Ultrathin (<4 nm) SiO$_2$ and Si–O–N gate dielectric layers for silicon microelectronics: Understanding the processing, structure, and physical and electrical limits

M. L. Green$^{a}$
Agere Systems, Murray Hill, New Jersey 07974

E. P. Gusev
IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

R. Degraeve
IMEC, Leuven 3001, Belgium

E. L. Garfunkel
Rutgers University, Piscataway, New Jersey 08854

(Received 1 March 2001; accepted for publication 21 May 2001)

The outstanding properties of SiO$_2$, which include high resistivity, excellent dielectric strength, a large band gap, a high melting point, and a native, low defect density interface with Si, are in large part responsible for enabling the microelectronics revolution. The Si/SiO$_2$ interface, which forms the heart of the modern metal–oxide–semiconductor field effect transistor, the building block of the integrated circuit, is arguably the world's most economically and technologically important materials interface. This article summarizes recent progress and current scientific understanding of ultrathin (<4 nm) SiO$_2$ and Si–O–N (silicon oxynitride) gate dielectrics on Si based devices. We will emphasize an understanding of the limits of these gate dielectrics, i.e., how their continuously shrinking thickness, dictated by integrated circuit device scaling, results in physical and electrical property changes that impose limits on their usefulness. We observe, in conclusion, that although Si microelectronic devices will be manufactured with SiO$_2$ and Si–O–N for the foreseeable future, continued scaling of integrated circuit devices, essentially the continued adherence to Moore’s law, will necessitate the introduction of an alternate gate dielectric once the SiO$_2$ gate dielectric thickness approaches ~1.2 nm. It is hoped that this article will prove useful to members of the silicon microelectronics community, newcomers to the gate dielectrics field, practitioners in allied fields, and graduate students. Parts of this article have been adapted from earlier articles by the authors [L. Feldman, E. P. Gusev, and E. Garfunkel, in Fundamental Aspects of Ultrathin Dielectrics on Si-based Devices, edited by E. Garfunkel, E. P. Gusev, and A. Y. Vul’ (Kluwer, Dordrecht, 1998), p. 1 [Ref. 1]; E. P. Gusev, H. C. Lu, E. Garfunkel, T. Gustafsson, and M. Green, IBM J. Res. Dev. 43, 265 (1999) [Ref. 2]; R. Degraeve, B. Kaczer, and G. Groeseneken, Microelectron. Reliab. 39, 1445 (1999) [Ref. 3]. © 2001 American Institute of Physics. [DOI: 10.1063/1.1385803]

TABLE OF CONTENTS

I. INTRODUCTION AND OVERVIEW ..................... 2058
   A. SiO$_2$ enabled the microelectronics revolution... 2058
   B. Fundamental limits of SiO$_2$: Is the end in sight? 2060

II. PHYSICAL CHARACTERIZATION METHODS
    FOR ULTRATHIN DIELECTRICS ..................... 2062
    A. Optical methods .......................... 2062
    B. X-ray based methods ...................... 2065
    C. Ion beam analysis ....................... 2067
    D. Electron microscopy ..................... 2072
    E. Scanning probe microscopy ................. 2073

III. ELECTRICAL CHARACTERIZATION
    METHODS AND PROPERTIES ...................... 2074
    A. Characterization methods ................... 2075
       1. C–V measurements ........................ 2075
       2. Gate tunnel current .................... 2076
       3. Charge pumping ........................ 2077
    B. Oxide degradation during electrical stress ... 2078
       1. Interface trap creation ................ 2078
       2. Oxide charge trapping .................. 2078
       3. Hole fluence .......................... 2078
       4. Neutral electron trap generation ......... 2078
       5. Stress-induced leakage current ......... 2080
       6. Discussion of trap generation mechanisms.. 2081

$^a$Electronic mail: mlg@agere.com

© 2001 American Institute of Physics
C. Oxide breakdown. ....................................... 2082
1. Breakdown modeling. ............................... 2082
2. Soft breakdown. ..................................... 2083
3. Breakdown acceleration models. ................. 2084
4. Temperature dependence of breakdown. ....... 2085
5. Oxide reliability predictions. ...................... 2086

IV. FABRICATION TECHNIQUES FOR ULTRATHIN SILICON OXIDE AND OXYNITRIDES. .................. 2086
A. Surface preparation. ............................... 2086
B. Fabrication of ultrathin oxide and oxynitrides. 2086
1. Thermal oxidation and oxynitridation .............. 2086
2. Chemical deposition. ............................... 2088
3. Physical deposition. .................................. 2090
C. Postoxidation processing and annealing. ......... 2090
D. Hydrogen/deuterium processing .................... 2091

V. THE Si/SiO₂ SYSTEM ..................................... 2093
A. The initial stages of oxygen interaction with silicon surfaces. ................................. 2093
1. The passive oxidation regime. ..................... 2094
2. The active oxidation regime. ....................... 2098
3. The transition regime. .............................. 2099
B. Growth mechanisms of ultrathin oxides, beyond the Deal–Grove model. ....................... 2100

VI. THE Si/SiO₂ N SYSTEM ................................ 2104
A. Oxynitride properties. .............................. 2104
1. Thermodynamics of the Si–O–N system. ........ 2104
2. Physical properties. ............................... 2105
3. Diffusion barrier properties of oxynitride layers. .............................................. 2105
B. Oxynitridation processes. ......................... 2106
1. NO processing. ..................................... 2106
2. N₂O processing. .................................... 2107
   a. Gas phase N₂O decomposition at high temperatures. .................................. 2107
   b. N incorporation and removal during N₂O oxynitridation. .......................... 2108
3. Nitridation in NH₃. ................................. 2109
4. Nitridation in N₂. ................................. 2109

VII. THE POST-SiO₂ ERA: ALTERNATE GATE DIELECTRICS ............................................ 2109
A. Si₃N₄ ............................................... 2110
B. Alternate (higher) dielectric constant materials. 2110

---

I. INTRODUCTION AND OVERVIEW

A. SiO₂ enabled the microelectronics revolution

Nature has endowed the silicon microelectronics industry with a wonderful material, SiO₂, as is shown in Table I. SiO₂ is native to Si, and with it forms a low defect density interface. It also has high resistivity, excellent dielectric strength, a large band gap, and a high melting point. These properties of SiO₂ are in large part responsible for enabling the microelectronics revolution. Indeed, other semiconductors such as Ge or GaAs were not selected as the semiconducting material of choice, mainly due to their lack of a stable native oxide and a low defect density interface. The metal–oxide–semiconductor field effect transistor (MOSFET), Fig. 1, is the building block of the integrated circuit. The Si/SiO₂ interface, which forms the heart of the MOSFET gate structure, is arguably the worlds most economically and technologically important materials interface. The ease of fabrication of SiO₂ gate dielectrics and the well passivated Si/SiO₂ interface that results have made this possible. SiO₂ has been and continues to be the gate dielectric par excellence for the MOSFET. Figure 2 is a transmission electron photomicrograph of an actual submicron MOSFET, showing the SiO₂ gate dielectric as well as the Si/SiO₂ interface.

In spite of its many attributes, however, SiO₂ suffers from a relatively low dielectric constant (κ=dielectric constant, or permittivity, relative to air=3.9). Since high gate dielectric capacitance is necessary to produce the required drive currents for submicron devices,¹⁶ and further since capacitance is inversely proportional to gate dielectric thickness, the SiO₂ layers have of necessity been scaled to ever thinner dimensions, as is shown in Fig. 3. This gives rise to a number of problems, including impurity penetration through the SiO₂, enhanced scattering of carriers in the

---

**TABLE I.** Selected properties of SiO₂ gate dielectric layers.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native to silicon (SiO₂ is the only stable oxide phase on Si)</td>
<td></td>
</tr>
<tr>
<td>Low interfacial (Si/SiO₂) defect density (∼10¹⁰ eV⁻¹ cm⁻², after H₂ passivation)</td>
<td></td>
</tr>
<tr>
<td>Melting point</td>
<td>1713 °C</td>
</tr>
<tr>
<td>Energy gap</td>
<td>9 eV</td>
</tr>
<tr>
<td>Resistivity</td>
<td>~10¹⁵ Ω cm</td>
</tr>
<tr>
<td>Dielectric strength</td>
<td>~1 x 10⁷ V/cm</td>
</tr>
<tr>
<td>Dielectric constant</td>
<td>3.9</td>
</tr>
</tbody>
</table>

---

**FIG. 1.** Schematic illustration of a submicron (channel length) CMOSFET (complementary metal–oxide–semiconductor field effect transistor) (courtesy of C. P. Chang, Agere Systems).
known as “high cost per bit. SiO2 gate dielectrics have decreased in thickness rent improvements in circuit speed, memory capacity, and decreasing minimum feature size, accompanied by concur-
describes the exponential growth of chip complexity due to scientific understanding of ultrathin dimensions of MOSFETs and other devices have been scaled through shrinkage of the circuit elements, the di-
resulting in physical and electrical property changes that impose limits on their usefulness. We will also discuss the near-future need for alternate gate dielectric materials such as Si3N4 and other metal oxides and nitrides, known as “high κ” materials.

In a continuous drive to increase integrated circuit performance through shrinkage of the circuit elements, the dimensions of MOSFETs and other devices have been scaled since the advent of integrated circuits about 40 years ago, according to a trend known as Moore’s law.5–8 Moore’s law describes the exponential growth of chip complexity due to decreasing minimum feature size, accompanied by concurrent improvements in circuit speed, memory capacity, and cost per bit. SiO2 gate dielectrics have decreased in thickness from hundreds of nanometers (nm) 40 years ago to less than 2 nm today, to maintain the high drive current and gate capacitance required of scaled MOSFETs. Further, as can be seen in Fig. 3, SiO2 thickness continues to shrink. Many ultrasmall transistors have been reported, with SiO2 layers as thin as 0.8 nm.9–15 The International Technology Roadmap for Semiconductors,16 excerpted in Table II, shows that SiO2 gate dielectrics of 1 nm or less will be required within 10 years. It will become obvious while reading this article that SiO2 layers thinner than ~1.2 nm may not have the insulating properties required of a gate dielectric. Therefore alternate gate dielectric materials, having “equivalent oxide thickness” less than 1.2 nm (for example), may be used. Equivalent oxide thickness, $t_{\text{ox(eq)}}$, is the thickness of the SiO2 layer ($\kappa_\text{eq}=3.9$) having the same capacitance as a given thickness of an alternate dielectric layer, $t_\text{dil (} \kappa_\text{dil} \text{)}$. Equivalent oxide thickness is given by the relationship: $t_{\text{ox(eq)}} = t_\text{dil}(3.9/\kappa_\text{dil})$. At these thicknesses, the Si/SiO2 interface becomes a more critical, as well as limiting, part of the gate dielectric. A 1 nm SiO2 layer is mostly interface, with little if any bulk character. It contains about five layers of Si atoms, at least two of which reside at the interface.17 Given its prominence, much of this article will focus on the physical and electrical properties of the Si/SiO2 interface.

Due to its commercial relevance, the Si/SiO2 system has received an enormous amount of scientific attention. It is daunting to count the number of scientific papers: using an INSPEC Database we found 36708 references (since 1969) devoted to this system. [We intersected the set “silicon” with the set “(SiO or SiO2 or SiOx)” and then subtracted the set “quartz.”] Only 2% of these references are cited in this article. Several excellent books18–23 and review papers on di-electric selection,24 atomic scale interactions between Si and O,25–27 oxidation of Si,28–30 SiO2 structure,31 interface structure and defects,32–34 reliability,35 metrology,36 Si–O–N2,37,38 and general growth, structure and properties39,40 have been published. However, some basic scientific issues at the forefront of the field remain unresolved. Among these issues are an understanding of the exact diffusion mechanisms and incorporation reactions of oxidizing and nitriding species in SiO2, an atomistic understanding of the initial stages of oxidation, the role of postoxidation processing on physical and electrical properties, the bonding structure at and near the Si/SiO2 interface, the relationship between local bonding/chemistry and electrical defects, and the failure mechanisms in ultrathin dielectrics. All of these topics will be addressed in this article.

It is amusing and instructive to learn that not only is SiO2 enabling to microelectronics, but also to some forms of life itself. Many forms of animal and plant life have cell membranes and exoskeletons composed of pure, crystalline (opaline) SiO2.41 This should not be too surprising, since Si and O are the two most abundant elements in the earth’s crust. In particular, one celled animals called diatoms fashion their cell membrane via self-assembly (molecule by molecule), from SiO2 dissolved in H2O, at ambient temperature.42 They exist in thousands of symmetric morphologies. Figure 4 is a scanning electron microscopy image of the SiO2 exoskeleton of a diatom, on which one can ob-

FIG. 2. Cross-section transmission electron photomicrographs of a 35 nm (channel length) transistor, and a detailed view of its 1.0 nm SiO2 gate dielectric and Si/SiO2 interface [from Timp et al. (Ref. 15)].

FIG. 3. Decrease in gate SiO2 (or equivalent oxide) thickness with device scaling (technology generation). Actual or expected year of implementation of each technology generation is indicated [adapted from ITRS (Ref. 16)].
serve SiO$_2$ features as small as 100 nm. It is humbling to discover that diatoms evolved with such “scaled dimensions” 400 million years ago, and suggests that in the future, SiO$_2$, as well as other electronic materials structures, might be self-assembled by biomimetic processes. In fact, such structures have already been achieved.

B. Fundamental limits of SiO$_2$: Is the end in sight?

The use of ultrathin SiO$_2$ gate dielectrics gives rise to a number of problems, including high gate leakage current, reduced drive current, reliability degradation, B penetration, and the need to grow ultrathin and uniform SiO$_2$ layers. We may ask when any of these effects will fundamentally limit the usefulness of SiO$_2$ as a gate dielectric.\

Due to the large band gap of SiO$_2$, $\sim$9 eV, and the low density of traps and defects in the bulk of the material, the carrier current passing through the dielectric layer is normally very low. For ultrathin films this is no longer the case. When the physical thickness between the gate electrode and doped Si substrate becomes thinner than $\sim$3 nm, direct tunneling through the dielectric barrier dominates leakage current. According to fundamental quantum-mechanical laws, the tunneling current increases exponentially with decreasing oxide thickness. Gate leakage currents measured on 35 nm transistors fabricated using advanced wafer preparation, cleaning, and oxidation procedures are shown in Fig. 5. The leakage current is seen to increase by one order of magnitude for each 0.2 nm thickness decrease.

Assuming a maximum allowable gate current density of 1 A/cm$^2$ for desktop computer applications, and $10^{-3}$ A/cm$^2$ for portable applications, minimum acceptable SiO$_2$ thicknesses (physical) would be approximately 1.3 and 1.9 nm, respectively.

Recent electron energy loss spectroscopy (EELS) experiments on ultrathin Si/SiO$_2$ interfaces, in a scanning transmission electron microscope (STEM), support these findings. Oxygen profiles across the interface were obtained and are shown in Fig. 6 for SiO$_2$ films 1.0 and 1.8 nm thick. The profiles consist of bulk-like regions and interfacial regions, and the interfacial regions are thought to be due to interfacial states. Based on this analysis, it has been calculated that the minimum oxide thickness, before leakage current becomes overwhelming, is 1.2 nm. This comes from the fact that a satisfactory SiO$_2$ tunnel barrier is formed when it is equal in thickness to six times the decay length of the interfacial states, about $6 \times 0.12$ nm $> 0.7$ nm, plus a 0.5 nm contribution from interfacial roughness.

Reduced drive (drain) current has been reported in small transistors with ultrathin gate dielectrics. Figure 7 shows that drain current increases with decreasing SiO$_2$ thickness, but then falls off; gate leakage current continuously increases, as expected, with decreasing SiO$_2$ thickness. Thus for SiO$_2$ layers thinner than about 1.3 nm there is no advan-

---

**TABLE II. MOSFET technology timetable, adapted from the International Technology Roadmap for Semiconductors (ITRS) (Ref. 16).**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM generation</td>
<td>1G</td>
<td>1G</td>
<td>4G</td>
<td>16G</td>
<td>64G</td>
<td>256G</td>
<td>1T</td>
</tr>
<tr>
<td>Minimum feature size, nm</td>
<td>180</td>
<td>160</td>
<td>130</td>
<td>100</td>
<td>70</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Equivalent oxide thickness, nm</td>
<td>1.9–2.5</td>
<td>1.5–1.9</td>
<td>1.5–1.9</td>
<td>1.0–1.5</td>
<td>0.8–1.2</td>
<td>0.6–0.8</td>
<td>0.5–0.6</td>
</tr>
</tbody>
</table>

---

**FIG. 4.** Scanning electron microscope image of a diatom exoskeleton, composed of opaline SiO$_2$ (courtesy of J. Aizenberg, Bell Laboratories/Lucent Technologies).

**FIG. 5.** Gate leakage current measured at 1.5 V as a function of oxide thickness (measured by TEM) for 35 nm NMOSFETs. Leakage current increases one order of magnitude for every 0.2 nm decrease in SiO$_2$ thickness. Horizontal lines indicate 1 A/cm$^2$ acceptable leakage current for desktop applications, and 1 mA/cm$^2$ acceptable leakage for portable applications [data of G. Timp, from Green et al. (Ref. 45)].
drive current for incurring the burden of an ever-increasing gate leakage current. This would suggest that SiO$_2$ layers thinner than 1.3 nm no longer deliver any performance advantage. The cause of the decreased drive current is not fully understood. One possibility is an additional scattering component from the upper (SiO$_2$/polycrystalline Si gate) interface. Some experimental evidence exists for this case\textsuperscript{49} but the observed effect is not enough to explain the data in Fig. 7. Another cause could be a universal mobility curve effect, i.e., lowered mobility due to enhanced scattering because of extreme carrier confinement in the inversion layer of the ultrathin oxide. The issue of long-range electrostatic interactions between charges in very heavily doped gate, source, and drain regions, and electrons traveling in the channel, was recently theoretically studied\textsuperscript{50} using both semiclassical two-dimensional self-consistent Monte Carlo–Poisson simulations and a quantum mechanical model based on electron scattering from gate–oxide interface phonons. It was shown that excitation and absorption of plasma modes in the gate region may result in a net momentum loss of carriers in the channel, thus decreasing their velocity, and leading to reduced drain current.

Reliability (lifetime to breakdown) of ultrathin SiO$_2$ is a major concern for oxide scaling into the sub-2 nm range and currently a contentious issue.\textsuperscript{35,40,51–62} Electrons traveling through the SiO$_2$ layer may create defects such as electron traps and interface states\textsuperscript{63–65} that in turn, upon accumulation to some critical density, degrade the insulating properties of the oxide. The accumulated charge the film can withstand before its breakdown ($Q_{bd}$) decreases with oxide thickness.\textsuperscript{54} Recently, it was predicted that oxide films thinner than about 2.2 nm would not have the reliability required by the industry roadmap.\textsuperscript{54} Data from another research group,\textsuperscript{62} shown in Fig. 8, indicates that acceptable reliability will be achievable for SiO$_2$ thicknesses as low as 1.4 nm. At about 1.0 nm thickness, the statistical probability of a percolation path may reduce reliability to an unacceptable level.\textsuperscript{66} One should mention that all reliability data is model dependent. Unlike directly measured parameters such as the gate leakage and drive currents, reliability studies always involve extrapolations from relatively high ($\sim 2.5–4$ V) stress voltages to real device operating voltages ($\sim 1.0–1.2$ V). The extrapolations are based on different, although most often percolation type, models with several parameters extracted from the high

![FIG. 6. Oxygen bonding profiles measured by STEM-EELS. The Si substrate is on the left and the gate polycrystalline Si is on the right. (a) 1.0 nm (ellipsometric) oxide, annealed at 1050 °C/10 s. The bulk-like O signal (y axis, arbitrary scale) yields a FWHM of 0.85 nm, whereas the total O signal yields a FWHM of 1.3 nm. The overlap of the two interfacial regions has been correlated with the observation of a very high gate leakage current, $10^3$ A/cm$^2$. (b) a thicker (1.8 nm ellipsometric) oxide, also annealed. The interfacial regions no longer overlap and the gate leakage current is $10^{-3}$ A/cm$^2$ from Muller et al. (Ref. 17).](image)

![FIG. 7. Drive current vs leakage current for two ultrasmall (gate lengths of 70 and 140 nm) NMOSFETs. In both cases it can be seen that gate current increases, as is expected, with decreasing SiO$_2$ thickness. However, drain current first increases and then falls off with decreasing oxide thickness. Decreasing the oxide thickness past the fall-off thickness confers no further advantage on the device from Timp et al. (Ref. 10).](image)

![FIG. 8. Predicted maximum allowable operating voltage as a function of oxide thickness (Ref. 16) (solid line), modeled so that requirements of power dissipation and circuit speed for successive technology generations are met. The Lucent data, at either 25 or 100 °C, estimates the safe operating voltages (for 10 year lifetime reliability) for sub-2.5 nm oxides, based on measured results and extrapolation principles (Ref. 62). The Lucent data clearly show that the 10 year reliability of sub-2.5 nm oxides meets the ITRS specifications for SiO$_2$ layers as thin as 1.4 nm.](image)
stress voltage experiments. Finally, the real impact of oxide breakdown on circuit performance, which ultimately is the critical issue, has yet to be understood.

