling channel hot-electron degradation effects in MOSFET characteristics [24, 58]. However, such a small nitrogen concentration is insufficient to reduce the effects of boron penetration from a p\(^+\) Si gate into the gate dielectric. The incorporation of large amounts of nitrogen leads to the reduction of boron penetration, but also causes threshold-voltage shifts, \( \Delta V_t \), and mobility and transconductance degradation which are dependent upon both the nitrogen concentration and its concentration profile. This is 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. To some extent these voltage shifts can be compensated by adjustments to the substrate doping. However, there are limits to the degree to which this charge may be compensated without significantly affecting device performance [82–84]. Threshold voltage shifts of \( \Delta V_t \leq 100 \, \text{mV} \) may be tolerated for some technologies beyond lithographic dimensions <0.25–μm CMOS [82–84]. As scaling continues and supply biases are reduced, less charge, and therefore less nitrogen, is tolerable.

It is not solely the concentration of nitrogen, but also its distribution, which leads to the aforementioned threshold voltage shifts. The relationship between the voltage shift and the charge distribution is given by

\[
\Delta V_t = \frac{1}{t_{\text{oxide}}} \int_{x=0}^{t_{\text{oxide}}} \frac{qN(x)}{\varepsilon_{\text{oxide}}} \, dx = \frac{\tilde{q}qN}{\varepsilon_{\text{oxide}}},
\]

where \( \Delta V_t \) is the threshold voltage shift, \( q \) is the electron charge, \( N(x) \) is the distribution of the charge density, \( \varepsilon_{\text{oxide}} \) and \( \tilde{\varepsilon}_{\text{oxide}} \) are the free space and relative oxide permittivities, respectively, and \( t_{\text{oxide}} \) is the oxide thickness, and
\( \bar{x} \) is the first moment or centroid of the charge distribution as measured from the gate contact. As the centroid of the nitrogen concentration and therefore the charge distribution moves away from the oxide/silicon interface, its effect upon the threshold voltage is reduced. Since only a small concentration of nitrogen is needed at the Si/SiO\(_2\) interface to reduce channel hot-electron effects, and a large concentration is needed at the poly-Si gate to reduce the effects of boron penetration, control of both the density and the distribution of the nitrogen is paramount.

**Figure 3** is a schematic illustration of the incorporation of nitrogen into SiO\(_2\) films using N\(_2\)O, NO, and O\(_2\) thermal processing [36–39, 67, 75–77, 85, 86]. It is shown in Figure 3(a) that the exposure of NO to silicon produces an oxynitride film with an approximately uniform distribution of nitrogen. At a given temperature, the oxidation of silicon with NO is essentially a self-limiting growth, since the increasingly nitrogen-rich oxide layer acts as a diffusion barrier. When N\(_2\)O is used, as in Figure 3(b), there is both incorporation and removal of nitrogen, which results in a profile peaked close to the Si/SiO\(_2\) interface. The peaked profile results from the reduction of N\(_2\)O in the gas phase to primarily NO and O radicals. The NO efficiently incorporates the nitrogen into the film as it grows, while the oxygen radicals remove nitrogen from the film, especially at its outer surface. When either N\(_2\)O or NO is used to anneal a previously grown SiO\(_2\) film [Figure 3(c)], the nitrogen concentration is peaked near the Si/SiO\(_2\) interface, with NO producing substantially higher nitrogen concentrations for a given thermal cycle. Only with the reoxidation in O\(_2\) of an oxynitride film can the peak of the nitrogen profile be displaced from the Si/SiO\(_2\) interface; this reduces interface degradation caused by channel hot electrons [87–89]. However, if too much nitrogen is incorporated at this interface, a threshold-voltage shift can occur [see Equation (2)], reducing device performance. For boron penetration, a large nitrogen concentration at the poly-Si/SiO\(_2\) interface is desired. Since any charge from the nitrogen would exist at the gate contact, little effect would occur on the threshold-voltage shift [see Equation (2)]. Therefore, a desirable nitrogen distribution profile would be one with an extremely high nitrogen concentration at the poly-Si/SiO\(_2\) interface and a very small concentration at the Si/SiO\(_2\) interface.

Thermal oxidation with O\(_2\), N\(_2\)O, and NO can tailor the nitrogen profile such that reasonable control of the nitrogen density and profile can be maintained. However, as thinner films are used, thermal techniques may no longer be able to produce sufficient nitrogen to reduce boron penetration and still maintain small enough concentrations at the Si/SiO\(_2\) interface to control the channel hot-electron degradation without substantially affecting device performance. In the limit, a silicon nitride film a monolayer or two deep would be desirable. These thermal growth and annealing techniques may be able to produce sufficiently high nitrogen concentrations far enough from the Si/SiO\(_2\) interface for films thinner than 30 Å.

