74. A. Bonora, "Flex-Mount Polishing of Si Wafers," *Solid State Technology*, 20, 55 (1977).
75. Y. Matsushita *et al.*, "Hydrogen Anneal of Silicon Wafer Formation of High Quality Device Active Layer," *Mtg. Abs. Electrochem. Soc.*, Spring 1998, p. 326.
76. F. Robertson and A. Allan, *Future Fab*, 1, No. 3, p. 27, 1997.
77. H.R. Huff, *Silicon Wafers for the Mesoscopic Era," Abs. Mtg. Electrochem. Soc.*, Spring 2000, Abs. 405.
78. *ibid. Ref. 13.*
79. K.V. Ravi, *Solid-State Phenomena*, 89-70, p. 103 (1999).
80. W.H. Reed, "Large Diameter Wafer Update," *Semiconductor International*, November 1994, p. 140.
81. D. Rose, "Wafer Makers Challenged by Many Aspects of 300 mm," *Solid State Technology*, February 1996, p. 79.
82. D. Anderson, "Stoking the Productivity Engine with New Materials and Larger Wafers," *Solid State Technology*, March 1997, p. 57.
83. Rose's Annual Forecast for Fab Materials, *Semiconductor International*, April 1995, p. 17.
84. Rose's Annual Forecast for Fab Materials, 1999.
85. J. Highfill *et al.*, "The Cost-Effective Challenges of 300-mm Silicon Wafers," *Solid State Technology*, October 2000, p. S-8.
86. C. Thomson and H. Mynster, "300-mm Wafers: Reclaimers Role and Challenges," *Solid State Technology*, August 1999, p. 40.

Chapter 3

GATE DIELECTRICS: THIN SILICON DIOXIDE FILMS

Thermally grown films of silicon dioxide (SiO_2) play many key roles in the operation and fabrication of silicon integrated circuits. However, the most critical application of such films in deep-submicron CMOS USLI circuits is serving as the MOSFET gate dielectric material. Although the many other roles of SiO_2 are also important (including diffusion masking, surface passivation, and functioning as a field oxide, interlevel dielectric, screen oxide, pad oxide, trench liner in STI, tunneling oxide in non-volatile memory devices, and sidewall spacer), the discussion here will focus on its role as the *gate dielectric of MOSFETs*.

It is commonly believed that crystalline-silicon is the main material responsible for the growth and success of the semiconductor industry. But others maintain that the real magic in silicon technology lies with silicon dioxide and its amazing ability to satisfy the myriad demands placed upon it as the gate dielectric of the MOSFET.

In this chapter we will explore the role of thin gate oxide films in deep-submicron MOSFETs, and address the following issues that relate to this subject:

1. The structure and role of silicon dioxide (and the Si/SiO_2 interface) as the gate dielectric of the MOSFET.
2. Dielectric breakdown and defects in SiO_2;
3. Leakage currents in silicon dioxide films;
4. Models and processes of the growth of ultra-thin oxide films;
5. Strengthening silicon dioxide films through the addition of nitrogen (oxynitrides), and the related problem boron penetration through gate oxides;
6. Technology of growing ultra-thin oxides on large diameter wafers (200-mm and 300-mm);
7. Limits of silicon dioxide as a gate dielectric material.

Despite the fact that we will restrict our discussion to SiO_2-films as gate dielectrics, it is informative to list the other IC applications in which oxide films are employed, and the oxide thickness used in such roles (see Table 3-1).

Table 3-1 RANGES OF THERMAL OXIDE THICKNESS USED IN ULSI PROCESSING

Application	SiO_2 Thickness (nm)
Gate oxide insulators for MOS devices	1.5–10.0
Tunnel oxides in EEPROMs and flash memories	6.0–10.0
Screen oxides for ion implantation	10.0–20.0
Side-wall liners for shallow trench isolation (STI)	15.0–40.0
Field oxide for LOCOS isolation	200–400
Re-oxidation of the etched gate stack side-wall damage	5.0-7.5
Inter-polysilicon dielectric	4.0–20.0
Masking film against ion implantation and diffusion	100–200
Sacrificial oxides for gate applications	6.0–15.0
Capacitor dielectric in DRAM circuits	5.0–10.0

3.1 REQUIRED CHARACTERISTICS OF GATE DIELECTRICS FOR DEEP-SUBMICRON MOSFETS

A dielectric material must satisfy a large number of important demands in order to successfully function as the gate dielectric of a MOSFET. Silicon dioxide has been used as the gate dielectric material in silicon-based MOS ICs ever since they were first introduced in the early 1970's. It continues to serve in this role, even as MOSFETs have been scaled down to the deep-submicron regime. We list the requirements of dielectric material for this application here, and provide more details about them in later sections of the chapter.

1. The most important requirement that the gate dielectric of a deep-submicron MOSFET must meet is the following: *It must be possible to continue to decrease the thickness of the oxide to the degree specified in accordance with the scaling of the gate length L_E of a MOSFET.* We should elaborate upon why being able to scale the gate oxide thickness is so fundamentally important for deep-submicron MOSFETs. Figure 3-1 shows that the ratio of the gate length to the oxide thickness (L_E/t_{ox}) has indeed remained constant at about a value of 45 for process technology at Intel for over 25 years.[1] The gate-oxide requirements for technology nodes for 0.18-μm and beyond (according to the ITRS Roadmap) are given in Table 3-2. At some point, *if SiO_2 cannot be scaled further to meet future demands, another material with higher reliability will be needed.*

As CMOS IC technology advances, the central focus of scaling the MOSFET device structure involves scaling the gate length L_E. That is, gate lengths of MOSFETs are scaled to obtain the significant benefits offered by such scaling. These benefits include: i) an increase in the drain current of the MOSFET I_D (which usually provides an improvement in the circuit

Table 3-2 Gate Oxide Equivalent Thickness for Deep-Submicron CMOS

	1999 (180 nm)	2001 (150 nm)	2003 (130 nm)	2006 (100 nm)	2009 (70 nm)	2012 (50 nm)
Equivalent Oxide thickness (nm)	3–4	2–3	2–3	1.5–2	<1.5	<1.0

speed); ii) a decrease in the gate area of the minimum-size MOSFET (which results in a reduction of the input capacitance that must be charged when logic levels are switched - and thus also improves the circuit speed); and iii) an increase in the density of devices that can be fabricated on a chip. (See Vol. 3 for more details on these issues.)

In deep-submicron MOSFETs, however, such reduction of the gate length increases the subthreshold drain current I_{Dst} of a MOSFET (which is one component of the total off-state leakage current of CMOS ICs). That is, in deep-submicron MOSFETs the short-channel effect known as *drain-induced barrier lowering* (DIBL) causes the subthreshold (off-state) drain current to rise significantly as the gate length is reduced (see Vol. 3 for more details). Two measures must be employed to prevent (or at least mitigate) this rise in the subthreshold leakage current as the gate length is

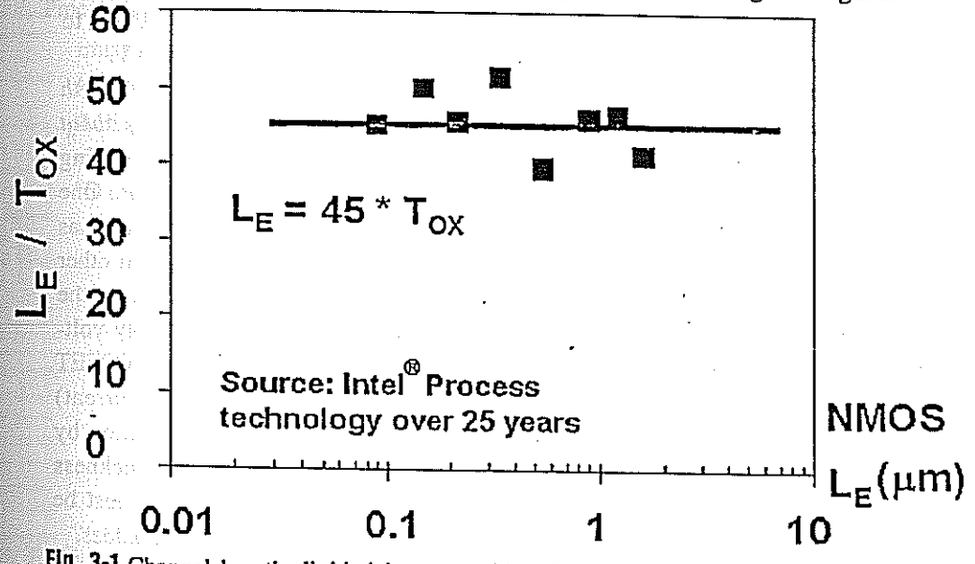

Fig. 3-1 Channel-length divided by gate-oxide thickness (L_E/t_{ox}) versus channel-length L_E for Intel's process technologies for the past 20 years.[1]

reduced. First, the junction depth X_j of the source/drain-extension-regions must be made shallower. Second, the depth of the MOSFET channel-depletion-region must also be made shallower. Both measures serve to keep the electric field of the drain from penetrating along the channel toward the source (which would worsen the DIBL effect).

The second measure (making the channel-depletion-region shallower as the gate length is shortened) is achieved by increasing the doping concentration in the channel (typically with a heavier-dose threshold-adjust-implant). While it is relatively easy to implement such an increase in the channel doping, doing so causes another undesirable effect. That is, if the gate-oxide-thickness of the MOSFET remains unchanged, the threshold-voltage V_T is increased as the doping concentration near the surface of the channel (boron) is increased (e.g., for an NMOSFET with a *p*-doped channel region). Since the threshold voltage is generally kept unchanged (or it may even be slightly *decreased*) as the CMOS technology is scaled, the increase in V_T due to the increased channel doping must be counteracted. This can only be done by reducing the gate-oxide-thickness (see Fig. 3-2). For example, as shown in Fig. 3-2, if the doping concentration of an NMOSFET with a 250-Å gate-oxide is increased from $6x10^{16}/cm^3$ to $1x10^{17}/cm^3$, V_T will increase from 0.6-V to about 1.4-V – which is far higher than the required V_T value of ~0.6 V.

In summary, this shows that in order to continue to scale the gate-length of MOSFETs, it will be necessary to decrease the oxide thickness by about the same scaling factor applied to the gate-length! However, such scaling of the oxide thickness brings with it both benefits and problems. One additional benefit is that the on-state drain current is further increased as the oxide thickness is scaled down (see Vol. 3). However, several other negative factors also ensue: the thinner oxide films cannot tolerate the increased gate electric fields, and this forces smaller power-supply voltages to be used (and such smaller supply voltages reduce the MOSFET on-state drain current); the tunneling leakage-current (gate current) through the oxide increases; the oxide is more prone to penetration by poly dopant (especially boron); and defects and oxide-layer thickness non-uniformities become bigger problems when thinner oxides must be manufactured. Figure 3-3 depicts these concerns for ultrathin gate oxides.

2. It must be possible to form the oxide layer with a thickness that closely matches the design specification of the MOSFET. This oxide thickness must also be sufficiently uniform across the entire wafer, and from wafer-to-wafer, and from run-to-run.

3. The oxide film and the Si/SiO_2 interface must maintain stable electrical characteristics. This includes exhibiting adequately small values of charge in the oxide and at the $Si-SiO_2$ interface (i.e., low Q_f, D_{it}, Q_{ot} and Q_m values).

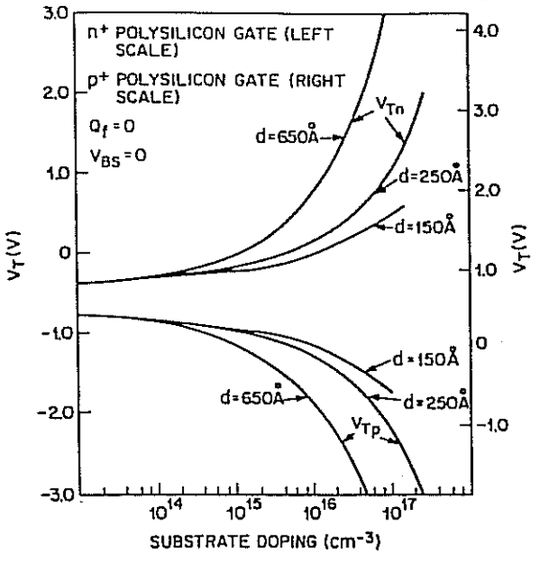

Fig. 3-2 Calculated threshold voltages of *n*-channel (V_{Tn}) and *p*-channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon gate (right scale). Curves for gate oxide thicknesses *d* of 150Å, 250Å, and 650Å are shown. From S.M. Sze Ed. *VLSI Technology*, 2nd Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Labs. Reprinted with permission.

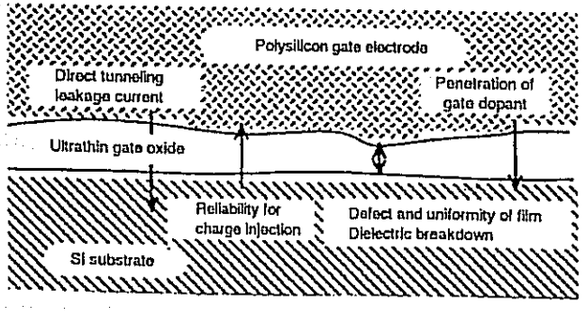

Fig. 3-3 Concerns in ultra-thin gate oxides.

4. The dielectric breakdown strength of the oxide must be sufficiently high (e.g., > 8-MV/cm), implying that the film is pinhole free and contains a negligible number of defects that would lead to oxide breakdown at lower electric fields.

5. The oxide film must exhibit sufficiently long lifetime under normal operating conditions (i.e., t_{BD} is adequate). This characteristic is related to 4.

6. The dielectric material (oxide) must be compatible with the other materials of the MOSFET structure, and be thermally, electrically, and chemically stable under the processes used to fabricate the ICs.

7. The leakage-current through the gate-oxide must be sufficiently small to meet the off-current leakage requirement of the IC in which it is used.

8. The oxide should exhibit high resistance to hot-carrier damage (i.e., device degradation should be low in the face of hot-carrier stressing).

9. If the oxide is to be used in a symmetric CMOS technology (i.e., in which both p^+ and n^+ poly are used), the oxide film needs to be resistant to the penetration of boron at the process temperatures used after gate doping.

10. The oxide must passivate the silicon surface.

3.2 THE STRUCTURE OF THERMALLY GROWN SiO_2 AND THE PROPERTIES OF THE Si/SiO_2 INTERFACE

3.2.1 The Microscopic Structure of Thermally Grown SiO_2

SiO_2 (silica) can be found in crystalline or amorphous (vitreous) forms. When the atomic structure of silica exhibits long-range order it is in *crystalline* form. Many varieties of crystalline-silica exist, including quartz, crystobalite, coesite, etc. However, the type of silica encountered in silicon ICs (i.e., silicon-dioxide films thermally grown under normal conditions) is not crystalline. Instead, it is silica in its *vitreous* (or *glassy*) form. Vitreous solids do not exhibit long-range order. Instead, their structure is ordered only over short ranges. This glassy state of SiO_2 is also often referred to as *fused-silica*. Since we are interested in thermally grown SiO_2, our discussion will focus on the glassy state of SiO_2 (fused-silica).

The basic structural unit of silica is centered around the *structural formula* (SiO_4). The spatial arrangement of Si and O atoms is due to their respective valence (+4 and -2), their relative sizes, and their electrostatic interactions. These factors give rise to elementary SiO_2 cells of tetrahedric configurations. That is, a silicon atom (with a valence of +4) is located at the center of the tetrahedron, with oxygen atoms (O^{2-}) at each of the corners (Fig. 3-4a).[3] In crystalline-SiO_2 (quartz), each oxygen atom belongs to two tetrahedra and is thus bonded to 2

silicon atoms. Such oxygen atoms are then known as *bridging oxygen atoms*. (Fig. 3-4b).[3] In vitreous- or amorphous-SiO_2 some of the vertices of the tetrahedra have *nonbridging oxygen atoms*, meaning they are not shared between two tetrahedra (Fig. 3-4d). The greater the ratio of bridging-to-non-bridging oxygens, the greater the cohesiveness of the SiO_2 structure.

The interatomic distances (from the center of one atom to the center of its nearest neighbor) have been measured and found to have the following mean values: 1.62-Å for the Si-O bond; 2.27-Å for the O-O distance; and 3.12-Å for the Si-Si distance. Note this implies that a 2.0-nm gate oxide is only 10-12 atomic layers thick!

The structure of fused-silica is a continuous random network of tetrahedra, where the Si-O-Si bond angle varies from 110° to 180°, with the most probable value being 144°. A two-dimensional representation of the vitreous-SiO_2 lattice is shown in Fig. 3-4c.[4]

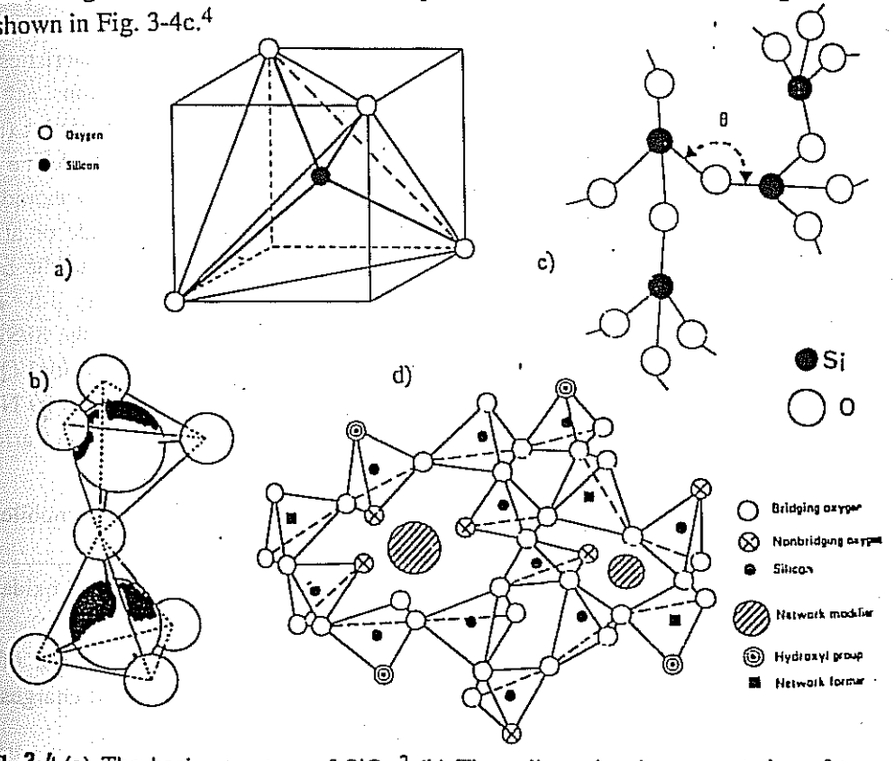

Fig. 3-4 (a) The basic structure of SiO_2.[2] (b) Three-dimensional representation of two neighboring SiO_4 cells, bridged by an oxygen atom. (c) Two-dimensional lattice representing vitreous-SiO_2.[2] (d) The structure of thermally grown SiO_2 showing bridging and non-bridging oxygen atoms and dopant (i.e., network modifier) atoms.[3] (© IEEE 1965)

The density of thermal-SiO_2 (fused-silica) is 2.20-g/cm^3 which is smaller than that of quartz (2.65-g/cm^3). In fused silica only 43% of the lattice space is occupied, making its structure much more open than that of quartz. Consequently, a large variety of impurity atoms can easily enter this oxide network and diffuse through it. Since it is difficult to directly assess the density of thin films, the refractive index of SiO_2 is usually measured instead, and the density is then inferred from it. A refractive index of 1.460 corresponds to a density of 2.20-g/cm^3. Generally, CVD oxides exhibit smaller densities than thermally grown SiO_2, and worse electrical and material properties correspond to less dense films. Hence, measurements of the index of refraction also provide a method for rapidly comparing the characteristics of deposited oxides.

3.2.2 The Si/SiO_2 Interface

In addition to the characteristics of the SiO_2 bulk structure, the characteristics of the Si/SiO_2 interface also play a critical role when SiO_2 functions as the dielectric layer of a MOSFET. Hence, a discussion of these characteristics (and the structure) of the SiO_2 interface is also relevant. First, it is useful to define the types of charges that exist in the oxide and at the Si/SiO_2 interface. (It should be noted that such charges are not expected to exist in an "ideal" oxide or at an "ideal" interface.) Four different types of charges are associated with the bulk oxide and "real" Si/SiO_2 interfaces, as shown schematically in Fig. 3-5. The *effective net* (i.e., *total*) *charge-per-unit-area* due to such charges at the Si/SiO_2 interface and within the oxide bulk (in C/cm^2) is denoted by the symbol Q_{tot}, whereas the *net* (*total*) *number of charges* per unit area (Q_{tot}/q) is given by N_{tot}. The standardized terminology and symbols used to differentiate these various individual charge types are as follows:

1. Mobile ionic charge per unit area Q_m (C/cm^2), and N_m (no. of mobile charges/cm^2);
2. Fixed oxide charge per unit area Q_f (C/cm^2), and N_f (no. of fixed oxide charges/cm^2);
3. Interface trapped charge per unit area Q_{it} (C/cm^2), and N_{it} (no. of interface trapped charges/cm^2); D_{it} - number of interface trapped charges per unit area *and* energy (no. of interface trapped charges/cm^2-eV);
4. Oxide trapped charge Q_{ot} (C/cm^2), and N_{ot} (no. of oxide trapped charges/cm^2).