Thus the fundamental limits imposed on SiO₂ are excessively high gate leakage current, reduced drive current, and reliability. The first two of these properties impose a limit of ~1.3 to 1.4 nm as the thinnest SiO₂ acceptable. Although potential improvements in interfacial roughness may push the leakage limit to ~1.0 nm, reliability may then be the limiting factor. Complementary metal–oxide–semiconductor (CMOS) applications of SiO₂ or Si–O–N gate dielectrics are limited neither by the ability to grow the ultrathin films in a manufacturing environment (see Sec. IV), nor to suppress the diffusion of B through them (see Sec. VI). Therefore, according to Table II, SiO₂ will have to be replaced in as little as 5 years if gate capacitance is to increase according to projected scaling. Technology generations that require little as 5 years if gate capacitance is to increase according to k

elliptically polarized reflected light
plane of incidence
sample

FIG. 9. Principles of an ellipsometric measurement.

According to Fig. 10, the measured parameters are the ellipsometric angles Ψ and Δ, defined from the ratio

\[ R_\parallel / R_\perp = \tan(\Psi) \exp(i\Delta) \]  

of the Fresnel reflection coefficients \( R_\parallel \) and \( R_\perp \) for the light polarized parallel and perpendicular to the plane of incidence. The reflection coefficients are determined by the optical properties and composition of the substrate and overlayers, their thicknesses, and morphology. The parameters Ψ and Δ can be measured either at a given wavelength of light, i.e., single wavelength ellipsometry (most often 633 nm), or as a function of photon energy, i.e., spectroscopic ellipsometry. The single wavelength configuration is often used for fast, nondestructive, on-line monitoring of film thickness, provided that the refractive index of the film is known. The spectroscopic mode allows determination of the refractive index (η) as well as the thickness. For SiO₂ films on Si, good agreement has been demonstrated, Fig. 10, between the ellipsometric oxide thickness and thickness values determined by transmission electron microscopy (TEM) and x-ray photoelectron spectroscopy (XPS), for films as thin as ~2 nm. 108 Since the very first publications on Si oxidation, ellipsometry

FIG. 10. Oxide thickness measured by x-ray photoemission spectroscopy, compared to measurements on the same films by ellipsometry, transmission electron microscopy, and capacitance–voltage techniques [from Lu et al. (Ref. 108)].

II. PHYSICAL CHARACTERIZATION METHODS FOR ULTRATHIN DIELECTRICS

In this introductory section we have systematically examined the repercussions of decreasing SiO₂ gate dielectric thickness on several physical and electrical parameters. Although the end of SiO₂ scaling may be in sight, it will be manufactured in integrated circuits for at least the next decade. This article is not intended to be an obituary for SiO₂, however, since its useful life will undoubtedly be extended by innovations in device design, e.g., dual gate dielectric thickness devices⁵⁷,⁶⁸ or vertical MOSFETs.⁶⁹

A. Optical methods

The most commonly used optical techniques are single wavelength and spectroscopic ellipsometry,⁵⁹,⁷⁰–⁸⁵ reflectance difference spectroscopy (RDS),⁸⁶–⁸⁸ second harmonic generation (SHG),⁸⁹–⁹² and Fourier transform infrared spectroscopy (FTIR).⁹²,⁹⁸–¹⁰⁷

Ellipsometry is based on the measurement and subsequent modeling of changes in the polarization state of a light beam reflected from a sample surface, as is illustrated in Fig. 9. The measured parameters are the ellipsometric angles Ψ and Δ, defined from the ratio

\[ R_\parallel / R_\perp = \tan(\Psi) \exp(i\Delta) \]  

of the Fresnel reflection coefficients \( R_\parallel \) and \( R_\perp \) for the light polarized parallel and perpendicular to the plane of incidence. The reflection coefficients are determined by the optical properties and composition of the substrate and overlayers, their thicknesses, and morphology. The parameters Ψ and Δ can be measured either at a given wavelength of light, i.e., single wavelength ellipsometry (most often 633 nm), or as a function of photon energy, i.e., spectroscopic ellipsometry. The single wavelength configuration is often used for fast, nondestructive, on-line monitoring of film thickness, provided that the refractive index of the film is known. The spectroscopic mode allows determination of the refractive index (η) as well as the thickness. For SiO₂ films on Si, good agreement has been demonstrated, Fig. 10, between the ellipsometric oxide thickness and thickness values determined by transmission electron microscopy (TEM) and x-ray photoelectron spectroscopy (XPS), for films as thin as ~2 nm. 108 Since the very first publications on Si oxidation, ellipsometry
has been the tool of choice to measure film thickness. Knowledge of the thickness is critical in modeling oxidation kinetics and determining growth mechanisms. An example of its application is shown in Fig. 11.

The ellipsometric parameters, $\Psi$ and $\Delta$, can be detected with very high accuracy, making ellipsometry one of the most sensitive thickness measurement techniques. However, the interpretation of the measurements is very model dependent, especially in the ultrathin regime where $\eta$ may become, as is the case for SiO$_2$, a function of thickness. For more complex gate dielectrics such as Si–O–N or SiO$_2$/Si$_3$N$_4$ multilayer stacks, the analysis is further complicated by the fact that $\eta$ changes with the composition of the film. For example, at a wavelength of 630 nm, $\eta$ varies linearly with the amount of N in the film, Fig. 12, from a value of 1.46 for SiO$_2$ to a value of approximately 2.0 for stoichiometric Si$_3$N$_4$.

Sensitivity to the Si/SiO$_2$ interface of ultrathin films is relatively low using conventional ellipsometric configurations. Interfacial sensitivity can be enhanced via an immersion ellipsometry technique, in which the Si/SiO$_2$ sample is placed in a liquid having an $\eta$ close to that of SiO$_2$, thereby effectively increasing the optical thickness of the oxide overlayer. Sensitivity to surfaces and interfaces may also be significantly enhanced using RDS. While in conventional ellipsometry the measurements are performed under oblique incidence, in RDS the primary photon beam is normal to the surface, as is illustrated in Fig. 13. The analysis is based on the determination of the reflectance differences of the light polarized along principle axes of the crystal surface. The anisotropic properties of surfaces and interfaces can therefore be determined. RDS has been applied to the study of Si oxidation as well as the wet cleaning of Si.

Another important method for the selective optical probing of surfaces and interfaces is SHG. In general, the polarization amplitude, $P$, of a photon beam of optical frequency $\omega$, interacting with a solid sample, can be expressed in terms of an expansion over the amplitude of the electromagnetic field ($E$), as

$$P \sim \alpha(\omega)E(\omega) + \beta(\omega, 2\omega)E^2(\omega) + \cdots. \quad (2)$$

The first term of the polynomial expansion represents the linear optical response, which is the basis of the ellipsometric and RDS techniques described above. In centrosymmetric solids such as Si, the second, nonlinear coefficient $\beta$ is non-zero only at interfaces, where crystal symmetry is broken. The electrical dipoles present at such interfaces give rise to the surface/interface selectivity of SHG. Recently, SHG generation methods have been used to investigate the structure of Si/SiO$_2$ and Si/Si$_3$N$_4$ interfaces, Fig. 14, as well as the initial stages of Si oxidation in N$_2$O and O$_2$.

FTIR is a powerful method to examine chemical (stoichiometry, bonding and impurities), and structural (stress, transitional layer) aspects of ultrathin dielectric films. FTIR, based on the absorption of light in the infrared region of the spectrum, is sensitive to
rotational, bending, and stretching vibrational modes, and therefore can probe local atomic configurations and compositions. It is one of a few techniques that can be used to study the behavior of H on Si surfaces and interfaces, as illustrated, for example, in Fig. 15. The spectrum of Si–O stretching vibrations in an amorphous SiO$_2$ oxide film is characterized by two bands, designated transverse optical (TO) and longitudinal optical (LO), polarized parallel and perpendicular to the plane of the film, respectively. This is exemplified in Fig. 16. Unfortunately, the bands may be broad, which complicates their interpretation. Detailed analysis of the relative intensities and TO/LO peak positions, based on recent theoretical and experimental findings is given in a recent review. It has also been shown that both TO and LO bands show a shift to lower wave numbers with decreasing oxide thickness, usually attributed to the presence of Si suboxides near the Si/SiO$_2$ interface. In the case of NO and N$_2$O grown Si–O–N films, in addition to the above TO and LO bands of the oxide, a peak at 960 cm$^{-1}$ has been observed, distinct from the vibrational frequency at 840 cm$^{-1}$ characteristic of Si$_3$N$_4$. This peak has been attributed to the asymmetric stretching vibration of double-bonded N atoms in the Si–N–Si structure. In Fig. 17 it is shown that IR methods have also been used to study composition and thickness of SiO$_x$ films at the interface of Ta$_2$O$_5$ and other oxide films deposited on Si, and the evolution of such layers upon thermal treatment.

B. X-ray based methods

Higher energy photon-based methods (x-ray analyses) are widely used to determine chemical and structural properties of ultrathin dielectrics and their interfaces. X-ray photoemission spectroscopy (XPS) is a common technique used to determine local chemical environment. XPS analy-
sis is based on the fact that core electron energies are characteristic of the emitting atom. For any given atom, exact core level positions, the so-called chemical shifts, change with local chemical environment. The analysis of photoemission from the Si 2p core level has been used to investigate the structure of the Si/SiO₂, Si/Si₃N₄, and Si/SiOₓNᵧ interfaces, 27,121,122,124,127–140 defect states in dielectric films, 141,142 valence band alignment, 143–145 and the initial stages of oxidation. 27,146–162 The Si 2p level for Si atoms in stoichiometric SiO₂, with all Si atoms in an Si(―O)₄ tetrahedral configuration, is shifted to ~4 eV deeper binding energy with respect to the position of the level in elemental Si. Photoemission yields of much lower intensity, at energies between the SiO₂ line and substrate Si, Fig. 18, have also been observed. These are usually interpreted in terms of suboxide (~Si₁₁, Si₁₂, and Si₁₃) states at the interface or in the film, 122,124,129,130,163–165 as can be seen in the compilation in Table III. More recently this traditional interpretation has been questioned. 133,136,139,166,167 It has been argued that second nearest neighbor effects can lead to additional shifts of Si 2p levels, making quantitative analysis of suboxide states even more complicated. In either case, current estimates of the total concentration of Si in the suboxide states are between one and two monolayers, whereas an atomically abrupt interface would result in one monolayer of suboxide. [A monolayer of a species on Si(100) represents a concentration of 6.8×10¹⁴ cm⁻², assuming no steric hindrance.] An example of a new approach to analysis of the Si/SiO₂ interface is depicted in Fig. 19. XPS is also useful in obtaining film thickness, determined by the Si(2p) mean free path and the ratio of Si 2p peaks corresponding to SiO₂ in the film and Si in the substrate. 121

![FIG. 16. Thickness dependence of the transverse optical and longitudinal optical modes of SiO₂ (on Si), as measured by Fourier transform infrared spectroscopy [from Queeney et al. (Ref. 119)].](image)

![FIG. 17. Fourier transform infrared spectra of the Ta₂O₅/Si system showing the progression of absorbance features in the Si–O stretch region, following various anneals [from Alers et al. (Ref. 103)].](image)

![FIG. 18. Synchrotron-based Si 2p x-ray photoemission spectra for Si(100)/SiO₂ and Si(111)/SiO₂ interfaces [from Himpsel et al. (Ref. 130)].](image)
TABLE III. Summary of XPS Si 2p analyses for the Si/SiO₂ system.

(1) Chemical shifts with respect to the Si substrate (Si⁰) peak, in eV, assuming five (Si⁰, Si¹⁺, Si²⁺, Si³⁺, Si⁴⁺) components of the Si 2p spectra. Second nearest neighbor effects [Banaszak-Holl et al. (Ref. 136)] are not considered.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Si¹⁺</th>
<th>Si²⁺</th>
<th>Si³⁺</th>
<th>Si⁴⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>0.95</td>
<td>1.75</td>
<td>2.48</td>
<td>3.90</td>
</tr>
<tr>
<td>132</td>
<td>1.00</td>
<td>1.65</td>
<td>2.50</td>
<td>4.00</td>
</tr>
<tr>
<td>743</td>
<td>0.95–1.00</td>
<td>1.72</td>
<td>2.50</td>
<td>3.60</td>
</tr>
<tr>
<td>134</td>
<td>0.98</td>
<td>1.82</td>
<td>2.65</td>
<td>3.84</td>
</tr>
<tr>
<td>138</td>
<td>0.97</td>
<td>1.73</td>
<td>2.58</td>
<td>3.57</td>
</tr>
<tr>
<td>744</td>
<td>0.94–0.95</td>
<td>1.78–1.80</td>
<td>2.50–2.56</td>
<td>4.00–4.04</td>
</tr>
</tbody>
</table>

Also, Tao et al. (Ref. 744), for a 0.9 nm film at different photon energies.

<table>
<thead>
<tr>
<th>Energy (eV)</th>
<th>I(Si⁰)</th>
<th>I(Si¹⁺)</th>
<th>I(Si²⁺)</th>
<th>I(Si³⁺)</th>
<th>I(Si⁴⁺)</th>
<th>I(SiO₂)/I(Si⁰)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>11.53</td>
<td>1.91</td>
<td>1.46</td>
<td>2.82</td>
<td>82.29</td>
<td>0.54</td>
</tr>
<tr>
<td>1.5</td>
<td>5.99</td>
<td>1.06</td>
<td>0.96</td>
<td>1.67</td>
<td>90.32</td>
<td>0.61</td>
</tr>
<tr>
<td>1.8</td>
<td>3.64</td>
<td>0.31</td>
<td>0.76</td>
<td>0.75</td>
<td>94.04</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Early XPS studies of Si–O–N films were limited to the detection of N at the interface of samples annealed in N. More recent high resolution photoemission studies of N(N 1s) core levels were useful in understanding N bonding and depth distribution, using either HF etch back or variable photoelectron takeoff angle methods. Assignments of N 1s peaks are indicated in Table IV, based on theoretical Si–N and Si–O–N bonding configurations depicted in Fig. 20. Depth analysis via XPS requires knowledge of the escape depths of the N 1s or Si 2p photoelectrons, which are on the order of 2.0–3.5 nm for conventional Al Kα or Mg Kα x-ray sources. A detailed analysis of the N 1s peak shape shows that the peak is asymmetric and consists of several components. Evidence for this can be found in Figs. 21 and 22.

X-ray scattering techniques have also been successfully used to study various structural aspects of ultrathin oxide films on Si, including interface roughness, variations in oxide density in the “bulk” of the film and near the interface, and ordered structures near the interface. The state of order at the interface is still a matter of debate and will be discussed in more detail in Sec. V.

X-ray diffraction crystal-truncation (CTR) profiling has been developed to measure roughness at buried interfaces, so that the SiO₂ layer need not be etched off. In the case of an amorphous overlayer, the variation of x-ray diffracted intensity along the truncated rod provides information about the root mean square value of roughness at the interface. Since the diffraction from the interface is rather weak, the

### Table IV. Summary of N 1s XPS analysis. Multiple (at least two) local electronic configurations of N atoms are observed in the films.

<table>
<thead>
<tr>
<th>N 1s peak at</th>
<th>Energy (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>397.6 eV</td>
<td>398.3–399 eV</td>
</tr>
</tbody>
</table>

located closer to the interface (Ref. 2)

N triple bonded to Si, i.e., as in Si₅N₄ (Ref. 137)

N 1s peak(s) at 400 eV
due to metastable bonding configurations of N (Ref. 172)

FIG. 19. Si 2p core level x-ray photoemission spectrum of the model interface derived from H₄SiO₄ clusters on Si, showing peak assignments [from Banaszak-Holl et al. (Ref. 136)].

FIG. 20. N 1s core level x-ray photoemission spectrum shifts at Si(100)/SiO₂ interfaces, calculated for N–Si₅ (circle), N–Si₂O (square), and N–Si₅O₃ (triangle) configurations, as a function of distance, z, from the interface [from Rignanese et al. (Ref. 748)].
experiment is best performed with a synchrotron radiation source. X-ray diffraction measures the Fourier transform of the electron density. By measuring and modeling intensity profiles, one can obtain statistical information about the smoothness of the interface. The technique has been used to explore the quality of the interface as a function of processing parameters such as growth temperature, Fig. 23, and interface nitridation.\textsuperscript{173–175,190} It has also been demonstrated that CTR results correlate well with roughness measurements deduced from the SHG technique.\textsuperscript{191}

X-ray reflectivity is another method for studying the interface, since the reflectivity represents the Fourier transform of the spatial derivative of the sample density in the perpendicular direction. Modeling x-ray reflectivity spectra with film thickness and density as fitting parameters yields information about structural properties of oxide films. In a recent study of thin rapid thermal oxides (RTO),\textsuperscript{184} Fig. 24, the best fit to the data was found using a two layer model, the thin (~1.5 nm) interfacial oxide layer being slightly denser than the overlying “bulk” oxide layer.

C. Ion beam analysis

The family of ion beam based techniques includes Rutherford backscattering spectroscopy (RBS),\textsuperscript{192–195} medium energy ion scattering spectroscopy (MEIS),\textsuperscript{2,103,196–209} nuclear reaction analysis (NRA),\textsuperscript{25,37,210–226} and secondary ion mass spectroscopy (SIMS).\textsuperscript{227–233} These techniques provide very accurate determinations of composition and distribution of Si, O, N, and other elements in dielectrics.

In RBS, a high energy ion beam (typically H\textsuperscript{+} or He\textsuperscript{+}, with greater than MeV energy), scatters off the target. The scattering event is described by a Coulomb interaction, the cross-section of which is well known. The energy of the backscattered ions is determined by the energy of the incident particles, the masses of the primary ions and the target atoms, and the scattering geometry, all based on a binary collision model. Typical backscattering data is illustrated in Fig. 25. The integrated area of a backscattered peak can be easily converted into a concentration of a particular element in the target, as is done for the case of SiO\textsubscript{2} films in Fig. 26. Experiments can be performed in either random or channeling scattering geometries. In the channeling geometry, the ion beam is aligned along a crystallographic direction of the target substrate. This significantly reduces the background scattering from the substrate and increases sensitivity to the overlayer. Since impinging particles in the MeV energy range lose relatively little energy upon scattering, energy shifts of the backscattered peaks will be small compared to the energy resolution of the solid state detectors used in RBS experiments. The result is that RBS has relatively poor depth resolution, i.e., 10 nm or greater, and is therefore not useful in the ultrathin dielectrics regime, although depth resolution
can be improved by utilizing glancing angle geometries. Despite this shortcoming, it is important to know the integrated elemental concentrations in multicomponent films, and RBS is a very accurate technique for measuring film composition and stoichiometry, as well as impurity concentrations.

MEIS is based on the same ion–solid interaction principles as conventional RBS. A schematic of a typical MEIS experiment is illustrated in Fig. 27. Compared to RBS, MEIS uses lower ion beam energies, typically 100–200 keV. In this energy range, \( \frac{dE}{dx} \), the stopping power, is at its maximum for protons or \( \alpha \) particles in solids, resulting in optimal depth resolution. For example, the stopping power for \(~100\) keV protons is 130 eV/nm for Si, 120 eV/nm for SiO\(_2\), 180 eV/nm for Ta\(_2\)O\(_5\), and 200 eV/nm for Si\(_3\)N\(_4\). More importantly, in MEIS backscattered ion energies can be measured with a high resolution toroidal electrostatic energy analyzer (\( \Delta E/E \approx 0.15\% \), i.e., \( \Delta E \approx 150\) eV for 100 keV ions), greatly increasing the energy resolution relative to the detectors used in conventional RBS. Taking into account the values of the stopping power and the travel length factor, this energy resolution translates into almost monolayer depth resolution in the near surface region. As an example, it has been demonstrated that 0.3–0.5 nm accuracy in N depth profiles, near the top surface or for ultrathin films, can be achieved,\(^{201}\) as is illustrated in Fig. 28. However, due to the statistical nature of ion energy loss in solids, known as the straggling effect, the
resolution decreases with increasing distance from the surface. We also note that a thorough understanding of straggling below the 2 nm range remains elusive.

