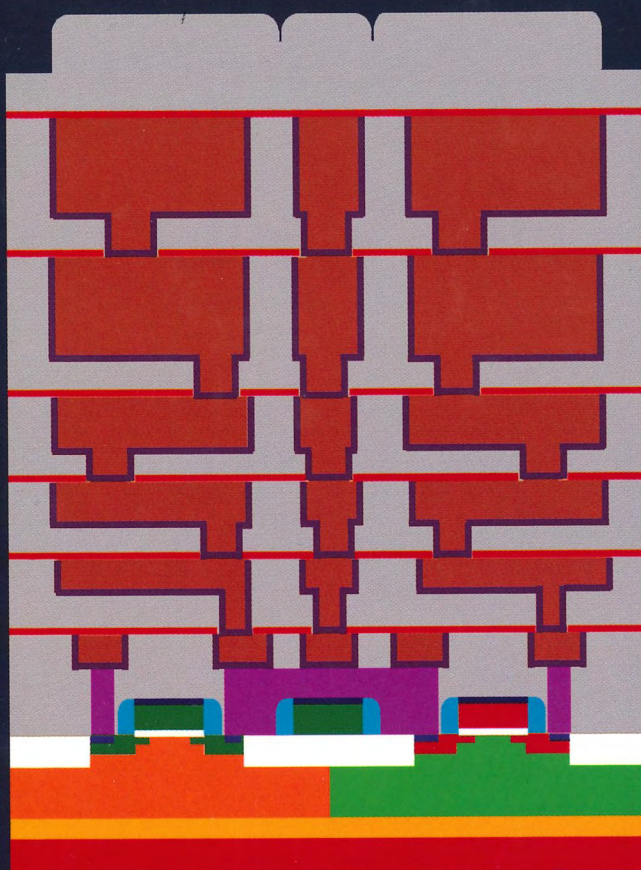


Silicon Processing

for the VLSI Era

Volume 4 - Deep-Submicron Process Technology



S. Wolf

**SILICON PROCESSING
FOR
THE VLSI ERA**

**VOLUME 4:
DEEP-SUBMICRON
PROCESS TECHNOLOGY**

STANLEY WOLF Ph.D.

**LATTICE PRESS
Sunset Beach, California**

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: S. C. Ross Wolf.

Copyright © 2002 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

Silicon Processing for the VLSI Era

Volume 4: Deep-Submicron Process Technology

Includes Index

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

ISBN 0-9616721-7-X

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

With the dawn of the new millennium, IC technology-nodes have evolved beyond 0.18-micron. New processes also had to be developed to meet the needs of this *Deep-Submicron Era*, including: copper/low-*k* dual-damascene interconnects; high-*k* MOSFET-gate and DRAM dielectrics; 300-mm wafers; chemical-mechanical polishing (CMP); excimer-laser-based DUV-lithography; chemically-amplified (CA) resists; silicon:germanium (Si:Ge); and silicon-on-insulator (SOI).

The semiconductor industry, however, also experienced an economic decline of unprecedented severity at the start of the 2000 decade. It seems clear that to fuel the recovery, chip manufacturers will have to exploit the above innovations. Such processes are a path to achieving the performance goals of future IC technology-nodes that will be needed for manufacturing new and profitable electronic products.

This is the *first available book dedicated exclusively to describing such deep-submicron process technologies* (including those listed above). It is intended to provide semiconductor engineers and researchers with a comprehensive, state-of-the-art reference about these emerging and leading-edge processes. While many other books have been written about the “traditional” IC fabrication processes, none offer what this volume contains. Its sole focus is those techniques needed to *manufacture chips for 0.18- μm (and smaller) generations*. Thus, it belongs on the shelf of every microelectronic professional whose livelihood depends on having a thorough grasp of the latest advances in IC fabrication techniques. There is no faster way to gain this information. As with our other volumes, owners will find themselves referring to this text again and again!

Readers can still consult Vol.1, 2nd Edition for background material on basic IC fabrication modules, as this book (Vol. 4) does not repeat that information. Instead, it presents over 800 pages of new content about the process technologies of the 2000's. It also has over 900 references (with 200 of these being published since 2000), and more than 500 illustrations. It is written in the same lucid, user-friendly style that has been widely praised in the other volumes of *Silicon Processing for the VLSI Era*.

As with the previous volumes, this text was produced using an accelerated book-production schedule. By employing revolutionary publishing techniques made available in the 1990's, Lattice Press can turn a finished manuscript into a published text in less than 3 months (a task that takes most publishers of technical books 15-18 months). This allows our books to be released faster, and to contain the most-up-to-date information at the time of publication.

SILICON PROCESSING FOR THE VLSI ERA is a multi-volume treatise designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The volume in hand is the fourth in the series. Volume 1 (*Process Technology*, 2nd Edition) was published in 2000 and deals with individual processes employed in the fabrication of silicon ULSI circuits (such as silicon wafering, thermal oxidation of silicon, CVD and PVD of amorphous and polycrystalline films, diffusion and ion implantation of dopants in silicon, microlithography, and patterning technology). Volume 2 (*Process Integration*) was released in 1990, and describes how the individual processes of Volume 1 are combined in various ways to produce MOS and bipolar VLSI and ULSI circuits. The task of integrating these various fabrication processes together is referred to as *process integration*. In Volume 3 (*The Submicron MOSFET*, published in 1995) the topics of submicron MOSFET device physics and the relationship between such device physics and submicron MOSFET fabrication are addressed.

