SILICON PROCESSING

FOR

THE VLSI ERA

VOLUME 2:

PROCESS INTEGRATION

STANLEY WOLF Ph.D.
Professor, Department of Electrical Engineering
California State University, Long Beach
Long Beach, California

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 Roy Montibon, Visionary Art Resources, Inc., Santa Ana, CA.

Library of Congress Cataloging in Publication Data
Wolf, Stanley

Silicon Processing for the VLSI Era
Volume 1 : Process Integration

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

86-081923

ISBN 0-961672-4-5

9 8 7 6 5

PRINTED IN THE UNITED STATES OF AMERICA

The primary disadvantages of the CDI process are the large collector-base and collector-substrate capacitances (as well as relatively low collector-base junction breakdown voltages, which limits them to applications which use small power-supply voltages). Consequently, advanced oxide-isolated structures that minimize these capacitances have largely replaced CDI. However, the benefits of oxide isolation are gained only at the cost of significantly increased process complexity. Work is therefore continuing to increase the performance of CDI so that its compact structure and process simplicity can be exploited for additional applications.[2]

2.2 BASIC ISOLATION PROCESS FOR MOS ICs (*LOCOS ISOLATION*)

The isolation requirements of MOS ICs are somewhat different than those of bipolar ICs. Because MOS transistors are self-isolated, as long as the source-substrate and drain-substrate *pn* junctions are held at reverse bias, the drain current (I_D) should only be due to current flow from source to drain through a channel under the gate. This also implies that no significant current between adjacent MOS devices should exist if no channel exists between them (Fig. 2-4). The buried n^+-p junction needed to isolate bipolar transistors is thus not necessary for MOS circuits.

The self-isolation property of MOS devices represents a substantial area savings for NMOS (and PMOS) circuits as compared to junction-isolated bipolar circuits. Hence, ICs having the highest component densities are fabricated with MOS technologies.

The method by which the components of an integrated circuit are interconnected involves the fabrication of metal stripes that run across the oxide in the regions between the transistors (the field regions, Fig. 2-4). However, these metal stripes form the gates of parasitic MOS transistors, with the oxide beneath them forming a gate oxide and the diffused regions (2) and (3) acting as the source and drain, respectively. The threshold voltage of such parasitic transistors must be kept higher than any possible operating voltage so that spurious channels will not be inadvertently formed between devices (see chap. 5 for a discussion of inversion in MOS transistors).

In order to isolate MOS transistors, then, it is necessary to prevent the formation of channels in the field regions, implying that a large value of V_T is needed in the field regions. In practice, the V_T of the field regions needs to be 3-4 V above the supply voltage to ensure that less than 1 pA of current flows between isolated MOS devices[108] (e.g., for 5 V circuit operation, the minimum field-region V_T must be 8-9 V). This condition must be maintained at the minimum isolation spacing used in a given technology, since decreases in field-region V_T arise for same reasons as in short-channel MOS devices (see section 5.5.1.1). Furthermore, V_T also decreases as the temperature rises; reductions of as much as 2 V in field-V_T values have been observed as the temperature rises from 25°C to 125°C.[109]

As will be discussed in chapter 5, several methods can be used to raise the threshold voltage. The two utilized to increase field-region V_T's involve increasing the field-

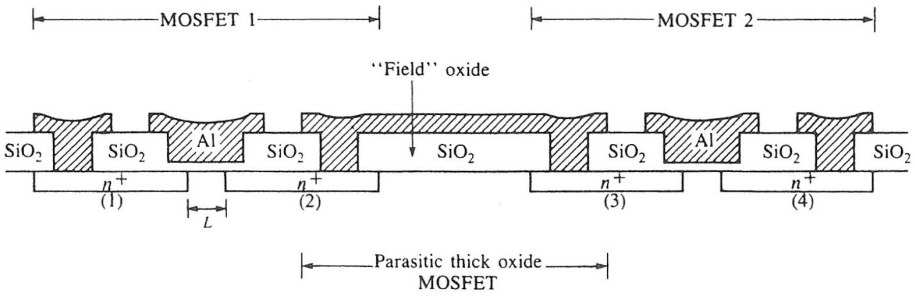

Fig. 2-4 No current should flow between source and drain regions of adjacent MOS transistors. This figure shows the parasitic field transistor with a possible channel under the field oxide if the substrate under this oxide is inverted.[128] From R. C. Jaeger, *Introduction to Microelectronic Fabrication.* Copyright, 1988, Addison-Wesley. Reprinted with permission.

