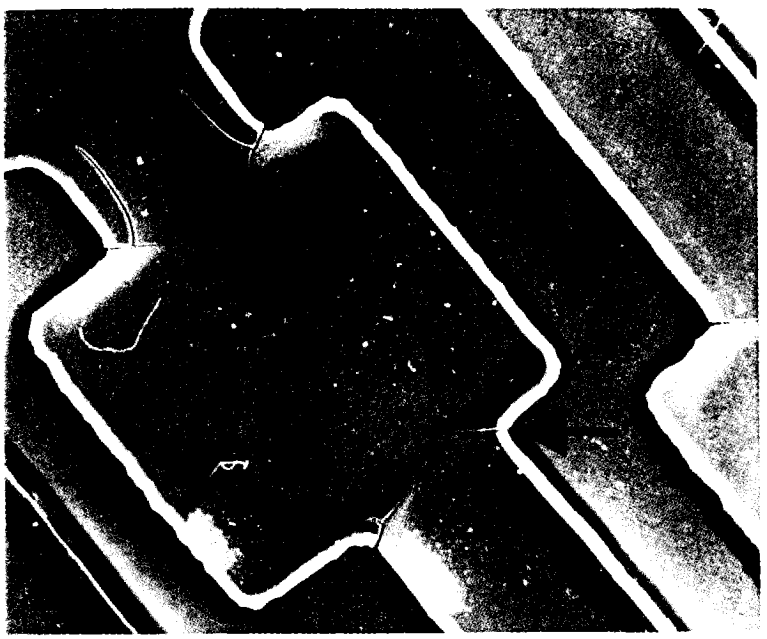


Fig. 6-8 Severe corner cracks in a nitride oxidation mask on silicon. The cracks occurred at concave corners. LOCOS thickness is 2-μm.[6] (© 1991 IEEE).

growing field oxide to lift the edges of the nitride. In some locations (e.g., 270° corners in the nitride film - see Fig. 6-8), the stresses in the brittle nitride film cause it to crack, defeating the purpose of the nitride to act as an oxidation barrier. The pad oxide flows during the field-oxide growth process in such a way that the stresses in the nitride can be kept small enough to prevent cracking. Stresses are also transmitted from the nitride to the silicon, and these can produce defects in the silicon crystal as well. The pad oxide also serves to prevent these defects. In general, the thicker the pad oxide, the fewer the defects in both the nitride films and the silicon substrate. However, a thick pad-oxide layer will render the nitride layer ineffective as an oxidation mask by allowing lateral oxidation to take place. Hence, the thinnest pad oxide that will serve as an effective stress relief layer is employed. Typically, pad oxide layers are 50 nm thick.

6.2.1.2 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

TSMC-1029
U.S. Patent 8,796,779

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 behaves as an integral oxidation-barrier layer during the entire field-oxide-growth step. Figure 6-9 compares the oxidation rate of silicon and silicon nitride, and shows that silicon oxidizes approximately 25 times faster than silicon nitride.[9] 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, thus becoming a source of fabrication-induced defects. The pad oxide, as described in the previous section, is used to combat these stresses and avoid dislocation generation.

6.2.1.3 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 areas on which active devices will later be formed. The nitride layer is then dry etched, and the pad oxide may be etched by means of either a dry- or wet-chemical process. In many cases, however, it is desirable to retain some of the pad oxide after the nitride etch. In this case, the dry nitride-etch process must be optimized to prevent all the pad oxide from being removed during the nitride overetch

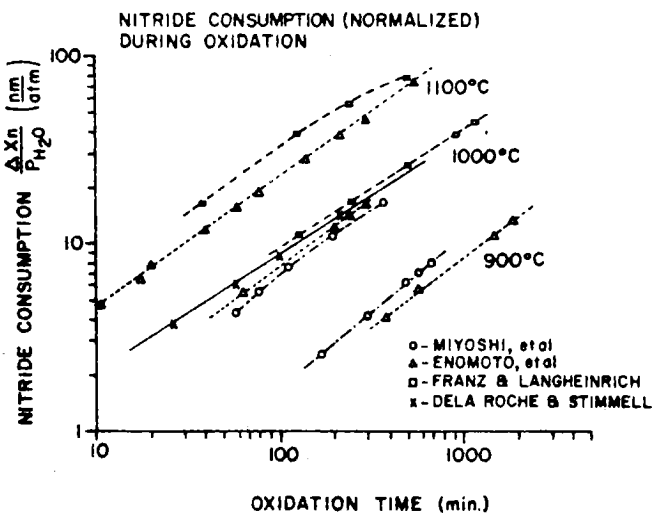


Fig. 6-9 Nitride consumption during oxidation.[9] Reprinted with permission of Semiconductor International.

that the resist is left in place to serve as a masking layer during the channel-stop-implant step, and must be stripped prior to the field-oxide growth step.

6.2.1.4 Channel-Stop Implant. An implant is 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.[11] After the implant has been completed, the masking resist is stripped.

6.2.1.5 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 ($5x10^{12}$-$5x10^{13}$ 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 typical). If the channel-stop doping is too heavy, it will cause high source/drain-to-substrate capacitances and reduce source/drain-to-substrate *pn* junction breakdown voltages.

The lateral diffusion of the boron also causes it to encroach into the NMOS active areas (Fig. 6-10). Such redistribution raises the boron surface concentration in the active regions near the edge of the field oxide, causing V_T to increase in that region of the active device. As a result, the edges of the channel will not conduct as much current as the central portion, making I_D smaller (narrow-width effect). This effect is also enhanced as the dose of the channel-stop implant is increased.

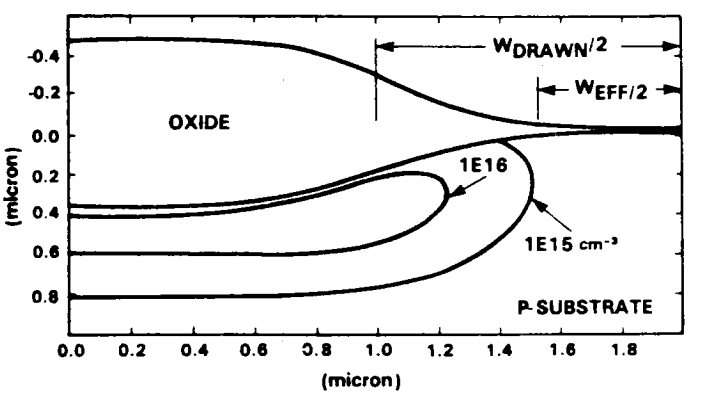


Fig. 6-10 Field oxide and boron encroachment in LOCOS.[10] From K. M. Cham et al., *Computer-Aided Design and VLSI Device Development.* Copyright 1986, Kluwer Academic Publishers. Reprinted with permission.

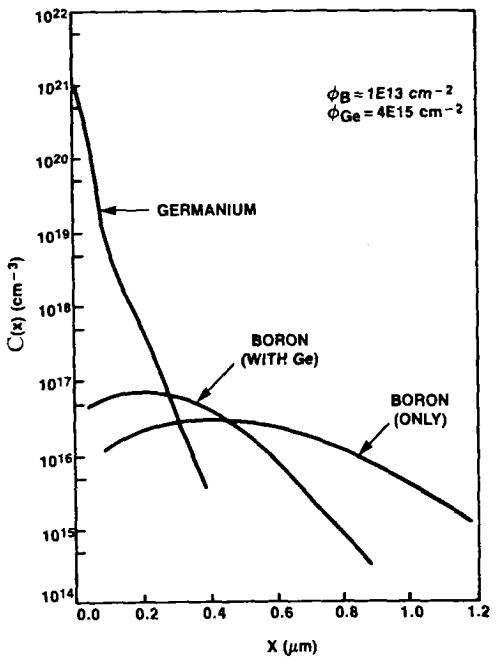


Fig. 6-11 Reduction of boron implant redistribution during field oxidation by co-implanting Ge. Figure shows implanted boron doping profile after field oxidation step with and without the presence of a co-implanted Ge dose. Boron doping profile is deeper without implanting Ge.[13] (© 1990 IEEE).

Several techniques have been studied to reduce the extent of the channel-stop dopant diffusion, including: (1) use of high-pressure oxidation (HiPOX) to grow the field oxide;[11,12] (2) co-implanting chlorine and boron into the field regions;[13] (3) use of a germanium-boron co-implant;[14] (4) use of a high-energy post-field-oxide channel-stop implant; and (5) implementation of a novel, field-implant-free well-drive process.[99]

HiPOX allows the oxide-growth temperature to be reduced, which reduces the diffusion length of the boron. The chlorine co-implant is performed in the field regions prior to field oxidation, causing the oxide to grow at a faster rate. Consequently, the same field oxide thickness can be attained in less time at the same temperature. The germanium-boron co-implant exploits the fact that boron diffuses with a lower diffusivity (by a factor of nearly three) in the presence of implanted germanium, and boron segregation effects are also reduced. 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. Figure 6-11 shows the effect of the Ge on the vertical boron profile in the field region after a field oxide step of 165 minutes at 1000°C in steam. The boron implant dose was $1x10^{13}$ cm^{-2} at 100 keV, and the Ge dose was $5x10^{15}$ cm^{-2}

at 150 keV. The Ge significantly enhances the final surface concentration of boron, and maintains a generally shallower profile with greater peak concentration.

A high-energy channel-stop implant following the field oxide growth greatly reduces the lateral penetration of boron because it no boron is present during the long, high-temperature field oxidation. In one reported process, a 400-keV boron channel-stop implant was implanted through a 0.8 μm field oxide (this same implant is also used in the process to create a retrograde *p*-well profile in the active areas).[15,118] Implanting the channel-stop dose after field oxidation appears to become even more attractive as device dimensions shrink toward 0.5 μm because the field-oxide thickness is also reduced in advanced, submicron CMOS processes. For example, for 0.5 μm devices, this thickness is projected to be ~0.4 μm. Such thinner field oxides require implant energies $\leq$250 keV for boron (and $\leq$600 keV for phosphorus). High-energy implanters, with implant energies exceeding 1 MeV, are now commercially available, making these latter processes manufacturable. The application of MeV implantation to well formation and isolation in CMOS technology is described further in sections 6.6 and 6.7 and in chap. 8.

The novel field-implant-free approach is reported to be applicable to a *p*-well or twin-well CMOS process, and is described more fully in section 6.7.1.1. Briefly here, a sufficiently high boron concentration at the Si surface is achieved without a channel-stop implant by using a unique well drive-in process. After implanting the boron well dopants, most of the well dopant distribution is accomplished in a non-oxidizing ambient *after* field-oxide growth. The time-at-temperature during field-oxide growth makes up the first part of the well drive-in, during which time the boron near the oxide surface is depleted by impurity segregation effects. However, during the after-oxide drive-in, boron dopants deep in the bulk diffuse back toward the surface and replenish the boron lost near the surface to the field oxide (Fig. 6-12). Since no channel-stop implant is used, boron does not encroach into the active regions, and the problem of lateral boron diffusion is eliminated. The associated narrow-width effect is also thereby avoided. Nominal V_{TF}s >10V at isolation spacing of 1.2 μm were reported.

6.2.1.6 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 0.8-1.0 μm). The oxide grows where there is no masking nitride, but at the edges of the nitride, some oxidant 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. 6-13). The bird's beak is a lateral extension of the field oxide into the active area of the devices. Although, as we shall discuss in the following section, 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) – for a typical 0.7-0.8-μm field oxide it is ~0.5 μm/side. This would make a lithographically defined 1-μm feature disappear on the chip following the field oxidation step (Fig. 6-14). 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

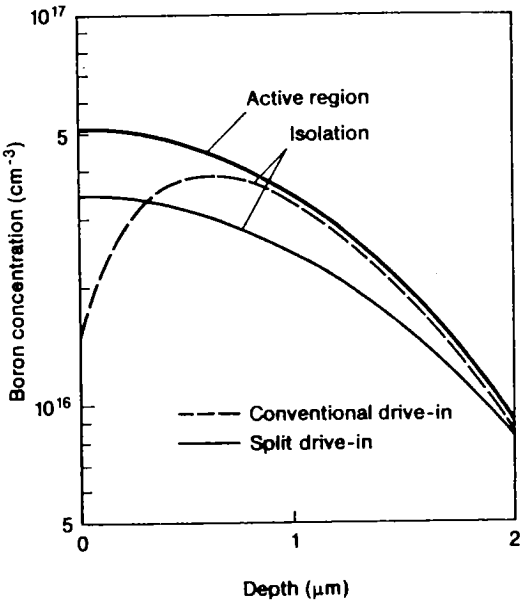


Fig. 6-12 SIMS profiles of boron well underneath field oxide and in the active area for split and non-split drive-in. Implanted boron dose: $8x10^{12}$ cm^{-2}. The active area profile is the same for both drive-in cases.[99] (© 1990 IEEE).

continue to be used for device-isolation spacings as narrow as 1.25-1.5 μm.[16,17] For dimensions smaller than this, modifications to basic LOCOS which reduce bird's beak had to be incorporated. One such modification is poly-buffered LOCOS, to be described later.