Consider producing a sufficiently high nitrogen concentration at the poly-Si gate to form a continuous monolayer or two of silicon nitride. This is possible with a plasma technique in which the high concentration of nitrogen ions, excited molecules, and radicals in the gas phase produce such a nitride film [59, 60]. In actuality, the plasma processing produces a heavily nitrided oxide film instead of a monolayer of nitride. If the nitrogen concentration at or near the poly-Si/SiO\(_2\) is high enough, this top layer might be considered an oxidized nitride rather than a nitrided oxide. The monolayer(s) may be enough not only to reduce or eliminate the boron penetration, but also to increase the film’s permittivity [see Equation (1)]. With adjustments to the material’s physical and equivalent thickness (\(t_\text{eq}\)) with the appropriate nitrogen concentrations, it is possible, at least in theory, to reduce the gate leakage current [59, 60, 90, 91].
The use of oxide/nitride stacks can eliminate boron penetration if the nitride film is of sufficient thickness. 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 ($t_{eq}$) [see Equation (1)], since the dielectric constant of silicon nitride is approximately twice that of silicon dioxide. Therefore, with the addition of silicon nitride to the stack, an increase in the effective thickness of approximately 30% is possible. A silicon nitride film is an excellent boron diffusion barrier, even with thermal processing up to temperatures of >1000°C for a few seconds. Silicon nitride should also be able to reduce the gate leakage current, at least in theory [59, 60, 90, 91]. 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 [40, 43, 92, 93], 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 [43–47, 50, 52, 53, 94, 95]. At present, the most promising nitride films produced are those using a remote plasma deposition technique [57, 60, 90, 91, 96–99].

These techniques for film deposition are not commonly used within the semiconductor manufacturing industry, which introduces questions as to their quality, uniformity, and reproducibility. The use of Si$_3$N$_4$ as part (or all) of a gate dielectric is in its infancy compared to other work on nitrided SiO$_2$ films. Despite some claims to the contrary [59, 60, 90, 91], a gate-quality nitride is still not available for manufacturing. Only time will tell which if any silicon nitride film will be used as a gate dielectric in a single layer or in a stacked dielectric structure with a silicon dioxide interfacial layer.

### Boron penetration

The use of p$^+$ gates for CMOS processes was first introduced to reduce short-channel effects and lower threshold voltages as the devices were pushed into the submicrometer regime [6]. However, with the use of boron as the dopant for the p$^+$ gates, dopant diffusion and its subsequent penetration into the gate dielectric have become a problem [24, 29, 55, 57, 58, 69, 70, 100–111]. The penetration of the boron into the gate dielectric causes a number of problems not only with the quality of the dielectric but especially with the device operation [24, 55–58]. Boron penetration shifts the threshold voltage of MOS devices to more positive values. Degradation of the MOSFET transconductance and the subthreshold slope is also correlated with boron penetration [24, 55–58].

Typically boron is implanted into the poly-Si gate using sufficiently high doses to ensure reasonable conductance of the poly-Si gate. Unfortunately, it is also necessary to activate the dopant with a high-temperature anneal, typically in the range of 950–1050°C in an inert ambient for a few seconds. Boron is an extremely small atom, and as such has a very high diffusion coefficient in both silicon and silicon dioxide at such elevated temperatures. During the high-temperature activation annealing, the boron penetrates into and through the gate dielectric. This penetration can be observed and measured using a number of techniques.

![Figure 4](image-url)

**Figure 4**

Boron distributions, as measured by secondary ion mass spectrometry (SIMS), in the polysilicon gate. The four traces show the boron profiles for four annealing conditions at a temperature of 1000°C for 1, 2, 5, and 10 seconds. The boron was implanted into a 1500-Å polysilicon gate. Part (a) shows the raw SIMS data of the four conditions prior to any normalization. Using the minima in the silicon peak as a reference, the boron distributions were normalized, as shown in part (b).
In Figure 4(a), the boron concentration in the poly-Si gate, the 30-Å gate dielectric, and the silicon substrate are shown. These profiles are measured using secondary ion mass spectroscopy (SIMS). Four profiles are shown demonstrating the diffusion of the boron under annealing conditions at a temperature of 1000°C for 1, 2, 5, and 10 seconds. It is easy to see from Figure 4(a) that with increasing time, significant diffusion of boron occurs, resulting in a buildup of boron at the poly-Si/SiO₂ interface. Note that these profiles are shown on a linear concentration scale, although they are normally plotted on a logarithmic axis, to show in some detail the movement of the boron from the poly-Si gate into the gate dielectric and crystalline Si substrate. Because these SIMS data were taken from different devices, the alignment along the thickness axis is not exact. For each annealing condition shown in Figure 4(a), the silicon signal is also shown. The minima of the silicon signals are approximately in the middle of the gate oxide. Using the silicon minima of each of the traces, the boron profiles were aligned to one another; these realigned profiles are shown in Figure 4(b).

With increasing time, the boron in the bulk of the poly-Si decreases, while the boron at the poly-Si/SiO₂ interface increases. The spatial resolution of the SIMS measurements is not high enough to determine whether the increasing boron is actually at the interface or in the oxide itself. However, one might speculate that this peak results from a defect-related trapping of boron.

As the boron penetrates into the oxide, a number of measurable effects can be noted. In Figure 5, capacitance–voltage curves are shown for a series of p⁺ boron-implanted poly-Si gate capacitors. These 10⁻⁴-cm² capacitors had a poly-Si gate 1500 Å thick and a 24-Å dry-O₂-grown silicon dioxide film on an n-type substrate. The C–V curves are shown to shift to higher, more positive values as the annealing time is increased. This C–V or flat-band voltage shift (ΔV₉₅), which is similar in magnitude to ΔV₉₅ of the MOSFET, degrades device performance. The flat-band voltage shift (or ΔV₉₅) can be used as a relative measure of the amount of boron.