oxide thickness and raising the doping beneath the field oxide. If the field oxide were made sufficiently thick, it alone could cause a high enough threshold voltage in the field parasitic device. Unfortunately, the large oxide steps produced by this approach would give rise to step coverage problems, and reduced field-oxide thicknesses are thus preferred. To achieve a sufficiently large field threshold voltage with such thinner oxides, the doping under the field oxide must be increased (see chap. 5, Example 5-5).

The field oxide is nevertheless typically made seven to ten times thicker than the gate oxide in the active regions. (This thick oxide also reduces the parasitic capacitance between the interconnect runners and the substrate, improving the speed characteristics of the circuits.) Normally, ion implantation is used to increase the doping under the field oxide. This step is called a *channel-stop implant.* The combination of thick oxide and channel-stop implant can provide adequate isolation for PMOS and NMOS (and as we shall see, for oxide-isolated bipolar) ICs. Additional isolation considerations that arise in the case of CMOS circuits will be described in chapter 6.

The thick field oxide can be fabricated in a number of ways. In one approach (used until about 1970), the oxide is grown to the desired thickness on a flat silicon surface and then etched in the active regions (leaving thick oxide in the field regions, Fig. 2-5b). Although this *grow-oxide-and-etch* approach also allows the isolation regions to be most sharply defined, two disadvantages have prevented it from being utilized in VLSI applications: (1) the field-oxide steps are high and have sharp upper corners, making them very difficult to cover with subsequent metal-interconnect lines; and (2) the channel-stop implant must be performed before the oxide is grown, and thus the active regions must subsequently be aligned to the channel-stop regions in the silicon with a lithography step. This imposes a severe packing density penalty.

In another approach, the oxide is selectively grown over the desired field regions. This is done by covering the active regions with a thin layer of silicon nitride that

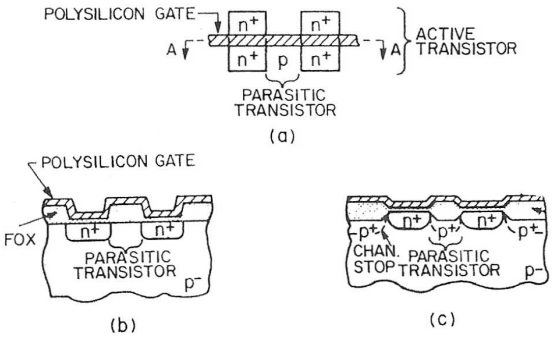

Fig. 2-5 MOSFET isolation. (a) Top view of adjacent MOS transistors with common polysilicon gate illustrating active and parasitic transistor conduction paths. Cross section through A - A for (b) Grow-oxide-and-etch isolation, and (c) LOCOS isolation structures.[130] From S. M. Sze, Ed., *VLSI Technology*, 2nd Ed., Copyright 1988 McGraw-Hill. Reprinted with permission.

prevents oxidation from occurring beneath them. After the nitride layer has been etched away in the field regions (and prior to field-oxide growth), the silicon in those regions can also be selectively implanted with the channel-stop dopant. Thus, the channel-stop region becomes self-aligned to the field oxide, overcoming one of the drawbacks of the grow-oxide-and-etch method.

