

**SILICON PROCESSING
FOR
THE VLSI ERA**

**VOLUME 1:
PROCESS TECHNOLOGY
Second Edition**

**STANLEY WOLF Ph.D.
RICHARD N. TAUBER Ph.D.**

LATTICE PRESS

Sunset Beach, California

91608-11

DISCLAIMER

This publication is based on sources and information believed to be reliable, but the authors and Lattice Press disclaim any warranty or liability based on or relating to the contents of this publication.

Published by:

LATTICE PRESS

Post Office Box 340

Sunset Beach, California 90742, U.S.A.

Cover design by Roy Montibon, New Archetype Publishing, Los Angeles, CA.

Copyright © 2000 by Lattice Press.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system without written permission from the publisher, except for the inclusion of brief quotations in a review.

Library of Congress Cataloging in Publication Data

Wolf, Stanley and Tauber, Richard N.

Silicon Processing for the VLSI Era

Volume 1: Process Technology

Includes Index

1. Integrated circuits-Very large scale integration. 2. Silicon. I. Title

ISBN 0-9616721-6-1

9 8 7 6 5 4 3 2

PRINTED IN THE UNITED STATES OF AMERICA

Chapter 8

THERMAL OXIDATION of SILICON

When a silicon surface at elevated temperatures is exposed to an oxidizing ambient (such as oxygen or water vapor) a stable and tenacious oxide forms. This process is known as *thermal oxidation*. The ability to form an oxide (SiO_2) that is chemically stable at high temperatures and adherent to the silicon is one of the cornerstones of planar silicon processing. The electrical stability of the Si/SiO_2 interface is also the basis of reliable MOS devices in silicon.

Note that germanium (Ge) is also a semiconductor, and it has some material properties that could be exploited to make electronic devices superior to those made using Si. However, Ge is not used to manufacture integrated circuits. One of the main reasons is that the oxide it forms, GeO_2 , is not thermally stable above 800°C and is also water soluble. SiO_2 is stable to temperatures above the melting point of Si ($\sim 1420^\circ\text{C}$) and does not dissolve in water. Note that there is another silicon oxide that also forms (i.e., silicon monoxide [SiO]), but like GeO_2 it is not stable at elevated temperatures.

When the oxidizing ambient contains steam or water vapor, the process that forms SiO_2 is known as *wet oxidation*, and when the ambient is pure O_2 , it is called *dry oxidation*. The term *oxide* will be used interchangeably with SiO_2 or *silicon dioxide* throughout this chapter.

Thermally grown SiO_2 is employed in ULSI applications in the thickness range of 3.0-400 nm. Some of the functions of such oxide films include: a) gate oxides and sacrificial oxides in MOSFETs; b) capacitor dielectrics in DRAMs; c) passivation of the silicon surface; d) tunnel oxides in non-volatile memory devices; e) screen oxides for masking ion-implantation; f) liner oxides for shallow-trench isolation (STI) trenches; and g) field isolation oxides in LOCOS (LOCAl Oxidation of Silicon) isolation structures. Table 8-1 lists the range of oxide thicknesses commonly used for these, and other, applications.

Although care must be taken when growing any thermal oxide, the most critical oxide application is the *MOSFET gate oxide*. The gate oxide thickness used in the most advanced processes has been getting thinner as the feature sizes are scaled. This trend will continue with each

Table 8-1 RANGES OF THERMAL OXIDE THICKNESS USED IN ULSI PROCESSING

Application	SiO_2 Thickness (nm)
Gate oxide insulators for MOS devices	3.0-8.0
Tunnel oxides in EEPROMs and flash memories	6.0-10.0
Screen oxides for ion implantation	10.0-20.0
Side-wall liners for shallow trench isolation (STI)	15.0-40.0
Field oxide for LOCOS isolation	200-400
Re-oxidation of the etched gate stack side-wall damage	7.5-5.0
Inter-polysilicon dielectric	4.0-20.0
Masking film against ion implantation and diffusion	100-200
Sacrificial oxides for gate applications	6.0-15.0
Capacitor dielectric in DRAM circuits	3.0-5.0

Table 8-2 EQUIVALENT GATE OXIDE THICKNESS VS. TECHNOLOGY NODE
(Extracted from NTRS, SIA [1997])¹

First year product shipped	1997	1999	2001	2003	2006	2009	2012
Linewidth in nm	250	180	150	130	100	70	50
Equivalent Gate Oxide, nm	4-5	3-4	2-3	2-3	1.5-2	<1.5	<1

new generation of technology. For example, the thickness of an advanced MOS gate oxide process in 1989 was about 25 nm and in 1998 was about 4 nm. One of the most significant challenges for ULSI process engineers is to develop and maintain a gate oxide process that results in uniform and reproducibly high quality oxides that are only several atomic layers thick. Significant advances in wafer cleaning, furnace technology, and the introduction of rapid thermal oxidation have gone a long way to accelerate this astounding capability.

Table 8-2 is extracted from the National Technology Roadmap for Semiconductors (NTRS)¹ published in 1997 by the Semiconductor Industry Association (SIA). It shows the equivalent gate oxide thickness that will be required for each new technology generation from 1997 to 2012. In general, the thinnest gate oxide requirements are driven by the performance needs of microprocessors rather than DRAM devices. As a result, the gate oxide thickness for microprocessors is 1–2 generations ahead of that for DRAMs. The decrease in the gate oxide thickness for future generations is evident from Table 8-2. One of the concerns is that SiO₂ may no longer be a suitable gate dielectric when the requirement for equivalent thickness is less than 1.5–2 nm. At that time an alternative gate dielectric material will have to be found. This issue is discussed in the concluding section of the chapter.

The properties of thermal SiO₂ and the technology for growing thin films of SiO₂ for IC fabrication are described in this chapter. These subjects will be discussed in the following order: a) properties of silica glass; b) kinetics of oxidation (Deal-Grove model); c) factors impacting oxide growth in one dimension; d) the initial oxidation stage and growth of thin oxides; e) the nature of the Si/SiO₂ interface; f) dopant impurity redistribution during oxidation; g) oxidation of polysilicon and silicon nitride; h) two-dimensional oxide growth; i) oxidation systems; k) thin oxide growth processes; j) the characterization of oxide films; and l) trends in thin gate oxides and alternative materials for gate dielectrics. Table 8-3 lists some typical values for selected physical properties of thermally grown SiO₂.

Table 8-3 SELECTED PHYSICAL CONSTANTS OF THERMAL SILICON DIOXIDE

Property (units)	Value	Property (units)	Value
Density (g/cm ³)	2.27	Melting point (°C)	≈1700
Dielectric constant	3.9	Molecular weight	60.08
DC resistivity @ 25°C (Ω-cm)	10 ¹⁶	Molecules/cm ³	2.3×10 ²²
Energy gap (eV)	≈9	Specific heat (J/g°C)	1.0
Thermal conductivity (W/cm ² °C)	.014	Film stress (at 25°C) dynes/cm ²	2–4×10 ⁹ compressive
Linear expansion coefficient (ppm/°C)	0.05	IR absorption peak (mm)	9.3
Refractive index	1.46	Etch rate BHF (49%) nm/min	100

8.1 THE PROPERTIES OF SILICA GLASSES

The amorphous (non-crystalline) state of SiO_2 is termed *fused silica*. The misnomer "fused quartz" or "quartz" is often seen when furnace hardware is being described, but such hardware actually consists of fused silica. The thermally grown oxide on Si is also fused silica. Quartz is the name of one of the crystalline phases of SiO_2 . Fused silica, being amorphous, is thermodynamically unstable below the melting point of quartz ($\sim 1710^\circ\text{C}$). That is, there is a thermodynamic driving force that attempts to transform fused silica from its amorphous state to that of a quartz crystal, a phenomenon called *devitrification*. The rate of this transformation is found to decrease with decreasing temperature, and approaches zero below 1000°C . Devitrification can be a problem for the fused silica tubes, boats, and pyrogenic torches used in high temperature processing, as the crystallized material breaks away and becomes a source of particles.

Stevels and Kats² developed a model that allows the structure of fused silica to be visualized. Their model acknowledges that fused silica is amorphous, and thus there is no long-range structure. However, they postulate that short-range order exists. This order is centered around the *structural formula* for the material (i.e., $[\text{SiO}_4]^{4-}$). A single Si atom (with a valence of +4) is located at the center of a regular polyhedron (in this case a tetrahedron), with an oxygen ion (O^{2-}) at each of its four vertices (Fig. 8-1a). The tetrahedral distance between the Si and O ions is 0.162 nm, and the oxygen-oxygen ion distance is 0.227 nm. The tetrahedra are joined by an oxygen ion called a *bridging oxygen*, which is shared between two touching tetrahedra. In crystalline quartz the tetrahedra are arranged to yield long range order and periodicity (Fig. 8-1b).

Thermally grown oxide (fused silica), on the other hand is considered to be amorphous, with a structure shown in Fig. 8-1c. That is, the oxygen atoms on some of the vertices of the tetrahedra are *nonbridging oxygen ions* (i.e., which are not shared). Thus, the tetrahedra are randomly distributed in a glassy *network*. This makes for a more open structure since the tetrahedra are not regularly arranged. Hence, the glassy state is less dense than the crystalline state. The density of fused silica is $2.15\text{--}2.25\text{ g/cm}^3$, while that of crystalline quartz is 2.65 g/cm^3 . The greater the ratio of non-bridging to bridging oxygen atoms, the poorer the quality of the SiO_2 glass. The lower density also implies a more open structure. This open structure is conducive to the interstitial diffusion of impurities through the network.

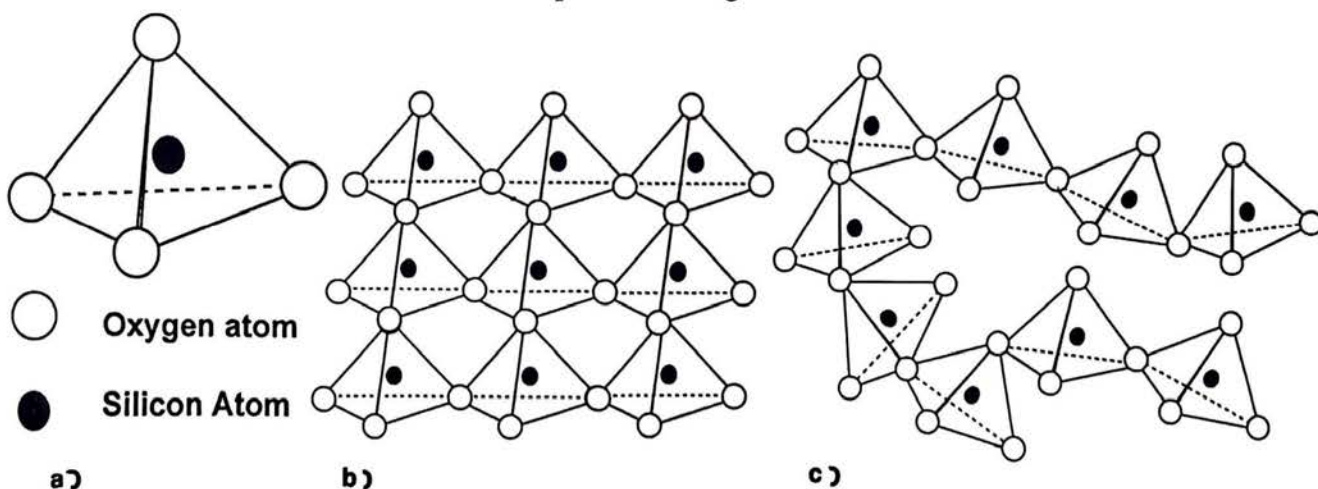


Fig. 8-1 Schematic representation of a) the $(\text{SiO}_4)^{4-}$ tetrahedron; b) $(\text{SiO}_4)^{4-}$ tetrahedra showing long-range order and periodicity (crystalline quartz); c) $(\text{SiO}_4)^{4-}$ tetrahedra arranged in a random pattern (amorphous - as in the fused silica form of SiO_2).

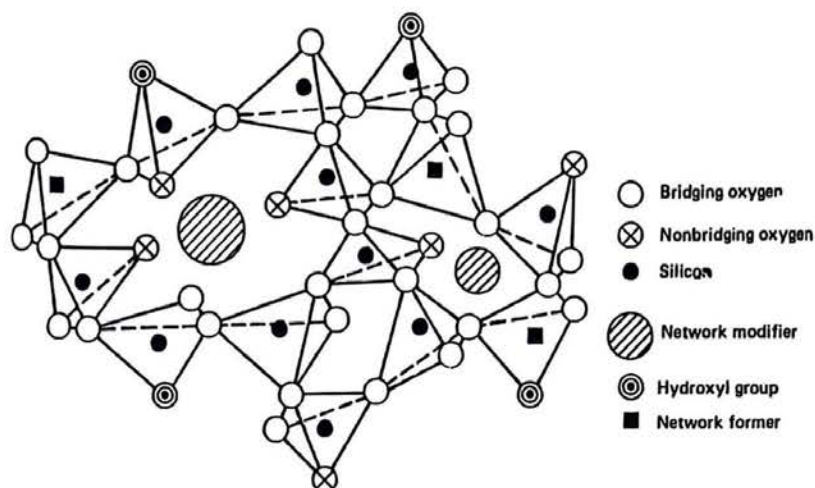


Fig. 8-2 Model of impurity incorporation in fused silica glass.³ (© 1965 IEEE).

Impurities introduced into fused silica radically change its properties. As is the case of silicon, both *substitutional* and *interstitial* impurities can exist. In Fig. 8-2 a schematic model is shown (developed by Revesz) of impurities in an SiO_2 network.³ The substitutional impurities replace silicon in the structure. The most important impurities of this type are boron (B) and phosphorus (P) atoms. Another term for these impurities is *network formers*, since they themselves can be the basis of a glassy structure (B_2O_3 or P_2O_5). Since these impurities have either one additional electron (P^{5+}), or one less electron (B^{3+}) to contribute to the covalent bonding than does silicon (Si^{4+}), an electrical unbalance develops. The missing (or extra) electrons in the network introduced with such impurities at the center of the tetrahedra (instead of Si) are accommodated by the elimination or formation of bridging oxygen ions, respectively. In the case of the elimination of a bridging oxygen bond, a non-bridging oxygen is created. This will tend to weaken the network (e.g., boron in SiO_2). The extra electron of phosphorous will *reduce* the number of non-bridging oxygen atoms, which will improve the oxide quality.

The oxides of Na, K, Pb, and Ba enter the structure as interstitial impurities. When this occurs the metal ion gives up its oxygen to the network, thereby producing two nonbridging oxygen ions, which replace the original bridging oxygen. Impurity oxides of this type are termed *network modifiers* since they do not form glasses themselves. Glasses heavily doped with network modifiers are most noted for their low melting points (e.g., leaded glass).

The last class of impurities that affect oxide quality are the *hydroxyl* (OH^-) group. These impurities arise when water (H_2O) enters the structure. The water reacts with a bridging oxygen to form a pair of stable non-bridging OH^- groups. The reaction is represented by:



The decrease in the number of bridging oxygen atoms weakens the structure resulting in "poorer" quality oxides. As shall be seen, OH^- can benefit the oxide breakdown field yet degrade the oxide reliability.

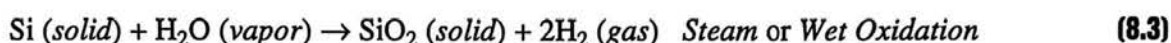
8.2 OXIDATION KINETICS: THE DEAL-GROVE MODEL

Silicon exhibits a propensity to form a stable oxide (SiO_2). This is demonstrated in nature as Si is never found in the elemental form, but instead is found in either silicate or silica (sand) compounds. When freshly cleaved Si is exposed to an oxidizing ambient (e.g., O_2 or H_2O) it will

form a very thin (1.0–2.0 nm) oxide layer, even at room temperature. When Si is exposed to oxidizing ambients at elevated temperatures, more rapid growth occurs and thicker oxides form.

As noted earlier, thermal oxide films of different thicknesses are needed for various device structure applications. It is therefore necessary to grow oxide films of such specific thicknesses. In this section a model which predicts the growth of SiO_2 for many of the conditions encountered in ULSI processing will be described. This model was proposed by Deal and Grove in 1965, and it remains the most important model for describing SiO_2 growth.⁴ Here their model is referred to as the *D-G model*, but in the literature it is also known as the *linear-parabolic (L-P) model*. Nevertheless, this model is not accurate for very thin oxides, as shall be discussed later.

There are two *oxidants* (or *oxidizing species*) used for oxidation - dry oxygen and water. Oxidation using O_2 is called *dry oxidation*, and that using H_2O is termed *steam oxidation*. Sometimes mixtures of O_2 and H_2O are used, and this process is known as *wet oxidation*. Recently there have been reports of using nitrous oxide (N_2O) as the oxidant to form a nitrogen-doped silicon dioxide. Such nitrided oxides are also discussed later. The overall chemical reactions governing the oxidation of Si are written as:



These oxidation reactions occur at the Si/ SiO_2 interface. Therefore, as the oxide grows, silicon is consumed and the interface moves into the silicon. Figure 8-3 shows the relationship between the thickness of the grown oxide film and the consumed silicon. Based on the relative densities and molecular weights of Si and SiO_2 it is found that the amount of silicon consumed is 44% of the final oxide thickness. Thus, if 1000 nm of oxide is grown, 440 nm of Si will be consumed. This relationship is important for calculating step heights that form in the silicon as a result of the varying oxidation rates at different locations on the silicon surface.

8.2.1 The Deal-Grove (or Linear-Parabolic) Model

The linear parabolic model developed by Deal and Grove results in an accurate representation of the growth of SiO_2 on Si over a wide range of thicknesses (30–500 nm), temperatures (700–1300°C), and oxidant partial pressures (0.2–25 atm).

The model assumes that the oxidation process occurs as a result of two fluxes that sequen-

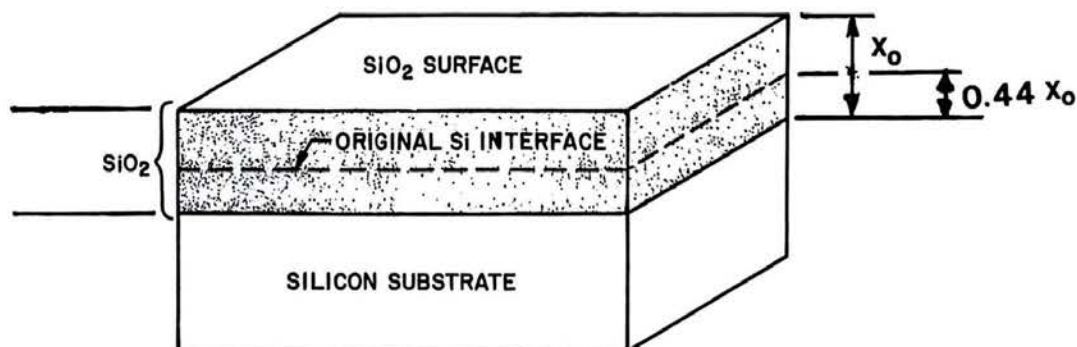


Fig. 8-3 The thermal oxidation of silicon model.⁴ Reprinted with permission of the American Physical Society.

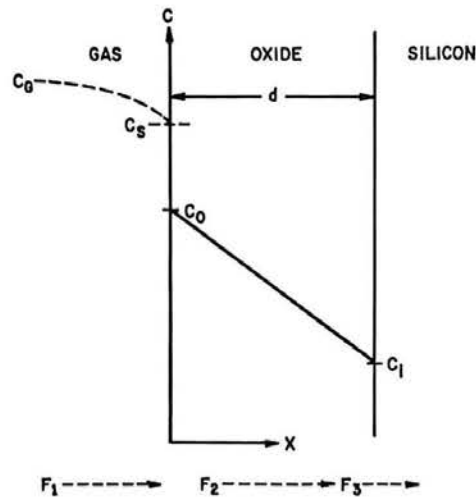


Fig. 8-4 The Deal-Grove model of thermal oxidation of silicon.⁴ Reprinted with permission of the American Physical Society.

ially transport the oxidizing species from the gas to the Si/SiO₂ interface, where a third flux is involved in the consumption of the oxidant by the reaction with the silicon. These fluxes are shown schematically in Fig. 8-4. (*Flux* is defined as the number of atoms or molecules crossing a unit area in a unit time, and has the dimensions of [number/cm²sec].) The three fluxes are formally identified as: a) the flux of the oxidizing species moving from the bulk of the gas phase to the gas/oxide interface, F_1 ; b) the flux of the oxidizing species as it diffuses through the existing oxide to the Si/SiO₂ interface, F_2 ; and c) the flux of the oxidizing species as it is consumed by reaction at the Si/SiO₂ interface, F_3 . In the steady-state condition, the three fluxes will be equal, i.e., $F_1 = F_2 = F_3$. The model continues by approximating each of these fluxes as they relate to physical quantities.

The oxidizing species are introduced into the reactor (e.g., a fused silica tube) as a gas, where they are then brought from the *bulk* of the gas stream to the *wafer surface* by a mass-transport process. The flux F_1 that occurs as a result of this transport is assumed to arise from the concentration difference between the oxidizing species in the bulk of the gas, C_g , and that at the oxide surface, C_s . It is further postulated that the magnitude of the flux is linearly proportional to the concentration difference between the bulk and the surface, and that h_g is defined as the *mass-transfer coefficient*: F_1 can then be written as:

$$F_1 = h_g (C_g - C_s) \quad (8.4)$$

The mass transfer coefficient h_g is related to the diffusivity of the oxidizing species D and the thickness of the boundary layer. (*Boundary layers* are discussed in Chap. 6.) Since the three fluxes are eventually equated, it is necessary to be able to express them all in terms of the same set of variables. Therefore, it is useful to express Eq. 8-4 in terms of the equilibrium concentration of the gas species in the oxide, instead of in terms of the concentrations in the gas. To do so, we use *Henry's Law*, which states that the concentration of a species dissolved in a solid is proportional to the partial pressure of the species in the gas phase at the solid surface, and the constant of proportionality is Henry's Law constant, H . Hence, we can write:

$$C_0 = HP_s \quad (8.5)$$

and

$$C^* = HP_g \quad (8.6)$$

where: P_s is the partial pressure of the oxidizing species at the oxide surface; P_g is the partial pressure in the bulk of the gas phase; C_o is the equilibrium concentration of the oxidizing species in the oxide at the outer surface; and C^* represents what the equilibrium concentration of the oxidizing species *would be* if the partial pressure at the oxide surface was P_g .

Assuming ideal gas behavior ($PV = kT$) in addition to Henry's Law, C_g and C_s can be rewritten in terms of the partial pressures as follows:

$$C_g = \frac{P_g}{kT} \quad (8.7)$$

$$C_s = \frac{P_s}{kT} \quad (8.8)$$

Combining Eqs. 8-4, 8-5, 8-6, 8-7, and 8-8, F_1 can be rewritten as:

$$F_1 = h (C^* - C_o) \quad (8.9)$$

where h is a newly defined mass-transfer coefficient, related to h_g by the constant $(1/HkT)$:

$$h = \frac{h_g}{HkT} \quad (8.10)$$

A limitation of this analysis is that Henry's Law does not apply under the conditions of molecular dissociation, hence Eq. 8-10 only holds if the oxidant is diffusing through the oxide in molecular form (i.e., as O_2 or H_2O). Indeed, most of the oxidation models assume molecular diffusion, although some models have postulated atomic oxygen as the diffusing species. Once inside the oxide layer, the oxidants diffuse towards the Si/SiO₂ interface, and this transport effect is described by the second flux, F_2 . This flux can be expressed as:

$$F_2 = -D \frac{dC}{dx_{ox}} = -D \frac{(C_o - C_i)}{x_{ox}} \quad (8.11)$$

in analogy with Fick's law, where D is the diffusion coefficient of the oxidizing species in the oxide, C_i and C_o are the concentrations of the oxidizing species at the growth interface and oxide surface, respectively, and x_{ox} is the thickness of the growing oxide.

Finally, the reaction-rate flux, F_3 , is expressed as being proportional to the concentration of the oxidizing species C_i at the interface:

$$F_3 = k_s C_i \quad (8.12)$$

where k_s is the first order reaction-rate constant of the reaction taking place at the interface.

Under steady state conditions (no build-up or depletion of oxidizing species) the three fluxes must be equal ($F_1 = F_2 = F_3 = F$). The result of setting Eqs. 8-9, 8-11, and 8-12 equal is a set of simultaneous equations, with solutions given by:

$$C_i = \frac{C^*}{1 + \frac{k_s}{h} + \frac{k_s x_{ox}}{D}} \quad (8.13)$$

$$C_o = \frac{C^* (1 + \frac{k_s x_{ox}}{D})}{1 + \frac{k_s}{h} + \frac{k_s x_{ox}}{D}} \quad (8.14)$$

Two important limiting cases occur for Eqs. 8-13 and 8-14. The first arises when the diffusion constant D is very small, and thus the flux of oxidant through the SiO_2 is small compared to the flux associated with the reaction at the Si/SiO_2 interface. This is referred to as the *diffusion limited* case, and under such conditions $C_i \rightarrow 0$ and $C_o \rightarrow C^*$. That is, as soon as any oxidant reaches the interface it is consumed by the reaction with silicon and the overall process is controlled by the diffusion of oxidant through the existing oxide.

The second case is encountered when D is very large, and the process is therefore *reaction-rate limited*. In that case:

$$C_i = C_o = \frac{C^*}{1 + \frac{k_s}{h}} \quad (8.15)$$

Under such conditions, a bountiful supply of oxidant is supplied to the Si/SiO_2 interface and the oxidation rate is controlled by the reaction rate of the oxidant.

By making use of these solutions together with the definition of the number of oxidant molecules incorporated into a unit volume of oxide, N_1 , the growth rate (dx_{ox}/dt) of the oxide can now be calculated. That is, since there are 2.2×10^{22} molecules of SiO_2 per cm^3 in the oxide, and since two oxygen atoms are incorporated per SiO_2 molecule, the value of N_1 for dry molecular oxygen (O_2) is $2.2 \times 10^{22}/\text{cm}^3$, and is $4.4 \times 10^{22}/\text{cm}^3$ for molecular water (H_2O).