The issues relating to the reliability of deep-submicron MOSFETs primarily involve Q_{it} and Q_{ot}. That is, the problems caused by Q_m and Q_f are not only well

controlled, but their impact does not get any more severe as oxides are scaled. On the other hand, those associated with Q_{it} and Q_{ot} are ongoing, scaling-dependent problems that must be addressed as MOSFETs are pushed further into the deep-submicron regime. Hence, the review provided here will examine only these two categories of interface/oxide bulk charges. This will include the origin of the charges and their effect on device behavior. A more detailed description of the characteristics of Q_m and Q_f can be found in Vols. 1 and 3, and in Refs. 1 and 7.

3.2.2.1 Interface Trapped Charge: The *interface-trapped charges* refer to charges which are localized at sites at the surface of the silicon (i.e., at the Si/SiO_2 interface). Such sites are believed to arise from energy states that exist in the forbidden gap of the silicon, and are referred to as *interface states*. Such surface states are extra allowed energy states (i.e., that electrons can occupy) present at the semi-conductor surface, but not within the bulk.

The term *surface state* was coined by Tamm in 1932.[5] Using quantum mechanical calculations, he showed that new electronic energy states arise in a crystal if the lattice is terminated at a surface. Such new states are confined to the region very close to the surface. He also calculated that these states will arise at interfaces (such as the Si/SiO_2 interface), as well as at free surfaces. Each of the states is associated with a single atom at the surface. Hence, an electron occupying one of these states is localized (i.e., it is forced to remain in a restricted region of space centered on that atom). Since such states thus effectively *trap* free carriers at the surface, they are also referred to as *interface traps*. The charge-per-unit-area stored in these traps is symbolized by Q_{it}.

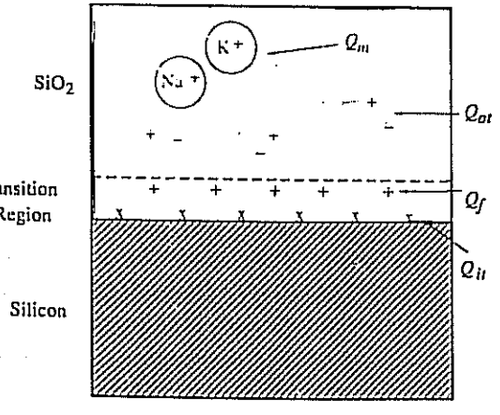

Fig. 3-5 Charges associated with the Si/SiO_2 system.

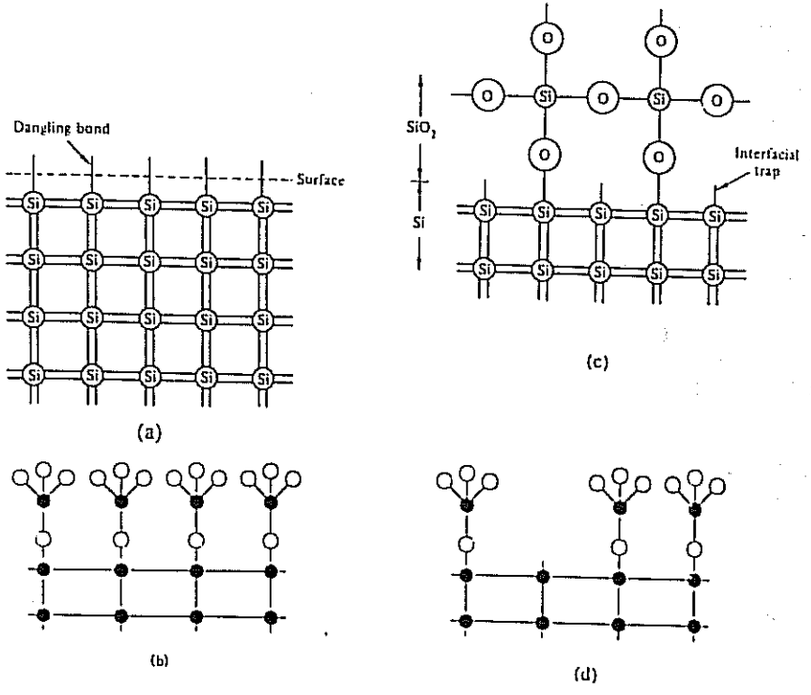

Fig. 3-6 Physical model for the interfacial traps. "Dangling bonds" which occur when the Si lattice is abruptly terminated along a given plane to form a surface. (b) Post-oxidation perfect (ideal) interface. (c & d) Post-oxidation dangling bonds that become interfacial traps.[7] Reprinted with permission of the publisher, the Electrochemical Society.

Although the theoretical basis of interface traps is well accepted and models exist that accurately detail the electrical behavior of the traps, the *physical origin* of these surface states has not been totally determined. The weight of experimental evidence supports the view that surface-states arise primarily from unsatisfied, or *dangling* bonds at the silicon surface (see Fig. 3-6). Note that a Si atom at a surface with a dangling bond can also be viewed as a *trivalent Si atom*.

From a qualitative perspective, the dangling bond model can be visualized with the aid of Fig. 3-6a. Here it is seen that if a Si lattice is abruptly terminated along a given plane to form a surface, one of the four bonds of each Si atom at the surface is left dangling. There will thus be extra states on this surface because the energy field of the crystal is one-sided (i.e., the electrons in the surface region are bonded only from the side directed toward the bulk). It is plausible that thermal formation of SiO_2 can tie up these surface Si bonds, and that along a perfect Si/SiO_2 interface all such bonds could be tied up (Fig. 3-6b). In this case,

surface-states at the interface would be suppressed.[6] However, it is more plausible that such oxidation will not tie up *all* the bonds (Fig. 3-6c). If even a very small fraction of the number of dangling bonds is left unsatisfied, a significant number of surface-states could exist. For example, on a (100)-Si-surface there are 6.8×10^{14} Si atoms per cm^2. If oxidation left 1/1000 of these bonds dangling, and each of them gave rise to a surface-state, the density of interface trapped charges would be $6.8 \times 10^{11}/cm^2$ (assuming one electronic charge is associated with each energy state). If a gate-oxide-thickness of 20-nm was being used, this would cause a threshold-voltage-shift of 0.63-V. This indicates that if the dangling-bond model correctly describes the origin of interface states, then only a relatively small number of residual-dangling-bonds can significantly perturb MOSFET device characteristics. Crystal-defects near the surface or foreign atoms bonded at the surface are other interface perturbations that have been identified as possible sources of surface-states.

Figure 3-7 represents the bond-stretching and broken-bond model of the Si/SiO_2 bulk of Verwey[2] and of the Si/SiO_2 interface of Sakurai and Sugano.[6] That is, Fig. 3-7a depicts the stretched and broken Si-O bonds in the bulk, and Figs. 3-7b-e at the interface. Figures 3-7f-h illustrate the possible defects in the oxide bulk. In Fig. 3-7b, the dangling-Si-bond at the surface is depicted as an *interfacial-trivalent-Si-atom,* while in Fig. 3-7f, a dangling-Si-bond in the bulk is portrayed as a *trivalent-Si-atom.*

Trivalent-Si-atoms introduce a deep trap-level near the Si midgap when located at the interface, and a deep trap-level in the SiO_2 bandgap when located in the oxide bulk. The weak Si-Si* and Si-O stretched bonds (Figs. 3-7c and 3-7d) at the interface introduce a continuum of deep trap levels, or if such stretched bonds exist in the bulk of the oxide (Fig. 3-7g), they produce shallow trap levels. If a dangling-Si-bond is tied up with an impurity (most commonly H or OH), this is thought to produce an electron-trap, while the oxygen vacancy (or weak Si-Si stretched bond) is considered to be the precursor of a hole-trap.

An anneal in a hydrogen ambient (100% H_2 or 4%-H_2 in N_2 [forming gas]) at approximately 450°C is normally the final step prior to assembly and packaging of an IC. This step is thought to allow hydrogen to penetrate the gate-oxide and then tie up the remaining dangling bonds at the Si/SiO_2 interface not tied up by thermal oxidation (as suggested in Fig. 3-8). However, the Si-H or Si-OH bond can be easily broken by injected hot-electrons, giving rise to interface traps. Thus, hydrogen introduced in the post-metal anneal or other process steps (i.e., during the steam reflow of an interlevel BPSG layer, or during the deposition of silicon nitride), can increase hot-electron degradation in a MOSFET.

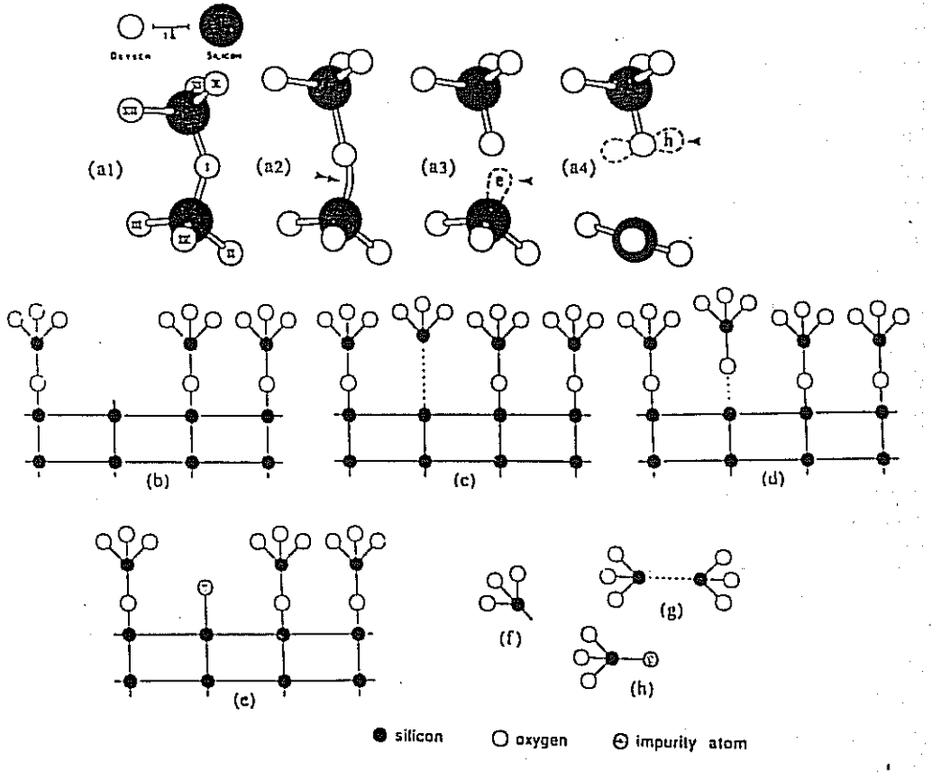

Fig. 3-7 (a) Strained and broken bonds: (a1) Normal Si-O bonds; (a2) Strained bond in vitreous silica; (a3) Broken bond represented with a trapped electron; (a4) Broken bond represented with a trapped hole. Possible interface defects (b-e): (b) Si dangling bonds. (c) Si-Si stretched bond (or oxygen vacancy). (d) Si-O stretched bond. (e) Impurity at interface. Possible oxide bulk defects (f-h): (f) Si dangling bond; (g) Si-Si stretched bond (or oxygen vacancy). (h) impurity in SiO_2.[6] (© IEEE 1988)

Consequently, it has been suggested that the ideal gate dielectric and all subsequent dielectric layers should contain as little hydrogen as possible to make them hot-electron resistant.

3.2.2.2 Effect of Interface Traps on IC Characteristics: The presence of interface traps has four major effects on the characteristics of ICs. First, the charge that is present in the interface traps interacts with the charge in the silicon near the surface and thus changes the silicon charge distribution and the surface potential. Second, interface traps act as generation-recombination centers, giving rise to leakage currents, as described below. Third, since the density of interface trap

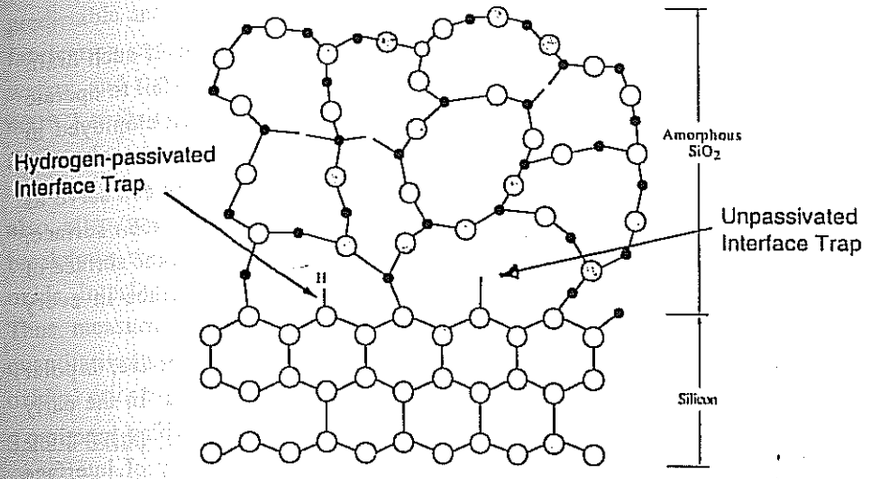

Fig. 3-8 How hydrogen can passivate the dangling bonds at the silicon surface shown in Fig. 3-7c and reduce the interface trap density.

charge changes with the surface potential, it gives rise to an additional capacitance component in parallel with the silicon gate capacitance. Fourth, an increase in interface-state-density has been correlated to weakening of the oxide, which appears to be a prelude to the second-stage breakdown event in gate oxides.

The issue concerning the *effect on the surface-potential of charge being trapped at the interface* is considered first. That is, such charge trapped at the interface has the same effect on the surface-potential as placing a sheet of charge Q_{ox} at the oxide-silicon interface. Such a sheet of charge will impact the threshold voltage, causing it to shift by an amount given by $\Delta V_T = -Q_{ox}/C_{ox}$. To keep this threshold voltage shift as small as possible, the value of Q_{ox} should also be kept as small as possible. The last paragraph of this section explains the measures taken to achieve this goal.

A secondary effect of interface-trap-charge on the surface-potential was introduced earlier. That is, injection of hot electrons into the gate-oxide is thought to produce interface traps (by breaking the bond between a silicon-surface atom and a hydrogen atom, which leaves a dangling bond on the Si atom). Since this process produces additional dangling bonds (and thus interface traps) as time goes by - and since these interface traps can trap electrons - this phenomenon will also cause the value of Q_{ox} to increase with time. In turn, this will cause the value of V_T to change with time. If V_T increases as result of the

buildup of Q_{ox}, the drain current of the MOSFET will decrease. If such drain current degradation becomes too severe the device will eventually fail to operate properly. Thus, measures to avoid hot-carrier-injection have been employed in MOSFETs as their gate lengths approached 1.0-μm and smaller.

The second effect of interface states involves the fact that they can serve as generation-recombination centers. Since they are present at the silicon surface, they will exist in the depletion regions of *pn* junctions that intersect the surface of the silicon. There are many such junctions at the Si/SiO_2 interface in every device, and therefore in every IC.

Before the invention of integrated circuits, only discrete diodes, bipolar transistors, and field-effect transistors (FETs) could be fabricated. In the early 1950s, these discrete devices exhibited relatively high reverse-bias junction-leakage and low breakdown-voltage (caused by the large density of interface traps at the surface of single-crystal silicon).

In 1958, a group of workers at Bell Telephone Laboratories, led by Atalla, found that when a thin layer of SiO_2 was grown on the surface of silicon where a *pn*-junction intercepts the surface, the leakage current of the junction was reduced by a factor from 10 to 100. It was later understood that the oxide reduces the number (or stabilizes many) of the interface traps. Not only did such oxide-passivation of the silicon surfaces allow diodes and transistors to be fabricated with significantly improved device characteristics, but the leakage path along the surface of the silicon was also effectively shut off. Thus, one of the fundamental isolation capabilities needed for planar devices and integrated circuits had also been developed. In his report on the evolution of the MOS transistor, C.T. Sah remarks that the successful effort by the Bell Labs group to stabilize Si surfaces was the most important technological advance in microelectronics during the 1950s, and that it blazed the trail that led to the development of the silicon integrated circuit.[8]

The third effect of interface traps involves the interface-trap capacitance. As noted earlier, the *interface-trap capacitance* C_{it} is in parallel with the gate capacitance. However, only those interface traps that can be filled and emptied at a rate faster than the capacitance measurement signal can contribute to C_{it}. Traps too slow to follow the capacitance-measurement signal will not contribute to C_{it}. Therefore, the observed C_{it} is not only a function of the interface-trap density but also depends on the MOS-capacitor gate voltage (which controls the surface potential and hence the probability of occupancy of the traps) and the frequency at which the capacitance measurement is made.

All of these observations point to the conclusion that for a MOSFET to have

predictable and reproducible capacitance characteristics, it is important to employ fabrication processes that minimize trap density. Using (100)-oriented silicon wafers, and a post-metallization hydrogen-annealing step (at temperatures around 400°C) at the end of the fabrication process, are the main techniques for keeping the interface-trap-density at a minimum.

It has been observed that the passage of tunneling current through a thin oxide also introduces interface traps at the Si/SiO_2 interface. Furthermore, while breakdown may not occur until about 10 C/cm^2 of charge has passed through the oxide, interface traps are detectable even after 0.001 C/cm^2 of charge passage. Thus, measuring interface traps has been proposed as a more sensitive means of quantitatively measuring oxide damage (due to Fowler-Nordheim stressing of oxides) prior to breakdown.

3.2.2.3 Oxide Trapped Charge: The oxide trapped charge Q_{ot} is due to holes or electrons trapped in the *bulk* of the oxide, and hence can be positive or negative. They become trapped at *trap centers* that are microscopic defects in the oxide structure *in the bulk of the oxide layer.* Such defects may be caused by ionizing radiation, hot-carrier injection, or high currents through the oxide (e.g., due to Fowler-Nordheim tunneling), or other reactions which either bend or break Si-O bonds in the oxide. Figures 3-7f-h show possible defects in the SiO_2 bulk that may give rise to such traps. For example, a trivalent Si atom in the SiO_2 will introduce a deep trap level in the SiO_2 bandgap ($.Si{\equiv}O_3$), and stretched Si-O or Si-Si bonds [in the latter, a missing bridging-oxygen atom leads to a defect ($O_3{\equiv}Si..Si{\equiv}O_3$)]. Figure 3-9 shows schematics illustrating the potential wells of electron traps in silicon dioxide. While, as-grown SiO_2 films contain few such microscopic oxide defects, electron and hole traps can be easily introduced by bombardment with high-energy photons or particles.[9] (The exception to this

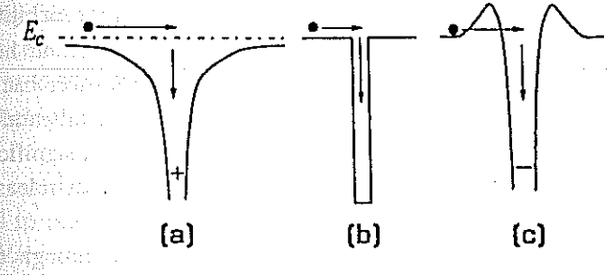

Fig. 3-9 Schematics illustrating the potential wells of electron traps in silicon dioxide: a) Coulomb-attractive trap; b) neutral trap; and c) Coulomb-repulsive trap.

statement is that as-grown wet oxides tend to contain some oxide traps because they also contain OH^- and H_2O species. The density of oxide traps in wet oxides can be significantly reduced with an appropriate post-oxidation-anneal.)

Q_{ot} can be annealed out by low-temperature treatments (above 550°C) although some neutral traps may remain. Therefore, oxide trapped charge was been considered relatively less important than the other oxide charges in just-completed MOSFETs. In early generations of MOSFET ICs, exposure of the devices to ionizing radiation encountered in space flights was the main concern involving Q_{ot}. However, in deep-submicron CMOS ICs, the problems involving Q_{ot} are becoming serious for several reasons. First, processing techniques such as ion implantation, sputter deposition, and plasma etching are potential sources for creating charged and neutral traps in the gate dielectric (all of which may not be completely annealed out at the end of the fabrication process). This may result in long-term reliability problems, especially if hot carriers are injected in these oxides and get trapped. Another reliability concern involves gate-oxide degradation due to high electric-fields in the oxide (and subsequent Fowler-Nordheim tunneling currents in the oxide). This has become more of a concern as oxide thickness has been scaled down.

3.2.2.4 Effect of Oxide Trapped Charge on Device Characteristics: The major effect of oxide trapped charge on MOSFETs is that they are a manifestation of the damage done to the gate oxide during device fabrication or by electrical stressing (either by the application of a high electric-field across the oxide, or by the injection of carriers into or through the oxide). We will discuss this issue in more detail in Sect. 3.3.

3.3 DIELECTRIC BREAKDOWN IN SILICON DIOXIDE FILMS

Dielectric breakdown of the gate-oxide layer is one of the major reliability issues that affects MOSFETs. We provide a review of this phenomenon here, and recommend that interested readers consult Vol. 3, Chap. 7 for more background information on the various breakdown measurements, statistical models, and burn-in tests used to screen out parts with weak oxides. However, significant new information that has been published since 1995 (when Vol. 3 was released) and this new information will be focus of the discussion here.