The more serious step-coverage limitation of the grow-oxide-and-etch isolation method is also overcome to some degree by the selective-oxidation approach. If the silicon is etched (to a depth of about half the desired field oxide thickness) after the oxide-preventing layer is patterned, the field oxide can be grown until it forms a planar surface with the silicon substrate. (Note: this occurs because the growing oxide film is about twice as thick as the silicon layer it consumes; see Vol. 1, chap. 6). This is known as a *fully recessed isolation oxide process.*

If the field oxide is selectively grown without etching the silicon (Fig. 2-5c) the resulting field oxide will be *partially (or semi) recessed.* In the semirecessed process, the oxide step height is larger than in the fully recessed process, but it is smaller than in the grow-oxide-and-etch process. In addition, the step has a gentle slope that is more easily covered by subsequent polysilicon and metal layers. Semirecessed LOCOS is also less complex than fully recessed LOCOS and results in fewer process-induced defects in the silicon substrate. The semirecessed oxide has become the workhorse isolation technology for MOS devices down to about 1.5-μm geometries, while the fully recessed oxide has primarily been used in bipolar circuits.

The selective oxidation process described above was introduced by Appels and Kooi in 1970,[3] and it has become the most widely used approach for forming the thick field oxide. We will first describe this process in detail, discussing its limitations for

submicron devices. Subsequently, alternative isolation schemes that have been proposed to overcome these limitations will be considered.

2.2.1 Punchthrough Prevention between Adjacent Devices in MOS Circuits

Parasitic conduction between adjacent devices due to *punchthrough* (Fig. 2-4c) must also be prevented (see chap. 5 for additional information on punchthrough). In MOS circuits, the source and drain regions of each transistor must be kept far enough from the source and drain regions of any neighboring devices so that the depletion regions do not merge together (i.e., this distance must be greater than twice the maximum depletion-region width). Substrate doping must also be considered, because lighter doping allows wider depletion-region widths, and punchthrough is more likely to occur at lower voltages in lightly doped substrates.

2.2.2 Details of the Semirecessed-Oxide LOCOS Process

Each of the steps used to fabricate conventional semirecessed LOCOS structures will now be discussed in additional detail (Fig. 2-6).

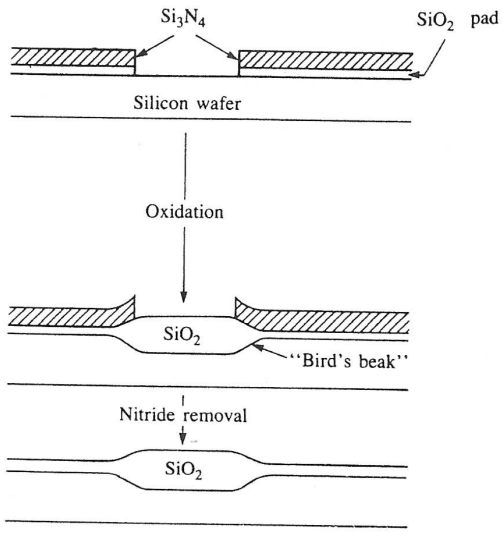

Fig. 2-6 Cross section depicting process sequence for semirecessed oxidations of silicon.[128] From R. C. Jaeger, *Introduction to Microelectronic Fabrication*. Copyright, 1988, Addison-Wesley. Reprinted with permission.

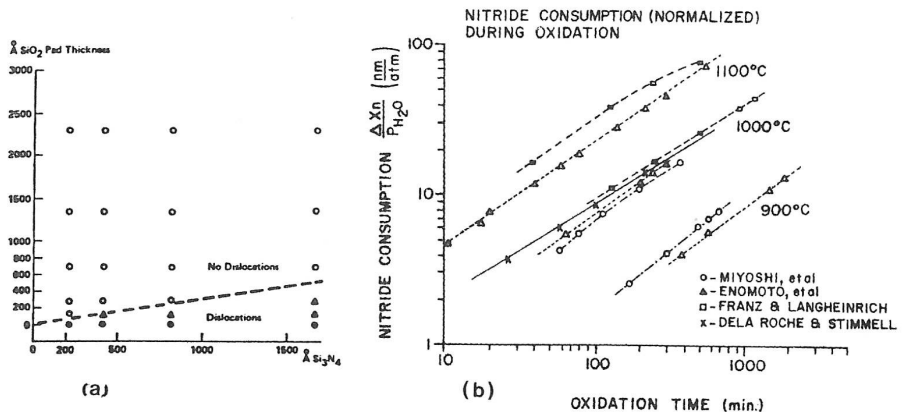

Fig. 2-7 (a) Dislocation generation at Si_3N_4 film edges versus CVD pad-oxide thickness, after 6 h of dry-wet-dry thermal oxidation at 950°C.[7] Copyright 1978, reprinted with permission of the AIOP. (b) Nitride consumption during oxidation.[131] Reprinted with permission of Semiconductor International.