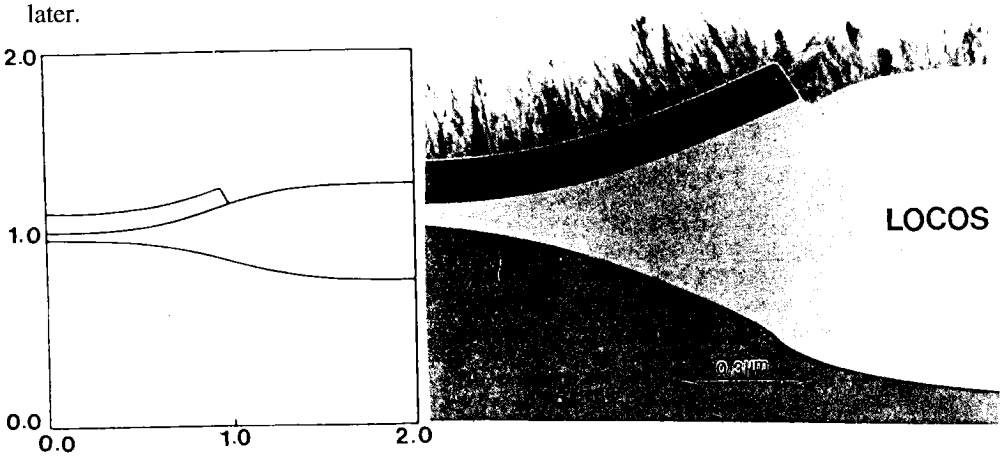


Fig. 6-13 Schematic of the bird's beak that occurs during semirecessed LOCOS.

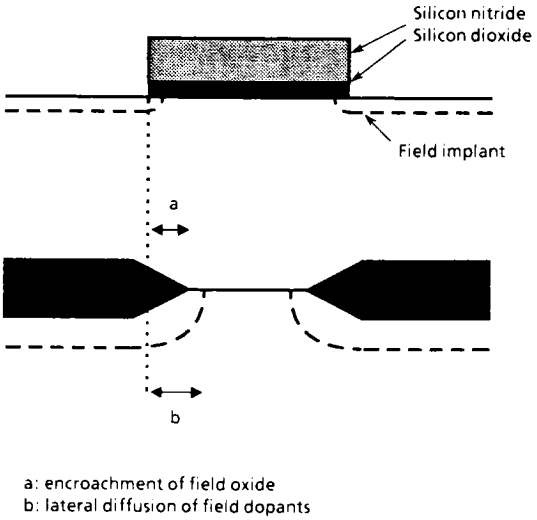


Fig. 6-14 Bird's beak encroachment limits the scaling of channel widths to about 1.2-1.5 μm.[16] Reprinted with permission of Semiconductor International.

Another limitation of LOCOS-based isolation schemes for submicron structures is the *oxide field-thinning* effect.[17] The field-oxide thickness in submicron-isolation spacings is significantly less than the thickness of field-oxides grown in wider spacings (Fig. 6-15).[17,18] 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. In 0.2 μm-wide windows, the thinning can be as great as 80%. Thin field oxides that result from this effect can have a great impact

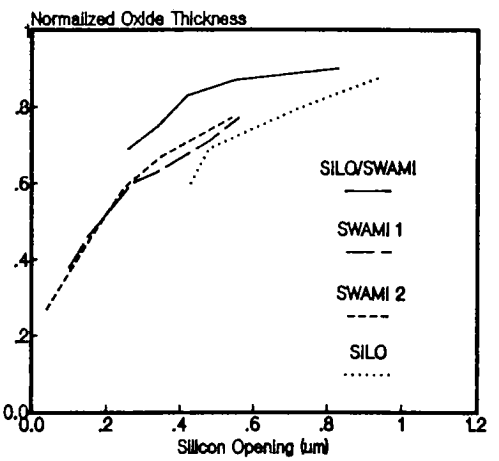


Fig. 6-15 Normalized field oxide thickness versus silicon opening for various advanced LOCOS processes.[17] (© 1985 IEEE).

on V_{TF} values and on the interconnect capacitances to substrate.

The field thinning effect is believed to be caused by higher compressive stress in the oxide growing in smaller openings.[19] The stress is sufficient to reduce the diffusivity of the oxidant through the oxide. Field oxides grown at 1200°C showed less thinning than those grown at 1000°C (in the same size openings). At 1200°C the viscosity of the growing oxide is reduced by roughly two orders of magnitude. This permits the oxide to flow and thereby reduce the growth inhibiting stress. Unfortunately, raising the growth temperature is not a feasible solution in most cases. Higher temperatures increase bird's beak length and exacerbate unwanted diffusion of channel-stop dopants in CMOS or BiCMOS. Indeed, the field-thinning effect and the need to maintain defect-free isolation structures (see next section), may limit the minimum space for LOCOS-type isolation structures with a field-oxide thickness of 550 nm to ~0.8 μm. Recently, however, it has been reported that use of HiPOX may be a way to reduce the severity of the problem (see section 6.4.1 for more details).

6.2.1.7. Factors Which Impact Bird's Beak Length and Shape. We mentioned earlier that the length and shape of the bird's beak is impacted by a number of factors. They include the following: pad layer thickness and composition; silicon crystal orientation; temperature of the field oxide growth process; thickness of the nitride layer; mechanical properties of the nitride layer; and mask stack geometry. Let us consider each of these factors in more detail:

1. *Pad layer thickness.* In general, lateral oxidation should be reduced by making the pad oxide thinner, leading to a shorter bird's beak (Fig. 6-16).[20] However, since the purpose of the pad oxide is stress reduction, it cannot be

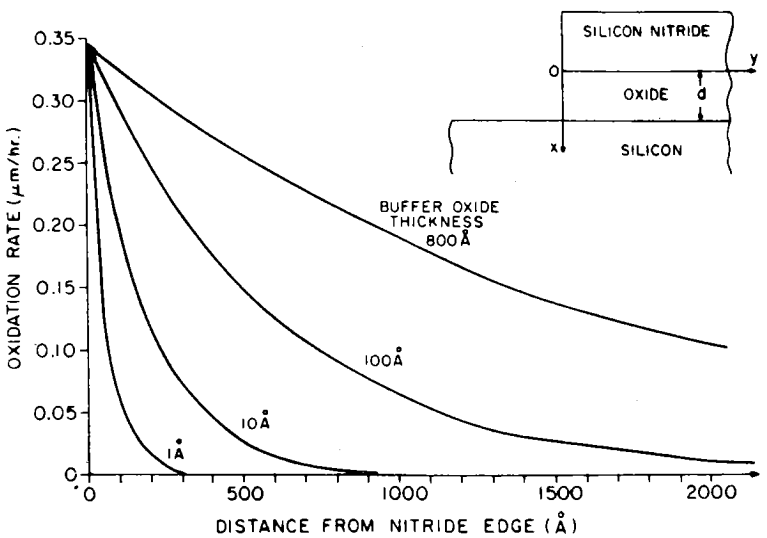


Fig. 6-16 Oxidation rates versus distance from the edge of the nitride edge with the buffer (pad) oxide thickness as a parameter.[20] (© 1982 IEEE).



Fig. 6-17 Suppressed LOCOS. Bird's beak reduction is achieved by replacing the pad oxide of the conventional LOCOS structure with an oxynitride. Note that the oxynitride is hardly oxidized. TEM of cross section.[3] (© 1991 IEEE).

made arbitrarily thin. The minimum pad oxide thickness will depend on the oxidation conditions. Also note that LOCOS structures with no pad oxide have been investigated, termed SILO structures (*sealed interface local oxidation*). In such cases, a native oxide of 10-20Å is still usually present, giving rise to some bird's beak formation.

2. *Pad layer composition.* Although direct contact of the nitride with the silicon can dramatically reduce bird's beak effects (i.e., pad oxide thickness is zero), defects may be thereby produced in the silicon. A variation that has been investigated is to employ a pad layer composed of CVD oxynitride (SiO_xN_y), rather than a thermally grown oxide. The material properties of such layers may be a compromise between those of pure nitride and oxide - with respect to the oxidation masking capability of nitride and the stress relief qualities of oxide. Kuiper et al., studied the oxidation behavior of such films as a function of their composition.[22] They concluded that oxynitride films with composition ratios of $0< x/y <0.5$ (nitride rich) show better oxidation resistance during field oxide growth than do oxygen rich films. Figure 6-17 shows a TEM photomicrograph of a LOCOS structure with a CVD oxynitride pad layer. Note that the bird's beak penetration is smaller in the latter structure, and that oxynitride pad layer is not noticeably oxidized.

3. *Silicon crystal orientation.* Since the oxidation rates of different crystal planes in silicon vary, the bird's beak lengths depend on the silicon crystal

orientation. Shorter bird's beaks are found on <111> silicon wafers than on silicon wafers with <100> surface orientation.

4. *Field oxide process temperature.* Shorter bird's beaks are found as the field oxidation temperature is increased. The oxide is less viscous at higher temperatures. As it forms under the nitride edge it is more easily squeezed out by the pressure of the nitride mask which is resisting being lifted.

5. *Thickness and mechanical properties of the nitride layer.* In general, all other factors being equal, the thicker the nitride layer, the less bird's beak penetration. The thicker the nitride, the stiffer it will be, and thus will resist the edge lifting that accompanies bird's beak penetration. However, if the nitride film is too thick, it is likely to crack. Consequently, the thickness of masking nitride layers is usually chosen to be 250 nm or less.

The mechanical properties of the nitride are also a major factor that influences the bird's beak. It has been reported that at elevated temperatures Si_3N_4 is a more viscous than elastic material.[23] As a result, once the nitride mask edge is lifted up by the growing bird's beak, it will have no further tendency to compress the oxide grown under it. The nitride seems to have a higher viscosity than SiO_2, but this also depends on the stoichiometry of the nitride. Silicon-rich nitride films were found to exhibit a higher resistance to flow, resulting in less nitride bending and less encroachment of the field oxide under the nitride edge.

6. *Mask stack geometry.* The shape of LOCOS structures is also dependent on the mask geometries. That is, the following effects have been observed: (1) the bird's beak penetration is significantly greater at convex corners (particularly at the end of long, narrow mask stack structures, e.g., mask widths $\leq 1.0\ \mu m$); (2) bird's beaks are shorter at concave corners; and (3) bird's beaks are shorter for narrow mask features.[6,18] (It should be noted that these effects are less pronounced in the poly-buffered LOCOS process discussed in section 6.3).

This is explained as follows: At the end of long, narrow mask stack lines (Fig. 6-18), the nitride layer is less resistant to bending than along the narrow dimension of the line, hence allowing greater lateral oxidation to occur as the nitride is lifted. The bird's beak length at the end of the line can be 2-3 times the normal bird's beak length. Convex (270°) corners are also less resistant to bending than concave (90°) corners, allowing further lifting of the nitride edge, and greater lateral penetration at the convex corner.

The extended bird's beak penetration effect can impose limitations on packing density, since more overlap area is needed for contacts to the end of small device areas. In addition, the region of the longer bird's beak is susceptible to junction leakage because etchback of the thin long beak after *pn* junction formation can easily expose the junction.