Combining Eqs. 8-12 and 8-13 together with the definition of flux:

$$N_1 \frac{dx_{\text{ox}}}{dt} = F_3 = \frac{k_s C^*}{1 + \frac{k_s}{h} + \frac{k_s x_{\text{ox}}}{D}} \quad (8.16)$$

This differential equation (i.e., Eq. 8-16) can be solved by applying the initial boundary condition, $x_{\text{ox}} = x_i$ at $t = 0$. The value of x_i may include the native oxide that forms on bare Si in room ambient, or any oxide present from a previous process. The solution to Eq. 8-16 is the Deal-Grove (or Linear-Parabolic) model and is given by:

$$x_{\text{ox}}^2 + Ax_{\text{ox}} = B(t + \tau) \quad (8.17)$$

where

$$A = 2D\left(\frac{1}{k_s} + \frac{1}{h}\right) \quad (8.18)$$

$$B = \frac{2DC^*}{N_1} \quad (8.19)$$

$$\tau = \frac{x_i^2 + Ax_i}{B} \quad (8.20)$$

In Eq. 8-20, τ represents the time displacement needed to account for the initial oxide layer x_i .

The relationship expressed by Eq. 8-17 is found to quite accurately model the experimental results over a wide range of conditions of temperature, pressure, and oxidizing species. However, it fails to predict the initial stage of oxidation ($x_{\text{ox}} < 30$ nm) where the dry oxidation growth rate is found to greatly exceed that predicted by Eq. 8-17. Thin-oxide growth kinetics are discussed in a subsequent section of the chapter.

To help understand the fundamental mechanisms underlying the model we examine two limiting cases of Eq. 8-17. For very long oxidation times $t \gg \tau$, and Eq. 8-17 reduces to:

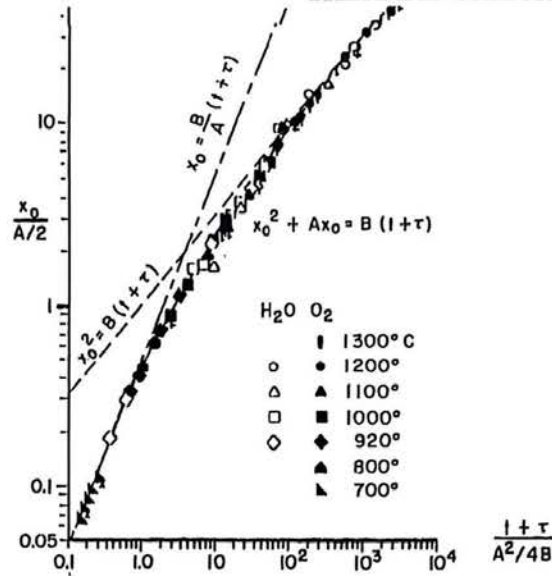


Fig. 8-5 The relationship governing the thermal oxidation of silicon and its two limiting forms.⁴ Reprinted with permission of the American Physical Society.

$$x_{\text{ox}}^2 = Bt \quad (8.21)$$

This is termed the *parabolic growth law* and B is the *parabolic rate constant*. From Eq. 8-19 one sees that B is proportional to the diffusion constant D , which represents the diffusion of the oxidant through the existing oxide. This implies that oxide growth in the parabolic regime is diffusion controlled. In other words, the rate limiting step involves the diffusion of oxidant across the growing oxide film. The thicker the oxide, the longer it takes for the oxidant to diffuse, hence the slower becomes the growth rate. As shown in Eq. 8-21 the oxide thickness is proportional to the square-root of the oxidizing time.

For very short times, on the other hand, $(t + \tau) \ll A^2/4B$. Thus, Eq. 8-17 reduces to:

$$x_{\text{ox}} = \frac{B}{A}(t + \tau) \quad (8.22)$$

Equation 8-22 is termed the *linear growth law* and (B/A) is the *linear rate constant*. From Eqs. 8-19 and 8-20 one sees that (B/A) is independent of the diffusion coefficient and is given by:

$$\frac{B}{A} = \frac{k_s h}{k_s + h} \left(\frac{C^*}{N_1} \right) \quad (8.23)$$

For most practical oxidation conditions the mass-transfer coefficient h is about 10^3 times larger than the reaction-rate constant, k_s . Hence, h can be eliminated from Eq. 8-23 and the linear rate constant, B/A , becomes:

$$\frac{B}{A} = k_s \left(\frac{C^*}{N_1} \right) \quad (8.24)$$

The key result of Eq. 8-24 is that for short oxidation growth times the growth rate depends on k_s , and the diffusion of the oxidant plays no role in limiting the growth rate.

Figure 8-5 shows a plot of the normalized oxide thickness (normalized for temperature and oxidant) for both dry and wet oxidation over the temperature range of 700–1300°C. The experimental results cluster very well around the curve predicted by Eq. 8-17. In the limits (short and long times) the results approach the lines given by $x_{\text{ox}} = B/A(t + \tau)$ and $x_{\text{ox}}^2 = Bt$ (representing the linear and parabolic limits, respectively). From Fig. 8-5, it can be seen that B/A is most influential at short oxidation times and thin oxides, while B is dominant at elevated

temperatures and long times. The transition region between the linear and parabolic dependence is also shown. For this case, the full D-G Model given by Eq. 8-17 must be used when calculating x_{ox} .

Next the physical significance of the rate constants is examined. As discussed earlier, the constant B is related to the diffusion of the oxidant through the oxide and the constant B/A is related to the reaction rate at growing oxide interface. In the Deal-Grove study the values of B/A and B were extracted from extensive oxidation experiments over a broad range of temperatures for both dry and wet processes,⁴ and these B and B/A values were plotted on an Arrhenius plot (see Appendix B for a discussion of Arrhenius plots) to determine the activation energies for these constants. Figure 8-6 shows the parabolic rate constant B plotted on an Arrhenius plot for both wet and dry oxidation. The results can be written as:

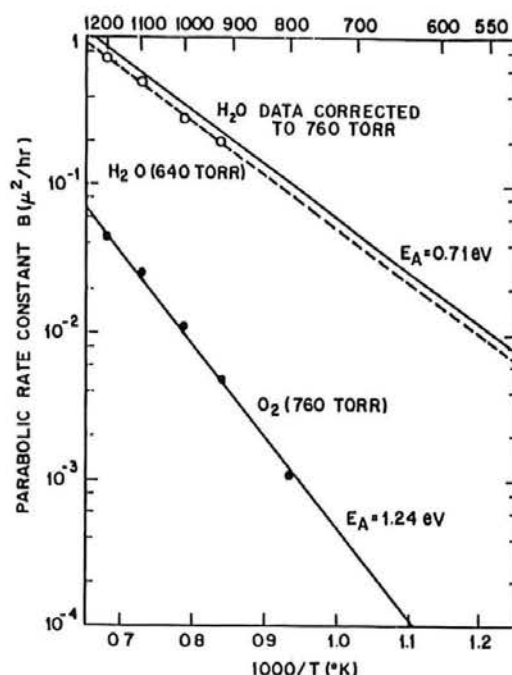
$$B = B_0 e^{-E_A/kT} \quad (8.25)$$

where B_0 is the pre-exponential term, E_A is the activation energy for the process, k is Boltzmann's constant, and T is the temperature, in K. The pre-exponential term is related to C^* and N_1 through Eq. 8-20. The values of activation energy obtained from the data in Fig. 8-6 are 1.24 eV for dry oxidation and 0.74 eV for wet oxidation. Since B and D (diffusivity of the oxidant through the oxide) are proportional (see Eq. 8-20), it is anticipated that the activation energy for both D and B should be similar. Unfortunately, there is no easy way to measure oxidant diffusivity in thin oxides so one must compare these results to those obtained from bulk fused-silica diffusion data. These bulk diffusivity values are available in the literature as 1.17 eV for O_2 and 0.80 eV for H_2O . These values are in good agreement with those obtained for B , verifying the model.

The linear rate constant B/A is now examined. Figure 8-7 shows an Arrhenius plot for the B/A constant. From the slope, the activation energies for dry and wet oxidation are calculated to be 2.0 eV and 1.96 eV, respectively. Unlike the parabolic case the activation energies are comparable for both wet and dry oxidation. When the activation energies for the different oxidants have similar values it is reasonable to assume both have the same controlling mechanism.

From Eq. 8-24 one finds that B/A is proportional to the reaction rate constant k_s . Assuming that free (unbonded) silicon is required to react with the oxidant, then it is reasonable to predict that the controlling mechanism for the reaction-rate constant (or the B/A constant) is the breaking of Si-Si bonds at the silicon surface. Once a bond is broken there is silicon available to react with the incoming oxidant to form SiO_2 . Under these conditions the activation energy for the process should be independent of the oxidant species, and dependent only on the energy necessary to break Si-Si bonds. To prove this conjecture the energy for Si-Si bond breaking was calculated to be 1.83 eV/molecule. This value is considered a good match to the measured value for B/A .

Tables 8-4 and 8-5 list measured values for A , B , B/A , and τ for wet oxidation and dry oxidation for several temperatures. Note that at any given temperature the B/A and B constants are much larger for wet oxidation than for dry oxidation. As a result, the oxidation rates for wet oxidation are much higher than for dry oxidation. There are two factors that contribute to the enhanced oxidation rates. The first is that C^* (the concentration of oxidant dissolved at the oxide), is much higher for H_2O than O_2 . ($C^*[\text{H}_2\text{O}] = 3 \times 10^{19}/\text{cm}^3$ and $C^*[\text{O}_2] = 5.2 \times 10^{16}/\text{cm}^3$). The second is that in the parabolic region the activation energy for diffusion of H_2O is considerably smaller than for O_2 . The smaller H_2O molecules can move more easily through the glassy network than can the larger O_2 molecules.



Note: Data is
for (111) Si

Fig. 8-6 The temperature dependence of the parabolic rate constant for dry and wet oxidations.⁴ Reprinted with permission of the American Physical Society.

Note also in Tables 8-4 and 8-5 that while $\tau = 0$ for wet oxidation when starting with bare silicon, it is non-zero for the case of dry oxidation when starting with bare silicon. The non-zero value of τ for dry-oxidation accounts for the fact that the Grove-Deal model under dry-oxidation underestimates the thickness of the oxide grown (as discussed in Sect. 8.4). Thus, the parameter τ must be adjusted in Eq. 8-17 (by graphically extrapolating the linear region back to the time axis, see Fig. 8-15a) to get good agreement between the predicted and measured values of oxide grown on bare silicon under dry oxidation conditions.

EXAMPLE 8-1: Calculate the step height in the silicon that would be measured at the edge of a window after the following sequence of process steps was performed. Use the rate constants given in Table 8-5.

1. A bare silicon wafer is oxidized in dry oxygen at 920°C for 10 minutes;
2. A window is etched into this oxide down to bare silicon;
3. The wafer is again oxidized in dry oxygen at 800°C for 30 minutes;
4. All the oxide is stripped.

SOLUTION: First calculate the amount of oxide grown in Step 1. To do so, employ the D-G model given by Eq. 8-17:

$$x_{\text{ox}}^2 + Ax_{\text{ox}} = B(t + \tau)$$

Referring to Table 8-5, obtain the values of the constants, A, B, and τ at 920°C.

$$A = 0.235 \mu\text{m}; B = 0.0049 \mu\text{m}^2/\text{hr}; \tau = 1.4 \text{ hr}; \text{ and } t = 10 \text{ min} = 0.167 \text{ hr}$$

Next put the constants into Eq. 8-17 and solve the quadratic equation for the value of x_{ox}

$$x_{\text{ox}}^2 + 0.235x_{\text{ox}} = 0.0049(1.4 + 0.167)$$

$$x_{\text{ox}}^2 + 0.235x_{\text{ox}} - 0.0049(1.4 + 0.167) = x_{\text{ox}}^2 + 0.235x_{\text{ox}} - 0.00768 = 0$$

or:

$$x_{\text{ox}} = 0.0291 \mu\text{m} = 291 \text{ \AA}$$

After the window is etched bare silicon is within the window and 291 Å of oxide around the window. The wafer is again oxidized at 800°C for 30 min in dry oxygen. Now calculate the amount of oxide that will grow on the bare silicon in the window and again employ Eq. 8-17 and Table 8-5. At 800°C:

$$A = 0.370 \mu\text{m}; B = 0.0011 \mu\text{m/hr}^2; \tau = 9 \text{ hr}; \text{ and } t = 30 \text{ min} = 0.5 \text{ hr}$$

$$x_{\text{ox}}^2 + 0.370x_{\text{ox}} = 0.0011(9 + 0.5)$$

Rearranging terms and applying the quadratic formula one obtain the value of x_{ox}

$$x_{\text{ox}} = 0.0263 \mu\text{m} = 263 \text{ Å}$$

Next calculate how much additional oxide has grown around the window during the 30 min oxidation at 800°C. To calculate the additional oxide grown, first find the equivalent time it would have taken if the original 291 Å oxide was grown at 800°C, and then add that to that the additional 30 min of Step 3 to find the new thickness of oxide around the etched window. Using Eq. 8-17 to calculate the time to grow 291 Å at 800°C:

$$(0.0291)^2 + 0.370(0.0291) = 0.0011(t + 9)$$

Solving for t

$$0.0011t = (0.00085 + 0.01077 - 0.00990); \text{ or } t = 1.56 \text{ hr}$$

Now calculate the total thickness grown at 800°C after the additional 0.5 hr. Applying Eq. 8-17 at 800°C, but for a total time of 1.56 hr + 0.5 hr (or 2.06 hr), we obtain:

$$x_{\text{ox}}^2 + 0.37x_{\text{ox}} = 0.0011(2.06 + 9)$$

Rearranging and solving for x_{ox} the final oxide thickness is found to be;

$$x_{\text{ox}} = 0.030 \mu\text{m} \text{ or } 300 \text{ Å}$$

Next calculate the additional oxide that was grown around the window

$$\Delta x_{\text{ox}} = 300 \text{ Å} - 291 \text{ Å} = 9 \text{ Å}.$$

To grow the 9 Å, $0.44(9 \text{ Å}) = 4 \text{ Å}$ of silicon is consumed. Inside the window 263 Å was grown and $0.44 \times 263 \text{ Å} = 155.7 \text{ Å}$ of silicon was consumed. The step height after stripping the final oxide is thus the difference in the amount of silicon consumed, or:

$$\text{Step height on the silicon} = 155.7 \text{ Å} - 4 \text{ Å} = 151.7 \text{ Å}.$$

8.3 FACTORS WHICH AFFECT THE ONE-DIMENSIONAL OXIDATION RATE

In this section other factors that affect the oxidation rate will be discussed. It is important to understand these effects when attempting to model oxidation kinetics. The oxidation rate of Si depends on various parameters, with the following being the most important: a) temperature; b) pressure; c) oxidant; d) silicon crystallographic orientation; e) dopant type and dopant concentration; f) the presence of a chlorine gas in the growth ambient; and g) the partial pressure of water present during dry oxidation. The effect of temperature (thermal activation) and the oxidant were described in detail above. This section discusses the effects of the others. All of these effects can be modeled using such process simulators as SUPREM3 and SUPREM4.

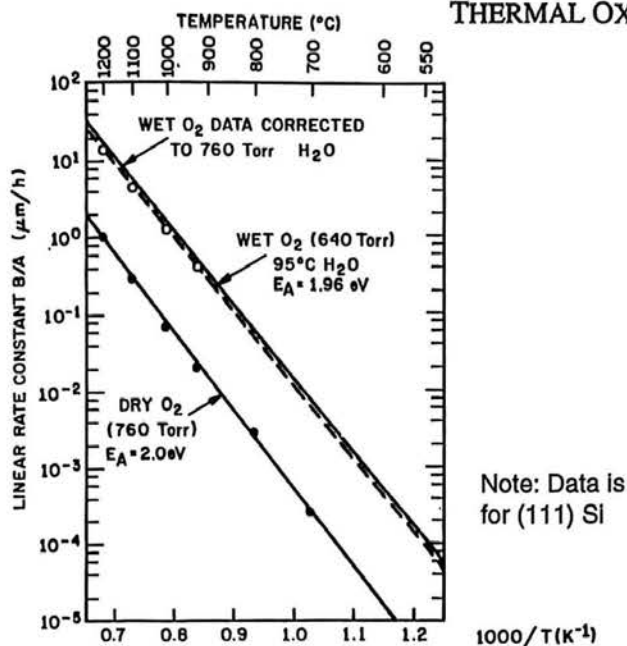


Fig. 8-7 The temperature dependence of the linear rate constant for dry and wet oxidation.⁴ Reprinted with permission of the American Physical Society.

8.3.1 Crystal Orientation Effects on Oxide Growth Rates

Although most ULSI device processing utilizes (100) silicon substrates, there are applications where other crystal orientations are oxidized. For example, during the oxidation of a trench, the side walls of the trench are generally not along $\langle 100 \rangle$ planes. Consequently, one needs to understand the effect of crystal orientation on the oxidation rate to be able to model such oxide growth conditions.

The parabolic rate constant B has been observed to be relatively independent of the crystal orientation. This is not surprising, since this constant depends on the diffusivity of the oxidant

Table 8-4 RATE CONSTANTS FOR WET OXIDATION OF SILICON

Oxidation Temperature (°C)	A (μm)	B (μm ² /hr)	B/A (μm/hr)	τ (hr)
1200	0.05	0.720	14.40	0
1100	0.11	0.510	4.64	0
1000	0.226	0.287	1.27	0
920	0.50	0.203	0.406	0

Table 8-5 RATE CONSTANTS FOR DRY OXIDATION OF SILICON

Oxidation Temperature (°C)	A (μm)	B (μm ² /hr)	B/A (μm/hr)	τ (hr)
1200	0.040	0.045	1.12	0.027
1100	0.090	0.027	0.30	0.067
1000	0.165	0.0117	0.071	0.37
920	0.235	0.0049	0.0208	1.40
800	0.370	0.0011	0.0030	9.0
700	—	—	0.00026	81.0

Table 8-6 RATE CONSTANT DEPENDENCE ON TEMPERATURE AND CRYSTAL ORIENTATION FOR WET OXIDATION

Oxide Temperature (°C)	Orientation (hkl)	A (μm)	B (μm ² /hr)	B/A (μm/hr)	Ratio (111)/(100)
900	100	0.95	0.143	0.150	1.68
	111	0.60	0.151	0.252	
950	100	0.74	0.231	0.311	1.68
	111	0.44	0.231	0.525	
1000	100	0.48	0.314	0.664	1.75
	111	0.27	0.314	1.163	
1050	100	0.295	0.413	1.400	1.65
	111	0.18	0.415	2.307	
1100	100	0.175	0.521	2.977	1.65
	111	0.105	0.517	4.926	

through the existing oxide and not the reaction at the Si/SiO₂ interface. On the other hand, the linear rate constant (B/A) is found to depend strongly on the crystal orientation. Table 8-6 lists values for several temperatures of A , B and B/A and the B/A ratio between (111) and (100) Si for wet oxidation. Over most of the temperature range the B/A constant for (111) is found to be 1.68 times the B/A for (100). Using very dry oxides, Irene found that the oxidation rate was highest for (110) followed by (111) and then (100).⁵ These results are similar to those obtained for steam oxidation by Ligenza⁶ and Pliskin.⁷ In a more recent study the ratio of the reaction rate constant on (110) surfaces and (100) was found to follow the relationship: $k_{s(110)} = 1.45 k_{s(100)}$.⁸

Figure 8-8 shows a plot of wet oxidation rate curves for both the (100) and (111) orientations.⁴ Two important points are illustrated. First, as the temperature increases, the orientation effect is diminished. At high temperatures the parabolic growth dominates, even at short times. Thus, there is little orientation dependence. Second, for very long times even at low temperatures the curves merge (since the parabolic regime begins to dominate the total oxide thickness).

Ligenza⁶ proposed a model to explain the orientation dependence for wet oxidation and Irene⁵ applied a similar model to dry oxidation. The basis of the models is that in the linear region the oxidation rate should depend on both the number of Si bonds per unit area available for reaction and the bond orientation relative to the silicon surface. The average number of silicon atoms in a unit area is greater for (111) than for (100). In addition, since the oxidants are thought to be molecular it is possible that these rather large molecules can shield additional incoming molecules from reacting with the Si surface. The bonds that make a larger angle with the surface have a less favorable position for the reaction to take place, which leads to an effectively larger activation energy. This shielding is termed *steric hindrance*. The (111) plane has more favorable bond angles (less steric hindrance) and a higher value for B/A .

8.3.2 Dopant Effects on Oxidation Growth Rates

The commonly used Group III (B) and Group V (P, As) dopants are known to enhance the oxidation rate when present in high concentrations at the silicon surface. However, the mechanism responsible for the increased rate is different depending on the particular dopant. In the case of boron (B), where the dopant tends to pile up in the growing oxide ($m < 1$), a high concentration of boron is incorporated in the oxide. This weakens the structure by increasing the

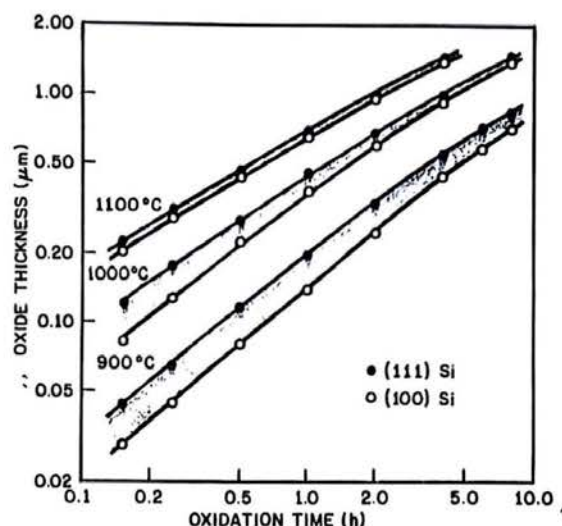


Fig. 8-8 Silicon dioxide thickness as a function of oxidation time for silicon in H_2O at 640 torr.⁹ Reprinted by permission of publisher, the Electrochemical Society.

number of non-bridging oxygen atoms. The weakened oxide provides a path for the oxidant to more easily diffuse through the oxide. More rapid diffusion should result in an increase in the parabolic rate constant, but should have no effect on the linear regime. Figure 8-9 shows a wet oxidation curve for three different boron concentrations.¹⁰ An increased oxidation rate for all temperatures and times is observed for $C_{\text{boron}} > 1 \times 10^{20}/\text{cm}^3$. This behavior is expected for an increase in the B constant. It is interesting to note that other Group III dopants such as Ga and Al which also pile up in the oxide, but rapidly diffuse away from the interface, do not cause the enhanced oxidation effect.

An enhanced oxidation rate is also observed for silicon heavily doped with phosphorous (P). In this case the mechanism is different from that occurring above. When phosphorus doped Si is oxidized the P piles up at the silicon surface ($m > 1$) and very little is incorporated into the oxide. Thus, one expects the heavy silicon surface concentration to affect only the linear rate constant and not the parabolic rate. This is indeed the case. In Fig. 8-10, a plot for wet oxidation at several P doping levels is shown. At low temperature, there is a very large increase in growth rate but the enhancement diminishes as the temperature increases.¹¹ This behavior is expected since the linear rate constant dominates at low temperature and shorter times. Fig. 8-11 shows the dependence of both the parabolic and linear rate constants as a function of the P doping level. Note that the B/A constant rapidly increases at high P concentration, while the B constant shows only a slight increase with heavy doping.

A model has been proposed to explain this behavior.¹² It is predicated on the conjecture that the high concentration of P at the surface causes a shift in the Fermi level, which in turn increases the surface Si vacancy concentration. The excess vacancy concentration increases the oxidation rate in the linear regime by providing additional sites for reaction of the oxidant with the silicon. There is some controversy about whether Si *vacancies* or *interstitials* are the primary defect formed during oxidation in the presence of heavily P-doped Si. Massoud *et al.*, have shown that in the initial stages of oxidation the presence of a heavily doped P surface results in only a moderate enhancement of the oxidation rate.¹³ They attribute this to the possibility that the controlling mechanism for thin oxidation may not actually involve the silicon surface.

During oxidation, the oxide thickness on the wafer surface over heavily doped regions may be greater than over lightly doped regions. Such variations in oxide thickness must be taken into

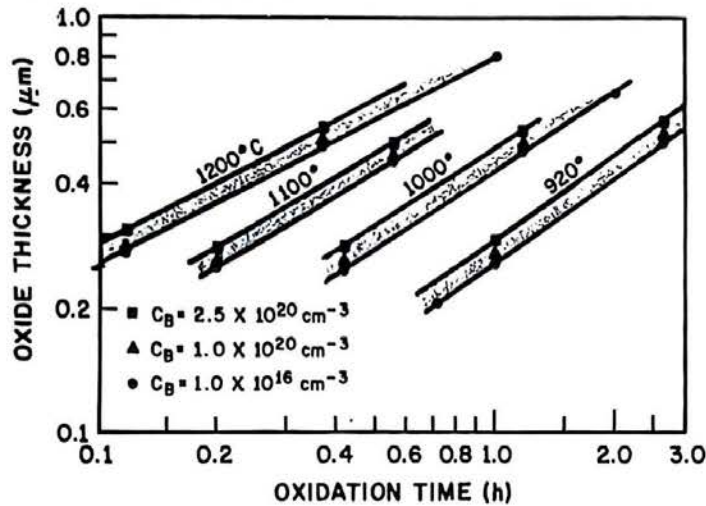


Fig. 8-9 Silicon dioxide thickness as a function of oxidation time in wet oxygen (95°C H₂O) for boron-doped silicon.¹⁰ Reprinted by permission of publisher, the Electrochemical Society.

account when specifying an etch process to insure that the thickest regions on the device are completely removed.

8.3.3 Water Effects During Dry Oxidation

Since the presence of water during dry oxidation can greatly increase the growth rate, and also increase the trap density, water should be treated as an impurity during this process. Irene has shown that as little as 25 ppm of water intentionally added to very dry oxygen increased both the linear and parabolic rate constants when compared against the baseline of 1 ppm.⁵ However, in practice it is extremely difficult to exclude all traces of water during dry oxidation.