A book of this length and diversity would not have been possible without the indirect and direct assistance of many others. From an indirect perspective, virtually all of the information presented here is based on the research efforts of countless numbers of scientists and engineers. Their contributions are recognized to a small degree in citations of their articles given in the references at the end of each chapter. The direct help was also provided by many people, and the resulting text is thus significantly better. The author expresses heartfelt thanks to those who gave their time, energy, and intellect.

Important direct help came from members of my immediate family. My wife, Carrol, read the manuscript for grammatical and stylistic correctness. My son, Ross, designed the cover and improved many of the graphics. My daughter, Jennifer, helped in marketing the book. Thus, it truly represents a family enterprise.

Stanley Wolf

For additional copies of the books contact:

LATTICE PRESS

Post Office Box 340

Sunset Beach, CA 90742 USA

www.latticepress.com

For your convenience an order form is provided on the final leaf of the book. It can also be ordered from the **LATTICE PRESS** website.

Chapter 3

GATE DIELECTRICS: THIN SILICON DIOXIDE FILMS

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

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

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

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

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

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

Application	SiO ₂ Thickness (nm)
Gate oxide insulators for MOS devices	1.5–10.0
Tunnel oxides in EEPROMs and flash memories	6.0–10.0
Screen oxides for ion implantation	10.0–20.0
Side-wall liners for shallow trench isolation (STI)	15.0–40.0
Field oxide for LOCOS isolation	200–400
Re-oxidation of the etched gate stack side-wall damage	5.0–7.5
Inter-polysilicon dielectric	4.0–20.0
Masking film against ion implantation and diffusion	100–200
Sacrificial oxides for gate applications	6.0–15.0
Capacitor dielectric in DRAM circuits	5.0–10.0

3.1 REQUIRED CHARACTERISTICS OF GATE DIELECTRICS FOR DEEP-SUBMICRON MOSFETS

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

1. The most important requirement that the gate dielectric of a deep-submicron MOSFET must meet is the following: *It must be possible to continue to decrease the thickness of the oxide to the degree specified in accordance with the scaling of the gate length L_E of a MOSFET.* We should elaborate upon why being able to scale the gate oxide thickness is so fundamentally important for deep-submicron MOSFETs. Figure 3-1 shows that the ratio of the gate length to the oxide thickness (L_E/t_{ox}) has indeed remained constant at about a value of 45 for process technology at Intel for over 25 years.¹ The gate-oxide requirements for technology nodes for 0.18- μm and beyond (according to the ITRS Roadmap) are given in Table 3-2. At some point, *if SiO₂ cannot be scaled further to meet future demands, another material with higher reliability will be needed.*

As CMOS IC technology advances, the central focus of scaling the MOSFET device structure involves scaling the gate length L_E . That is, gate lengths of MOSFETs are scaled to obtain the significant benefits offered by such scaling. These benefits include: 1) an increase in the drain current of the MOSFET I_D (which usually provides an improvement in the circuit

3.8.5 Impact of Process Induced Damage

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

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

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

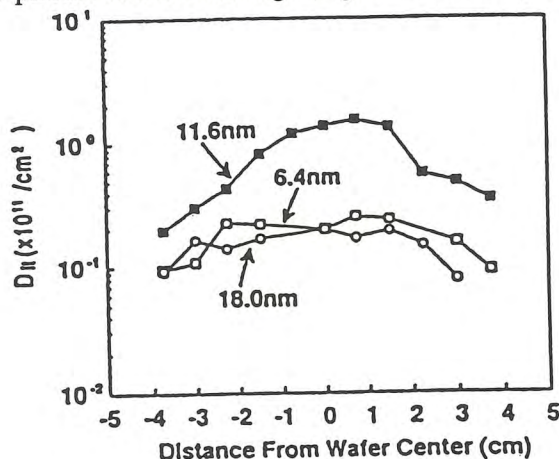


Fig. 3-47 Plasma-process-antenna-effect actually causes less damage to 6.4-nm oxides than to 11.6-nm oxides.¹¹

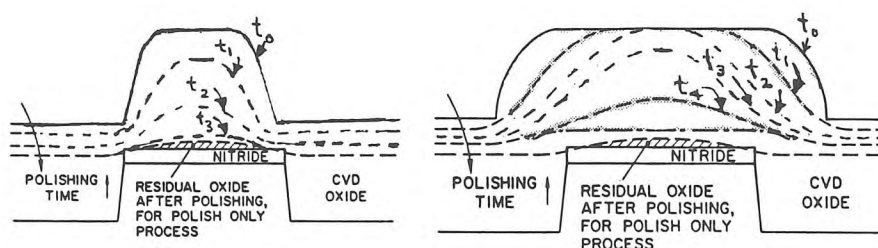


Fig. 8-28 Schematic cross-section of a narrow and wide feature of oxide being polished by CMP, as a function of time. Since the leading edges are polished most rapidly (and the feature continually rotates, so that all edges become the leading edge for the same amount of time), the narrow feature is planarized more rapidly than the wide feature.