2.2.2.1 Pad-Oxide Layer. A wafer with a bare silicon surface is cleaned, and a 20-60 nm SiO_2 layer is thermally grown on the surface. The function of this layer, called a *pad* or *buffer* oxide, is to cushion the transition of stresses between the silicon substrate and the subsequently deposited nitride. In general, the thicker the pad oxide, the less edge force is transmitted to the silicon from the nitride (Fig. 2-7a).[110] On the other hand, a thick pad-oxide layer will render the nitride layer ineffective as an oxidation mask by allowing lateral oxidation to take place. Therefore, the minimum pad-oxide thickness that will avoid the formation of dislocations should be used. Figure 2-7b indicates that the minimum thickness of a thermally-grown pad oxide should be at least one third the thickness of the nitride layer.

Alternative techniques to allow the use of thinner pad oxides have also been reported. One method involves the use of CVD SiO_2 in place of thermal SiO_2.[4] Because CVD SiO_2 is more effective than thermal SiO_2 for avoiding edge defects, such pad layers can be about 25% as thick as thermal-oxide pad layers. The second method utilizes a pad layer consisting of a thin thermal SiO_2 layer and a buffer polysilicon layer.[5,6,34,35]

2.2.2.2 CVD of Silicon Nitride Layer. Next, a 100-200 nm thick layer of CVD silicon nitride which functions as an oxidation mask is deposited. Silicon nitride is effective in this role because oxygen and water vapor diffuse very slowly through it, preventing oxidizing species from reaching the silicon surface under the nitride. In addition, the nitride itself oxidizes very slowly as the field oxide is grown (typically,

only a few tens of nm of nitride are converted to SiO_2 during the field-oxide growth process). Thus, the nitride should remain as an integral oxidation-barrier layer during the entire field-oxide-growth step. Figure 2-7b shows the thickness of the nitride converted to oxide during a wet silicon oxidation step. Silicon oxidizes approximately 25 times faster than silicon nitride. One criterion for selecting the nitride thickness is that it should be greater than the thickness that will be converted to SiO_2 during the subsequent field oxidation step.

Silicon-nitride films, however, have the well-known drawback of exhibiting a very high tensile stress when deposited by CVD on silicon (on the order of 10^{10} dynes/cm^2). The termination of intrinsic stresses at the edge of a nitride film gives rise to a horizontal force that acts on the substrate. Under some circumstances, this stress can exceed the critical stress for dislocation generation in silicon, and will thus become a source of fabrication-induced defects. For example, at 1000°C in oxidizing ambients, defects can be generated at the edges of nitride films as thin as 21 nm if they are deposited *directly* on silicon.[7] Pad oxides are used to combat these stresses and avoid dislocation generation. The effect of the pad layer is to reduce the force transmitted to the silicon at the nitride edge, and to relieve the stress of the nitride via the viscous flow of the pad oxide.

2.2.2.3 Mask and Etch Pad-Oxide/Nitride Layer to Define Active Regions. The active regions are now defined with a photolithographic step. A resist pattern is normally used to protect all of the areas where active devices will be formed. The nitride layer is then dry etched, and the pad oxide is etched by means of either a dry- or wet-chemical process. After the pad oxide has been etched, the resist is not removed but instead is left in place to serve as a masking layer during the channel-stop implant step.

2.2.2.4 Channel-Stop Implant. An implant is next performed in the field regions to create a channel-stop doping layer under the field oxide. In NMOS circuits, a p^+ implant of boron is used, while in PMOS (and in the n-tubs of CMOS circuits) an n^+ implant of arsenic is utilized. Although this normally requires two masking steps in CMOS circuits, a single mask process for implanting both p and n channel stops has been reported.[8] After the implant has been completed, the masking resist is stripped.