The presence of water vapor during oxidation can arise from several sources: a) adsorbed water on the incoming wafers; b) residual water in the oxygen source gas; c) residual hydrogen or hydrocarbons which react with oxygen at elevated temperatures to form water; d) the diffusion of water vapor from the fab environment into the oxidation furnace; and e) from the reaction of intentionally added hydrogenated chlorine sources (e.g., HCl, TCA) that combust with oxygen to form water.

Stringent steps can be taken to exclude water from fab oxygen, however in practice this is not done. For example, the residual hydrogen and hydrocarbons in the oxygen can be removed by passing the oxygen through a pre-burner to induce combustion of these materials to form

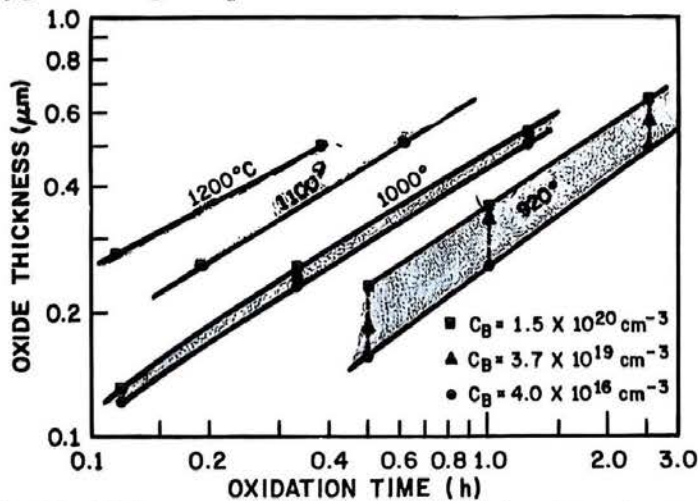


Fig. 8-10 Silicon dioxide thickness as a function of oxidation time in wet oxygen (95°C H₂O) for phosphorus-doped silicon.¹¹ Reprinted with permission of the Electrochemical Society.

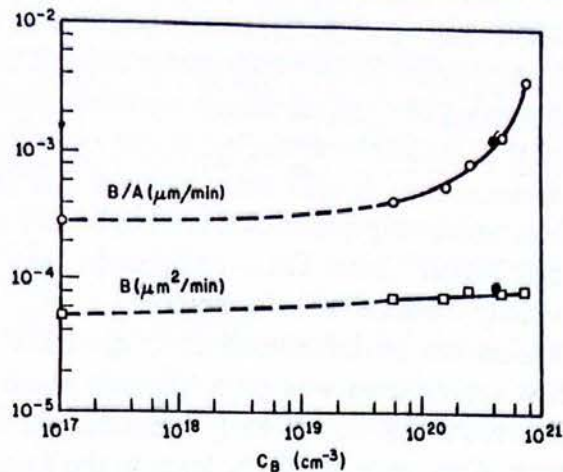


Fig. 8-11 Oxidation rate constants for dry oxidation of silicon as a function of phosphorus doping level at 900°C.¹¹ Reprinted with permission of the publisher, the Electrochemical Society.

water and then through a drier or cold trap to remove the water. In general, the purest fab oxygen (from a liquid oxygen tank) is supplied to the oxidation system. The water *backstreaming* from the fab environment can be minimized by using double-wall furnace tubes and flowing an inert gas between the walls to sweep out the water. Recently, some furnaces systems and rapid thermal oxidation systems have incorporated a loadlock (that is nitrogen-purged or kept under vacuum) to keep water vapor from entering the high temperature portion of the reactor. It should be noted that in production environments it is not necessary to employ such water reduction procedures. As long as reasonable care is exercised to maintain a reproducible level of water and as long as the maximum level remains in the 25–50 ppm range, high quality, dry oxides can be produced.

8.3.4 The Dependence of Oxidation Rates on Chlorine

Chlorine (Cl) gas mixed with the oxygen during all or part of the oxidation cycle causes two things to happen. First, during dry oxidation the chlorine increases the oxidation rate. Second, and more beneficially, by introducing Cl several key material properties of the oxide are improved. The reported improvements include: a) reduction in mobile ionic charge in the oxide (due to Na); b) an increase in the *minority carrier lifetime* (MCLT) in the underlying silicon; c) reduction in the number of oxide defects and improvement in the oxide breakdown characteristics; d) reduction in the interface and fixed charge density; and e) reduction in the number of oxidation-induced stacking faults in the underlying silicon. The chlorine reacts with metallic impurities (e.g., Na, Fe) on the wafer during oxide growth, to form volatile products which are swept out in the exhaust. This is important, since Na contamination is known to adversely affect the stability of a device and transition metal impurities such as Fe are known to decrease the gate oxide integrity (GOI) and minority carrier lifetime. It should be pointed out that excellent quality gate oxides can be produced without the use of chlorine if careful pre-furnace wafer cleaning and wafer handling techniques are employed. It is interesting to note that chlorinated dry oxidation for gate oxide production is primarily used in U.S. fabs, while Japanese fabs have tended to use wet oxidation without Cl, for the gate oxide process.¹⁴ Note that the chlorine provides a level of insurance against poorly cleaned or heavily contaminated wafers.

Since chlorine can be beneficial, most furnace systems are equipped with the capability to introduce some chlorine species into the growing oxide. Even when Cl oxidation is not used it is common practice to clean the furnace tube *in situ*, using Cl and steam at elevated temperatures.

The direct introduction of Cl_2 gas into the furnace tube is not the preferred Cl source. The use of gaseous HCl is still the predominant method for chlorine introduction, although many device manufacturers have switched to the carbon-containing (chlorocarbon) precursors in order to avoid corrosion problems associated with HCl delivery to a furnace.¹⁵ The chlorocarbon precursors are *volatile* liquids at room temperature and need to be stored in a clean fused silica ampoule and delivered through Teflon® lines. These compounds come pre-packaged in ultra-high purity ampoules from specialty chemical manufacturers.¹⁶

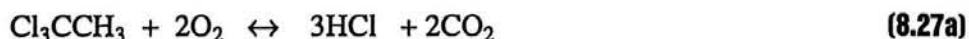
As previously stated, chlorine can be introduced using a variety of precursors. From a historical perspective, the first source used was pure Cl_2 gas. There were severe problems associated with delivery of Cl_2 since it was easy to carry corrosion products from the tank or the stainless steel lines to the wafer. Also, very small Cl_2 leaks in the furnace area could result in severe corrosion of the expensive process tools in the adjacent areas.

The next source (which is still widely used), was anhydrous HCl. Unlike Cl_2 , HCl contains hydrogen, which reacts with the oxygen during oxidation (combustion) to form Cl_2 and H_2O . The equilibrium reaction is given by Eq. 8-26. It is also critical to ensure there is an adequate supply of O_2 , otherwise it is possible that the silicon surface will be etched with the unreacted HCl. If silicon etching does occur it tends to roughen the surface and lead to reduced gate oxide integrity. HCl, like Cl_2 , can react with the stainless steel lines in the presence of water vapor, and thereby carry contamination onto the wafer surface. Great care must be exercised to ensure that delivery lines are leak tight and the storage tank is changed on a regular basis.



The water formed by the reaction in Eq. 8-26 can increase the oxidation rate. However, Hess *et al.*, show that the water made by the reaction is not adequate to explain the increased oxidation rate.¹⁷ It appears that the Cl_2 itself increases the oxidation rate. In their study they found that the linear rate constant increased by a factor of 2, when the HCl concentration increased from 0 to 3%. The Cl_2 enhancement mechanism is still not well understood.

Since Cl_2 and HCl can carry corrosion products to the wafer, alternative Cl precursors have been investigated. Much interest has focused on the chloroorganic sources, which also form free Cl_2 when they combust with oxygen. Eqs. 8-27a and 8-27b represent the reaction of TCA ($\text{C}_2\text{H}_3\text{Cl}_3$) with oxygen to form Cl_2 , H_2O , and CO_2 .



The first chlorocarbon source used was trichloroethylene (TCE). This material worked admirably as a Cl_2 source, but it was found to be a *carcinogen* (cancer causing agent) and was thus banned from fabs. A new source, trichloroethane (TCA) quickly replaced TCE and was used for more than 10 years. Recently TCA was identified as an *ozone depleting* gas (also termed *global warming* gas) and its production, uses, and/or transportation have been severely restricted, and even banned in some of the countries that are signatories to the Montreal Protocols.¹⁸ The move away from TCA was accelerated by the fact that there is a dwindling supply of semiconductor grade TCA and many IC manufacturers want to employ environmentally friendly practices. The problem with TCA is its long atmospheric half life (6.2 years) and its ability to reach the ozone layer. Consequently, the industry has been looking for alternatives. Two promising compounds have been explored, namely, trans-1,2-dichloro-ethane (DCE) and oxalyl chloride (OC), and both are commercially available in high purity form.

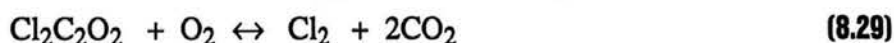
Table 8-7 SOME PROPERTIES OF THE CHLORINE PRECURSORS

Chlorine Precursor	Common Name	Chemical Formula	Atmos. Half-Life	Vapor Pressure @ 20°C (Torr)	Flash Point (°C)	Why (or Why Not) Use
Chlorine	Cl ₂	Cl ₂			None	Corrosion of delivery lines
Hydrogen Chloride	HCl	HCl			None	Corrosion of delivery lines
Trichloroethylene	TCE	Cl ₃ CCH			None	Carcinogen
1,1,1,Trichloroethane	TCA	Cl ₃ CCH ₃	6.2 years	104	None	Global warming gas
Trans-1,2 Dichloroethane	DCE	Cl ₂ CCH ₂	3.6 days	273	< 2	Good substitute
Oxalyl Chloride	OC	Cl ₂ C ₂ O ₂	4 min	173	None	Better substitute?

The combustion reaction for DCE is shown in Eqs. 8-28a and 8-28b



Although DCE has made the largest penetration as a TCA replacement, there are good reasons why OC is considered a better alternative. One is that OC (unlike TCA and DCE), does not contain hydrogen, and therefore does not produce HCl when combusted. With no HCl present, no water is formed, and all the Cl in the OC is liberated as Cl₂ (which is the active species required for cleaning). The reaction of OC with oxygen is shown in Eq. 8-29:



As it turns out only 10–50% of the total chlorine liberated by DCE combustion is present as Cl₂ and the remainder as ineffective HCl. Since OC forms Cl₂ directly, without forming HCl, more Cl is incorporated into the growing oxide for equivalent amounts of the precursor in the oxygen. Both OC and DCE are effective when removing K and Ca metallic impurities from intentionally contaminated wafers. On the other hand, Al and Ti, which tend to form oxides, are not removed by either compound.¹⁹ Table 8-7 lists some properties of the Cl precursors described above.

8.3.5 Effect of Pressure on Oxidation Rates

In this section the effect of pressure on oxide growth kinetics is covered. The D-G Model predicts a linear dependence of both the *B* and *B/A* constants. While *B/A* does increase linearly with pressure, the parabolic constant follows a $P^{0.6-0.8}$ dependence. Several models have been proposed to explain the less than linear dependence, but none fully accounts for all of the experimental evidence. Since there are applications where oxides are grown at either high or low pressure, it is necessary to understand the growth kinetics in both of these regimes.

8.3.5.1 High-Pressure Oxidation (HIPOX): *High-pressure oxidation* is used when it is necessary to increase the growth rate at a fixed temperature or keep the growth rate constant at a lower temperature. By growing the oxide at a lower temperature (or for a shorter time) the *thermal exposure* (or *thermal budget*) that the wafer experiences, is reduced. Other advantages of using HIPOX are discussed later in the chapter. HIPOX has been routinely used (in the pressure range of 10–25 atm) to grow LOCOS field oxide, STI trench liner, and for other applications using

either dry oxygen or steam as the oxidant. High pressure oxidation using nitrous oxide (N_2O) to grow oxynitrided gates has also been reported.²⁰

The use of high pressure increases the oxidation rate by increasing both the parabolic and linear rate constants. The increase in the rate constants is expected since C^* depends on the partial pressure of the oxidant. Fig. 8-12 shows growth curves for wet oxidation at 900°C .²¹ The growth rate increases rapidly as the pressure increases. For example, a 120-nm oxide can be grown in 60 minutes at 1 atm on (100) silicon, while at 20 atm and also in 60 minutes, a 1200-nm oxide can be grown. It is also possible to maintain the same growth rate at a lower temperature by increasing the pressure. A good rule of thumb to estimate the temperature-pressure trade-off is: *a constant growth rate can be maintained, if for each increase in pressure of 1 atm, the temperature is decreased by about 30°C* . HIPOX process tools and some applications where HIPOX has advantages will be discussed in a subsequent section.

8.3.5.2 Low Pressure Oxidation: Reduced pressure oxidation is mainly used when it is necessary to decrease the oxidation rate. For example, to grow ultra-thin gate oxides in a furnace one needs to reduce the oxidation rate to better control the process. If the rate is too rapid the oxidation time can become too short for needed control. By lowering the oxidation temperature the oxidation time can be increased. The problem with low temperature oxidation is that it results in oxides of poorer quality than those produced at high temperature. Another way to slow down oxidation is to maintain the temperature but reduce the partial pressure of oxygen or steam in the furnace. There are two ways to accomplish reduced pressure in a furnace system. The first requires sealing the fused silica tube and evacuating it with a mechanical pump. Thus, the furnace is operated at a reduced pressure. The second is to flow oxygen *diluted* with an inert gas such as argon or nitrogen. For example, by diluting oxygen to 10% in N_2 , a 3.0 nm oxide can be grown at 850°C in 10 minutes.²² This is an adequate time for a controllable process.

Figure 8-13 shows the oxide growth rate normalized to growth at 1 atm, as a function of oxygen partial pressure (in the range 10^{-2} –1 atm) for oxides grown in mixtures of dry O_2 diluted with argon. The oxidation rate dependence on partial pressure is given in Eq. 8-30:

$$\text{Oxidation rate} \propto P^{0.8} \quad (8.30)$$

Note that a $P^{0.8}$ dependence is similar to the dependence reported by other investigators studying oxidation at elevated pressure.²³

8.4 THE INITIAL OXIDATION STAGE AND THE GROWTH OF THIN OXIDES

Since the D-G Model is not accurate for oxide thicknesses smaller than 30 nm, and since the majority of gate oxides grown in ULSI devices are less than this, there has been an enormous amount of study dedicated to understanding the growth kinetics in the thin oxide regime. The D-G model underestimates the oxide growth rate for *dry oxides*, but accurately predicts the growth for *steam oxides*. Figure 8-15a shows the region of thin oxide growth in dry oxidation, and how a linear curve must be extrapolated back to the time axis to extract τ . In this section some of the models that have been proposed to explain the enhanced oxidation in the initial dry oxide growth regime are described. After more than 30 years of study, a good understanding of the mechanism(s) that control the initial regime is still a subject of keen debate. Fortunately, the enhanced growth rate can be accounted for empirically, and correction terms have been developed. When these are added to the D-G model, relatively good estimates of the experimental values of all oxide thicknesses are obtained. Despite the controversy surrounding

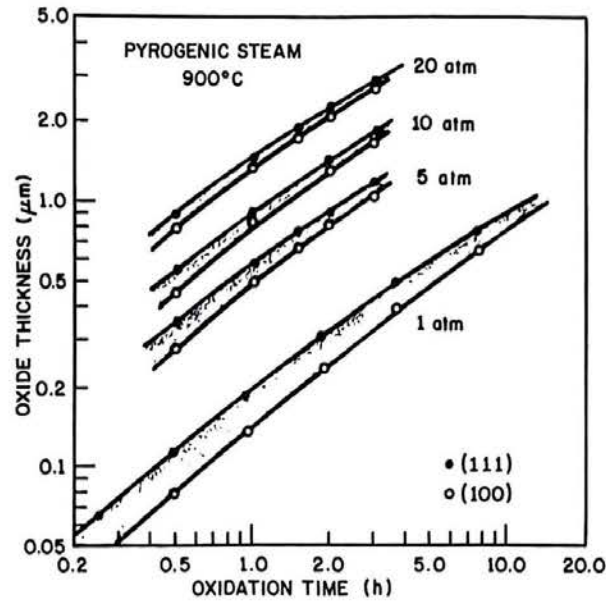


Fig. 8-12 Silicon dioxide thickness as a function of oxidation time and pyrogenically produced steam pressures at 900°C.¹⁹ Reprinted by permission of the publisher the Electrochemical Society.

the mechanisms, process engineers have been able to reproducibly grow high quality gate oxides as thin as 3.0 nm in the manufacturing environment. As discussed earlier a *native oxide* will grow on a freshly cleaved silicon surface when that surface is exposed to a room ambient. Note that the native oxide growth is not continuous, but progresses in a stepwise manner. The growth also tends to saturate at about 0.8 nm for lightly doped silicon and 1.3 nm for heavily doped silicon. The growth of a native oxide is depicted in Fig. 8-14.²⁴ It is believed that the steps correspond to the growth of a single atomic layer. The more rapid growth for the higher doping levels is expected from the enhanced growth rate for high levels of doping.

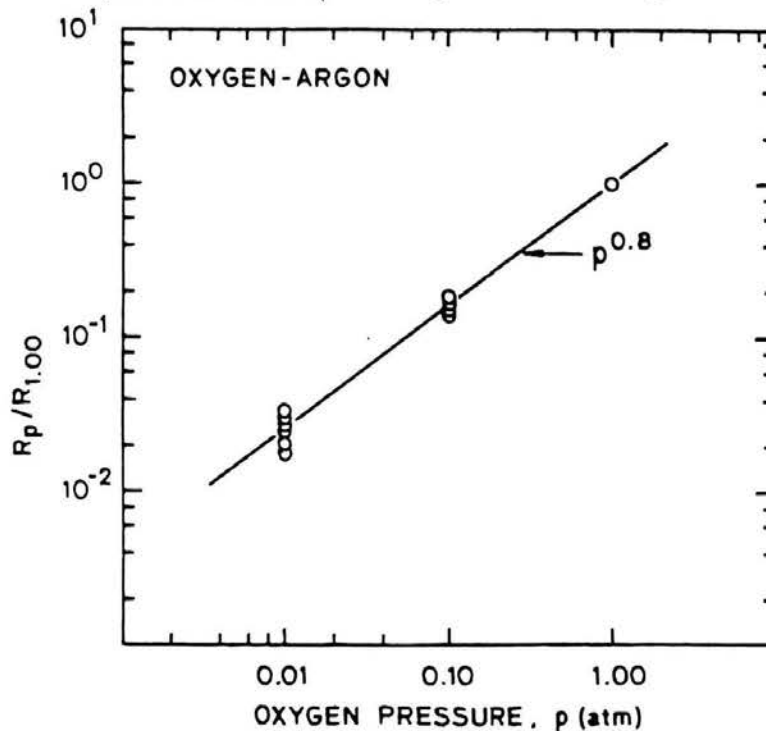


Fig. 8-13 Normalized oxidation rate as a function of oxygen partial pressure for (100), (110), and (111), lightly doped silicon in dry O_2 -Ar mixtures at 800–1000°C.¹³

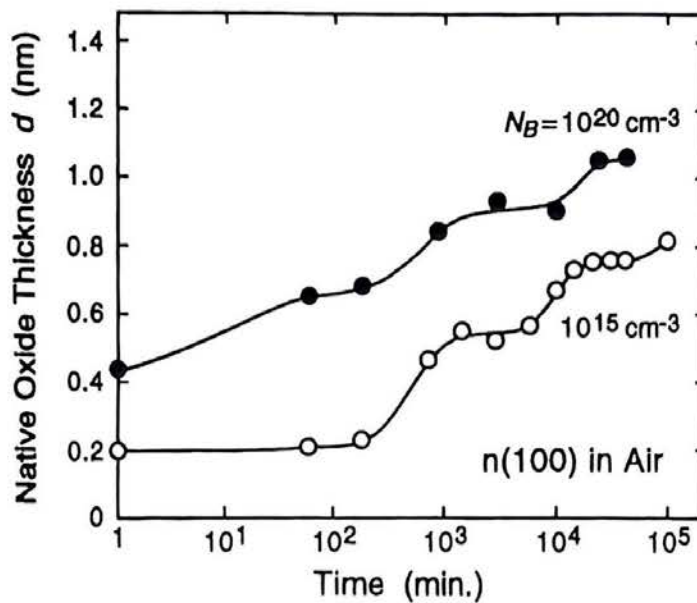


Fig. 8-14 Growth of a native oxide film on exposure to air at room temperature at 42% RH.¹⁴

The kinetics of oxidation in the thin film regime are known from experimental findings to depend on many parameters, some of which do not affect the growth of thicker oxides. The most important parameters are: a) the nature of the chemical cleaning performed prior to oxidation; b) the formation of a sub-oxide (e.g., SiO) at the interface; and c) the amount of interstitial oxygen dissolved in the substrate. Of course, other factors discussed previously also affect oxidation in the thin region. There has been excellent experimental work performed in the growth of thin oxides and several excellent critical review articles exist on this subject. Readers are referred to them for more details.^{25,26}

8.4.1 Experimental Results for Thin Oxide Growth

Massoud *et al.*, monitored the growth rate of thin oxides by placing an automated high-temperature ellipsometer in the furnace (*in-situ*). In this way they were able to measure the increase in oxide thickness with time during growth. These measurements were performed over the temperature range of 800–1000°C, for the pressure range of 0.01–1.0 atm, and for (100), (110), and (111) crystal orientations. Fig. 8-15b shows a plot of the oxidation rate of (100) silicon as a function of the oxidation thickness. Also shown for comparison (solid line) is the curve given by the D-G model. For thicknesses below 25–30 nm (250–300 Å) the oxidation rate is increased over that predicted by the D-G model.

The authors then plotted the excess oxidation rate (above the predicted D-G rate), against the oxide thickness and found that the data could fit an expression that has two distinct regions. In the first region, the excess oxidation rate was found to decay exponentially with thickness with a characteristic decay length of about 5.0 nm (50 Å). In the second region the excess rate could also be fitted to an exponential term that decays exponentially with thickness and has a characteristic length of 7.0 nm. From these data they were able to fit all the growth rate to an analytical expression given by:

$$\frac{dx_{ox}}{dt} = \frac{B}{2x_{ox} + A} + C_1 e^{-x_{ox}/L_1} + C_2 e^{-x_{ox}/L_2} \quad (8.31)$$

The first term on the right side of the equation is just the time differential of the D-G equation, while the other two terms are related to the excess rate. After further analysis they found that the

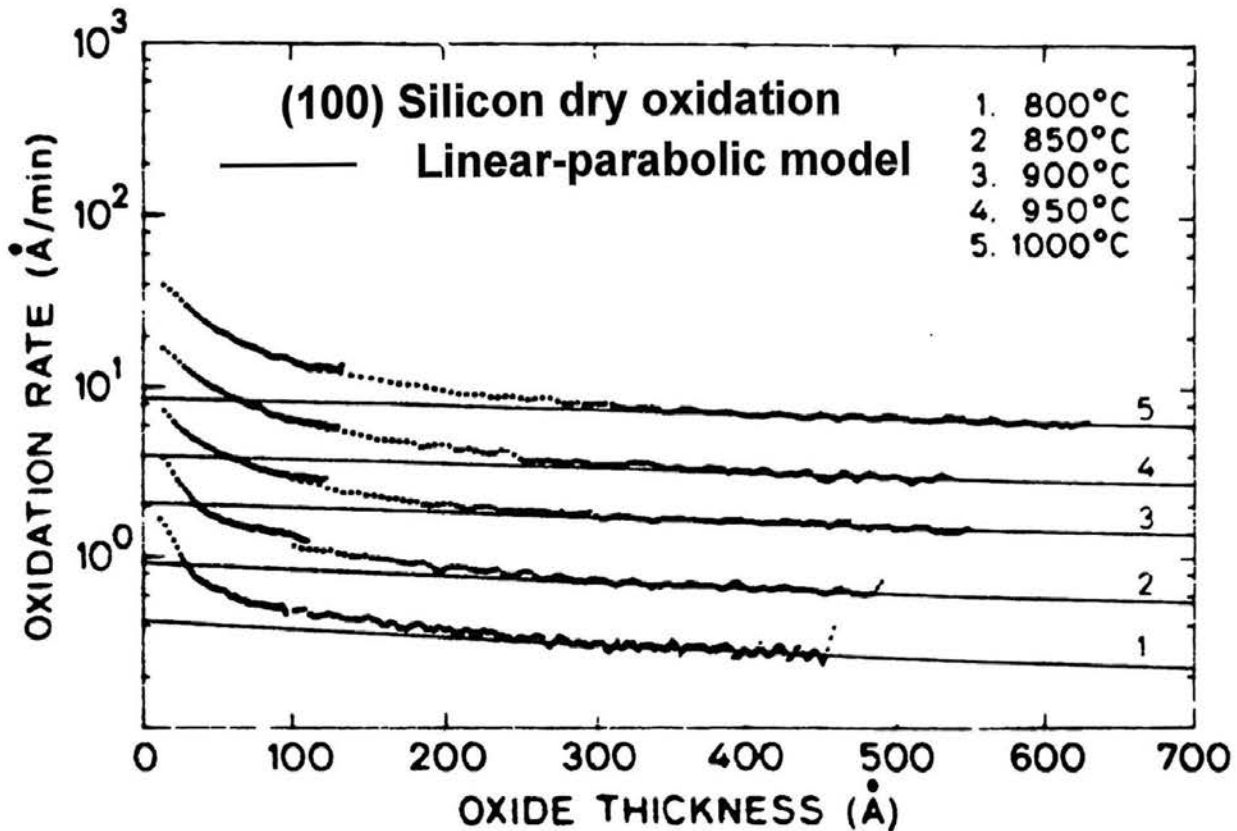


Fig. 8-15 a) Thickness of thin dry oxides versus time showing a portion of the curve not fitting the D-G model. b) Oxidation rate as a function of oxidation thickness for lightly doped (100) silicon in dry oxygen at 800–1000°C. The solid lines represent the linear/parabolic (or D-G) model.¹³

first exponential term could be eliminated from the rate equation and, as a result, introduce only about a 5% error in the first 5–10 nm of growth. In addition they discovered the following dependencies for C_2 and L_2 .