Figure 3-10 shows the time dependence of the gate current through an oxide under a large and constant voltage stress applied across the oxide layer. One can see that there is a gradual decrease in the current with time until the oxide suffers *dielectric breakdown* (Fig. 3-10). At that point the current rises rapidly (*catastrophic breakdown*).[10] After such breakdown, the oxide is irreparably damaged,

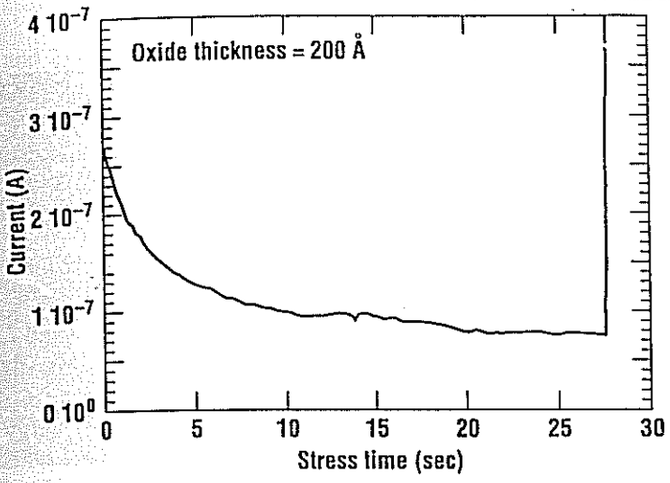

Fig. 3-10 Constant-voltage stress measurements of a 20-nm-thick oxide film. The sharp increase in current near the end of the trace indicates that irreversible breakdown has occurred.

and can no longer function as the dielectric layer in the MOSFET. Since such breakdown occurs after the oxide is subjected to a voltage stress for some extended period of time, this phenomenon is also referred to as *time-dependent dielectric breakdown* (TDDB). Anomalous "soft breakdown" of oxides also occurs for films < 5.0-nm thick, and this will be discussed in Sect. 3.8.

A different metric of the quality of oxide films is expressed in terms of the electric field at which dielectric breakdown occurs. To measure this value, a voltage (that increases linearly with time) is applied between the gate and substrate of a MOS-C test structure until the oxide breaks down. The "best" (apparently defect-free) oxides break down at fields greater than about 10 MV/cm, and failures at such field strengths are referred to as "intrinsic breakdown." Less-than-perfect-oxides (i.e., those that contain some form of defect) may fail at lower voltages (i.e., between 4-6-MV/cm).

In any case, dielectric breakdown in ICs must be avoided. Extensive work has demonstrated that 20-year oxide-lifetimes can be achieved if the oxide field remains below about 8-MV/cm.[11] Thus, for devices used in CMOS logic and memory chips, the maximum electric field during normal device operation is typically restricted to the 5 MV/cm range, and this limit on the electric field is projected to remain at about this value for each technology generation. That is, the 5-MV/cm value is expected to give an acceptable safety margin for 20-30-year oxide-lifetimes. Chips are also normally subjected to burn-in tests to screen out those chips that contain oxides that would fail under the normal operating conditions during their time of service.

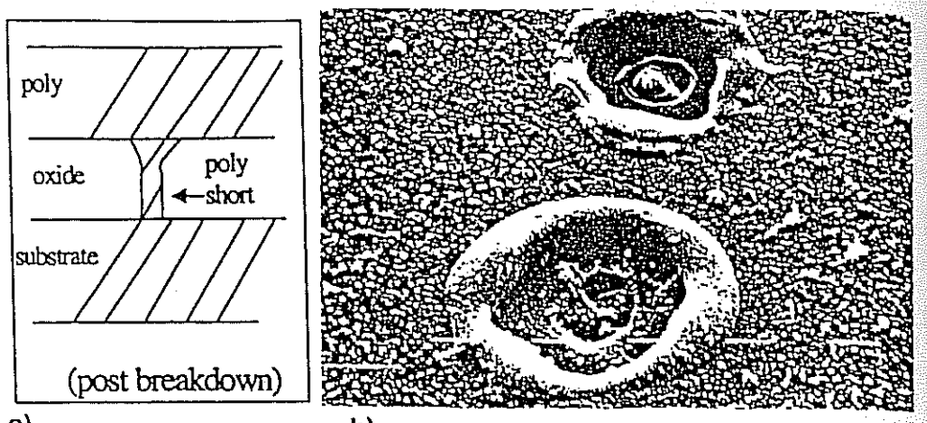

Fig. 3-11 a) Localized poly filament shorting the gate and substrate after gate oxide breakdown; b) A SEM photograph of two breakdown spots. The polysilicon electrode has melted away over 20 μm^2. The polysilicon is spread over a large area in the form of dust and little globules. The size of a dash is 1-μm. From D.R. Wolters and J.F. Verwey, "Breakdown and Wear-Out Phenomena in SiO_2 Films," Chap. 6, p. 329, in *Instabilities in Silicon Devices*, G.M. Barbottin and A. Vapaille Eds., © 1986 Elsevier Science.

Because dielectric breakdown is such a critical reliability issue in CMOS ICs, breakdown of thin oxide films has been (and still is) a subject of intense research. The physical mechanisms involved in the dielectric breakdown process are quite complex. We will discuss some of the latest proposed models to shed light on this phenomenon.

In general, the catastrophic breakdown of a dielectric is believed to be a two-stage process. First, the dielectric is slowly degraded over time (perhaps years) due to being subjected to high-electric fields or due to the passing of current through the oxide. This degradation takes the form of creating defects in the oxide (e.g., electron/hole traps and/or bond-breakage). These defects form a localized conductive path through the oxide during this time. The second stage is the very short runaway breakdown process itself (it may occur in microseconds). During the runaway breakdown phase the oxide is shorted (due to severe joule-heating that causes a conductive filament to form in the dielectric) and a very large current flows between the gate and substrate in the MOS device (Fig. 3-11).

There are numerous forms of damage to the oxide that are postulated to occur during the oxide degradation process, including lattice damage, trap creation by electrons, and trap creation by holes. The two physical mechanisms that are blamed for producing these types of damage are: 1) externally applied electric

fields; and 2) current that passes through the oxide. Thus, there are two classes of models that have been developed to predict TDDB: 1) models that assume the oxide degradation is electric-field-driven (the so-called *E-models*); and 2) models that assume the oxide degradation is current-driven (the so-called [1/*E*]-models). Here we briefly describe both models. Also note that although the physical basis for oxide degradation is still under debate, definitions and practices have been developed to describe and test the integrity of the oxide.

3.3.1 Electron Trapping in Silicon Dioxide: (and why the Current in the Oxide Decreases with Time During Constant-Voltage Stressing)

As described in Sect. 3.2.2.3, silicon-dioxide layers contain electron and hole traps (either as a result of residual damage inflicted during device processing, or caused by the electric fields applied across the oxide - or gate current that flows - during normal device operation). If these sites trap carriers which are transported in the oxide conduction-band (electrons), they can give rise to oxide-trap charge Q_{ot}. Electrons can enter the conduction band of the oxide in several ways, including: 1) ionizing radiation can generate electron-hole pairs within the oxide itself; 2) electrons can enter from the silicon substrate or the gate polysilicon either through tunneling or through hot-electron injection. Some of these electrons may get trapped at oxide trap sites. These trapped electrons will then cause the electric field in the oxide to be modified in the following manner: The field near cathode (i.e., the electrode that acts as an electron source) is decreased, while the electric field near the anode (the electrode that acts as the electron sink) is increased (as depicted in Fig. 3-12). A decrease in the field near the cathode causes Fowler-Nordheim current in the oxide to decrease. Since the gate current is due primarily to Fowler-Nordheim tunneling, this explains why the gate current declines with time under a constant-voltage stress (as shown in Fig 3-10).

3.3.2 The Electric-Field-Driven Model of Oxide Degradation

In the electric-field -driven models of oxide degradation (*E*-models), the cause of low-field (< 10-MV/cm) TDDB is postulated to be field-enhanced thermal bond breakage. That is, at a given temperature, chemical bonds in the SiO_2 structure can be broken by thermal vibrations. Such bond breakage produces local defects that weaken the oxide, and when the concentration of such defects reaches a critical value, breakdown can occur. An externally applied electric field serves to reduce the activation energy required for such thermal-bond breakage, and therefore increases the degradation-reaction rate for failure exponentially. The *time-to-breakdown* t_{BD} which is the inverse of degradation rate, decreases exponentially

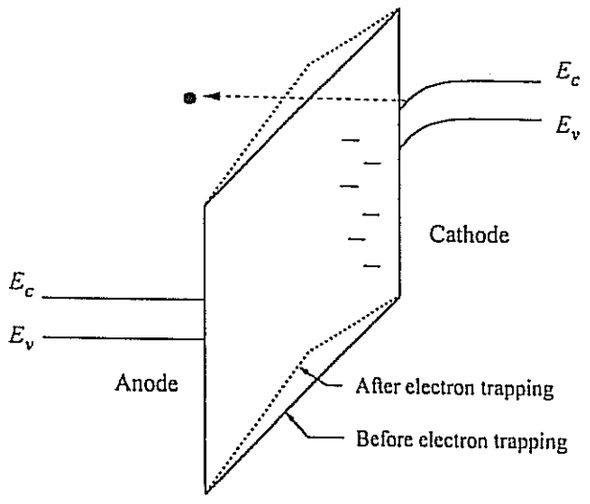

Fig. 3-12 Schematic illustrating the trapping of tunneling electrons. As electrons are trapped, the oxide field near the cathode (electron source) is decreased, while the oxide field near the anode (electron sink) is increased.

with field according to:

$$t_{BD} = A_0 \exp(-\gamma E_{ox}) \exp(Q/k_B T) \quad (3.1)$$

where γ is the field acceleration-parameter, E_{ox} is the electric field in the oxide (usually expressed in MV/cm), and Q is the enthalpy of activation for bond breaking in the absence of an external electric field. Generally, for oxide thicknesses greater than 90-Å, γ ~2.5-3.5-cm/MV.

3.3.3 The Current-Driven Model of Oxide Degradation (1/E Model)

In the so-called 1/E model for TDDB, degradation is postulated to be due to current flow through the dielectric caused by Fowler-Nordheim conduction. Electrons, which are F-N injected from the cathode, may cause damage to the dielectric due to impact ionization as the electrons are accelerated through the oxide. Also, when these accelerated electrons reach the anode, hot holes may be produced (see Sect. 3.3.4 below) which can tunnel back into the dielectric causing damage. Since both the electrons from the cathode and the hot holes from the anode are the result of F-N conduction, the time-to-failure is expected to show an exponential dependence on the reciprocal of the electric field, $1/E$:

$$t_{BD} = \tau_0(T) \exp[G(T)/E_{ox}] \quad (3.2)$$

where $\tau_0(T)$ is a temperature dependent pre-factor, G is the $1/E$-model field acceleration-factor, and k_B is Boltzmann's constant.

The temperature dependence of G has been expressed as a 1/T power series expansion, given by:[12]

$$G = G_0 [1 + (\gamma/k_B)\{(1/T) - (1/300K)\}] \quad (3.3)$$

where:

$$\gamma = (k_B/G_0) [dG/d(1/T)]_{300K} \quad (3.4)$$

and where the derivative is evaluated at $300K$. At room temperature, G_0 ~350-MV/cm and γ ~0.017-eV.

$\tau_0(T)$ is also usually represented as a 1/T expansion:

$$\tau_0(T) = \tau_0 \exp[(-E_a/k_B)\{(1/T) - (1/300K)\}] \quad (3.5)$$

where τ_0 ~1×10^{-11} sec, and E_a ~0.3-eV.

3.3.4 The Hole-Trapping Model that Describes How Holes are Injected and Trapped in SiO_2

The transport of holes through the oxide has been shown to precede breakdown, indicating that holes cause damage to the oxide (namely by creating traps in the bulk oxide).[13] A model (the *anode-hole injection model* [*AHI model*]) has been published that seeks to explain how such holes are generated and injected into the oxide during high electrical field stress.[14] The mechanisms that occur are depicted in Fig. 3-13.

When an electric field is applied across an oxide layer using an MOS-C structure, the energy-band diagram of this situation is depicted in Fig. 3-13. If an electron tunnels from the cathode into the conduction band of an oxide layer (by F-N tunneling), it can then gain kinetic energy from the electric field applied across the oxide, and get accelerated toward the anode. When the electron arrives at the anode it has a large amount of kinetic energy (E_{gain} in Fig. 3-13). Some fraction of such electrons that arrive at the anode are able to transfer this energy to an electron deep within the valence band of the anode silicon near the oxide-anode interface. This valence band electron is excited to the bottom edge of the silicon conduction band in the anode. The excitation event also produces a hole deep within the valence band, and this "hot" hole can tunnel into the valence band of the oxide (again by F-N tunneling, as depicted in Fig. 3-13). These holes can be trapped in the oxide and also create the damage described earlier. This damage accumulates with time, degrading the oxide.[35]

In addition, the trapped holes in the oxide layer cause an increase in the oxide field near the cathode, and a decrease in the field near the anode (Fig. 3-14). A small increase in the oxide field near the cathode can cause a large increase in the

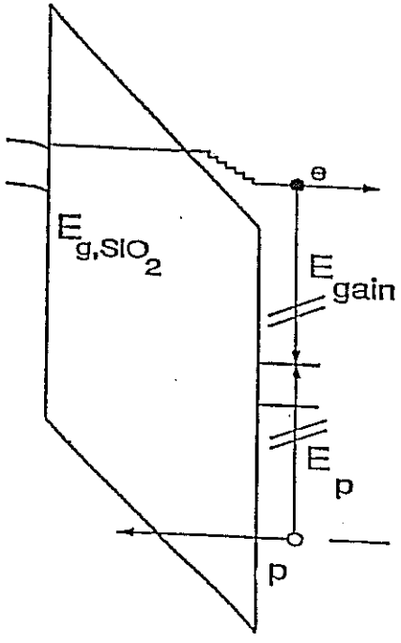

Fig. 3-13 Energy-band diagram of the anode-hole injection (AHI) model, illustrating the generation of an electron-hole pair by a tunneling electron. The hole thus generated can then be injected (by tunneling in this model), into the oxide layer.

tunneling current. Thus, hole trapping in the oxide near the cathode provides positive feedback to the electron tunneling process. It should be pointed out that the increase in tunneling current due to hole trapping is usually not readily observable, except near breakdown, because it is masked by the decrease in tunneling current due to electron trapping.

It should be noted that recent work has called some conclusions of this model into question. That is, until recently it was generally accepted that the substrate current originates from the anode-hole-injection (AHI) mechanism, which was schematically depicted in Fig. 3-13. Part of the holes created by the phenomena postulated by this model tunnel into the oxide, and the oxide degradation is related to the hole fluence (which is found to be constant at breakdown, independent of the stress current or stress voltage). However, the explanation of the substrate current by anode-hole-injection has never been completely verified, and other mechanisms such as minority-carrier generation due to F-N-induced photons have been suggested to be a cause as well. Two recent reports claim that the F-N-induced photons in the gate are the dominant source of the substrate hole

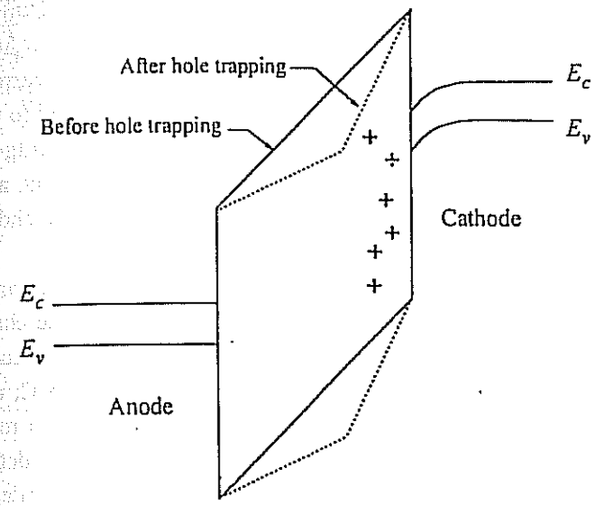

Fig. 3-14 Schematic showing the trapping of holes in the oxide layer. The trapped holes enhance the electric field near the cathode, and decrease the electric field near the anode.

current, and thus the generally accepted explanation of oxide degradation based on anode-hole-injection may have to be revised.[31,32]

3.3.5 Comparing the Electric-Field-Driven and the Current-Driven Oxide Breakdown Models

There has been a great deal of disagreement as to which is the dominant degradation mechanism for low-field TDDB in thin oxide films; i.e., is it field-induced or current-induced? The (1/*E*)-model fits the TDDB data very well at high fields (in which high F-N-currents exist), but falls short at lower fields (< 10-MV/cm). In fact, there have been several low-field/long-term TDDB studies carried out, with each showing that TDDB data were described much better by the *E*-model.[15,16,34] The good fit of the physics-based *E*-model to low-field/long-term TDDB data suggests strongly that it is field (not current), which is the dominant degradation mechanism at low fields because the F-N current is simply so vanishingly small at these lower fields. However, for very-thin oxides (< 40-Å) the direct tunneling-current is very high, which could mean that in these oxides the dominant degradation mechanism may again be current-driven.

3.3.6 Time-to-Breakdown (t_{BD}) and Charge-to-Breakdown (Q_{BD})

The breakdown characteristics of an oxide film are described in terms of the *time-to-breakdown* (as was discussed earlier) and in terms of its *charge-to-breakdown* Q_{BD} (which measures the *integrated* total tunneling current [i.e., total charge] just before breakdown). Recent publications have tended to use the value of Q_{BD}. It appears that it is simpler to develop models relating to the other physical processes, such as hole current, trapping, trap generation, and interface-state generation, than it is to develop models relating time-to-breakdown to these processes. Thus, Q_{BD} needs to discussed further.

As mentioned above, evidence indicates that tunneling-electron-current can lead to hole-injection into the oxide. Thus, Q_{BD} is the sum of the charges passing through the oxide due to electrons and holes. If a MOS-C structure is used to measure Q_{BD}, then due to the two-terminal nature of the MOS-C, only the total charge can be measured. However, if an NMOSFET is used to measure Q_{BD}, then both the total charge and the hole-charge component can be determined. For the case of an NMOSFET, the bias configuration for such measurements is shown in Fig. 3-15. Integration of the gate current gives the total charge, while integration of the substrate current gives the charge due to the holes. For a given oxide film, Q_{BD} is often plotted as a function of oxide voltage. Figure 3-16 is such a plot for oxide thickness in the 2.5-10-nm range.[15] It shows that for a given oxide thickness, Q_{BD} decreases with increasing oxide voltage.

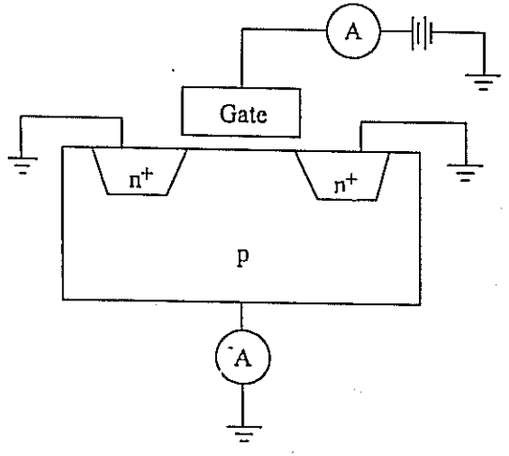

Fig. 3-15 Schematic depicting the bias configuration of an *n*-channel MOSFET for measuring the charge-to-breakdown Q_{BD} and its hole-charge component.

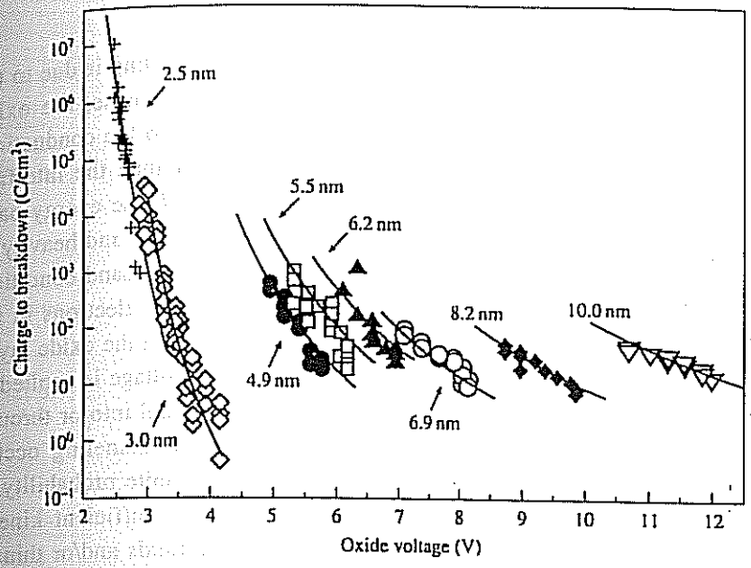

Fig. 3-16 Typical plot of charge to breakdown versus oxide voltage for several oxide thickness values.