MEIS allows one to simultaneously monitor absolute concentrations and depth profiles of N, O, Si, and other elements in dielectric films.\textsuperscript{199–201,203,204,234–238} As in RBS, the detection limit is proportional to the square of the atomic number. Thus sensitivity is very high for heavy elements, but decreases with atomic number. For example, the detection limit of N in SiO\textsubscript{2} is \(\sim 2 \times 10^{14}\) N/cm\textsuperscript{2}, depending on film thickness and the width of the N distribution in the film. The detection of very light elements such as H and B can be enhanced in the elastic recoil configuration, as has been demonstrated.\textsuperscript{198,239} Fig. 29. Light elements can also be detected in a forward scattering geometry, usually using a heavier element as the incident particle to increase the scattering cross section and to facilitate detection. This method is called elastic recoil detection (ERD),\textsuperscript{240,241} or in the case of H detection, H forward scattering (HFS). Habraken and colleagues examined N, O, Si, and H using ERD of primary MeV ions.\textsuperscript{37,225} More recently, Brijs\textsuperscript{242} used a 60 MeV Ni beam, in a grazing angle geometry with a magnetic sector spectrometer, to look at forward scattering from ultrathin Si–O–N layers. An example of ERD to detect trace H in a Si\textsubscript{3}N\textsubscript{4} film\textsuperscript{241} is shown in Fig. 30.

It is worth noting that contrary to other depth profiling techniques such as SIMS, both the sensitivity to a given element in the film and the depth resolution increase with decreasing dielectric film thickness. Another strength of MEIS and most other ion scattering methods is mass sensitivity, which enables isotopic labeling experiments. The use of O isotopes \((^{18}\text{O}\text{ and } ^{16}\text{O})\), in combination with high resolution depth profiling, has been used to study the mechanism of Si oxidation in the ultrathin regime.\textsuperscript{199,200,204,238} The principle of such an \(^{18}\text{O}\) labeling study is shown in Fig. 31. MEIS has also proven itself useful in the arsenal of tools used to analyze ultrathin high-\(\kappa\) metal oxides.\textsuperscript{103,206–208} Fig. 32.

In addition to elastic scattering during ion–solid interactions (the bases of RBS and MEIS), high energy ions can induce nuclear reactions in the target.\textsuperscript{25,216} NRA techniques are based on the detection of protons, \(\alpha\) particles, and \(\gamma\) rays generated in nuclear reactions induced by the collision of high energy charged particles with N, O, Si, H, D, and other elements. Table V is a compilation of those nuclear reactions that have applicability for the study of dielectrics. The cross sections for these reactions can be determined in independent calibration experiments, and the signal-to-background ratio is in many cases quite favorable. Since the nuclear reaction rates are only sensitive to the number of nuclei in the films, the technique yields absolute concentrations of the species investigated. For example, the technique allows one to determine the absolute concentration of \(^{14}\text{N}\) with 7\%–10\% accuracy at a detection limit of \(\sim 5 \times 10^{13}\) N/cm\textsuperscript{2} via the reaction \(^{14}\text{N}(d,\alpha_{0.1})^{12}\text{C}\), induced by 1.1 MeV deuterons.\textsuperscript{243} The O

![FIG. 27. Principles of a medium energy ion scattering experiment. MEIS is essentially a low energy, high resolution variant of Rutherford backscattering (RBS). \(E\) and \(E'\) are the incident and scattered particle (\(H^+\)) energies, \(M_1\) and \(M_2\) are the particle and target atomic masses, and \(\theta\) is the scattering angle.](Image 97x677 to 252x739)

![FIG. 28. A medium energy ion scattering study of a thin (\(\sim 2\) nm) Si–O–N film grown on Si(100) in a furnace at 950 °C for 60 min in NO ambient. The upper panel shows the scattering spectrum measured at a scattering angle of 125°. The inset shows the corresponding N and O profiles as a function of distance from the oxide surface. The lower panel shows a closeup of the spectrum in the N region along with the results of simulations for different N distributions: (1) all N (determined to be \(8.8 \times 10^{14}\) N/cm\textsuperscript{2}) is located at the Si/SiO\textsubscript{2} interface, (2) all N is uniformly distributed in the film, (3) N profile shown in the inset (best fit), and (4) N profile same as (3) but extended 0.3 nm into the substrate. Therefore profiles (1), (2), and (4) can be eliminated from consideration [from Lu et al. (Ref. 201)].](Image 319x259 to 557x739)
content in films can also be measured by NRA, using the \( ^{16}\text{O}(d,p)^{17}\text{O} \) reaction at 850 keV.\(^{212}\) Knowledge of these concentrations allows one to calculate film thickness, assuming that film density is known.\(^{244}\) An additional advantage of the NRA technique is the detection of \( ^{15}\text{N} \) and \( ^{18}\text{O} \) isotopes with the help of the following reactions: \( ^{15}\text{N}(p,\alpha)^{12}\text{C} \) at 1000 keV, \( ^{18}\text{O}(p,\alpha)^{15}\text{N} \) at 730 keV, and the resonant reactions of \( ^{15}\text{N}(p,\alpha)^{12}\text{C} \) at 429 keV and \( ^{18}\text{O}(p,\alpha)^{15}\text{N} \) at 151 keV. An example of NRA analysis is shown in Fig. 33. Due to the narrow resonances of the latter two reactions, widths of 120 and 100 eV, respectively, they can be used for depth profiling of \( ^{15}\text{N} \) and \( ^{18}\text{O} \) on about a 1 nm depth scale, under favorable conditions.\(^{219,245,246}\) Nitrogen depth profiling can also be performed by HF acid etch back and subsequent NRA measurement of \( ^{15}\text{N} \) remaining in the film.\(^{173}\) Isotopic labeling, in combination with NRA mea-

### TABLE V. Compilation of useful nuclear reactions for gate dielectrics analysis.

<table>
<thead>
<tr>
<th>Target element</th>
<th>Nuclear reaction</th>
<th>Energy, keV</th>
<th>Sensitivity, atoms/cm(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>(^{15}\text{N},\alpha)^{12}\text{C} )</td>
<td>6420</td>
<td>(-10^{14})</td>
</tr>
<tr>
<td>D</td>
<td>(^{15}\text{N}(d,\alpha)^{12}\text{C} )</td>
<td>1100</td>
<td>(-5\times10^{13})</td>
</tr>
<tr>
<td>(^{15}\text{N})</td>
<td>(^{15}\text{N}(p,\alpha)^{12}\text{C} )</td>
<td>1000</td>
<td>(-10^{12})</td>
</tr>
<tr>
<td>(^{16}\text{O})</td>
<td>(^{16}\text{O}(d,\alpha)^{12}\text{C} )</td>
<td>429</td>
<td>(-10^{13}) (estimated)</td>
</tr>
<tr>
<td>(^{17}\text{O})</td>
<td>(^{17}\text{O}(p,\alpha)^{12}\text{C} )</td>
<td>850</td>
<td>(-10^{13})</td>
</tr>
<tr>
<td>(^{18}\text{O})</td>
<td>(^{18}\text{O}(p,\alpha)^{15}\text{N} )</td>
<td>730</td>
<td>(-10^{12})</td>
</tr>
<tr>
<td>(^{19}\text{Al})</td>
<td>(^{19}\text{Al}(p,\alpha)^{12}\text{C} )</td>
<td>151</td>
<td>(-10^{13}) (estimated)</td>
</tr>
</tbody>
</table>

### FIG. 29. Medium energy ion scattering elastic recoil spectra for a H terminated Si(100) surface followed by deposition of an ultrathin amorphous Si overlayer (upper panel), and B segregated to the Si(111) surface (lower panel) [from Copel et al. (Ref. 198)].

### FIG. 30. An elastic recoil detection spectrum collected from a Si\(_3\)N\(_4\) layer –300 nm thick. A 28 MeV Si beam was used. The H is present as a minor constituent of the film, a remnant of chemical vapor deposition at 800 °C. The H signal is sitting on the N signal background [from Barbour et al. (Ref. 241)].

### FIG. 31. Principles of an isotopic labeling (in this case \(^{18}\text{O}_2\) reoxidation of an Si\(^{16}\text{O}_2\) film) medium energy ion scattering experiment. \(^{18}\text{O}_2\) is detected at the interface (new oxide growth) and at the oxide surface (surface exchange reaction).
surements, has proven useful for studying transport and incorporation of N and O during thermal oxynitridation in NO or N2O.25 A further useful application of NRA, Fig. 34, is based on its ability to detect H/D using the reactions H($^{15}$N,αγ)$(^{12}$C at 429 keV and D($^3$He,p)$(^4$He at 700 keV, since H and D are believed to be important to device performance.220,221,225,226

SIMS is an industry standard technique, used to monitor concentration profiles in semiconductor structures. It has been used very effectively to study growth mechanisms of oxides,228,230,247 as well as the incorporation of N227,248 and H231,249,250 in oxides. The SIMS experiment involves continuous sputtering of the sample and accurate measurement of the sputtered species by mass spectroscopy. Relative to RBS or MEIS, the technique has a very high sensitivity, on the order of 0.001 at. %, can be performed rapidly, and is more commonly available. However, the technique begins to reach its limiting depth resolution (estimated to be ~2 nm) for sub-5 nm films. Another more important complication of SIMS analysis is the matrix effect, in which the sputtered ion yield depends strongly on the local chemistry around the element. For example, in Si–O–N films, the ion yield of CsN+ (Cs+ being the sputtering ion) from N in bulk Si is about six times smaller than N in SiO2. Further, the ion yield for N near the interface is about three times that observed in the bulk of an SiO2 film.227 This effect is shown in Fig. 35. The use of CsN+ ions as the detected species seems to minimize the matrix effect, since for negative ions matrix effects are even more severe. In many cases, SIMS analysis is also complicated by surface contamination and initial sputtering effects, which make quantitative measurements of the com-

FIG. 32. Medium energy ion scattering depth profiles for a thin (7 nm) Ta2O5 film as-deposited on Si (upper panel) and after an ultrahigh vacuum anneal at 950 °C for 30 min (lower panel). The annealed sample shows significant intermixing near the interface [from Alers et al. (Ref. 103)].

FIG. 33. Isotopic labeling during N2O oxynitridation of Si, studied in combination with nuclear reaction analysis. Excitation curves are shown for the nuclear reaction $^{18}$O(p,α)$^{15}$N, for SiO2 reoxidized in $^{18}$O2 (top), and Si oxynitrided in N2O and reoxidized in $^{18}$O2 (middle). $^{18}$O depth profiles are shown in the insets. The bottom figure is the excitation curve of the nuclear reaction $^{15}$N(p,αγ)$^{12}$C for a Si–O–N film obtained by thermal annealing of SiO2 in $^{15}$N2O. The inset shows the $^{15}$N depth profile [from Ganem et al. (Ref. 217)].

FIG. 34. Nuclear reaction analysis detection scheme and a typical spectrum for D in SiO2 films [from Baumvol et al. (Ref. 220)].
position near the top ~1 nm surface layers difficult. However, high depth resolution SIMS analysis has proven very useful for investigating ultrathin Si–O–N films. One more aspect of SIMS analysis is that the peak concentration of an element, and details of its profile in ultrathin films, may not be accurate, due to ion mixing. The measured areal density of the element should be used instead. Reference samples calibrated by other quantitative methods ~NRA, MEIS, etc. allow for more accurate SIMS analysis.

Finally, we mention one more high energy technique, positron spectroscopy, that has potential for the study of defects in MOS structures. An excellent example of its application to the study of the Si/SiO$_2$ interface is shown in Fig. 36. Initially, TEM was used as a direct way to measure oxide thickness and film roughness. However, SiO$_2$ films thinner than ~2 nm the use of TEM to measure oxide thickness becomes problematic due to uncertainty in defining the Si/SiO$_2$ and polycrystalline Si/SiO$_2$ interfaces, since the image is averaged over the travel length of electrons in the sample.~$^{40}$ Over the past decade, electron microscopy has advanced in sophistication along two main directions: atomic resolution imaging and real-time observation of processes such as oxidation. Further, atomic column scanning in TEM has been made possible by generating very small (~0.2 nm) electron probes. Using such probes, atomic number ($Z$) contrast structure imaging and electron energy loss spectroscopy (EELS) can be performed. A schematic diagram of an atomic probe microscopy apparatus with EELS is shown in Fig. 37. The $Z$-contrast method allows one to obtain direct information about atomic positions, since scattering power varies with atomic number. In EELS, a fraction of the transmitted electrons lose energy by exciting a target atom’s core electrons, and are thus characteristic for that atomic species, as well as its chemical state. EELS can therefore probe variations of composition as a function of distance from the interface, even for amorphous materials (see Fig. 6).

Real-time measurement capabilities are de rigueur for studies of dynamic processes. TEM and SREM have been used to understand the details of Si oxidation and etching (from SiO desorption at elevated temperatures). Real-time studies of the interaction of O with the Si(111) surface as a function of temperature, pressure, and time have been performed. A layer-by-layer oxide growth mode, illustrated in Fig. 38, was clearly observed at high pressure and low temperature. At high temperature and low pressure, the SiO desorption regime, step flow was observed. The intermediate regime was characterized by surface roughening. Layer-by-layer oxidation of Si(001) was also observed by SREM, a technique illustrated in Fig. 39. The initial stages of Si oxidation, Fig. 40, as well as nitridation have also been studied by low energy electron microscopy (LEEM).
E. Scanning probe microscopy

Scanning probe microscopy techniques such as atomic force microscopy (AFM) and scanning tunneling microscopy (STM) are important methods for imaging the surface morphology of ultrathin dielectric films. Since the invention of STM almost 20 years ago, probe microscopy methods have been successfully applied to study surface structure, interfacial roughness, wafer surface morphology after various pregate oxidation cleaning steps, initial stages of Si oxidation, initial stages of nitridation, and oxide decomposition. An example of the applicability of STM is shown in Fig. 41. Spectroscopic mode STM experiments were used to study oxide breakdown mechanisms, and also to identify O bonding configurations during initial oxidation. A fundamental limitation of probe microscopy methods is related to the probe tip size and shape. STM usually affords the best lateral and vertical resolution. Since the radius of the tip is typically on the order of ~10 nm, it is difficult to observe smaller features, especially if they have a high aspect ratio. In general, the lateral resolution is better for STM than AFM, in part because in AFM the tip tends to broaden due to contact with the sample. A fundamental limitation for STM as it applies to SiO2-based dielectrics is the large band gap of SiO2 (~9 eV), which usually limits STM studies to films less than ~1.5 nm thick. Alternate (high-κ) dielectrics have smaller band gaps and are often O deficient, resulting in energy gap states. These properties result in a higher conductivity (at a given voltage and equivalent thickness), than SiO2, enabling the use of STM for the examination of thicker, high κ dielectric films.
Ballistic electron emission microscopy (BEEM), scanning microwave microscopy (SMM), scanning capacitance microscopy (SCM), scanning tunneling spectroscopy (STS), and hot electron emission microscopy (HEEM) are recent variations of scanning probe techniques for the study of thin dielectric films. The BEEM technique, shown in Fig. 42, is based on the local ballistic transport of electrons emitted from an STM tip through a thin metal electrode/thin oxide/Si structure. It has been used to better understand the mechanism of defect generation in thin oxides, and therefore oxide breakdown. In HEEM, hot electrons emitted into vacuum from biased MOS structures form a direct image of spatial emission distribution and can be used to investigate prebreakdown phenomena with less than 20 nm spatial resolution. SMM measures the response of a thin dielectric film to a microwave cavity scan. It allows one to measure the dielectric constant of the film, making the technique especially attractive for high \( k \) studies, as is shown in Fig. 43. However, poor lateral resolution precludes its use for studying small features. SCM has recently been used to study the gate region of ultralarge scale integration (ULSI) transistors, but its low lateral resolution limits its ability to image the gate. SCM allows one to image the tunneling current through an oxide at a given bias, while simultaneously recording the AFM image.

III. ELECTRICAL CHARACTERIZATION METHODS AND PROPERTIES

In this section we will discuss the electric characterization of ultrathin oxide layers. As the thickness is scaled from 4 to 1 nm, an increasing complexity in the interpretation of electrical data is observed. We will show that capacitance–voltage \((C-V)\) measurements require quantum mechanical corrections, gate leakage current is dominated by direct, rather than Fowler–Nordheim tunneling, and charge pumping measurements need corrections for excessive leakage current. When electrically stressed, the basic oxide degradation mechanisms described for thicker oxides \((t_{ox}>4 \text{ nm})\) are still applicable. Oxide wearout can be described as the continuous creation of trapping centers. These traps cause stress-induced leakage current, and can trap charge until breakdown is finally triggered. The occurrence of soft breakdown
A. Characterization methods

In this section, the application of electrical characterization techniques to sub-4 nm oxides will be discussed. We will focus on the three most commonly used measurements: $C-V$, leakage current, and charge pumping (CP).

1. $C-V$ measurements

The most frequently used electrical technique to assess the properties of both the thin oxide layer and its interface with Si is the $C-V$ measurement. In thicker oxide layers $C-V$ curves can be fitted satisfactorily with classical models, described in textbooks. The $C-V$ technique can be used to determine flatband and threshold voltage, fixed charge, and interface state density. It is also often used to determine the oxide thickness.

In sub-4 nm oxide layers, $C-V$ measurements provide the same information, but the interpretation of the data requires considerable caution. The assumptions needed to construct the “classical model” are no longer valid, and quantum mechanical corrections become mandatory, thus increasing the complexity of the analytical treatment. First, several authors have demonstrated that for ultrathin layers, Maxwell–Boltzman statistics no longer describe the charge density in the inversion and accumulation layers satisfactorily, and should be replaced by Fermi–Dirac statistics. In addition, band bending in the inversion layer near the semiconductor–insulator interface becomes very strong, and a potential well is formed by the interface barrier and the electrostatic potential in the semiconductor. This potential well may be narrow enough to give rise to electron confinement at discrete energy levels, as illustrated in Fig. 44(a). The correct analytical treatment requires solving the coupled effective mass Schrödinger and Poisson equations self-consistently. Closed form analytical treatments require a simplification of the problem, e.g., replacing the actual potential well by a triangular well and/or by considering only the lowest subbands. One of the main effects of the quantum mechanical treatment of the inversion layer is a considerable shift of the inversion charge centroid away from the semiconductor–insulator interface, as illustrated in Fig. 44(b). This effect can be modeled as an additional capacitance in series with the oxide capacitance.

A similar effect is generated by polycrystalline Si depletion on the gate side of the capacitor of a MOS transistor. This effect is related to both the high fields at the insulator surface as well as the incomplete activation of the dopants near the polycrystalline Si/SiO$_2$ interface. A carrier concentration profile with a finite width, having a centroid several tenths of a nanometer away from this interface, results. This effect can also be modeled as an additional capacitance in series with the oxide capacitance. As a consequence of quantum mechanical effects and polycrystalline Si depletion, the measured capacitance is smaller than the expected (physical) oxide capacitance, and the difference becomes very significant for ultrathin layers. This also implies that oxide thickness extraction from $C-V$ measurements becomes more difficult, but not impossible. If $C-V$ curves are fitted properly, a good agreement between $C-V$ extracted...
oxide thickness and physical oxide thickness, measured by ellipsometry, for example, may be obtained, as illustrated in Figs. 10 and 45.60,108,317,318,320

For very thin oxides ~ typically sub-2 nm!, the huge leakage current through the oxide, due to direct tunneling of electrons ~see Sec. III A 2 and Fig. 5!, creates an additional complication in the interpretation of C–V curves. A sharp drop in the capacitance is observed as the voltage increases.321,323,324 This effect is illustrated in Fig. 46 and can be modeled by taking into account the tunnel conductance and an additional series resistance.

2. Gate tunnel current

When a voltage, V_{ox}, is applied across an oxide layer with thickness t_{ox}, the resulting oxide field, E_{ox} = V_{ox} / t_{ox}, gives rise to a current flow through the oxide. This current originates from electrons that quantum mechanically tunnel through the Si/SiO₂ potential barrier from the Si conduction band to the SiO₂ conduction band, as is illustrated in Fig. 47. When the tunneling occurs through a triangular barrier, Fig.

\[
J_{FN} = A \cdot E_{ox}^2 \exp \left( -\frac{B}{E_{ox}} \right),
\]

(3)

A and B are two constants, and B is related to the electron effective mass in the oxide conduction band and the Si/SiO₂ barrier height. In thin oxide layers, oscillations in the gate

FIG. 44. (a) Schematic representation of a potential well in strong inversion. In the state of weak inversion the triangular shape is adequate. In the quantum mechanical picture the energy spectrum consists of a discrete set of energy levels. The first energy level does not coincide with the bottom of the conduction band. (b) Electron density as a function of distance from the interface. The average distance to the interface is larger in the quantum mechanical framework. Therefore a larger band bending is needed for a given population in the conduction band ~from Van Dort et al. ~Ref. 319!.

FIG. 45. Comparison between capacitance–voltage extracted oxide thickness and thickness measured by ellipsometry. Every ellipsometric thickness represents the average of nine points and every C–V extracted thickness represents the average of 64 capacitors ~area 10⁻⁴ cm²! over a 200 mm wafer. The C–V extracted thickness obtained using quantum mechanical models agrees well with the ellipsometric thickness ~from Lo et al. ~Ref. 318!.

FIG. 46. Experimental metal–oxide–semiconductor capacitance–voltage curves for 1.3, 1.5, and 1.8 nm SiO₂ films, respectively. The sharp decrease of the gate capacitance results from high gate tunneling current ~from Choi et al. ~Ref. 321!.