2.2.2.5 Problems Arising from the Channel-Stop Implants. During field oxidation, the channel-stop boron experiences both segregation and oxidation-enhanced diffusion. Thus, relatively high boron doses are needed (mid 10^{12}-10^{13} atoms/cm^2) in order for acceptable field threshold voltages to be achieved. This also implies that the peak of the boron implant must be deep enough that it is not absorbed by the growing field-oxide interface (implant energies in the 60-100 keV range are used). If the channel-stop doping is too heavy, it will cause high source/drain-to-substrate capacitances and will reduce source/drain-to-substrate *pn* junction breakdown voltages.

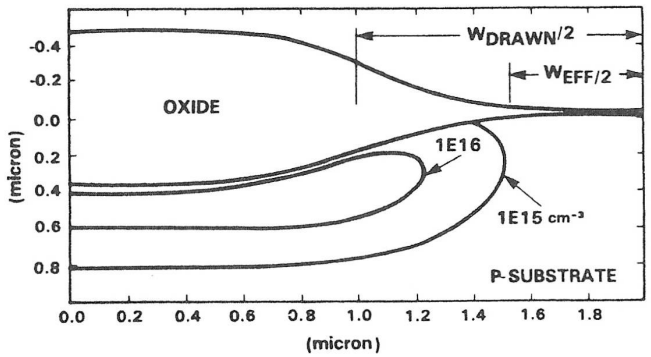

Fig. 2-8 Field oxide and boron encroachment in LOCOS.[96] From K. M. Cham et al., *Computer-Aided Design and VLSI Device Development*. Copyright 1986, Kluwer Academic Publishers. Reprinted with permission.

The lateral diffusion of the boron also causes it to encroach into the NMOS active areas (Fig. 2-8). Such redistribution raises the boron surface concentration near the edge of the field oxide, causing the threshold voltage to increase in that region of the active device. As a result, the edge of the device will not conduct as much current as the interior portion, and the transistor will behave as if it were a narrower device. (This effect is also enhanced as the dose of the channel-stop implant is increased.)

Finally, dislocations generated during the channel-stop implant step (which would ordinarily climb out to the surface during appropriate annealing conditions) can be caused to glide under the nitride-covered regions by the stresses at the nitride edges. The penetration of such dislocations through nitride-edge-defined junctions can cause increased emitter-base leakage in bipolar devices.[10]

The extent of the channel-stop dopant diffusion can be reduced by using high-pressure oxidation (HIPOX) to grow the field oxide,[8] by use of a germanium-boron co-implant,[9] or by use of a chlorine implant.[51] HIPOX allows the oxide-growth temperature to be reduced, which reduces the diffusion length of the boron. The germanium-boron co-implant exploits the fact that boron diffuses with a lower diffusivity in the presence of implanted germanium, and boron segregation effects are also reduced (see Vol. 1, chap. 7). The result is a 40% increase in field threshold voltage for the same dose of implanted boron, and a corresponding decrease in lateral-boron encroachment. The chlorine implant is performed in the field regions prior to field oxidation, and this causes the oxide to grow at a faster rate. Consequently, the field oxide can be grown in less time at the same temperature.

2.2.2.6 Grow Field Oxide. After the channel-stop implant has been performed, the field oxide is thermally grown by means of wet oxidation, at temperatures of around 1000°C for 2-4 hours (to thicknesses of 0.3-1.0 μm). The oxide grows where there is

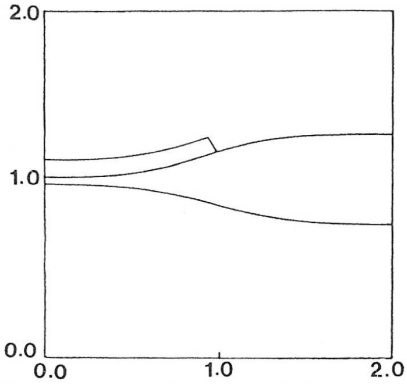

Fig. 2-9 Schematic of the bird's beak that occurs during semirecessed LOCOS.