1. The characteristic length L_2 is independent of temperature in the range of 800–1000°C, and has a value of 6.9, 7.8, and 6.0 nm for (100), (110), and (111), respectively.
2. The constant C_2 follows an Arrhenius temperature dependence with an activation energy of 2.37 eV and 2.32 eV for (100) and (111), respectively, and an activation energy of 1.80 eV for (110) silicon.

In the second paper of Massoud *et al.*, some of the models that have been previously advanced to explain the enhanced oxidation rate were examined.¹³ The existing models were classified into the following areas: a) space charge effects where the enhancement is electro-chemical in nature; b) oxide structural defects that allow enhanced diffusion of the oxidant; c) stress effects in the oxide that enhance the transport of the oxidant and the reaction rate at the interface; and d) oxygen solubility enhancement (where the growth would be affected by oxygen concentrations exceeding the solid solubility limit). Their arguments for and against each of these models are lengthy and the reader is referred to their paper for details. Since none of the models fit their data, they proposed a new model. This model postulates that the silicon contains

additional sites for oxidation. It assumes that the concentration of the sites decreases exponentially into the silicon and has a decay length of 3 nm. The model goes on to assume that these additional sites are responsible for the enhanced oxidation in the early stages of growth. Once the oxide has reached the thickness where the additional sites are consumed, the oxidation rate returns to the predicted D-G behavior. The characteristic length of 3 nm is calculated from the enhanced oxide growth characteristic length which is measured to be 7 nm. In order to grow 7 nm oxide, 3 nm of silicon must be consumed.

In a later paper, Massoud proposes that the presence of a *blocking layer*, perhaps SiO, exists at the interface during the initial stages of oxidation. The blocking layer limits the diffusion of oxygen to the interface and prevents the formation of SiO₂.²⁷ As additional oxygen reaches the blocking layer, the oxygen concentration continues to increase up to the point where SiO₂ formation is preferred. When this point is reached the SiO and Si are rapidly transformed to SiO₂. This rapid transformation is responsible for the anomalously high growth rate.

Mott *et al.*, also proposed a new model to explain the initial fast growth rate.²⁵ These authors proposed that the initial rate is not really enhanced but rather retarded. They speculated that all oxidation is limited by diffusion of the oxidant and there is no surface-reaction-limited regime, as postulated by D-G. During the initial stages of oxidation there is an *altered oxide layer* adjacent to the silicon. This altered layer may be too thin (or even nonexistent) during wet oxidation, and that is why the process is not operative in that case. They then propose that oxygen is easily incorporated into the network of the altered oxide, but the molecular oxygen does not readily penetrate through the layer to reach the silicon surface. The O₂ in the network continues to increase and by doing so generates a stress in the layer. At some point the stress is relieved by a viscoelastic relaxation mechanism and oxidation then proceeds by the diffusion mechanism.

8.4.2 Growing Thin Oxides

There are several known approaches to growing thin oxides for high volume production processes. Two of them, namely low temperature growth and reduced pressure growth, have already been discussed earlier in this chapter. A third approach, *rapid thermal oxidation (RTO)* can be used to grow well-controlled thin and ultra-thin oxides (1.3–3.0 nm). Thin RTO oxides have a distinct advantage over furnace grown oxides since they can be prepared at relatively high temperatures, but for very short times. High temperature oxides have been found to have a smoother silicon interface and improved breakdown characteristics.²⁸ A 4 nm RTO oxide can be grown in pure oxygen at 1 atm and 1050°C in about 40 sec. The use of RTO oxides is now becoming a production worthy and tested process.

8.5 THE NATURE OF THE Si-SiO₂ INTERFACE

The interface between SiO₂ and silicon (and the silicon dioxide itself) contain both charges and traps. These charges and traps have a profound effect on the properties of the devices fabricated in the underlying Si. In MOS devices, for example, the behavior of the transistor is dominated by the thin gate oxide grown on the silicon. In this section we describe the nature of the Si/SiO₂ interface, the types of charges and traps, and the effect of these on the device parameters and reliability. For more information on MOS device operation, readers are referred to Vol. 3.

Electrical measurements on the actual device structures are the most effective method available to characterize the properties of the SiO₂ and the Si/SiO₂ interface. The device structures most widely employed to evaluate such properties are the *metal-oxide-semiconductor*

capacitor (which we will refer to as the *MOS-C*) and the *MOS field-effect transistor* (which we will refer to as the *MOSFET*). It turns out, that for thin gate oxides (<5 nm), the capacitor measurements provide less pertinent information and the actual MOSFETs must be scrutinized.

There are four types of charges associated with the oxide and interface and they are shown schematically in Fig. 8-16. The total net *charge per unit area*, Q_{tot} which is given by the algebraic sum of all the charges present in the oxide, has units of (C/cm^2), while the *number of charges per unit area* (Q_{tot}/q) is denoted by N_{tot} . The standardized terminology and symbols used to represent the various individual charge types are as follows:²⁹

1. Fixed oxide charge per unit area, Q_f (C/cm^2), and number of fixed charges per unit area, N_f ($\#/\text{cm}^2$).
2. Mobile ionic charge per unit area, Q_m (C/cm^2), and number of mobile charges per unit area, N_m ($\#/\text{cm}^2$).
3. Interface trap charge per unit area, Q_{it} (C/cm^2), and number of interface trap charges per unit area, N_{it} ($\#/\text{cm}^2$), number of fixed charges per unit area and energy, D_{it} ($\#/\text{cm}^2\text{eV}$).
4. Oxide trapped charge per unit area, Q_{ot} , and number of oxide trapped charges per unit area N_{ot} ($\#/\text{cm}^2$).

8.5.1 Fixed Oxide Charge

Fixed oxide charge, Q_f , is usually *positive*, although in a few circumstances it can be negative. The *fixed charge* polarity does not vary with the surface potential (applied voltage) nor with time, and hence is termed fixed charge. The Q_f is considered to reside as a sheet of charge near the Si/SiO₂ interface and is commonly believed to result from the structural changes associated with the transition region between Si and the oxide. In thicker oxides this charge is conjectured to be about 3.0 nm from the interface. However in ultra-thin oxides (< 3.0 nm) the charge must be closer to the interface or distributed through the oxide. One model assumes that the origin of Q_f is from excess ionic Si that has broken away from the Si substrate during oxidation, and

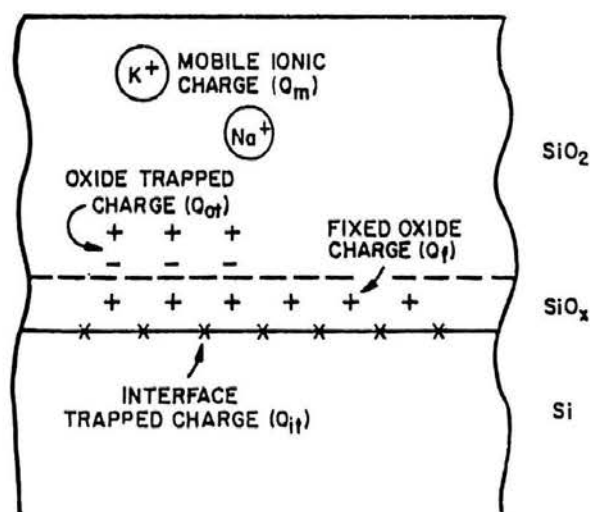


Fig. 8-16 Standardized terminology for oxide charges associated with thermally oxidized silicon.⁴⁴ (© 1980 IEEE).

moved into the oxide. The ionic Si remains in the oxide in the vicinity of the interface when the oxidation is terminated.

Regardless of the origin of the fixed charge, it has been experimentally established that the value of Q_f depends on several factors. These factors are: a) the silicon crystal orientation; b) the oxidation temperature and ambient; and c) the *post-oxidation anneal* (POA) conditions. Even though fixed charge will shift the threshold voltage V_T of the MOS transistor, the influence of the charge is small for today's devices, since V_T is dominated by the channel doping (which is set by ion implantation). As long as the value of N_f is in the order of a few times $10^{10}/\text{cm}^2$, the presence of fixed oxide charge is not considered an important processing limitation.

The Q_f value is determined by measuring the voltage shift of a high frequency capacitance-voltage (C-V) curve of an MOS-C test device. Details of the measurement are discussed in Reference 33. For the case of Q_f at the interface, the flat-band voltage, V_{FB} of the C-V curve is related to the oxide charge, N_f and oxide thickness, x_{ox} . The amount of fixed charge at the interface can be calculated from ΔV_{FB} which is given by $(V_{FB})_{meas} - (V_{FB})_{ideal}$. The value of C_{ox} is obtained from the C-V curve and the value of N_f can be calculated from Eq. 8-32:

$$\Delta V_{FB} = -Q_f/C_{ox} = q N_f/C_{ox} = q N_f x_{ox}/\epsilon_{ox}\epsilon_0 \quad (8.32)$$

where C_{ox} is the oxide capacitance in F/cm^2 , ϵ_{ox} is the dielectric constant of SiO_2 , ϵ_0 is the permittivity of free space, and q is the charge of an electron. As shown in Eq. 8-32, the change in the value of V_{FB} is directly dependent on the oxide thickness for any amount of fixed charge. For thin oxides the shift in V_{FB} may be too small to detect. For example, for oxide charge of $N_f = 10^{10}/\text{cm}^2$ in a 5 nm gate oxide, the V_{FB} shift would only be about 0.02V.

The factors that control fixed charge were studied in detail and the results summarized in the well known *Deal Triangle* shown in Fig. 8-17.³⁰ Using the triangle one can study how temperature and annealing in inert gas (Ar, N_2) affect fixed charge, N_f . Plotted on the y-axis is N_f , and on the x-axis, oxidation (or post-oxidation anneal [POA]) temperature. The hypotenuse of the triangle represents the impact of the oxidation (dry O_2) temperature on N_f . The higher the oxidation temperature the lower the value of N_f . The vertical lines represent the effect of cooling down from a high temperature oxidation in an inert gas. These results were obtained for (111) silicon, but are known to also apply to (100) silicon. It should be noted that the lowest values of N_f are obtained for oxidation or anneals at the highest temperatures. However, during furnace cooling processes, the wafers spend considerable time at temperatures lower than the oxidation temperature. Therefore, oxidation continues at lower temperatures. As a result, the value of N_f increases during cooling. There are several ways to circumvent this problem. First, if the wafers are rapidly cooled (quenched) from the elevated temperature to low temperatures they should retain the value of N_f they had at the highest temperature. However, quenching is not practical in a furnace since the rapid cooling can cause wafer warpage, breakage, or slip. But, it is practical to rapidly cool a wafer grown in a rapid thermal oxidation (RTO) system without causing slip or wafer breakage. The key to avoiding slip or breakage is to ensure that the radial thermal gradient across the wafer does not exceed $0.2^\circ\text{C}/\text{mm}$ during the heating and cooling cycles.

In place of quenching, the same result can be achieved by cooling the wafer in an inert ambient. This causes the oxidation to stop at the higher growth temperature and to "freeze-in" the lower N_f value. This process is termed *post oxidation annealing* (POA) and is generally carried out in a nitrogen or argon ambient in the furnace. It has been shown that not all the fixed charge can be annealed out at the same temperature as the oxidation temperature, and therefore the POA is often performed at a slightly higher temperature than the oxidation temperature.³¹

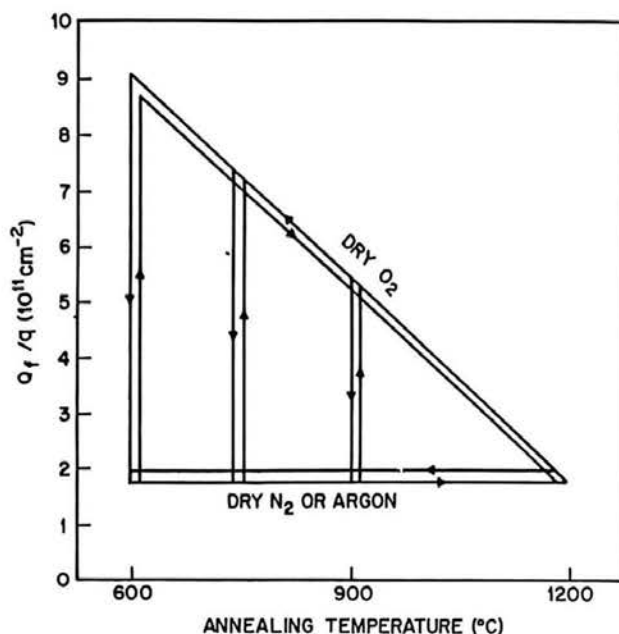


Fig. 8-17 Oxide fixed charge density as a function of annealing temperature for times one hour or less (The Deal Triangle).³⁰ Reprinted by permission of the publisher, the Electrochemical Society.

Figure 8-18 shows the dependence of Q_f and Q_{it} on Si crystal orientation,³² with (100) Si resulting in the lowest values for each type of charge. Thus, this orientation is almost exclusively used for MOS VLSI and ULSI circuits. Values of N_f in the low $10^{10}/\text{cm}^2$ range are achievable on (100) material after careful oxidation, anneal, and cooling in an inert environment.

8.5.2 Mobile Ionic Charge

The *mobile ionic charge* Q_m is commonly caused by the presence of ionized alkali metal atoms (Na^+ , K^+). These ions possess a very high diffusivity in an oxide film, even at temperatures as

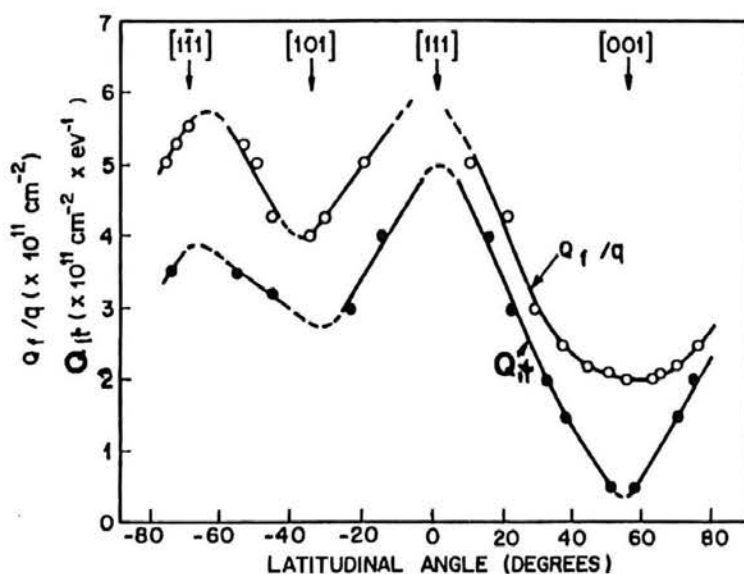


Fig. 8-18 Oxide fixed charge density and interface trap density as a function of the silicon orientation.³² Reprinted with permission of the American Physical Society.

low as 200°C. Since these elements are present in ionic form their transport is greatly enhanced by the presence of an electric field across the oxide. As a matter of fact, these ions can drift through thin oxides even at room temperature when a positive bias is applied to the gate electrode. The drift of the mobile ions from the gate electrode to the silicon interface can result in MOSFET threshold voltage instability.

The amount of mobile charge incorporated into the oxide depends on the cleanliness of the oxidation process. In today's modern fabs measures have been instituted to keep the level of ionic contamination well below the limits where instabilities can occur ($Q_m \leq 10^{10}/\text{cm}^2$). A value of Q_m in the low 10^{10} cm^{-2} range will result in a V_T shift of only a few hundredths of a volt for an MOS device with a gate oxide of 100 nm thickness. On the other hand, values of Q_m in the $10^{12}/\text{cm}^2$ range will result in V_T shifts of several volts.

Such low levels of sodium can be maintained by employing special precautions during processing, including: a) using a chlorinated oxidation; b) periodically cleaning the furnace tubes and associated glassware with chlorine; c) using ultra-high-purity chemicals for wet processing and for gaseous sources; and d) ensuring that the gas delivery lines remain clean. It is also important that the gate material (usually polysilicon) be free of these harmful contaminants. The use of heavily doped BPSG and PSG glasses as the interlayer dielectric serves to minimize the effect of gate electrode contamination, since the glasses effectively *getter* (trap) ionic impurities. Plasma deposited silicon nitride and/or PSG can be used to encapsulate the finished device and help keep ionic impurities away from the gate oxides during the operation of the devices.

The very thin oxides that are used to fabricate the gates for advanced devices make it difficult to detect the presence of the low levels of ionic contamination. The reason for this was discussed earlier for fixed charge. When it is necessary to qualify a furnace for ionic contamination, it is customary to grow a thick *monitor oxide* (100 nm) in the furnace and to perform the C-V analysis on this so-called monitor oxide. The monitor will exhibit a V_{FB} shift of 0.4V for a $10^{10}/\text{cm}^2$ level of contamination.

The amount of mobile ionic charge in the oxide and in the gate material is measured by the use of a C-V technique called a *bias-temperature test* (B-T test). This technique is performed by first plotting the C-V curve on a freshly prepared MOS capacitor. The next step is to apply a

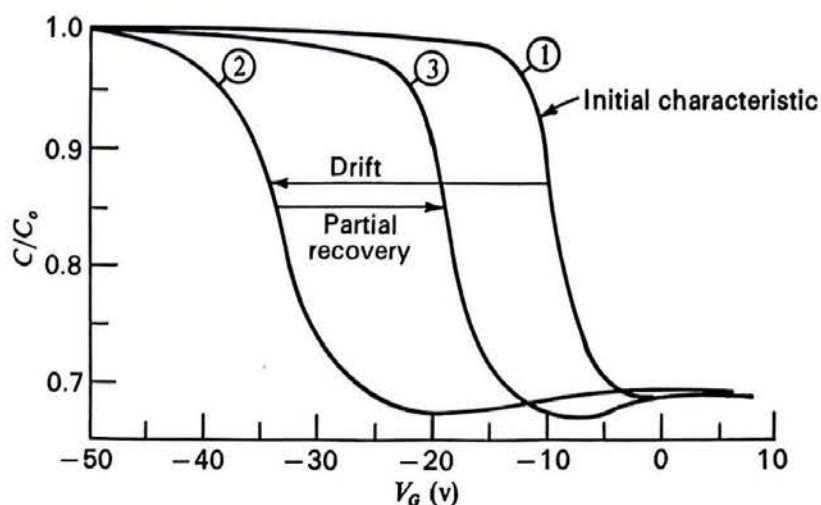


Fig. 8-19 High frequency C-V curves showing the shift after B-T stressing as a result of ionic contamination.³⁴ Reprinted with permission of the American Physical Society.

positive voltage (bias) of 1 MV/cm to the gate electrode while heating the device to 200–300°C. Once at the required temperature, the bias is applied to the device for an additional 10–30 min to ensure the mobile charge has been driven to the interface. The device is then cooled back to room temperature while maintaining the bias. A new C-V measurement is now made. The change in ΔV_{FB} before the B-T test and after the test is equal to N_m/qC_{ox} or $N_m x_{ox}/q\epsilon_{ox}\epsilon_0$. Figure 8-19, shows the results of a B-T test performed on a heavily contaminated device.³⁴ As the positive charge is swept to the silicon interface, the C-V curve shifts to more negative voltages. If one applied a negative voltage to the gate and performed the B-T test again, the C-V curve might be expected to recover to its original position. However, this does not always occur and sometimes only a partial recovery is observed. This is the result of charge injection and trapping in the oxide during the first B-T test.

8.5.3 Interface Trap Charge

The *interface trap charge*, Q_{it} refers to charge which is localized at sites near the Si/SiO₂ interface. Such sites are believed to arise from energy states that somehow exist in the forbidden gap of silicon, and are referred to as *interface states*. These interface states are distributed across the silicon band gap (unlike the other charges discussed) and therefore it is advantageous to define a new term, the *interface trap density per unit energy*, D_{it} , which has units of #/cm²eV. Figure 8-20 shows a plot of D_{it} from two sets of data as a function of energy in the forbidden gap.³⁵ The value of D_{it} has a well known U-shape with the minimum trap density at the mid-gap and the maximum at the band edge. The minimum trap level at mid-gap is typically the concentration level used to characterize their presence. The mid-gap value of D_{it} for (100) silicon is about $2-3 \times 10^{10}/\text{cm}^2\text{eV}$ and increases rapidly to $10^{13}/\text{cm}^2\text{eV}$ at the band edges. The trap density for (111) silicon is at least a factor of 5 higher at mid-gap than for (100) silicon.

Interface states above mid-gap are acceptor-like and are negative when filled with electrons, and neutral when empty, while those below mid-gap are donor like and are neutral when filled and positive when empty. As a result the charge state of the trap will depend on the bias condition (Fermi level) of the MOS capacitor. The details of how the applied bias affects charge states is discussed in Chap. 3, Vol. 3 of this series. The change in threshold voltage arising from the presence of interface traps is given by Eq. 8-33. The V_T shift associated with Q_{it} depends on the surface potential ϕ_{surf} , which in turn depends on the applied voltage.

$$\Delta V_T = -Q_{it} \phi_{surf}/C_{ox} \quad (8.33)$$

It is difficult to determine D_{it} from high-frequency C-V measurements, since the interface charge tends to distort the curve, rather than shift the flatband voltage. The threshold voltage may shift with increased interface charge. But, the shift will be small for moderate values of interface state density in the case of thin oxides. More insidiously, the presence of a large amount of interface state charge can significantly reduce the channel mobility of the transistor and thereby degrade the device performance. The value of D_{it} is found to be about 10^{11} – $10^{12}/\text{cm}^2\text{eV}$ after dry oxidation of a (100) surface, and the value tends to decrease (similar to the case of Q_f) as the oxidation temperature increases. On the other hand, annealing in an inert atmosphere does not further reduce D_{it} . The level of D_{it} can be brought to an acceptable level ($3-5 \times 10^{10}/\text{cm}^2\text{eV}$) by applying a post-metallization anneal (PMA). The PMA is performed after the device has been metallized with the aluminum interconnect. The procedure is to anneal the completed device in a hydrogen ambient, such as forming gas (95%N₂–5%H₂) or even pure hydrogen, in the temperature range of 400–450°C for 5–30 minutes. It is believed that atomic hydrogen (H) is

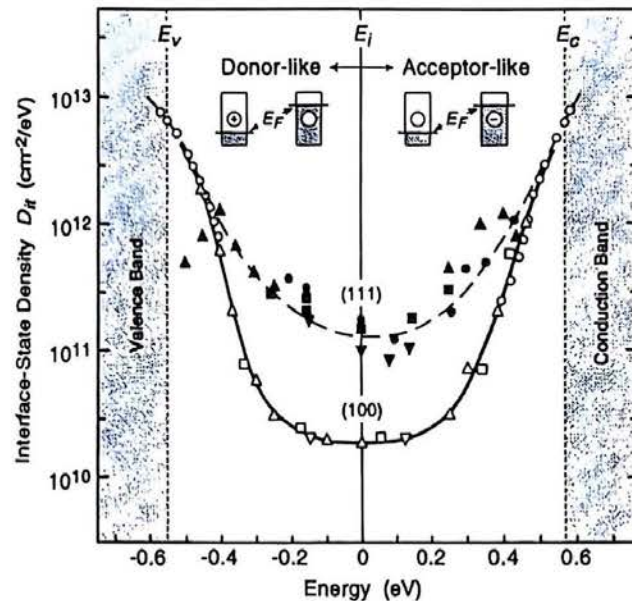


Fig. 8-20 Interface state density D_{it} measured on thermally grown oxide on (100) and (111) silicon.³⁵

formed in the presence of the Al, and the atomic H diffuses to the Si/SiO₂ interface where it reacts with dangling bonds (the putative source of interface states), passivating the bonds and reducing the interface states. The presence of significant H concentration at the interface can abet a reliability problem called *hot carrier instability*. This phenomenon is discussed in detail in Vol. 3. Recent results have shown that by using deuterium (D₂, an isotope of hydrogen) as the PMA annealing gas, the passivation of the dangling silicon bonds is more resistant to the onslaught of hot carriers.³⁶

Although there are models to explain the electrical properties of the interface states, their physical origin is still not fully understood. What is known is that there is a non-equilibrium transition region between the oxide and the silicon substrate. This transition region may only be 1 nm thick but the disruption that results to the periodicity of the silicon lattice can have a profound effect on the electrical properties of the interface. The weight of the experimental evidence suggests that the interface traps arise from unsatisfied chemical bonds (e.g., a triply bonded Si atom) at the interface called *dangling bonds*.

The value of D_{it} can be measured by several techniques including: a) conductance (I-V) method; b) low-frequency C-V measurements; c) charge-voltage Q-V method; d) charge-pumping method; and e) MOSFET sub-threshold current. The details of these methods is beyond the scope of this volume and the reader is referred to the reference text by Schroder.³⁷

8.5.4 Oxide Trapped Charge

The last of the oxide charges to be discussed is *oxide trapped charge* Q_{ot} . These trap centers are located in the bulk of the oxide, but are *neutral* until they interact with injected carriers and *trap* them. As a result the traps can become charged either positive or negative depending on whether they trap holes or electrons. The traps are generally associated with defects in the oxide. Some of the defects that can cause a trap include: a) dangling bonds; b) interface traps; c) Si-Si bond stretching; d) non-bridging oxygen atoms (oxygen dangling bonds); e) presence of a weak Si-Si bond which may be broken and become electrically active; f) strained Si-O bonds; and g) the presence of Si-H or Si-OH bonds. Some traps can be *amphoteric*. That is, they can trap either holes or electrons. Oxide traps are important since their presence can diminish the reliability of

devices. The device failure mode from trapped charge is either V_T drift outside the operating range of the device or, as traps become filled, premature and catastrophic dielectric breakdown.