7. *Isolation Lifting* (or *Bird's Beak Punchthrough*). If the bird's beak on either side of a mask-stack-isolation line penetrates under the line such that the

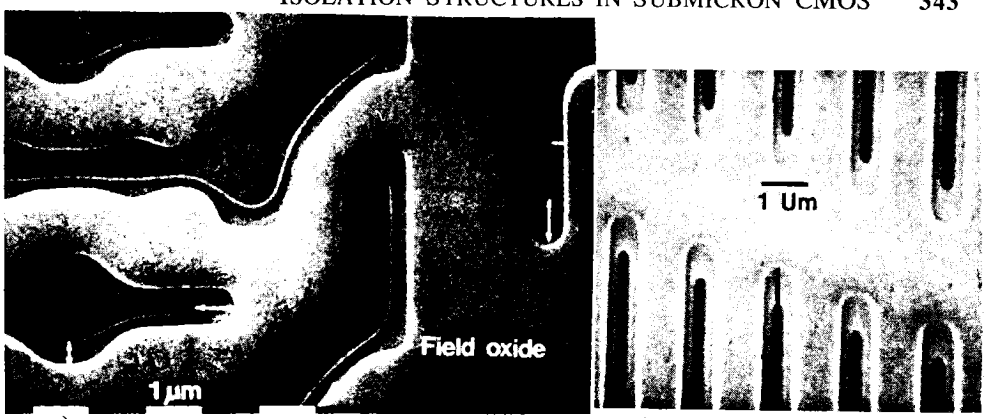


Fig. 6-18 Bird's beak dependence on mask geometry: (a) Top view of a CMOS RAM cell. The bird's beak length (white arrows) is different in concave and convex corners; (b) Enhanced bird's beak (EBB) at the end of narrow active areas. BB refers to the normal bird's beak.[3] (© 1990 IEEE).

active-area width reaches zero, it is said that *isolation lifting,* or *bird's beak punchthrough* has occurred. Figure 6-19a shows this effect.[131] The effect occurs for two reasons: (1) Wider isolation spacings permit the mask stack to deform at the edges, relieving the stress that is lifting the lines. As the isolation space is made narrower, the mask stack becomes more rigid with respect to edge bending, and thus will not deform. Instead the entire mask stack line will lift. Thus, an optimum nitride thickness must be used. That is, the nitride must be thick enough to suppress bird's beak penetration by resisting upward deformation at the mask line edges, yet it must not be so thick that

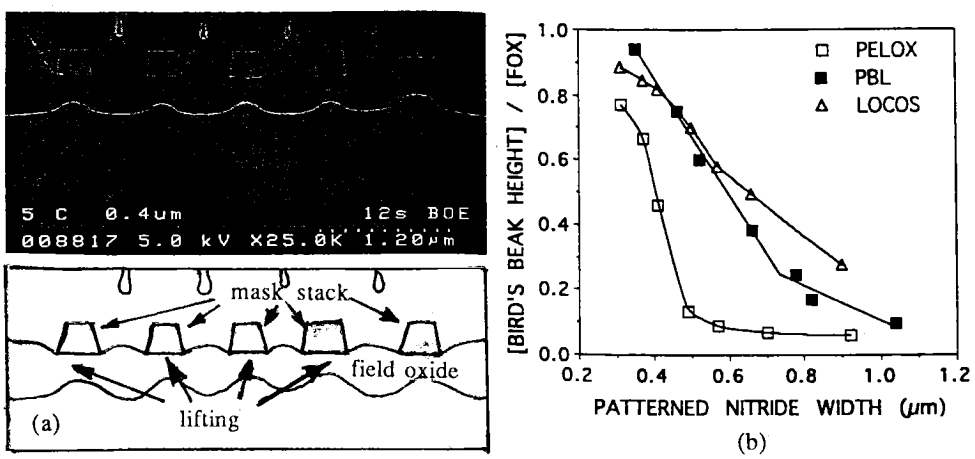


Fig. 6-19 (a) Bird's beak lifting[131]; (b) Bird's beak height under the center of the isolation line as a function of isolation line width, and for three different isolation technologies.[132] (© 1994 IEEE).

deformation cannot occur at all, instead giving rise to lifting. A suggested PBL mask stack with optimized layer thicknesses for both isolation widths and isolation spacings of 0.6 μm consists of 2500Å/500Å/100Å-thick nitride/amorphous silicon/pad oxide layers, respectively.[131]; (2) As the active area dimension is decreased, the concentration of the oxidizing species merges at the center of the isolation line at a very early stage of field oxidation. In this case, as the field oxidation proceeds, the oxidizing species concentration quickly reaches supersaturation at all locations beneath the line, resulting in a surface-reaction limited growth condition. A uniform oxide punchthrough profile beneath the isolation line thus results (see Fig. 6-19a). This suggests that if the lateral diffusion of the oxidizing species under the nitride can be limited, oxide lifting can be reduced. In the PBL and PELOX processes (which are advanced variations of LOCOS described in section 6.4), nitride cladding of the mask-stack sidewalls is used to reduce this lateral diffusion for the purpose of suppressing the oxide lifting effect.[122]

Figure 6-19b shows the bird's beak height under the center of the isolation line as a function of isolation line width, and for three different isolation technologies.[132] Note that the advanced LOCOS isolation processes can reduce the effect for isolation spacings down to ~0.5 μm.

However, if isolation lifting occurs, the thicker oxide underneath the nitride mask must then be etched back, resulting in a reduced field oxide thickness elsewhere, which degrades the isolation of the small active spaces. In addition, increased field etchbacks are limited by the susceptibility of LOCOS to the field thinning effect in small active spaces. As a solution, a post-field-oxide implantation has been proposed in order to locally increase the well doping and compensate for the thinner field oxide. However, achieving acceptable MOSFET performance with this solution requires that a very critical alignment-sensitive photomasking step be used to keep the implant out of neighboring active regions. Reference 132 describes the integration of a PELOX isolation structure and a split-well drive-in scheme to realize the same benefits, but without the need for such an extra post-field oxide implant, allowing this scheme to provide adequate isolation for CMOS technology down to 0.35 μm.

From this list of factors, we can see that accurately predicting the length and shape of the bird's beak is a formidable task. We shall see in a later section the attempts that have been made to model such 2-D and 3-D oxidation effects. In practice, some experimentation is generally required to get the optimum conditions for making the desired device structures.

6.2.1.8 Use of High-Energy Ion Implantation to Introduce the Channel-Stop Dopant After Field Oxide Growth. As noted earlier, the commercial availability of high-energy implanters now makes it possible to implant the channel stop dopants (chiefly boron) through the field oxide, rather than having to

implant them into the substrate prior to field oxide growth. This approach offers several attractive benefits. First, a higher concentration of boron at the surface under the oxide can be achieved with the same implant dose than if the implant is done prior to field-oxide growth. This is possible because the leaching of the boron that takes place during field oxide growth is avoided, and this leaching causes the boron surface concentration of at the silicon surface to be reduced (see section 6.2.1.5.) Thus, the field oxide thickness needed to attain the target V_{TFiso} value can be reduced. Second, such thinner field oxides exhibit smaller bird's beak. This means that smaller isolation spacing design rules can be used which, in turn, translates to a reduced chip size. Third, the thinner field oxide makes the topography of the wafer after the semirecessed field oxide growth step smoother. Smoother topography may extend the use of existing stepper tools for one or more device generations, and will delay the need to introduce chemical-mechanical polishing for back-end processing. Finally, the redistribution of channel-stop dopants by lateral diffusion during the field oxide growth step is largely replaced by lateral straggle of the dopant occurring from the high-energy implant step.[112] Since the latter effect results in approximately ~25% less lateral dopant spread, this also contributes to smaller isolation spacing design rules, and thus reduced chip size.

6.2.1.9 Strip the Masking Nitride/Pad-Oxide Layer. Following field oxidation, the masking layer is removed. For several reasons, this removal process must be optimized (and during manufacture, carefully controlled) in CMOS technologies smaller than 2.0 μm.

First, since 20-30 nm of the top of the nitride is converted to SiO_2 (and oxynitride) during field oxidation, the removal of the nitride must in fact begin with a step to remove this layer. (The very top surface actually consists of a layer of approximately stoichiometric SiO_2. Beneath this is an oxynitride region that is graded in composition from SiO_2 to Si_3N_4 i.e., SiO_xN_y.) If wet chemical removal of the nitride is used, an HF "deglaze" step is first usually used to remove the SiO_2/mixed SiO_xN_y layer. (Note also that vapor HF is reportedly replacing wet HF deglaze processes in some advanced processes.[24]) Hot (180°C) phosphoric acid (H_3PO_4) is then used to etch the remaining nitride. The etch rates of SiO_2 in 5% HF, 0.5% HF and hot H_3PO_4, respectively are 293±73, 70±17, <1 Å/min. The etch rates of silicon nitride in these same chemicals are, respectively, 8, nil, and 50 Å/min. This indicates that etching in hot phosphoric acid offers excellent selectivity with respect to the underlying oxide and considerable overetching during this step can be used.[25]

However, during the deglaze step, the HF also etches the field oxide. Thus, an optimum deglaze etch time must be established. (If the deglaze does not completely remove the oxynitride layer, the remaining residue may delay the onset of nitride etching by the H_3PO_4 - as hot H_3PO_4 is ineffective at removing oxide. As a result, the nitride may not all be removed in the nominal time allotted for the nitride stripping process. Such inadequate removal of the nitride can sometimes result in scrapping of wafer lots. On the other hand, too long a deglaze step will excessively etch the field oxide.) To assist in the development of an optimal deglaze process, a thorough characterization of the material properties of the nitride film after field oxidation should thus be conducted.

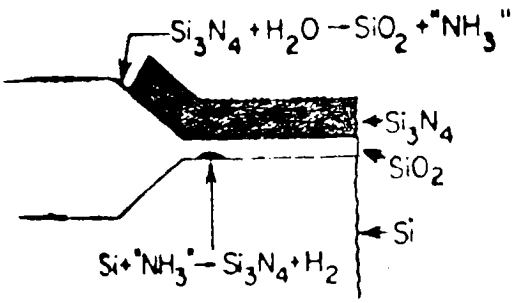


Fig. 6-20 Kooi's model for explaining nitride growth under the gate oxide.[26] Reprinted by permission of the publisher, The Electrochemical Society, Inc.

If a dry process is used to remove the oxidized nitride, the variation in the etch rates of the two layers must still be taken into account when choosing the etching conditions. The oxynitride layer generally etches more slowly than the underlying nitride. The amount of field oxide removed during this dry etch step also must be considered.

After stripping the nitride, the pad oxide must also be removed, so that a gate oxide can be grown over the active device regions. The etchant used to remove the pad oxide will also attack the field oxide, hence the conditions of this step must also be carefully determined and controlled. In fact, the field-oxide thickness-reduction is cumulative to the deoxidation of the nitride layer, the removal of the pad oxide, and removal of the sacrificial gate oxide (see next section). Typically, 100-200 nm of field oxide may be removed by the sum of these three etching steps. Some of the thinner part of the bird's beak will also be removed, reducing lateral oxide penetration.

6.2.1.10 Regrow Sacrificial Pad Oxide and Strip (Kooi Effect). During the growth of the field oxide, another phenomenon occurs that can cause defects later 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) close to the border of the active regions, as a result of the reaction between the oxidizing species (O_2 or H_2O) and the silicon nitride (Fig. 6-20).[26] That is, NH_3 (or some other NH compound or nitrogen) is generated from the reaction between H_2O and the masking nitride during the field-oxidation step. This NH_3 then diffuses through the oxide and reacts with the silicon substrate to form silicon-nitride spots or ribbon (these regions are also sometimes called *white ribbon* because they have this appearance when viewed under an optical microscope due to light scattering, see Fig. 6-21). Under the field oxide the reaction of NH_3 with Si would not be very effective due to competition with the oxidation reaction of H_2O with Si. However, at some distance from the field oxide edge, under the nitride mask, the concentration of H_2O might be assumed to be low enough to allow an effective nitridation reaction to occur. This would allow some silicon nitride (or possibly, silicon oxynitride with oxidation masking capability), to form. Direct observation of this masking layer by transmission electron microscopy has also been reported.[139]

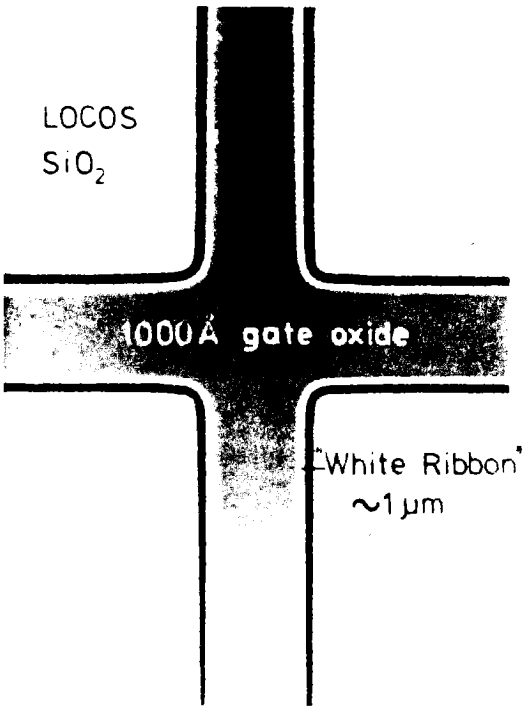


Fig. 6-21 Top view of a LOCOS structure with a "white ribbon" effect.[6] (© 1991 IEEE).

When the nitride mask and pad oxide layer are removed, there is a possibility that the nitride ribbon would remain present. When the gate oxide is subsequently grown, the growth rate becomes impeded at the locations where such silicon nitride remains. The gate oxide is thus thinner at these locations than elsewhere (Fig. 6-22), giving rise to problems of low-voltage breakdown of the gate oxide. One way to overcome this problem would be to etch the pad oxide long enough to also ensure that the ribbon is removed. Unfortunately, too much field oxide is removed at the same time.

Instead, the most widely method to eliminate the white ribbon problem is to grow a wet "sacrificial" gate oxide (typically 25 - 50 nm thick) after stripping the masking nitride and pad oxide. This sacrificial oxide layer is removed by wet etching before growing the final gate oxide.[27,28] The chemical removal of the sacrificial oxide will again consume part of the field oxide and, as noted above, must likewise be optimized.