The next question is: Why are densely-spaced narrow features planarized more slowly than isolated, narrow features? As first discussed by Burke,²⁴ the time required to planarize an “up-area” structure as shown in Fig. 8-26 (or to reduce its height to some fraction of its initial value) is dependent on the “down-area” adjacent to it. Specifically, the time to planarize a step depends on the width of the “up-area,” the width of the “down-areas” between “up-areas,” and the ability of the polishing-pad to rigidly span that distance between “up-areas.” The stiffer the pad, the more rigid the span - and the slower the rate of material removal from “down-areas.” This gives good planarization. The softer the pad, the less able it is to bridge larger distances. Thus, more material is removed by softer pads in “down-areas” between adjacent “up-areas.” This gives poorer planarization. When features to be polished are separated by distances greater than the leveling-distance of the pad, there is no selectivity between “up-areas” and “down-areas,” and planarization does not occur. In practice, some selectivity in RR occurs between “up-” and “down-areas,” but this selectivity is not complete. Ideally, the leveling-distance of a pad should be large enough to planarize the widest “down-areas” on a chip. On the other hand, the greater the leveling-distance, the more a pad will respond to wafer-scale non-flatness parameters, such as TLV and bow. This results in non-uniformity of the polished film thickness across the wafer (a subject discussed in more detail in Sect. 8.12.2).

In summary, planarization will occur most slowly for step heights in an oxide over the largest or densest interconnect structures adjacent to wide unpatterned

field areas, scribe lines, and wafer edges. It will occur most rapidly for step heights in the oxide over smaller arrays and narrow, isolated features. It is also necessary to specify an acceptable degree of planarization, because in the latter stages of the CMP process, material is removed nearly as fast from "down-areas" as from "up areas." Step heights no longer decrease at appreciable rates.

Like metal-CMP processes, oxide-CMP in a production environment is often a multi-step process. When in-line endpoint-detection is employed (as will be discussed in Sect. 8.6.9), two or more polishing steps may be needed to obtain the correct oxide thickness. In addition, a separate final-polish (or buff) procedure is frequently used to provide light polishing and to remove slurry particles and small wafer-defects. A small amount of oxide can also be buffed away with this buffing procedure (typically 30–40-nm), which aids in removing defects (especially residual slurry particles and the small scratches created by the primary polish). Use of a buff procedure is one of the reasons why CMP can actually *reduce* defect-densities (including eliminating some types of defects created in other processes), rather than increase them.²⁵

8.5.4 Metal-CMP Mechanisms

The mechanisms involved in the CMP of metals which cause planarization of the metal film surface, follow a relatively simple three-step sequence proposed by Kaufman.²⁶ This three-step sequence is as follows (Fig. 8-29):

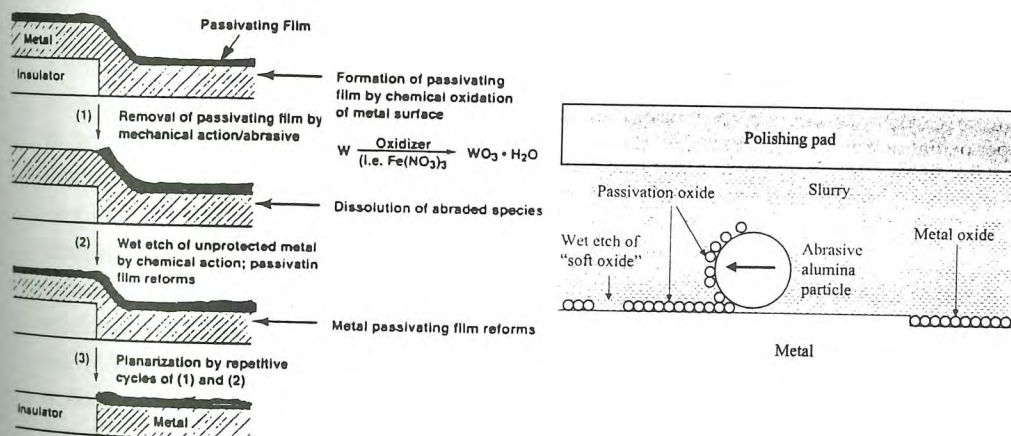


Fig. 8-29 Schematic representation of the model of CMP of metal films.²⁶

Chapter 14

LOW- k DIELECTRICS

For the purposes of IC microelectronic applications, dielectrics are categorized into two classes: 1) *low- k dielectrics* and; 2) *high- k dielectrics*. Those having a dielectric-constant k lower than that of thermal silicon dioxide (i.e., $k < 3.9$) are classified as *low- k dielectrics*, while those having a value higher than that of silicon nitride ($k > 7$), are classified as *high- k dielectrics*. Note that the minimum value of k is 1.0 for air. In this chapter low- k dielectrics are discussed, and in Chap. 4 high- k dielectrics are covered.

14.1 INTRODUCTION TO LOW- k DIELECTRICS

Low- k dielectric films are important in advanced IC applications because their use can significantly improve circuit performance characteristics, primarily with respect to interconnect delay. As detailed in Chap. 12, interconnect delay can be reduced not only by decreasing R of the conductor structures in ICs, but also by decreasing C of the dielectric layers (Fig. 14-1). The value of C , in turn, can be reduced by using dielectric materials with smaller permittivity values (i.e., *low- k dielectric materials*).¹ As semiconductor device geometries shrink below 0.25- μm , the parasitic capacitance in the intermetal-dielectric between metal lines becomes more and more important, in terms of the RC -delay in switching.