Although as-grown oxides usually have little oxide trapped charge, there are several ways in which charge carriers can subsequently get into the oxide and be trapped. The most common are: a) electron-hole pairs created in the oxide due to plasma-process-induced radiation (or radiation during device operation, e.g., device operation in nuclear or outer-space environments), and b) hot carrier injection from the silicon substrate during device operation.

Electron trap densities are measured by injecting electrons into the oxide and measuring the shift in the C-V curve. Since the traps are not a sheet of charge, but are instead distributed throughout the oxide, the observed shift is inversely related to the moment arm of the charge density in the silicon. The charge closer to the interface has a larger contribution than charge further away. Oxides that contain H^+ , OH^- , and H_2O have more electron traps. Accordingly, one might expect wet oxides to contain more electron traps than dry oxides, a conjecture sustained by published results.³⁸ Electron traps are also reported to depend on the POA (*post-oxidation anneal*), PMA (*post metallization anneal*), and gate preparation techniques used during device fabrication. The density of electron traps can be significantly reduced with a proper POA.³⁵ As a result wet oxides can be grown with trap densities nearly equal to those obtained in dry oxides.

Hole traps, on the other hand, differ substantially from electron traps. For instance, the concentration of hole traps can exceed $10^{12}/\text{cm}^2$ for even the best quality oxides, and hole traps are much more efficient at capturing holes than are electron traps at capturing electrons. It is believed that the extremely low carrier mobility of holes in oxides (6–8 orders of magnitude less than electrons) is why they have such a high capture cross-section. The hole traps also tend to be localized near the Si/SiO₂ interface as opposed to electron traps which tend to be more uniformly distributed throughout the oxide. Fortunately, holes are less likely (due to the larger barrier height for hole injection) to be injected into the oxide. But once injected, they have a more insidious effect on the oxide properties. Devices operating at high voltage levels and those subjected to charged particle irradiation (e.g., plasma etching) usually are subjected to both hole and electron injection. Thus, it is important to minimize the number of hole traps in a grown oxide. The use of high temperature POA in an inert ambient *increases* the density of hole traps. However, the low hole trap value can be restored by a brief high temperature anneal in oxygen.³⁹ Hole traps are theorized to arise from an oxygen deficiency at the Si/SiO₂ interface. Furthermore when such traps are formed, it is believed that the oxide network undergoes a structural modification, making hole traps more stable than electron traps. It is found experimentally that injected holes produce 2–3 times as many interface states as do injected electrons.

During high field stressing of a device it is possible to inject *hot* or *energetic carriers* into the oxide where they become trapped at the interface. For example, such carriers can break Si-H bonds that are passivating dangling silicon bonds. The released hydrogen can escape from the oxide, resulting in the creation of a new interface state. Another way to increase interface states is through radiation damage. Radiation damage can occur during subsequent processes such as reactive ion etching, ion implantation, and in some cases sputtering. If the incoming radiation has an energy greater than the bandgap of the oxide an electron-hole pair is formed. Keeping in mind that hole trapping is much more efficient than electron trapping, it is not surprising that the excess holes generated near the Si/SiO₂ interface are easily trapped. The trapped holes can create interface states which adversely affect the device performance. Some of the radiation damage can be annealed away during the PMA, but it is still important to minimize the radiation levels since the hole traps can result in reduced breakdown fields.

8.6 STRESS IN SILICON DIOXIDE

There are two main sources of stress in silicon dioxide grown on silicon. The first is the *residual stress* that occurs when the oxidized wafer is cooled from the oxidation temperature to room temperature. This stress originates from the difference between the thermal coefficients of expansion (TCE) of the silicon and the oxide (the TCE of Si is about 5 times larger than that of SiO₂). When the wafer cools from the oxidation temperature, the silicon contracts more than the oxide, building in a compressive stress in the oxide on the order of 10⁹ dynes/cm².

The second type of stress is that which develops in the SiO₂ at the Si/SiO₂ interface during growth of the oxide. This stress develops as a result of the way the oxide grows. When silicon is oxidized the new oxide forms at the Si/SiO₂ interface. The silicon atom that is consumed during the growth has a volume of about 0.02 nm³, while the growing SiO₂ molecule has a much larger volume of 0.04 nm³. The volume expansion causes a stress in the SiO₂ at the Si/SiO₂ interface. This stress has been experimentally measured by EerNisse.⁴⁰ The stress in the oxide was determined by measuring the wafer bending as the oxide was being grown (at high temperatures) on only one surface of the silicon. The stress was calculated from the radius of curvature of the silicon. The results of the study showed that for temperatures up to 950°C the SiO₂ was in state of compressive stress with a value of 7x10⁹ dyne/cm². As the temperature was raised to 975°C the stress essentially disappeared. Above 1000°C the stress shifted to be slightly tensile. These results were interpreted to mean that at about 965°C the oxide undergoes *viscous flow* which can relieve the built-in stress. Models of viscous flow in SiO₂ have been incorporated into the process simulator SUPREM4.

8.7 DOPANT IMPURITY REDISTRIBUTION DURING OXIDATION

The dopant impurity concentration near the silicon surface changes during oxidation. The concentration change is called *dopant redistribution*, and it is important since the surface concentration controls a variety of device parameters, including V_T. The analysis of dopant redistribution depends on three main factors. The first is called the *segregation coefficient*, *m*, which is defined as:

$$m = \frac{\text{equilibrium concentration of impurity in Si}}{\text{equilibrium concentration in SiO}_2} = \left(\frac{C_{\text{Si}}}{C_{\text{SiO}_2}} \right) \quad (8.34)$$

As the oxide grows it consumes silicon along with the impurities that were originally in the silicon. However, the oxide phase will only be able to accommodate the equilibrium concentration appropriate to the oxide. If *m* < 1, the impurity concentration in the oxide will be higher than the concentration in the silicon, and *C*_{Si} < *C*_{SiO₂}. If *m* > 1, the inverse condition will prevail, and *C*_{Si} > *C*_{SiO₂}.

The dopant can also be redistributed by a second factor, namely the *diffusivity of the dopant impurity in the oxide*. If the dopant quickly diffuses away from the interface (and then through the oxide), the concentration in the Si can be further reduced. Even if impurities segregate into the oxide, they will escape through the SiO₂ surface, reducing the concentration in the oxide. Gallium has a high diffusivity and so does boron when diffusion takes place in a H₂ ambient.

The third factor that can affect the redistribution is the *rate of the Si/SiO₂ interface movement* into the underlying silicon. Since the volume of the SiO₂ is more than twice that of the silicon consumed, and assuming the oxide were to incorporate the same impurity concentration as the silicon, some impurities must be depleted from the Si to compensate for the volume gained

during oxide growth. In other words, the oxide needs more than twice the amount of impurity to maintain the same concentration level as the silicon.

Grove *et al.* developed a pictorial representation of four impurity redistribution cases, reproduced here in Fig. 8-21.⁴¹ To begin, two circumstances are possible: a) the oxide takes up the impurity ($m < 1$); and b) the oxide rejects the impurity ($m > 1$). The redistribution profile then depends, to a large extent, on how rapidly the impurity can diffuse through the oxide. In Case #1 (shown in Fig. 8-21a), $m < 1$ and the diffusion through the oxide is slow (as is the case of boron in an oxidizing ambient). Since the boron segregates into the oxide, the silicon surface becomes depleted of boron. If in subsequent processing, however, the oxide is subjected to a hydrogen atmosphere (whereupon the diffusivity of boron in SiO_2 rapidly increases), the boron at the silicon surface is depleted even more than in Case #1 (i.e., see Case #2, shown in Fig. 8-21b). In Case #3 (see Fig. 8-21c) the oxide rejects the impurities ($m > 1$), and the impurity diffuses slowly in the oxide (e.g., phosphorus). Here, the phosphorus accumulates at the silicon surface. In the Case #4 where $m < 1$ but the impurity rapidly diffuses through the oxide (e.g., gallium), the Ga leaves the solid, and the oxide may become devoid of dopant (Fig. 8-21d).

Table 8-8, shows the range of m values, both calculated and measured. The value of m can be calculated by fitting the measured diffusion profiles to a model for redistribution. The model involves the solution of the diffusion equations with a moving boundary (i.e., taking into account that the silicon is being consumed). The details of the model are beyond the scope of this text and are discussed elsewhere.⁴¹

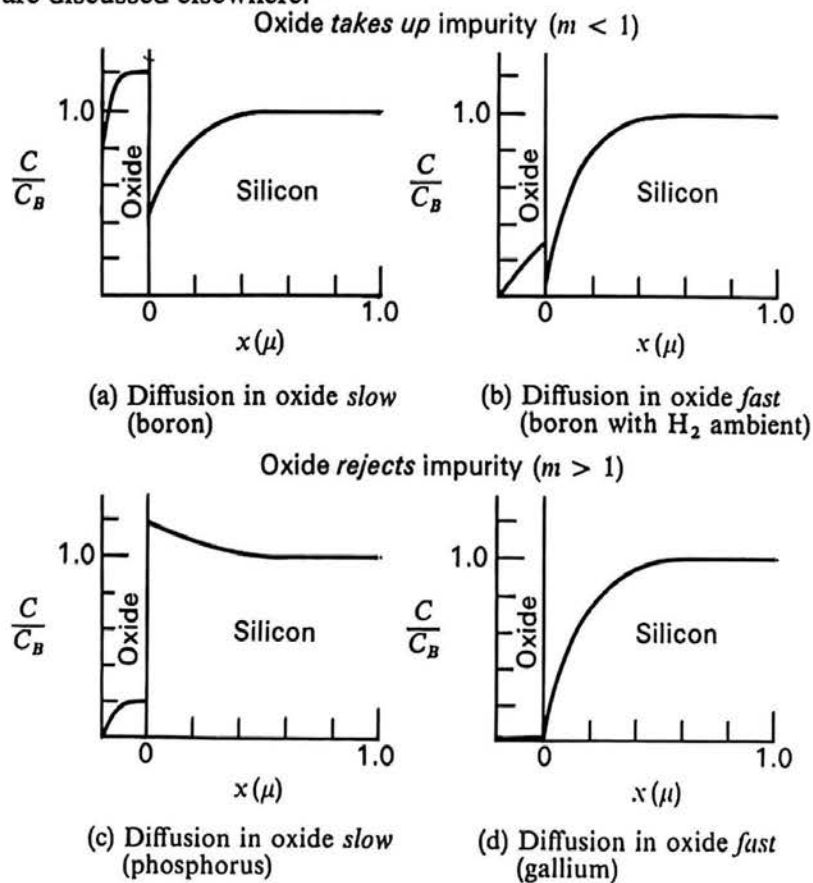


Fig. 8-21 The effect of thermal oxidation of silicon on the impurity segregation at the Si/SiO₂ interface: (a) slow diffusion in oxide, $m < 1$ (boron); (b) fast diffusion in oxide, $m < 1$ (boron in H_2 ambient); (c) slow diffusion in oxide, $m > 1$ (phosphorus); and (d) fast diffusion in oxide, $m > 1$ (gallium).⁴¹ Reprinted with permission of the American Physical Society.

Table 8-8 VALUES OF THE SEGREGATION COEFFICIENT OF IMPURITIES AT THE Si/SiO₂ INTERFACE⁴¹

Impurity	Thermodynamic Est.	Redistribution Expts.	Oxide Masking Expts.
Ga	> 1000		20.0
B	0.001–1000	0.3, 0.1	0.01
In	> 1000		
P, Sb, As	>1000	≈10	

The value of m , however, may be experimentally measured. Secondary ion mass spectroscopy (SIMS), can be used to directly determine the segregation coefficient by measuring the concentration of the impurity both in the oxide, and in the silicon near the Si/SiO₂ interface. Using the SIMS method, the value of m for boron is found to increase exponentially with temperature,⁴² with an activation energy of 0.33 eV for dry oxidations and 0.66 eV for wet or steam oxidations. The value of m is greater for (100) than for (111) oriented silicon.

An example where the redistribution of boron creates a device related problem is during LOCOS field oxidation. Since a very thick oxide is grown in the field region over boron-doped p -type silicon (in n -channel MOSFETs), and since boron piles up in the oxide, it is possible to deplete the silicon surface of boron. Adding in the effect of positive fixed charge in the oxide (which tends to further deplete the silicon surface of holes), it is possible to invert (make n -type) the silicon surface in the field region. If this inversion is significant, the leakage currents in the device can become large enough to make the circuit fail. In order to minimize the leakage and compensate for the boron depletion, additional boron is ion implanted (channel-stop implant) in the field region.

8.8 THE OXIDATION OF POLYSILICON

Polysilicon, like single-crystal silicon, readily forms an oxide layer when exposed to an oxidizing ambient. The oxide of polysilicon is an important device component in such applications as the capacitor dielectric in DRAMs. The oxidation rate of polysilicon depends on the presence of a preferred grain orientation, grain size, dopant type, and concentration. In general, lightly-doped polysilicon oxidizes more rapidly than does lightly-doped single-crystal silicon. It is postulated that enhanced oxidation results from stress created by the presence of the grain boundaries. Detailed studies of oxidation of polysilicon show that for undoped polysilicon in wet and dry ambients, at low temperature (<1000°C), the poly oxidation rate, at short times, is greater than either the oxidation rate of (100) or (111) silicon. At longer times, however, the rates converge. This is not surprising since the parabolic rate constant depends only on the diffusion through the existing oxide. It was also found that heavily-doped polysilicon oxidizes faster than lightly-doped polysilicon but not as fast as heavily-doped single-crystal silicon.⁴³

When polysilicon oxides are used in device applications the quality and reliability of the oxide is paramount. The breakdown strength of the oxide is heavily influenced by the smoothness of the polysilicon surface. The rougher the surface the lower the breakdown strength, since the roughness can cause electric field enhancement and breakdown at lower voltages than expected for the same oxide thickness on single-crystal silicon. In some applications, such as in DRAM capacitors, the polysilicon is deposited under conditions to increase the surface roughness (in order to increase the surface area of the electrode and the capacitance of a DRAM cell). High temperature growth can improve the quality of the oxide by allowing viscous flow to

relieve the stress in the film. Unfortunately, the thermal budget allowed for ULSI devices does not afford the luxury of high temperature processing.

8.9 THE OXIDATION OF SILICON NITRIDE

During the field oxide growth in the LOCOS isolation process (see Chap. 17), some oxide will form on the surface of the silicon nitride film. The oxide grows as a result of conversion of nitride to oxide. The conversion is a slow, self-limiting process. During field oxidation, when 400 nm of oxide grows on silicon, only 1–2 nm of oxide will grow on the nitride. But, the presence of the oxide can cause a process issue during the nitride strip step. If the nitride strip is done in hot phosphoric acid, it is necessary to first strip the thin oxide on the nitride in HF. Unless this oxide is removed, the phosphoric acid that etches the nitride will not remove the oxide, and stripping of the nitride will be retarded. If the nitride strip is done in a downstream plasma system, the problem is mitigated (since the oxide will etch in the CF_4 chemistry). However, the process must still include provisions for removing any oxide formed on the nitride.

8.10 THERMAL NITRIDATION OF SILICON AND SILICON DIOXIDE

Thermal nitridation (with ammonia) directly on silicon to produce silicon nitride was first aimed at generating a gate quality dielectric to replace silicon dioxide. The hope was that silicon nitride would have less defect density and higher breakdown strength than the oxide. These lofty goals were never achieved. As a matter of fact, devices made with directly nitrided gates have poor charge trapping and very high interface state densities. Thermal nitridation is achieved by reacting NH_3 with silicon according to:



The growth of the nitride is extremely slow and self-limiting, since nitrogen cannot easily diffuse through the growing nitride film to react with silicon. (Silicon nitride is one of the best barriers to diffusion used in ULSI processing, see Chap. 6.) It is possible to grow only 3–4 nm of silicon nitride at high temperatures. There is a DRAM capacitor application where 1.5–2.0 nm of thermal nitride is grown on polysilicon at 900°C for 1 min using RTP. This nitride layer serves as a barrier to prevent further oxidation of the polysilicon during subsequent processing.

The nitridation of thermal oxides (the nitrided oxide film being called an *oxynitride* or even a *nitrioxide*), has resulted in improvements for thin gate oxides (<4 nm) and tunnel oxides (<10 nm). Some of the benefits realized include: a) improved resistance to boron penetration; b) increased V_{BD} and Q_{BD} ; c) improved resistance to radiation damage; and d) improved endurance in flash memories. Some of the disadvantages reported for nitrided oxides include: a) increased fixed charge and interface state density and b) reduced channel mobility.

The techniques available for forming nitrided oxides are too numerous to discuss in detail here. This field is a very fertile area of research with new articles being published each month. Hence, only a brief number of the current techniques can be listed. Interested readers should consult the post-1995 journals and meetings publications of various technical societies for the latest work in this area. (Some of these include: *Applied Physics Letters*; *Electron Device Letters*; *Technical Digest of the IEDM*; the *Digest of the Papers of the VLSI Technology Symposium*; and the *Proceedings of the Spring Meetings of the Materials Research Society*).

There are two main ways to form a nitrided oxide. The first is to grow a silicon dioxide layer and then anneal that layer in either ammonia (NH_3), nitrous oxide (N_2O), or nitric oxide (NO).

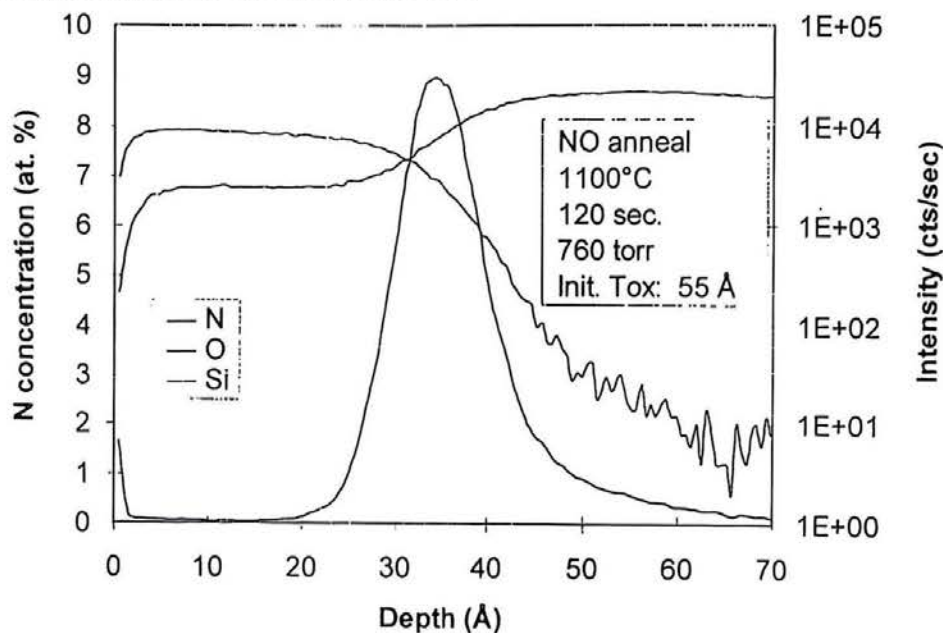


Fig. 8-22 SIMS profile showing N incorporation at the SiO_2 interface after annealing a thermal oxide in NO at 1100°C for 120 sec. Courtesy of Applied Materials.

The annealing is often performed in an RTP system, but a furnace can also be used. The annealing technique puts nitrogen into the oxide near the Si/SiO₂ interface. The nitrogen concentration and distribution can be measured using either SIMS or medium ion backscattering. Fig. 8-22 shows a SIMS plot of the nitrogen concentration for a NO annealed 3.3 nm oxide. The peak nitrogen concentration is 9% atomic, but the value depends on the temperature and time of the anneal. The second technique is to grow the oxide directly in N₂O or NO, or mixtures of those gases, with O₂. When the oxide is grown this way the nitrogen tends to be more uniformly distributed within the oxide. When using N₂O growth only about 1.5–2.0% nitrogen can be incorporated in the oxide.

8.11 TWO-DIMENSIONAL OXIDE-GROWTH EFFECTS

It is important to consider two-dimensional oxide growth when considering such effects as birds beak encroachment, oxide thinning, trench corner rounding, and gate bird's beak.

8.11.1 Birds Beak Encroachment

During the isolation process called *LOCAl Oxidation of Silicon* (LOCOS), a thick oxide is grown in the field region of MOSFETs (while the active device region is protected by a masking layer of silicon nitride, see Chap. 16). Since oxidation can occur in two-dimensions, some oxidant diffuses under the edges of the nitride where unwanted oxide grows. The growth under nitride decreases as oxidant moves inward under the nitride edge. As a result, a gradually tapering oxide wedge merging into the pad oxide forms under the nitride. The oxide grown under the nitride can lift the nitride edges. The shape of the field oxide as it penetrates under the nitride has been given the name *bird's beak*. The bird's beak is a lateral extension of the field oxide into the active area of the device. Figure 8-23 shows an example of the bird's beak.⁴⁴ The length of the bird's beak depends on the thickness of the pad oxide and nitride layers, and on the oxidation temperature and pressure. The length of the bird's beak for a 500-nm-thick field oxide is about 500 nm on a side. Thus, a lithographically defined active region of 1 μm will virtually disappear after field oxidation. To accommodate the bird's beak it is necessary to make the active region larger. This uses extra chip real estate and is inefficient for ULSI circuits.

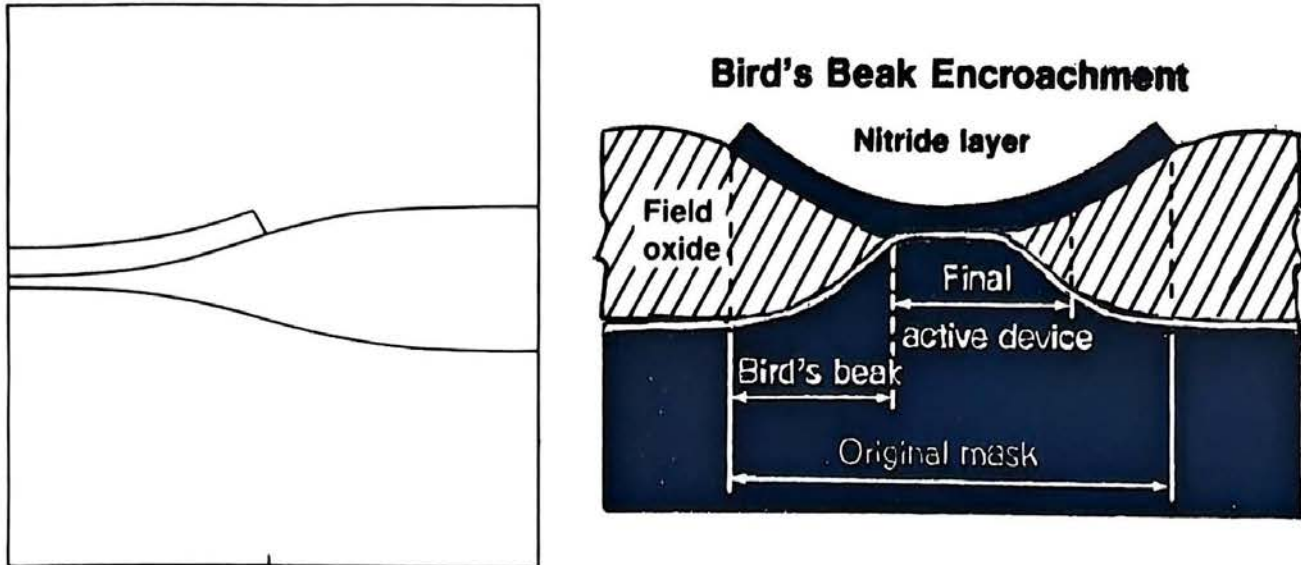


Fig. 8-23 Bird's beak encroachment limits the scaling of channel widths to about 1.5 μm .⁴⁴

There are numerous processing approaches to minimize the bird's beak, and some are described in Volumes 2 and 3 of this series.

HIPOX can be used to minimize bird's beak encroachment since vertical growth is more rapid than lateral growth during high pressure oxidation. As a result, the target field thickness can be reached with less lateral growth; accordingly resulting in a reduced bird's beak.

8.11.2 LOCOS-Induced Gate Oxide Thinning

In LOCOS processing the sequence of steps after growing the field oxide is stripping the protective nitride, growing and stripping a sacrificial oxide, and then growing the gate oxide. After the gate oxide is grown it is found that the edges of the gate can be substantially thinner than the center region. The oxidation in the gate region near the LOCOS edge is reduced by the presence of the thick oxide. The thick field oxide can limit the availability of oxidant in the region at the edge of the gate, resulting in gate oxide thinning. The use of HIPOX to minimize this effect is discussed below.

8.11.3 Trench Corner Rounding

There are two major applications of trench technology in modern devices. In both applications it is necessary to oxidize the inside and top of a trench etched into silicon. As a result, oxidation takes place along two different crystalline planes and two-dimensional effects must be considered. The first use of such trenches is in DRAM capacitors. In this application a deep trench (5-10 μm) is etched into the silicon. A capacitor structure is built in the deep trench. The capacitor dielectric can be either an oxide or a thermal oxide/CVD-nitride combination. The second use of trenches is called the *shallow trench isolation* (STI) structure. In the STI application individual MOS devices are isolated from each other by a shallow trench etched into the silicon. The trench is thermally oxidized and filled with a deposited oxide to make it non-conducting. The device is then built on the plateaus of silicon between the trenches. The use of STI is the most popular isolation technology for advanced devices (i.e., for 0.25 μm technologies, and smaller). See Chap. 16 for more details on STI.