FIG. 47 ~left!, the conduction mode is called Fowler–Nordheim ~FN! tunneling325 and the measured current density, J_{FN}, can be described by the well-known formula:
current occur. These arise from the interference of incident and partially reflected electron waves propagating between the conduction band edge and the anode interface.326–328

When the oxide voltage drops below 3.7 V, electrons no longer enter the oxide conduction band, but tunnel directly from the anode to the cathode, as illustrated in Fig. 47 (right). In state-of-the-art CMOS technologies, direct tunneling is the dominant current conduction mechanism at operating voltage, and for oxide layers less than ~3 nm it is also the conduction mode for accelerated oxide wearout and breakdown tests. The direct tunneling current density cannot be described easily in a closed analytical form, but several approximate formulas and computer simulations can be found in the literature.48,329–333 The general trend in these analytical formulas is that precision is bought at the expense of complexity. The simplest analytical expressions fail in the low voltage range, the very range of interest to forecast the leakage current at operating voltage. In Fig. 48, the $I_G-V_G$ characteristics for oxides between 9.7 and 2.2 nm thick are shown. The change from a FN tunneling to a direct tunneling mechanism is indicated. The huge increase due to direct tunneling, as shown in Fig. 5, poses a major leakage current problem in VLSI technologies and severely limits the scaling of the oxide thickness.

A special case arises when a $p^+$ polycrystalline Si/oxide/Si structure is biased with negative gate voltage. Since the inversion layer cannot form in the highly doped $p^+$ polycrystalline Si, electrons are injected from the valence band. This is evidenced by the higher potential barrier extracted from the $I_G-V_G$ characteristic, as well as by experiments using a polycrystalline SiGe electrode.334–336 Charge separation experiments demonstrated that in this case, for ultrathin oxides at low voltage, hole tunneling from the anode dominates the gate current.337

3. Charge pumping

Charge pumping is a technique well suited to electrically characterize the Si/SiO$_2$ interface, and it has been used to study both interface quality and subsequent degradation under electrical stress.338,339 Charge pumping consists of measuring the dc current at the Si substrate, due to recombination processes at interface defects, while pulsing the interface from accumulation to inversion under the action of periodic gate pulses. When ultrathin oxide layers are investigated, a parasitic current component, resulting from direct electron tunneling, is added to the charge pumping current, as illustrated in Fig. 49.340 This effect is very similar to that observed in $C-V$ measurements (see Fig. 46) and can lead to large errors in the extracted interface trap density and the flatband and threshold voltages. However, since the charge pumping current is proportional to the measurement frequency, whereas the leakage current is not, the charge pumping signal can be easily extracted by combining low and high frequency measurements.340 Consequently, charge pumping remains a simple and accurate method that allows the determination of interface trap density, without the quantum mechanical modeling required for $C-V$ measurements.

B. Oxide degradation during electrical stress

In this section oxide degradation during electrical stress, ultimately leading to breakdown, is discussed. We define oxide degradation as the continuous, gradual deterioration of the oxide properties, resulting from structural damage generated in the oxide by electrical stress. Breakdown is triggered when the accumulated damage reaches a critical level. Dur-

---

FIG. 47. Schematic illustration of Fowler–Nordheim (left) and direct (right) tunneling mechanisms of electron flow through an oxide potential barrier of height $\Phi_B$.

FIG. 48. The gate current density as a function of gate voltage for SiO$_2$ films with thicknesses between 2.2 and 9.7 nm. The transition from Fowler–Nordheim to direct tunneling is indicated [from Degraeve et al. (Ref. 3)].

FIG. 49. Frequency dependence of the charge pumping current as gate pulse base level on a 10×3 mm$^2$ N-channel transistor having 1.8 nm gate oxide. Measurement conditions: sinusoidal gate pulses, 2 V pulse amplitude, 0 V reverse bias. The charge pumping signal is superimposed on the gate oxide leakage current [from Masson et al. (Ref. 340)].
ing oxide stressing, several properties can be monitored, such as interface trap creation, negative and/or positive charge trapping, hole fluence, neutral electron trap creation, and the generation of a stress-induced leakage current (SILC). These properties are important indicators of oxide stress and can help one understand degradation mechanisms.

Several research groups have proposed breakdown models that directly correlate one of these properties with breakdown. One line of thought is common to all models: some damage related parameter exceeds a critical threshold at the moment breakdown is triggered. An important observation is that a relationship between accumulated oxide damage and breakdown is only found for intrinsic breakdown. Extrinsic breakdown is determined by localized, defect related physical processes that do not influence global degradation phenomena.

1. Interface trap creation

During high field oxide stressing, interface traps are created at the Si/SiO₂ interface. Their density can be obtained either from $C-V$ or charge-pumping measurements on transistors. Recently, low voltage SILC measurements on sub-3.5 nm oxides have been used to quantify the interface trap density. The mechanisms proposed for interface trap generation correspond to those for bulk trap generation, and will be discussed later in this section.

It has been claimed that the interface trap density, $D_{it}$, reaches a critical density, $D_{it,crit}$, at the moment of oxide breakdown. In earlier work, the trigger of breakdown was thought to be caused by a local interface softening due to trap accumulation, but more recently the critical interface trap density is merely viewed as a monitor for the total density of traps in the oxide. By means of a percolation model (to be discussed in detail later), the interface trap density can be related to this total density.

2. Oxide charge trapping

In oxides of thickness greater than 4 nm, a typical observation during high field constant current stressing (CCS) is the initial decrease of the applied voltage needed to achieve the required current, followed by a voltage increase which can become larger than the initially applied voltage. The voltage shifts are caused by charge trapping of initially positive then negative charges, leading to oxide field distortion and subsequent change in the tunnel current density. During constant voltage stressing (CVS), exactly the opposite current shifts are measured, i.e., an initial increase of the current followed by a decrease.

In sub-4 nm oxides the oxide buildup almost completely disappears. Typically, as illustrated in Fig. 50, a very small increase of the stress current during CVS is measured, which is attributed to positive charge trapping and the gradual generation of a SILC. In some previous work it was claimed that positive charge trapping in the oxide was responsible for triggering the breakdown event. During CCS, for example, locally enhanced charge trapping will not influence the total current density in the capacitor; however, it may lead to a local current density increase, resulting in an increased stress, which in turn leads to further positive charge trapping. In this way, a positive feedback mechanism is initiated that finally results in breakdown. Two arguments oppose this idea: (i) in ultrathin sub-4 nm oxides the measurable positive charge trapping is extremely small and yet the oxides break down, and (ii) the possible role of the negative charge in the oxide is completely ignored, while for $t_{ox}>4$ nm the net charge trapped at breakdown is negative. Other work claims that the net negative trapped charge in the oxide exceeds a critical threshold value at breakdown. Vincent et al. related their observation to charge trapping kinetics, without assuming any additional trap creation. This is in contradiction with most work on oxide degradation that clearly show an increase in the trap density.

3. Hole fluence

When the gate oxide of an NMOSFET is stressed with a positive gate voltage, while source and drain are grounded, electrons tunneling through the oxide are provided from the source and drain and injected from the transistor channel. In this configuration, a positive current can be measured at the substrate (charge separation technique). The substrate current density has a similar oxide field dependence as the FN current density, as is shown in Fig. 51. However, it should be noted that the curves in Fig. 51 are not parallel; the ratio between gate and substrate current depends on the oxide field. A well-known and widely accepted explanation for the physical origin of the substrate current has been given, and is schematically illustrated in Fig. 52. When the injected electrons enter the anode (the polycrystalline Si gate), they lose their energy by creating high energy holes, possibly through the excitation of some intermediate state, and these holes can then be injected back into the oxide. The hole fluence reaches the cathode and is measured as a positive substrate current, $J_p$. In ultrathin oxide layers the valence band injection of electrons can generate a hole current that may dominate the anode hole-generated current.

Another possible explanation for the substrate current is the creation of holes in the cathode by photons generated in the anode. This has been suggested by DiMaria et al., and
recently experimental evidence supporting this model has been presented.359 The hole current density can be related to the electron current density, $J_n$, as follows

$$J_p = \alpha(E_{ox}) J_n,$$

where $\alpha(E_{ox})$ is the field dependent hole generation efficiency. In the anode hole injection model, $\alpha$ is interpreted as the probability for a tunneling electron to generate an anode hole which is injected back into the oxide towards the cathode.356 Under constant voltage stress, $\alpha$ is found to be almost constant,330 while under constant current stress, $\alpha$ increases slightly. However, for thin oxides the difference between the initial and final values of $\alpha$ is very small. Therefore an approximate equivalent relation to Eq. (4) also holds for the integrated values of $J_p$ and $J_n$, the hole fluence, $Q_p$, and the electron fluence, $Q_n$:

$$Q_p = \alpha(E_{ox}) Q_n.$$

Chen et al.354 observed for the first time that the hole fluence reaches a critical value at breakdown, designated $Q_{p,crit}$. This result has been confirmed360,364 and further, substrate hot electron injection experiments indicate that a critical hole fluence is needed to trigger oxide breakdown.362,363 In Fig. 53, $Q_{p,crit}$ and $Q_{BD}$, measured over a wide field range of 8–14 MV/cm, are plotted. Clearly, $Q_{p,crit}$ remains constant over the entire field interval, while $Q_{BD}$ decreases with increasing field. A satisfactory physical explanation for the experimentally observed invariance of $Q_{p,crit}$ cannot be found in the literature, but the various suggested possibilities will be discussed in detail in Sec. III B 6.

It should be noted that in the work of Satake et al.,360 $Q_{p,crit}$ was observed to be constant only at 300 K; at lower temperatures it decreased as a function of the field. This observation is consistent with the findings of Kaczer et al.,364 where an increase as a function of field is found for temperatures above 300 K. These temperature effects indicate that the hole fluence at breakdown is possibly not the determining factor for breakdown.

4. Neutral electron trap generation

During oxide stressing, neutral electron traps are generated in the oxide.365,366 Although many researchers have related such trap creation to the breakdown process, direct measurements of the oxide neutral trap density, $D_{ot}$, as a function of the applied stress conditions, are not commonly done. In order to measure this degradation phenomenon, the neutral traps have to be made electrically visible. This can be accomplished by periodically interrupting the stress to fill the generated traps with electrons, without creating any additional traps. A technique ideally suited for this purpose is uniform substrate hot electron (SHE) injection.367–369 In some cases368,369 the necessity of a trap filling step is demonstrated. Indeed, immediately after stress, a majority of the available oxide traps are neutral, and further, the probability of occupancy depends on the applied stress field. Therefore the trapped charge, measured immediately after stress, is not a good indication of the neutral trap density.

Substrate hot electron injection can also be used to stress the oxide in a field range below the practically accessible FN field range.345,370 In Fig. 54, the neutral trap density increase is shown as a function of the injected fluence. Degradation and breakdown in the field ranges of 6.0–8.5 MV/cm with SHE injection and 8.5–11.0 MV/cm with constant voltage FN injection are presented. The fraction of filled traps after SHE injection is found to be constant for all stress fields.
In the field range under consideration, breakdown occurs when $D_{\text{el}}$ reaches a critical value, within some small statistical fluctuation. This result supports the idea that a critical electron trap density is necessary to trigger breakdown.

A breakdown model based on neutral trap generation has been proposed by several authors.55,345,365,371–373 It is assumed that at some place on the capacitor the local trap density becomes sufficiently large to allow the formation of a conductive chain of traps connecting the anode with the cathode interface. This model will be investigated in further detail in Sec. III C 1.

In Fig. 55 the generated neutral trap density has been plotted as a function of hole fluence for different oxide thicknesses and FN stress conditions. A unique relationship is observed, independent of oxide field and thickness.345 Thus it can be concluded that the critical hole fluence, $Q_{\text{p,crit}}$, corresponds to a critical generated density of neutral electron traps, $D_{\text{el,crit}}$, proving that both breakdown criteria are equivalent.

5. Stress-induced leakage current

The fifth phenomenon that occurs during oxide degradation is the generation of a SILC through the gate. Stress induced leakage current is illustrated in Fig. 56, where $I_g - V_g$ curves from fresh and stressed samples are shown. The SILC rises continuously with injected fluence, and its $V_g$ dependence can be empirically fitted with a FN expression using a barrier height of 1 eV.374–377 When the SILC is continuously monitored as a function of time after a given stress, two components can be distinguished.366,378 Initially, a decaying transient component is observed, leading to a steady state SILC after some time. Both components depend on oxide thickness, as is illustrated in Fig. 57. Thick oxides have a large transient component and low steady state component, whereas ultrathin oxides have a very small transient component and a large steady state component.

The SILC itself is in some cases an important reliability problem. Leakage current through the gate translates to a power waste problem in MOSFETs, resulting, for example, in having to constantly charge the batteries of portable devices. Such leakage is especially detrimental to the performance of nonvolatile memories. When the charge stored on the floating gate leaks off, the threshold voltage of the cell shifts and the stored information is lost after some time.366,375,379,380

The transient component of the SILC results from employing negatively charged traps immediately after stress. A simple tunneling model predicts a $1/t$ decay of this current.365 In recent publications, however, it has been shown that, apart from the electron component, a hole component is also present in the transient current. Consequently, the electron transient current was found to follow a $1/t^n$ dependence with $n < 1$.381 Substrate hot electron381 and channel hot electron injection382 experiments further confirm the role of the hole component in the transient SILC. The steady state component of the SILC is caused by trap-assisted tunneling from cathode to anode. In thick oxides the probability that an electron can use traps as “stepping stones” for tunneling is very small, resulting in a small steady state SILC component. In thinner oxides, fewer traps are necessary for the electrons to move from cathode to anode, and therefore the steady state

FIG. 54. Occupied electron traps at a filling field of 7 MV/cm, $D_{\text{el}}$, as a function of the injected electron fluence measured during oxide stressing at fields between 6.2 and 11.1 MV/cm. Oxide stressing was performed either by Fowler–Nordheim (open symbols) or substrate hot electron injection (closed symbols). For each stress condition, the mean breakdown value is indicated by an asterisk [from Degraeve et al. (Ref. 345)].

![FIG. 54. Occupied electron traps at a filling field of 7 MV/cm, $D_{\text{el}}$, as a function of the injected electron fluence measured during oxide stressing at fields between 6.2 and 11.1 MV/cm. Oxide stressing was performed either by Fowler–Nordheim (open symbols) or substrate hot electron injection (closed symbols). For each stress condition, the mean breakdown value is indicated by an asterisk [from Degraeve et al. (Ref. 345)].](image)

FIG. 55. Generated neutral electron traps, $D_{\text{el}}$, as a function of the hole fluence for different oxide thicknesses (7–14 nm), stress types (constant voltage and constant current), and oxide fields (9.5–12 MV/cm). A unique relationship is observed [from Degraeve et al. (Ref. 345)].

![FIG. 55. Generated neutral electron traps, $D_{\text{el}}$, as a function of the hole fluence for different oxide thicknesses (7–14 nm), stress types (constant voltage and constant current), and oxide fields (9.5–12 MV/cm). A unique relationship is observed [from Degraeve et al. (Ref. 345)].](image)

FIG. 56. Gate current density as a function of applied voltage for an unstressed device (open symbols) and after various high field stresses have been applied (solid symbols). A gradually increasing leakage current, called the stress induced leakage current, appears at low voltage [from Degraeve et al. (Ref. 3)].

![FIG. 56. Gate current density as a function of applied voltage for an unstressed device (open symbols) and after various high field stresses have been applied (solid symbols). A gradually increasing leakage current, called the stress induced leakage current, appears at low voltage [from Degraeve et al. (Ref. 3)].](image)
component will dominate. Modeling of the SILC trap-assisted tunneling process has been the topic of recent publications. In ultrathin oxides, a new SILC component, arising from electrons that tunnel directly into interface traps, is observed. This leakage current is relatively large in the low voltage range, close to the flatband voltage, as illustrated in Fig. 58. In Fig. 59 the close relationship between SILC and the oxide trap density is illustrated. To gather this data, the FN stress was periodically interrupted to measure the leakage current density \( J_E \) at a fixed low oxide field \( E \), while on the same device the density of neutral oxide traps was determined. Plotted against each other, a stress-independent, one-to-one relationship is revealed between the steady state SILC increase and the generation of neutral traps. Several other authors have emphasized the relationship between the neutral trap density and the SILC, but very recently this relationship has been questioned. Despite these recent observations, many authors consider SILC to be a measure of the neutral trap density. Consequently, SILC, either bulk trap or interface trap related, has been used as a degradation monitor and time-to-breakdown predictor.

6. Discussion of trap generation mechanisms

It is clear from the results shown in the previous sections that trap generation is the key factor determining oxide degradation and breakdown. Here, we compare three trap generation models: the “anode hole injection model,” the “electrochemical model,” and the “H release model.”

In the anode hole injection model, it is assumed that holes tunneling back to the cathode can create electron traps in the oxide, probably in conjunction with an electron in the SiO\(_2\) conduction band. The physics of the trap creation process is still speculative. There have been several studies demonstrating that the interaction of electrons and holes in an oxide results in trap creation. However, the precise role of electrons and holes in the trap creation process and the details of the microscopic mechanism are still uncertain. The most important difficulty in studying this effect is the inability of most techniques to separately control the hole and the electron injection.

Electrons injected in the oxide may lower their energy by light emission. Experimental evidence has been presented that photoexcitation of valence band electrons in the cathode by light generated in the anode, and not anode hole injection, is the dominant source of the measured holes in the substrate. If this is correct, or if there exists a voltage limit to anode hole injection, it could be questioned whether the anode hole injection model will correctly predict low voltage oxide reliability. Recent modified simulations of anode hole injection show the existence of an exponentially decaying impact ionization rate at low voltage, which corresponds to the exponentially decaying trap generation rate measured by SILC.

According to the anode hole injection model, the observation of a unique relation between hole fluence and neutral electron trap generation (Fig. 55) is interpreted as a causal relation, i.e., the holes are necessary to create the traps. This mechanism is outlined in Fig. 60(a). However, other explanations are possible. As is outlined in Fig. 60(b), energy release of the incoming electrons at the anode can, beside hole creation, also activate some other mechanism that is
C. Oxide breakdown

Thus far we have discussed oxide degradation phenomena during electrical stress. In this section we focus on the breakdown event itself. First, the modeling of the breakdown event will be discussed. Then, the occurrence of so-called soft breakdown in ultrathin oxides, and its relationship to device failure, will be addressed.

1. Breakdown modeling

As early as 1990 a “weakest link” breakdown model had been proposed.\textsuperscript{373,409,410} In this model, a capacitor is divided into a large number of small cells. It is assumed that during oxide stressing neutral electron traps are generated randomly over the capacitor area. The number of traps in each cell is counted, and at the moment the number of traps in one cell reaches a critical value, breakdown occurs, by definition. In other words, in that critical cell the number of traps is sufficiently large to create a conductive path from anode to cathode.

A disadvantage of this model\textsuperscript{373} is its two-dimensional nature. Therefore a new “weakest link” model based on percolation theory principles\textsuperscript{411} has been proposed that can accurately describe the intrinsic breakdown distribution. The use of the percolation concept for oxide breakdown modeling has been suggested\textsuperscript{412} and thoroughly elaborated upon.\textsuperscript{53,55,344,345} The percolation model for breakdown exists in two versions: (i) the sphere model, where each generated defect in the oxide is characterized by a sphere with radius $0.9$ nm, and (ii) the cube model, where each defect is represented by a cube whose edge is $1.3$ nm in a three-dimensional frame. Both models are implementations of the same concept and provide, therefore, similar results. As an example, the sphere model\textsuperscript{344} is schematically illustrated in Fig. 61. It is assumed that electron traps are generated inside the oxide at random positions in space. Around these traps a sphere is defined with a fixed radius $r$, the only parameter of the model, Fig. 61(a). If the spheres of two neighboring traps overlap, conduction between these traps becomes possible. Further, the two interfaces are modeled as an infinite set of traps, Fig. 61(b). This mechanism of trap generation and coalescence continues until a conducting path is created from one interface to the other, which is the definition of breakdown in this model, Fig. 61(c). In a computer simulation, the total electron trap density needed to trigger breakdown, $D_{\text{crit}}$, can now be calculated. It is found that the simulated $D_{\text{crit}}$ distribution can be fitted with a Weibull function.\textsuperscript{53}

The percolation model of breakdown is able to quantitatively explain two important experimental observations: (i) as the oxide thickness decreases, the density of oxide traps needed to trigger breakdown decreases,\textsuperscript{36,345,369} and (ii) as the oxide thickness decreases, the Weibull slope, $\beta$, of the breakdown distribution decreases, i.e., a larger spread of the $t_{\text{BD}}$ values is observed.\textsuperscript{55,344,345,413} The latter effect is illustrated in Fig. 62. An important consequence of the decreasing Weibull slope for thinner oxides is the strongly enhanced area dependence of $t_{\text{BD}}$ or $Q_{\text{BD}}$.\textsuperscript{345,413} Indeed, based on the random nature of the breakdown location, it has been shown that for the scale factors $\eta_1$ and $\eta_2$ of two Weibull distribu-
tions (either \( t_{\text{BD}} \) or \( Q_{\text{BD}} \) distributions) of capacitors with identical oxide thickness, but area \( A_1 \) and \( A_2 \), respectively, the following relationship holds: \(^{314,415} \)

\[
\frac{\eta_1}{\eta_2} = \left( \frac{A_2}{A_1} \right)^{(1/\beta)}.
\]

Equation (6) predicts that as \( \beta \) decreases, the area dependence of \( Q_{\text{BD}} \) will be enhanced. This is illustrated in Fig. 63 for films of varying thickness. It can be concluded that for thick oxides (\( t_{\text{ox}} > 10 \) nm), the intrinsic \( Q_{\text{BD}} \) value is, within acceptable approximation, constant in the range of capacitor areas commonly available for experimental purposes. However, for thin oxides \( Q_{\text{BD}} \) can no longer be considered area independent. This has important implications when the \( Q_{\text{BD}} \) value is used as the figure of merit for the oxidation process, a common industrial practice. It is meaningless to specify a \( Q_{\text{BD}} \) value without specifying the area, and it is further necessary to specify the area when \( Q_{\text{BD}} \) values of different oxide thicknesses are compared.