The primary dielectric film material for interconnects has been SiO_2 . The choice of SiO_2 was based on its good dielectric and mechanical strength, as well as its ease of processing. However, SiO_2 has a dielectric-constant ranging from $k = 3.9$ to 4.5, depending on the method of its formation. This value is too high to be applicable in devices below about 0.18- μm . That is, as indicated in Fig. 14-1 only by using Cu as a conductor together with a low- k dielectric as the insulator, can the sufficiently-low values of RC -delay required for deep-submicron ICs be obtained. Thus, there has been a push to integrate new dielectric materials with k -values smaller than that of SiO_2 into IC fabrication. It should be pointed out that it has become apparent that the tacitly-hoped-for dielectric with $k \sim 1$ is not realizable, and that some of the qualities of SiO_2 will also no longer be attainable in actual low- k dielectrics. Evaluations of various dielectric materials for low- k thin-film applications will provide the empirical data needed to determine the trade-off limits of k -value versus mechanical performance.

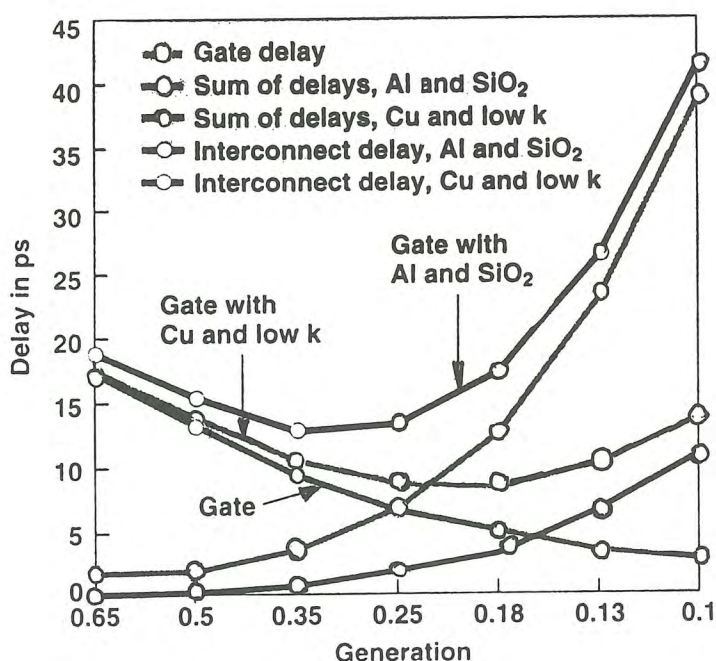


Fig. 14-1 RC-delays caused by interconnects are charted as a function of feature size for both Al/SiO₂ and Cu/low-*k* approaches. (© 1997 IEEE)

The conclusions from Fig. 14-1 can be justified by noting that the SiO₂ molecule is a pyramid-shaped tetrahedron (see Chap. 3), and its lattice structure thus fits together rather tightly. Although thermal SiO₂ does not possess long-range-order, it is nevertheless a relatively dense material (~2.3 g/cm³). Since the goal is to find dielectric films with *k*-values smaller than that of oxide (~4.0), the factors that impact the dielectric constant should first be considered.

In general, no dielectric films as dense as thermal oxide can achieve *k*-values significantly smaller than 3.5. In fact, even CVD-oxides (which are also silica-based films) have about the same *k*-values, even though they are somewhat less dense. While CVD-oxides are more disordered than thermal-SiO₂ (which lowers their density), such disorder also makes them contain more silanol (OH) groups (i.e., at sites where Si-O-Si bonds are unable to form). Thus, they are prone to absorb more moisture than thermal oxides, and the absorbed water raises their *k*-value. Furthermore, materials that are less dense than thermal-oxide are also generally mechanically weaker. Nevertheless, to achieve a *k* < 2 in any dielectrics (including silica films), free space must be added, even though the strength of the material diminishes according to some power of the solids content.

The approaches used to produce dielectrics with k -values smaller than that of thermal-oxide and CVD-oxides therefore fall into three categories: 1) reduction of the film density; 2) reduction of ionization in the material, and 3) reduction of the film polarization. Details of how these approaches are exploited in each low- k material-type are provided later, but a few general comments on them are given here.

First, the most dramatic technique for reducing the k -value is to *decrease the density* of the film material. Figure 14-2 shows the density of various low- k materials, as well as their k -values. It shows that the lower the material density, the lower the value of k . In all ultra-low- k dielectric-materials, the density is significantly lower than that of oxide films. This density reduction can occur either at the *microscopic level* (i.e., through the formation of a lattice structure that is more open than that of silicon oxide), or at the *macroscopic level* (i.e., through the formation of macropores in the film material - with sizes in the nm-range).