The oxidation profile in a trench will now be examined. When a trench is oxidized, oxidation occurs on the flat regions at the top and bottom of the trench and at both the convex and concave

corners. Marcus and Sheng showed dramatic examples of oxide thinning at the corners of oxidized etched steps in silicon.⁴⁵ Low temperature oxidations (900–950°C) displayed as much as 30% decrease in the oxide thickness at both the concave and convex corners of the step. A later study by Kao *et al*, examined the retardation of oxidation on cylindrical structures with varying radii of curvature.⁴⁶ They measured the oxidation rates using steam and dry oxidation, over the temperature range of 800–1200°C on flat regions and corners of various radii of curvature, and used SEM photographs to determine the oxide thickness. At low temperature they found significant oxide thinning at the corners while at 1100°C, the oxide thickness was uniform and the corners were well rounded. Fig. 8-24 shows a SEM of such a structure with substantial thinning at both corners. Figure 8-24b, shows a schematic representation of the results. Also shown in Fig. 8-24 are oxides grown on the outside of a cylinder of silicon (concave) and the inside of a silicon hole (convex). Both corners show reduced oxidation, while the oxide on the concave appears to be thinner than the convex. The circular structures reveal that oxidation on the concave is less than on the convex structure. Fig. 8-25 shows a plot of normalized oxide thickness vs. reciprocal radius of curvature for both concave and convex surfaces. That plot shows that the smaller radius of curvature results in thinner oxides, as does oxidation at lower temperature. In all cases the concave surface produced a thinner oxide than the convex surface.

The retarded oxidation observed at both types of surfaces was attributed to the reduced chemical reaction rate (k_s) at the SiO_2 /silicon interface.⁴⁷ The reduced oxidation rate is believed to arise from viscous stress. The different stresses at the concave (compressive) and convex (tensile) corners account for the increased reduction on concave surfaces. The temperature dependence can be explained by the decrease in oxide viscosity at higher temperatures and less stress. The effect of radius of curvature is explained by noting that as the radius is made smaller (sharper corner) the deformation of the oxide increases. Thus, higher stress occurs in the oxide.

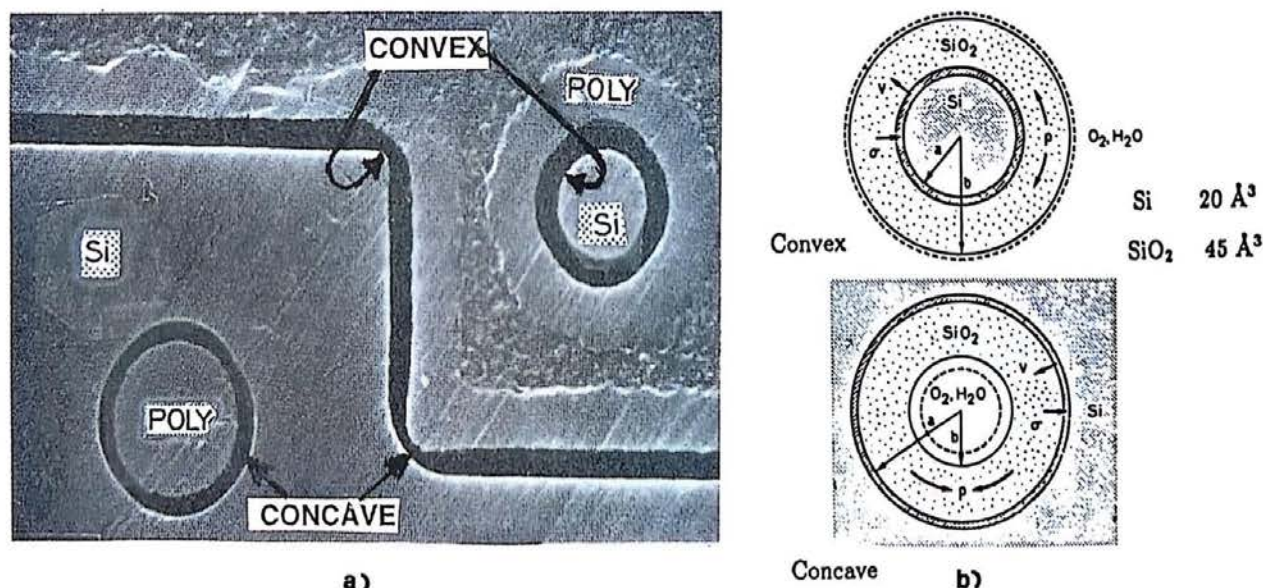


Fig. 8-24 Comparison of convex and concave silicon surfaces a) SEM on top and b) schematic on bottom. The oxide on the hole (concave) is thinner than the oxide on the cylinder (convex).⁴⁶ (© IEEE 1987).

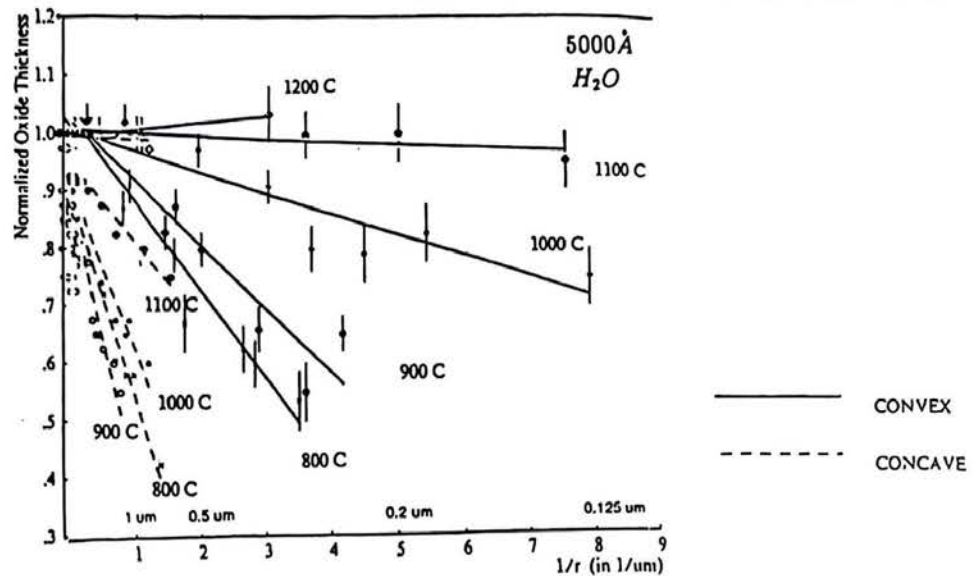


Fig. 8-25 Experimental data for wet oxidation showing the relative thickness on concave and convex surfaces as a function of radius of curvature and temperature.⁴⁶ (© IEEE 1987).

8.11.4 Gate Bird's Beak

After the etching of the gate stack in a CMOS process it is necessary to re-oxidize the structure in order to heal plasma etch damage at the edge of the gate oxide. This oxidation step is termed *polysilicon re-oxidation*. Without the polysilicon re-oxidation step the yield of the circuits will degrade as a result of low breakdown voltage of the damaged gate oxide. At the same time new oxide is grown at the edge of the gate region, the polysilicon sidewalls are also oxidized. The growth of the polysilicon oxide forms a *gate bird's beak* (or GBB), under the polysilicon edge. The GBB encroachment can degrade some device characteristics of MOSFETs (see Chap. 9, Vol. 3), hence the re-oxidation process must be optimized for performance and yield.

8.12 OXIDATION SYSTEMS

In this section we discuss the role of *thermal processing tools* (what were once termed furnaces) in the manufacture of ULSI circuits is discussed. Conventional vertical and fast ramp furnaces, rapid thermal processors (RTP), and high-pressure oxidation systems will be described, as well as the concept of *thermal budget*.

8.12.1 Thermal Budget

The *thermal budget* or *thermal exposure* of a process can be thought of as the amount of time the wafer is at a particular temperature during the whole process. The thermal budget can be quantified as the area under the time-temperature (t-T) curve or time-diffusivity (t-D) curve.⁴⁸ In order to fabricate MOSFETs with shallow-junction source and drain junctions it is necessary to minimize the thermal budget of all the process steps after the source-drain implant step. For example, to fabricate such shallow junctions one must be able to activate the implanted atoms and remove the lattice damage, while at the same time minimize the atom movement due to diffusion. Typically all the steps are done in one high-temperature thermal process. This minimizes the total thermal exposure of the wafer. A challenge facing the process engineer is how to best accomplish thermal processing. One option is to choose a low-temperature process for longer times, while another is to select a high temperature process for shorter times (and still maintain the same t-T budget). To optimize both diffusion and implant damage annealing, one

needs to consider not just the t-T exposure but the t-D exposure. The t-D approach takes into account the activation energy for the different processes. Using the t-D methodology, the best results are often obtained by employing annealing at *high temperature* and *short times*. As a result, RTP has replaced furnace annealing when fabricating shallow source/drain junctions.

8.12.2 Thermal Processing Equipment (Tools)

Thermal processing tools provide the vehicle for numerous applications in today's fabs. Among the more common uses are: a) thermal oxidation and POA; b) activation of ion-implanted impurities; c) dopant drive-in; d) post-implantation damage annealing; e) reflow and densification of CVD glasses; f) formation of Ti and Co silicides; g) annealing of contacts; h) post-metallization anneal (PMA); and i) deposition of a diversity of non-plasma CVD films such as doped and undoped polysilicon, thermal silicon nitride, and doped glasses. The total market for thermal processing tools in 1998 was \$1.8 billion and is projected to grow at a rate of 10–20% per year.

8.12.3 Furnaces

The hot-wall *horizontal diffusion furnace* was the workhorse of the IC industry for over 30 years (until the mid-1980s), when *vertical* furnaces began to proliferate into production fabs. Vertical furnaces offer several technological advantages over horizontal furnaces, although horizontal furnaces offer high productivity at low cost. A 4-stack horizontal furnace tool in 1998 cost about \$0.8–\$1.0 million, while one vertical furnace costs just as much. More recently *rapid thermal processors* (RTP) have replaced vertical furnaces in some critical applications. RTP is widely used when control of the processing atmosphere is crucial (e.g., oxygen exclusion), when very short time anneals are required, or when reduced thermal budget is paramount. RTP continues to find applications where it performs better than furnaces. On the other hand, vertical furnace technology continues to evolve, and now, fast-ramp and small-batch furnaces compete with RTP systems in some applications where thermal budget and throughput are important.

Furnace tubes can be classified as operating at *atmospheric pressure* and at *low or sub-atmospheric pressure*. There are also furnaces that operate at high pressure and they will be discussed below. The atmospheric furnaces are more widely used for oxidation and anneal, while the reduced-pressure systems are employed for CVD applications. However, with the requirements of growing very thin oxides for gate applications, the use of reduced pressure oxidation has become more widespread. In the late 1990's there were a large number of conventional furnace suppliers. Those with the largest world-wide market share (in approximate descending order) were Tokyo Electron (Japanese), Kokusai (Japanese), Silicon Valley Group (U.S.), ASM International (Holland), and Semitherm (U.S.).

8.12.4 Horizontal Furnaces

Most horizontal furnaces are configured as a 4-stack, typically with a single robot to deliver the wafers to the tube. The horizontal furnace consists of the following modules: a) a furnace cabinet; b) heating elements; c) measuring and control thermocouples; d) fused-silica process tubes; e) fused-silica paddles and boats; f) a temperature control system; and g) a load station. Since horizontal furnaces are rarely used in producing ULSI devices they will not be discussed further here. For additional information, readers are referred to the first edition of this text (Vol. 1., published in 1986) and to product literature distributed by furnace suppliers.

8.12.5 Vertical Furnaces

Vertical furnaces have replaced horizontal furnaces in many critical applications, despite a 4x cost factor. The reason is that by employing vertical furnaces device manufacturers can get: a) improved process control; b) improved particle performance; and c) improved automation capabilities. These three factors are the major economic drivers of the modern fab.

The advantages of vertical furnaces over horizontal furnaces are now examined in more detail. First, the radial temperature gradient across the wafer is smaller in a vertical furnace since the wafer is lying parallel to the horizontal plane. Therefore it is less likely to be in a temperature gradient (keep in mind that heat rises). Also, the wafers are not partially shielded from the heat source by the cantilever arm that is used to put the wafers and boats in a horizontal furnace. Since the wafers can be made to sit flat and well centered in the closed fused-silica pedestal (called a *tower*) they can be more easily centered in the furnace itself. With the wafers and tower centered, it is easier to optimize the gas flow dynamics in the tube. Another factor is that wafer spacing is equal between all wafers in the load. Each wafer sits on the bottom of a slot in the tower, rather than balancing on its edge in a slot of the horizontal boat. Finally, the tower can be rotated in the furnace, and this tends to average out temperature and flow fluctuations.

The second set of advantages of vertical furnaces involves the fact that they can provide better ambient and particle control than can horizontal tubes. It is well known that the ambient around the wafer during the loading and unloading steps can greatly influence the film properties. Vertical furnaces can easily automate a controlled loading and unloading atmosphere. Horizontal furnaces have used so-called *white elephants* in an attempt to control the atmosphere during cooling. These devices are attached to the end of the furnace process tube to allow the wafer to cool in a controlled ambient. White elephants, however, are cumbersome at best, and extremely difficult to automate.

The third advantage of vertical furnaces is that it is relatively easy to integrate an ULPA filter air-handling system to the load station, so that these furnaces tend to have superior particle performance. Many furnace manufacturers also offer an automated wafer stocking feature where up to 18 cassettes of 25 wafers each can be maintained close to the furnace in an ultra-clean ambient. This provides for easy access to work-in-process (WIP).

Finally, vertical furnaces can be integrated into a *cluster tool* configuration where multi-step processing is possible without removing the wafer from the controlled environment. A major disadvantage of clustered vertical furnaces is their very low throughput.

Despite the wide acceptance of the vertical furnace some fabs that are processing chips with feature sizes greater than $0.5\ \mu\text{m}$ may still use horizontal furnaces - particularly if they are manufacturing cost-sensitive products. Yet even in such fabs, vertical furnaces are employed for

Table 8-9 TYPICAL SPECIFICATIONS FOR 200-mm VERTICAL FURNACES⁴⁹

Process	Target Thickness (nm)	Product Wafers in Load	Process Temp (°C)	Within-Wafer Uniformity (nm)	Wafer-to-Wafer Uniformity (nm)	All Point 3 σ (nm)	Particle D ₀ /cm ² $\geq 0.2\ \mu\text{m}$
Thin Cl "dry" oxides	5.0	150	800	± 0.25	± 0.25	± 0.40	< 0.01
Thin "wet" oxides	5.0	150	800	± 0.25	± 0.25	± 0.40	< 0.01
Thick "wet" oxides	100.0	150	950	± 0.25	± 0.25	± 0.40	< 0.01

the more critical processing steps. Table 8-9 shows typical process specifications for an advanced vertical furnace.

8.12.6 Anatomy of a Vertical Furnace

A schematic of a vertical furnace is shown in Fig. 8-26. The key features of the furnace are briefly described. The *heater element* usually consists of 5-7 independent heating zones. The heating elements are fabricated from Kanthal, MoSi_2 , or advanced powder metallurgical wires. By proper control of the zones it is possible to obtain a *flat temperature zone* of $\pm 0.5^\circ\text{C}$ for up to 30 inches in length. Such a long flat zone allows for up to 150 wafers per batch. A water-cooled, *stainless-steel jacket* surrounds the furnace and keeps the exterior cool. The *process tube* used is usually fused silica. Some furnace processes use a double-walled tube so inert or chlorinated gases can flow between the walls to help keep impurities out of the process region. The *fused-silica tower* is used to hold the wafers in a horizontal position. *Temperature measurement and control* is accomplished by placing thermocouples in the furnace. The control algorithm can be either PID or model-based control, both methodologies will be briefly discussed below. Finally, the *gas injection hardware* is required to bring controlled levels of process gas into the tube. The gas system (sometimes called the *gas jungle*) is usually fabricated from special stainless-steel components.

When executing a process, wafers are loaded into the fused silica tower from a cassette by a robotic mechanism. When the loading is complete either the tower moves into the furnace, or in the case of one manufacturer, the tower is stationary and the furnace moves over the wafers. The wafers are inserted into the furnace hot zone at low temperature ($600\text{--}700^\circ\text{C}$) to prevent crystal slip. Once the wafers are in the furnace, the temperature is slowly ramped to the process temperature. After reaching the process temperature the wafer batch is stabilized before the process begins. After the process is finished the wafers are slowly cooled until they reach a safe temperature for withdrawal from the furnace. Table 8-10, shows the steps of a typical furnace process.

The temperature control system used in furnaces has recently been improved. In the past, temperature control was executed by comparing the temperature measured by control thermocouples in the furnace to the set-point temperature on a controller. Corrections to temperature were made using a *proportional/integral/differential* (PID) control algorithm. This algorithm works by adding together three terms that are functions of the error between the measurement and the set point. PID constants are tuned from experience or some rules of thumb. More sophisticated strategies use the signals from both the spike and profile thermocouples. The PID approach assumes no coupling between adjacent heating zones, but in reality that is not the case. As a result profiling a furnace can take up to 12 hours and it may be required as often as once per week. Despite careful profiling, the PID control is reactive and has a tendency to overshoot (or undershoot) the control point, particularly during ramp. A newer technique called *model based temperature control* (MBTC) has recently been introduced.⁵⁰ This MBTC relates to a broad range of control strategies that uses systematic process modeling to determine the control algorithm. To control furnace processes, an advanced mathematical thermal model of wafers and wafer environments (including the furnace) is used to estimate and control within-wafer and wafer-to-wafer temperatures. There is a commercial software package, marketed by RelMan, Inc., that provides the advanced models necessary for such control. The use of MBTC, when measured against PID control, gives as much as a 50% improvement in oxide film uniformity for a furnace oxidation process.

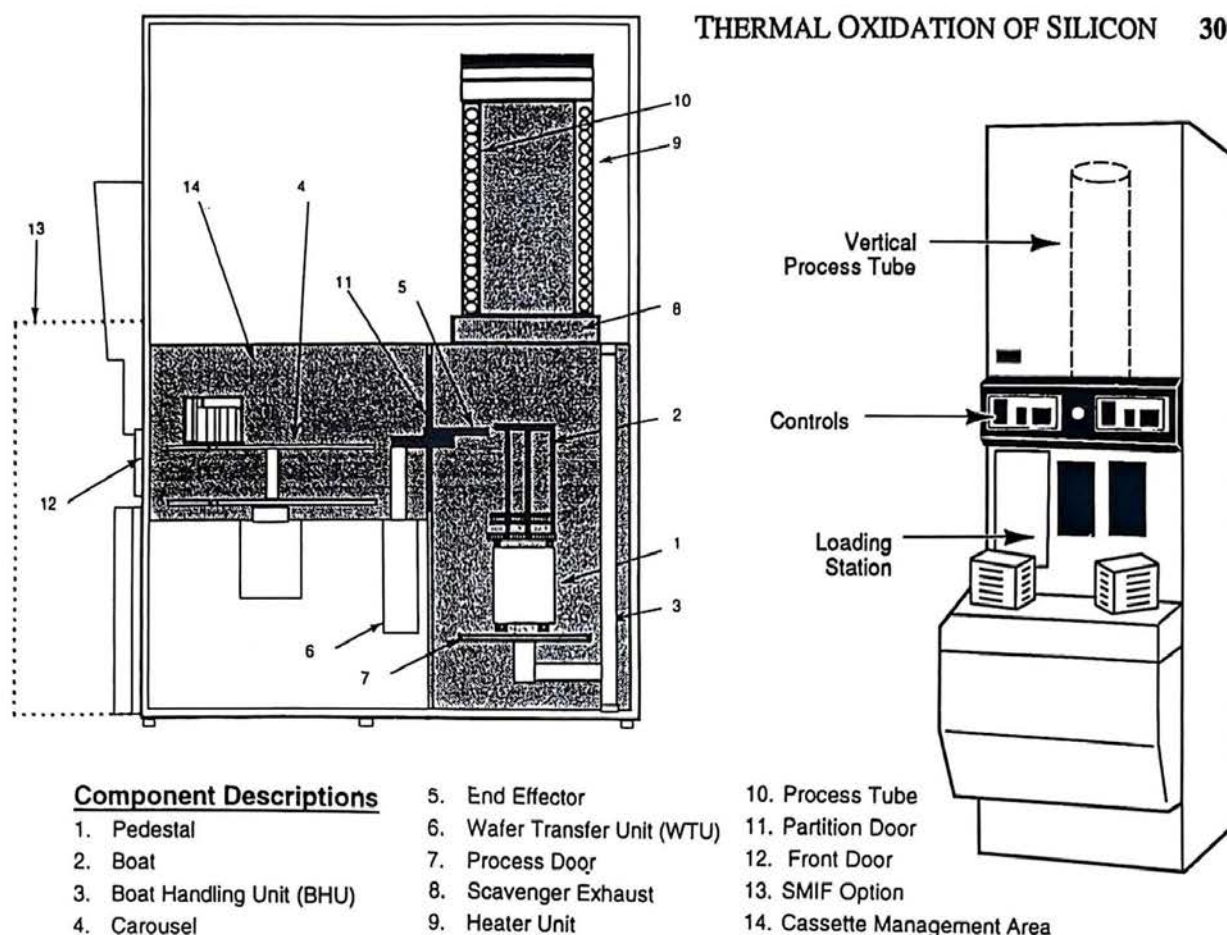


Fig. 8-26 Schematic of a vertical furnace showing its individual components. Courtesy of Silicon Valley Group.

8.12.7 Fast-Ramp Furnaces

The cycle time in a conventional vertical furnace can be as much as 4.5–5 hours. As a result, it is necessary to use large batch sizes (150 wafers) to achieve acceptable *throughput* (i.e., *wafers per hour*, or WPH) and *cost of ownership* (CoO). One of the major limitations is the slow ramp-up and cool-down from the load/unload temperature to the process temperature. In a conventional vertical furnace the ramp-up rate is 10°C/min and the cool-down rate is about 3°C/min, with the loading temperature being 650°C, and that for unloading being 700°C.

As a practical example, consider a cycle time comparison for a 900°C thermal oxidation process. It takes 25 min to reach the process temperature (based on the ramp-up rate of 10°C/min), and 100 min to cool down (based on a cool-down rate of 3°C/min). The cool-down rate is limited by the ability of the furnace to lose heat, and the ramp-up rate by the ability to maintain a small temperature gradient across the wafers. It is crucial to maintain a small gradient to keep wafers from undergoing thermal-induced slip (Chap. 2). In addition to the long cycle time for the process, wafers are exposed to a large thermal budget (area under the t-T curve).

Several practical solutions are available to reduce the cycle time and thermal exposure. One of the most successful solutions has been introduction of the single-wafer ultra-fast-ramp (50–75°C/sec) *rapid thermal processor* (RTP). The RTP system is discussed in detail later in this chapter. In this section two techniques that use modified vertical furnaces are described. The first is called *Extended Ramp* (ER) while the second is termed *Small Batch Fast Ramp* (SBFR). The ER uses a large batch of wafers (150) but the ramp rates are increased to 15°C/min on heating and to 15°C/min on cooling. For the same gate oxide process discussed above, the

ramp-up time is now 16.7 min, and the cool-down time is 13.3 min. If all other process times remain the same, the cycle time is reduced by 62 min. Since the wafer load size is identical to the standard furnace the throughput (wafers/total process time) will improve. The t-T product is also substantially reduced. Despite the higher ramp rates the temperature gradient across the wafer still remains below the value that will result in crystal slip. The challenge for making the ER furnace is to improve the heating and cooling rates. This has been accomplished by modifying the hardware. The modifications include: a) improved heaters with higher power-handling capability; b) built-in high-coolant air flow to achieve high cool-down rates, and c) special "ring" boats that increase the thermal mass of the wafers to minimize thermal gradient across the wafers. Notwithstanding these improvements there is still the possibility of crystal slip for process temperatures above 850°C. In addition, as the wafer size increases to 300 mm the problem of maintaining thermal uniformity across the wafer is exacerbated. Furnace manufacturers have had to increase the wafer pitch for 300-mm wafers, forcing a reduction in the load size to less than 100 wafers.

The second approach, SBFR, employs a different technique to achieve even greater ramp rates. First, the batch size is reduced from 150 to about 50 wafers and the spacing between wafers (pitch) is increased to 15–25 mm. The reduced batch size and the increased wafer pitch allows the furnace to ramp-up at 100°C/min and cool-down at 50°C/min. The thermal gradients that develop can still be quite severe and care must be exercised not to exceed the slip limit of the wafers. Furnace manufacturers have developed special fused-silica towers and have designed the furnace heaters and furnace mass to minimize thermal gradients. Although the thermal gradients may be below the slip limit they can still pose a problem with process non-uniformity across the wafer.⁷⁰ Semitool introduced a vertical SBFR with a ramp rate of 75°C/min.

Now examine the cycle-time savings by using a SBFR furnace for the same 900°C oxidation process considered above. To minimize thermal shock, in the SBFR furnace the wafers are inserted and removed at a temperature 200°C lower than in standard or ER furnaces. On ramp-up the wafers reach process temperature in 4 min, and on cool-down in 4 min. This results in an overall savings of 83.5 min of ramping, a respectable improvement. However, the ramps do not account for all of the process time since loading/unloading, stabilization, and wafer-transfer takes the same time, regardless of the ramp rate. Since the SBFR furnace can only accommodate 50 wafers rather than 150, the throughput will actually decrease for SBFR furnaces.

Table 8-10 shows the time for a full 10-minute oxidation process, comparing conventional, ER, and SBFR furnaces. From Table 8-10 it can be seen that the throughput increases for an ER process but decreases for SBFR process. SBFR furnaces, however, have other benefits. The t-T budget is reduced and the cycle time for a small wafer lot (25 wafers) can be reduced. Cycle time reduction is important during process development and for IC companies that run low-volume production.