Boron penetration also affects the quality of the oxide. In the legend of Figure 5, ΔV₉₅ and the yield percentage are shown. The yield decreases with increasing boron penetration, which is coincident with increases in ΔV₉₅. The increase in ΔV₉₅ and decrease in device yield are shown more clearly in Figure 6. The yield represents the number of good devices, as measured by their current–voltage characteristics. A failed device was one in which the current density had risen at least two orders of magnitude relative to the characteristic of the device annealed for the shortest time. Even small amounts of boron penetration can have significant effects on device characteristics. The reduction in yield resulting from boron penetration could become a substantial impediment to the manufacture of future CMOS devices, especially in light of the decreasing dielectric layer thickness.
As mentioned previously, nitrogen incorporated into an oxide film may be used to retard the effects of boron penetration. In Figure 7 a series of $C-V$ curves is shown for dielectrics grown in $N_2O$. This reduction in $\Delta V_{FB}$, compared to that shown for a non-nitrided oxide in Figure 5, demonstrates that boron penetration has been significantly reduced. Figure 8 shows the effects of boron penetration for oxide films with and without nitrogen. Even a small amount of nitrogen, such as that present when the oxide is grown in $N_2O$, can cause a substantial reduction in boron penetration, which can have a major effect on the final device properties.

It is desirable that all of the ions implanted into a poly-Si film be ionized. With continued scaling of MOSFET dimensions, higher active dopant concentrations are required. To compensate for the effect of incomplete ionization in the poly-Si gate, it is desirable to increase the dopant implant dose. For a given anneal condition, the increased dose will also increase the penetration of the dopants into the gate dielectric. It has been suggested that as little as 50% of the implanted dopant becomes electrically active after annealing [29–31, 112]. This is in part due to the limited thermal budgets that must be maintained to restrict the diffusion of other dopants, especially for those devices with the shallow junctions. Increasing the thermal cycle will, of course, also increase the boron penetration. This problem thus becomes one of design and optimization. The highest active dopant concentration is required while maintaining a minimum of boron penetration with the lowest possible thermal cycle.

Since the dopant implantations currently used for the poly-Si gate are quite close to the solid solubility limits of silicon, an increase in doping levels may not be possible [113–115]. To increase solid solubility, a gate alloy containing silicon and germanium could be used [113–115], but this leads to another optimization problem. If the germanium concentration within the gate is too large, the electron affinity of the gate changes. A change in the electron affinity of the gate also adjusts the threshold voltage, and therefore the device performance, of a MOSFET. Too little germanium will not increase the solid solubility limit to any significant degree.

If the poly-Si gate has not been implanted to a high enough concentration, or if the dopant that was implanted is not sufficiently activated, a significant voltage is dropped across the gate; this effect is known as polysilicon depletion. When the active carrier concentration in the poly-Si is not high enough to pin the Fermi level at the poly-Si/SiO$_2$ interface, the band bending in the poly-Si becomes voltage-dependent. As the device is biased such that the silicon substrate is inverted and a channel is formed, the poly-Si gate becomes depleted of free carriers. Polysilicon depletion, caused by an insufficient electrically active doping concentration within the gate, affects device performance. As the poly-Si is driven into depletion, part of the applied voltage is dropped across the gate, reducing the field at the Si/SiO$_2$ interface and decreasing the
channel carrier concentration. The drive current of the MOSFET, which is proportional to the carrier concentration, is therefore reduced as well. The reduction of the on current has a direct impact on the device switching speed. Practical CMOS circuits can be quite complicated, but the basic CMOS building block is the inverter, containing one p-FET and one n-FET. If, for example, the inverter drives just one other device (although in most circuits it would drive more than just one), it has a reduced on current owing to the polysilicon depletion effect. The driven inverter has a reduced capacitive load resulting from the series capacitance of its partially depleted poly-Si gate. And although polysilicon depletion is an undesirable effect, for certain types of circuits the loss of drive current may not represent an overwhelming degradation of device properties. It has been suggested that for advanced sub-0.1-μm CMOS, polysilicon depletion represents only a 10–20% loss in performance, which, though undesirable, is tolerable [84, 116].

Alternative (high-k) dielectrics
One of the more fundamental limits to the scaling of the gate dielectric is the exponential increase in tunnel current with decreasing film thickness. For films as thin as 20 Å, leakage currents can rise to 1–10 A/cm² [29, 112]. This incredibly high current can conceivably alter device performance, to say nothing of the difficulties associated with dissipating 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 of gate leakage current is an important reason, if not the primary one, for replacing SiO₂-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, is exponentially dependent upon the dielectric thickness, while the capacitance is only linearly dependent 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 is exponentially dependent upon the barrier height. Therefore, not only is a material with higher dielectric constant required, but this material must also 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 silicon dioxide or silicon nitride as a possible gate dielectric. The most common of these are the simple metal oxides such as Ta₂O₅ [117–128] and TiO₂ [129–133]. Other materials have been suggested, including Y₂O₃ [134–138], CeO₂ [139–142], ZrO₂ [143, 144], HfO₂ [145, 146], and Al₂O₃ [147]. Ferroelectric materials such as BST (barium strontium titanate) [130, 148, 149] have also been suggested as gate dielectrics. Unfortunately, a number of these materials, specifically Ta₂O₅, TiO₂, and BST [121, 122, 127], are not thermally stable on silicon. The formation of SiO₂ and/or metal silicides often occurs when these materials are deposited upon silicon. Further growth of silicon dioxide or a silicide takes place during subsequent annealing, which is usually found to be necessary to reduce leakage currents [124, 127, 128, 131, 132, 150–153]. An underlying SiO₂ layer reduces the effectiveness of any high-k material, since SiO₂ has a lower dielectric constant and thus reduces the effective capacitance of the film. For example, if a film of thickness $t_{eq} \sim 20$ Å is required and a substantial proportion of that is SiO₂ (for example, 10 Å), this leaves only $t_{eq} \sim 10$ Å for the high-k material. As the thickness of the high-k material decreases, the leakage current through the film increases substantially. Therefore, any underlying film with a dielectric constant less than that of the high-k film is not desirable.