3.4 LEAKAGE CURRENTS IN SiO_2 FILMS (TUNNELING PHENOMENA)

When considering the ideal MOS capacitor structure, it is assumed that no current flows through the oxide. However, in real MOS capacitors, a current from the gate to the substrate through the oxide can flow (termed *gate current*, I_g). Although the carriers that constitute such current can enter the conduction band (electrons) and valence band (holes) of silicon dioxide by hot-carrier injection, here we only discuss the other way that they can enter, namely by tunneling. This is because few carriers in deep-submicron MOSFETs enter the oxide as a result of hot-carrier injection during normal device operating conditions; hence they do not play a major role in the observed gate current.

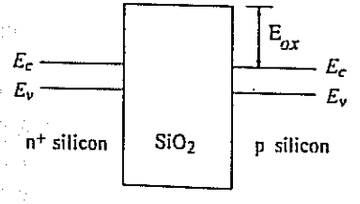

Fig. 3-17 Energy-band diagram of an *n*-type polysilicon-gate MOS structure at flat band.

The phenomenon of carrier injection into the oxide by tunneling is due to the quantum-mechanical nature of the electron. In the classical sense, the oxide represents an impenetrable barrier to injection of electrons into the conduction-band of the silicon if they possess kinetic energies smaller than the interface energy barrier for the electrons (E_{ox} = 3.1-eV). In Fig. 3-17, the energy-band diagram of a MOS capacitor with silicon dioxide as the insulator and heavily *n*-doped-polysilicon as the gate is shown when biased at the flat-band condition. When a large positive voltage is applied to the gate electrode, electrons in the strongly inverted Si substrate surface can tunnel into or through the oxide layer, and give rise to gate current. (Similarly, if a large negative voltage is applied to the gate electrode, electrons from the n^+-polysilicon can tunnel into or through the oxide-layer and again give rise to gate-current.) Such tunneling occurs because the wave nature of an electron (or hole) allows a finite probability of crossing the barrier, even if the electron does not possess sufficient kinetic energy. This probability increases with larger gate electric fields and/or thinner barriers. Electrons injected by tunneling are generally considered to be in equilibrium with the lattice, and are thus characterized as "cold" (i.e., not "hot").

3.4.1 Fowler-Nordheim (F-N) Tunneling (Tunneling *Into* Silicon Dioxide)

The first type of tunneling is called *Fowler-Nordheim tunneling*. Here electrons are injected by tunneling into the conduction band of the oxide through the triangular energy barrier. Once injected into the oxide conduction-band, electrons are accelerated by the oxide field toward the anode (gate), causing a gate current. Figure 3-18a sketches the energy band diagram for this case assuming a positive bias is applied to the gate.

To determine the probability of an electron exhibiting F-N tunneling through a barrier, Schroedinger's equation is solved for a triangular barrier. The WKB approximation (discussed in most quantum mechanical textbooks) is invoked, and the following expression for the tunneling current-density expression is derived:

$$J = A_F E_{ox}^2 \exp(-B/E_{ox}) \qquad (3.6)$$

where A_F = 1.25x10^{-6} A/V^2, B ~240-MV/cm, J is the current-density in A/cm^2, and E_{ox} is the oxide field in V/cm.[17] According to this equation, the oxide current is exponentially dependent on the electric field and experimental studies have found that the equation accurately describes the oxide current. This can be seen in Fig. 3-20, which shows the measured gate current and the gate current calculated using Eq. 3-6 in oxides 6-nm- and 8.3-nm-thick. The tunneling probability rapidly rises as E_{ox} increases because the tunneling distance (i.e., the

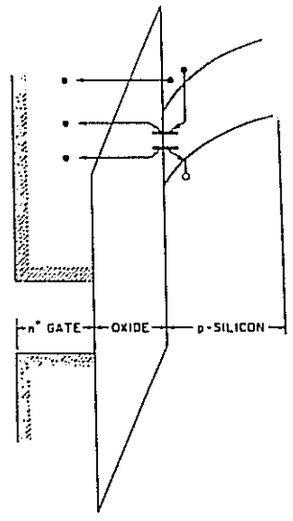

Fig. 3-18 Energy-band diagram for the phenomenon of Fowler-Nordheim tunneling in the MOSFET gate oxide. Also shown are some possible interface-trap assisted leakage paths, with trap levels indicated by short, solid bars.

distance from the Si-surface through the triangular barrier to the oxide conduction-band) becomes smaller as E_{ox} increases.*

It should also be noted that it may be possible for electrons to be "hot" but still not possess sufficient energy to surmount the barrier. In this case, carrier injection may occur by allowing these hot-carriers to tunnel through a smaller triangular-barrier distance. Thus, a synergistic combination of processes which individually would not have caused injection may yet enable carriers to be injected. This effect was modeled by Ning and Yu.[18]

At an oxide electric-field of 8-MV/cm, the measured Fowler-Nordheim tunneling-current-density is about 5x10^{-7} A/cm^2. Thus, for normal device operation, Fowler-Nordheim tunneling does not produce enough gate-current leakage to be significant. Hence, Fowler-Nordheim tunneling does not represent an issue when gate-oxides get thinner as the gate length of MOSFETs are scaled, even to deep-submicron gate-lengths.

3.4.2 Direct Tunneling (Tunneling Through Silicon Dioxide)

If the oxide is very thin (say 4-nm or less), electrons from the inverted silicon

* Because the distance of tunneling between Si surface and the oxide conduction band decreases with increasing electric field.

surface can tunnel *directly* through the forbidden-energy-gap of the oxide layer instead of tunneling into the conduction-band of the SiO_2 layer. In direct-tunneling, electrons tunnel through a trapezoid-shaped barrier as shown in Fig. 3-19. (Because of the differences in the height of barriers for electrons and holes, and because holes have a much lower tunneling-probability in oxide than do electrons, the tunneling limit will be reached earlier for NMOS than PMOS.)

The theory of direct-tunneling is even more complex than that of Fowler-Nordheim-tunneling, and there is no theoretical simple dependence of the tunneling-current-density on voltage or electric field. (While an expression for the tunneling probability can be derived, no analytical expression is available for direct tunneling (as exists for the case of F-N tunneling). Instead, direct-tunneling-current must be calculated from numerical analysis, and the results of such calculations are plotted in Fig. 3-20.[19]) Nevertheless, experimental data indicates that tunneling current exhibits a weaker dependence on oxide fields than is shown by the equation that governs Fowler-Nordheim tunneling (Eq. 3-6). This means that direct-tunneling dominates in thin oxides at low voltages.

For oxide thicknesses smaller than 3-nm, the direct tunneling-current becomes very large. Figure 3-21 is a plot of the measured and simulated thin-oxide tunneling-current versus voltage in polysilicon gate MOSFETs. For the gate-voltage range shown in Fig. 3-21, the current is primarily a direct-tunneling current.[20] This tunneling represents a limit on the oxide thickness that can be used. This limit will be reached approximately when the gate-to-channel tunneling-current becomes equal to the off-state source-to-drain subthreshold

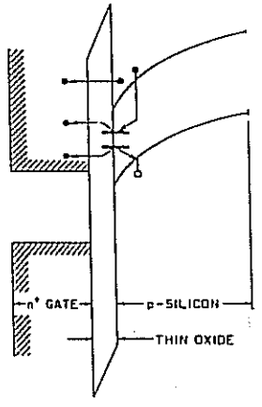

Fig. 3-19 Energy-band diagram for the phenomenon of direct tunneling through the gate oxide for thin oxides. Also shown are both unassisted-tunneling and some possible interface-trap-assisted leakage paths, with trap levels indicated by short, solid bars.

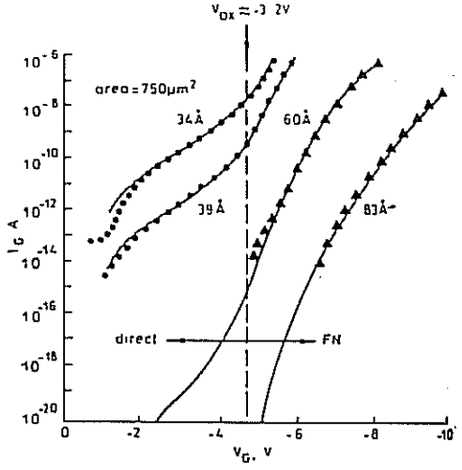

Fig. 3-20 Theoretical (and experimental) tunneling I-V curves of Al-gate *n*-channel MOS structures under negative gate-bias, illustrating the I-V characteristics of Fowler-Nordheim tunneling in 6-nm- and 8.3-nm-thick oxides and the calculated direct tunneling- currents in 3.9-nm- and 3.4-nm-thick oxides.[19] (© IEEE 1983)

leakage current (e.g., I_{off}/W = 10 nA/μm). That is, a MOSFET with a 1-μm gate length and width, would have a channel area of 1-μm^2. Thus, to keep its value equal to the subthreshold source-to-drain leakage current, the tunneling leakage current density should not exceed 10^{-8} A/μm^2 (which is equal to 1 A/cm^2).

From Fig. 3-21, the minimum oxide thickness that could be used to achieve a gate current density of 10^{-8} A/μm^2 is 1.4-nm.* Since this is the thickness of the oxide needed for the 70-nm generation (according to the L_E/t_{ox} guideline given at the beginning of Sect. 3.1 of this chapter), it appears that alternative gate-dielectric materials with reduced gate-leakage will be required for 50-nm MOSFETs and beyond. (Another, more detailed analysis of the leakage-currents in CMOS technology in the 100-nm regime arrives at the same oxide-thickness limit due to tunneling-leakage as the one given above [i.e., 1.4-nm].[30]) This topic is considered in Chap. 4. It should also be noted that logic circuits can usually withstand higher leakage currents than can DRAMs. Therefore. the gate oxide

* Note that the data shown in Fig. 3-21 are the results of work by researchers at Agere Systems.[20] They used an *in-situ* cleaning of the silicon surface (with HF gas, followed by ultra-violet [UV] cleaning) before the rapid-thermal-oxidation. The ultra-thin oxides they grew using this cleaning sequence resulted in oxides with very-low gate leakage currents (Fig. 3-21). Ultra-thin oxides grown by wet rapid-thermal-oxidation (RTO) processes have also reportedly shown smaller gate leakage currents (see Sect. 3.6.2).

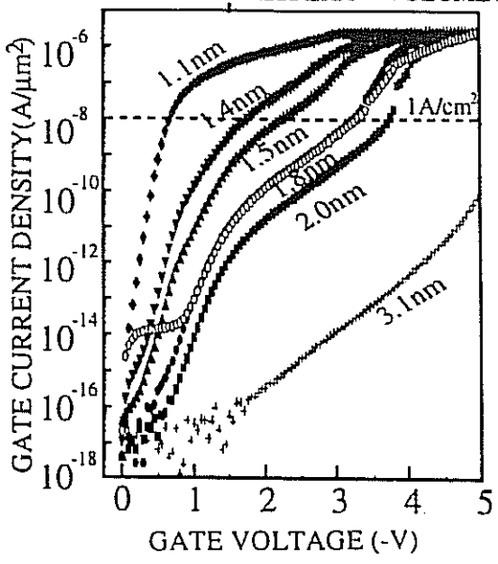

Fig. 3-21 Negative gate-current-density as a function of negative gate-voltage for oxide thicknesses from 1.1-nm to 3.1-nm, showing gate tunneling-leakage-current-densities.[20]

limit for DRAMs is expected to be larger (about 3-nm will be the minimum oxide thickness for DRAMs).

3.5 MODELS OF THIN-OXIDE GROWTH

The Deal-Grove model described in Vol. 1 for the growth of SiO_2 provides excellent agreement with experimental observations for thick oxides.[21] That is, the rate-constants derived in the Deal-Grove model allow the thickness of oxide films >350-Å to be well predicted as a function of temperature, furnace-ambient, silicon-doping-concentration, and silicon-crystal-orientation. However, the model does not give a detailed understanding of the mechanisms that produce these dependencies, nor is it valid when the oxide is thinner than 200-Å. Specifically, the experimental measurements of oxide-growth-rates in dry-O_2 are not accurately predicted by the Grove-Deal for oxides less than about 200-Å (20-nm). This is the so-called *anomalous regime* of the Deal-Grove model. This is particularly troublesome because gate-oxide-thicknesses of $\leq$ 100-Å are used for sub-micron MOSFETs. Figure 3-22 depicts a computer simulation of the formation of the first few layers of oxide being grown on a silicon surface.[22]

Several physical mechanisms have been proposed as models for the enhanced oxidation that is observed when oxides thinner than 200-Å are grown in dry-O_2, including:[23]

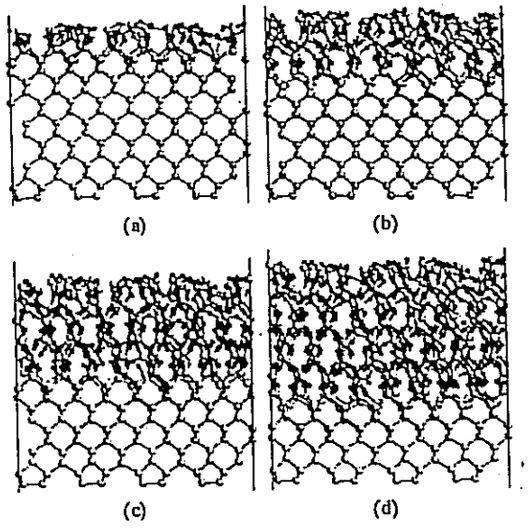

Fig. 3-22 Si(100)/SiO_2 interface structures after oxidation of a) 2nd b) 4th c) 6th d) 8th layer.[22]

1. Models based on space-charge effects where the oxidation-enhancement in the thin regime is due to field-assisted oxidant-diffusion (arising from O_2^- coupled diffusion with holes).

2. Models based on the existence of structural defects in the oxide, such as micropores or channels (~10-Å in diameter). These would allow the oxygen molecules to diffuse more rapidly when the oxide is thin.

3. Models based on parallel O_2 and O reactions at the Si/SiO_2 interface.

4. Models which postulate that a blocking layer exists near the Si/SiO_2 interface which forms as the oxide grows and slows down the rate for thicker layers.

5. A model by Massoud *et al.*, based on the presence of a thin surface layer in the silicon where additional reaction sites are available. The concentration of these sites decays exponentially with a characteristic decay length L_2 in Eq. 3-9.[23]

Despite the large number of models proposed to explain oxide growth in this very-thin regime, none has yet been shown to be clearly correct, and thus none have been convincingly accepted as more valid than the others. We are therefore left with several empirical models that yield reasonable agreement with measured values of oxide thickness in the thin regime, but with no completely satisfactory

physical explanation. Additional work in the future may help clarify some of these issues.

One *empirical* model for thin oxide growth was proposed by Reisman and Nicollian in 1987.[59] By analyzing a vast amount of data, they derived an empirical model that calculates the oxide thickness versus time using a general power law of the form:

$$t_{ox} = a\,(t_g + \tau)^b \qquad \textbf{(3.7)}$$

where a and b are constants, and t_g is the growth time measured in a given experiment, and τ is the time to grow an oxide of thickness t_{ox_i} already present on the surface. Their model could fit all published dry-O_2 data. In fitting this equation to experimental data, they extracted b values between 0.25 and 1.0, depending on temperature and oxidant-partial-pressure. This equation has two fitting parameters (a and b), just as the Deal-Grove model has two (A and B/A). However, there are significant differences between the models. First, the Reisman-Nicollian model is claimed to be able to fit data down to oxide thicknesses of essentially zero thickness. There is no anomalous regime, as exists with the Deal-Grove model. Second, each model has its own physical basis for the oxide growth process. The Deal-Grove model is based on the idea of oxidant-diffusion and an interface-reaction - with each process dominating the growth under different conditions. Reisman suggests that the interface-reaction actually controls the oxidation process at all times, and that the volume expansion necessary at that interface to accommodate the growing oxide was provided by viscous-flow (relaxation) of the oxide layer. The time-dependent viscous-flow of the oxide in the Reisman model, is used to explain the extracted pressure- and temperature-dependence of the parameters in that model (a and b). A third model by Han and Helms (which we will not consider in detail here), has also been proposed to predict oxide growth. We will, however, compare the results of the Reisman and the Han and Helms model to the Grove-Deal model in upcoming paragraphs.

A comparison between the Reisman-Nicollian model of oxide thickness and the Deal-Grove model is shown in Fig. 3-23. Data from Deal and Grove at 700°C are plotted on linear scales and Eq. 3-7 is plotted with measured values of a and b. Shown for comparison is Eq. 3-7 with the appropriate constants obtained by a fractional weighted least squares fit to the data.[24] Kouvatsos extracted the a and b parameters of Eq. 3-7 at 950°C and 1100°C for oxide growth in O_2 diluted in He, and found very good fit to the Reisman-Nicollian power-law model.[25]

Figures 3-24 and 3-25 again compare the predictions of the Reisman and the Han and Helms models with those of the Deal-Grove model. The model parameters for these plots for the Deal-Grove model are the standard A and B/A values for 1 atmosphere. For the Reisman model at 800°C, a = 0.302 nm, b = 0.704, and τ = 13.1 min. At 1000°C a = 3.02 nm, b = 0.701, and τ = 1.26 min. At 800°C (Fig. 3-24), the differences in these models for very-thin oxides is clear.

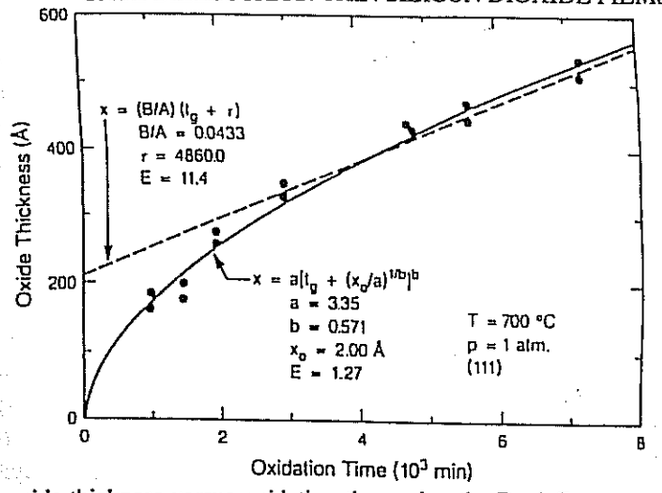

Fig. 3-23 Fits to oxide thickness versus oxidation time using the Deal-Grove model and the Reisman-Nicollian power-law model expression.[24]

The Reisman (and the Han and Helms) model agree fairly well with each other and are a good match to experimental data in the thin regime. The Deal-Grove model does not do a good job for thin oxides, either with τ set at 8 hours

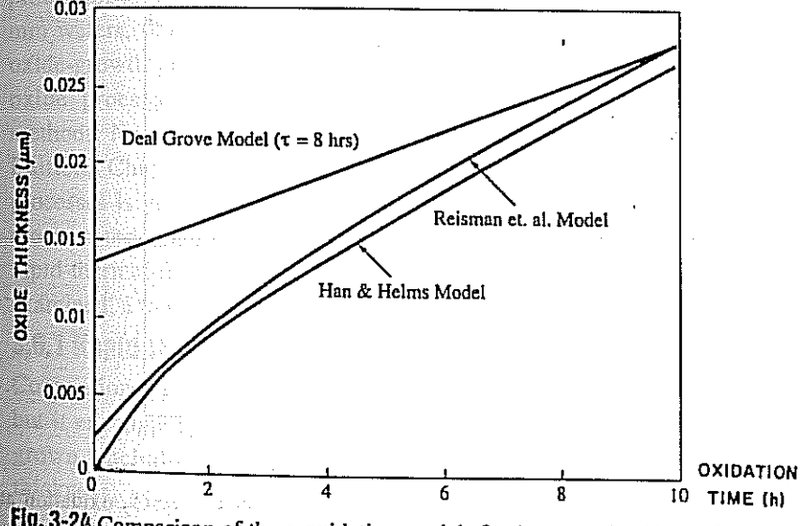

Fig. 3-24 Comparison of three oxidation models for 1 atmos, dry-O_2 oxidation at 800°C.

(the value Deal and Grove extracted in their original work), or with $\tau = 0$ (which predicts even thinner oxide than do the Reisman or Han/Helms models). At 1000°C (Fig. 3-25), the differences in the models in the thin regime are less apparent since much thicker oxides grow at the higher temperature on the same time scale. For the particular parameters chosen for this plot, the Reisman-Nicollian model diverges from the predictions for thicker oxides.