### 2. Soft breakdown

Breakdown in oxides of thickness greater than 5 nm proceeds by the creation of a localized conductive path, followed immediately by the propagation of thermal damage leading to a highly conductive short between anode and cathode. On the other hand, it has been known for several years that ultrathin oxides can exhibit anomalous failure,\(^{374} \) characterized by the creation of a more resistive breakdown path. This phenomenon has been termed soft, quasi, early, or non-destructive breakdown, or B-mode SILC. Only recently has this breakdown mode gained much attention in sub-4 nm oxide layers.\(^{61,416–428} \) Soft breakdown (SBD) can be defined as oxide breakdown without lateral propagation of the breakdown spot due to thermal damage.\(^{418} \) It is generally accepted that soft and hard breakdown originate from the same precursor defects,\(^{427} \) show the same stress current dependence,\(^{430} \) and can be described by the same Weibull statistics,\(^ {429,430} \) although the latter point has recently been called into question.\(^ {431} \)

It has been shown that the \( I_g \) after soft breakdown has a unique relationship with \( V_{\text{gs}} \), i.e., it is independent of area, suggesting that soft breakdown is a localized effect.\(^ {421} \) Figure 64 shows a typical example of an \( I_g - V_{\text{gs}} \) plot, after soft breakdown. A dramatic rise compared to SILC is observed, but the SBD is less destructive as compared to hard breakdown. This conclusion has been confirmed by emission microscopy experiments.\(^ {422} \) However, for very small area devices, a lower, unstable current is measured.\(^ {421} \) This might be explained by the difference in energy available for discharge at the moment of breakdown. Typically, when the applied voltage is plotted as a function of time during CCS, a very small drop in the voltage (or equivalently a small jump of the current during CVS), followed by noisy behavior, signals soft breakdown. However, the small voltage jump observed during FN stress is not a characteristic feature of soft breakdown, although often claimed as such in the literature.
fact, when the test structure area is scaled down, the voltage jump is of the order of volts. At typical large area test structures, the small voltage or current jumps indicating soft breakdown become very difficult to detect with automated measurement systems. Therefore other breakdown detection methods have been proposed. The most effective one is the use of the sudden increase of the gate current noise as the detection monitor. This noise has been studied extensively. At low gate voltage, multilevel random telegraph signals are observed with current amplitudes that depend on the applied gate voltage, shown, for example, in Fig. 65. It has been demonstrated that the I–V characteristics of the on and off states are shifted over a constant voltage interval. This observation can be explained by electron capture-emission-induced local field fluctuations in the breakdown path. With this model, the area of the soft breakdown spot is estimated to be 2 × 10⁻¹³ cm².

The current conduction mechanism through the softened region in the oxide has been modeled in several ways, but the physical picture remains unclear. Models based on variable range hopping of carriers, point contact conduction, energy funnels, resonant tunneling through strategically placed traps, direct tunneling through a thinned oxide region, and electrode controlled conduction have been proposed.

Some researchers have claimed that soft breakdown does not occur in short channel length transistors; instead, hard breakdown is immediately observed. However, soft breakdown has been demonstrated in 50 nm gate length devices and has been shown to depend on applied voltage, not gate length. It has been further claimed that soft breakdown is more pronounced in MOSFETs and can be observed and therefore, for some applications, soft breakdown does not necessarily imply device failure. In determining the relationship between soft breakdown and device failure, the position of the breakdown spot is extremely important. Breakdowns which occur in the drain/gate overlap region have been shown to more profoundly affect transistor function than breakdowns in the middle of the transistor. However, the most recent evidence points to devices being operational after soft breakdown if operated at low voltage and current, such as would be the case in an actual circuit.

3. Breakdown acceleration models

The most important issue for the analysis of tBD data, in order to predict oxide reliability, is the choice of the proper, i.e., field or voltage, extrapolation law. Indeed, oxide reliability has been pinpointed as one of the possible showstoppers in scaling the oxide thickness; however, this prediction was based on tBD values measured at very high fields, whereas actual devices operate at much lower voltages and fields. The correctness of the oxide reliability prediction at actual operating conditions depends completely on the validity of the extrapolation law.

There have been contradictory opinions on the exact field dependence for tBD. Some research groups claim that based on the anode injection model, the logarithm of tBD scales with 1/Eox, whereas others find better results using an Eox dependence. Most recently, a Vg dependence has been proposed for ultrathin oxides where ballistic transport of electrons through the oxide occurs. According to the anode injection model, the field dependence of α, the probability of creating a hole which can tunnel back into the oxide (see Sec. III B 3), can be described as

\[ \alpha = \alpha_0 \exp \left( -\frac{H}{E_{ox}} \right) \]

and

\[ Q_{BD} = Q_0 \exp \left( \frac{H}{E_{ox}} \right) \]

(7)

with \( \alpha_0 \) and H constants (for a fixed oxide thickness) and \( Q_0 = Q_{p,crit}/\alpha_0 \). With this equation, \( t_{BD} \) becomes

\[ t_{BD} = \frac{Q_{BD}}{J_{FN}} = \frac{Q_{BD}}{A \pi \exp \left( \frac{B^*}{E_{ox}} \right)} = \frac{Q_0}{A \pi \exp \left( \frac{B^* + H}{E_{ox}} \right)} = \tau_0 \exp \left( \frac{G}{E_{ox}} \right) \]

(8)
with \( \tau_0 \) a constant. Reported values for \( G = B^+ + H \) vary from 290 to 350 MV/cm, depending on oxide thickness and stress type (constant voltage or current stress). Equation (8) expresses the 1/E model, because the logarithm of \( t_{BD} \) depends linearly on the reciprocal oxide field. The E model, on the other hand, predicts a linear relationship between the logarithm of \( t_{BD} \) and the oxide field:

\[
t_{BD} = t_0 \exp(-\gamma \cdot E_{ox}),
\]

where \( t_0 \) and \( \gamma \) are constants. The E model had been used long before there existed any physical data to support it. In the literature, publications trying to prove the correctness of either the E or the 1/E model can be found, as is illustrated in Fig. 66.

All attempts to provide the E model with a sound physical basis assume that a direct correlation exists between the electric field and oxide degradation. Therefore proponents of this model ignore the role of injected electrons as an intermediate step for generating oxide traps [see Sec. III B 6, Fig. 60(c)]. Recent experiments clearly demonstrate that the oxide degradation process is fluence driven. Furthermore, detailed simulations of anode hole injection, including minority carrier ionization, no longer link a 1/E model directly with the anode hole injection concept. Instead, a mixed model, with approximate 1/E dependence at high voltage and E dependence at low voltage, is used. Other groups have proposed similar unified models.

The E vs 1/E model discussion is mainly valid for oxides thicker than 5 nm, where the injection of electrons is dominated by nonballistic FN injection; the injected electrons enter the conduction band of the SiO₂ and interact with it. The oxide field mainly determines the electron energy at the anode and consequently, the oxide degradation. Since there exists a unique relationship between the FN current density and oxide field, \( Q_{BD} \) should be measured using CCS. For ultrathin oxides, however, the injected electrons travel ballistically through the oxide without interacting with the SiO₂ network. This can be either by FN tunneling above 3.5 V, typically in oxides with thicknesses of between 5 and 3.5 nm, or by direct tunneling below 3.5 V, typically in oxides with thicknesses of below 3.5 nm. The electron energy at the anode is determined by the voltage difference between the cathode and the anode, which corresponds to the applied gate voltage. As a consequence, \( Q_{BD} \) should be measured using CVS. This also means that for an ultrathin oxide the gate voltage determines the breakdown, and CCS methodology needs to be replaced by CVS analysis.

4. Temperature dependence of breakdown

Since advanced CMOS integrated circuits operate somewhat above ambient, the temperature dependence of oxide degradation and breakdown has received considerable attention. The temperature dependence of \( t_{BD} \) in ultrathin oxides is especially strong, as is illustrated in Fig. 67. In most research, it is assumed that an Arrhenius law can describe the temperature dependence of \( t_{BD} \), and many authors have determined the activation energy for their oxides. The activation energies depend, however, on the oxide thickness, the voltage or field range, and the temperature range.

![FIG. 66. Two data sets from the literature, one supporting the E (field) extrapolation model [two top figures, from Suehle et al. (Ref. 444)] and the 1/E model [lower figure, from Schuegraf et al. (Ref. 445)].](image)

of the measurement. From the huge spread of the observed values it can be concluded that the Arrhenius relationship is not suitable as a description for the temperature dependence of \( t_{BD} \). However, both the trap density at breakdown as well as the trap generation rate depend on temperature, although the oxide traps created at different temperatures are not completely equivalent and consequently, oxide damage generated during electrical stress at different temperatures is not simply cumulative.

5. Oxide reliability predictions

Accurate predictions of oxide reliability can only be obtained if a correct and complete reliability specification is defined. This specification must contain four elements: the
allowed failure rate, expected lifetime, the oxide area, and the voltage and temperature conditions. A typical specification reads: 0.01% failures are allowed after 10 years of dc operation, at 1.5 V and 70 °C, on a device of area 0.1 cm². With similar specifications, it has been reported that a reliability limit for oxide thickness scaling is encountered at about 2.2 nm (optical thickness) at room temperature.443 Recent work has shown these predictions to be overly pessimistic. Improved accuracy of the voltage acceleration law, as well as oxide uniformity improvements, have shifted the reliability limit to oxides as thin as 1.5 nm.425

IV. FABRICATION TECHNIQUES FOR ULTRATHIN SILICON OXIDE AND OXYNITRIDES

The gate dielectric’s ultimate electrical performance is determined not only by its composition (SiO₂ vs Si–O–N) and fabrication method (growth, deposition, or implantation), but also by pre-growth surface preparation and postfabrication processing such as plasma etching of the gate stack. The interdependence between the various steps, especially surface preparation, becomes more prominent for ultrathin gate dielectric layers, since the Si/SiO₂ interface is a more significant part of the layer as it gets thinner.

A. Surface preparation

Surface preparation is a more appropriate term than cleaning, since preparation of the Si surface for subsequent oxidation is far more involved than merely removing contamination.461 In fact, engineering of the surface, i.e., conditioning it to result in its smoothest, cleanest, and most unperturbed state, is just as important a step as the actual dielectric fabrication. Silicon surface preparation has been the subject of exhaustive work.461–468 Among the important physical attributes of ultrathin SiO₂ layers and the Si/SiO₂ interface that can be influenced by surface preparation are interfacial roughness,469–471 interfacial transition layer width,180 contamination level of the SiO₂ and the interface,472–474 and chemical bonding structure at the interface.104,124,475 Figure 68 is an example of the effect of various cleans on the surface roughness and reliability of ultrathin oxides.

Wet cleaning currently dominates pre-gate oxidation clean applications.476 The so-called “RCA” chemistry is the most widely used clean for removing organic compounds and metals from Si wafers.461,464 Subsequent processing in HF, which removes the chemical oxide that results from the RCA clean, is optional. Although much research has gone into understanding wet cleaning chemistry, there is still much to learn about the effects of the various chemistries on the electrical properties of the resultant dielectrics. This is a fertile area for continued research.

B. Fabrication of ultrathin oxide and oxynitrides

Fabrication of ultrathin dielectric layers may be accomplished by growth, deposition, or implantation. Growth refers to thermal oxidation or oxynitridation of the Si. Deposition usually refers to chemical or physical generation of the layer, not involving a reaction with the Si substrate. Some processes are combinations of deposition and growth, e.g., low energy implantation of Si with N, followed by thermal oxidation in O₂. Table VI is a compilation of published fabrication techniques for SiO₂ and Si–O–N gate dielectrics.

1. Thermal oxidation and oxynitridation

The utter simplicity of growing thermal SiO₂ by exposing Si to O₂ at elevated temperatures, as well as the perfection of the resulting interface, are in large part responsible for the success of Si as the integrated circuit material of choice. Of course, the quality of the SiO₂ and the Si/SiO₂ interface is in the growth details. Virtually all commercial SiO₂ gate dielectrics are grown by thermal oxidation, using O₂ or H₂O as the oxidant species. Since oxidation in H₂O enhances oxidation kinetics,481 it is not generally used for the growth of ultrathin films. However, some argue that H₂O-oxidized SiO₂ results in dielectrics with lower defect densities and perhaps enhanced reliability.484,485 Thermal oxides consume Si during growth, thereby continuously creating a new and fresh interface. Thermal growth usually takes place at a higher temperature than chemical or physical deposition, and higher fabrication temperature has been associated with improved dielectric properties.486–489 Chemically and physically deposited gate dielectrics usually require post-deposition annealing at temperatures higher than the deposition temperature to attain properties similar to thermal
oxides.\textsuperscript{490,491} Thus attempts to reduce thermal budget by de\-positing dielectrics at temperatures lower than thermal growth temperatures may be misguided.

Silicon oxidizes according to a linear-parabolic (Deal–Grove) kinetic law, well-described in the literature.\textsuperscript{483,492,493} However, this formalism breaks down in the ultrathin film regime, and the details of the first stages of SiO\textsubscript{2} formation are still not well-understood, in spite of much research.\textsuperscript{211,494,495} Oxidation mechanisms will be discussed in detail in Sec. V B.

There are currently two primary thermal techniques for growing SiO\textsubscript{2} or Si–O–N, furnace or rapid thermal oxidation (RTO). Oxidation usually takes place in the temperature range of 750–1100 °C. Typical apparatus for both techniques are shown in Figs. 69 and 70. Although there are many ways to input thermal energy into Si to oxidize it, at present vertical furnace technology is the manufacturing standard for

\begin{table}[h]
\centering
\caption{SiO\textsubscript{2} and Si–O–N gate dielectric fabrication techniques.}
\begin{tabular}{|l|}
\hline
\textbf{Thermal oxidation/oxy\-nitr\-idation} \\
\hline
O\textsubscript{2}, H\textsubscript{2}O, O\textsubscript{3} \\
N\textsubscript{2}O, NO, NH\textsubscript{3}, N\textsubscript{2} \\
\hline
\textbf{Chemical deposition} \\
\hline
Chemical vapor deposition (CVD) \\
Plasma enhanced chemical vapor deposition (PECVD) \\
Jet vapor deposition (JVD) \\
Atomic layer deposition (ALD) \\
\hline
\textbf{Physical deposition} \\
\hline
Low energy ion implantation \\
Remote plasma nitridation \\
\hline
\end{tabular}
\end{table}
Attain high temperatures during operation; therefore, integrated processing is readily accomplished. An RTO process, clustered to a pregate oxidation clean, is a likely scheme for growing ultrathin dielectrics in the near future. As dielectrics get thinner, the advantages of in situ processing should become apparent, since interfacial effects due to contamination and surface roughness will start to dominate electrical characteristics. Such processing has already succeeded in producing high performance 30 nm (minimum feature size) transistors. Figures 72 and 73 are examples of oxidation kinetics attainable in RTO and furnace systems. Thinner oxides, grown at higher temperatures for shorter times, are typical of the RTO process.

Various oxidation gases can be used to grow SiO$_2$, although the common practice is to use O$_2$ or H$_2$O. CO$_2$ will also oxidize Si to form SiO$_2$, with some retardation of oxidation kinetics. Ozone (O$_3$) is sometimes added to the O$_2$ gas stream to enhance growth kinetics, since atomic O, an ozone decomposition by-product, is very reactive. Low temperature growth ($\leq$400 °C) is achievable in O$_3$, and is possible for growing ultrathin Si–O–N with precise N distributions. Oxynitrides, essentially SiO$_2$ containing small (less than 5%–10% locally, and less than 1% if averaged over the film thickness) but significant amounts of N, can readily be grown using N$_2$O or NO. This important body of work will be covered in Sec. VI B.

2. Chemical deposition

Chemical deposition processes are usually used when a lower thermal budget for the dielectric growth step is desired. Since deposition kinetics are slow at such temperatures (typically 350–600 °C), a plasma source is commonly used to activate the reaction. This technique has been used very effectively to deposit ultrathin Si–O–N with precise N distributions. Chemical deposition methods do not consume the substrate, unlike thermal oxidation, and interfacial properties are usually inferior to those of thermal oxides. High temperature anneals ($\geq$750 °C) are usually necessary to bring the electrical performance up to the level of thermal oxides.
SiO$_2$ and Si–O–N layers have been deposited by chemical vapor deposition (CVD), especially for the preparation of “stacked” oxide multilayers. Such CVD/thermal oxide stacks may improve dielectric reliability. However, their application to ultrathin gate dielectrics will be limited due to difficulties in controlling deposited layer thickness uniformity across large (200–300 mm) wafers. Thermal oxidation is inherently an easier process to implement uniformly. Atomic layer deposition (ALD), in which films are grown approximately one monolayer at a time, has been used to grow ultrathin SiO$_2$ layers. This may be an important technique to grow ≤0.5 nm SiO$_2$ layers, useful as buffer layers between Si and high-$\kappa$ gate dielectrics. ALD has the outstanding advantages of superb conformal coverage as well as precise thickness control. Figure 74 illustrates one cycle of this layer-by-layer process. Jet vapor deposition (JVD) has been used to deposit high N-containing Si–O–N. Although these layers have properties similar to Si$_3$N$_4$, the interface, and to a lesser extent the bulk, contain significant amounts of O. Gate dielectrics produced by this room temperature, supersonic gas flow, remote plasma technique are sufficiently intriguing to warrant further research.

Wet-chemically prepared oxides, e.g., the ultrathin 1.0 nm oxides resulting from chemical oxidation of the Si in oxidizing solutions ordinarily used to clean Si or in H$_2$O or ozonated solutions, may be appropriate precursors for ultrathin gate dielectric layers. When annealed at temperatures as low as 350 °C they exhibit FTIR spectra identical to high temperature SiO$_2$.

3. Physical deposition

Physical growth techniques for SiO$_2$ or Si–O–N include ion implantation, usually followed by an anneal or oxidation treatment, and high density plasma oxynitridation or nitridation. Both techniques afford great control over SiO$_2$ or Si–
O–N thickness, composition, and structure. Using plasma nitridation, for example, one can incorporate much more N in an ultrathin layer than by thermal oxynitridation in NO or N2O. Physical growth techniques are distinct from plasma or ion induced chemical deposition, in that in the former no deposition takes place. The dielectric layers are grown through the incorporation and subsequent reaction of the energetic species with the substrate. These techniques may induce damage in the substrate.