Second, *reduced ionization* and *reduced polarization* are only able to reduce the k -value to a much smaller extent. Reduced ionization is a common feature of the partially-organic low- k -materials. Such organic films contain carbon, often in the form of methyl groups (CH_3). These methyl groups replace oxygen atoms to create Si-CH_3 bonds in place of Si-O bonds. A smaller percentage of Si-CH_3 bonds ionize than do Si-O bonds (and they are less polarizable than are Si-O bonds). Reduced polarization is exploited in such dielectrics as fluorinated silicate glass (FSG). That is, the fluorine atoms introduced into the oxide lattice

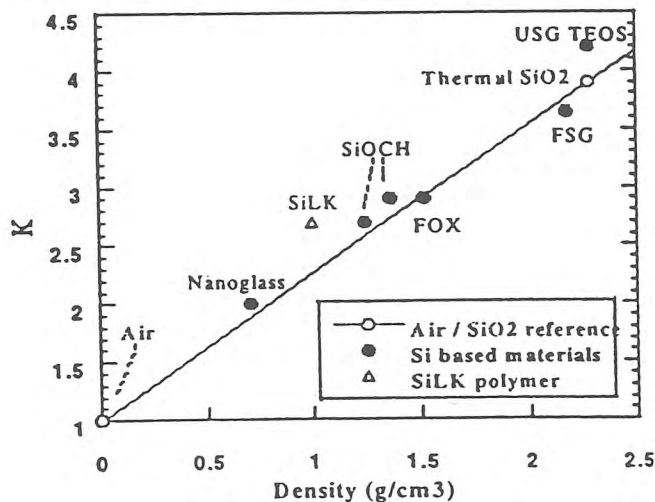


Fig. 14-2 Value of k is shown vs. density for various dielectrics in ULSI interconnects.

TABLE 14-1 ITRS Roadmap of the Effective Dielectric Constant Needed at Each Device Generation - Out to 2012

Table 1. SIA Roadmap								
1997								
Year of first product	1997	1999	2001	2003	2006	2009	2012	
Shipment technology node	250 nm	180 nm	150 nm	130 nm	100 nm	70 nm	50 nm	
Interlevel metal insulator effective k	3.0-4.1	2.5-3.0	2.0-2.5	1.5-2.0	1.5-2.0	<1.5	<1.5	
1999								
Year of first product		1999	2001	2002	2005	2008	2011	
Shipment technology node		180 nm	150 nm	130 nm	100 nm	70 nm	50 nm	
Interlevel metal insulator effective k		3.5-4.0	2.7-3.5	2.7-3.5	1.6-2.2	<1.5	<1.5	
2000?								
Year of first product			2001	2003	2005	2007	2009	2011
Shipment technology node			180 nm	150 nm	130 nm	100 nm	70 nm	50 nm
Interlevel metal insulator effective k			3.5-4.0	2.7-3.5	2.7-3.5	2.2-2.7	2.2-2.7	1.6-2.2

replace O atoms and form Si-F bonds instead of Si-O bonds. The shorter bond length of Si-F bonds makes the molecular structures containing a fraction of Si-F bonds less polarizable than those consisting of only Si-O bonds.

A few materials that were identified as good low- k on-chip insulators for ICs were first integrated into two-level-metal structures in the 1995-96 timeframe. These first-generation low- k films had k -values somewhat smaller than 4.0 ($2.8 < k < 3.5$). In 1999 the *International Technology Roadmap for Semiconductors* (ITRS-99) was published.² It urged IC manufacturers to identify, characterize, and integrate interlevel dielectrics with values of $k = 2.5-3.0$ for 0.18- μm devices, and $k = 2.0-2.5$ for 0.15- μm devices (Table 14-1). However, the industry proved to be more conservative. As Table 14-1 shows, there was a two-year delay in the implementation of the various low- k solutions in the years from 1997 to 1999. The delay in the adoption of low- k films was due to numerous technical challenges that had to be overcome before low- k materials could be successfully integrated into existing processes. As a result, the 0.25- μm and 0.18- μm technology generations were generally implemented with dielectrics having ranges of $k = 3.0-4.0$, and $k = 2.5-3.0$, respectively. As a result, ICs with low- k dielectrics did not appear in significant volume until about 1999. However, the consumption of materials used to produce low- k films is expected to gradually increase, and by 2003 the global low- k material market should reach \$150 million and expand rapidly. In addition, it is quite possible that the expected

change from a k range of 2.5-3.0 to a k range of < 2.0 will be stretched out beyond the 70-nm generation in 2005, due to a significant increase in integration challenges for materials with dielectric constants below 2.0.

As mentioned, there are many process-integration issues that must be satisfied before a new material is successfully incorporated into an IC fabrication sequence.³ Candidate low- k materials must also satisfy these requirements to be useful in IC fabrication. Here a number of low- k materials being considered for use in deep-submicron IC processes are described, as well as some of the process integration challenges they offer.

14.1.1 Spin-On versus CVD Methods for Forming Low- k Dielectrics

The two most common ways of depositing low- k dielectric films are: 1) spin coating followed by furnace annealing (*spin-on deposition*); and 2) chemical vapor deposition (*CVD deposition*). As will be seen, there are many varieties of low- k films within the spin-on and CVD-deposition groups being studied for use in advanced IC interconnects. Note that spin-on films are all applied in liquid solution form, and are then cured in nitrogen at temperatures as high as 450°C. Curing is necessary for: **a)** driving-off solvent; **b)** completing chemical reactions; **c)** driving-off unreacted monomers and oligamers; and **d)** densification.

Proponents of the CVD deposition approach point to the reuse of existing toolsets and simpler integration due to the silicon-dioxide-like structure of CVD siloxanes as advantages of their method. However, the cost of depositing films by CVD is generally higher than the cost of a spin-on deposition. Hence, this is a factor that must also be considered when choosing a low- k material. Manufacturers of CVD-tools for depositing low- k films include Applied Materials, Novellus, ASMI, and Trikon Technologies.