Oxygen diffusion barriers may also be used to reduce oxidation of the silicon, but at the expense of the gate capacitance. Some other metal–oxide alloys are much more stable on silicon but have lower dielectric constants, thus reducing the thickness and increasing the gate current. For most of these high-k materials, the bandgap is inversely proportional to the dielectric constant [132]. As the dielectric constant increases, the bandgap decreases. Thus, even if there is a reduction of the leakage current due to the increased thickness, this may be nullified by the reduction of the barrier height. Since both the thickness and barrier height have an exponential influence upon the leakage current, it is an optimization problem to discover which will have the larger effect.

Note that there is not necessarily a direct correlation of the barrier height or band offset between a material and a cathode contact to the bandgap of the material, although any material with a large bandgap yields a greater chance of a large band offset. Therefore, one must be somewhat careful in assuming that a material with a high bandgap will have a reduced gate leakage current.

Let us assume for the moment that we now have a material with a reasonably high $k$ that reduces the leakage current for a given equivalent oxide thickness. There is still the problem of integrating this material into ever
more complicated device processing. The easiest integration would occur if this material could simply be substituted for SiO$_2$ in a given CMOS process. Unfortunately, most of the materials studied are not sufficiently stable thermally. They can change phase (and therefore their material properties) at much lower temperatures than that needed for normal CMOS processing. It is also not clear that any of these materials could be used with poly-Si gates because of the materials’ chemical reactivity with the poly-Si. If a suitable barrier layer to reduce reaction with the poly-Si could not be found, at least one if not two gate materials would have to be metals. Since most of the materials would not be compatible with the high-temperature processing required for most CMOS devices, the gate structure might have to be moved to the end of the device process instead of remaining at the beginning. Aside from the issues related to the thermal stability of these high-$k$ materials, in order to fabricate CMOS devices, other processes including the chemical and reactive ion etching (RIE) of these materials and their gates will also have to be developed. Therefore, the use of these high-$k$ materials would require not only the discovery and development of a new material, but possibly a complete re-engineering of the CMOS manufacturing process.

**Thin-oxide reliability and device lifetime**

A number of different theories have been used to describe the physics of dielectric reliability and to predict the lifetime of MOS devices. In two of the most notable of these methods, the oxide reliability, typically denoted by charge or time to breakdown ($t_{BD}$ or $Q_{BD}$) scales with either the oxide electric field or its inverse [154, 155]. These methods suggest that breakdown, at least with very small devices, is related to the “intrinsic” breakdown of the oxides, where intrinsic breakdown is not related to process-induced or other extraneous defects. For most applications, it is generally accepted that MOS devices are required to work continuously for about ten years. Understanding the behavior of oxide breakdown, particularly as a function of the applied bias and temperature, is extremely important for device lifetime prediction. Voltage and temperature are often used to accelerate the breakdown process, since waiting for ten years to measure failure would not be possible. With the use of such acceleration techniques, it is implicitly assumed that the mechanism responsible for the device degradation under the accelerated conditions is the same as that when the voltage and temperature are at the true device operating conditions.

Much work in the past has demonstrated that the creation of defects in SiO$_2$, whether they are hole traps, electron traps, or interface states, is directly related to breakdown [7, 28, 42, 44, 45, 50, 52, 156–166]. Which particular mechanism is responsible for the creation of defects and the manner in which they are responsible for breakdown are still under debate. Most researchers [7, 29, 42, 44–46, 50, 52, 94, 112, 157–161, 164–167] believe that the creation of defects is related to the passage of electrons (and/or holes) through the SiO$_2$ dielectric. It is also well accepted that the energy of the electrons has a direct effect upon the generation of such defects and specifically the rate at which such defects are produced. An understanding of electron transport is therefore a key to discovering the nature of dielectric breakdown.

In the past, many studies have been published [44, 45, 159] that suggest that there are perhaps four regimes in which the relationship between the generation of defects and the electron energy during transport is well understood. In the regime where the electron energy is greater than $\sim 9$ eV, the bandgap of SiO$_2$, impact ionization occurs. Generally this occurs only in relatively thick films and extremely high electric fields, in which case electron–hole pairs are created. The subsequent trapping of holes and electrons, and the recombination of electrons on trapped hole sites, generate defects in the form of electron traps and interface states. For a bias condition in which the electron energy is less than the SiO$_2$ bandgap, holes may still be created in the anode of a MOS device. (The anode is the poly-Si gate or the silicon substrate for positive and negative bias conditions, respectively.) Some of these holes may be injected back into the oxide film, where they can be trapped and create defects. This mechanism of creating defects is usually referred to as anode hole injection. The exact mechanism for the production of holes is still under some debate, but it is accepted that the injection of these holes back into the oxide film leads to trapped holes and the production of interface states. In this case, defects are produced by the trapped holes themselves, the recombination of electrons on the trapped holes, and electron traps created directly by the transport of the electrons. Anode hole injection usually occurs in relatively thick films ($>100$ Å) if the electric field across the oxide is greater than $\sim 5$ MV/cm. Anode hole injection can also occur in thin films independent of the oxide field, but only if the applied bias is greater than $\sim 7$–8 V.