Conventional ion implantation has been used to grow ultrathin SiO2,68 Si–O–N,518,519 and Si3N4.520 Because it is desirable to have the implanted species close to the substrate surface to limit damage and facilitate subsequent incorporation during a high temperature step, the implantation should be done at low energies, for example <25 keV. Most conventional implanters do not afford high ion fluxes at these energies, so implantation time may become problematic for production. However, selective ion implantation is a unique scheme for growing dual oxide thicknesses on a single device, as is illustrated for the case of O implantation in Fig. 76. In this case the thicker oxide is created by the reoxidation of an area previously implanted with O. The same effect can be achieved by preimplanting N into selected areas.67 Most of the implanted N is incorporated into the growing Si–O–N during subsequent oxidation; the retarding effect of incorporated N on oxidation kinetics521 is utilized to result in a thinner oxide in the N implanted areas. Although these schemes are very attractive from a device design standpoint, it remains to be seen if the implanted SiO2 and Si–O–N films have the required reliability for ULSI devices.522

Plasma immersion implantation is a high ion flux rate technique that allows rapid incorporation of the implanted species at low energies. Thus ultrathin layers can be grown without the time constraint of conventional implantation. However, there is no mass selectivity using this method, and other species present in the chamber may be extracted from the plasma and implanted as well. For example, since H2O is ubiquitous in many plasma chambers, Si–O–N films usually result from growth in a nominal N plasma.524 Since N ions may be extracted from the plasma at very low energies, i.e., the floating potential of ~10 V, very shallow implantation is possible. Thus the surface of previously grown thermal oxides can be lightly implanted with N525 using, for example, the apparatus depicted in Fig. 77. Unlike incorporation of N during thermal oxynitridation in NO or N2O, where the N is confined to the Si/SiO2 interface, mobility degradation is not observed with the plasma process because the N is at the SiO2 surface.526 In addition, reliability of these oxides is acceptable, in spite of the damage caused by the plasma implantation, if proper annealing treatments are used.527 However, for the case of sub-2.0 nm plasma immersion oxides, both mobility and reliability may be degraded. Further studies of plasma immersion nitridation have been carried out.528–530 Higher temperature favors nitridation, over oxidation caused by residual species such as H2O, and compositions closer to Si3N4 are possible.531,532

C. Postoxidation processing and annealing

Crucial electrical performance parameters such as mobility and interface state density are directly related to physical structure and chemical bonding at the Si/SiO2 interface. This interface does not reach its final configuration after oxidation, but rather after all postoxidation processing has been completed.686 Since the interface is defined by the last SiO2 to form and the last thermal and/or physical (e.g., irradiation) treatment it is exposed to, postoxidation processing, which involves among other steps implant activation annealing, polycrystalline Si deposition, and plasma etching and deposition, greatly impacts the properties of the gate dielectric. Plasma processing in particular has been studied as a major issue in determining the final electrical quality of gate dielectrics.533–535

A fundamental understanding of the “processing window” for postoxidation treatments is absent, making the pos-
D. Hydrogen/deuterium processing

It has long been realized that H plays an important role in MOS device performance.\textsuperscript{21,33,225,543–551} The relatively high background level of H-containing molecules in silicon microelectronics processing ambients (e.g., NH\textsubscript{3} and SiH\textsubscript{4} in CVD reactors), and the high diffusivity of H in SiO\textsubscript{2}, supports the belief that H is ubiquitous in the Si/SiO\textsubscript{2} system.\textsuperscript{544,552–556} In fact, nuclear reaction analysis experiments have revealed that H\textsuperscript{557} or D\textsuperscript{220} concentrations in SiO\textsubscript{2} films can be as high as the 10\textsuperscript{21} cm\textsuperscript{-3} range. Figure 79 shows D depth profiles for a thin (~5 nm) SiO\textsubscript{2} film annealed in D at 450 °C.\textsuperscript{220,221} One can see that D is located in an extended region close to the interface.

Hydrogen, as well as D, in their many bonding and isotopic configurations including atomic/molecular H, Si–H\textsuperscript{+}, OH groups, and bridging H, influence the electrical behavior of the Si/SiO\textsubscript{2} system through various defect processes, as can be seen in Fig. 80.\textsuperscript{64,65,403,453,549} Hydrogen annealing below ~600 °C has been proven to passivate dangling bonds at the interface, which in turn further reduces the concentration of surface states.\textsuperscript{64,65,403,453,549} Hydrogen, for example, can respond to as-grown layers grown in 760 Torr of O\textsubscript{2}, and the open triangles in Fig. 78. Excess scattering strength in the interfacial layer, as measured by x-ray reflectometry, as a function of oxide thickness. The open squares correspond to as-grown layers grown in 760 Torr of O\textsubscript{2}, and the open triangles for layers grown in 40 Torr of O\textsubscript{2}. Annealed layers are indicated by the closed squares. Annealing is seen to remove the high density interfacial layer, since the excess scattering decreases [from Kosowsky et al. (Ref. 184)].

Several postoxidation annealing studies have concerned themselves with interfacial roughness. \textit{In situ} TEM studies of the interface have shown that postoxidation annealing improves roughness for the (100), but not the (111) interface.\textsuperscript{540} TEM, STM, and electrical measurements have been correlated in another study, in which it was found that annealing in Ar improved roughness for low index [e.g., (100) or (110)] interfaces, but degraded higher index interfaces.\textsuperscript{541} Spectroscopic ellipsometry has been used to show that roughness is reduced by reoxidation of suboxides at lower temperatures (<900 °C), and by viscous flow at higher temperatures.\textsuperscript{75} Longer anneals have been shown, through EPR studies, to result in irreversible degradation of the interface due to formation of silicon monoxide, SiO.\textsuperscript{542,543}

![FIG. 78. Excess scattering strength in the interfacial layer, as measured by x-ray reflectometry, as a function of oxide thickness. The open squares correspond to as-grown layers grown in 760 Torr of O\textsubscript{2}, and the open triangles for layers grown in 40 Torr of O\textsubscript{2}. Annealed layers are indicated by the closed squares. Annealing is seen to remove the high density interfacial layer, since the excess scattering decreases [from Kosowsky et al. (Ref. 184)].](image)

![FIG. 79. Integrated amount of D, measured by nuclear reaction analysis, in a thin (~5.5 nm) deuterated SiO\textsubscript{2} film, as a function of vacuum annealing temperature and time [from Baumvol et al. (Ref. 221)].](image)

![FIG. 80. Schematic energy band diagram showing defect generation processes in the Si/SiO\textsubscript{2} system involving H [from Stathis et al. (Ref. 750)].](image)
although this mechanism is not well-understood. More details on the behavior of H in the Si/SiO₂ system can be found in recent review papers. Hydrogen also plays an important role in thin Si₃N₄ and Si–O–N films deposited on Si, as summarized in recent reviews.

Replacing H by its heavier isotope, D, offers advantages both for studying fundamental aspects of H/D behavior in ultrathin SiO₂ films, as well as for improving the electrical properties of the devices. Several groups have successfully used D to mimic basic aspects of H reactivity and diffusivity in SiO₂. The practical benefits of D processing include improved immunity of the D-passivated Si/SiO₂ interface against electrical stressing. In particular, D annealed MOSFETs showed a retardation of the rate of interface trap buildup, under gamma irradiation, by factors of 2.6–4.5 compared to conventional H processing. Lyding and coworkers demonstrated that final annealing of NMOS devices in D, instead of H, resulted in improved hot electron degradation immunity, as is shown in Fig. 82. It was further observed that the hot electron reliability lifetime of transistors was improved by factors of 10–50. It is clear that the difference in mass between H and D is not enough to explain the significant isotopic effects. Electronic effects also need to be taken into account. One can draw an analogy between the electrical findings and results of experiments on H and D desorption from Si surfaces induced by tunneling current from an STM tip, in which a large isotopic effect was observed. It was argued that H and D desorption from the interface under hot electron electrical stress can be caused by multiple excitations of the Si–H or Si–D vibrational modes. The desorption yield is governed by competing processes: the excitation rate, which in turn depends on current density and electron energy, and the excitation decay via substrate phonons. It was postulated and then deduced from first principles calculations that the coupling between the substrate phonon modes and Si–H and Si–D vibrations is much stronger for the Si–D case. This enables more efficient energy dissipation of the excited Si–D bond, whereas in the Si–H case it is easier to pump the energy up until the bond breaks. As a
result, the D at the interface exhibits much longer lifetime under electron stressing, as was seen in Fig. 83. One model of hot electron degradation, alluded to above, invokes H desorption during excitation. Further elucidation of this point requires an understanding not only of all relevant bulk and H-defect vibrational frequencies, but also direct measurements of the vibrational lifetimes of the Si–H and Si–D modes themselves. Ultrafast optical experiments offer direct evidence for the lifetimes, thereby significantly increasing our understanding of H/D effects in defective solids. Recent bulk studies, combined with previous work on surfaces, is leading to a much deeper understanding not only of vibrational lifetimes, but more importantly of MOSFET transconductance degradation.

The isotopic effect is very sensitive to processing history, especially to the way the D is incorporated into the gate dielectric, and subsequent thermal processing. The magnitude of the effect seems to be correlated with the amount of D left after all processing steps. Deuterium anneals in the 400–600 °C range are usually used to dose the gate dielectric and the interface with D, as can be seen from Fig. 79. The amount of incorporated D is strongly dependent on the materials structure of the device. Deuterium diffuses easily through SiO₂ layers, whereas Si₃N₄ layers significantly retard D penetration. It is also observed that impurities, for example dopants in polycrystalline Si gate layers, may affect the diffusivity, since undoped polycrystalline Si appears to be an efficient barrier for D transport. Ference and coworkers employed a deuterated PECVD Si₃N₄ capping layer for the MOSFETs, both as a huge reservoir of D close to the gate, as well as an efficient diffusion barrier to prevent its escape under subsequent thermal treatment, and observed a significant effect on hot electron immunity. As far as the thermal anneals are concerned, one can see from Fig. 79 that D in SiO₂, like H, is not very stable and tends to escape at temperatures higher than approximately 550–600 °C.

V. THE Si/SiO₂ SYSTEM

A. The initial stages of oxygen interaction with silicon surfaces

The interaction of O with the Si surface is known to depend on temperature, time, and O₂ pressure. Lander and Morrison, and subsequently others, observed two distinct regions in pressure-temperature phase space, Fig. 86, that represent completely different O–Si interactions. In the low temperature-high pressure half of the diagram, O interaction with the surface results in oxide growth:

$$\text{Si}_{(s)} + \text{O}_2(g) = \text{SiO}_2(s).$$

In the high temperature-low pressure regime, on the other hand, surface etching via volatile SiO formation, also known as disproportionation, occurs:

$$2\text{Si}_{(s)} + \text{O}_2(g) = 2\text{SiO}(g).$$

These oxide growth and etching modes are often referred to as “passive” and “active” oxidation, respectively. Recently, several groups have shown that there is a transition regime in pressure-temperature space between the active and passive oxidation modes (also shown in Fig. 86), where the surface morphology differs dramatically from that of either region. In particular, it has been shown that the surface becomes very rough in this region, as is shown in Fig. 87, whereas under normal active and passive oxidation the surface may remain flat. Below, we summarize the details of the interaction of O with Si surfaces. More details on O adsorption and the initial stages of Si oxidation may be found elsewhere.
1. The passive oxidation regime

Basic questions about the first few monolayers of oxide growth, such as whether the initial oxidation proceeds by a layer-by-layer or three-dimensional island growth mechanism, have been intensely debated over the past decade.\[^{26,147,268}\] Research tools brought to bear on this problem include conventional XPS,\[^{146,147,156,583}\] synchrotron-based XPS,\[^{130,159,593}\] HRTEM,\[^{261,262}\] LEEM,\[^{272}\] and MEIS.\[^{199,204}\] Further, STM results for submonolayer O coverages have been reported.\[^{284,285,289,290,300,301,587,594–596}\] Some STM images show nucleation and growth of oxide islands on Si\[^{111}\] and Si\[^{100}\] surfaces, as is illustrated in Fig. 89. The islands are thought to be one monolayer high. However, since STM sees a local density of states at the surface rather than a real geometrical map of surface atoms, changes in the local electronic configuration of the Si and O atoms must be considered when interpreting STM images. For example, in one report the difference in grayscale level in the images, Fig. 89, was argued to represent different height variations ~0.2 nm! in oxide islands grown at 873 K.\[^{290}\] However, larger contrast variations, observed for a higher temperature oxide, could easily be due to a change in the local electronic configuration, e.g., local stoichiometry, rather than a simple height variation.

Island growth was also suggested from both core level and valence band photoemission experiments.\[^{149,159,597}\] These studies show nonuniform oxidation behavior at submonolayer coverages, with phase separation of oxidized and nonoxidized areas on the surface, as seen in Fig. 89. Oxidation kinetics at high temperatures was also described within a nucleation and growth model.\[^{147}\] and island growth of the surface oxide has been predicted from first principles calculations.\[^{598,599}\] Some researchers suggested that oxide island growth at elevated temperatures proceeds via surface diffusion of O to the growing oxide islands.\[^{147,290}\] However, relatively little is known about O diffusivity on Si surfaces.\[^{278}\] STM studies have suggested that O diffusion actually occurs as SiO bound to the surface, which is chemically reasonable given the known stability of SiO. We also note that STM images taken during the interaction of O with Si at room temperature show a more uniform O overlayer in comparison to those taken after elevated temperature anneals,\[^{285,287,290}\] offering further evidence for the thermodynamic stability of SiO\(_x\) islands. Photoemission, scanning tunneling spectroscopy, and EELS experiments are consistent with the interpretation that the stable bonding configuration for O at submonolayer coverages is the bridging position between two Si atoms.\[^{122,129,159,161,285,289,600–602}\] Some alternative metastable structural models have been suggested for surface bound O, notably singly bound O with a dangling
bond, peroxy Si–O–O–Si species, and singly bound O$_2$ species.$^{104,272,312,603,604}$

Oxidation behavior, after the first monolayer, has been studied by XPS, AFM, TEM, SREM, and MEIS. Stepwise evolution of the net concentration of Si$^{n+}$ ($n = 1,2,3,4$) oxidation states, as measured by Si 2$p$ XPS during the oxidation of Si$(100)$,$^{159}$ Fig. 90, and antiphase oscillations of Si$^{1+}$ and Si$^{3+}$ intensities as a function of oxide thickness on Si$(111)$,$^{156}$ Fig. 91, were cited as support for layer-by-layer oxidation. Although the term "layer-by-layer growth" has become popular in the literature to describe the initial oxidation of Si, one should keep in mind that, thermally grown SiO$_2$ on Si surfaces is amorphous and, therefore, not a layered system, and (ii) traditionally this term refers to thermodynamically stable structural behavior during film growth (i.e., the Frank/van de Merwe growth mode), whereas the initial oxidation of Si is likely kinetically limited, at least under some conditions. In contrast, XPS results$^{130}$ were interpreted as implying discontinuous, island-like film growth for oxide thicknesses less than 0.8 nm.

When interpreting XPS results, one should keep in mind that experimental data are subject to different interpretations and analyses. An analysis of XPS (Al $K\alpha$, 1487 eV) and synchrotron XPS (130 eV) data taken on the same Si$(100)$/SiO$_2$ sample yielded differences in the concentration of the Si$^{n+}$ states by as much as a factor of 2, and a difference of about 30% in the oxide thickness.$^{134}$ This discrepancy was attributed to differences in data analysis caused by uncertainties in cross sections, photoelectron mean free paths, and takeoff angles, rather than fundamental differences between the two techniques. Other XPS$^{146}$ and synchrotron XPS$^{593}$ experiments performed on samples oxidized under similar conditions also yield different results. In addition, the traditional interpretation of the Si$^{n+}$ suboxide states observed in photoemission, Fig. 18, as a Si atom with $n (n = 1,2,3)$ O bonds in the first coordination sphere,$^{122,124,125,130,161}$ is currently under debate.$^{133,136,165,167,605}$ An alternate interpretation derives from experiments illustrated in Fig. 19. It has been argued that the local electronic configurations, and corresponding Si 2$p$ chemical shifts, around each Si atom depend not only on the nearest neighbor configuration but also on the O–Si bonding in the second coordination sphere.$^{133,136,166,167}$ Two different first principles electronic structure calculations support the traditional interpretation of the Si 2$p$ photoemission spectra, but differ as far as the second nearest neighbor effects are concerned.$^{163–165,605}$

Hattoni and colleagues$^{144}$ used in situ noncontact mode AFM to examine the evolution of SiO$_2$ surface microroughness during the initial oxidation of Si$(100)$ and Si$(111)$. It was observed that the surface roughness varies periodically with the progress of oxidation, Fig. 92. The change of the microroughness was found to correlate with periodic oscillations of the XPS signal, illustrated in Fig. 91, of the suboxide states, in particular Si$^{1+}$. This observation was interpreted in

---

**FIG. 88.** Schematic representation of the three different mechanisms active during the interaction of O with Si surfaces at high temperatures: (a) active oxidation (SiO desorption), (b) roughening regime, and (c) passive oxidation (SiO$_2$ growth) [from Feltz et al. (Ref. 301)].

**FIG. 89.** Series of scanning tunneling microscopy images (36×36 nm$^2$) during Si$(111)$ passive oxidation (850 K, 1×10$^{-4}$ Pa) as a function of time: (a) clean Si$(111)$ 7×7 surface to (1) 1205 second exposure. Inhomogeneous oxide nucleation at steps and homogeneous oxide nucleation on terraces can be seen [from Feltz et al. (Ref. 301)].
terms of the formation of Si$^{1+}$ at the interface which, in turn, causes the formation of protrusions at the oxide surface, Fig.
93, further supporting the layer-by-layer model of growth.\textsuperscript{144}

Surface composition and morphology on Si$^{111}$ have been studied by MEIS for passive oxidation conditions.\textsuperscript{591} It was shown, Fig. 94, that the width of the O backscattering peak, as a function of O coverage at elevated temperatures, has a plateau for coverage less than 3.2±0.4 monolayers. This result is consistent with a model in which the surface is covered by a nonstoichiometric oxide before “bulk” oxidation begins. As oxidation proceeds the O peak broadens, reflecting oxide growth in the subsurface layers.\textsuperscript{591} Surface oxide formation at elevated temperatures was compared with the interaction of O with Si at 300 K. These results imply that the difference in oxidation between elevated and room temperatures is due to a difference in local stoichiometry.

One can distinguish between two different kinds of layer-by-layer oxidation. The first refers to a single atomic layer being oxidized at a time, which can be called monolayer growth. A second case refers to a bilayer growth mode, with the unit being a Si bilayer. Ross and coworkers\textsuperscript{266,268} showed, via real-time HRTEM experiments, that surface

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig90}
\caption{(Left) Peak heights of the Si$^{1+}$, Si$^{2+}$, and Si$^{3+}$ oxidation states [measured by synchrotron based photoelectron spectroscopy during the initial oxidation of Si(111) at different temperatures] as a function of O coverage, measured by Auger spectroscopy. The right panel shows possible bonding configurations and corresponding peak intensities. One can see that at room temperature all Si atoms are oxidized at more or less the same rate, whereas at high temperatures some Si atoms are oxidized faster. The distribution of adsorbed O is therefore highly nonuniform [from Tabe et al. (Ref. 159)].}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig91}
\caption{(Left) Si$^{2p}$ photoelectron spectra showing the evolution of the Si$^{1+}$, Si$^{2+}$, and Si$^{3+}$ suboxide states during the initial oxidation of Si(111). The right panel summarizes the density of the suboxide states as a function of oxide thickness [from Oishi et al. (Ref. 156)].}
\end{figure}
steps are immobile during the initial oxidation of Si(111), suggesting that the oxidation follows a monolayer growth mode, depicted in Fig. 38. Photoemission results, Fig. 91, also imply a monolayer growth mode.\textsuperscript{144}

The kinetics of the initial oxidation of the Si(001) surface was recently explored by SREM, in combination with Auger measurements of O and Si signals.\textsuperscript{271} Very similar to the pioneering microscopy work by Ross and Gibson,\textsuperscript{267,268} a periodic reversal of the SREM contrast and immobility of the surface steps was observed, as can be seen in Fig. 39. Based on these observations, the authors concluded that the most likely layer-by-layer oxidation scenario is random nucleation of nanometer-scale oxide islands, followed by their subsequent lateral growth.

Shklyaev \textit{et al.}\textsuperscript{588,589,606} used optical methods, including ellipsometry and second harmonic generation, to study the very early stages of oxide formation as a function of pressure, temperature, and ambient (O\textsubscript{2} vs N\textsubscript{2}O). A decrease in the rate of oxide formation was observed with decreasing oxidant pressure as the transition to the Si etching regime is approached, shown in Fig. 95. These results were interpreted with an oxide island nucleation and growth model, according to which the nucleation proceeds through the interaction of the layer of intermediately adsorbed species, with the critical size of the islands decreasing with O pressure.

Infrared absorption spectroscopy and density functional cluster calculations were recently used to study the initial oxidation of Si using the H\textsubscript{2}O-exposed Si(100)-(2\times1) surface as a model.\textsuperscript{104,607} The Si−Si dimer bond was found to dissociate during initial oxidation, followed by O incorporation into the Si backbonds. It was also found that high temperature annealing of this surface results in the formation of Si-epoxide intermediate states, consisting of three-membered Si−O−Si rings.\textsuperscript{104}

Finally, we should mention the change of the electronic band gap of ultrathin SiO\textsubscript{2} layers on Si as a function of oxidation, as observed by valence band spectroscopy.\textsuperscript{143,144} One can see that while for thicker films the band offset does not change with thickness, Fig. 96, for thinner films the top of the SiO\textsubscript{2} valence band decreases by approximately 0.2 eV.
as the films become thicker than $\sim 0.9$ nm, Fig. 97. This result is consistent with local density cluster calculations of ultrathin SiO$_2$ layers, in which it was found that the band gap decreases for films thinner than 0.7 nm. Muller and coworkers used STEM to study the physical and electronic structure of ultrathin SiO$_2$ layers and recently followed this up with \textit{ab initio} electronic structure calculations. Muller’s results, Fig. 6, have been interpreted to show that more than a monolayer of Si at both interfaces of a Si/SiO$_2$/Si sandwich are in the suboxide configuration, and that this suboxide has wave function tails that decay slowly into the SiO$_2$. The result of this analysis places limitations on how thin such a structure can be and still limit charge transport to acceptable levels in devices. Although the concept of a scaling limit for SiO$_2$ films, in the range 1 to 2 nm, is agreed upon by most workers in the field, the authors’ provocative interpretation of their STEM data has yet to be substantiated by other methods.