Manufacturers of spin-on materials and spin-on dielectric (SOD) tracks contend that the spin-on dielectrics allow better extendibility to future generations (especially in the sub-0.25- μm regime, when porous, low- k materials are likely to be used). Manufacturers of *spin-on dielectric (SOD)* coaters include: Dainippon Screen (DNS, Tokyo), Tokyo Electron (TEL America, Austin, TX), FSI International (FSI, Chaska, MN), Semix (Fremont, CA), and SVG (San Jose, CA). Spin-on dielectric material vendors include: Honeywell, Dow Chemical, Dow Corning, Hitachi Chemical, and Tokyo Ohka Kogyo. Each SOD equipment vendor has optimized the coating and curing modules specifically for SODs, making them significantly different from the photoresist tracks on which they were originally based. Note that one cost factor in SOD use is the quantitative dispense of the SOD. Due to the high cost ($> \$2000/\text{liter}$) and potential sensitivity

underlying, plug-containing layer. This overlying dielectric film therefore also remains planarized. A second lithography step is then used to form the pattern of a trench, which is aligned to the previously fabricated via plug. With this aligned trench pattern acting as a template, a trench or groove is etched into the planarized overlying ILD. After the resist pattern is stripped, the trench is filled with a metal (such as aluminum or copper).² This metal is polished back by CMP to create a metal line embedded in a dielectric trench, and in contact with the underlying plugs. Finally, another dielectric layer is deposited onto the dielectric layer containing the metal-filled trenches. This dielectric layer will be the one through which the next level of vias (or holes to the contact pads) will be created.

It is worth emphasizing several important aspects of the single-damascene process: 1) If CMP is used to form both the plug and the metal in the trench, no metal etch process is needed; 2) No gaps have to be filled by the dielectric-film-deposition processes; 3) The process still contains about the same number of steps as the subtractive interconnect process (i.e., the two polish steps replace the via and metal etch steps of the subtractive process, and the two metal-deposition steps replace the metal-plug-deposition step and main-metal-line deposition step of the plug/metal-line subtractive interconnect process); and 4) Since the first-level interlevel dielectric layer is planarized before the contact holes are etched, *contact holes* will generally have different depths (e.g., those going to the substrate will be deeper than those going to the gates). Thus, to allow all contact holes to be opened, the shallower ones will expose the underlying material to the plasma etching reactants during the time the deeper contacts are being opened. To prevent the underlying material in the shallow contact holes from being attacked during this time, an etch-stop layer is often deposited after the salicide process is completed (typically using Si_3N_4 as such an etch-stop film). After the contact etch is completed, the nitride etch-stop layer is removed. Since the thickness of this layer is uniform on all regions, it can be etched away with a much shorter overetch time. This method allows all of the contact holes in a single-damascene process to be opened without damaging the underlying material on any part of the die.

15.4 INTRODUCTION TO DUAL-DAMASCENE INTERCONNECT PROCESSES

In a *dual-damascene (DD) structure*, only a single metal deposition step is used to simultaneously form the main metal lines and the metal in the vias. That is, both trenches and vias are formed in a single dielectric layer. The vias and trenches are defined by using two lithography steps (Fig. 15-3). Trenches are typically etched to a depth of 4000-5000-Å, and the vias are typically 5000-

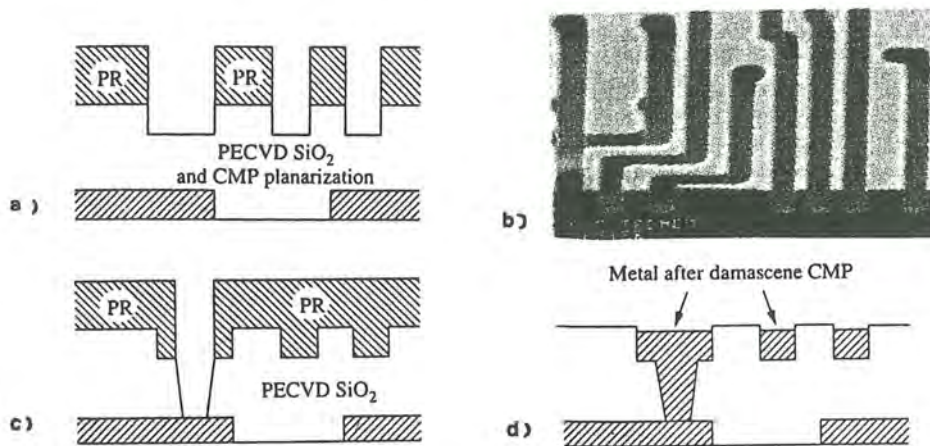


Fig. 15-3 Dual-Damascene process: (a) IMD is deposited by CVD and planarized by CMP. Trench is defined by PR #1 and then etched; (b) SEM of the etched trench after PR #1 has been stripped; (c) Vias are defined by PR #2 and then etched, using PR #2 to protect other regions of the IMD from etching; (d) Metal is deposited to simultaneously fill trenches and vias. CMP is used to remove excess metal. The sequence is repeated for the next level of metal.

7000Å-deep. After the via and trench recesses are etched, the via is filled in the same metal-deposition step that fills the trench. After filling, the excess metal that is deposited outside the trench is removed by a CMP process, and a planar structure with metal inlays is achieved. As in the single-damascene process, once a planarized surface is achieved, it is no longer necessary to perform CMP on the dielectric layers. Thus, one of the CMP steps needed for each metal level in the subtractive interconnect process is also eliminated.