A double sacrificial-oxide process has also been reported. Its benefit is a smoother active-area silicon surface on which the final gate oxide is grown. This results in smaller gate leakage currents (due to the onset of Fowler-Nordheim tunneling at higher fields), a lower interface state density, and a lower gate-oxide defect density.[29]

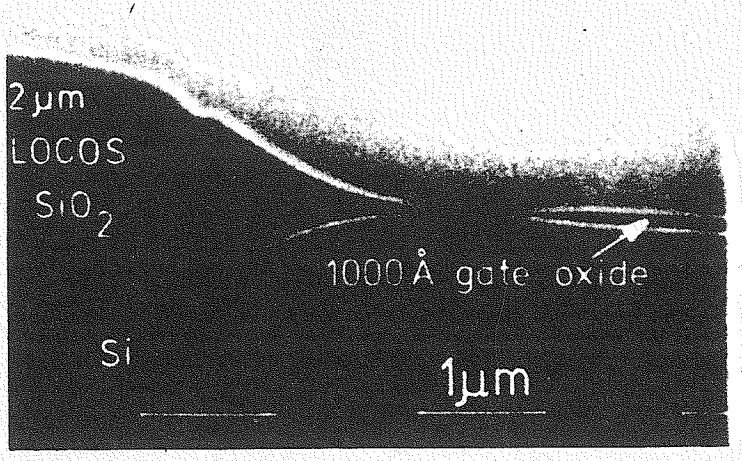


Fig. 6-22 SEM picture of a cross-section of a LOCOS structure with a "white ribbon" effect.[6] (© 1991 IEEE).

6.2.2. Planarizing Semirecessed LOCOS Structures

As described in chap. 4, Vol. 2, very planar surfaces are required for accurate lithography and for implementing multilevel interconnects in submicron ICs. However, semi-recessed LOCOS structures produce non-planar conditions (i.e., smooth steps) on the wafer surface. LOCOS structures modified to reduce bird's beak (such as PBL described in section 6.4), produce topographies with steeper steps than conventional LOCOS (even though the step heights are comparable). To return the wafer surface to a flat condition after semi-recessed LOCOS, efforts have been made to planarize semi-recessed LOCOS structures.

One example of such a planarization scheme is shown in Fig. 6-23.[15] A mask stack consisting of a 50 nm oxynitride pad layer and a 100 nm nitride film is first formed. An 800 nm field oxide is grown. Prior to planarization (by RIE etchback), a 200 nm thick layer of CVD oxide is deposited to fill the gaps between the nitride edge and field oxide (Fig. 6-23d). Then, when the field oxide is etched back, the gap will not be etched into the remaining planarized field oxide structure. (Fig. 6-23e). This procedure is reported to produce bird's beak free field oxide for thicknesses below 500 nm.

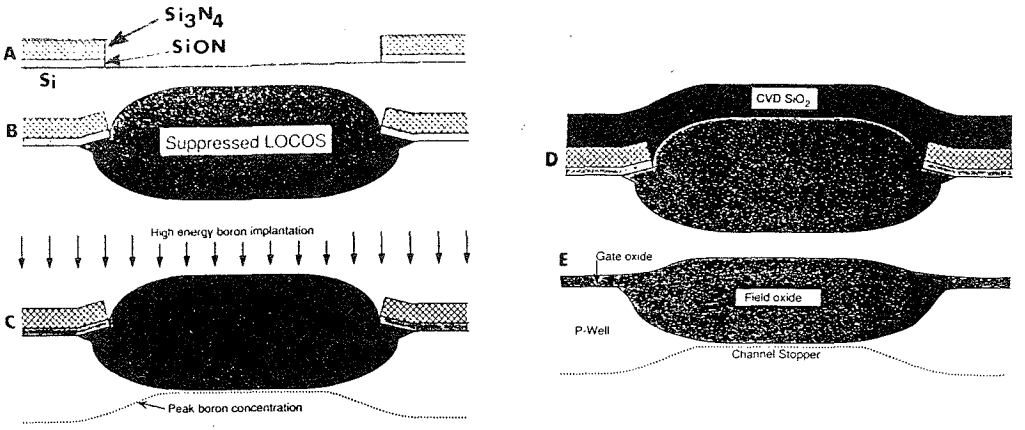


Fig. 6-23 A scheme to planarize semi-recessed LOCOS. In this example SLOCOS is combined with a high energy ion implantation scheme to provide channel stoppers and retrograde wells to make high density CMOS devices.[15] (© 1987 IEEE).

6.3 MODELING THE TOPOGRAPHY OF LOCOS STRUCTURES

For the case of 1-D oxidation, it is possible to accurately simulate oxide growth under virtually all practical conditions. When oxidation occurs in two dimensions, (e.g., "bird's beak" growth beneath the nitride-mask edge in LOCOS), simulation of oxide growth is more difficult. Complexities associated with 2-D oxidation include: (1) nonplanar geometries; (2) moving boundaries; and (3) volume expansion arising from the formation of new oxide at the Si surface. In the latter effect, the newly formed oxide expands and pushes out the old oxide, which rearranges itself through viscous flow.

Nevertheless, both empirical and physically based models have been developed to provide two-dimensional simulations of oxide growth in LOCOS structures. (Some process simulators such as SUPREM IV contain both types, and the user is able to select the desired model for a particular simulation). The *physical models* solve the complete set of differential equations governing both the oxide growth and the simultaneous viscous flow. Although such models can give accurate results, they are complex and require long computation times. The *empirical models* – which are expressed in terms of relatively simpler analytical functions – involve much less computational effort.

6.3.1 Empirical Models of 2-D Thermal Oxidation

Empirical models have been implemented to simulate the "bird's beak" of LOCOS structures. The 2-D model in SUPRA for local oxidation uses an empirically determined shape function, based on such input parameters as pad-oxide thickness, silicon-nitride thickness, oxidation temperature, and pressure. That is, the contour of the "bird's beak" is approximated with a complementary-error function whose parameters are defined by the particular LOCOS process conditions. The empirical model in Stanford's SUPREM IV is essentially the same as that used in SUPRA.

Another empirical model that also treats semi-recessed LOCOS oxide shapes was developed by Guillemot et al. and has been incorporated in the TSUPREM-4 simulator, available from TMA, Inc.[30] This model again uses a simple parametric relationship that is based on the analysis of experimental data. It characterizes the bird's beak by three main parameters: (1) the bird's beak length l_{bb}; (2) the extent the nitride edge has been lifted, H_2, and the depth of the bird's beak below the surface at the nitride edge, H_1. It also classifies the shapes of the bird's beaks that occur in semi-recessed LOCOS into two groups, depending on the stress exerted by the nitride masks (Fig. 6-24): (1) If the nitride layer is relatively thin, low mask stress exists and smooth birds-beak shapes occur (Case 1 in Fig. 6-24). For such shapes, two analytical complementary-error functions are used to fit the oxide contours: one for the oxide-silicon interface, and the other for the oxide-nitride interface. In Case 2, the nitride layer is thicker and the pad layer thinner. This causes stronger mask stresses, in which case the oxide becomes "pinched" beneath the nitride edge. Therefore, two functions are used to simulate the bird's beak contours at each interface: one on the left side of the pinch and the other on the right side. Figure 6-25 shows that in each case the bird's beak shape predicted by the simulation closely duplicates the shapes observed in SEM photos.

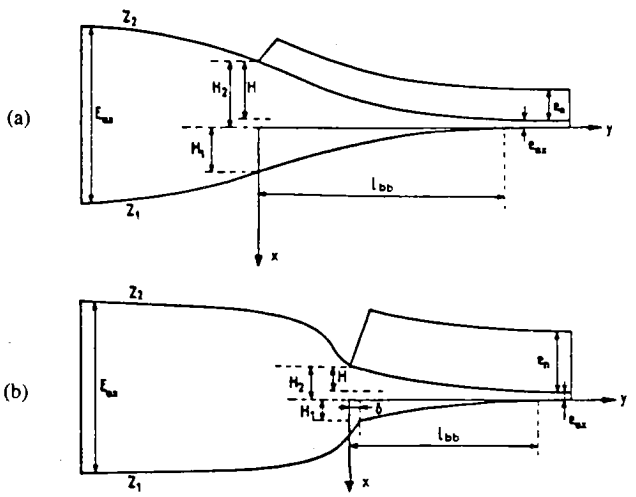


Fig. 6-24 The two shapes of the bird's beak with characteristic lengths (E_{ox}, l_{bb}, and H) and processing parameters (ε_{ox} and ε_n). (a) Shape 1 and (b) shape 2.[30] (© 1987 IEEE).

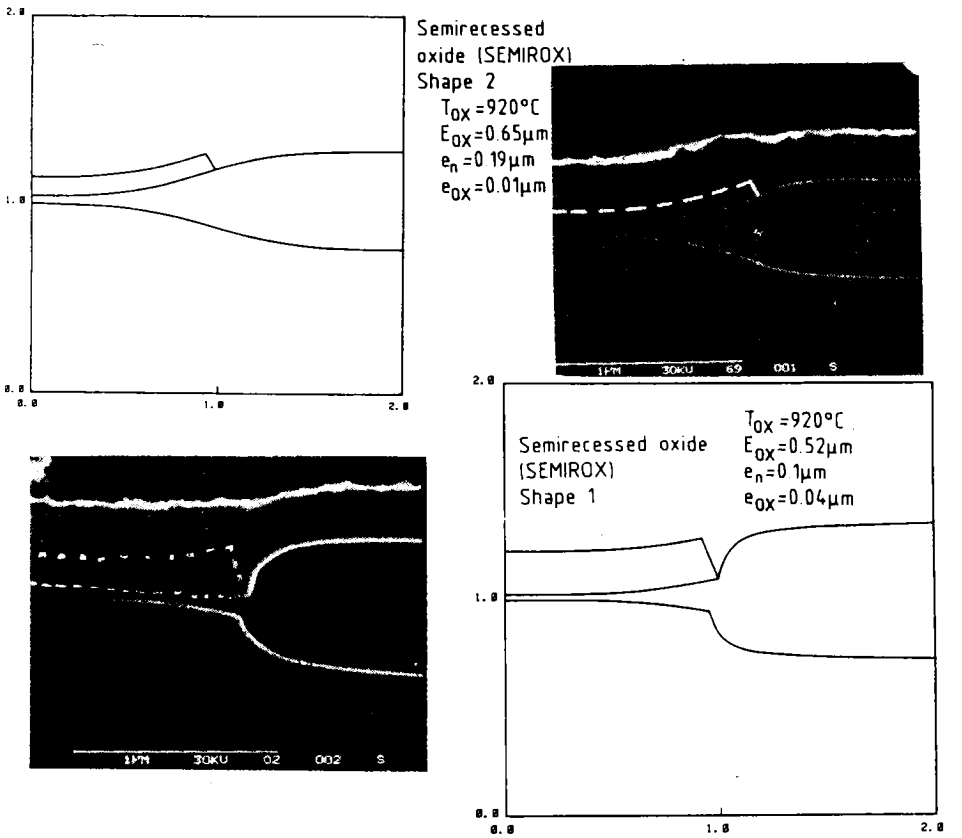


Fig. 6-25 Comparison between experimental and simulated field oxide shapes.[30] (© 1987 IEEE).

6.3.2 Physically Based Models of 2-D Oxidation

As mentioned above, the physically based models of 2-D oxidation solve the equations of the Deal-Grove linear-parabolic model (or the equivalent steady-state diffusion model) and at the same time account for the volumetric expansion due to oxide formation and viscous flow. The viscous-flow phenomenon was first quantitatively studied by Eer Nisse, using silicon wafers that had a thermal-oxide film on only one side.[46] Since the thermal-expansion coefficient of SiO_2 differs from that of silicon, the wafer was bowed when the oxidation temperature was less than 960°C. If the temperature was increased above 960°C, however, the wafers remained flat, which suggests that the stress due to thermal mismatch was relieved by the viscous flow of the oxide.

The first physically based model of 2-D oxidation that became available in a public-domain simulator was designed to handle LOCOS processes and was incorporated in the

Stanford Oxidation Analysis Program (SOAP).[31] It assumed that the growing SiO_2 films behave like viscous fluids that flow when the thermal-oxidation temperature exceeds 960°C, with the viscosity of the oxide determined by the curvature of the wafers. The viscosity of thermally grown SiO_2 films decreases exponentially with increasing temperature. Above 960°C the oxide may flow with relatively little viscous resistance.[32]

The oxidation process is modeled by assuming that two mechanisms occur: (1) oxidant diffuses through the SiO_2 to the SiO_2/Si interface, and (2) the oxide flows due to the volume expansion and associated stress buildup. The oxidant diffusion in the SOAP model is described by a generalized diffusion equation (rather than by the flux-conservation equations of the 1-D model given in Volume 1, chapter 7):

$$D \nabla^2 C = \frac{\partial C}{\partial t} \cong 0 \qquad (6 - 1)$$

The boundary conditions for this equation for LOCOS structures are nonhomogeneous, and hence cause the oxidant distribution to be different from place to place along the silicon/oxide interface. The volume-expansion rate therefore also varies with position. The flow of the oxide is described by a simplified Navier-Stokes equation as

$$\mu \nabla^2 V = \nabla P \qquad (6 - 2)$$

where μ, V, and P are the viscosity, the velocity, and the pressure of the oxide, respectively. To model the localized oxide structures well, one needs information on the mechanical properties of the materials used in LOCOS processing, namely SiO_2 and Si_3N_4.