2. The active oxidation regime

Walkup and Raider used laser induced fluorescence to study SiO(g) evolution during O interaction with Si(111). SiO molecules were detected in the active oxidation regime.

On the other hand, they noted that in the passive oxidation regime, the SiO density in the gas phase was below the detection limit, $3 \times 10^6$ molecules/cm$^3$. SiO desorption was found to be thermally activated and linearly dependent on the O flux (pressure). SiO evaporation under active oxidation conditions was also observed using mass spectrometry. Starodub and coworkers developed a sensitive XPS method for detecting volatile SiO species, based on its condensation on an Ag foil. As an example, Fig. 98 shows how SiO desorption increases with temperature (the amount of Si condensed on the Ag foil is proportional to the area under the Si 2s XPS peak). They also found that small but measurable yields of SiO(g) desorb from the oxide film during the initial stages of passive oxidation, even when the oxide film continuously covers the surface.

SiO desorption implies that Si atoms leave the surface, i.e., that surface etching is occurring. Surface etching has also been observed as a result of Si interaction with other reactive gases, for example, I$^{611}$ and Cl, at high temperatures. Surface morphology in this oxidation regime has been studied by conventional electron microscopy in the
and by STM. It has been shown that etching in the “pure” active oxidation regime at high temperatures proceeds via a step-flow mechanism for both Si(111) and Si(100), Fig. 38, and Si(100), Fig. 39. Both the step velocity and the “vertical” etch rate increase with O2 pressure and appear to be independent of the substrate temperature. The origin of the step movement across the surface has been explained in two different ways. Ross et al. suggested that the step flow is caused by preferential evaporation from step edges with O diffusion to the step. However, Felz et al. using AES, did not detect O on the surface and argued, therefore, against this mechanism. They and other workers favor a model where the step flow is caused by the formation of surface vacancies, taking place uniformly over the surface terraces, with their subsequent diffusion to the steps.

Vacancy diffusion to, and disappearance at steps gives rise to step movement across the surface. The vacancy diffusivity has been shown to be much higher than the surface mobility of O. For the Si(100)-(2×1) surface, the diffusing species is believed to be a divacancy, preferentially diffusing along the dimer rows. The surface vacancy diffusion length and terrace width are important factors that affect surface morphology during active oxidation. If the diffusion length is greater than the terrace width, only step etching occurs. At the opposite extreme, surface vacancies can form an island or a two-dimensional vacancy hole, as observed experimentally by REM, STM, and LEEM. Vacancy island formation processes during active oxidation are similar to nucleation and growth of surface holes during Si sublimation at high temperatures and during low energy ion bombardment at high temperatures.

The fact that SiO desorbs under high temperature-low pressure conditions can be used for in situ precleaning of Si wafers covered with ultrathin native, chemical, or thermal oxide films. To accomplish this, one must be very careful about using the lowest possible pressure during the high temperature anneals, as interfacial SiC formation and/or surface roughening may otherwise occur. An interesting modification of this approach was demonstrated recently by Wilk et al. They showed that the temperature of SiO desorption can be lowered by supplying a flux of Si atoms during the in situ preclean. Their process also resulted in a smoother surface, as can be seen from the STM images in Fig. 99.

3. The transition regime

For both the active and passive oxidation regions the surface remains relatively smooth. However, this is not the case if oxidation is performed in the transition regime, or during active oxidation at low temperatures, as was shown in Fig. 87. SEM experiments showed microscopic growth features, interpreted as Si islands as large as microns in size, covered by a thin oxide layer. Ross and coworkers used the term “roughening regime” to describe this unusual oxidation behavior, ascribed to surface roughening with a characteristic lateral scale of about 5.0 nm. Significant surface roughening in the transition regime was independently found by STM for Si(111) and Si(100). Fig. 100. Experiments by Sutherland et al. showed the existence of a roughening regime during the initial oxidation of Si(111) with N2O, but at much higher pressures than are observed for O2 roughening.

The origin of surface roughening is not fully understood, and the shape of the transition regime in pressure-temperature space is only partially documented. Ross and coworkers found a relatively narrow transition regime, whereas recent STM experiments suggest that this regime covers several orders of magnitude in pressure.
Roughening is thought to be caused by competition between the two simultaneous processes of oxide growth, and surface etching from the areas not covered by oxide islands.\textsuperscript{268,301,591} However, neither HRTEM nor STM can probe the surface O directly, especially when the surface is very rough. Feltz and coworkers\textsuperscript{301} used AES to complement their STM results and showed that O is present on the surface. This observation was later supported by MEIS experiments,\textsuperscript{591} in which both the Si and O backscattering peaks were measured, Fig. 101. MEIS provides clear evidence for the presence of surface O in the transition region of pressure-temperature space, and both Si and O peaks demonstrate surface roughening. From the MEIS experiments the vertical roughness was estimated to be as large as 2 to 3 nm.

These observations suggest that once the oxide islands form, they passivate the Si under the oxide and hinder etching of the steps near them, although the etching reaction still takes place in between the oxide islands, as shown in Fig. 87. This process results in the formation of three-dimensional conical islands with heights as large as 4.0 nm,\textsuperscript{281} shown in Fig. 100. The O atoms, in the form of a surface oxide, are assumed to at least partially cover these islands, whereas areas in between the islands are assumed to be free of O.

Oxide island growth and etching in the transition regime is kinetically limited. Thus time plays an important role in the roughening behavior. Therefore surface morphology should be described in three-dimensional pressure-temperature-time phase space, rather than by the conventional pressure-temperature phase diagram.\textsuperscript{620–622} Seiple and Peltz showed that the roughness is not simply a function of O exposure, but depends on O\textsubscript{2} pressure or, equivalently, time, at a constant exposure.\textsuperscript{289,587} It should also be mentioned that observed discrepancies in growth modes in the passive oxidation regime may be related to how close the oxidation parameters in such studies were to the transition regime.

B. Growth mechanisms of ultrathin oxides, beyond the Deal–Grove model

While the seminal paper by Deal and Grove\textsuperscript{483} brought our understanding of the mechanism of thermal oxidation of Si to a new level of sophistication, much still remains under debate, especially in the limit of thin (sub-10 nm) films.\textsuperscript{28,29,31,199,585,623} The Deal–Grove (DG) model\textsuperscript{483} treats Si oxidation as the reaction of Si and O at the planar Si/SiO\textsubscript{2} interface, accomplished by O\textsubscript{2} transport to the interface of the growing film. The model is schematically depicted in Fig. 102. For thick films, the model predicts a parabolic dependence of oxidation time on oxide thickness because the growth is limited by diffusion. Further, in the limit of thin

![FIG. 100. Top and three-dimensional views of 300×300 nm\textsuperscript{2} scanning tunneling microscopy images after Si(100) exposure to O\textsubscript{2} in the roughening regime (600 °C, 6×10\textsuperscript{-3} Torr) as a function of exposure: (a) 100 L, (b) 200 L, (c) 400 L, and (d) 800 L. 1 L = 1 langmuir = 10\textsuperscript{-6} Torr s [from Seiple et al. (Ref. 281)].](http://ojps.aip.org/japo/japcr.jsp)
films, DG shows that the reaction rate at the interface governs the growth and results in a linear relationship between the thickness and the oxidation time for a given pressure and temperature, although the microscopic mechanism of the reaction which determines the rate constant is still unclear.

The original assumption of the DG model was that the reaction was first order between O and Si species. This approximation is not entirely appropriate in the solid state where the reaction takes place at the dense, lattice mismatched, and defect rich Si/SiO₂ interface. Another assumption of the model that becomes questionable in the thin film regime is the steady state assumption, in which all fluxes are assumed constant, and defined by the incoming flux of O₂ from the gas phase. Recently, it has been demonstrated that without the steady state assumption it becomes possible to describe Si oxidation kinetic data very adequately, including the very initial regime, as is seen in Fig. 103.

The DG model yields oxide thickness as a function of the processing parameters temperature, oxidant pressure, and time. This kinetic approach works well for the parabolic growth stage. In the “reaction-limited” region, it is not clear that the linear dependence can be applied to the zero-thickness limit. Ellipsometric, Fig. 104, as well as other studies, showed that there is a region of “faster” initial growth. These observations were the basis of several phenomenological models based on the DG formalism. A review of these models can be found elsewhere. In one model, for example, two exponential terms were added to account for the faster growth. In another model, reactive sites in the near-interfacial oxide, whose concentration exponentially decays with time, were introduced to account for the faster initial growth.

The breakdown of the DG model in the important ultra-thin regime motivated intense experimental and theoretical work aimed at elucidating diffusion and reactions on the atomic scale. Much recent progress in the experimental arena came from the use of powerful isotopic labeling methods, using both O and Si, starting from pioneering work by Rigo and his group. A comprehensive review of isotopic methods applied to Si systems has recently been published. Many studies involved the use of sequential oxidation of Si in ¹⁶O₂ and ¹⁸O₂, or in isotopically labeled H₂O, with subsequent analysis of the isotopes and their depth profiles by ellipsometry. A fast initial oxidation regime is observed. In dry O₂, measured by ellipsometry. A fast initial oxidation regime is observed [from Massoud et al. (Ref. 625)].
reaction at a sharp interface, and (ii) more studies are needed to understand the exact structure and composition of the transition region. One needs to know how the transition region is affected by the oxidation process, as well as how the properties of the transition region are correlated with electrical defects, many of which are known to be localized at the interface and near-interface regions.

Several models include the near-interfacial region as an element of the oxidation process. For example, Tiller and subsequently others suggested that Si interstitials formed at the reaction interface diffuse, or are injected into the near-interfacial oxide, where they may subsequently react with the O diffusing towards the interface. Si injection was directly observed by STM during the very initial stages of O interaction with Si, as is shown in Fig. 107. A similar scenario, but with SiO formation at the interface and diffusion towards the surface, has been suggested by others. Some researchers postulate Si fragments, or clusters, in the oxide and their subsequent reoxidation. Another reactive layer model further considers a thin reactive SiO layer near the interface. However, according to this model, the reaction takes place on top of, and not throughout, the reactive layer, inconsistent with MEIS results. One can propose a more general model in which the near-interfacial reaction is caused by O interaction with incompletely oxidized Si, i.e., suboxides, Si interstitials, clusters, and/or SiO. Again, we stress that the near-interface reactions become most important for ultrathin films. For thicker oxides, the near-interfacial reactions still probably occur; however, since the reaction region in this case is much thinner than the thick-

![FIG. 105. Oxidation of an isotopically enriched $^{29}$Si epitaxial layer Si surface. (a) $^{29}$Si nuclear reaction analysis excitation curve (solid circles) and simulated results (solid line) for an oxidized sample, with the $^{29}$Si profile shown in the inset, and (b) $^{29}$Si excitation curves of the initial $^{29}$Si episolp-sample (solid circles) and after oxidation in O$_2$ (empty circles, same data as solid circles in (a)). The arrows show the energy position of surface $^{29}$Si atoms. One can see that the $^{29}$Si stays at the surface after oxidation [from Baumvol et al. (Ref. 223)].](image)

![FIG. 106. Schematic model of the different regions where oxidation reactions occur: interface reaction (traditional Deal–Grove), near-interface reaction, and surface exchange reaction. In thicker (>3 nm) films, the regions are spatially separated, but in thinner films (<2 nm) they overlap [from Gusev et al. (Ref. 199)].](image)
The surface exchange reaction is perhaps even more complicated than the interface growth reaction. The O$_2$ exposure produced protrusions, seen as white spots. One monolayer of Si(100) is 0.14 nm high [from Cahill et al. (Ref. 647)].

The mechanisms of O uptake, reaction, and loss from the growing oxide film are not well understood. One group claimed that the reaction was related to an exchange of atomic O with the SiO$_2$ network. This is consistent with the provocative suggestion that O, and not O$_2$, is the primary diffusing species in SiO$_2$. Others tried to correlate the O exchange with H$_2$O traces present during oxidation. It has also been claimed that surface defects, e.g., peroxo-bridged O, may be involved in the surface reaction. Despite the fact that the surface exchange reaction does not directly contribute to the growth of the SiO$_2$ layer, it may be important for sub-2.0 nm films, where both surface exchange and near interface reactions overlap in space, Fig. 106, and may therefore influence each other. A recent study shows that an O exchange mechanism is also operative at the Si/SiO$_2$ interface.

Pasquarello et al. performed first principles molecular dynamics calculations, in which electronic structure evolves self-consistently with atomic motions, to explore structural rearrangements at the interface between crystalline Si and disordered SiO$_2$ during high temperature oxidation. The calculations revealed interesting features. The upper layer of the Si substrate contained about a quarter of a monolayer of extra Si atoms, and they formed intralayer Si–Si bonds. It was also found that changes in the bonding network near the interface occur via transient exchange events in which O atoms are momentarily bonded to three Si atoms. This metastable configuration enables the interface to evolve without leaving dangling bonds. Based on these calculations, the following mechanism for the near-interfacial reaction was proposed. Incoming O atoms react with Si–Si bonds in a reaction layer approximately 0.5–1.0 nm thick, near the interface. At high reaction temperatures, network rearrangement, via the formation of metastable threefold coordinated O atoms within the reaction layer, leads to a constant exchange motion of O atoms through to the interface. This process disturbs the Si lattice, which in turn promotes the incorporation of O into the disturbed network, resulting in the formation of new interfacial oxide. This mechanism drives some excess Si atoms into the Si and SiO$_2$ layers adjacent to the interface. The new Si–Si bond “injected” into the reaction layer reacts again with incoming O as the oxidation reaction proceeds. This scheme, based on network motion in the reaction layer, is consistent with isotopic labeling MEIS experiments, such as are illustrated in Fig. 108, which show evidence for motion of O atoms near the interface and provide further evidence and a visualization of the excess Si near the interface.

The formation of the threefold coordinated O atoms was also deduced from density functional theoretical calculations. Using first principles, total energy calculations it was shown that the key process in the atomic scale reactions during the formation of thin thermal SiO$_2$ films on Si, as well as SiO$_2$ precipitation in O-rich Si crystals, is the emission of Si interstitials. This process helps to eliminate dangling bonds and therefore explains the low concentration of defects at the Si/SiO$_2$ interface.

The structural and electronic properties of the (100) Si/SiO$_2$ interface and O diffusion in amorphous SiO$_2$ were studied using first principle calculations by Ng and Vanderbilt. The calculations showed an O deficient, i.e., Si suboxide, transition region at the interface, Fig. 109, of thickness ~2.0 nm. Calculations of the oxidation kinetics in SiO$_2$ suggested that the peroxy linkage configuration may be responsible for O diffusion in the growing silica film, as is suggested by Fig. 110. This mechanism is similar to Ha-
mann’s peroxy-linkage diffusion model637 deduced from density functional calculations.

Unfortunately, still unresolved is a fundamental and quantitative understanding of (i) the O interaction with Si either at the clean surface or the Si/SiO₂ interface, (ii) what fraction of the oxidation takes place at the “formal” interface as compared to the near interfacial region, and (iii) what is the steady state concentration of Si–Si bonds, describable, for example, as suboxides and interstitials, as a function of distance from the interface. This level of information is needed if one attempts a more accurate, atomic scale revision of the DG model.

VI. THE Si/SiO₂Nₓ SYSTEM

Ultrathin Si–O–N layers, with essentially the same dielectric constant as SiO₂ due to their low N concentration, may replace SiO₂ for thicknesses ≲ 2 nm, since the N offers process latitude and protection against B and other impurity penetration through the gate dielectric. Oxynitrides may also enhance reliability and reduce hot electron induced degradation. Several excellent reviews of Si–O–N layers have been written.2,37,38

A. Oxynitride properties

1. Thermodynamics of the Si–O–N system

The bulk phase diagram of the Si–O–N system653 is shown in Fig. 111. The diagram consists of four phases: Si, SiO₂, Si₃N₄, and Si₂N₂O. The three compound phases are composed of similar structures comprising SiO₄, SiN₄, and SiN₃O tetrahedrals, implying that SiO₂ can be converted to Si₂N₂O and finally to Si₃N₄ by replacing O with N in a 2N/3O ratio. However, the Si₃N₄ and SiO₂ phases never coexist in the bulk under equilibrium conditions. Si₂N₂O is the only thermodynamically stable form of Si–O–N.653–657

At chemical equilibrium, N should not incorporate into an SiO₂ film as long as almost any partial pressure of O, i.e., > 10⁻²⁰ atm, is present, as is implied in Fig. 111. A puzzling question is why N atoms are incorporated at all in SiO₂ as a result of the thermal decomposition of N₂O173,226,243,658–660 or NO158,169,201,203,661,662 since the partial pressures of oxidizing species, which are decomposition by-products, are very large. At least two reasons for the presence of N in the SiO₂ film can be suggested. First, N atoms may simply be kinetically trapped at the reaction zone near the interface, and thus the N is present in a nonequilibrium state. In this model it is assumed that the N is incorporated into the film during oxynitridation and reacts only with Si–Si bonds at or near the interface, not with Si–O bonds in the bulk of the SiO₂ overlayer. Alternatively, the N at the interface may indeed be thermodynamically stable, due to the presence of free energy terms not represented in the bulk phase diagram. For example, N may lower the interfacial strain, known to exist at the Si/SiO₂ interface226,663–665.
Even when N is implanted into Si it tends to incorporate into SiO$_2$ and migrate to the Si/SiO$_2$ interface after subsequent oxidation. Therefore there is some evidence to suspect that the N plays a specific role at the interface, but also that it is not stable away from it, in the SiO$_2$ bulk.

2. Physical properties

Figure 112 shows the change in dielectric constant and band gap of Si–O–N films as a function of N content. In this case, the dielectric constant ($\kappa$) was assumed to increase linearly with increasing fraction of Si$_3$N$_4$, from $\kappa$(SiO$_2$) $= 3.9$ to $\kappa$(Si$_3$N$_4$) $= 7.5$. A very recent study shows a departure from linearity; $\kappa$ is found to be higher than the predicted values.

Due to the higher $\kappa$ of Si$_3$N$_4$, an Si–O–N film having the same capacitance as an SiO$_2$ film will be physically thicker. However, at the same time the band gap, and therefore the barrier height for electron and hole tunneling, decreases with increasing N content. In addition, there is a possible change of the effective mass of the carriers. All of these factors must be taken into account in understanding the tunneling properties of ultrathin Si–O–N films. Finally, the refractive index ($\eta$) also increases with the amount of incorporated N from $\eta$(SiO$_2$) $= 1.46$ to $\eta$(Si$_3$N$_4$) $= 2.0$, as is shown in Fig. 12.

3. Diffusion barrier properties of oxynitride layers

Boron penetration through the gate dielectric causes threshold voltage shifts and degrades reliability. An important property of Si–O–N layers is their ability to act as a diffusion barrier to impurities, notably B. Oxygen, N$_2$O, and NO diffusion rates are also lowered in Si–O–N films, significantly retarding the rate of further oxidation or nitridation, as can be seen, for example, in Fig. 113. One explanation for the enhanced diffusion resistance is that the density of Si$_3$N$_4$ and Si–O–N is higher than SiO$_2$ ($\rho$(SiO$_2$) $= 2.3$ g/cm$^3$ and $\rho$(Si$_3$N$_4$) $= 3.1$ g/cm$^3$). Therefore the diffusivity of N$_2$O, NO, O$_2$, N$_2$, as well as other nonreactive molecular species such as the noble gases should be lower in the N containing films. Studies of B penetration in SiO$_2$ films of slightly varying density have shown enhanced diffusion resistance to B with excess film density. However, the rigidity of the N bonded lattice might be the basis of an equally important explanation. The three bonds connected to each N atom, such as in Si$_3$N$_4$, are more constrained than the two bonds connected to each O atom in SiO$_2$. Whereas the Si–O–Si bond angles can vary from 120° to 180° with little change in energy, the Si$_3$N$_4$ lattice is much more constrained. This may contribute to a decrease in the ability of the nitrided lattice to permit diffusion of atoms and small molecules.

To explain the “stopping power” of the N in Si–O–N films, a model assuming B diffusion via peroxy linkage defects (Si–O–O–Si bonds), whose concentration changes under different processing conditions and film thicknesses, has been suggested. For a 1.5 nm SiO$_2$ film, the diffusivity at
900 °C would increase by a factor of 24 as compared with 10 nm oxides. According to this model, the role of N is to compete with B for occupation of the defect sites. Another model, in which B diffuses substitutionally for Si atoms, has also been suggested.678 The role of the Si–N bond is to impede substitution for that Si atom. The model incorporated a Monte Carlo simulation and showed good agreement with experimental data. Several groups have also suggested that a dative B–N bond may be formed, thus hindering the diffusion of B in the dielectric.