no masking nitride, but at the edges of the nitride, some oxidant also diffuses laterally. This causes the oxide to grow under and lift the nitride edges. Because the shape of the oxide at the nitride edges is that of a slowly tapering oxide wedge that merges into the pad oxide, it has been named a *bird's beak* (Fig. 2-9). The bird's beak is a lateral extension of the field oxide into the active area of the devices. Although the length of the bird's beak depends upon a number of parameters – including the thicknesses of the buffer oxide, nitride, and isolation oxide (as well as the oxidation temperature and oxygen partial pressure) – the length for a typical 0.5-0.6-μm field oxide is ~0.5 μm/side. This would make a lithographically defined 1-μm feature disappear on the chip following the field oxidation step (Fig. 2-10). Although this effect led to predictions that LOCOS isolation would have to be replaced for device dimensions smaller than 2 μm, optimization of process steps has allowed conventional LOCOS to continue to be used for device-isolation spacings as narrow as 1.25-1.5 μm.[11,14]

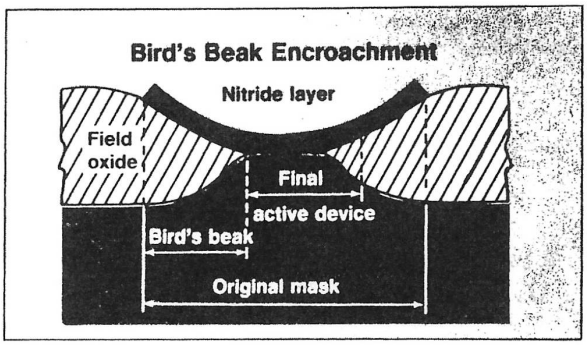

Fig. 2-10 Bird's beak encroachment limits the scaling of channel widths to about 1.2-1.5 μm.[11] Reprinted with permission of Semiconductor International.

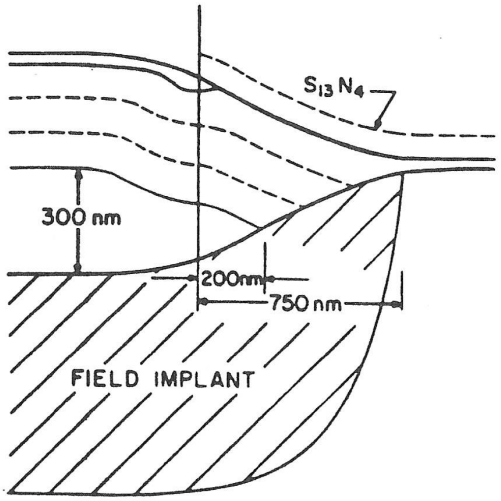

Fig. 2-11 Eventual exposure of the *p*-substrate region when etching back the LOCOS field oxide in an NMOS device (in an attempt to regain some of the active device area lost by bird's beak encroachment).

The LOCOS bird's beak also creates another problem during the later process step of contacting metal to the source and drain regions of an MOS device. Because of the shape of the bird's beak, any overetching that must be performed during the contact-window-opening step will etch away part of the bird's-beak oxide. This may expose the substrate region under the source or drain region (Fig. 2-11). If this occurs, the source will become shorted to the well region when the metal interconnect film is deposited, impairing or destroying device operation.

The problem worsens in CMOS as shallower junctions are used, due to exposure of the well region. (If deeper junctions are used, the lateral diffusion distance is also longer and etching back part of the bird's beak may not expose the well region.) In NMOS processes, this problem was counteracted by redoping of the contacts with phosphorus after they were opened (and before the metal was deposited). This created a deep junction in the contact areas only. The source/drain junction near the gate oxide was kept shallow in order for good device performance to be to retained. In CMOS processing, not only are shallow junctions used, but redoping of the contacts presents an additional problem because both n^+ and p^+ contact openings are defined at the same masking step. As a result, in CMOS processes enclosed contacts are commonly employed. This solution, however, reduces packing density, as area must be provided in the source and drain regions to keep the contact from overlapping the bird's beak. Another solution would be to use an alternative isolation process in which the contacts are able to coincide with (or overlap) the active areas.