The goal of simulating 2-D-oxidation is to determine the position of the oxide boundaries at the conclusion of the oxide-growth cycle. Since these boundaries are continuously moving, it is very difficult to obtain this information from the above equations. The SOAP program incorporates a boundary-value numerical technique to solve the integral form of Eqs. 6-1 and 6-2 by using Green's function. Details of this method are given in reference 31.

Figure 6-26 shows the simulation of two semirecessed LOCOS field oxides using SOAP. In Fig. 6-26a, the nitride mask is 25 nm thick. Since the stress is small, the oxide moves in a direction normal to the interface throughout its entire growth. In Fig. 6-26b the nitride mask is much thicker (175 nm), and it thus prevents the oxide from moving normal to the interface by forcing it to move laterally during growth. The compressive stress due to the thicker nitride film reduces oxide growth by making the SiO_2 flow toward the open surface.

At temperatures below 960°C, oxides no longer exhibit viscous behavior, but instead deform as nonlinear viscoelastic materials. This has consequences for mechanical stress and bird's beak. Another model must therefore be used to treat such conditions. Such a model was developed by Kao et al. and has been incorporated into SUPREM IV.[33, 34] This model takes into account the stresses that build up in the growing oxide film and how such stresses impact the growth rate. Experimental evidence to support the model shows that these stresses cause the oxide to grow most slowly at concave corners. The

more general nature of this model allows it to simulate the growth of oxide in trenches as well as in LOCOS processes. Figure 6-27 shows a SUPREM IV simulation of an oxide grown in a trench in dry oxygen for 240 minutes at 950°C.[34] Thinning of the oxide in the lower corners is predicted for this structure. In addition, the stresses induced by growing oxide will be lower at higher temperatures. This implies that bird's beak dimensions will then be smaller, because the nitride edge can more successfully resist lifting if the oxide can flow aside. Figure 6-28 compares the bird's beak length and edge lifting of various experimental LOCOS structures with simulation results obtained using SUPREM-IV. Relatively good agreement is displayed.

CREEP is another physically based 2-D oxidation simulator developed to incorporate the stress effects associated with the growth of thermal SiO_2.[35,36] This simulator incorporates a modified form of Kao's 2-D thermal-oxidation model. The major change from the earlier model is the use of a shear-stress-dependent viscosity model in lieu of the hydrostatic-stress-dependent model. CREEP simulations of 2-D oxide thicknesses on both convex and concave cylindrical silicon structures over a wide range of temperatures and curvature conditions show better agreement with experimental

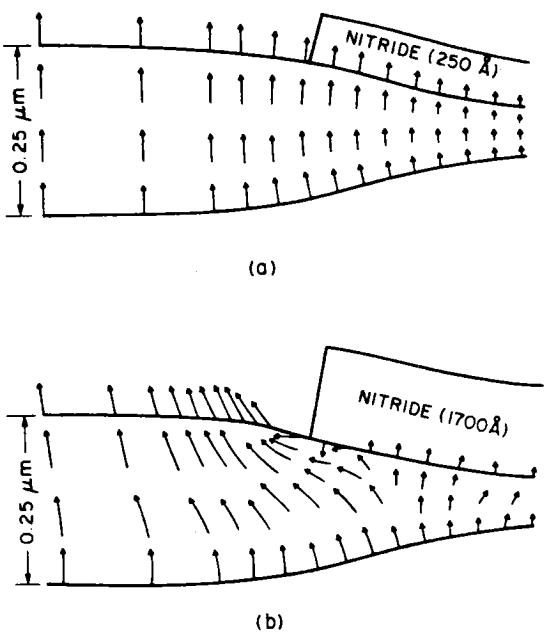


Fig. 6-26 Two-dimensional oxide growth simulated using SOAP with (a) a thin nitride (25 nm) masking layer, and (b) a thicker (170 nm) nitride masking layer.[31] (© 1983 IEEE).

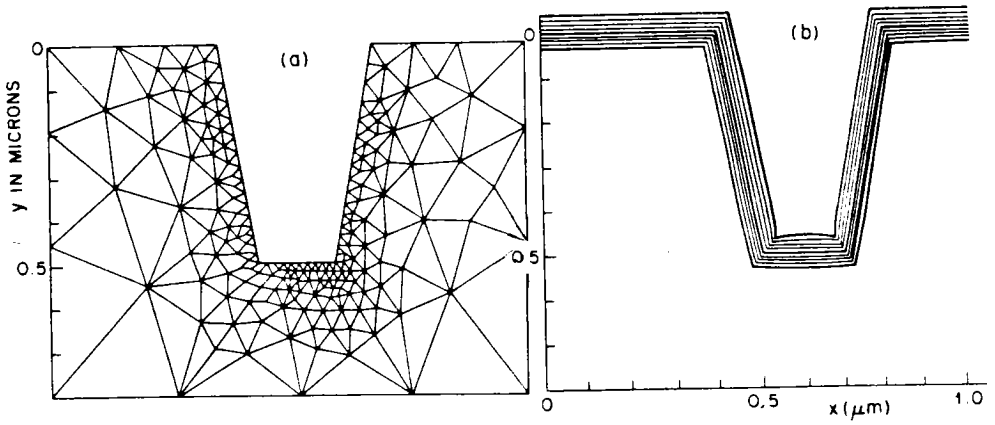
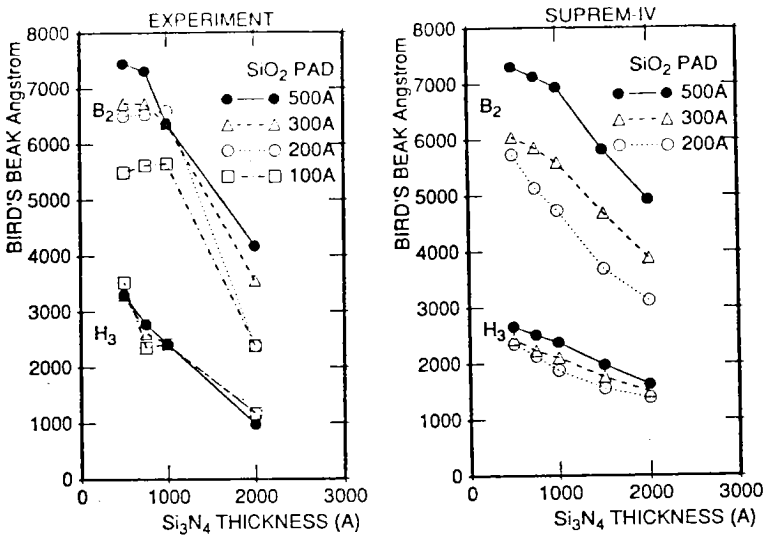


Fig. 6-27 Oxidation of a trench structure using SUPREM IV. (a) Initial grid. (b) Final oxide profile.[136]

measurements than do simulations based on previous models. The model used indicates that the effect of stress during oxidation on the linear-growth-rate constant dominates the two-dimensional profile-shape evolution.

Figure 6-29a shows the CREEP simulation of an oxide grown on a vertical step in wet oxygen for 3 minutes at 1000°C. This simulation clearly predicts the thinning of the oxide in the lower and upper corners of the step, as is observed in experimentally



6-28 Comparison of SUPREM-IV simulations to parameterized data on LOCOS structures.[3] (© 1990 IEEE). Experimental data is from Bassous.[135]

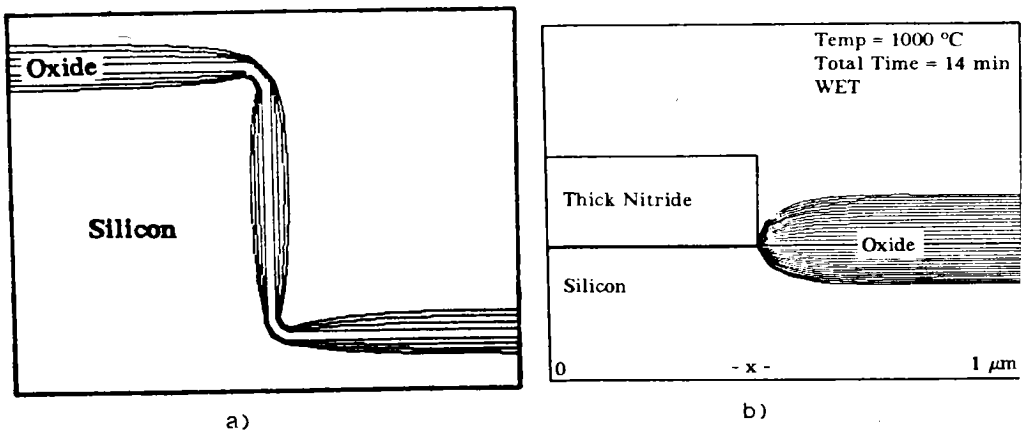


Fig. 6-29 (a) CREEP simulation of an oxide grown on a vertical step in wet oxygen for 3 minutes at 1000°C. (b) CREEP simulation of a SILO process using a 2-nm-thick pad oxide, a thick nitride mask, and an oxide-growth temperature of 1000°C.[35] (© 1986 IEEE).

grown oxides on trench structures. Figure 6-29b shows a CREEP simulation of a SILO process using a 2-nm-thick pad oxide, a thick nitride mask, and an oxide-growth temperature of 1000°C. CREEP is also capable of calculating the stress along the SiO_2/Si interface as a function of time and to simulate reflow of deposited oxides.

Oxidation-induced stress in LOCOS structures have been calculated by Isomae et al.,[37] and others.[123] Compressive stress in the bird's beak region is presumed to cause retardation of the oxidation rate (both the oxidant diffusion and the interface oxidation rate may be impacted) and thus influence the bird's beak shape. It is also reported that plastic deformation is initiated at the edge of the nitride mask, and expands into the silicon substrate as oxidation proceeds.[123]

6.3.3. Simulating Thermal Oxidation in 3-D Using Physical Models

Earlier we discussed that corners in the patterns of the mask stack can have a significant effect on the length of bird's beak: longer bird's beak arises at convex corners and shortened bird's beak at concave corners. Simulations with 3-D thermal oxide simulators confirm this effect. Results of 3-D numerical simulations with the aid of a supercomputer and the 3-D process simulator SMART-P[38] have been published by Unimoto and Odanaka.[39] In 3-D calculations of the stress in LOCOS structures, they assume elastic beam bending of the edges of the nitride mask during local oxidation though, as discussed earlier, according to Griffin[23] this may not be totally correct. Their simulation of the corner effects mentioned above quantitatively considered two factors: (1) the stress distribution in the LOCOS oxide; and (2) the local flux of oxidant. Figure 6-30 shows an example of calculations around corners in semi-recessed oxide patterns. In

the case of a square "hole" structure (oxidation mask surrounded by LOCOS), the formation of bird's beak is significantly enhanced at the corners of the hole because the stress at the corners is lower and the flux of oxidant is larger at such a 270° corner. In the case of a "LOCOS island" structure, with concave (90°) corners, it is more difficult to bend the nitride edge at the corners, resulting in enlarged pressure at these locations. Hence, the bird's beak is reduced there.

These 3-D simulations also support the idea that it is important to limit the bending of the nitride edge by making the mask sufficiently thick in submicron isolation structures. That is, Odanaka's simulations confirmed that at the end of long, narrow (<2 μm-wide) nitride lines the bird's beak can be surprisingly long because the nitride edge can be easily lifted at such locations.