Despite the fact that the Deal-Grove model in its original form does a poor job at predicting oxide growth in dry-O_2 in the thin-oxide regime, their model can be "fixed" to do a much better job. Such a"fix" was developed by Massoud *et al.*, who demonstrated that a much better fit can be accomplished than with the Deal-Grove model alone (by adding an additional term that decays exponentially with thickness to the oxidation rate.). The expression for the oxide growth-rate in the classical Deal-Grove model dt_{ox}/dt (see Vol. 1, Chap. 7) given by:

$$(dt_{ox}/dt) = B/(2t_{ox} + A) \tag{3.8}$$

then becomes

$$(dt_{ox}/dt) = \{B/(2t_{ox} + A)\} + C\exp(-t_{ox}/L_2) \tag{3.9}$$

where

$$C = C^0 \exp(-E_A/k_BT) \tag{3.10}$$

where C^0~3.6×10^8-μm/hr, E_A ~2.35-eV, and L_2 is the characteristic decay length described in paragraph 5 above. For dry-oxidation of lightly doped substrates in

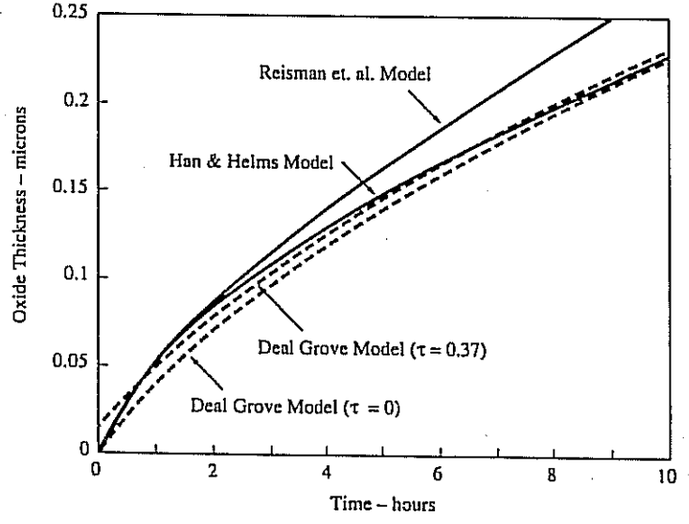

Fig. 3-25 Comparison of three oxidation models for 1 atmosphere, dry-O_2 oxidation at 1000°C. The dashed lines are the Deal-Grove model.

the 800-1000°C range, L_2 was found to be ≅7-nm, independent of surface orientation [i.e., the above numbers apply to either (100)- or (111)-oriented silicon substrates]. Heavy doping was found to only slightly affect the oxidation-rate-enhancement in the thin regime. Equation 3-9 has been inserted into SUPREM III and SUPREM IV as the model for thin-oxide growth in dry-oxygen. With Massoud's addition to the Deal-Grove model, its predictions match the data well over the whole thickness range. Note that it was decided to implement the Massoud model (but not the Reisman model) in SUPREM III and IV. This is because many other oxidation-effects have also been tied to the basic Deal-Grove model, and these effects are also modeled in SUPREM III and IV.

While the debate continues about the physical mechanisms responsible for thin oxide kinetics, the problem has taken on a new importance. This is because deep-submicron MOSFET gate-oxides are routinely grown with thicknesses in the anomalous regime. This is an example of an industrial application outpacing basic scientific understanding. From the industrial view, it is more important to be able to grow thin oxide layers reproducibly and uniformly (and with good electrical properties), than it is to understand the governing physical principles.

3.6 SINGLE-WAFER TECHNOLOGY OF THIN OXIDE GROWTH

Thin oxides can be grown using a batch process or a single-wafer process. Batch processes currently use vertical furnaces (200-mm wafers), and such tools are discussed in more detail in Vol. 1, 2nd Ed. With the advent of 300-mm wafer processing and the need to grow uniformly-thick 20-nm oxides across such 300-mm wafers, the single-wafer process may begin to make inroads against batch-oxide-growth processes. Here we will limit the discussion to such single-wafer oxide processes, which are also based on rapid-thermal-oxidation (RTO) methods. Note that single-wafer RTP processes are also described in Vol. 1, 2nd Edition, Chap. 7, and a more comprehensive review of RTP systems and processes is given in Ref. 26. However, in both of these references, the focus is more on the general aspects of RTP systems. Here our focus is on the details of rapid-thermal-oxidation (RTO) tools and processes. Wet-RTO processes are also examined, a topic not covered in Vol. 1, 2nd Ed.

3.6.1 Rapid Thermal Oxidation Tools

Conventional furnace-based oxidation has been the workhorse of the industry, but it encounters some problems when gate oxides thinner than 2.0-nm must be grown, including: 1) furnace oxidation is difficult to scale below 2.0-nm thickness; and 2) the lower temperatures used during such ultra-thin oxide growth processes degrade the oxide quality.

RTO is able to offer better quality ultra-thin oxides when compared to those grown in furnaces for two reasons: 1) RTO grows the oxides at higher temperatures; and 2) RTO is more conducive to cluster tooling. RTO can grow thin oxides at high temperatures because the oxidation time is kept down to just a few seconds (whereas the minimum oxidation time in a batch furnace is at least a few minutes).

There are several advantages to growing oxides at higher temperatures. First, oxides exhibit smaller gate-leakage-currents as their growth temperature increases (up to a maximum of 1050°C, beyond which leakage-currents increase).[27] Second, they show higher resistance to boron penetration. In one study, 2-nm-thick oxides grown in a furnace at 800°C were compared to RTP-grown oxides at 1000°C. The results indicated that boron penetration was suppressed better by the RTP-grown oxides.[28] The third advantage is that MOSFETs built with the RTP-grown oxides had higher effective carrier mobilities (for both electrons and holes) as the temperature of the oxide growth process was increased from 800°C to 1000°C.[29] This was attributed to the smoother Si/SiO_2 interface that results from growing the oxides at higher temperatures.

Cluster tools can potentially produce better-quality oxides because they allow process sequences to be carried out in an integrated fashion. For example, in oxide growth processes, a cleaning procedure is done prior to the growth step. If this can be done *in situ,* the quality of the oxides can be improved. Clustering is easily achievable with single-wafer RTP systems, but is impractical with batch furnace tools. By integrating RTO chambers with other processing modules it may be possible to extend the use of oxides to additional technology generations.

As noted above, with the introduction of single-wafer RTO-systems, it becomes possible to carry out *in situ* surface preparation, wafer rotation, ambient control, and process integration. Because such systems can grow oxides at such rapid rates (i.e., more than 1000-Å/min), it also makes RTO capable of competing economically with vertical hot-wall furnaces.

Figure 3-26 is a cross-sectional view of an RTO reaction-chamber depicting the necessary components of a lamp-heated RTO-system. In this reactor, the wafer is inserted into the reaction chamber by a robot wafer handler onto a rotating SiC support ring. The wafer is rapidly heated at rates of 50-100°C to the process temperature (800-1050°C). Process gases are introduced at reduced pressure in an axial flow pattern where they react to form water vapor, hydroxyl ions, oxygen ions, and other species (depending on the H_2:O_2-ratio and whether N_2, NH_3, HCl, or other reactant gases are introduced). At the conclusion of the oxidation and annealing steps, the wafer is cooled rapidly to a temperature where

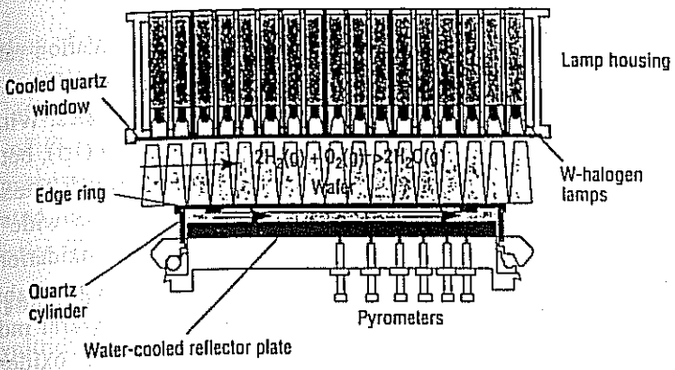

Fig. 3-26 Cross-section of an Applied Materials ISSG RTO reactor.

it can safely be removed by the wafer handler from the reaction chamber to an additional cool-down chamber that can be used to enhance system throughput. Cool-down can be enhanced further by providing a wafer backside cooling gas, such as helium or hydrogen.

3.6.2 Wet RTO Processes

As will be discussed in Sect. 3.8, traditional dry oxides are expected to be limited to minimum thicknesses somewhere between 1.6-nm[36] and 2.5-nm.[37] Various techniques are being explored to reduce this thickness limit. One of these methods employs wet-RTO. A technique pioneered by Applied Materials called *in-situ steam generation* (*ISSG*) has reportedly allowed oxides in the 2.0-nm thickness range to be formed with improved reliability compared to dry oxides of the same thickness.[38] ISSG-oxidation is a low-pressure process (typically below 20 torr, for safety reasons) performed in a cold-wall RTP reactor.

Unlike hot-wall furnaces that employ a pyrogenic torch for growing wet oxides, no hot quartz is used to generate the steam ambient (eliminating the problem of quartz devitrification). Instead, in ISSG-oxidation, H_2O, OH and atomic oxygen are generated directly in the reaction chamber without pre-combustion. The hot wafer acts as an ignition switch, causing reactions among electronic-grade H_2 and O_2 gases (that are pre-mixed in a plenum before being injected into the RTP chamber). This configuration reduces risk of metal contamination, since no metal catalyst is used in the in-situ steam generation method.

The oxidation-rate rapidly *increases* as the pressure is decreased below 20 torr (unlike in conventional oxidation processes where oxidation-rates decrease with decreasing pressure). The growth-rate at 10-torr is nearly twice that of atmospheric dry-rapid-thermal-oxidation. Modeling studies indicate that the oxidation-rate exhibits a strong correlation to the atomic-oxygen-concentration

and not to any other atomic or molecular species. The oxidation-rate is also strongly temperature-dependent compared with wet-atmospheric-oxidation.

Oxides grown with the ISSG process show good characteristics. Within-wafer uniformity and wafer-to-wafer repeatability are well below 1% (1σ), both for thin (< 40-Å) and thick (> 100-Å) oxide-films. The faster growth-rates improve productivity while offering a thermal-budget reduction. The wide range of temperatures and hydrogen concentrations allow controlled oxidation for any thickness of oxide. The ISSG-oxides also demonstrate better reliability than comparable furnace-oxides, or those created by dry-RTO. Several independent measurements of charge-to-breakdown of 30-50-Å wet-RTP oxides show a threefold (or more) improvement over furnace-grown-oxides. The wet-RTO-oxides also exhibit about an order-of-magnitude less gate-leakage-current than dry-RTO oxides, which are comparable to furnace-grown-oxides (Fig. 3-27).[76]

3.7 NITRIDED AND FLUORINATED OXIDES AS MOSFET GATE DIELECTRICS

One modification to the basic SiO_2 material that has recently been intensely examined is the addition of nitrogen to the oxide. That is, oxynitrides have been studied as replacements for silicon dioxide as the gate dielectric of MOSFETs for thicknesses below 4.0 nm. The advantages of oxynitrides over silicon dioxide for this application are the following: 1) they exhibit improved boron penetration resistance; 2) they have better breakdown characteristics; 3) they show greater immunity to hot-carrier-injection effects; and 4) MOSFETs made with such layers exhibit improved high-field channel-electron-mobility. Improved performance of nitrided silicon-dioxide gate-dielectrics has been associated with the accumulation of a small amount of N at the Si/SiO_2 interface. Interfacial N-accumulation is thought to reduce interfacial defects by relieving strained Si-O bonds and "dangling" bonds. An interfacial peak-concentration of about 1-2 atom % is typically associated with higher breakdown electric fields and lower fixed-oxide charge densities. But, a higher N-concentration (> 5 atom %) is needed to effectively suppress boron penetration through the dielectric layer.

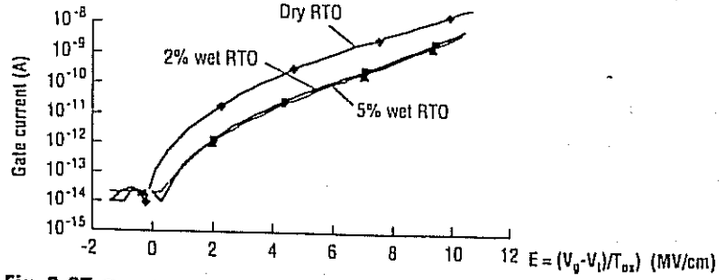

Fig. 3-27 Gate leakage current comparing dry-RTO to 2% & 5% ISSG RTO for 30Å oxides.[76]

Nitridation of oxides is considered to be a variation of silicon oxidation, and such films are formed mainly though the nitridation of silicon oxide or the oxidation of silicon in a nitrogen containing ambient such as NH_3, N_2O, NO (or mixtures of O_2 and NO). Such films have been formed in high-temperature furnace processing or rapid-thermal-processing tools. Fluorine can also be incorporated into gate oxides as a method for improving its dielectric characteristics. This subject is also discussed in a following section.

3.7.1 Oxynitridation of Silicon in N_2O

Early oxynitride films were formed by nitriding a thermally grown oxide film by exposing it to an NH_3 ambient at high temperatures, and then re-oxidizing the nitrided oxide to reduce the amount of hydrogen that remains in the film after NH_3 nitridation. Such oxynitrides were thus called *reoxidized nitrided oxide* (*ROXNOX*) films.

Although ROXNOX films exhibit better hot-carrier resistance than do pure oxides, some practical issues make them troublesome to manufacture. First, the ROXNOX formation sequence is complex and difficult to optimize, especially for CMOS applications. For example, the combination of nitridation and reoxidation steps used in the ROXNOX process may leave excess hydrogen in the film. In that case, the density of electron traps will be higher than in a pure oxide film. Since electron trapping is the dominant degradation mechanism in PMOS devices, PMOSFETs with such films will exhibit worse hot-carrier degradation.[39] On the other hand, if the reoxidation step is excessive, the film can lose its strength against interface-state generation by hot-carriers. Thus, NMOSFETs containing such films would be prone to more hot-carrier degradation. The gate dielectric layer might also grow too thick. Another drawback is that the best NH_3-nitrided films always exhibit higher densities of positive fixed charge (Q_f) than do the best SiO_2 films. The fixed charge arises as a result of the nitridation process. Finally, the ROXNOX process generally involves multiple high-temperature steps, which are undesirable for ULSI CMOS applications. As a result, ultra-thin oxynitrides are not formed in production by using NH_3 as the nitrogen source.*

A solution to some of the problems of NH_3-nitrided films is to avoid the incorporation of hydrogen in the film in the first place. One way to accomplish this is to replace NH_3 with N_2O as the nitriding ambient. Such N_2O-based techniques have indeed proved to be more fruitful because of the hydrogen-free nature of the processing.

* Interested readers can find more details on ROXNOX films in Vol. 3.

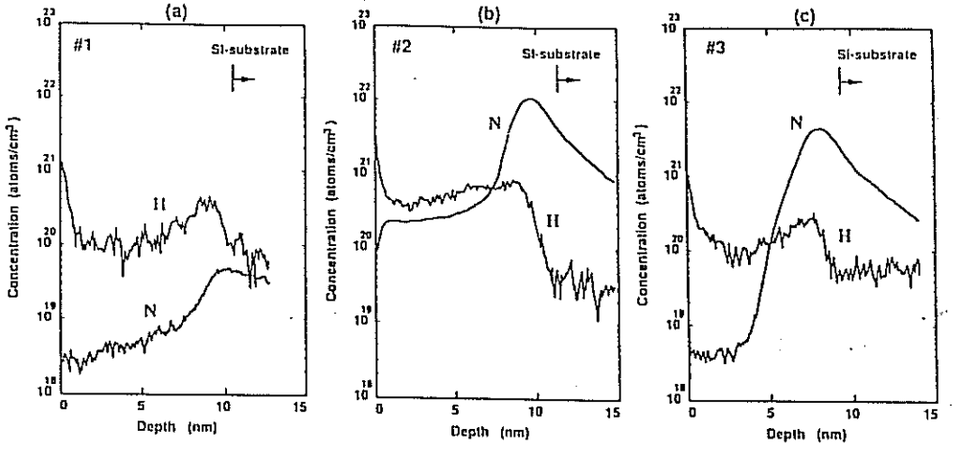

Fig. 3-28 SIMS depth profiles of N and H in three kinds of oxide films: (a) pure SiO_2 (#1); (b) NH_3-nitrided SiO_2 (#2); and (c) N_2O-nitrided SiO_2 (#3).

This approach was first used to grow a gate dielectric by directly reacting N_2O at the silicon surface in an RTP system.[40,41,42,43] Since there is no hydrogen in N_2O, this process results in films with significantly less incorporated hydrogen than NH_3-nitride films. Figure 3-28 shows SIMS profiles of dielectric films prepared by: (a) thermally oxidizing silicon; (b) nitriding an oxide film in NH_3; and (c) growing a film on silicon by RTP in an N_2O ambient.[41] It can be seen that the hydrogen concentration in the N_2O film is significantly smaller than in the NH_3 film. It is also evident that the nitrogen concentration in the N_2O film is highest at the Si-SiO_2 interface. Early studies of such films concluded that they are self-limiting in thickness, as is shown in Fig. 3-29a.[40,44] The build up of the nitrogen concentration was thought to block the oxidation reaction that occurs simultaneously at the interface. Later work showed that the thickness was not self-limiting (see Fig. 3-29b),[45,46] and that the thickness in fact, increases approximately as (time)$^{1/2}$. This seems to indicate that the growth is instead limited by the diffusion of N_2O through the oxide bulk (rather than through the nitrogen-rich layer).[47] The discrepancy in the apparent growth mechanisms may have been explained by Okada, who observed significant retardation of the growth only at relatively high (~0.9 at%) interfacial nitrogen concentrations, which is higher than observed in some N_2O furnace processes.[59] However, the interfacial reactions may still dominate growth rates for short times and thin oxides.

Nevertheless, the nitrogen profile in the N_2O film is similar to that in the ROXNOX film, yet this is achieved with only a one-step process (i.e., no

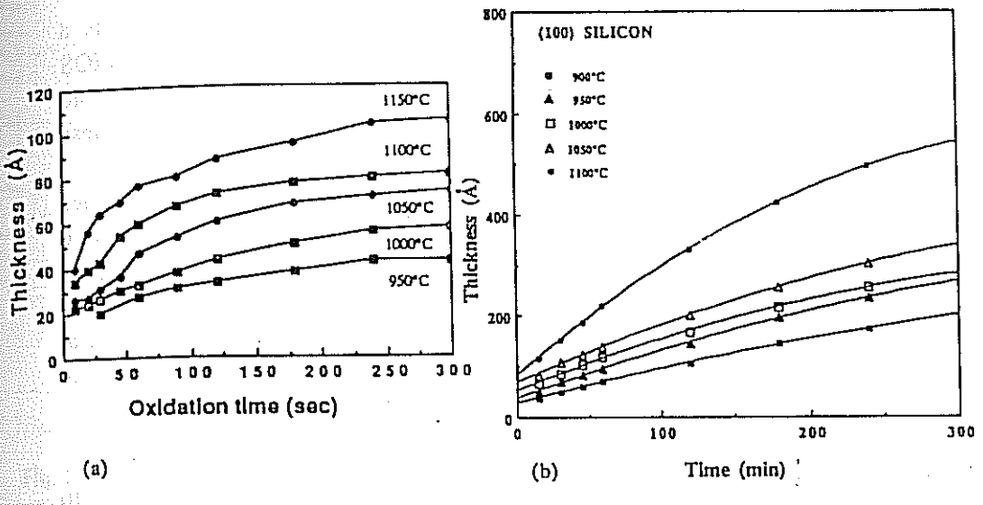

Fig. 3-29 (a) N_2O oxynitride thickness versus oxidation time using RTP for various process temperatures. Ramp-up rate was 200°C/sec. Oxide thickness was measured at a fixed refractive index (1.46)[40] (© IEEE 1990); (b) N_2O oxynitride thickness versus time in a furnace-anneal process. (© Electrochemical Society 1993)[45]

reoxidation is necessary). The peak nitrogen concentration (at the Si/SiO_2 interface) is ~0.5%, which is less than that found (2-10 atom%) in SiO_2 furnace-nitrided in NH_3 under atmospheric pressure. Furthermore, residual hydrogen incorporation into the film is no longer a concern. This is evidenced by the low electron trapping density exhibited by such films. NMOSFETs fabricated with such films also exhibited less tendency toward hot-carrier degradation than those made with control oxides (i.e., they showed smaller values of ΔD_{it}), and were also less susceptible to the creation of electron traps throughout the oxide under hot-carrier stressing.[48]

However, several limitations were still observed in these N_2O-grown oxynitrides. First, larger initial values of D_{it} were observed than in pure oxide films (Fig. 3-30), even though the rate of their generation under hot-carrier stressing was much smaller (see also Fig. 3-30).[48] Second, the initial Q_f values were also somewhat higher. Third, the slow growth-rate of such films make them only practical for ULSI applications which use dielectric films thinner than 100Å (see Fig. 3-29). Fourth, the maximum mobility of electrons in MOSFETs made with such films is ~5% less than in MOSFETs with oxide gate dielectrics (but the mobility is actually higher at large lateral electric fields - see Fig. 3-31).[49] The most severe limitation, however, is that such RTP N_2O-grown films exhibit

unacceptably large nonuniformities in composition and thickness across a wafer.[50] Such non-uniformities would limit the use of such films in ULSI MOS devices. Temperature nonuniformities in the RTP chamber, gas depletion, and heat-transfer effects were some of the phenomena suspected of leading to these nonuniformities.

Nevertheless, it was found that growth of such N_2O dielectrics in a furnace (rather than by RTP) resulted in films with thickness uniformities comparable to those of pure oxide films, and with adequate compositional uniformity as well.[51,52] The process described in these reports was a 950°C furnace step in N_2O, which grew 60-Å-thick dielectric films having a nitrogen concentration at the Si/SiO_2 interface of ~3%. Hot-carrier degradation of PMOSFETs with such films was excellent since electron-trap-densities in these films was also low.