In the ultrathin-oxide regime, the transport of the electrons is approximately ballistic. The term *ballistic* is meant to refer to the fact that electrons traverse the film without losing energy. For an applied bias, $5 < V_{\text{applied}} < 7$ V (for a poly-Si/SiO$_2$/Si structure), where the transported electrons have an energy at least 2 eV above the SiO$_2$ conduction band edge, only electrons, not holes, are present in the dielectric. These hot electrons create damage in the form of electron traps and interface states.
This mechanism of defect generation is often referred to as trap creation [44, 45]. Trap creation also occurs in the impact ionization and anode hole injection regimes, but in those regimes it is not the dominant mechanism for defect creation, at least for relatively low injected electron fluence, \(Q_{\text{inj}}\). It was originally assumed [44, 45] that below a 2-eV threshold (relative to the SiO\(_2\) conduction-band edge) electrons would no longer have sufficient energy to create significant numbers of defects. This 2-eV threshold was approximately the energy required to liberate hydrogen bonded in a SiO\(_2\) matrix. Hydrogen, whether artificially introduced or released by hot electrons, has been shown to cause damage in SiO\(_2\) [43–50, 52, 53, 94]. In this bias regime, a link between the release of hydrogen and the creation of defects within oxide films is well established [43, 50, 52, 53]. However, for ultrathin films and biases below the trap-creation threshold (\(V < 5\) V), defect generation still occurs. The rate at which these defects are created (\(P_{\text{gen}}\)) is statistical in nature, and the rate drops orders of magnitude as the bias is decreased [43, 50, 52, 53, 94, 157, 158, 161, 164, 167–170]. For thin films, even at relatively low bias conditions, there is still a finite, though small, probability of an electron producing a defect, although its energy is well below what was thought to be the threshold for trap creation. This defect generation can still occur when the electron energy is below the SiO\(_2\) conduction-band edge, where transport through the dielectric occurs by direct quantum-mechanical tunneling [159, 164, 167, 168].

In the ultrathin regime, which for the sake of argument is arbitrarily assumed to be where oxide thickness is less than 40 Å, the transport of electrons through the dielectric at normal operating biases is dictated by direct tunneling. The creation of defects in this regime is also assumed to result from the transport of the electrons through the oxide, and such creation of defects is statistical in nature [159, 164, 167, 168]. The generation of defects can therefore be described by a defect-generation rate, \(P_{\text{gen}}\), where the rate is the number of defects produced per electron that passes through a film of a given thickness for a given electron energy, applied bias, and temperature [43–45, 50, 52, 53, 85, 157–159, 161, 164–166, 168, 170–173].

The discovery of how \(P_{\text{gen}}\) changes as a function of the applied bias, oxide thickness, and temperature, and the statistics associated with such generation, are required for the prediction of the lifetime of MOS devices. Let us begin with the statistics of the breakdown of these ultrathin-oxide films. It has been shown that the breakdown of thin dielectrics can be described using critical failure analysis [157–159, 161, 164, 167, 168]. This is also referred to as the weak link theory, in which the breakdown or failure is determined by the weakest link in a chain. The statistical distribution relating to such critical failure analysis is often referred to as a Weibull distribution [174]. Let us also assume that a critical defect density (\(N_{\text{BD}}\)) is required to cause the breakdown of a silicon dioxide film in which such a density of defects produces a connecting path from one side of the film to the other through which electron transport can freely occur [157, 158, 161, 164, 168]. For a given defect-generation rate, whether that be as a function of time or charge, the lifetime of a device can be given [159, 160, 164, 168] by

\[
Q_{\text{BD}} = \frac{N_{\text{BD}}}{P_{\text{gen}}},
\]

where \(Q_{\text{BD}}\) is the electronic charge that passes through the film to cause breakdown, \(N_{\text{BD}}\) is the critical defect density, or the density of defects required to cause breakdown, and \(P_{\text{gen}}\) is the rate at which defects are created. This relationship has been shown to be a valid description of the breakdown of silicon dioxide in the ultrathin regime [159, 164, 168]. Since the breakdown of thin oxides is statistical, the distribution of the breakdown events as a function of either time or charge must be determined. Critical failure analysis or Weibull statistics can be used to describe such a distribution. This distribution may be characterized by an average \(Q_{\text{BD}}\) value (i.e., at 50% of the distribution) with a slope \(\beta\) [157–161, 164, 168].