Another report studied the sensitivity of the bird's beak profile to variations in the

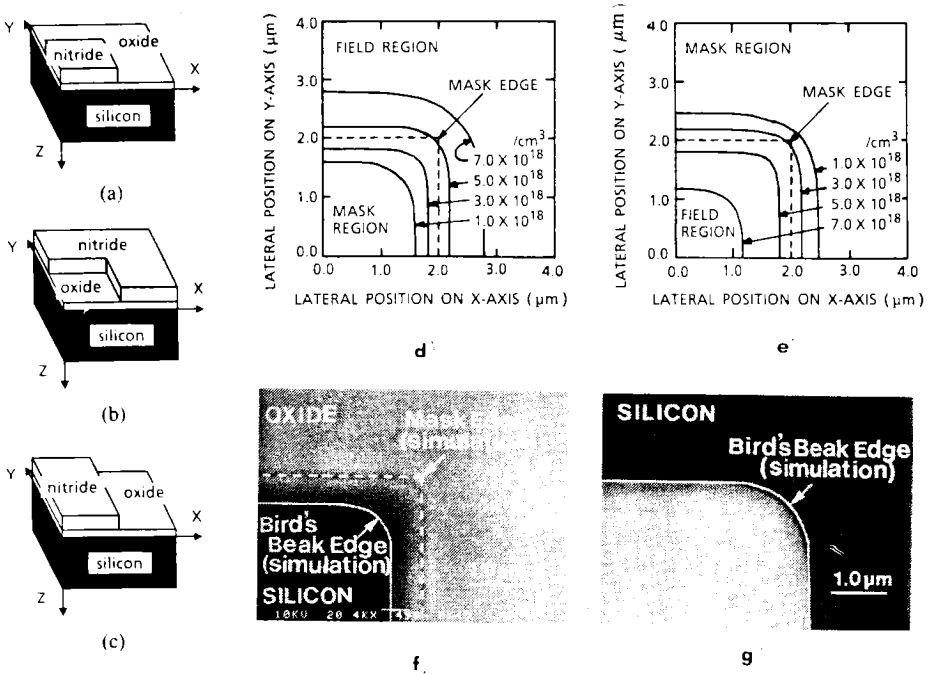


Fig. 6-30 Three typical nitride mask structures before field oxidation: (a) hole structure; (b) island structure; and (c) line structure. Calculated position of the bird's beak edge: (d) hole structure; (e) island structure. Top view SEMs of LOCOS structure after nitride and pad oxide stripping: (f) Hole structure; (g) Island structure. Solid and dashed lines show the bird's beak edge and the mask edge by simulation, respectively.[38] (© IEEE 1988).

active pattern geometries.[122] They used experimental data of bird's beak heights and widths for various active pattern widths for conventional LOCOS, poly-buffered LOCOS and PELOX (described in section 6.4). They also compared the data to simulation of the bird's beak profile obtain from SUPREM4 simulations.

6.4 POLY-BUFFERED LOCOS

To reduce the undesirable encroachment of bird's beak, various semirecessed- and fully-recessed LOCOS bird's-beak compaction schemes have been developed, including *polybuffered-LOCOS* (PBL), SILO, SWAMI, and OSELO. These have allowed LOCOS-based isolation structures to continue to be used in MOS ICs down to 0.6 μm. Of the above structures, we elect to discuss only PBL here, as it appears to be the most widely adopted LOCOS-based isolation structure for 1.0-0.5 μm ICs. (An example of a study favorably comparing PBL to SILO is given in ref. 40.) SILO (*sealed-interface local oxidation*) relies on interface sealing and SWAMI (*sidewall masked isolation*) on sidewall masking to reduce bird's beak to almost zero. However, these approaches are less attractive than PBL because they are susceptible to defect formation leading to either excessive diode leakage or gate-oxide degradation. SWAMI also involves a much more complicated process sequence. Fabrication details of these and other LOCOS-based structures can be found in Vol. 2, chap. 2.

All bird's-beak compaction schemes exploit the fact that a thinner pad oxide and a thicker nitride produce a shorter bird's beak. As the pad oxide is reduced and the nitride layer thickened, increasingly greater stresses arise during field oxidation. In conventional LOCOS structures this will eventually lead to defects in the silicon. In PBL, however, a polysilicon layer is deposited on the pad oxide before depositing the nitride layer. The polysilicon absorbs some of the stress produced during field oxidation, permitting the use of a thinner pad oxide and a thicker nitride film without the producing defects. A typical mask stack in a PBL process (Fig. 6-31a) consists of a very thin pad oxide (8-10-nm), an LPCVD poly layer (40-60 nm), and a thicker LPCVD nitride layer (e.g., 180-250 nm).[41,42,43]

PBL exhibits two other significant advantages. First, the dependence of bird's beak length on mask stack geometry discussed in section 6.2.1.7 appears to be less severe than in conventional LOCOS. A study of the mask geometry effects on bird's beak in PBL structures by Wils et al., reports this conclusion.[44] Second, the white ribbon problem seems to be mitigated by the PBL approach. The underside of the nitride mask is shielded from oxidation and thus less NH_3 is produced. With less NH_3 available, there is less propensity to form Si_3N_4 on the Si surface under the gate oxide.

A detailed explanation of how the poly provides stress relief is given in Ref. 45, but briefly, the poly apparently gets squeezed out of the edge of the mask stack by compressive forces arising from volume expansion (caused by the bird's beak oxide growth) and the rigidity of the thick top nitride. The small-grained nature of the poly allows some of it to be squeezed out, thereby relaxing the stress that would otherwise be applied to the silicon substrate. The subsequent oxidation of this squeezed-out poly also produces a feature on the field oxide named a *crab-eye*. Such features have been observed

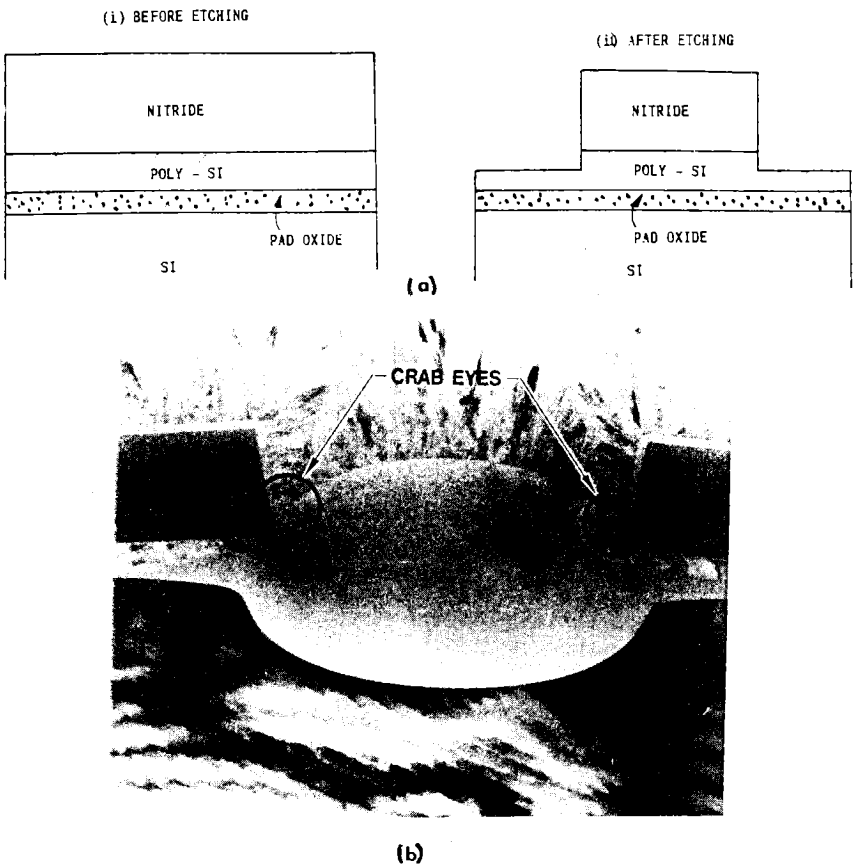


Fig. 6-31 (a) Poly-buffer LOCOS mask stack. (b) 'Crab eyes' at edges of poly-buffer LOCOS.[45] Reprinted by permission of the publisher. The Electrochemical Society.

in field oxides grown using PBL mask stacks (Fig. 6-31b). The formation of the crab-eye thereby also retards the oxidation rate near the mask edge by providing a longer oxide path to the substrate for the oxidizing species. As a result, the degree of lateral oxidation (which is responsible for bird's beak), is even further reduced. Conventional PBL processes yield a bird's beak of only 0.2-0.3 μm/side for field oxides of 400-600 nm thickness (compared to 0.5 μm per side for LOCOS). Leakage currents of the n^+-p junction comparable to those of a conventional LOCOS process are reported, provided the bird's beak penetration is not reduced below ~0.3 μm per side.

Reference 46 provides an extensive description of a PBL isolation technology optimized for an 0.8-μm CMOS process. By carrying out a full-factorial experimental plan the effects on defect formation, bird's-beak penetration, and isolation-leakage currents as a result of varying the pad-oxide-layer and nitride-layer thicknesses were studied. An additional description of a conventional PBL technology developed for a

high-volume 0.8 μm CMOS process is given in ref. 47. Included is a discussion of how this process can be enhanced to provide wider manufacturing process-latitude. These reports also address some of the issues unique to the fabrication of PBL structures, including:

1. Special precautions should be taken to minimize oxidation of the poly layer prior to the deposition of the nitride layer (such as inserting the wafers into the nitride deposition tube more rapidly than the usual ~25 cm/min, and allowing no wet chemical cleanups of the poly between poly and nitride deposition). The presence of a significant poly-oxide layer obstructs the removal of the sandwich layers after field oxidation, and provides a more efficient path for lateral oxidation - leading to enhanced bird's beak encroachment.

2. When the PBL stack layer is patterned, it may be desired to leave approximately one-half of the poly layer (as well as the pad oxide layer) unetched. The remaining poly-plus-pad-oxide film serves as a screen layer to minimize damage and contamination during the channel-stop implant. The poly remaining on the pad oxide in the field regions is oxidized during the field oxidation step;

3. After field oxidation, the removal of the PBL stack is similar to the removal of the mask stack in conventional LOCOS, with two exceptions:

a. It has been reported that the action of the H_3PO_4 with respect to nitride etching is inconsistent when a poly layer is below the nitride. This can lead to inadequate removal of the nitride, and may even lead to scrapping wafer lots.[25]

b. The poly layer must be carefully stripped. In ref. 38, the poly is dry-etched in a manner that leaves the pad-oxide intact. The pad-oxide is then removed in a 10% HF solution, with a lengthy overetch used to remove a portion of the bird's beak that penetrates under the nitride. This reduces the final bird's beak encroachment to ~0.13 μm. In ref. 48, the removal of the poly by dry etching and by steam oxidation (followed by wet etching in HF) was studied. MOS capacitors fabricated after dry-etch removal of the poly showed worse gate-oxide endurance under high-field stressing than did those in which the poly was removed by steam oxidation.

c. In two other reports, wet stripping of the nitride/poly mask layers was compared with an all dry stripping process.[128,138] It was found that the gate oxides grown after the dry process exhibited worse oxide integrity behavior than those grown after the wet strip was performed. This degradation was blamed on damage (in the form of increased interface states in the substrate) created in the silicon substrate.

4. Voids can occur in the polysilicon layer, during field oxide growth (Fig. 6-32). In ref. 49 Lin et al, propose that these voids arise by a Kooi-like effect (see section 6.2.1.9) that causes nitridation of some of the poly grains, and oxidation of others during field oxide growth. They argue that the nitrided poly

grains are removed when the nitride is stripped in hot H_3PO_4, resulting in voids in the poly along the mask stack edges. After the nitride is stripped, the oxided grains remain behind as a residue on the poly surface. If this residue is removed using dip in a weak (150:1) H_2O:HF solution, the etchant will also attack the pad oxide exposed by the poly voids. Then, when the poly is removed, the pad oxide may be breached, and the Si substrate attacked by the poly etch (assuming the selectivity of poly-to-SiO_2 is not high enough during poly removal). This will produce pits in the Si substrate. PBL isolation structures in which pitted substrates occur exhibit higher leakage currents than conventional LOCOS. It was suggested that the dip in weak HF be replaced with a short dry process that attacks the ribbon residue much faster than the pad oxide to prevent Si-substrate pitting.

Yang et al., however, observed that such voids occur in the poly before the nitride is removed, and also that they increase in density as the width of the active-area (nitride mask width) decreases.[50] A significant number of voids begin to appear when the width shrinks below 1.0 μm. Furthermore, as the nitride/poly thickness ratio is increased more voids are observed. Thus, it was

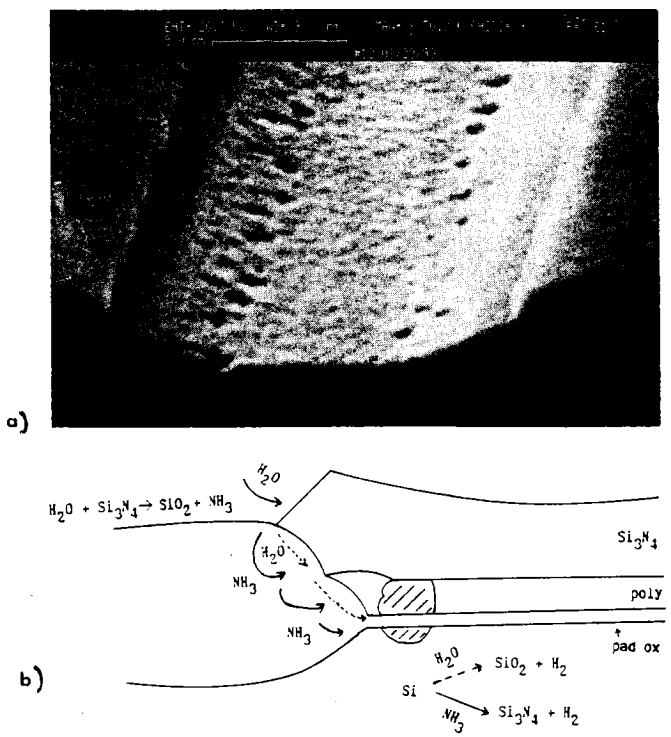


Fig. 6-32 Voids (or pits) are observed on polysilicon around the field oxide edge after nitride removal by hot phosphoric acid.[49] Reprinted by permission of the publisher. The Electrochemical Society.