It should also be mentioned that in early dual-damascene processes, the trench dimension was somewhat wider than that of the vias. This made the alignment of the via to the trench somewhat easier. However, with time, the via and trench widths have been made equal. (At the 0.18- μm node, the via and trench widths are 0.22-0.24- μm wide, while at the 0.13- μm node they will probably be ~0.18- μm wide.) Thus, modern dual-damascene processes are more sensitive to misalignment between the trench and via. However, this overlay sensitivity problem is mitigated to some degree by the nature of the damascene process, making it possible to maintain high interconnect packing density. That is, the alignment of the trenches to the vias in the dual-damascene sequences is an easier task than in the subtractive interconnect process. In the latter, aligning metal lines

to the underlying contact holes (or vias) must be done after the opaque, shiny, blanket-metal film has been deposited. In damascene approaches, aligning the trenches to the vias is done through the transparent dielectric film, which allows tighter design-rule tolerances to be used when performing this alignment.

When comparing the single-damascene process to the simple dual-damascene process described above, one observes one metal-deposition step and one CMP step are eliminated in the latter (as well as one dielectric-deposition step). The reduction in the number of processing steps is another of the benefits that have driven the development of dual-damascene processes. Note that in single-damascene structures, plugs are typically filled with W, but in the dual-damascene-interconnects the metal that fills the holes is Al or Cu. The technology of filling of vias and trenches with Cu is described in Chap. 16. A dual-damascene process using Al was described by workers at IBM in 1998.³

15.5 THE THREE DUAL-DAMASCENE PROCESS SEQUENCES

Three different fabrication sequences have been developed to produce dual-damascene structures:

1. Trench-First Dual-Damascene
2. Via-First Dual-Damascene
3. Self-Aligned (or Buried-Via) Dual-Damascene

It should be noted that the *via-first* and the *trench-first* sequences are the processes currently being used in mainstream production. The trench-first sequence was probably the first one to be adopted, but the via-first sequence is currently the approach that has gained the greatest acceptance (and will likely remain so going forward). As a result, we will focus on these two sequences, and will point out the advantages and drawbacks of each one. While the third approach (the self-aligned process), has not been implemented in production (for reasons we will also explain), we will provide a brief description of it as well. However, whichever sequence is used, the final damascene structure is the same (see Fig. 15-4).

Before outlining the three dual-damascene processes, it should be mentioned that the dielectric layer in which the trenches and vias are etched actually consists of at least three layers of dielectric (see Fig. 15-8b). That is, SiO₂ (or a low-*k* dielectric is deposited to a depth of 5000-7000Å (and this will be the layer in which the vias will be etched). Next, a thin silicon nitride layer - or other material (which will serve as an *embedded etch-stop layer*), is deposited (about 30-nm thick). Finally, another SiO₂ layer (or a low-*k* dielectric layer) is deposited

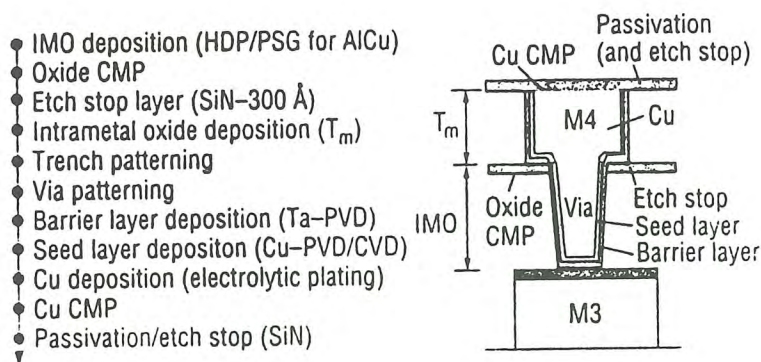


Fig. 15-4 A suggested dual-damascene-based Cu metallization flow and resulting cross-sectional structure.⁸ Reprinted with permission of Solid State Technology. Published by PennWell.

(~4000-5000-Å thick). This will serve as the layer in which the *trenches* will be etched.

In the *trench-first* sequence (Fig. 15-5), the trench patterns are defined in the ILD first. That is, after spinning on the resist, the trench-pattern-mask is used to expose the resist. The trench is then produced by etching the dielectric down to the embedded etch-stop layer. After the trench-etching process, the first resist layer is stripped. A second resist layer is then spun on, and the via-pattern-mask is used to create openings in this resist layer - aligned to the trench that was etched previously. The resist protects the other parts of the wafer surface (including the etched regions of the trenches) so that the vias can be etched without further etching the dielectric in the trenches.

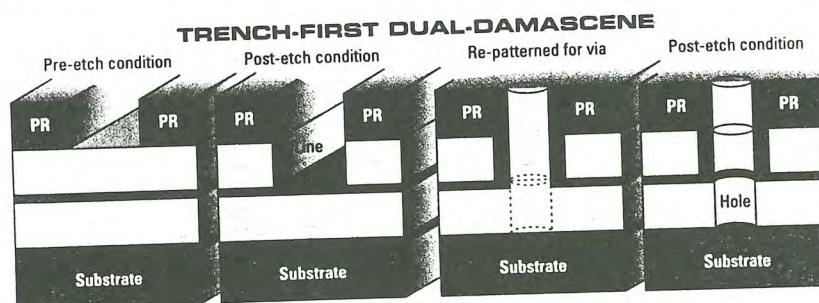


Fig. 15-5 In the trench-first approach, the vias are patterned & etched after trenches are etched.