The critical defect density (\(N_{\text{BD}}\)) is dependent upon the thickness of the oxide film but is relatively independent of the applied bias [157–161, 164, 168]. Within the context of the relationship between \(N_{\text{BD}}\) and \(Q_{\text{BD}}\), some models suggest that breakdown occurs when the defect density is great enough to cause a connecting path between anode and cathode. As the gate dielectric becomes thinner, fewer defects are required to create such a path. Therefore, \(N_{\text{BD}}\) decreases as a function of decreasing oxide thickness. From the statistics, \(N_{\text{BD}}\) continues to decrease with decreasing dielectric thickness until only one defect is required to create this leakage path. At this point, \(N_{\text{BD}}\) remains relatively constant. Since \(N_{\text{BD}}\) is at best weakly dependent on the applied bias, it has been suggested that \(N_{\text{BD}}\) is an “intrinsic” property of the dielectric.

DiMaria and Stathis [159, 164, 168] showed that for 50-Å oxides the created defect density prior to breakdown, measured in this case as an electron-trap density, is a strong function of the injected charge and the applied bias. In a similar experiment with 30-Å films, they showed a similar relationship, but used the relative change in the stress-induced leakage current (SILC), \(\Delta J/J_0\), as a measure of the increasing defect density during stressing. The values of \(\Delta J/J_0\) were calibrated to electron-trap and interface-state generation in films whose thickness overlapped both thickness regimes. Schematically, the
The relative change in the SILC, $\Delta J/J_0$, is useful as a description of the generated defects within the oxide film and has been calibrated to the more traditional measurement of either electron traps or interface states [159, 164, 168, 175]. The slope of a plot showing $\Delta J/J_0$ as a function of either charge or time is the defect-generation rate $P_{\text{gen}}$. An example of $\Delta J/J_0$ as a function of total injected charge is shown in Figure 10. The data shown are from a capacitor with an $\sim 25$-Å N$_2$O-grown silicon dioxide film, stressed at $-4.0$ V. The three curves represent the change in the stress-induced leakage current measured at $-1.5$, $-2.0$, and $-2.5$ V. Note that the slope of each of the curves is approximately the same, with the only difference being the absolute value of $\Delta J/J_0$. In previously published works [112, 158, 159, 161, 164, 167, 168, 175], it has been established that $\Delta J/J_0$ at breakdown is relatively independent of the applied stress bias and represents a measure of $N_{BD}$. As the SILC bias ($V_{\text{sense}}$) is changed, the position of the Fermi level at the Si/SiO$_2$ also changes. The value of $V_{\text{sense}}$ dictates the energy of the tunneling electrons; therefore, $\Delta J/J_0$ as a function of $V_{\text{sense}}$ is a measure of the energy distribution of the generated defects.

Defect generation is the density of defects created per injected electron. Therefore, the defect-generation rate may now be given by

$$P_{\text{gen}} = \frac{\Delta J}{J_0 \cdot Q_{\text{inj}}}$$

(4)

where $Q_{\text{inj}}$ is the integrated injected electron fluence. The slope of the curves shown in Figure 10 represents the defect-generation rate. As demonstrated schematically in Figure 9, $P_{\text{gen}}$ (or slope) is a strong function of the applied bias. In this ultrathin regime, $P_{\text{gen}}$ is at best a weak function of the film thickness. For silicon dioxide films thinner than 40 Å, $P_{\text{gen}}$ values have been observed from as high as 100 at a 5-V bias to $\sim 10^{-10}$ at a bias of 2.0 V. This represents a change of 12 orders of magnitude in $P_{\text{gen}}$ for only a 2.5 $\times$ change in the bias [159, 164, 168]. A plot of $P_{\text{gen}}$ as a function of the applied bias can be extrapolated to the operating voltage, yielding $P_{\text{gen}}$ at the operating bias. From the values of $P_{\text{gen}}$ and $N_{BD}$, $Q_{BD}$ can be determined by using Equation (2).

Figure 11 shows an example of a typical Weibull distribution of $Q_{\text{inj}}$ and $t_{\text{BD}}$ for a series of capacitors. The Weibull function is given by $\ln[-\ln(1 - F)]$, where $F$ is the cumulative probability of failure. The capacitors shown in Figure 11 had an area of $10^{-4}$ cm$^2$, a dielectric thickness of $\sim 25$ Å, an n$^+$ poly-Si gate, and a p-type substrate. The devices were stressed using a constant voltage starting with voltage pulses of 3-ms duration. The data shown in Figure 11 are typical of sub-30-Å oxides and oxynitrides. The $Q_{BD}$ value for these capacitors at the mean of the distribution (i.e., 50%) was $\sim 6400$ C/cm$^2$, and
the slope of the distribution yields $\beta \approx 1$. The importance of this $Q_{BD}$ average and the slope of the distribution are discussed below.

In Figure 12 Weibull distributions of an ~26-Å film that was grown in O$_2$ and annealed in NO are shown for three different stress voltage conditions. These $Q_{BD}$ distributions are a strong function of the applied stress bias, changing three orders of magnitude in charge for a 0.5-V change in the stress voltage. In Figure 13, the 50% values of these distributions (and others) are shown as a function of applied bias. To understand the voltage-acceleration effects upon $Q_{BD}$, data such as those shown in Figures 12 and 13 are required for lifetime prediction.

For industrial applications, the device or chip lifetime is normally specified by a maximum acceptable probability of failure for given operating conditions for a certain length of time, the device or chip lifetime. For example, a specification for sub-0.25-μm CMOS could be described as a maximum of one hundred parts per million fails at 1.25 V and 125°C for ten years. To establish whether a certain technology will pass given a lifetime specification, these requirements can be reworded to establish the maximum applied bias or minimum oxide thickness needed to meet the failure criteria.