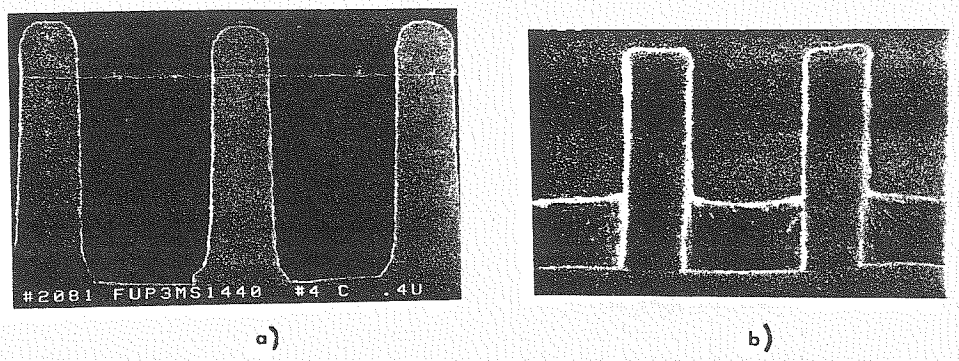


Fig. 6-33 "Microfoot" at the patterned resist/nitride interface: (a) Examples of 0.40 μm photoresist imaged over 2500Å nitride/500Å polysilicon/100Å pad oxide, in which the microfoot can be seen; (b) 0.40 μm lines imaged over 2000Å nitride/500Å polysilicon/100Å pad oxide. Microfoot does not occur. [129] This paper was originally presented at the Spring 1994 Meeting of the Electrochemical Society, Inc. held in San Francisco, CA.

> proposed that the voids arise from stress between the nitride and poly layers. A thicker, narrower nitride feature does not bend upward as bird's beak oxide forms under the edges, producing more stress in the poly film. Yang postulates that the stress is relieved by void formation. It was also reported that voids were not observed if the oxidation temperatures exceeded 1000°C or if the poly deposition temperature was <500°C.

It has also been reported that some unique issues arise in patterning PBL mask layers. That is, standing waves introduced by the interaction of the nitride layer and photoresist can produce a "micro-foot" at the patterned resist/nitride interface (see Fig. 6-33a, which shows the presence of the micro-foot on resist features with 0.4 μm lines and spaces).[129] Because image variation and subsequent substrate etch profiles are critical in PBL processes to maintain and control active area encroachment, this foot must be eliminated. This is done by minimizing substrate reflectivity; i.e., by proper choice of the nitride film thickness. The substrate reflectivity is found to be a maximum when the nitride is 2500 Å thick, and a minimum when 2000 Å thick. Thus, use of a 2000 Å nitride film eliminates the micro-foot (as shown in Fig. 6-33b).

6.4.1. Modified PBL Processes

Several *modified* PBL processes have also been proposed. In the first, a thinner field oxide is grown, and the channel-stop implant is performed through this 220-nm-thick field oxide (Fig. 6-34).[51] The 100-keV boron implant that forms the channel stop also serves as the punchthrough implant in the active regions. Since the channel-stop implant is done after the field-oxide growth, much less boron encroachment occurs. In addition, a smaller field-oxide step is produced, thereby helping the planarity requirement. A thinner field-oxide region can be used because the process is designed for

0.7-μm (and smaller) device applications. At these short channel lengths, the doping of the substrate must be increased to prevent punchthrough. As a result, the dopant concentration of the channel stop also ends up being high enough to prevent field inversion, even with the use of such a thin field oxide.

6.4.1.1 Framed-Mask PBL (FMPBL). Another advanced PBL process forms nitride spacers on the pad-oxide-poly-nitride stack after the nitride layer of the stack is patterned in the conventional PBL manner (Fig. 6-35).[52,53] These spacers are formed by depositing a second 1000Å-thick nitride layer and performing an anisotropic etch which stops on the underlying poly layer. Since the nitride spacers frame the masking stack (and also reduce the field isolation width to dimensions smaller than can be achieved lithographically), this process has been named *framed mask PBL* (FMPBL). Field encroachment of FMPBL isolation can be reduced by increasing the width of the nitride sidewall or the thickness of nitride-II, as opposed to increasing thickness of nitride-I (which has the potential for defect generation at high stress). A field-oxide etch-back step can be used to further reduce field-oxide encroachment. FMPBL provides excellent isolation characteristics (e.g., nominal V_{TF} values>10V) down to isolation spacings of 0.8-μm.

A variation of the FMPBL approach called a *reverse L-shaped sealed PBL* (RLS-PBL) technology has recently been reported.[54] It provides comparable bird's beak encroachment to the FMPBL method, but with improved reverse leakage currents of NMOS submicron devices. In this process the pad oxide of the (pad oxide/poly/nitride-1) mask stack is laterally undercut by 300-500 Å using a 15:1 HF etch for several minutes as a part of

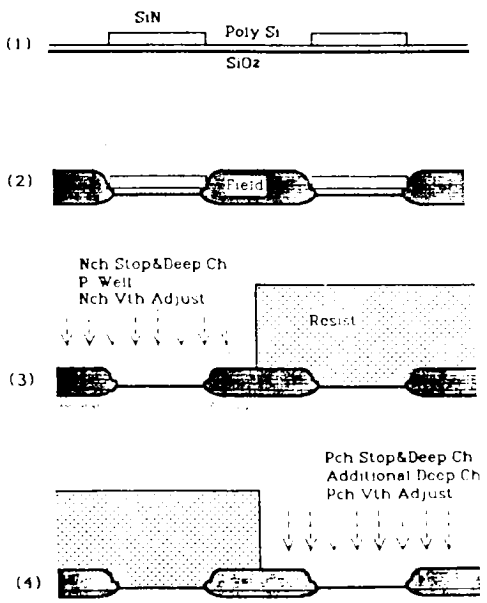


Fig. 6-34 Process flow of Advanced Poly Silicon Pad LOCOS (APPL).[51] (© 1988 IEEE).

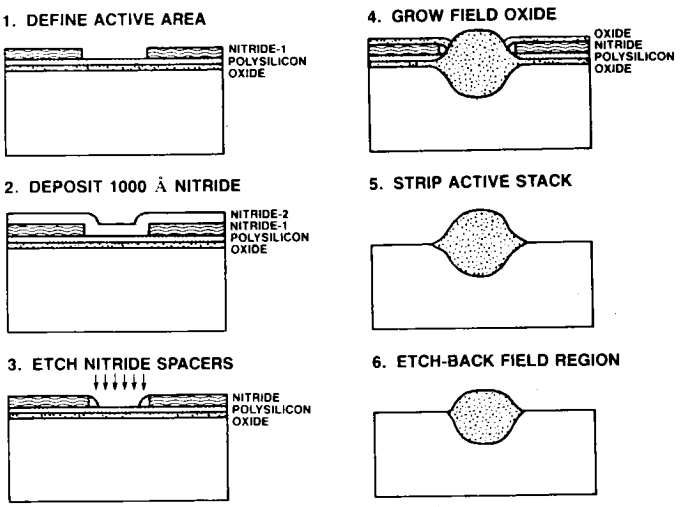


Fig. 6-35 Process flow for the advanced or etched-back FMPBL isolation scheme.[53] (© 1991 IEEE).

the mask stack patterning method (Fig. 6-36). The exposed portions of Si including the undercut portions were reoxidized to grow a very thin oxide as a new buffer layer. Next,

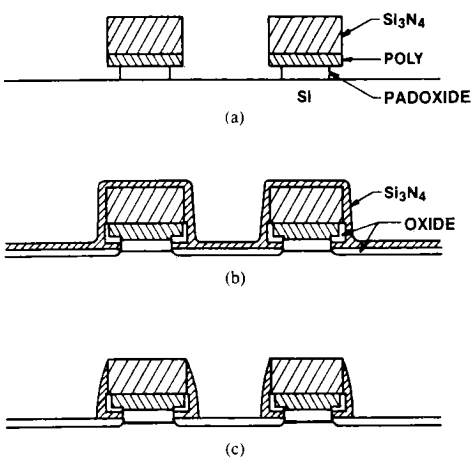


Fig. 6-36 Reverse L-shape sealed PBL process flow: (a) formation of side etched portions of pad oxide; (b) sealed nitride film deposition; and (c) RIE nitride etch. The pad oxide (100Å), polysilicon film (500Å), nitride cap (2000Å), and regrown oxide (60Å) are not shown to scale.[54] (© 1990 IEEE).

a thin (500 Å) LPCVD SiN film is deposited over the whole surface (nitride-2). The conformally deposited SiN layer also fills the etched portions of the pad oxide film (Fig. 6-36b). Finally, nitride spacers are formed by anisotropic etching of the nitride-2 film (Fig. 6-36c, which also shows the L-shaped nature of these interface-sealing spacers). Following a 600-nm field oxide growth virtually no bird's beak is observed. It is conjectured that this approach results in reduced transistor and diode leakage currents because the stresses created in the silicon lattice are absorbed by the polysilicon film which is sandwiched between the nitride and pad oxide via the specially designed L-shaped sealed nitride spacer.

An advanced form of the RLS-PBL technology, reportedly suitable for devices with 0.3 μm dimensions (e.g., 256-Mbit DRAM cells), was recently described, called PBR-LOCOS (Fig. 6-37).[55] During the mask-stack etch, the Si substrate is also anisotropically etched to a depth of 25-200 nm. Then, the exposed sides of the pad oxide (SRO-1) are selectively etched to form lateral undercuts of 30 nm using buffered HF. The exposed Si substrate and polysilicon surfaces are reoxidized to form a new stress-relief oxide (SRO-2), which extends into the lateral undercut areas as well. Next, a thin nitride layer is deposited and etched so that it remains on the sidewalls of the stack and

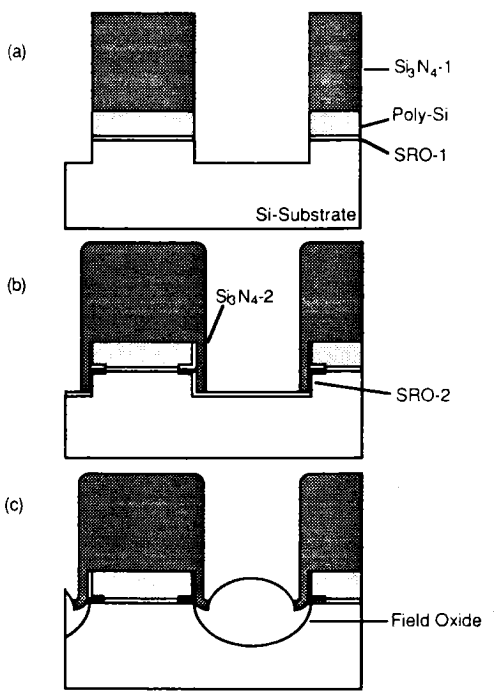


Fig. 6-37 PBR LOCOS process sequence. (a) After recess etching. (b) After nitride sidewall formation. (c) Field oxidation.[55] (© 1992 IEEE).

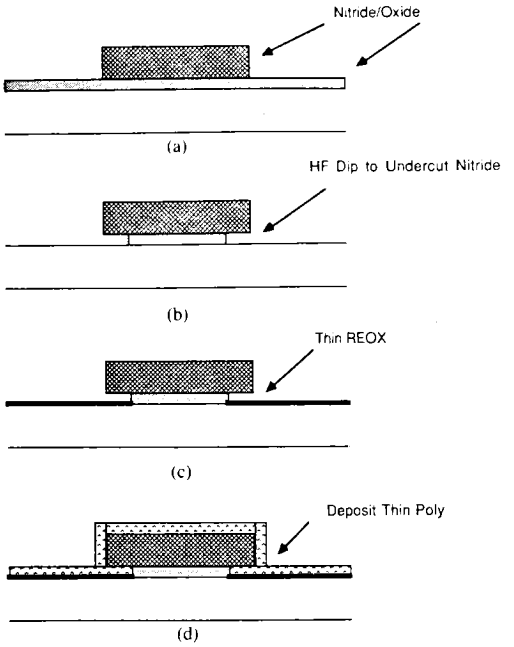


Fig. 6-38 PELOX process flow: (a) RIE nitride; (b) HF dip to undercut nitride and form cavity; (c) grow reoxidation; (d) deposit polysilicon encapsulation layer.[56] (© 1991 IEEE).

the trenched Si areas (SiN-2). Field oxide is then grown. Using PBR-LOCOS, bird's beak encroachment is reportedly suppressed to 0.15 μm for a 50 nm-deep Si trench. Junction leakage is almost identical to that of conventional semirecessed LOCOS and the punchthrough voltage of the parasitic field-oxide transistors exceeds 5V at isolation spacings of 0.3 μm. Furthermore, no anomalous leakage currents are observed in the subthreshold characteristics of MOS transistors formed using this technique, and gate oxide integrity in these devices is also excellent, implying that defect generation in the Si is not significant.