3.7.2 Oxynitridation of Oxides in N_2O or NO

Another method used to form thin oxynitride layer involves a two-step operation. Oxides are first grown in O_2, and are then subjected to an oxynitridation step in N_2O or NO. When N_2O is used as a post-oxidation-annealing ambient, a wide range of final oxide thicknesses is achievable with much smaller thermal budgets. Processes using this two-step N_2O sequence have been described using an atmospheric-pressure-furnace,[46,53,54,55] a low-pressure-furnace,[56] and RTP conditions,[57] Figure 3-32 shows an example of such a two-step process performed in a conventional furnace at atmospheric pressure.[54]

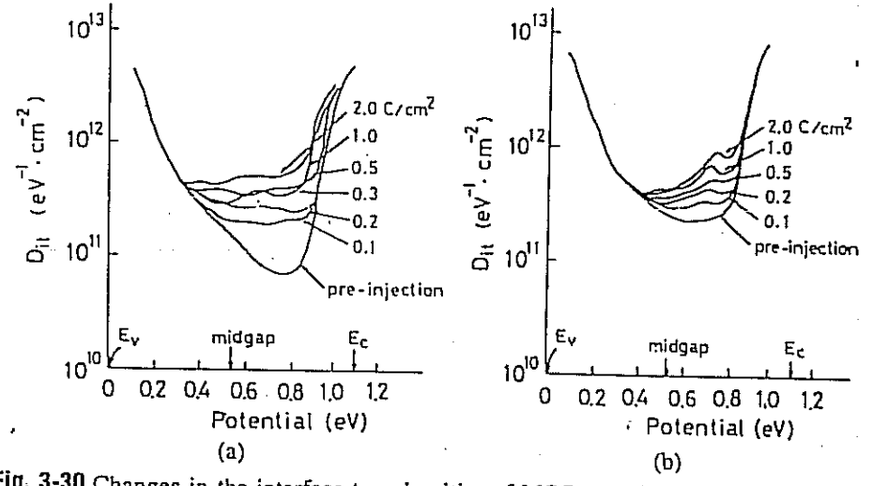

Fig. 3-30 Changes in the interface trap densities of MOS capacitors with: a) pure SiO_2; and b) SiO_xN_y by Fowler-Nordheim electron injection. Electrons were injected from the gate electrodes into the oxide films at a current density of ~10 mA/cm².[42] (© IEEE 1991)

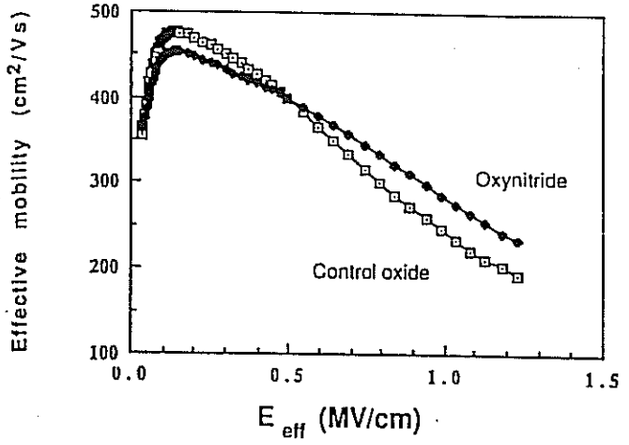

Fig. 3-31 Effective electron mobility versus effective normal field of NMOSFETs with a 65-Å-thick gate dielectric. The mobility values were calculated using the linear region I_D versus V_{GS} curves (V_{DS} = 100 mV), with a Z/L ratio of 75-μm/60-μm. The maximum effective mobility of the oxynitride is 5% lower than that of the control oxide. However, the high-field mobility of the oxynitride shows a 10% improvement.[49] (© IEEE 1991)

Typically, the oxide is grown in dry-O_2 in a furnace at 850-900°C. The N_2O step is usually carried out at 800-850°C for 5-40 minutes. In the process described in Ref. 55, 50-Å SiO_2 films are grown in dry-O_2 at 850°C for 30 minutes. During the second step in N_2O at 950°C, an additional 35-Å of oxide is grown to form films with a final thickness of 85-Å.

PMOSFETs made by oxynitridation of SiO_2 in N_2O also exhibit enhanced

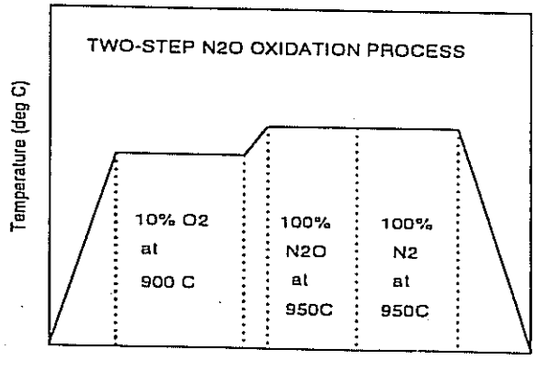

Fig. 3-32 Two-step oxidation & N_2O nitridation process in a conventional furnace.[54] (© IEEE 1993)

hot-carrier immunity compared to devices made with pure oxide. This is thought to be due to the fact that electron trapping is suppressed in devices with such N_2O dielectrics.[58] Consequently, it has been suggested that the optimum N_2O oxynitride process is oxidation in O_2 followed by nitridation in N_2O, rather than oxidation in N_2O alone.

In the low-pressure approach[56] it was reported that a minimum pressure of between 1 and 50 torr of N_2O was needed before a measurable amount of nitrogen was incorporated into the oxide. However, at low pressures of N_2O (i.e., just beyond the minimum needed for the onset of nitrogen incorporation), the concentration of nitrogen at the interface could be well controlled, but at the same time, minimal additional oxide was grown during the oxynitridation step. Thus, tight control of the final film thickness could be maintained.

As noted in a previous paragraph, the oxide film continues to grow during the second (i.e., N_2O) step. However, if re-oxidation of N_2O oxides is carried out after they are formed, unusual behavior occurs. That is, new SiO_2 grows at the Si/SiO_2 interface, and the existing N-rich layer remains intact and is displaced away from the interface by growth of the new oxide (see Fig. 3-33).[59]

Using N_2O as the nitrogen source during the second step of this procedure was found to have some disadvantages. The concentration of N-incorporation may be

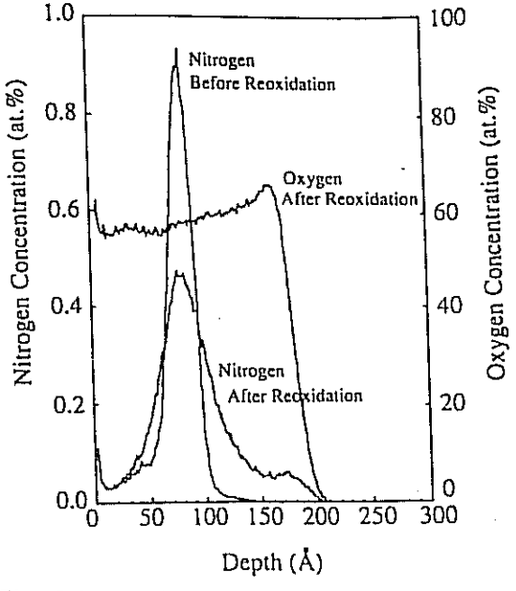

Fig. 3-33 SIMS profile of the oxynitride film grown from 100% N_2O at 1100°C before reoxidation, and after 85 min reoxidation.[59] By permission of the Electrochemical Society.

too low for some applications, especially suppression of boron penetration, unless a much higher thermal budget is permitted.

More recently oxynitrides formed with the two-step process using NO as the nitrogen source have been reported. Generally, the results for NO oxynitrides overcame the problems of using N_2O as the nitrogen source, in that they offer a lower thermal budget process for incorporating higher concentrations of N, as well as better thickness control for scaled dielectrics.[60,61] It is thought that nitrogen incorporation in the N_2O process is driven by NO species, which is a product of N_2O molecule dissociation. The N_2O molecule likely dissociates into NO (4.7%), N_2 (64.3%), and O_2 (31%), according to calculations by Tobin *et al.*[62] Such by-products act in competition. That is, while the NO incorporates nitrogen, O_2 continues the oxidation by reacting with the silicon substrate, and N_2 reduces the partial-pressure of the nitridation species (which increases the thermal budget). The parallel oxidation explains the observed increase of oxide thickness during annealing in N_2O. The high temperature needed for N_2O dissociation (> 850°C) mandates the high thermal budget needed to get acceptable results. Good quality gate-dielectric-films have reportedly been produced with RTP processes using $NO + O_2$ mixtures.

The percent nitrogen incorporation also depends on the relative amount of NO in the gas mixture. Figure 3-34 shows the dependence of N incorporation on NO concentration for both direct (1-step), and 2-step oxynitride growth carried out in a $NO + O_2$ mixtures using RTP processes. This indicates that NO can produce

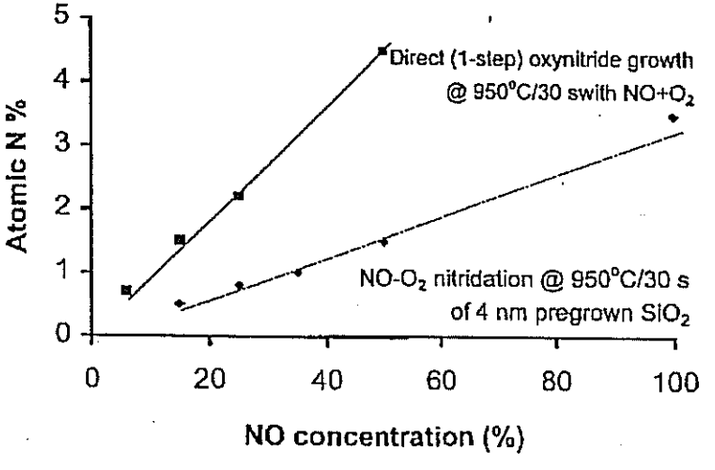

Fig. 3-34 Dependence of nitrogen incorporation in oxide on NO concentration in $NO + O_2$ mixtures.

oxynitride films with a wide range of incorporated N, allowing the films to be tailored for specific applications.

In another reported process, 4.0-nm-thick oxynitrides were formed in a two-step furnace operation. First, 4.0 nm oxides were grown in O_2 at 800°C with a 6% HCl, and a post-oxidation anneal in N_2. The second step involved annealing this oxide in a diluted NO ambient at 800°C for 45 min.[60] Finally, a process that uses a mixture of N_2O and NO in the reaction chamber has been described. The postulated advantage of this method is that the vertical positions of the nitrogen concentration peaks within the oxynitride layer can be adjusted by changing the N_2O + NO mixture.[64]

We should also mention that adding nitrogen to gate oxides has also been reported by several other novel techniques, including implantation of nitrogen through the polysilicon gate and drive-in annealing,[64] reoxidation of films formed by growing the oxynitride in N_2O or NO,[65] and nitridation of the Si surface by using a nitrogen remote plasma.[66,67,68]

3.7.3 Fluorinated Gate Oxides

The presence of fluorine in the gate oxide has also been reported to increase the hot-electron resistance of devices fabricated with such oxides.[69,70] Such fluorine may inadvertently be introduced from a BF_2^+ source/drain implant, or from an LPCVD-W or LPCVD WSi_x process. (In the W-CVD and WSi_x-CVD processes fluorine is produced during the reaction of WF_6 and SiH_4.[71]) Other reports confirm that deliberate incorporation of fluorine into the gate oxide (from implanting fluorine into a polysilicon film, and then diffusing it into the gate oxide) produces a more hot-electron resistant interface.[72,73] The increased hot-carrier-resistance of the F-containing-oxides is thought to be due to fluorine atoms which form Si-F bonds. These are stronger than Si-H bonds. The presence of the fluorine implies that some of the Si-H bonds at the Si/SiO_2 interface are replaced by Si-F bonds, thus yielding an interface that has better hot-carrier-resistance.

Unfortunately, while the above benefits make fluorinated oxides attractive for NMOSFETs, the presence of fluorine in the gate oxide may increase the penetration of boron (from p^+-poly gates) though the gate oxides, and this represents a potential problem for deep-submicron PMOSFETs. That is, BF_2^+ is used when the shallow source/drain-extension-region is formed, and this implant enters the poly gate. Thus, the F from this implant will diffuse through the poly and into the gate oxide during the implant activation steps. However, a recent study indicated that if the fluorine dose implanted into the poly is restricted to a medium dose ($1x10^{14}/cm^2$), no noticeable enhanced boron penetration effect is

observed. In addition, the charge-to-breakdown (Q_{BD}) characteristics of the PMOSFET capacitors fabricated with 4-nm-thick oxides are also significantly improved.[74]

Recently, a new form of "anomalous" leakage was discovered in oxides used in Flash-memory devices. Such leakage occurs at temperatures below 150°C, and thus cannot be screened by a traditional 250°C high-temperature retention bake. However, by incorporating a high concentration of fluorine in the tunnel oxide (introduced in a WSi_2 capped-control-gate process), a much higher resistance to this anomalous leakage is obtained.[75]

3.7.4 "Dual-Gate-Oxide-Thickness" Structures

An increasing number of mixed-signal and system-on-a-chip applications require dual on-chip power-supply-voltages. The higher-voltage may be used for the analog and input/output (I/O) circuitry, and the lower-voltage for the core digital-logic-devices. However, a serious reliability problem would be created if the same thin oxide needed for the high-performance logic MOSFETs was used in the devices subjected to the higher voltage (i.e., the analog and I/O circuitry MOSFETs). That is, the analog and I/O devices require thicker gate oxides than the core logic devices. (The thicker oxide also makes the I/O devices more robust against transient, off-chip electrostatic-discharge-events.) Consequently, it must be possible to fabricate gate oxides with different thicknesses on such ICs.

Early "dual-gate-oxide-thickness" processes had two oxidation steps, with a masked etch-step in between to remove the first oxide from regions intended to have thinner oxide. Newer processes use a different, simpler approach. That is, it has been discovered that the rate of thermal-oxidation on silicon is reduced by nitrogen implants.[88] Hence, nitrogen implantation into the silicon substrate prior to the oxidation process can be used to facilitate the implementation of such "dual-gate-oxide" ICs. Namely, selective implantation of nitrogen will cause oxides grown on such areas to be thinner than those simultaneously grown on non-implanted areas (Fig. 3-35a). Typical implant conditions for the nitrogen implant are energies of 10-30 keV, at doses of $2x10^{14}/cm^2$-$1x10^{15}/cm^2$. Figure 3-35b shows how the oxide thickness from an 800°C, dry-O_2 oxidation is reduced with increasing nitrogen dose.[89] It is suggested that nitrogen suppresses oxidation by forming Si-N bonds at the interface of the growing oxide, thereby constraining further growth.

3.8 PROJECTIONS OF THICKNESS LIMITS OF GATE OXIDES

In the late 1970's it was projected that the scaling limit of MOSFETs would be 0.5-μm. In the mid-1980's, this limit was reduced to 0.25-μm. In 2001, there

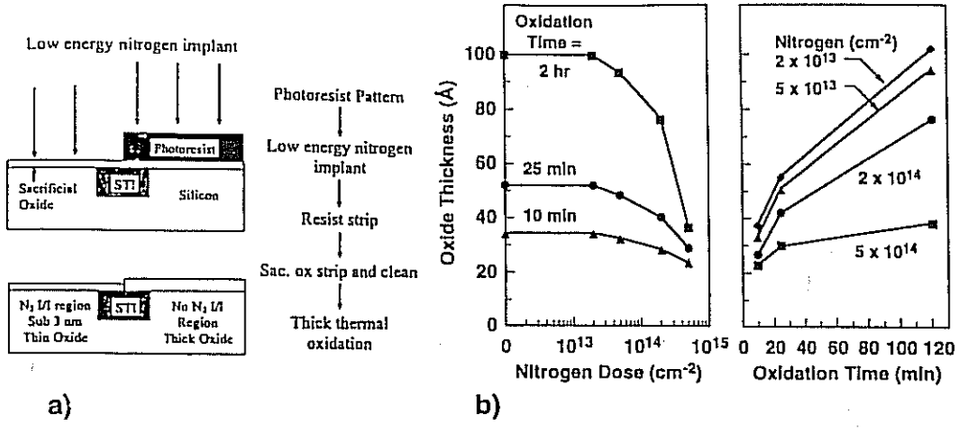

Fig. 3-35 a) Schematic of the process flow for sub-3-nm gate oxide formation on N_2-implanted substrates. b) Oxide thicknesses obtained in dry-O_2 at 800°C for various nitrogen implant doses (energy = 25 keV) and oxidation times.[89] (© IEEE 1997)

does not seem to be any device physics barrier for scaling at least to 0.05-μm (50-nm). The factor that kept the predicted limit on MOSFET gate lengths in previous decades higher than it is today was the uncertainty in the minimum thickness of the gate-oxide-layer that could be successfully implemented in MOSFETs as they were scaled. Thus, the question of the minimum-oxide-thickness that can be implemented in MOSFETs remains an important one. Here we address the predictions of the minimum-oxide-thickness as viewed in 2001.

There are a number of factors that play a role in the minimum-oxide-thickness that can be used, including: oxide-reliability, tunneling-leakage-current, stress-induced leakage-currents (SILC), polysilicon-depletion, process-induced oxide-damage, and hot-carrier-induced damage. The type of application in which the oxide is to be used also plays a role. That is, different minimum-oxide-thickness limits exist for logic devices, DRAMs, and nonvolatile memory devices. We will discuss each of these topics here, with the exception of hot-carrier-induced damage. The latter phenomenon is not treated because hot-carrier-reliability is expected to greatly improve as the gate voltage is decreased below 3-V (since at such low voltages, few carriers can gain sufficient energy to create interface-traps or shallow oxide-traps). This is one of the positive aspects of scaling devices below 0.25-μm.

3.8.1 Minimum-Oxide-Thickness Due to Defects and Tunneling

The first two (and probably the most important) phenomena that impact the minimum-oxide-thickness that will be usable in full-scale IC production are: 1) defect densities in the thin-oxide films and; 2) the direct-tunneling-current through these films. We discuss these two effects together here because approximately the same minimum-oxide-thickness limit is set by either of them.

In 1996, Hu predicted what the minimum-oxide-thicknesses would be in production ICs, depending on the power-supply voltage.[11] He based these values on the following assumptions: First, the oxides will be used in ICs designed to operate for at least 10 years. Second, the time-to-breakdown (t_{BD}) that oxides of various thicknesses will exhibit can be extrapolated to the 10-year limit by using the anode-hole-injection (AHI) model discussed in Sect. 3.3.4. Figure 3-36 shows a plot of such t_{BD} data for oxides of various thicknesses and intrinsic failure behavior (i.e., the oxides depicted in this plot are assumed to be defect-free, and thus fail according to intrinsic-breakdown failure mechanisms). For such defect-free oxide films, the following thicknesses could be expected to exhibit a 10-year time-to-breakdown behavior for respective power supply voltages: a) 80-Å for 5.5 V (5 V power supply + 10%); b) 45-Å for 3.6 v (3.3 V + 10%); c) 33-Å (3.3-nm) for 2.75 V (2.5 V + 10%).

To allow for defects in the oxide that would weaken them (i.e., the B-mode defects, which locally thin, but do not create catastrophic pinholes in the oxide), Hu suggested that a safety margin be employed. That is, he used a 30% increase

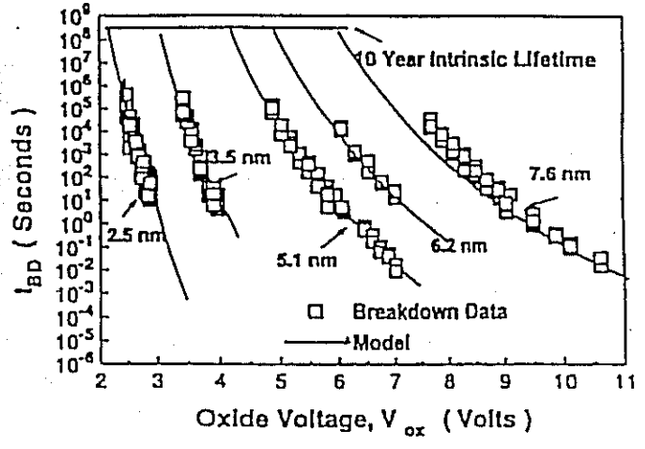

Fig. 3-36 Oxide lifetime as described the anode-hole-injection model.[11]

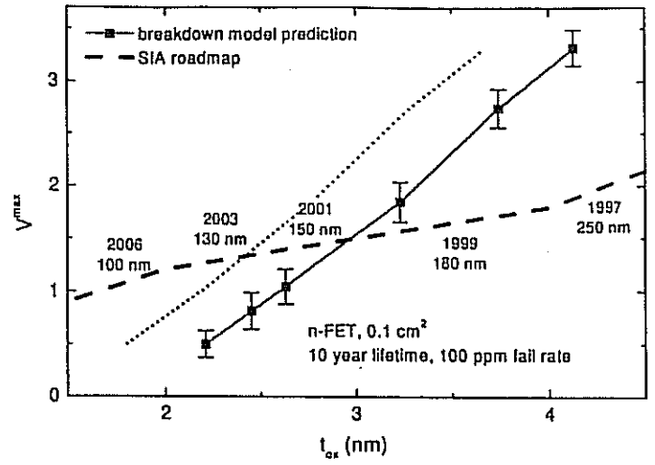

Fig. 3-37 The largest allowable supply (gate) voltage, V^{max}, as a function of oxide thickness, consistent with a failure rate of 100-ppm for a total gate area of 10 cm^2 on a chip operating at room temperature for 10 years. Solid and dotted curves correspond to t_{ox} measured by CV-extrapolation or quantum mechanical methods, respectively. The heavy dashed line is the SIA Roadmap (1997 version) for t_{ox} and V_{DD}, corresponding to the indicated year of first manufacture and technology generation.[37]

in the oxide thickness to allow for such defects. Thus, the minimum oxide thicknesses that he predicted that could be used in production were: a) 110Å for 5 V; b) 65Å for 3.3 V; and c) 45Å for 2.5 V.