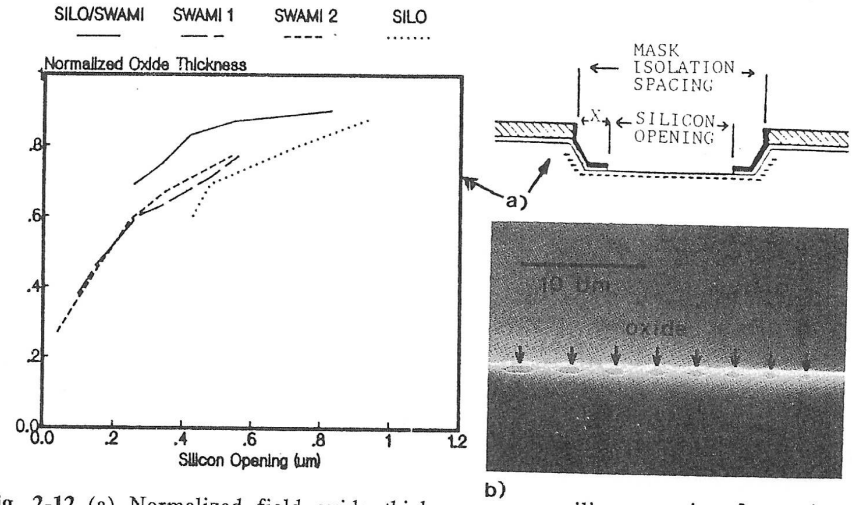

Fig. 2-12 (a) Normalized field oxide thickness versus silicon opening for various advanced LOCOS processes.[14] (© 1985 IEEE). (b) SEM photo of the thinning of the field oxide as the dimension of the exposed silicon gets smaller.[121] This paper was originally presented at the Spring 1989 Electrochemical Society Meeting held in Los Angeles, CA.

Modifications to the basic process have allowed the bird's beak to be reduced in length. These include etching back a portion of the field oxide after it is grown; using silicon nitride without a pad oxide; and using a thin pad oxide covered with polysilicon under the masking nitride layer. These advances will be described in more detail in a later section.

One last limitation of LOCOS-based isolation schemes for submicron structures is the oxide field-thinning effect.[14] The field-oxide thickness in submicron-isolation spacings is significantly less than the thickness of field-oxides grown in wider spacings (Fig. 2-12).[14,121] The narrower the width of the exposed substrate silicon region, the thinner the field oxide. For example, a field oxide that is grown to a thickness of 400 nm above a region of exposed Si that is 1.5-μm wide would be only about 290 nm thick if grown above a 0.8-μm-wide region of Si.

This effect is believed to be caused by the reduction in the oxidants available in the submicron opening compared to those available in wider openings.[15] Thin field oxides that result from this effect can have a great impact on field-threshold voltages and on the interconnect capacitances to substrate. It is predicted that as a result of this field-thinning effect and the need to maintain defect-free isolation structures (see next section), the minimum space that must be allowed for a LOCOS-type isolation structure with a field-oxide thickness of 550 nm will be 0.75 μm.

2.2.2.7 Strip the Masking Nitride/Pad-Oxide Layer. Following field oxidation, the masking layer is removed. Since 20-30 nm of the top of the nitride is converted to SiO_2 during the field oxidation, this layer must be etched off first. The

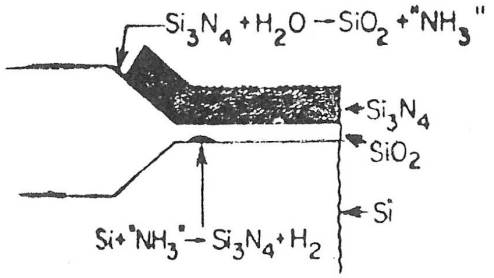

Fig. 2-13 Kooi's model for explaining nitride growth under the gate oxide.[16] Reprinted by permission of the publisher, The Electrochemical Society, Inc.

remaining nitride and pad oxide are then etched. These steps can be carried out by means of wet-chemical etching, since there is no need to maintain dimensions or to otherwise tightly control the etching. Since chemical methods that remove nitride offer excellent selectivity with respect to the underlying oxide, considerable overetching can be used.