Two more pieces of this lifetime extrapolation puzzle are required to establish whether a given reliability specification will be met: the statistical distribution of the charge to breakdown for a series of devices of a given size and thickness, and the total gate-dielectric area of a chip. The statistical distribution is important for a number of reasons. First, it allows an extrapolation from measured results that are usually in the range of a few parts per hundred, to the reliability specification at ~100 parts per
million. The Weibull distribution can be described by the absolute value of \( Q_{\text{BD}} \), at one point and the slope, \( \beta \). The value of \( \beta \) is also important when scaling the statistical distribution from the area of one device to the area of the whole chip, or, for that matter, between devices with different areas. For thicker films, where \( \beta > 1 \), the area effect is much smaller than it is for thinner films, where \( \beta \sim 1 \). The charge to breakdown, \( Q_{\text{BD}} \), and area are related \([157–159, 161, 164, 168]\) by
\[
\frac{Q_{\text{BD}}}{Q_{\text{BD}}} = \left[ \frac{A}{A} \right]^{1/\beta},
\]
where \( A \) is the device area and \( \beta \) is the slope of the Weibull distribution.

Finally, the leakage current at specific operating or lifetime conditions must be known. The time to breakdown \( t_{\text{BD}} \) (from which a lifetime may be calculated) is given \([157–159, 161, 164, 168]\) by
\[
t_{\text{BD}} = \frac{Q_{\text{BD}}}{J} = \frac{N_{\text{BD}}}{J \times P_{\text{gen}}},
\]
where \( J \) is the current density.

From the previously mentioned methods and equations, a room-temperature lifetime \( t_{\text{BD}} \) may be calculated. The \( Q_{\text{BD}} \) at a specific failure rate (e.g., \( 10^{-4} \) being 100 parts per million) is determined by extrapolating the \( Q_{\text{BD}} \) distribution to the \( Q_{\text{BD}} \) failure rate \( [Q_{\text{BD}}]_t \). This value of \( Q_{\text{BD}} \) at the specified failure rate \( [Q_{\text{BD}}]_t \) may be extrapolated to the gate dielectric area of the entire chip by
\[
[Q_{\text{BD}}]_{\text{chip}} = [Q_{\text{BD}}]_t \times \left[ \frac{A_{\text{chip}}}{A_{\text{device}}} \right]^{1/\beta},
\]
where \( [Q_{\text{BD}}]_{\text{chip}} \) is the \( Q_{\text{BD}} \) value at a specific failure rate for the chip.

From the relationship between \( Q_{\text{BD}} \) and \( V_\phi \) (i.e., the gate voltage acceleration) that was discussed earlier in the paper, one may then extrapolate to the operating voltage. From this extrapolation, a final value of \( [Q_{\text{BD}}]_{\text{chip}} [V_\phi] \) may be determined, and a chip lifetime (for a given set of operating conditions and device design) may be determined from
\[
t_{\text{lifetime}} = \frac{[Q_{\text{BD}}]_{\text{chip}} [V_\phi]}{[J][V_\phi]},
\]
where \( J \) is the gate leakage current at the prescribed operating conditions and \( [Q_{\text{BD}}]_{\text{chip}} [V_\phi] \) is the failure rate for the entire chip gate area at a specified gate bias \( V_\phi \).

The temperature during normal operating conditions and during preconditioning or “burn-in” may also have a substantial effect upon the \( Q_{\text{BD}} \) and lifetime of a device. In the extreme, burn-in conditions may be specified at temperatures greater than 100°C and voltages greater than the operating bias. For thicker oxide films, the acceleration of \( Q_{\text{BD}} \) as a function of temperature was found to be a thermally activated process, with an activation energy as low as 0.1 eV \([176]\). For thinner films, it is not clear whether this process is still thermally activated, although some have stated that it should be related to the activation of the electron-induced release of hydrogen and its subsequent diffusion within the thin oxide film \([160, 164, 167, 168]\). Using an activation energy of 0.1 eV, \( Q_{\text{BD}} \) would be reduced by a factor of only about 2.5× from room temperature to 125°C. Others have suggested that the temperature-acceleration factor is significantly higher and could be in the range 10× to 100× \([160, 164, 167, 168]\). For the thinnest films of \( \text{SiO}_2 \), no data are available in the literature that explicitly and independently give this acceleration factor with any accuracy. Especially for thin films under low bias conditions, the temperature dependence of the lifetime could be controlled by the temperature dependence of the electron source distribution at the cathode of a MOS device. As the temperature is increased, the \( Q_{\text{BD}} \) could decrease because of an increase in \( P_{\text{gen}} \) or a decrease in \( N_{\text{BD}} \). Since \( N_{\text{BD}} \) appears to be an intrinsic property of these thin dielectric films (for a given thickness), it is more likely that this temperature dependence arises from a change in \( P_{\text{gen}} \). However, as temperature is increased, the distribution of the electron energies at the cathode/insulator interface also changes. Therefore, it is not clear which of these effects dominate the temperature acceleration of \( Q_{\text{BD}} \) \([177]\).