Since that time, however, MOSFET scaling has been projected to even smaller gate lengths, mandating the use of even thinner gate-oxides (and smaller power-supply voltages). That is, projections for MOSFETs scaled to 50-nm (and using oxides below 2.0-nm) have been subsequently published. In 1998, Stathis and Di Maria of IBM published a study in which they projected the minimum oxide thickness that could be used in production was 2.6-nm for a 1.0-V power supply in order to meet the 10-year operating lifetime.[37] This was based on the conclusion that intrinsic oxide-degradation and breakdown would become the most significant reliability concerns, and therefore the dominant limiting factors that would set the allowable oxide thickness. Figure 3-37 shows the graph reported in that study, indicating that at 1.0 V, the minimum oxide thickness to meet a 10-year lifetime would be about 2.6-nm. This plot also implies that a 2.2-nm-thick gate oxide would no longer meet the ITRS Roadmap requirements.

In 1999, however, another study by Weir *et al.*, of Bell Labs (Lucent Technologies) indicated that if the oxide uniformity could be significantly improved, the reliability projections of an oxide at a given thickness can also be increased. That implies that thinner oxide films than were projected by Stathis and Di Maria may be possible. In fact, Weir showed that oxides with a thickness of 1.6-nm could meet the 10-year operating lifetime goal with a power-supply of 1.2 V - if they were produced with a process that allowed the film-thickness-uniformity to be increased.

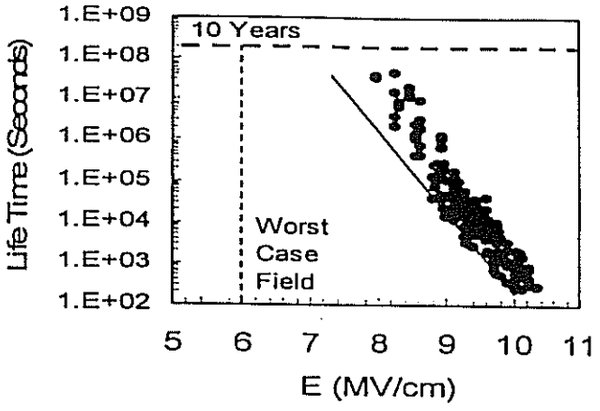

Fig. 3-38 TDDB data for 2.0-n- thick oxides at 125°C from MOS-capacitor data.[79]

In 1999, a 100-nm CMOS-technology that uses an oxide with a thickness of 2.0-nm and a power-supply voltage of 1.5 V was reported by Intel.[79] Figure 3-38 indicates that this oxide is able to meet the 10-year operating lifetime. In 2000, another report by the same group indicated it could fabricate 130-nm CMOS ICs with a gate oxide that was 1.5-nm-thick and with a 1.3-V power supply. These recent results appear to exceed the predictions of the minimum gate oxide thicknesses made just 2 years ago. At this point (circa 2002), the limit of oxide scaling does not yet seem to have reached.

The tunneling gate-leakage-current is another factor that has been predicted to put a limit on the minimum-oxide-thickness that can be used. In Sect. 3.4.2 we mentioned that at oxide thicknesses below 1.4-nm, the tunneling-leakage-current would become excessive.* This about the same thickness of oxide that has been reported to be used in the most advanced CMOS processes by 2000. Thus, both gate oxide reliability and excessive gating tunneling currents both limit the gate

* At some point this leakage will grow to be too large. That is, for SiO_2 at a gate-bias of −1 V, the leakage-current changes from $1x10^{-12}$ A/cm^2 at 35-Å, to 1x10 A/cm^2 at ~15-Å (see also Fig. 3-21). This represents a change of *twelve orders of magnitude* in leakage-current for an oxide-thickness-change of little more than a factor of 2!

oxide thickness at about the same value. (Although, it has been reported that even high levels of tunneling current do not seem to affect device reliability.[87])

3.8.2 Minimum-Oxide-Thickness Due to Stress-Induced Leakage-Current (SILC)

When thin oxides are exposed to high electric-field-stress (and consequently Fowler-Nordheim gate current), it causes them to subsequently exhibit low-field leakage. This effect, termed *stress-induced leakage current* (SILC), was first reported in 1987 (Fig. 3-39).[80] Apparently, such high-field stress generates neutral oxide-traps that facilitate electron tunneling. In addition, SILC is a transient phenomenon in thicker oxides, but is a continuous current in thinner oxides.[81,82] That is, in thicker oxides the SILC is observed to decay as traps are filled without significant tunneling out of traps. In thinner oxides, a steady-state current flows when there is an equilibrium between trap-filling and -emptying processes.[83]

The SILC is a particular problem for non-volatile memory devices because such oxide leakage reduces data retentivity. As shown in Fig. 3-40, it also tends to increase under the high-field program/erase cycles that the tunnel-oxides of such devices experience. The leakage makes nonvolatile-memory tunneling-oxide scaling much below 8-nm difficult (and perhaps impossible), unless the 10-year retention requirement is relaxed.

3.8.3 Soft-Breakdown in Thin Gate Oxides

In Sect. 3.3 we described how oxide films undergo sudden, catastrophic breakdown in which the I-V characteristic becomes essentially ohmic after being

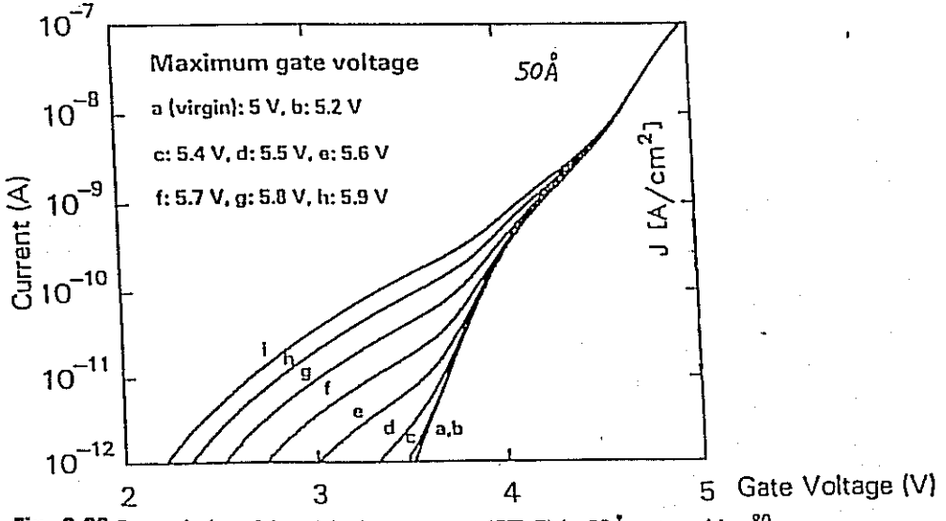

Fig. 3-39 Stress-induced (gate) leakage current (SILC) in 50Å gate oxides.[80]

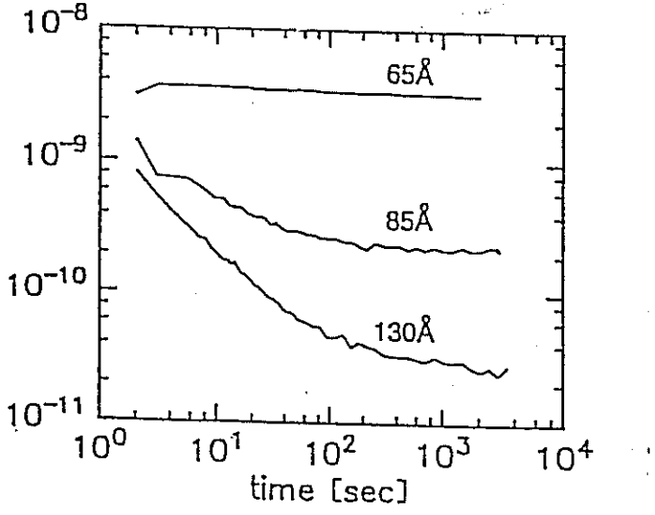

Fig. 3-40 Stress-induced leakage current is a transient current in thicker oxides but is a continuous current in thinner oxides, preventing nonvolatile memory scaling.[81]

subjected to sufficient electrical stress. However, for oxide films thinner than 5.0-nm, the breakdown mechanism may be fundamentally different. That is, another "anomalous" breakdown mode, termed *soft breakdown* or *quasi-breakdown*, occurs. Soft breakdown is detected by a much smaller change of voltage or current during stress, and a post-breakdown I-V characteristic that is given by a power law. In fact, while hard breakdown dominates for thicker oxides, it is often not observed in thin gate oxides. Instead, thin oxides appear to undergo soft breakdown. Also, soft breakdown is exhibited when oxides are stressed using lower and more realistic voltages or current densities. Soft break-down becomes "softer" and even less abrupt as the stress and oxide thickness is decreased.[90]

Figure 3-41a shows how the gate current in MOSFETs with thin (2.5-nm) oxides changes with time during voltage stressing (V_G = 4.1V and T = 140°C). Two breakdown regimes are observed: 1) soft breakdown - where the current rises suddenly to an intermediate level of ~10^{-4} A in some devices, and 2) hard breakdown – where the current rises suddenly to about 10^{-2} A. The I-V curves of a pre-breakdown oxide (before), and a post-breakdown oxide (after) are shown in Fig. 3-41b. In fact, when the thin oxide is first stressed with excess voltage, it undergoes SILC. However, further stress beyond a critical oxide-field-value causes the SILC to transition to soft breakdown. That is, soft breakdown is observed to be triggered once the oxide electric-field exceeds 10.7-11.0-MV/cm.[91]

After the oxide has undergone soft breakdown, the gate voltage becomes noisy (Fig. 3-42), but the threshold voltage V_T and transconductance g_m do not change significantly (as observed in MOSFETs with an oxide thickness of 2.5-nm). While MOSFETs with thicker oxides (5.5-nm-thick) undergo hard breakdown, their V_T and g_m also undergo little change. However, the current through the gate becomes so large that the transistor characteristics are distorted (making the MOSFETs that undergo hard breakdown unusable). Nevertheless, the only measurable effect on MOSFETs after soft breakdown is an increase in noise level in the gate voltage signal. While such noise fluctuations can couple into the drain current, for many applications this effect will not prevent MOSFETs from functioning properly. This suggests the possibility of a low-voltage technology which does not have gate-oxide breakdown as a failure mode.[90]

A later report, however, indicated that the soft breakdown mode in MOSFETs with channel lengths smaller than 1.0-μm is not observed. That is, such small MOSFETs undergo hard-breakdown.and immediate device failure.[92] Thus, being able to operate deep-submicron MOSFETs in a mode in which gate oxides do not undergo hard breakdown is not a likely scenario. This same study also demonstrated that hard- and soft-breakdown events are triggered by the same type of defect paths. The impact of soft breakdown on the reliability of ICs fabricated with deep-submicron MOSFETs is still being actively examined.

3.8.4 Impact of Polysilicon Depletion

Polysilicon depletion becomes a significant effect in MOSFETs when gate

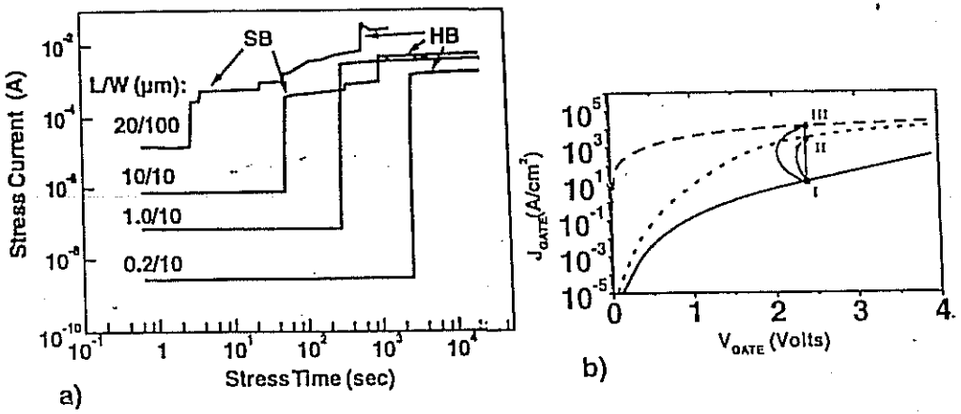

Fig. 3-41 a) Observed soft and hard breakdown in stress current curves vs. time for t_{ox} = 2.5-nm with V_G = 4.1 V at 140°C.[93] b) Oxide I-V characteristics before and after soft-breakdown.[94] (Pre-breakdown – solid line; Post-soft (short-dash); Post-hard (long-dash).

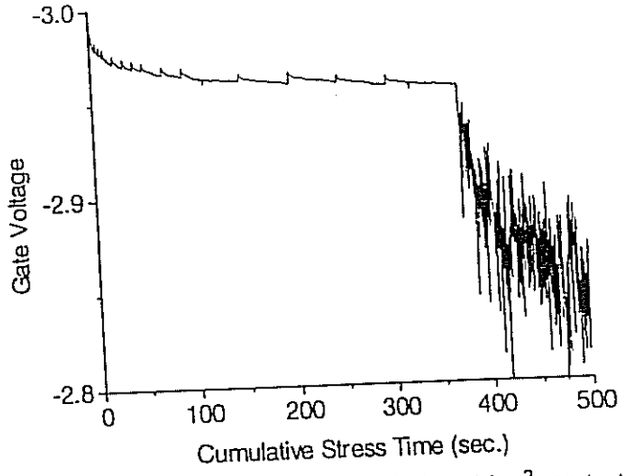

Fig. 3-42 Gate voltage measured during sequential –10 mA/cm^2 constant-current stresses of a 2.0-nm gate oxide. Soft breakdown occurs after about 360 seconds. Spikes are visible in the pre-breakdown trace because the stress was removed periodically.[90]

lengths are scaled below 0.25-μm. However, while the effect is widely discussed as being a problem, its physical basis on a fundamental level has not been carefully addressed in most texts. Thus, it is useful to explain the origin of this effect before describing its ramifications on the characteristics of deep-submicron MOSFETs.

While the polysilicon-depletion-effect has recently become a topic of attention in the technical literature, in truth it is an effect that *always* exists when a MOS-C (or MOSFET) is biased into depletion or inversion. Although the polysilicon-depletion-effect always occurs under such biasing conditions, it can be ignored by treating the polysilicon gate layer as being an ideal, equipotential region. In fact, this idealized perspective is *merely an approximation, that can only be invoked as long as doing so has insignificant effects on the device behavior.* Once the effect becomes significant, this idealization must be removed. (Note that most book authors are probably oblivious to the fact that this approximation is even invoked, as the correct energy-band diagram is only mentioned in one or two of the dozens of device physics texts that purport to treat MOSFETs.)

Let us consider an MOS-C which has an n^+-doped polysilicon gate and a p-doped silicon substrate. When a positive bias is applied between the gate and substrate, this causes depletion or depletion/inversion in the silicon substrate (depending on the bias voltage), and the charges induced in the silicon substrate by this biasing are *negative*. In any case, an equivalent *positive* charge must exist

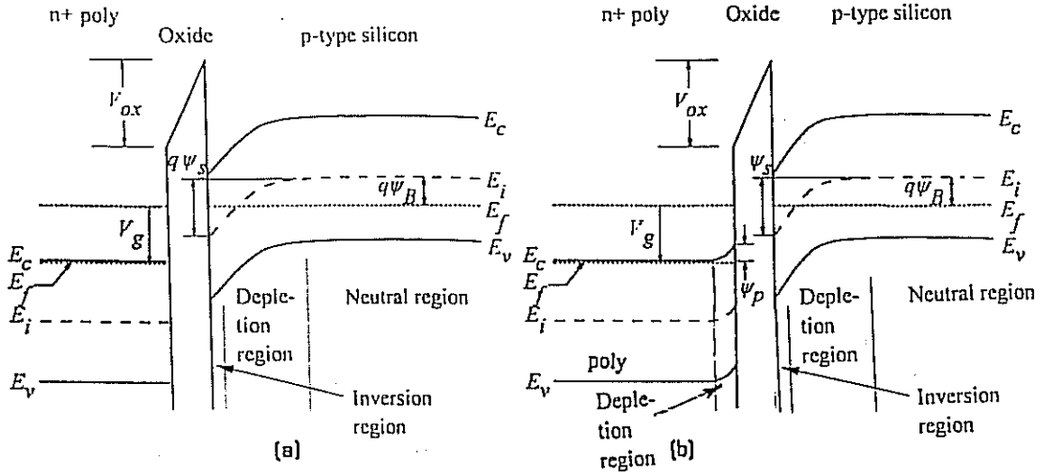

Fig. 3-43 a) Energy-band diagram of an NMOS-C biased to inversion which ignores the depletion-region formed in the polysilicon gate; b) Energy-band diagram of an NMOS-C biased to inversion showing the polysilicon depletion-region that actually always forms.

on the gate (polysilicon electrode).. Donor ions in the n^+-doped poly constitute this positive charge (i.e., some of the mobile, negatively charged electrons leave the poly gate when a positive bias is applied to it). In other words, the biasing causes a *depletion region* of some finite thickness to be formed in the poly gate (at the poly/oxide interface). This depletion region in the poly is the basis of the *polysilicon-depletion-effect (PDE)*. While most energy-band diagrams depict the situation of a MOS-C under depletion (or inversion) as in Fig. 3-43a, this is not strictly correct. That is, this diagram ignores the polysilicon depletion-region. The more correct energy-band-diagram is shown in Fig. 3-43b. Here, the presence of the poly depletion-region is indicated by band-bending (which implies the presence of an electric field) in the poly gate near the oxide.

The existence of an electric field means that a voltage-drop exists across the polysilicon depletion-region (which represents some fraction of the applied gate voltage). This means that the remaining voltage (i.e., the sum of the voltage across the gate oxide and the silicon substrate) is smaller than if no voltage was dropped across the poly depletion-region. The reduced remaining voltage will induce a smaller inversion-layer charge-density than if there was no voltage dropped across the polysilicon depletion-layer. (The effect is the same as if the gate oxide thickness was increased; i.e., PDE increases the "equivalent oxide thickness" $t_{ox_{eq}}$ – see Chap. 4, Sect 4.1). The decreased inversion-layer charge-density results in a degradation of the drain-current I_D of the MOSFET.

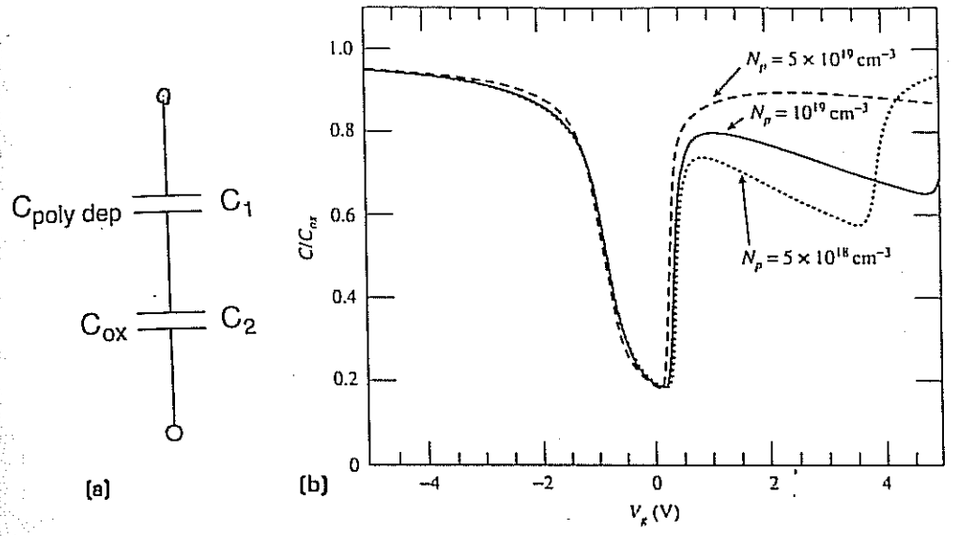

Fig. 3-44 a) Equivalent-circuit diagram of the polysilicon depletion-capacitance in series with the gate-oxide-capacitance. b) Low-frequency C-V curves of *p*-type MOS capacitor with an n^+-polysilicon gate doped at several different concentrations.[96]

The polysilicon-depletion-effect (PDE) becomes more significant as device scaling continues because the oxide capacitance increases as the oxide is made thinner, but the capacitance due to the polysilicon-depletion-layer remains constant (assuming there is no increase in the poly doping). At some point (i.e., at about when the gate-oxide thickness is scaled below 10.0-nm and the gate voltages are 2.5-V [or lower]), the depletion effect starts to adversely impact the device performance of MOSFETs made with polysilicon gates to a significant degree. We will now explain quantitatively why this is so.