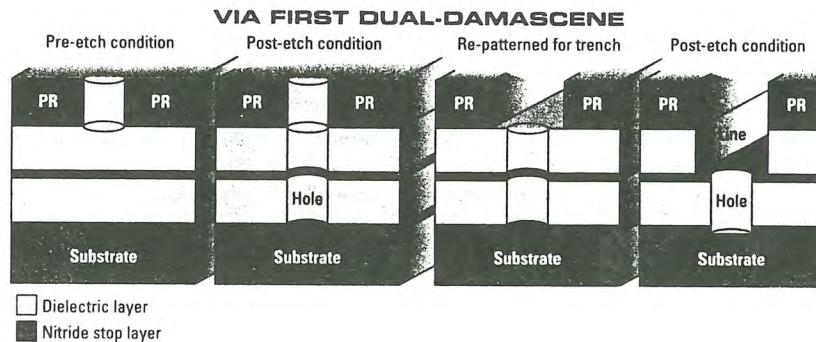


Fig. 15-6 In the via-first approach, the trench is patterned and etched after the via is formed.

The chief disadvantage of the trench-first sequence involves the fact that the vias must be patterned after the trench etch. That is, since resist is applied as a liquid onto the wafer surface, it fills recessed regions as water fills a lake. Hence, the top surface of the resist is planar. This means that the regions of resist over the damascene trenches is quite thick. Resolving fine features in thick resist is harder than in thinner resists. For features smaller than $0.25\text{-}\mu\text{m}$ the process latitude becomes too small for this to be a practical manufacturing process. Thus, while the trench-first sequence could be used for early dual-damascene processes ($0.35\text{-}\mu\text{m}$ and some $0.25\text{-}\mu\text{m}$ processes), it is difficult to extend it to smaller technology nodes (i.e., in order to be able to use it at such small dimensions, it may be necessary to implement a thin imaging-resist-layer process, as discussed later).

In the *via-first approach* the vias are defined first in the ILD, followed by patterning the trenches. As noted above, this is currently the most common method of fabricating dual-damascene interconnects (and is schematically shown in Fig. 15-6). The sequence of forming the damascene recesses in this approach begins by exposing the *via patterns* with the first mask. After etching the vias completely through the entire dielectric stack (except not through the barrier layer at the bottom of the dielectric stack) and stripping the resist, a second mask is used to pattern the trenches. The trenches are then created by etching the dielectric down to the embedded etch-stop layer. The barrier layer at the bottom of the vias is protected from further etching during the trench-etch either by resist or a BARC layer (applied for the trench-etch step) that floods the vias, or by using an etch process that is highly selective to the bottom etch-stop material (i.e., so that the etch-stop layer at the bottom of the via is not removed during the trench-etch step). After the resist is stripped and the etch-stop layer at the bottom of the via is removed by dry-etching, the metal that fills both the vias and the

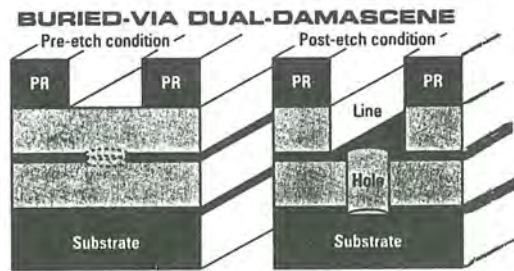


Fig. 15-7 In the self-aligned, or buried-via approach, a hard mask is patterned and etched. The trench and via can be etched in one step.

trenches can be deposited. After deposition, it is polished back to create the dual-damascene structure.

In the *self-aligned* dual-damascene process (Fig. 15-7), the via pattern is created in the embedded etch-stop layer with a lithography-and-etch sequence that is performed before the top dielectric layer of the stack is deposited. A special etch process is required for this step. After the via pattern is thus etched, the top dielectric layer is deposited. Finally, the trench mask is aligned to the via openings in the embedded nitride layer, and both the trench and via are opened with a single etch step. This sequence requires nearly perfect trench-to-via alignment, otherwise the via may no longer be round (but instead only half-moon shaped), resulting in high via resistance. In addition, an etch-process with high nitride-to-oxide selectivity is required. For these reasons this scheme has not found use in production.

15.6 THE EVOLUTION OF DUAL-DAMASCENE FROM SiO_2 TO LOW- k DIELECTRIC STACKS

In the first generation of dual-damascene processes (i.e., at the $0.25\text{-}\mu\text{m}$ node), Cu was used together with a conventional undoped PECVD- SiO_2 dielectric (with a PECVD-SiN layer beneath the SiO_2 , which acts as both a Cu diffusion barrier and an etch-stop when the vias are etched, Fig. 15-8a).^{*} The nitride and oxide layers may be deposited sequentially in the same tool (e.g., a PECVD reactor).

^{*} This SiN layer is deposited first, and as noted, serves as a barrier against the diffusion of Cu between metal levels and also as an etch-stop in the dielectric etch that defines the vias. Because SiN has a high value of k , this layer should also be as thin as possible to avoid adding significantly to the interlevel capacitance. At the same time, the SiN layer must be low-stress, dense, and pinhole-free, so that its diffusion-barrier properties are not compromised. High-density is also critical to the nitride layer's role as an etch-stop.