8.12.8 Rapid Thermal Processing

Rapid thermal processing (RTP) is a *single-wafer, fast-ramp* thermal processing capability that can be used to heat a wafer from room temperature to 1100°C in a matter of seconds. Typical ramp rates used in RTP are 50–75°C/sec, compared to < 2°C/sec for the fastest-ramp furnaces. More recently, RTP systems have boasted controlled ramp rates exceeding 150°C/sec.⁵¹ The use of RTP as the tool of record for thermal processes has grown significantly since 1996. It is projected to grow even faster in the next decade. The advantages of RTP over furnaces have been

**Table 8-10 HYPOTHETICAL RECIPE FOR A 900°C, 10 MIN GATE OXIDATION
COMPARING: 1) CONVENTIONAL; 2) ER; AND 3) SBFR FURNACES⁵²
(ALL TIMES IN MINUTES UNLESS OTHERWISE NOTED)**

Step Number	Process Step	Standard Furnace	ER Furnace	SBFR Furnace
1	Load Cassettes	12.0	12.0	12.0
2	Load Wafers	22.1	22.1	7.2
3	Insert Wafers	11.2	11.2	2.6
4	Ar/N ₂ Purge	5.0	5.0	5.0
5	Ramp to Process Temperature	25.0	16.7	4.5
6	Stabilization at Process Temperature	5.0	10.0	10.0
7	Oxygen Stabilization	0.5	0.5	0.5
8	Oxidation Process	10.0	10.0	10.0
9	POA (at Same Temperature)*	30.0	30.0	30.0
10	Ramp to Pull Temperature	66.7	13.3	4.0
11	Pull Load	11.2	11.2	11.2
12	Cool Down	30.0	30.0	30.0
13	Unload Wafers	22.1	22.1	7.2
14	Unload Cassette	12.0	12.0	12.0
Total Process Time (hrs)		268 (4.46 hr)	207 (3.44 hr)	121 (2.03 hr)
Throughput (Wafers/hr)		33.6	43.6	24.6

* If POA is done at higher temperature, as it often is, the time for such ramp-up and cool-down must be added.

enumerated in various articles.⁵² Some of the more noteworthy advantages are listed below.

1. Reduced thermal budget: As a result of the rapid ramp rates in an RTP system the wafer experiences a smaller thermal exposure than a wafer in a batch furnace, despite the higher process temperature required for RTP.

2. Single wafer processing: RTP uses single-wafer processing, which is more compatible with the rest of the fab tools. Therefore, it is easier to run a balanced production facility. Furnaces, on the other hand, require a batch of 150 wafers for economical processing. For low volume fabs or during technology development, the use of RTP processing can significantly reduce the total cycle time of the whole process.

3. Higher temperature processing: Since RTP uses short process times, most of the anneal steps require a higher temperature than in a furnace. For example, a boron source/drain implantation anneal in a furnace can be 30 min at 900°C, while equivalent annealing in a RTP system is 10 sec at 1050°C. An advantage of using higher temperature/short time anneal steps is that *transient-enhanced diffusion* effects (*TED*, see Chap. 9) are reduced. Growing a gate oxide at higher temperatures in a rapid thermal oxidation system (RTO) is known to result in an improved oxide (since the higher temperature produces a smoother SiO₂/Si interface).²⁸

4. Improved gas distribution: RTP systems are small and symmetrically designed. As a result, process gas distribution into the system is more uniform than in a furnace.

5. *Better control of process ambient:* RTP systems are very easy to integrate onto a vacuum-tight central handling system that could bring wafers in and out through a vacuum loadlock. This approach makes it unlikely that ambient gases will enter the process chamber. The advantage of this is during silicide formation, where small quantities of oxygen adversely affect the resistivity of the silicide.

6. *Wafer size scaling:* As wafer sizes reach 300 mm and larger, processes developed in RTP do not need to undergo radical change to be successful. If a furnace is used, the process must be redesigned (since the wafer pitch for the larger wafers is usually increased).

7. *Integrated Processing:* It is extremely easy to *cluster* and integrate several processes on a single central handling system if the thermal processes use RTP. The advantages of integrated processing include: a) no exposure to the ambient between process steps; b) reduced particles; and c) less wafer handling. An example of a clustered process that is under development is an integrated gate stack. This tool includes: a) a vapor HF pre-clean to remove native oxide; b) a rapid thermal oxidation unit to grow the thin gate oxide; and c) a doped polysilicon deposition unit to deposit the gate electrode material.

Despite these advantages RTP has not replaced furnaces in all applications. Even though RTP may perform better in some applications, furnaces perform well enough at 0.25 μm feature size. Moreover, furnaces have a lower cost of ownership (CoO) for some applications. It is conjectured that as the feature size shrink to 0.18 μm and below, and the wafer size increases to 300 mm, the use of RTP in production processes will continue to expand.

Some of the applications where RTP has been accepted as the process tool of record (PTOR) in major fabs throughout the globe are: a) source/drain anneal; b) BPSG reflow; c) titanium and cobalt silicide formation; d) contact alloying; and e) WSi_2 annealing. Other applications where RTP is emerging as the tool of choice are: a) growth of STI liner oxides to enhance corner rounding; b) growth of high quality thin gate oxides and oxynitrides; and c) polysilicon re-oxidation. The recent introduction of steam oxidation in RTP tools opens this technology to many new applications.⁵⁴

8.12.9 RTP Systems

In the past, RTP systems have suffered from two design shortcomings. First, it was difficult to maintain uniform temperature across the wafer during the rapid ramp-up and cool-down steps. Second, it was difficult to accurately measure wafer temperature, particularly as the emissivity of the wafer varied. The emissivity of silicon can vary widely depending on the type and thickness of films on the silicon. These design issues have been solved for some of the commercially available RTP systems, and as a result RTP has grown in popularity. AG Associates (U.S.) has been, a long time supplier of RTP furnace chambers. By the late 1990's Applied Materials (U.S.) had the largest RTP market share, with Steag AST Elektronik (Germany) and AG Associates as the next largest suppliers. To illustrate the design solution the Applied Materials RTP Centura® will be discussed as the model.⁵¹ Other commercial systems have met some of these challenges in different ways and readers are directed to the product literature of the companies manufacturing those systems.⁵⁶

Figure 8-27 shows a cross section of the Applied Materials system. The major sub-systems are: a) the chamber; b) the lamp assembly; c) the wafer support system; and d) the fiber optic temperature probes and temperature control. Each of these sub-systems will briefly be discussed. The *chamber* is constructed from stainless steel and is water cooled to maintain a low temperature. Process gases flow into the side of the chamber from an injection port located at

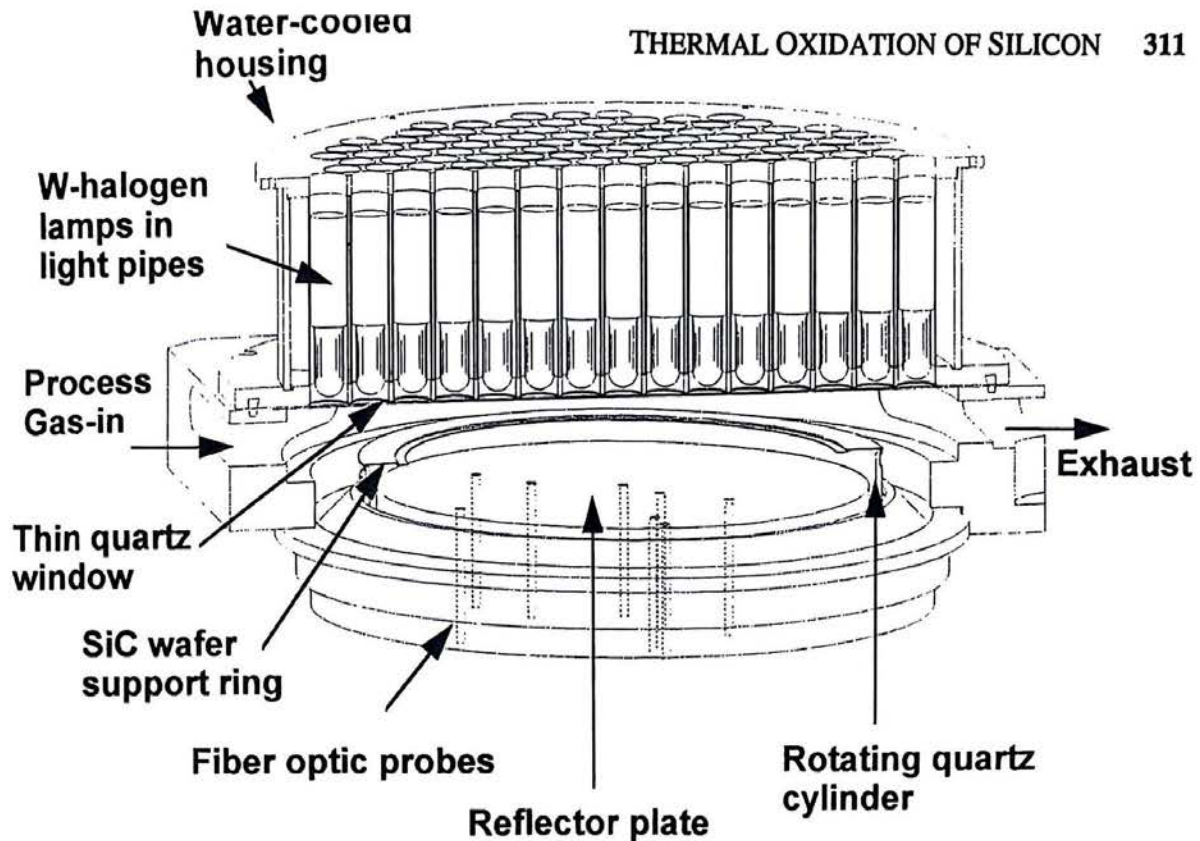


Fig. 8-27 Schematic cross-section of a rapid-thermal processor. Courtesy of Applied Materials, Inc.

the wafer loadlock. The system can be operated at atmospheric or reduced pressure. Reduced pressure is achieved by pumping with a mechanical or turbo pump. The *lamp housing* sits on top of the chamber and consists of 187 W-halogen lamps inside collimating light pipes. These direct the energy to the wafer. The lamps are arranged in a honeycomb structure to give the closest packing of radiant energy and fine spatial resolution of the lamp output. A thin (4 mm) fused-silica window separates the lamp head from the process area of the chamber. For reduced-pressure applications the window is replaced by a thin adapter plate, consisting of a sandwich arrangement with a water-cooled honeycomb in the middle of two quartz windows. The area between the windows is pumped so the honeycomb supports the windows when the system is at reduced pressure. The honeycomb matches the lamp-head honeycomb to ensure good transmission of the radiant energy to the wafer. The lamps are arranged into twelve independently powered circular groups combined into six independently controlled radial zones. Independent control permits the precise power level to be directed to a particular lamp zone to adjust for temperature deviation from the set point. Temperature measurement is discussed below.

The wafer rests on a Si-coated SiC support ring. The support ring is placed on a fused-silica cylinder which extends to the chamber bottom where it is supported by a bearing. The bearing is coupled to an external motor which rotates the cylinder, and therefore the wafer. Wafer rotation smooths out thermal and gas flow variations and results in improved process uniformity.

The *temperature measurement and control* is accomplished by using six pyrometers mounted on the bottom of the chamber, opposite the light source. A schematic of the control system is shown in Fig. 8-28. The pyrometers measure the temperature from the backside of the wafer and are spaced along the bottom plate (reflector plate) of the system so each one is looking at a different point on the wafer. As the wafer rotates each pyrometer samples an arc on the wafer at various radii out to the edge of the wafer. The pyrometers sample temperature at a rate of 20 Hz. The temperature measurement from each of the pyrometers is fed into a multi-input/multi-output controller. The measured temperature is compared with the recipe (input) temperature in

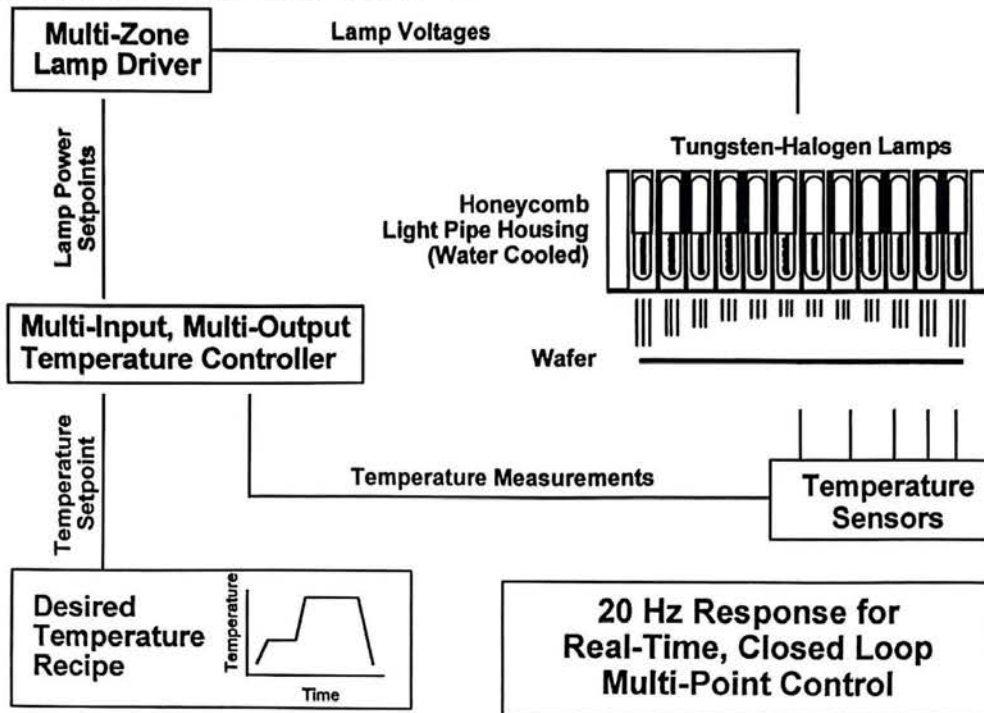


Fig. 8-28 Schematic of a closed-loop temperature control system. Courtesy of Applied Materials, Inc.

the controller, which uses a sophisticated algorithm to analyze the data and correct the output to each lamp zone. As a result, any particular lamp zone can receive more or less power in order to correct the wafer temperature to the recipe temperature. Figure 8-29 shows a recipe temperature profile and the measured temperature profiles for the 6 pyrometers for a 1050°C process with a 50°C/sec ramp rate. The temperature control is better than $\pm 2^\circ\text{C}$ on the ramp and $\pm 0.5^\circ\text{C}$, during the process (soak time). The control of temperature during the ramp is extremely important for good process control. A general discussion of temperature measurement in RTP systems is given in Ref. 67.

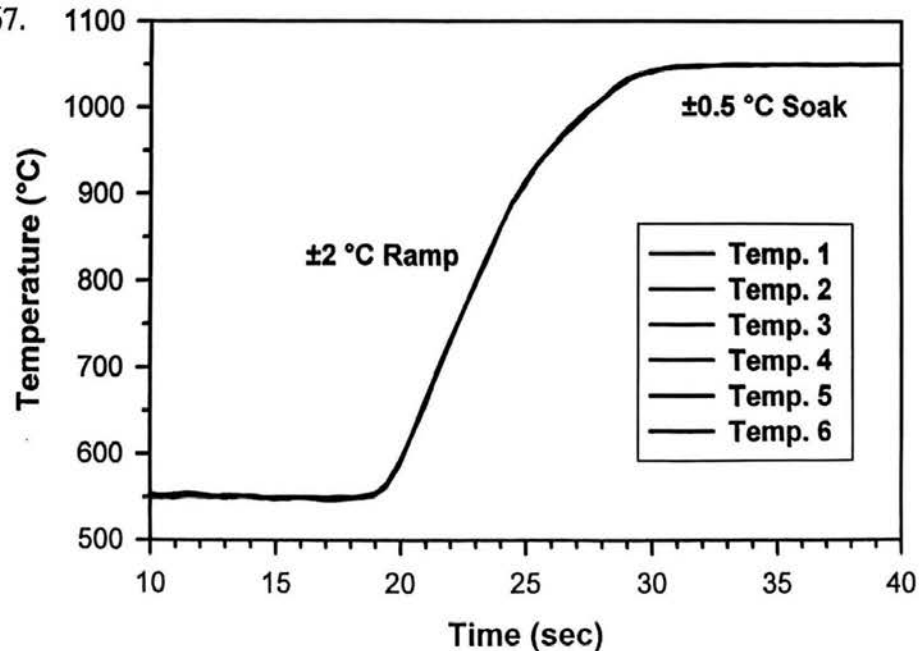


Fig. 8-29 Temperature-time profile of a rapid thermal processor during ramp and soak (process time). Six temperature measurements are made across the wafer. The temperature variation is within $\pm 2^\circ\text{C}$ during the ramp and within $\pm 0.5^\circ\text{C}$ during the soak. Courtesy of Applied Materials, Inc.

The next issue involves the sensitivity of the temperature measurement to the wafer backside emissivity. The emissivity of the wafer can vary depending on the surface conditions or films on the wafer backside. For example, at 1000°C, the measured temperature can vary by as much as 100°C at the wavelength used by the pyrometer (0.93 μm) as the emissivity varies between 0.3 and 1. The Applied Materials tool addresses the emissivity variation by using a *virtual blackbody effect*. This effect is achieved by making the reflectivity of the cavity equal, as nearly as possible, to unity. A perfect blackbody has an emissivity value of 1. Since the chamber has a finite length this is not possible. However, a good approximation to a blackbody can be achieved by making the bottom plate (which contains the pyrometers) highly reflecting in the spectral region used for pyrometer measurement. This is done by coating the reflector plate with a film stack that results in high reflectivity. The virtual blackbody significantly improves emissivity sensitivity, but does not completely eliminate the effect. Further improvements can be achieved by a technique that uses a second pyrometer near one of the measuring pyrometers. The second pyrometer has a different shape to the hole that allows radiant energy to enter. Although the second pyrometer is sampling the same region of the wafer as the measuring pyrometer, it will record a different temperature. The reason for this is that it receives reflected energy from a wider angle of acceptance. An algorithm has been developed which uses the apparent temperature difference between these two pyrometers to calculate the emissivity of the wafer. Once the emissivity is known, the true temperature can be easily calculated and adjustments can be made. The reader is referred to the paper by Peuse *et al.* for additional details on the temperature measurement techniques used in the Applied system.⁵⁵

8.12.10 High Pressure Oxidation (HIPOX) Systems

The most widely used high-pressure oxidation (HIPOX) system is supplied by GaSonics International.⁵⁷ The system is configured in the same way as a vertical furnace, and is referred to as a *vertical high pressure (VHP) furnace*. The VHP operates to pressures ranging up to 25 atm, and in the temperature range from 600–1200°C. The main components of the system include: a) the pressure vessel; b) the zone heaters; c) the fused-silica process tube; and d) the fused-silica wafer tower. There are four independently controlled heaters, with two additional plug heaters to help achieve improved temperature uniformity. The fused-silica tube and wafer cassette are similar to the vertical furnace hardware as previously described.

After the wafers are loaded into the furnace and sealed, the system is pressurized to the set point. When either filling or venting the tube, a differential pressure is maintained between the tube and the vessel. The vessel is pressurized with nitrogen, while the process tube can be pressurized with either nitrogen, oxygen, steam, or a mixture of any of these gases. Steam is generated in the reactor at high pressure by introducing deionized water onto a hot fused silica *flash boiler*. Hence, wet, dry, and dilute steam oxides can be grown at high pressures.

As mentioned earlier in the chapter, HIPOX can improve the field oxidation process for LOCOS and more advanced isolation techniques (such as *poly-buffered LOCOS* [or *PBLOCOS*]). When HIPOX is used in this application two problems occurring in normal atmospheric oxidation are alleviated: 1) bird's beak encroachment, and 2) gate oxide thinning.

Bird's beak encroachment occurs because the oxidation proceeds laterally under the protective nitride layer while the field grows vertically. Since the field oxide is typically grown to 200–400 nm, the growth kinetics are for the most part diffusion-limited (parabolic) and take a long time to reach the desired thickness. On the other hand, the lateral oxide growth tends to be less self-limiting and continues to grow during the field oxidation. High-pressure oxidation

provides a significant increase in the vertical growth rate without a concomitant increase in the lateral growth rate. An added benefit is that with HIPOX a field oxide can be grown with dry oxidation, which is known to eliminate the Kooi effect.

The second improvement of HIPOX involves the *field-oxide thinning effect* in LOCOS isolated structures. That is, the field oxide growth rate becomes progressively smaller as the isolation spaces decrease below 1 μm . This can result in a 30% reduction in the field-oxide thickness for 0.4 μm -wide isolation spaces. As previously described, several models have been proposed to explain field-oxide thinning. Whatever the cause, use of HIPOX tends to reduce such thinning. For example, a field oxide grown in a 0.5 μm space at 1000°C and at 1 atm results in about 16% thinning, while at 20 atm only about 11% thinning occurs.

8.13 THICKNESS MEASUREMENTS OF OXIDE FILMS

The accurate measurement of oxide film thickness is an important process-control tool used during ULSI wafer fabrication. For instance, the thickness of gate oxides in ULSI MOSFETs (which is currently in the range of 2–4 nm) must be tightly controlled, as it is one of the parameters that directly determines the device threshold voltage V_T . Various techniques are available for measuring oxide thickness, the most important being: a) optical interference; b) ellipsometry; c) capacitance; and d) physical thickness measurements. Microprocessor-controlled optical interference and ellipsometry tools are the most widely used for thickness measurements in a production environment.

8.13.1 Optical Measurements of Oxide Film Thickness

The *optical interference* method is simple to use for a single dielectric film on silicon.⁶⁶ Oxide is the most frequently measured film, but the technique works for all transparent films as long as the index of refraction is well known. In the ensuing discussion the focus will be on the measurement of oxide films. The technique is nondestructive and can be used to routinely measure oxide thickness from less than 10 nm to more than 1000 nm. The method is based on the interference that occurs between light reflected from the air/SiO₂ interface and the Si/SiO₂ interface. The equation governing this interference is:

$$x_{\text{ox}} = \lambda (g - P)/2n^* \quad (8.36)$$

where

$$n^* = (n_i^2 - \sin^2\theta)^{1/2} \quad (8.37)$$

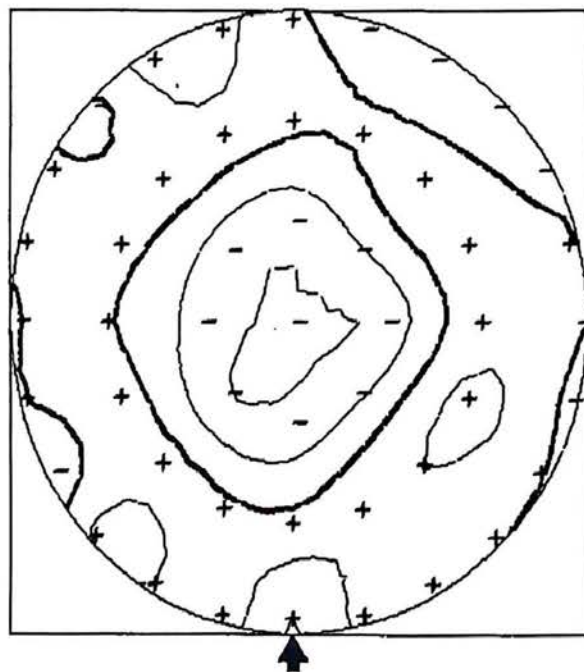
and n_i is the index of refraction of the oxide film; θ is the angle of incidence of the light relative to the substrate; g is the order of the interference; λ is the wavelength of the incident radiation; and P is the net phase shift (i.e., $P = [\phi_s - \phi_f]$, where ϕ_f is the phase shift at the air/SiO₂ interface, and ϕ_s is the phase shift at the Si/SiO₂ interface). Equation 8-37 assumes that n_i is independent of wavelength, and that multiple reflections can be neglected. The equation predicts that energy minima in the reflected light occur when $g = m + 1/2$, where m is an integer ($m = 1, 2, 3, \dots$), and energy maxima result when $g = m$.

The most widely used implementations of this technique vary the wavelength of the light, while holding the incident angle constant. A spectrophotometer that supplies the incident light in the ultra-violet-to-visible range is used in performing the measurement. The intensity of the reflected light from the sample is measured as a function of wavelength. The intensity minima are then fitted to Eq. 8-36 to calculate this film thickness.

In the early days of IC technology, film thickness was calculated using a *fringe chart*,⁵⁸ in which the film thickness is plotted as a function of wavelength minima for various orders of reflection, g. Film thickness was measured independently by multiple-beam interferometry, an accurate but complex measurement technique. Thus, such charts account for the phase shift terms (ϕ_f and ϕ_s), multiple reflections, and dispersion effects. Fringe charts have been replaced by complex algorithms that calculate the thickness directly from the reflected intensity spectra.

The measurement is easily automated and rapid, and numerous points can be quickly measured across a wafer. A common output of such a measurement is a *thickness contour map*, which shows the variation of thickness across the wafer. Such a map is shown in Fig. 8-30 for a thin oxide layer on a 200-mm wafer. The mean thickness line is shown along with deviations from the mean. The standard deviation in percentage of mean is also calculated. The user can specify any value of edge exclusion. (Typical values of edge exclusion are 3 mm.) Such automated optical interference instruments are widely used to measure the thickness of SiO_2 on Si, Si_3N_4 on Si, photoresist on Si, and poly-Si on oxide. The system is also capable of measuring the thickness of multiple-film stacks.

Another optical technique for measuring the oxide thickness (as well as the optical index of refraction), is *ellipsometry*. Archer was the first to use this method, and a detailed analysis of the method is found in his original paper,⁵⁹ or a shorter discussion in Schroder.⁶⁸ In general, the index of refraction, n_i , of thermally grown SiO_2 is a constant at a given light wavelength (e.g., with a value of $n_i [\text{SiO}_2] = 1.46$ at $\lambda = 546 \text{ nm}$), while the value of n_i for deposited films depends strongly on deposition conditions. As a result, for process control only a thickness measurement is usually needed for thermal oxides, but for deposited films both a measurement of the thickness and the index of refraction are needed. Using ellipsometry it is possible to reproducibly measure oxide films to 2 nm thickness, and in some cases to as thin as 1 nm.



Process:
 900 °C, 30 seconds
 Dry RTO
 Atmospheric Pressure

Thickness: 22.52 Å
Std. Deviation: 0.5%
Range: 0.6 Å
49pt contour map
3mm edge exclusion
Contour spacing: 0.15Å

Fig. 8-30 Contour plot showing the thickness uniformity across a 200-mm wafer. Dark spaces represent average thickness. Each contour represents only 0.15 Å. Courtesy of Applied Materials, Inc.



Fig. 8-31 Photograph of an automated thin film thickness measuring system. Courtesy of KLA-Tencor.