With careful measurements of the dependence of \( Q_{\text{BD}} \) (and its statistical variations) upon the physically independent variables (applied bias, oxide thickness, temperature, etc.), a lifetime may be calculated with the method described above. Implicit within this method is the assumption that acceleration of \( Q_{\text{BD}} \) upon these variables does not change over the regime of extrapolation. The best data at present \([160, 164, 167, 168]\) extend down to a thickness of \( \sim 15 \) Å and voltages as low as 2.0 V. From these data, for a specific failure rate, it appears that \( \text{SiO}_2 \) may cease to be useful at about 25 Å because of dielectric breakdown \([168]\). This is significantly thicker than the usually accepted range of 15–20 Å \([29–31, 83, 84, 112]\), where the gate leakage current is the only criterion.

**Conclusions**

Continued scaling of the gate dielectric has precipitated the need for a greater and more detailed understanding of the issues pertaining to integration and reliability. In this work the more critical issues have been outlined and discussed, including the different materials that may be used as the gate dielectric of the future. As an interim
solution, nitrided and/or oxide/nitride stacks may serve to reduce boron-penetration effects, assuming that leakage current and reliability can be improved. The feasibility of such structures remains to be proven in manufacturing.

Even if an oxide/nitride combination that would allow at least some scaling of CMOS were to be found, it is not clear that the thin oxide layers within could pass the reliability criteria needed to produce reasonable devices. Some experimental results suggest that this too may be of concern for films thinner than 25 Å [168]. Although a detailed description of one technique for estimating the lifetime of MOS devices is presented here, the validity of the extrapolations from voltage and temperature acceleration down to the lower voltage and temperature remains to be fully demonstrated [161, 168].

High-$k$ dielectrics are being studied as alternatives to SiO$_2$, which may permit further MOS scaling. The integration of such dielectrics into current device processing may be an extremely difficult task. The easiest path would involve a simple substitution of the higher-$k$ film for the silicon-based one. Unfortunately, most of the current materials, of which Ta$_2$O$_5$, TiO$_2$, and BST seem to be the forerunners (according to DRAM experience), are not stable on silicon and tend to form an underlying SiO$_2$ film which reduces their effective capacitance. This path is almost self-defeating unless suitable barrier layer(s) for both interfaces can be found. If these new gate materials are not simply substituted for the SiO$_2$ film, metal gates must be used in place of the doped poly-Si. However, since most designers agree that a mid-gap metal such as tungsten or molybdenum would yield a threshold voltage that is too high, not one but two metals will be required to manufacture CMOS devices. If metals are reintroduced into a MOS process, the formation of the gate stack must be moved to the end of the process for thermal reasons. This would require a redesign of the process—an extremely difficult but surmountable task. Although there may be a solution to integrating high-$k$ dielectrics into device processing, it appears that it will be anything but straightforward.

From the viewpoint of standby power, the dielectric thickness limit appears to be 15–20 Å. If the device designers tolerate higher leakage currents and power dissipation, gate-dielectric thickness may be reduced even further. However, with the reduction of gate-dielectric thickness, other processing problems will certainly arise. Boron penetration will be a primary concern unless better dielectrics (or composites) are found. Finally, the issue of dielectric reliability still remains a detractor for thinning the gate dielectric. It remains paramount that reasonable measurement and analysis techniques be developed to accurately predict the lifetime of fully integrated devices. At present, a gap exists between our understanding of the relationship between current accelerated testing techniques and device lifetime. This gap, however, is becoming smaller [168].

Acknowledgments
The author wishes to acknowledge the other members of the Advanced Gate Dielectric group, specifically P. Varekamp and E. Gusev, for their assistance in fabricating and measuring the devices and for many fruitful discussions; and S. L. Cohen for the critical reading of this manuscript. For the many discussions on oxide reliability, it is necessary to thank E. Cartier, J. Stathis, and D. DiMaria. Tutoring on the physics of quantum-mechanical tunneling received from M. V. Fischetti, F. Stern, and S.-H. Lo is greatly appreciated. The author wishes to thank R. Dennard, Y. Taur, and C. Wann for many informative sessions on device design and scaling. For the high-resolution transmission electron (HRTEM) work that produced the images in Figures 1 and 2, the author is indebted to L. Gignac and D. Klaus. Finally, a debt of gratitude goes out to the workers in the Advanced Semiconductor Technology Center (ASTC) in East Fishkill, New York, and the Advanced Semiconductor Technology Laboratory/South (ASTL/S) in Yorktown Heights, New York, for fabricating the devices without which this work would not have been possible.

References
8. Card No. 5-0565, International Centre for Diffraction Data Card File, 1601 Park Lane, Swarthmore, PA 19801.


Received September 4, 1998; accepted for publication January 26, 1999

Douglas A. Buchanan IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (buchan@watson.ibm.com). Dr. Buchanan received his B.Sc. and M.Sc. degrees in electrical engineering from the University of Manitoba, Winnipeg, Canada, in 1981 and 1982, respectively. In 1986 he received his Ph.D. from Durham University, Durham, England. Following a two-year postdoctoral fellowship at the IBM Thomas J. Watson Research Center, he spent three years in the CVD thin-films technology group in the IBM Microelectronics Division. Currently Dr. Buchanan works at the Thomas J. Watson Research Center on issues that relate to the growth, characterization, and integration of ultrathin and high dielectrics. He is a member of the American Physical Society and the Institute of Electrical and Electronics Engineers.