Another interesting aspect of the diffusion barrier properties of Si–O–N films is that Si interstitials generated at the interface during oxynitridation are blocked from diffusing into the oxide. This results in an enhanced flux of the interstitials into the Si substrate, which in turn yields an increased density of oxidation induced stacking faults. These may then influence the extent of oxidation induced transient diffusion.684,685

B. Oxynitridation processes

Nitrogen may be incorporated into SiO2 using either thermal oxynitridation or annealing, or chemical and physical deposition methods, as is shown in Table VI. Thermal nitridation of SiO2 in NO or N2O generally results in a relatively low concentration of N in the films, on the order of $5 \times 10^{14}$ N/cm$^2$, the equivalent of <1 monolayer of N in the thin SiO2 film. Whether annealing preoxides, or oxynitriding the Si surface directly, the N content increases with temperature.173,203 To attain higher N concentrations, other methods such as CVD,688 JVD,313 ALD,587 or nitridation by energetic N particles can be used.525,688 These nitridation methods can be performed at lower temperatures, ~300–400 °C. However, the deposition methods are nonequilibrium and subsequent thermal processing steps are often required to improve film quality through minimization of defects. Because the thermodynamics of the Si–O–N system and the kinetics of N incorporation are rather complex, these different methods produce Si–O–N films with different total N concentrations and depth distributions.

1. NO processing

Oxynitridation of Si, or annealing of SiO2, in nitrous (N2O)689,690 or nitric (NO)662,691 oxide is the most popular processing method for making Si–O–N films by conventional thermal routes. NO is believed to be responsible for N incorporation in both cases.2,158,233,662,667,692 If NO is the species responsible for N incorporation, oxynitridation in pure NO should be considered for ultrathin dielectrics, since it results in nearly self-limiting growth behavior.157 From the practical viewpoint, the slower growth of Si–O–N relative to pure SiO2 facilitates good thickness control in the ultrathin regime during high temperature processing. Compared to N2O, oxynitridation in NO results in more N incorporation at a given temperature.169,170,201 In addition, NO Si–O–N films exhibit lower leakage currents and interface defect densities, as well as improved electrical stress properties.157,170,662,691,693 However, excessive N concentration at the interface may lead to mobility degradation.694

Figure 114 shows the total amount of O and N in NO-annealed ultrathin SiO2 films on Si(100), after 1 h anneals in the temperature range 700–1000 °C.203 As the temperature increases, the total amounts of both N and O increase, as well as the ratio of N to O in the film. It is believed that NO diffuses to the interface and dissociates there, thereby oxidizing and nitriding the Si.692 Quantitative MEIS N depth profiles203 of these Si–O–N films are shown in Fig. 115, in which it can be seen that the width of the N-containing region increases with temperature. As the position of the interface, deduced from the O depth profile and marked by arrows, propagates deeper into the Si substrate with increasing temperature, the N follows its movement. In all cases, the N is distributed almost evenly in the films except for the very near-surface SiO2 region. These results are inconsistent with a model of a continuous Si3N4 layer near the interface. However, multiple features in angularly dependent core level (N 1s and Si 2p) photoemission spectra,137,169,201 and the multiphase nature of electronic paramagnetic resonance (EPR) signals from Si–O–N films,695 suggest more complex local chemical bonding in the film.
Due to the relatively high concentration of N incorporated into NO films, the kinetics of reoxynitridation are very slow. To make a thicker NO film, an SiO2 preoxide of desired thickness can first be grown, followed by an NO anneal. Typically NO annealed preoxides yield N distributions different from the NO grown films, in that the N becomes concentrated near the interface. The N pileup near the interface results from NO molecular diffusion, and subsequent incorporation, at the interface. The kinetics of N incorporation in a 4.5 nm SiO2 film annealed in NO at 850 °C are shown in Fig. 116. One can see that after a fast initial accumulation, the rate of N incorporation saturates. The total concentration of N for an NO-annealed preoxide is comparable with the concentration of an NO grown film oxidized under equivalent conditions, at least for long oxynitridation times. Also noteworthy from Fig. 116 is the excellent agreement between MEIS and NRA N data.

2. N2O processing

Oxynitridation in N2O is particularly attractive because it allows one to incorporate a small but significant amount of N near the Si–O–N/Si interface, typically ~5 × 10^{14} N/cm², and because of its processing similarity to O2. However, among other factors, oxynitridation in N2O is complicated by the fast gas phase decomposition of the molecule into N2, O2, NO, and O at typical oxidation temperatures, 800–1100 °C.

a. Gas phase N2O decomposition at high temperatures.

In contrast to NO, a relatively stable molecule, N2O rapidly decomposes at high temperatures. The decomposition is a complex process. Gas flow rate, temperature, partial pressure of N2O, reactor type (e.g., rapid thermal processing versus furnace) and geometry, and gas phase impurity levels all affect the kinetics and final distribution of products. An additional complication is that N2O decomposition is strongly exothermic and can cause temperature and therefore film thickness nonuniformities across the reactor.

N2O chemistry has been studied extensively for many years by environmental scientists. Of the many possible NOx reactions known to occur, Table VII, the first five key steps seem to dominate the N2O decomposition to NO. The rate limiting reaction is the first step

\[
\text{N}_2\text{O} \rightarrow \text{N}_2 + \text{O}.
\]

This decomposition obeys first order kinetics, and thus the initial rate law for N2O decomposition is

\[
R_{\text{init}} = 2k_1[N_2\text{O}],
\]

which rapidly changes to

\[
R_{\text{late}} = k_1[N_2\text{O}]
\]

as the reaction proceeds, where \(k_1\) is the reaction constant of the initial reaction. The apparent activation energy for the decomposition of N2O is 2.5 eV/molecule. The characteristic decay time of the N2O concentration is on the order of ~20

---

**TABLE VII. Reaction scheme for N2O decomposition [from Gupta et al. (Ref. 699)].**

<table>
<thead>
<tr>
<th>Reaction Step</th>
<th>Reaction Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>N2O → N2 + O</td>
</tr>
<tr>
<td>2.</td>
<td>N2O + O → NO + NO</td>
</tr>
<tr>
<td>3.</td>
<td>N2O + O → N2 + O2</td>
</tr>
<tr>
<td>4.</td>
<td>O + O → O2</td>
</tr>
<tr>
<td>5.</td>
<td>NO + O → NO2</td>
</tr>
<tr>
<td>6.</td>
<td>NO2 + O → NO + O2</td>
</tr>
<tr>
<td>7.</td>
<td>NO2 + NO2 → NO + NO</td>
</tr>
<tr>
<td>8.</td>
<td>NO2 → NO + O</td>
</tr>
<tr>
<td>9.</td>
<td>NO2 + NO2 → NO + NO</td>
</tr>
<tr>
<td>10.</td>
<td>NO + NO → NO2 + NO</td>
</tr>
<tr>
<td>11.</td>
<td>NO2 + O → NO3</td>
</tr>
<tr>
<td>12.</td>
<td>NO3 + O → NO2 + O2</td>
</tr>
<tr>
<td>13.</td>
<td>NO3 + O → O2 + NO2</td>
</tr>
<tr>
<td>14.</td>
<td>NO3 + NO2 → N2O4</td>
</tr>
<tr>
<td>15.</td>
<td>N2O3 → NO2 + NO</td>
</tr>
<tr>
<td>16.</td>
<td>NO3 + NO2 → N2O5</td>
</tr>
<tr>
<td>17.</td>
<td>N2O3 + NO2 → N2O5</td>
</tr>
<tr>
<td>18.</td>
<td>O3 + O → O2</td>
</tr>
<tr>
<td>19.</td>
<td>O3 + O → O2</td>
</tr>
<tr>
<td>20.</td>
<td>O3 + O → O2</td>
</tr>
<tr>
<td>21.</td>
<td>O3 + NO → NO2 + O2</td>
</tr>
</tbody>
</table>
s at 1000 K and decreases with increasing temperature. Oxygen atoms then react further by the following two key reactions:

$$\text{N}_2\text{O} + \text{O} \rightarrow 2\text{NO},$$  

$$\text{N}_2\text{O} + \text{O} \rightarrow \text{N}_2 + \text{O}_2.$$  

The branching ratio for these two reactions lies between 0.1 and 0.5 and varies with the specific conditions. The rate law for NO formation is given by Eq. (15), and the apparent activation energy for the formation of NO is 2.4 eV/molecule.

The main products of N$_2$O decomposition are N$_2$, O$_2$, and NO, and the equilibrium concentration of NO increases with temperature, as shown in Fig. 117. Since N$_2$ is much less reactive than O$_2$ and NO, one might expect that a properly chosen mixture of NO and O$_2$ would produce a Si–O–N film similar to that grown in N$_2$O, but this is not the case. Oxynitridation in various mixtures of NO and O$_2$ is more similar to NO oxynitridation than to N$_2$O oxynitridation. This suggests that there are other active species present during N$_2$O decomposition which are important in the nitridation process. Atomic O is believed to play a key role in N$_2$O oxynitridation. Though the equilibrium concentration of atomic O is rather low, the transient concentration of atomic O created in the first step of the decomposition process may be relatively high, especially if N$_2$O decomposition takes place near the wafer. Since O radicals are very reactive, even a small partial pressure of O may have a significant effect on oxynitridation.

b. N incorporation and removal during N$_2$O oxynitridation. Under equivalent conditions, N$_2$O results in less N incorporation than NO. The total concentration of N increases with film thickness as well as with N$_2$O oxynitridation temperature, as can be seen from Fig. 118. This implies that N$_2$O oxynitridation may not be an efficient N incorporation mechanism for films thinner than ~3 nm. The N distribution in the films is sensitive not only to the processing conditions, but also to the reactor type. Rapid thermal (RT) N$_2$O films show N pileup near the interface, while the distribution of N in the furnace-grown films is broader. In the RTN$_2$O case the decomposition takes place near the hot wafer, and most of the decomposition products, including atomic O, can diffuse to the wafer. In a conventional furnace N$_2$O may decompose in the hot inlet of the furnace before it reaches the wafer. This depletion effect also explains the observation that the slower the input flow, the lower the N incorporation rate.

The fundamental difference between oxynitridation in N$_2$O and NO is that while both incorporate N via NO reactions near the interface, in the N$_2$O case the N incorporation occurs simultaneously with N removal from the upper layers of the film. The final N concentration and distribution is influenced by a competition between N incorporation and removal. The removal reaction likely involves O. Since the reactivities of NO, N$_2$O, and O$_2$ with Si, SiO$_2$, and Si–O–N are quite different, properly chosen sequences of thermal reactions with Si can lead to Si–O–N films with different N concentrations and profiles, and therefore electrical properties. Based on an understanding of the kinetics and thermodynamics of N incorporation in SiO$_2$, one can “engineer” the N profile in an ultrathin dielectric. Nitrogen profile engineering has also been accomplished using an O and N radical deposition technique.

3. Nitridation in NH$_3$

Nitridation in NH$_3$ was one of the first methods used to incorporate relatively high (~10–15 at. %) concentrations of
Nitrogen piles up both at the interface and the outer surface during the initial stages of nitridation, unlike the case of N2O nitridation, and the N containing region near the interface is about 3 nm wide. The concentration of N slowly increases with NH3 nitridation time, and the N distribution becomes more uniform in the film. The thickness of the film remains essentially unchanged during nitridation. Nitrogen incorporation is enhanced by decreasing the SiO2 film thickness. NH3, directly reacted with Si, has also been used to form thin Si3N4 layers, although they may contain substantial O if either the ambient or the NH3 contain H2O or O.

Ammonia nitridation introduces high concentrations of H into SiO2 films, which can act as electrical traps. SIMS analysis shows that the H concentration increases monotonically with nitridation time and temperature. The H tends to pileup near the interface. Based on the detrimental effects of H on device performance, lighter nitridation, i.e., at lower temperature for shorter times, is desirable since it introduces less H. Detailed isotopic studies of the behavior of H, as well as O and N in NH3-nitrided SiO2 have been carried out.

Ammonia nitridation of SiO2 for gate dielectric applications has generally fallen into disfavor due to the deleterious effects of the incorporated H. However, reasonable electrical results have been reported and there are still applications for NH3 nitridation in the growth of some stacked SiO2/Si–O–N/SiO2 gate and tunnel dielectric layers.

4. Nitridation in N2

Although N2 is relatively inert, annealing of SiO2 films in N2 results in an initial reduction of fixed oxide charge. This behavior was attributed to the formation of an Si–O–N layer near the interface, as detected by XPS. This work was performed in a furnace at high temperatures (>925 °C), for long times (>1 h). To reduce the thermal budget, one can form Si–O–N in nominally pure N2 by rapid thermal processing. The term “nominally” means here that all elements in the film. Other researchers have also seen the formation of Si–O–N layers during high temperature RTP.

VII. THE POST-SiO2 ERA: ALTERNATE GATE DIELECTRICS

SiO2 gate dielectric thickness is continually decreasing with device scaling, as is illustrated in Fig. 3. SiO2 or lightly nitrided Si–O–N gate dielectrics are expected to be useful to thicknesses of about 1.3 nm. Thinner films will exhibit excessive gate leakage current, Fig. 5, and diminished drive current, Fig. 7, thus requiring the implementation of alternate gate dielectrics. It will not be easy to replace SiO2, since with the exception of its lower dielectric constant, it is an ideal material (Table I). Alternate gate dielectrics will introduce many complicated materials science and process integration issues, but many of these will be addressed first for back-end capacitor applications, before the necessary introduction into the more demanding gate dielectric application, perhaps 5 years away.
A. Si₃N₄

Si₃N₄ has been suggested as an alternative gate dielectric. However, typically its interface properties with Si are poor. Further, it is difficult to fabricate ultrathin films without incorporating some O near the interface, since Si has a greater affinity for O than for N. Indeed, the incorporation of O at the interface may be the reason that high N-containing Si–O–N films have any reasonable electrical properties at all. Some progress has been made in growing electrically acceptable pseudo-Si₃N₄ and Si₃N₄ films but since the dielectric constant, even for the best case of pure Si₃N₄, is only twice that of SiO₂, it may be preferable to use metal oxides, some of which have much higher dielectric constants. If the interface properties of Si₃N₄ are perfected, it may be useful as an interim gate dielectric solution.

B. Alternate (higher) dielectric constant materials

Table VIII is a list of requirements for alternate gate dielectrics. We will use this table to frame our discussion. A more comprehensive treatment of high-k gate dielectrics will be found in a companion review article. Alternate gate dielectric materials must meet a variety of chemical, physical, electrical, and manufacturing criteria. The gate dielectric must be thermodynamically stable on Si, with respect to the formation of both SiO₂ and MSiₓ. At the appropriate process temperatures to which the wafer will be exposed, Si should not reduce the gate dielectric to form SiO₂ or silicide. If SiO₂ were formed, the capacitance “budget” would be consumed by a low dielectric constant material. If the silicide were formed, a conductive path across the channel might be created. Stability criteria for all simple, as well as some multi-component metal oxides, have been thoroughly explored. The imposition of the thermodynamic stability criteria substantially reduces the field of acceptable alternate gate dielectrics. Figure 120 shows the elements whose oxides are thermodynamically stable on Si, at 1000 K, with respect to both SiO₂ or silicide formation.

Table IX is a partial list of candidate high-k materials suggested by Fig. 120. No simple oxide seems to have the right combination of dielectric constant, thermodynamic stability on Si, and stability of the amorphous phase at reasonable processing temperatures. Zirconium, and other metals including notably Ti, will readily form silicides if the processing ambient is not controlled properly. Al₂O₃ has a reasonable dielectric constant, and has therefore been the

<table>
<thead>
<tr>
<th>TABLE IX. Some candidate high-κ simple metal oxides.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO₂</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Dielectric constant</td>
</tr>
<tr>
<td>Band gap, eV</td>
</tr>
<tr>
<td>Free energy of reaction: Si+MOₓ→M+SiO₂ at 727 °C, kcal/mole of MOₓ</td>
</tr>
<tr>
<td>Stability of amorphous phase</td>
</tr>
<tr>
<td>Silicide phase formation possible?</td>
</tr>
<tr>
<td>Oxygen diffusivity at 950 °C, cm²/s</td>
</tr>
</tbody>
</table>

FIG. 120. Elements whose oxides are stable on Si, with respect to the formation of both SiO₂ and MSiₓ (silicide), at 1000 K. The question mark indicates lack of thermodynamic data for the silicide of the particular element [adapted from Hubbard et al. (Ref. 725)].
subject of considerable research.\textsuperscript{207,726–728} Since SiO\textsubscript{2} and Al\textsubscript{2}O\textsubscript{3} have very stable amorphous phases, silicates\textsuperscript{729–731} and aluminates\textsuperscript{732,733} may be promising candidates. It should be mentioned that some metal oxides, not stable on Si, e.g., Ta\textsubscript{2}O\textsubscript{5}, may be suitable with the use of a barrier layer, or due to the slow kinetics of the reaction, with the Si substrate. For that reason, Ta\textsubscript{2}O\textsubscript{5} is included in Table IX; however, the use of these materials involves compromises that will not lead to the most robust solutions for high-\(\kappa\) gate materials.

Clearly, amorphous gate dielectrics (as is SiO\textsubscript{2}), offer many benefits and are the most desirable. However, most metal oxides exhibit a strong tendency to crystallize. It remains to be seen to what extent some degree of crystallinity may degrade the reliability of metal oxides. Figure 121 is an example of amorphous phase stability in a silicate phase.\textsuperscript{729} SiO\textsubscript{2} readily forms a stable amorphous phase due to the SiO\textsubscript{4} tetrahedral subunits that comprise it. Methods of stabilizing the amorphous phases, either through the formation of silicates, or other alloying schemes,\textsuperscript{733} may have to be found. The microstructure of the dielectric layer plays a significant role in its performance. Polycrystalline dielectrics are not desirable because grain boundaries will facilitate mass and electrical transport, and thereby adversely affect reliability. Crystallization may also give rise to interfacial roughness, which will lower channel mobility. Although epitaxial dielectrics have been demonstrated,\textsuperscript{734} and promise very low defect density interfaces, they too present problems. The requirement for an epitaxial match reduces the field of possible candidates yet further. Also, the domain boundaries in the epitaxial dielectric that will result from growth on the stepped Si surface may behave just like grain boundaries.

Further, epitaxial dielectrics will have anisotropic properties, not necessarily desirable. Finally, commercial growth techniques for epitaxy of these materials would have to be developed.

From an electrical standpoint, the dielectric constant of the alternate gate dielectric should be \(\sim 9–25\). If the dielectric constant is too high, the gate dielectric layer will be too thick relative to the channel dimensions, and fringing field effects will affect performance.\textsuperscript{735,736} The dielectric should also have a large band gap, so that gate leakage current may be minimized. Few alternate gate dielectric materials have band gaps approaching that of SiO\textsubscript{2}, 9.0 eV, but the thicker gate dielectric layer (compared to its SiO\textsubscript{2} equivalent) will also result in lower leakage current. Further, the gate dielectric should form a high quality, low defect density (\(< 5 \times 10^{10} \text{cm}^{-2} \text{eV}^{-1}\) interface with Si. There are many other considerations that must be taken into account when selecting a high-\(\kappa\) material, including etchability, O diffusivity, and electronic defect structure. These will not be known in detail until the field of candidates has been narrowed down.

Processing requirements for alternate gate dielectrics are also important. Without a manufacturable process, high-\(\kappa\) materials will never be introduced into technology. It is very critical that adventitious SiO\textsubscript{2} not be formed during alternate gate dielectric growth. Although a few monolayers of SiO\textsubscript{2} at the interface might prove beneficial for good interface properties, a thicker layer of this low dielectric constant material would not allow for the growth of an alternate gate dielectric with \(t_{\text{ox}}\) (eq) of \(< 1.0 \text{ nm}\). Reactive sputtering in particular results in an adventitious SiO\textsubscript{2} layer at least 1.5 nm thick,\textsuperscript{726} and is not a suitable deposition technique. Low pressure CVD,\textsuperscript{727,737,738} atomic layer deposition (ALD),\textsuperscript{739} and molecular beam epitaxy (MBE),\textsuperscript{734} all of which have the potential for growth with sufficient interface control, are candidates for deposition of the alternate gate dielectrics. Of these, MBE may be the most problematic in that this technique has not made commercial inroads in Si processing. Further, in the future new gate formation (“replacement gate”) technologies,\textsuperscript{740,741} might be used to make integrated devices. Conformal film coverage will be of paramount importance for such technologies, since nonplanar topographies will be introduced, and processes such as ALD\textsuperscript{208,207,512,739,742} or LPCVD\textsuperscript{738} will be required.

ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful discussions and comments from D. A. Buchanan, E. Cartier, T. W. Sorsch, J. H. Stathis, and B. E. Weir.