The ellipsometry technique makes use of the change of state of the polarization of light when it is reflected from a surface (in this case the Si/SiO₂ interface). The state of polarization is determined by the relative amplitude of the parallel and perpendicular components of the radiation, and by the phase difference between these two components. The polarization change depends on the optical constants of the silicon, the angle of incidence of the light, the optical constants of the film (i.e., the index of refraction n_i and the extinction coefficient k), and the film thickness. If n_i and k of the substrate are known and if the film is non-absorptive at the wavelength being used (i.e., $k = 0$), the state of polarization of the reflected beam depends only on n_i and the thickness of the transparent film.

In the past, ellipsometry measurements required a delicate and experienced skill to determine the extinction points (which allow the state of polarization to be directly calculated). Automated equipment, which provides values of n_i and thickness has since been developed (Fig. 8-31). Such instruments are capable of measuring the oxide thickness and n_i in films 2-nm to 2000-nm thick. Thickness measurements of ± 0.003 nm in accuracy and ± 0.007 nm in repeatability are possible, and the index of refraction can be determined to within an accuracy of ± 0.005 units. These types of film thickness measuring systems are quite versatile and thus they are also routinely used to measure the thickness and index of refraction of other dielectric films on silicon substrates (such as silicon nitride [Si₃N₄], silicon oxynitride [SiO_xN_y], and tantalum pentoxide [Ta₂O₅]). They can also be used to determine the thickness of multi-layer film structures such as polysilicon on SiO₂ on Si, or photoresist on SiO₂ on Si.⁶⁹

8.13.2 Electrical Measurement of Oxide Film Thickness

The *capacitance method* can also be used for measuring the electrical thickness of a dielectric film, but requires the fabrication of a MOS capacitor. For an oxide film, the oxide thickness x_{ox} is given by:

$$x_{ox} = C_{ox}A/\epsilon_{ox}\epsilon_0 \quad (8.38)$$

where C_{ox} is the measured oxide capacitance in accumulation (in F/cm²), A is the area of the

gate (in cm^2), ϵ_{ox} is the dielectric constant of SiO_2 , and ϵ_0 is the permittivity of free space. The area of the gate electrode of the capacitor must be large compared to the oxide thickness in order to minimize capacitive fringing effects (which introduce errors in the measurement). The area of the gate electrode needs to be accurately known, since it also enters into the calculation. The capacitance is measured with the capacitor biased so that the Si is in accumulation (negative voltage on the gate for p -type and positive voltage for n -type). This helps ensure that the silicon capacitance does not contribute to the measurement. For thin oxides the electrical thickness tends to be larger than the optical thickness as a result of a small amount (0.4 nm) of depletion that does occur in the silicon surface. Wafer maps of electrical thickness can be obtained by measuring 49 points across the wafer. Electrical thickness is mostly used for oxide measurements where the dielectric constant of the film is well known.

8.13.3 Physical Thickness Measurement of Oxide Films

The most accurate way to measure the film thickness is to prepare a cross-section of the film on the silicon substrate and take a high resolution transmission electron microscope (HRTEM) picture of the stack. Since the silicon lattice spacing can readily be seen in such a picture, the distance between lattice planes can serve as a built-in thickness standard. This standard can be used to accurately measure the thickness of the oxide (or any other film) on the wafer. In addition, the roughness of the interface between silicon and the film can be observed at the same time. HRTEM sample preparation is tedious and costly so it can only be used sparingly. Therefore it is not used for determining uniformity across a wafer, except on a very limited sampling basis. Figure 8-32 shows such a HRTEM cross-section.

8.14 TRENDS IN THIN OXIDE GROWTH

8.14.1 Limitations of Silicon Dioxide

The SIA Roadmap for gate oxide thickness is given in Table 8-2. As the feature size scales the gate oxide thickness scales to less than 1.5 nm and eventually to 1 nm (less than 3 atomic layers!). The gate oxide is scaled in order to maintain equivalent drive current I_D as the MOSFET

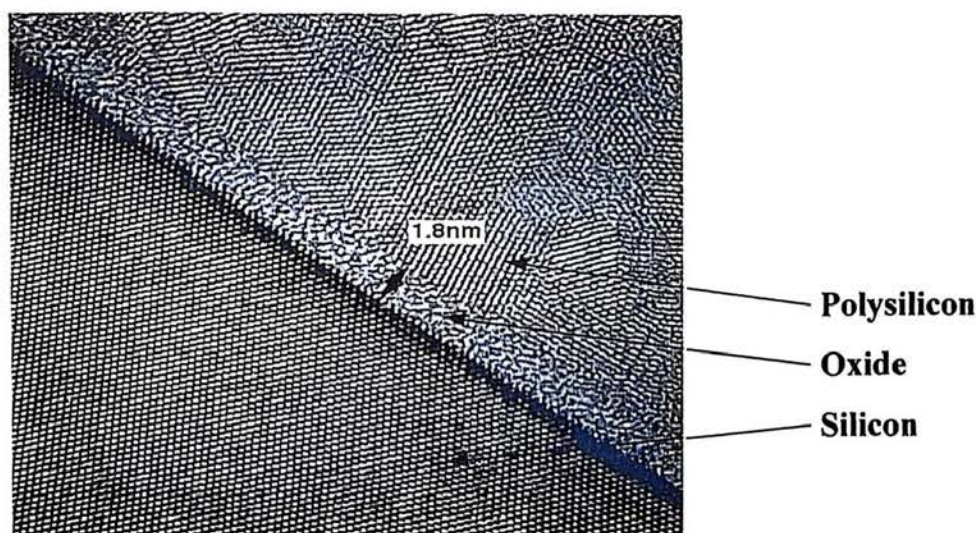


Fig. 8-32 High-resolution TEM showing a crystalline silicon/polysilicon structure. The lattice planes in the crystal are defined. Courtesy of Applied Materials.

gate length is decreased. Higher drive current directly relates to improvements in circuit speed, operating voltage, and power. There are numerous issues related to using thinner oxides in production. When looking at the limitations of SiO_2 as a gate oxide the fundamental physical limits (as well as manufacturing constraints) need consideration.

One important issue for ultra-thin oxides in a manufacturing environment is the ability to maintain *thickness control*. Since advanced devices require high channel doping, small changes in oxide thickness will result in large shifts in the threshold voltage. For an oxide thickness of 6.5 nm a variation of 4% (3σ) will cause the n -channel V_T to increase by 30 mV, which can decrease the circuit performance by 10%. A second concern is that boron from the p -doped polysilicon gates may diffuse through the oxide to the Si substrate, an effect known as *boron penetration*. When such boron enters the silicon surface it changes the doping in the transistor channel, causing a change in V_T . This problem is exacerbated by the use of ultra-thin gate oxides, because B diffusion increases as the thickness decreases. The presence of hydrogen or fluorine in the oxide is known to further enhance boron penetration. Boron diffusion can be retarded by introducing nitrogen into the gate dielectric, as discussed previously.

Polysilicon depletion is another factor that will ultimately limit how thin a gate oxide can be made. The polysilicon depletion effect increases the "electrical" equivalent thickness of the gate oxide. To compensate for this effect the oxide must be scaled even more aggressively than required from the scaling rules. It has been shown that polysilicon depletion may require an additional reduction of 10-40% of the gate oxide thickness.⁶¹ A proposed solution to this problem is to replace the polysilicon gate with a metal gate. The integration of metal gates into advanced ULSI processes, however, is not well established at this time.

8.14.2 Reliability Limits for Thin Oxides

Hot carrier degradation depends on the generation and trapping of hot carriers in the oxide. The number of hot carriers increases as the feature size of the device is scaled since the operating voltage does not scale with each new technology generation. Fortunately, however, the thinner gate oxide provides less trapping volume, thereby effectively reducing the hot carrier effect. Thus, it is believed that hot carriers will not become a limitation until device dimensions shrinks below 0.1 μm . *Plasma processing* can induce various types of traps and damage in thin oxides. These can cause premature breakdown and high leakage currents. Conveniently, as the oxide becomes thinner the induced interface trap density D_{it} also decreases for the same level of radiation. It is still prudent to design plasma tools that minimize plasma damage to the oxide.

8.14.3 Fowler Nordheim (F-N) Tunneling

Thinner oxides tend to operate at higher electric fields that can easily reach the *Fowler-Nordheim tunneling* (FN) regime ($\sim 6 \text{ MV/cm}$). FN-tunneling injects charge into the oxide, which can then be trapped and cause *stress induced leakage current* (SILC). The SILC can result in reduced *charge-to-breakdown* (Q_{BD}) and early breakdown failures. Scaling of the gate voltage helps this problem. The best approach to achieve continued useful scaling of oxides is to reduce the defect level. The relative thinning effect from the contemporary level of defects may limit the oxide thickness to only 3.4 nm for a device operating at 2.5V and 120°C.⁶¹ Therefore, it is necessary for the device manufacturers to improve the defect levels of the oxides. *Intrinsic pinholes* or *micropores* that form naturally as the oxide grows, may become the limitation since their density rapidly increases as the oxide becomes thinner. A potential solution is to use stacked thermal/CVD oxides to reduce the pinholes to acceptable levels.

The decisive limitation to oxide scaling may be the *direct (or quantum mechanical) tunneling* of carriers between the gate and the silicon substrate. Direct tunneling is a strong function of oxide thickness and the WKB approximation indicates that 1.8–2.5 nm may be the lower thickness limit for oxides arising from direct tunneling.⁶² That of course depends on the allowed leakage current of the device. Gate leakage is particularly important in DRAM devices, where the leakage current of the device affects the data retention time. The general leakage current limit for DRAM devices is less than 10^{-8} A/cm². These restrictions are not as severe for logic devices, where gate-current densities as large as 1 A/cm² are acceptable for some applications. It is generally assumed that SiO₂ as a gate dielectric will no longer be effective below 1.5–2.0 nm oxide thickness for DRAMs and below 1 nm oxide thickness for logic devices. Recently Momose *et al.*, reported that quite acceptable devices operating at room temperature have been fabricated with 1.5 nm gate oxides and 0.09 nm gate length.⁶³ It is not clear if making such thin gate oxides can become a high yielding, reliable, manufacturing process.

8.14.4 Alternative Gate Insulator Materials for MOSFETs

If SiO₂ eventually becomes unsuitable as the gate insulator in deep-submicron MOSFETs, what can be used instead? Regardless of what alternative material is suggested, process engineers will do everything possible to stay with oxide. This is because oxide is not only compatible with (poly)silicon and well behaved and understood, but it has also been used since the inception of MOSFET technology some 35 years ago. A replacement material will have to be thicker than oxide to reduce the direct tunnel current and yet maintain the same drive current capability. This can be achieved by using an insulator that has a higher dielectric constant (ϵ_{ins}) than does oxide ($\epsilon_{\text{ox}} = 3.9$). That is, it must be possible to achieve the same equivalent thickness t_{oxeq} as required for SiO₂, where the oxide equivalent thickness t_{oxeq} , is given by:

$$t_{\text{oxeq}} = t_{\text{ins}} (\epsilon_{\text{ox}} / \epsilon_{\text{ins}}) \quad (8.39)$$

For example, a gate insulator that is 10-nm thick and has a dielectric constant of 20 will behave in the same way in the device as an SiO₂ layer 2.0 nm thick. There are other stringent requirements placed on an alternative dielectric, since the material must be compatible with the process flow. There are currently four materials being investigated as a replacement for SiO₂: 1) silicon nitride (Si₃N₄); 2) tantalum pentoxide (Ta₂O₅); 3) titanium dioxide (TiO₂); and 4) barium strontium titanate (BST). Table 8-11 lists some key properties of these materials. Silicon nitride does not have as large a value of ϵ as do the alternative oxides, but the industry has worked with nitride for a long time. Nitride is being studied as the earliest replacement for oxide.

All of the materials investigated to date, however, suffer from manufacturing problems. If they are deposited directly on Si they tend to have a high level of interface states, D_{it} . The interface state density can be improved by growing a thin layer of oxide (1–2 atomic layers) on the silicon to passivate the surface. Ta₂O₅ and TiO₂ are commonly prepared by MOCVD from organic based precursors reacting with oxygen. As deposited, these films are typically amorphous, deficient in oxygen, and they contain some residual carbon. As a result, the as-deposited films have relatively high conductivity (i.e., they are leaky) and must be annealed in oxygen (or nitrous oxide) to improve their film properties. When annealed, commonly at elevated temperature (e.g., 800°C), the films tend to crystallize, which may also result in higher leakage currents. In addition, when Ta₂O₅ is deposited directly on silicon (or when it is subsequently annealed in oxygen), SiO₂ grows under the Ta₂O₅. As much as 2 nm of oxide can

Table 8-11 SELECTED PROPERTIES OF ALTERNATIVE DIELECTRICS

Material	Dielectric Constant (Thin film)	Breakdown Strength (MV/cm)	Preparation Methods	Leakage current	Maturity of Technology
SiO ₂	3.9	12-15	Thermal	low	Very mature and reliable
Si ₃ N ₄	7-9	10-11	LP, PECVD, Thermal nitride, JVD*	moderate	Experience with DRAM capacitors
Ta ₂ O ₅	20-26	3-5	MOCVD	mod-high	Some DRAM experience
TiO ₂	30	0.5	MOCVD	mod-high	Little experience
BST	190		MOCVD, PVD	low (thick film)	Some experience but complex material

* JVD is a jet vapor spray technique that directly deposits silicon nitride on the silicon wafer. For additional details readers should access the review paper on this subject by T.P. Ma⁶⁴

grow, thereby increasing t_{oxeq} . In summary, there is still a significant amount of work required to turn these alternative dielectrics into production-worthy materials.

REFERENCES

1. The National Technology Roadmap for Semiconductors Technology Needs, Semiconductor Industry Association, p. 46 (1997).
2. J.M. Stevels and A. Kats, "The Systematics of Imperfections in Silicon-Oxygen Networks," *Philips Research Repts.*, **11** p. 103 (1956).
3. A.G. Revesz, "The Defects Structure of Grown Silicon Dioxide Films," *IEEE Trans. Electron Dev.*, **ED-12**, 97 (1965).
4. B.E. Deal and A.S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, **36**, 3770 (1965).
5. E.A. Irene, "The Effect of Trace Amounts of Water on Thermal Oxidation of Silicon in Oxygen," *J. Electrochem. Soc.*, **121**, 1613 (1974).
6. J.R. Ligenza, "Effect of Crystal Orientation on Oxidation Rates in High Pressure Steam," *J. Phys. Chem.*, **65**, 2011 (1961).
7. W.A. Pliskin, "Separation of the Linear and Parabolic Terms in Steam Oxidation of Silicon," *IBM J. Res. Dev.*, **10**, 198 (1966).
8. D-B Kao, J. McVitte, W. Nix, K. Saraswat, "Two-Dimensional Thermal Oxidation of Silicon-I," *IEEE Trans. Elect. Dev.*, **ED-34**, 1008 (1987)
9. B.E. Deal, "Thermal Oxidation Kinetics of Silicon in Pyrogenic H₂O and 5% HCl/H₂O Mixtures," *J. Electrochemical Soc.*, **125**, 576 (1978).
10. B. E. Deal and M. Sklar, "Thermal Oxidation of Heavily Doped Silicon," *J. Electrochem. Soc.*, **112**, 430 (1965).
11. C.P. Ho, *et al.*, "Thermal Oxidation of Heavily Phosphorus Doped Silicon," *J. Electrochem. Soc.*, **125**, 665 (1978).
12. C. Ho, J.D. Plummer, "Si-SiO₂ Interface Oxidation Kinetics: Physical Model for the Influence of High Substrate Doping Levels I: Theory", *J. Electrochem. Soc.*, **126**, 1516 (1979), "II: Comparison with Experiments and Discussion," *J. Electrochem. Soc.*, **126**, 1523 (1979).
13. H.Z. Massoud, J.D. Plummer, and E.A. Irene, "Thermal Oxidation of Silicon in Dry Oxidation: Growth Rate Enhancement in the Thin Regime," I. Experimental Results, II. Physical Mechanisms," *J. Electrochem. Soc.*, **132**, 2685,(1985).

14. T. Hori, *Gate Dielectrics and MOS ULSI*, Springer-Verlag, Heidelberg (1997) p. 155.
15. G. Ereecke, *et al.*, *Proc. 3rd Internatl Symposium on Ultra Clean Processing of Silicon Surfaces*, Acco, Lueven, 1996, p. 69.
16. Schumaker, a Division of Air Products and Chemicals.
17. D.W. Hess and B. Deal, "Kinetics of the Thermal Oxidation of Silicon in O₂/HCl Mixtures," *J. Electrochem. Soc.*, **124**, 735 (1977).
18. The Clean Air Act Amendments of 1990, 40 CR Part 82, Protection of Stratospheric Ozone
19. M.J. McGeary, *et al.*, "Chlorine Precursors for Gate Oxidation Processes," *Mat'l Res. Soc.* (Pittsburgh, PA), **447**, (1996).
20. R. Ganesh, S. Kim, D. McClosky "Growth Kinetics & N₂ Profiles of Ultrathin Gate Oxides Grown Using High Pressure N₂O," *4th Internatl Conf on Adv. Thermal processing of Semiconductors*, RTP 1996.
21. R.R. Razouk, L.N. Lie, and B. E. Deal, "Kinetics of High Pressure Oxidation of Silicon in Pyrogenic Steam," *J. Electrochem. Soc.* **128**, 2214 (1981).
22. M. Depas, *et al.*, "Ultra-Thin Gate Oxides Below 3 nm Grown in a Cluster Tool Furnace," *Proc. 3rd Internatl Symposium on Ultra Clean Processing of Silicon Surfaces*, Acco, Lueven, 1996, p. 291.
23. L.N.Lie, R.R. Razouk, and B. E. Deal, *J. Electrochem. Soc.*, **129**, 2828 (1982)
24. M. Morita, T. Ohmi, E. Hasegawa, M. Kawakami, *Appl Phys. Letts.*, **55**, 562 (1989).
25. N.F. Mott, S Rogo, F. Rochet, A.M. Stoneham, "Oxidation of Silicon," *Phil. Mag. B*, **60**, 189 (1989)
26. S.J. Sofield and A. M. Stoneham, "Oxidation of Silicon: the VLSI Gate Dielectric?" *Semicond. Sci. Technol.*, **10**, 215 (1995).
27. H.Z. Massoud, *Extend Abs. of the 22nd Internatl Conf. on Solid State Devices and Materials*, Sendai, 1990, p. 1083.
28. S.A. Ajura *et al.*, "Re-Evaluation of the Benefits of Post-Oxidation Annealing on Sub-100Å Gate Oxide Quality," *IEEE Electron Device Letts.*, **17**, 282 (1996).
29. B.E. Deal, "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon," *IEEE Trans. Electron Dev.*, **ED-27**, 606 (1980).
30. B.E. Deal, M. Sklar, A.S. Grove, and E.H. Snow, "Characteristics of the Surface State Charge of Thermally Oxidized Silicon," *J. Electrochem. Soc.*, **114**, 266 (1967).
31. F. Montillo and P. Balk, *J. Electrochem. Soc.*, **118**, 1463 (1971).
32. E. Arnold, J. Ladell, and G. Abowitz, *Appl. Phys. Letts.*, **13**, 413 (1968).
33. E.H. Nicollian and J.R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*, John Wiley & Sons, Inc., New York, 1982, Chap. 12.
34. E.H. Snow, A.S. Grove, B.E. Deal, and C.T. Sah, "Ion Transport Phenomena in Insulating Films," *J. Appl. Phys.*, **36**, 1664 (1965).
35. T. Hori, *Gate Dielectrics and MOS ULSIs*, Springer-Verlag, Heidelberg (1997) p. 41.
36. Z. Chen *et al.*, "Deuterium Process of CMOS Devices: New Phenomenon and Dramatic Improvement," *Proc. Symposium on VLSI Tech.*, paper T18-3, June 9-11, 1998.
37. D.K. Schroder, *Semiconductor Materials and Device Characterization*, John Wiley, New York, (1990).
38. D.R. Young, *et al.*, *J. Appl. Phys.* **50**, 6366 (1979).
39. M. Aslam P. Balk, "Insulating film on Semiconductors" *Electrochem. Soc. Ext. Abs.* 1983 p. 103.
40. E.P. EerNisse, "Stress in Thermal SiO₂ During Growth", *Appl. Phys. Letts.*, **35**, 8 (1979).
41. A.S. Grove, O. Leistiko, and C.T. Sah, "Redistribution of Acceptor and Donor Impurities During Thermal Oxidation of Silicon," *J. Appl. Phys.*, **35**, 2629 (1964).
42. R.B. Fair, and J.C.C. Tsai, "Theory and Direct Measurement of Boron Segregation in SiO₂ during Dry, Near Dry, and Wet O₂ Oxidations," *J. Electrochem. Soc.*, **125**, 2050 (1978).
43. Y. Wang, J. Tao, S. Tong, T. Sun, A. Zhang, S. Feng, "The Oxidation Kinetics of Thin Polycrystalline Silicon Films," *J. Electrochem. Soc.*, **138**, 214 (1991).
44. J. Manoliu, *Semiconductor International*, April 1988, p. 90.

45. R.B. Marcus and T.T. Sheng, "The Oxidation of Shaped Silicon Surfaces," *Extended Abstracts, Electrochem. Soc. Mtg.*, **129**, p. 1728 (1982).
46. D.-B. Kao, J.P. McVittie, K.S. Saraswat, *IEEE Trans Electron Devices*, **ED-34**, 1008 (1987).
47. D.-B. Kao *et al.*, "Two-dimensional silicon oxidation, experiments and theory," *tech. Digest IEDM*, p. 388 (1985).
48. E.G. Seebauer, R. Ditchfield, "Fixing Hidden Problems with Thermal Budget," *Solid State Technology*, p. 111, October 1997.
49. Extracted from the SVG/Thermco Data Sheet for 8000AVP System.
50. C. Porter, *et al.*, "Improving Furnaces with Model Based Control," *Sol. State Technol.*, Nov. 1996, p. 119
51. Applied Materials Centura RTP System.
52. M.F. Pas and S.D. Pas, "Rapid Thermal Processing: When Will it Replace Batch Processing?" in *Rapid Thermal & Integrated Processing VI*. Eds. T.J. Riley *et al.*, Materials Res. Soc. (1997) p. 147.
- 54 G. Miner *et al.*, "Advanced Technologies for Rapid Thermal Oxidation," *5th Internatl. Conf. On Advanced Thermal Processing of Semiconductors, RTP' 97*, Sept. 1997, New Orleans, p. 31.
55. B. Peuse, M. Yam, S. Bahl, C. Elia, "Advances in Temperature Measurements and Control for RTP," *Proceedings of the RTP 1997* (New Orleans, LA) Sept. 1997.
56. Steag-AST Elektronik and AG Associates.
57. GaSonics International, San Jose, CA.
58. F. Reizman and W. VanGelder, *Solid State Electronics*, **10**, 625 (1967).
59. R.J. Archer, "Determination of the Properties of Films on Silicon by the Method of Ellipsometry", *J. of the Optical Soc. of America*, **52**, 970 (1962).
60. C.Y. Lu, *et al.*, *IEEE Electron Device Letters*, **10**, 192 (1989).
61. K.F. Schuegraf and C. Hu, *IEEE Trans Elect. Dev.*, **ED-41**, 761, (1994).
62. J. Maserjian, *Physics and Chemistry of Si/SiO₂ Interfaces*, Eds. C.R. Helms, B. E. Deal, Plenum Press p. 497 (1989)
63. H.S. Momose, *et al.*, "Direct Tunneling Gate Oxides Si MOSFET," *IEEE Trans. Electron Dev.*, **ED-43**, 1233 (1996).
64. T.P. Ma, "Making Silicon Nitride Film a Viable Gate Dielectric," *IEEE Trans. Elect. Dev*, **ED-45**, 680 (1998).
65. P. Munzinger *et al.*, "Pyrogenic Wet Oxidation in RTP," *Solid State Technology*, June 1998, p. 121.
66. R. DeJule, "Advances in Thin Film Measurement," *Semicond. International*, May 1998, p. 52.
67. P.J. Timans, "Temperature Measurement in RTP," *Solid State Technology*, April 1997, p. 63.
68. *ibid.* Ref. 37, p. 459-66.
69. M. Haller, *et al.*, "Multidomain Ellipsometry for Thin-Film Process Control," *Semiconductor International*, July, 1998, p. 269.
70. A. Dip, "Fast Thermal Processing," *Semiconductor International*, June 1996, p. 112.

PROBLEMS

1. Explain the difference between mass-transport limited oxidation, and reaction-rate limited oxidation.
2. Explain why steam oxidation proceeds at a faster rate than dry oxidation.
3. A silicon wafer of 150 mm diameter has a 0.2 μm thick layer of SiO_2 . Calculate the weight of the SiO_2 in this layer.
4. Show that in Eq. 8-17, $x_{\text{ox}}^2 + A x_{\text{ox}} = B (t + \tau)$, reduces to $x_{\text{ox}}^2 = Bt$ for log times, and $x_0 = B/A (t + \tau)$ for short times.
5. A silicon wafer has a 0.3 μm layer of thermally grown SiO_2 on its surface. Find the time required to increase the SiO_2 to a thickness of 0.5 μm by oxidation in wet oxygen at 1100°C. Repeat the calculation for dry oxygen at 1100°C.
6. If the wet oxidation of Problem 5 is conducted with 3% HCl added to the oxidizing ambient, determine the time that would be necessary for the film thickness to reach 0.5 μm .

7. Determine the SiO_2 thickness which would result from the following oxidation sequence: 30 min dry O_2 , followed by 30 min in wet O_2 , both at 1050°C . Assume (100)-Si is being used.
8. A cut is made in an SiO_2 film that is $0.4\text{ }\mu\text{m}$ in thickness. An oxide of 150 nm is grown over this cut. Sketch a cross section of the resulting Si/ SiO_2 structure, showing the thickness of the oxide in all locations, and Si/ SiO_2 interface at all locations.
9. List several techniques that have been found to allow an oxide to be initially grown on a silicon surface without forming oxidation-induced stacking faults.
10. Give some reasons why vertical furnaces have advantages over horizontal furnaces for growing thermal oxides, and also list some possible disadvantages.