2.2.2.8 Regrow Sacrificial Pad Oxide and Strip (Kooi Effect). During the growth of the field oxide, another phenomenon occurs that causes defects in when the gate oxide is grown. Kooi et al., discovered that a thin layer of silicon nitride can form on the silicon surface (i.e., at the pad-oxide/silicon interface) as a result of the reaction of NH_3 and silicon at that interface (Fig. 2-13).[16] The NH_3 is generated from the reaction between H_2O and the masking nitride during the field-oxidation step. This NH_3 then diffuses through the pad oxide and reacts with the silicon substrate to form silicon-nitride spots or ribbon (these regions are sometimes called *white ribbon)*. When the gate oxide is grown, the growth rate becomes impeded at the locations where the silicon nitride has formed. The gate oxide is thus thinner at these locations than elsewhere, causing low-voltage breakdown of the gate oxide. One way to eliminate this problem is to grow a "sacrificial" gate oxide after stripping the masking nitride and pad oxide, and removing it before growing the final gate oxide.[17,18]

2.2.3 Limitations of Conventional Semi-Recessed Oxide LOCOS for Small-Geometry ICs

The limitations of conventional LOCOS for submicron technologies can be summarized as follows:

- The bird's-beak structure causes unacceptably large encroachment of the field oxide into the device active regions (Fig. 2-10).

- Boron from the channel-stop implant of *n*-channel MOSFETs is excessively redistributed during the field-oxide growth and other high-temperature steps, leading to unacceptable narrow-width effects (see chap. 5).
- The planarity of the surface topography is inadequate for submicron lithography needs (see section 2.9.2).
- The thickness of the field oxide in submicron regions of exposed silicon is significantly thinner than that grown in wider spacings (Fig. 2-12). This field-oxide-thinning effect can produce problems with respect to field threshold voltages, interconnect-to-substrate capacitance, and field-edge leakage. Finally, it introduces additional nonplanarity across the surface of the wafer.

2.3 FULLY RECESSED OXIDE LOCOS PROCESSES

Several conventional, fully recessed oxide LOCOS methods have been developed for bipolar integrated-circuit applications, including the ISOPLANAR,[19] ROI,[20] and OXIS[21] processes. Such fully oxide-isolation processes allow higher performance bipolar ICs to be fabricated than if junction isolation is used. The collar of SiO_2 that forms the sidewall isolation eliminates the sidewall contribution to collector-to-

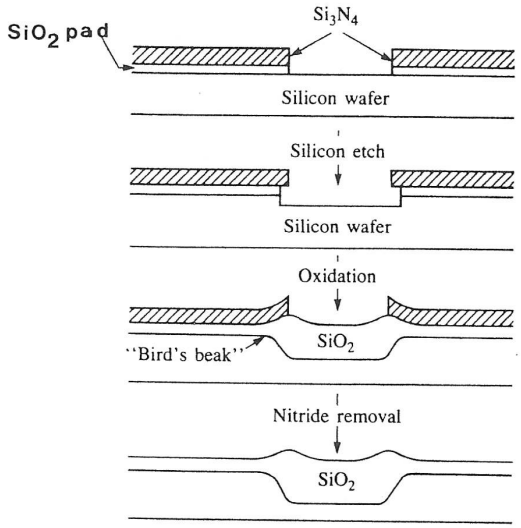

Fig. 2-14 Cross section depicting process sequence for fully-recessed oxidation of silicon.[128] From R. C. Jaeger, *Introduction to Microelectronic Fabrication.* Copyright, 1988, Addison-Wesley. Reprinted with permission.