The depletion-capacitance due to the PD-region is in series with the gate-oxide-capacitance (and the silicon-substrate depletion-region-capacitance, as well). An equivalent circuit diagram of the poly-depletion-capacitance in series with the gate-oxide-capacitance is shown in Fig. 3-44a.* The total capacitance of two capacitors in series is given by:

$$1/C_{tot} = (1/C_1) + (1/C_2) \qquad (3.11)$$

Thus, the smaller capacitor largely determines the total capacitance value. For the

* Note that although no capacitor electrodes physically exist at the poly/SiO_2 interface, such "virtual" electrodes are shown in Fig. 3-44a. This is nevertheless a valid equivalent-circuit representation because the two "virtual" electrodes each have the same amount of charge (although of opposite polarity). Thus, the effect of one cancels that of the other.

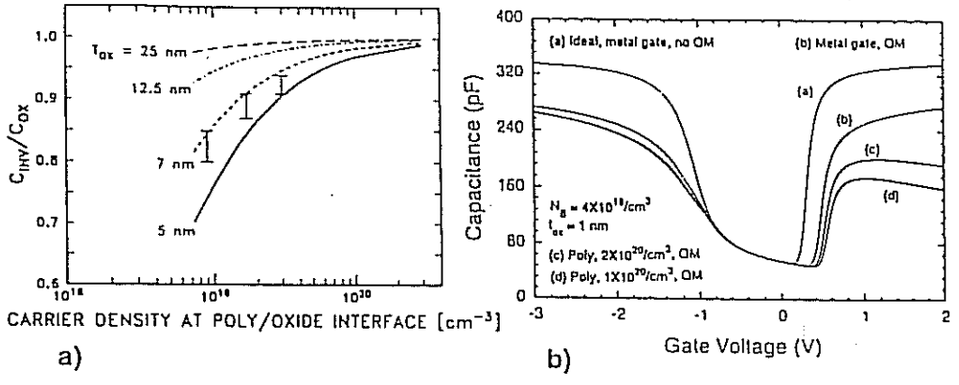

Fig. 3-45 a) Theoretical plot of C_{inv}/C_{ox} as a function of carrier concentration at the polysilicon/oxide interface for different gate oxide thickness.[101] b) Theoretical C-V curves for a thin gate oxide (1-nm-thick) including quantum mechanical effects and polysilicon depletion.[96]

case of the polysilicon-depletion-capacitance and the gate-oxide-capacitance, the gate-oxide-capacitance is much smaller than the polysilicon-depletion-capacitance until the oxide thickness gets below about 10.0-nm. For such thin oxides (assuming the poly layers doping remains constant), the oxide-capacitance starts to get large enough that the poly-depletion-capacitance can begin impacting the total capacitance. This is also shown in Fig. 3-44b in which the low-frequency C-V curves are shown for an MOS capacitor (with a 7-nm gate oxide, a p-substrate, and an n^+-poly gate with several doping concentrations).[86] It can be seen that the total capacitance at inversion (C_{inv}) does not return to the full oxide capacitance (i.e., the total capacitance in inversion is reduced by the PDE). The degree of reduction in C_{inv} depends on the poly doping concentration (Fig. 3-45a). This is because the higher the doping concentration, the narrower is the PD region, and thus the smaller is effect of PD. For the 7-nm gate-oxide thickness used in Fig. 3-44b, the poly-depletion effect becomes negligible only when the poly doping reaches about $1x10^{20}/cm^3$ (C_{inv} ~.98C_{ox}, see Fig. 3.45a).

As the oxide thickness gets smaller, the problem gets worse.[95] Figure 3-45b depicts computer simulated C-V curves for an MOS-C with a 1-nm-thick oxide, and an n^+-poly on a p-type substrate having a doping density of $4x10^{18}/cm^3$ (to allow for a V_T of several tenths of a volt). Curve a) is the C-V curve for an ideal (infinite conductivity) gate material, ignoring quantum effects. Curve b) is for a metal gate but including quantum effects. Curves c) and d) are for n^+-poly gates

doped to $2x10^{20}/cm^3$ and $1x10^{20}/cm^3$, respectively, and including quantum effects. Curves c) and d) illustrate the effects of finite polysilicon doping.[96]

The above discussions indicate that if the polysilicon doping density could be increased without limit, the polysilicon depletion effect could be minimized. However, it appears that it will be very difficult to get electrically active doping densities above $10^{20}/cm^3$ for n-type polysilicon and above the mid- to upper-$10^{19}/cm^3$ for p-type polysilicon. This will result in reductions in drive-current capability that might not be acceptable.

To summarize, the impact of polysilicon-depletion on the minimum-oxide-thickness is the following: As the oxide is made thinner, the capacitance of the MOS gate structure is more impacted by the polysilicon-depletion-layer capacitance (assuming the doping in the poly remains constant). The effect is to reduce the total MOS-gate-capacitance to a lower value than if there was no poly depletion-layer-capacitance. For the same gate voltage, the potential at the Si surface is thus reduced, which in turn decreases the inversion-layer charge-density. Thus, the drain current actually decreases as the oxide is made thinner, causing the speed of the circuits made with such MOSFETs to decrease. Figure 3-46 shows how the propagation-delay in a ring oscillator increases with decreasing oxide thickness due to this effect.[84] This implies that it may not be appropriate to use the thinnest oxide thickness that is allowed by the 5-MV/cm breakdown limit.

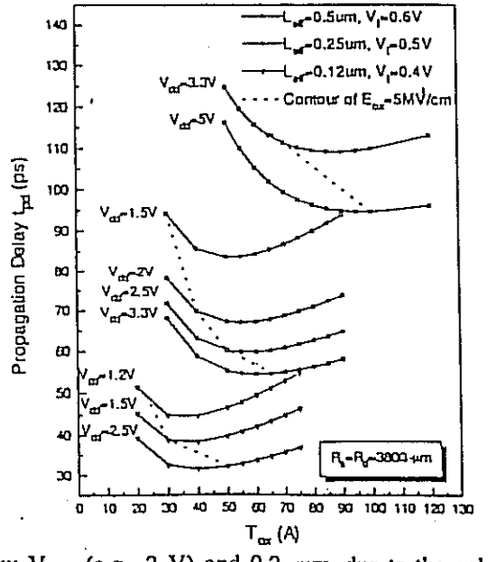

Fig. 3-46 At low V_{DD} (e.g., 2 V) and 0.2- μm, due to the polysilicon-depletion-effect, optimum speed may require a thicker t_{ox} than that allowed by the 5-MV/cm limit.[84]

3.8.5 Impact of Process Induced Damage

In 1996, Hu suggested that process -induced damage would have less impact on gate-oxide-reliability as oxides less than 10-nm in thickness were used.[78] That is, oxides that were 6.4-nm-thick showed an order of magnitude smaller interface-trap-density than did oxides that were 11.6-nm-thick (Fig. 3-47). A model that sought to explain this effect treats plasma charging-current as a constant-current-source, rather than a fixed voltage-source. Thus, for the same plasma process conditions, the same amount of current will pass through oxides as they are scaled below 10-nm, regardless of their thickness. In addition, the thinner oxide (e.g., 6.4-nm) is more tolerant of the same current stress than the thicker oxide (11.6-nm-thick). This data led Hu to conclude that plasma-process damage would not become a more serious issue with oxide scaling below 5-nm.

However, more recent information was not as optimistic. In 2000, at the 5th International Symposium on Plasma Process-Induced Damage, P.W. Mason of Lucent Technologies reported that plasma damage did, in fact, increase as the gate oxide thickness was decreased from 11.5-nm to 5.5-nm. The main damage source was the rf-backsputter via-clean process used to prepare vias for metal deposition. The plasma damage was found to correlate with long-term oxide reliability degradation, although not with short-term reliability problems. Another report from K.P. Cheung of Lucent at the same conference indicated that currently available measurement techniques cannot determine whether or not a low-level of plasma-induced damage is occurring.

The fact that plasma-induced damage in gate oxides may still remain a relia-

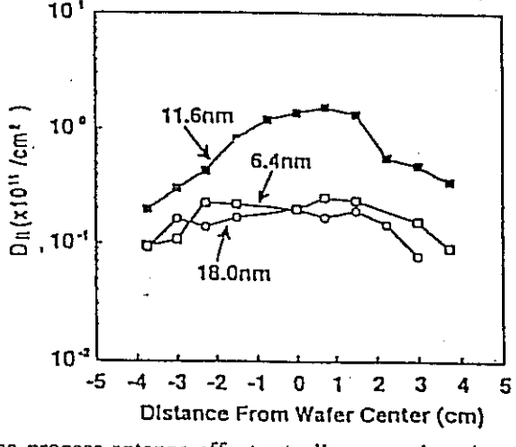

Fig. 3-47 Plasma-process-antenna-effect actually causes less damage to 6.4-nm oxides than to 11.6-nm oxides.[11]

bility problem, has led work to continue on ways to limit such damage. In July 2000 it was reported that using 3.0-nm-thick oxides nitrided in N_2O produced CMOS transistors which showed significantly improved resistance to charging damage, especially for PMOS devices.[85]

3.9 MEASURING ULTRA-THIN GATE OXIDES

Ultra-thin oxide films can be measured in a number of ways, as described in detail in Ref. 98. Two of the more common methods are *ellipsometry* and *high-resolution transmission electron microscopy* (HRTEM). Ellipsometry is a powerful and accurate optical method. The technique uses plane-polarized monochromatic light to illuminate the oxidized silicon surface at an angle. (The name *ellipsometer* comes from the use of elliptically polarized light to measure film properties.) An elliptically polarized laser beam is directed at the Si surface at an angle, and is then reflected from both the silicon substrate and oxide surface (Fig. 3-48). The technique makes use of the change of state of polarization when light is reflected from a surface. If the index of refraction (n_i) and the extinction coefficient (κ) of the substrate are known, and if the film is non-absorptive at the wavelength being used (so it's $\kappa = 0$), the state of polarization of the reflected beam then depends only on the thickness of the transparent film (and n_i).

The reflected beam passes through an analyzer drum and onto a detector. The analyzer drum is rotated to produce a minimum value in the light intensity reaching the photodetector. By reading the polarizer and analyzer settings, the

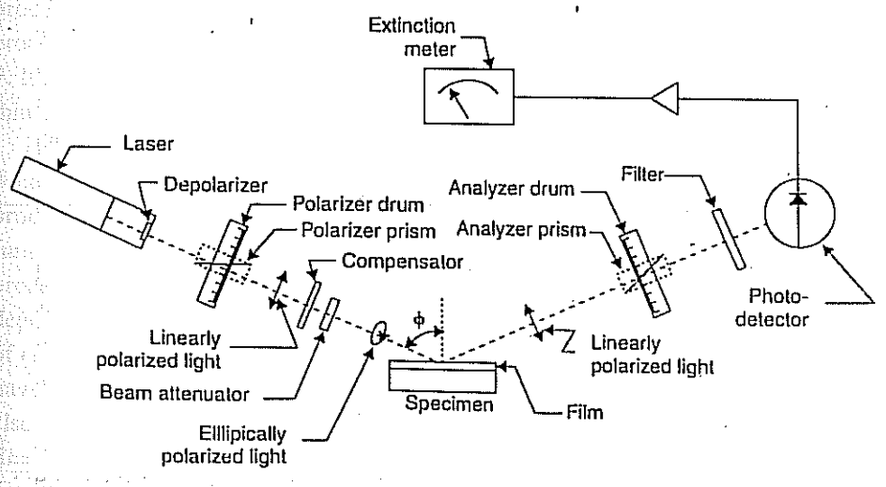

Fig. 3-48 Schematic of an ellipsometer.

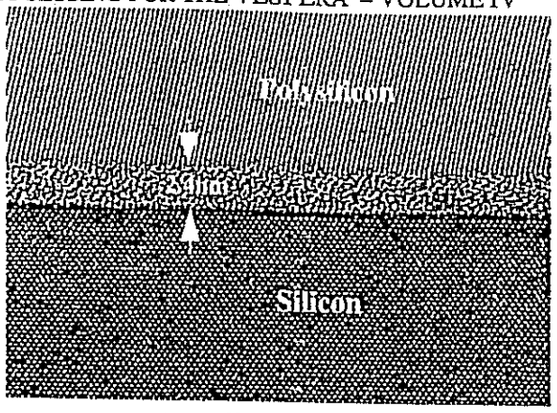

Fig. 3-49 HRTEM of 2.4 nm-thick oxide.[90] (© 1997 IEEE)

film thickness and index of refraction can be determined. The direct outputs of an ellipsometric measurement are the angles ϕ and ψ, which are related to the refractive index of the silicon substrate, the refractive index of the silicon oxide, and the thickness of the film.

If a single reflectance angle is used, multiple film thicknesses can produce a null, and the user must thus start with some idea of the expected film thickness to select the correct value. However, if multiple angles are used, an unambiguous single thickness value can be obtained. In any case, ellipsometry can be used to measure oxide thicknesses down the very thin regime (< 50-Å). However, inaccuracies in this regime may be present. Therefore, more complicated methods, such as multiple-wavelength ellipsometry or multiple-angle ellipsometry may be used to improve the accuracy of these measurements.[97] One recently described technique for measuring ultra-thin oxides involves a modification to conventional ellipsometry. In this method, an oxide *under* a polysilicon gate film is measured with a multiple-wavelength, multiple-angle, laser ellipsometer (the Rudolph Technologies S200-*ultra*® model). This allows the thickness of oxides fabricated *in-situ* in the single-wafer cluster tools described in Sect. 3.8.6, to be determined.[99]

Thin-oxide film-thickness can also be measured from micrographs obtained by HRTEM (Fig. 3-49). Preparing a sample for HRTEM requires the tedious tasks of slicing the wafer, mounting the slice in epoxy, and mechanically polishing and ion milling it to the proper thickness. However, unlike other techniques, no physical values such as index-of-refraction or dielectric constant are needed.

As gate-oxides scale to the ultra-thin regime, issues associated with measuring these films becomes increasingly critical. A transition region is thought to exist between the crystalline-silicon and bulk-silicon-dioxide that extends from

several angstroms up to about 10-Å from the silicon. For ultra-thin oxide films this could constitute 20-50% of the film. The structural differences in the transition region can change the properties of the film, including its refractive index, static dielectric constant, band gap, and effective mass. These property changes can effect many measurement techniques used to determine the oxide thickness. Thus, it is necessary to examine the extent to which this interlayer must be included in optical models to obtain the required measurement precision for ultra-thin oxide films. Surface roughness will also play a role since many of the measurements assume an atomically smooth transition between silicon ad silicon dioxide. Leakage and tunnel currents through the oxide will also affect many metrology aspects, including capacitance and conductance measurements.

The presence of contamination on the surface of the dielectrics to be measured is another complicating factor. Such airborne contaminants can form an ad-layer on the surface (see Sect. 3.10.1). Under these circumstances, the ellipsometric response will include the effect of the ad-layer. To eliminate these effects, it is necessary to develop a reproducible and easily implemented cleaning procedure. The analysis of ultra-thin stacked dielectrics consisting of individual layers of oxide and nitride may also be made more complex by parameter correlation effects.

3.10 MANUFACTURING ULTRA-THIN GATE OXIDES IN EARLY-2001

3.10.1 Process Control Issues of Growing Ultra-Thin Gate Oxides

In the 2001 timeframe, gate oxides were still largely being grown with batch processes that employ vertical furnaces. Details of such batch oxide-growth processes (and the tools used to carry them out) are found in Vol.1, 2nd Ed., Chap. 7. It has been demonstrated that these types of processes can grow oxide films as thin as 2.5-nm, and with a uniformity of ±0.5-Å across a 200-mm wafer.[100]

However, when oxide thicknesses have to be made even thinner (≤ 2.0-nm), it becomes very difficult to maintain this kind of control using such conventional processing methods (i.e., in which the gate oxide is grown in an oxidation furnace and the polysilicon gate electrode is subsequently deposited in a separate processing tool). In the time between gate-oxide-formation and polysilicon-deposition, the gate oxide is exposed to the fab environment, allowing adsorption of *molecular airborne contamination* (MACs – see also Chap. 6).[102] Such MACs have been shown to have a negative effect on gate-oxide integrity, and they are also known to cause the measured thickness of the gate oxide to increase over time. In traditional batch processing, wafers may be stored for hours (or even days) after the gate oxide is grown, before the poly gate electrode is deposited.

Newer production methods thus form both the gate oxide and the polysilicon in the single vacuum environment of a cluster tool (with separate process chambers linked by an isolated vacuum robot that transfers wafers between chambers). This not only speeds the formation of the gate stack but also does not MACs the opportunity to become incorporated MACs between the two layers.[99]

3.10.2 Extending the Use of Ultra-Thin Silicon Dioxide Layers as Gate Dielectrics: Stacked Oxide Films Which Incorporate Nitrogen

As argued earlier, growing gate dielectrics with a thickness of < 2.0-nm in a manufacturing environment may not be possible, due to several limits: reliability; tunneling leakage-current; and boron penetration. However, work has continued to try and extend the use of oxides into this thickness regime. The main thrust in the 2001 timeframe for doing this involves the formation a dielectric *stack*, still mainly based on SiO_2. However, nitrogen is incorporated into the oxide such that it is present in several places: 1) in a very-thin layer near the Si/SiO_2 interface at a concentration of < 1atom%; and 2) in another very thin layer at the poly/SiO_2 interface, at a percentage of greater than 5 atom% (or, more commonly, in the form of a very thin SiON layer on top of the SiO_2 layer). To make such a gate-dielectric stack-structure effective, it is critical to be able to control the amount, position, and concentration-profile of the nitrogen throughout the stack.

Excess nitrogen in the bulk of the film may harm the oxide properties, and thus it is desirable that little (or none) exists in the bulk of the oxide. However, nitrogen at a concentration of < 1% close to the Si/SiO_2 interface has been found to be useful in strengthening the bottom Si/SiO_2 interface against hot-carrier degradation and interface-state generation, without degrading the electrical properties of the stack. On the other hand, a nitrogen concentration of at least 5 atom% is needed at the top oxide/poly interface to prevent boron penetration into the oxide bulk during subsequent high-temperature processing steps. Such boron penetration into the oxide (i.e., from the p^+-doped poly of deep-submicron PMOSFETs) has two detrimental effects: 1) If not prevented from doing so, the boron can enter the Si substrate, causing unacceptable shifts in the V_T of the PMOSFETs; and 2) If the boron does enter the oxide, but is prevented from entering the silicon substrate, it will nevertheless pile up in the oxide. Such boron-pileup in the oxide has been shown to degrade the oxide reliability, likely by causing defects that weaken the oxide structure.

Thermal oxidation with O_2, N_2O, and NO (as described in Sects. 3.7.1 and 3.7.2) can tailor the nitrogen profile such that reasonable control of the nitrogen density and profile at the Si/SiO_2 interface and within the bulk oxide can be

maintained. However, it may not be possible to form a layer with enough nitrogen at the oxide/poly interface to prevent B penetration (i.e., > 5 atom% is needed) using thermal techniques in oxide films ≤ 2.0-nm thick. Instead, a low-temperature plasma-technique (e.g., using a remote nitrogen plasma) can be used to form a very-thin nitride film on the top oxide surface (actually a heavily-nitrided oxide-film is formed with this method).[103] Such thin SiON-layers may be enough not only to reduce or eliminate boron penetration, but also to increase the stacks permittivity. The result is that by adding an oxynitride film to the stack may cause an increase in its effective thickness of approximately 30%. This, in theory, may also make it possible to reduce the gate leakage-current. The nitride film formed by such a remote plasma nitridation process has also recently been shown to improve the reliability of the dielectric stack.[104] These improvements reportedly arise from reduced leakage and higher voltage acceleration factor.

3.10.3 Use of Ultra-Thin Oxide Layers Under High-*k* Gate-Dielectric Films

Even when it comes time to use high-*k* dielectrics in MOSFETs, an ultra-thin silicon oxide layer may still need to be formed on the silicon surface under the high-*k* dielectric film (to produce a reliable interface between the high-*k* material and silicon substrate). Such an SiO_2 layer would preserve the interface state characteristics and channel mobility. However, since the interfacial layer contributes to the equivalent oxide thickness, the oxide thickness will still have to be minimized the without compromising reliable interface properties. Thus, it seems that processing ultra-thin SiO_2 films may remain a part of silicon IC fabrication for yet a long time!

REFERENCES

1. S. Thompson *et al.*, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal*, 3rd Quarter 1998.
2. D. Wolters and J.F. Verwey, "Breakdown and Wearout Phenomena in SiO_2 Films," Chap. 6, in *Instabilities in Silicon Devices - Vol. 1*, Ch. 1 Eds. G. Barbottin and A. Vapaille, (Elsevier) North-Holland, Amsterdam, 1986.
3. A.G. Revesz, "The Defect Structure of Grown Silicon Dioxide Films," *IEEE Trans. Electron Dev.*, **ED-12**, p. 97 (1965).
4. S. Rigo, "Silica Films on Silicon," in *Instabilities in Silicon Devices - Vol. 1*, Ch. 1 Eds. G. Barbottin and A. Vapaille, (Elsevier) North-Holland, Amsterdam, 1986.
5. I.E. Tamm, *Phys. Z. Sowjetunion* **1**, p. 733 (1932).
6. P. Roblin, *et al.*, "Simulation of Hot-Electron Trapping and Aging in *n*MOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2229.
7. B.E. Deal, *J. Electrochemical Society.*, **121:** 198C (June 1974).