It has been noted that in the RLS-PBL approach the etching process that is used to form the nitride-2 spacers may represent a critical problem when attempting to implement this process. That is, the anisotropic nitride-spacer etch must stop on the nitride-1 layer of the mask stack, and thus there is no selectivity. A secondary concern is that the nitride-2 spacers will worsen the field thinning problem, as they will further narrow the spacing of the field oxide area.[133]

6.4.1.2 Polysilicon-Encapsulated LOCOS (PELOX). Another variation of PBL, in which the poly layer does not extend over the entire pad-oxide/nitride interface is called *polysilicon-encapsulated local oxidation* (PELOX).[56,57] As shown in Fig. 6-38 the poly layer is "inserted" into a cavity between a thin pad-oxide layer (at the

mask stack edges) and the nitride. That is, the initial mask consists of a conventional 50-nm-thick pad oxide and a 140-nm-thick CVD nitride. After patterning the nitride film, the uncovered pad-oxide layer is etched in diluted hydrofluoric acid. During this step, some of the pad oxide under the nitride is undercut, forming a cavity beneath the nitride edges that penetrates on each side to a depth of ~0.12 μm. Next, the exposed portions of the silicon are slightly re-oxidized. When a thin layer of poly (~300Å thick) is next deposited by CVD, it covers all the exposed surfaces, and fills the cavities under the nitride edges created when the original pad was undercut. Unlike the RLS-PBL process, no spacer-formation etch process is needed after the cavities have been filled. Instead, the wafers are transferred directly into the field-oxidation furnace. After the field oxide is grown, the oxidized polysilicon on top of the nitride mask is removed by wet etching, and finally the nitride mask is stripped.

By varying the thickness of the re-oxidized pad-oxide layer (called REOX in Fig. 6-38c), both the height and length of the bird's beak oxide may be minimized. The authors claim that if no re-oxidation is used, less than 0.1 μm bird's beak is observed. With 25 nm of regrown SiO_2 the bird's beak length is ~0.3 μm. Although some reoxidation is required to prevent defect formation in the silicon crystal, about 3 nm seems to be enough for this purpose. A second report combined PELOX with a shallow trench prior to growing the REOX pad layer and depositing the poly. The field oxide is more recessed than in either conventional-PBL or PELOX structures. Almost fully planar isolation with slightly larger encroachment than non-trenched PELOX was achieved with a 140 nm deep trench and 720 nm field oxide.[58] In addition, the small bird's beak in PELOX significantly reduces the degree of bird's beak punchthrough in 0.5 μm-wide active regions.

The enhanced field oxidation substrate recess of PELOX compared to conventional PBL also makes PELOX an attractive candidate for technologies smaller than 0.5 μm. That is, conventional PBL appears to be sufficiently robust for 0.8-0.5 μm technologies, but its susceptibility to field-oxide punchthrough at active spacings of $\leq$0.5 μm limits its ability to scale further.[59] On the other hand, PELOX, or nitride-clad-LOCOS (NCL) have been reported to be attractive candidates for 0.35 μm[132] and 0.25 μm CMOS processes, respectively.[133] The integration of PELOX into an existing PBL 1-MB DRAM baseline process is also described by Roth et al.[60]

6.4.1.3 Nitride-Clad LOCOS (NCL). Despite the fact that PELOX appears to be suitable for 0.35 µm CMOS technology, it has a problem that may prevent from being implemented in the next generation technology (i.e., 0.25 µm CMOS). That is, the oxidized poly mask layer must be stripped by a wet etch after the field oxide is grown. This etch steps thins down the field oxide, and may potentially degrade the isolation in narrow field regions.

To circumvent this difficulty, a modification to the PELOX process has been suggested that may be suitable for 0.25 μm CMOS.[134] That is, instead of using polysilicon to fill the oxide cavity under the nitride mask, very thin (~100Å thick) CVD-nitride-2 layer is deposited instead. Thus, the process has been dubbed *nitride-clad LOCOS* (NCL). Unlike in the RLS-PBL process, however, no nitride spacer-formation

etch step is needed here. Instead, the wafers are immediately subjected to the field oxidation step. The thin nitride-2 layer is oxidized away during the initial portion of the field oxidation cycle. Then, only the relatively short etch step used in the conventional LOCOS process to remove the oxidized nitride is required prior to stripping the nitride (see section 6.2.1.9). Hence, less field oxide is removed in the NCL process, allowing it to be scaled to 0.25 μm technologies.

6.4.1.4 Mitigating the Field-Thinning Problem of LOCOS.

Recently, the field thinning problem has also been reported to be reduced in PBL by growing the field oxide under high pressure (HiPOX).[61,62,63] Field oxides were grown at 8-20 atm and 900-1000°C in steam with isolation spacings of 2.0-0.8 μm. Field oxide thinning as a function of decreasing isolation spacing was significantly reduced especially at 1000°C and 20 atm. Increases in nominal-V_{TF} of 30% were also observed in parasitic field MOSFETs in which $L = 1.0$ μm, compared to those in which the oxide was grown at 1 atm. Part of the V_{TFiso} increase was also attributed to the reduced encroachment of channel-stop boron during HiPOX field oxidation, since less time (at constant T) is needed to grow the same thickness of field oxide under HiPOX. Bird's beak length was also decreased by the HiPOX step.

Another technique for reducing the field thinning problem (and to reduce the encroachment of the field oxide at convex corners, as discussed in section 6.2.1.7) is to etch only the nitride layer of the PBL mask (i.e., leave the poly-buffer layer unetched) prior to the field oxidation growth step.[130] In addition, if a thinner field oxide (e.g., 150 nm thick) can be tolerated, these two effects are reduced in magnitude even further, especially if combined with the nitride-only etch approach.

6.5 SHALLOW TRENCH AND REFILL ISOLATION

Despite the advances made to decrease bird's beak and channel-stop encroachment in LOCOS isolation by PBL and other approaches, it appears that LOCOS-based techniques will eventually become inadequate for deep-submicron MOS technologies. The most probable successor to LOCOS in CMOS is the *shallow trench and refill* isolation structure. Here, a relatively shallow trench is first etched in the Si substrate (0.3-0.5 μm deep), and is then refilled with an insulator material (following a short thermal oxidation growth that is first grown on the trench walls to control the Si-SiO_2 interface quality). The surface is planarized after refill to complete the isolation structure.

Since bird's beak is entirely eliminated and channel-stop dopant redistribution is reduced, smaller isolation spacing is possible than with LOCOS. In addition, the structure is fully recessed, offering the potential of a completely planar surface. Furthermore, the problem of field-oxide thinning in narrow isolation spaces does not occur and the threshold voltage is constant as a function of channel width because of the elimination of the narrow-channel effect found in LOCOS (see chap. 5). Also, MOSFETs made with fully recessed oxide isolation exhibit larger g_m and smaller S_t values than those using semirecessed LOCOS. Finally, fully recessed structures with sharp lower corners provide improved isolation efficiency due to the so-called "corner

effect."[64] That is, the field lines at the bottom corners of the trench spread out at the vicinity of the corner, leading to a higher threshold at the corners. Until the silicon surface at these corners is inverted, isolation is maintained.

Despite the impressive advantages of shallow trench isolation listed above, a variety of new difficulties associated with such structures have slowed their acceptance as replacements for LOCOS isolation. As we shall show, however, significant progress has recently been made toward overcoming these problems.

6.5.1. Early Shallow Trench and Refill Isolation

The earliest example of shallow-trench and refill is the so-called *buried-oxide* (or BOX) isolation technology. The original BOX technique (BOX-1) was introduced in 1981 (see Fig. 6-39i),[113] and an improved version was presented in 1983 (BOX-II, Fig. 6-39ii).[65] In the latter, shallow trenches (0.5-0.8 μm deep) are etched with an orientation-dependent Si etch to provide an inclination of the trench sidewalls of ~60°. Next a channel stop implant step is performed. (The inclination ensures that the sidewalls also receive sufficient doping during the channel stop implant step to prevent weak inversion of the sidewall Si surface. Thus, subthreshold leakage from source to drain along this parasitic conduction path is suppressed.) A thin thermal oxide (20-50 nm) is then grown to passivate the trench sidewalls. Next, CVD oxide is deposited to fill the trenches. This layer is etched back, leaving the oxide only in the recesses. Ideally, the top of the oxide

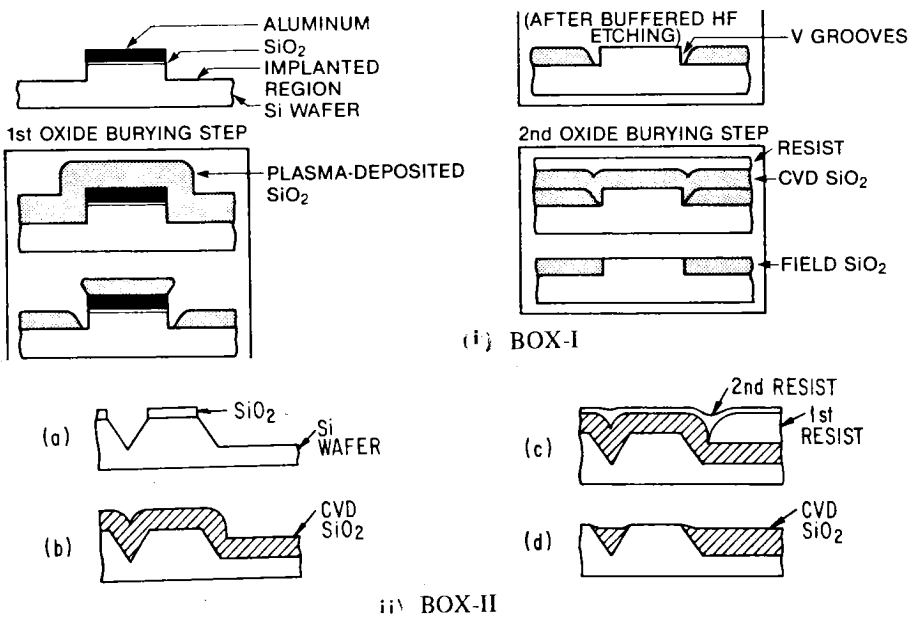


Fig. 6-39 (i) The BOX-I isolation process. (ii) The BOX-II isolation technique. (a) Si is etched using an oxide transfer mask. (b) Oxide is deposited by CVD. (c) Two layers of resist are applied, with the first being flowed. (d) The oxide is planarized using an RIE etchback technique which etches oxide and resist at the same rate.[65] (© 1983 IEEE).

should be level with the original silicon surface. Etchback was performed by eroding photoresist and SiO_2 at the same rate in a plasma (see chap. 4, Vol. 2). In the improved process, two layers of resist were applied, to try to produce a globally planarized surface prior to etchback. Active regions are those protected from etch when the trenches are formed.

These early BOX processes, however, exhibited several drawbacks, including:

1. Void formation occurred if the CVD oxide was deposited into 0.5 μm-wide trenches that are narrower than about 2 μm. However, more recently developed CVD-deposition processes have mitigated this problem.

2. Inversion of the silicon at the sidewalls of *p*-type active regions may occur, especially if vertical trench sidewalls are used to reduce the isolation spacing. This problem is made worse by boron segregation into the thin sidewall-passivating thermal oxide (which lowers its concentration at the sidewall surface, as discussed in following section, which defines isolation in CMOS and gives further details of this problem).

3. Despite complex planarization schemes, the early BOX processes were unable to yield adequately planar surfaces. (In particular, global planarity was poor.) More recently, new planarization approaches have been announced which may overcome this difficulty - e.g., chemical-mechanical polishing. Note that the topography of the trench-surface after etch should be smooth to enable a lower particle size detection limit and to provide a favorable surface for subsequent film nucleation.

4. This isolation structure can negatively impact the device operating behavior.[67,68,69,141] That is, in a conventional LOCOS-isolated MOSFET, a polysilicon gate that passes across the device will curve upward and away from the device once it reaches the isolation oxide, as shown in Fig. 6-40a. (This figure represents the cross section of the region under the gate of the MOSFET, where the source and drain of the device are situated above and beneath Fig. 6-40, which lies in the plane of the paper.) For this particular configuration of the polysilicon gate, the portion of the gate over the LOCOS region will have little effect upon the device behavior. Consequently, a simple 2-D simulation will give an adequate description of the behavior of the device, where the 2-D region is taken in the plane perpendicular to Fig. 6-40 and along the axis of the device, so that it includes the source and drain regions.

In the case of a shallow-trench-isolated MOSFET, the gate region runs flat across the isolation, as shown in Fig. 6-40b. Consequently, the part of the gate over the field region creates more of a two-dimensional fringing field on the corner and sidewall of the silicon mesa than in the LOCOS case of Fig. 6-40a. This establishes a *corner parasitic MOSFET* in parallel with the main device.[140] The parasitic device turns on at gate voltages lower than those required by the main channel, resulting in a "hump" in the drain current versus gate voltage

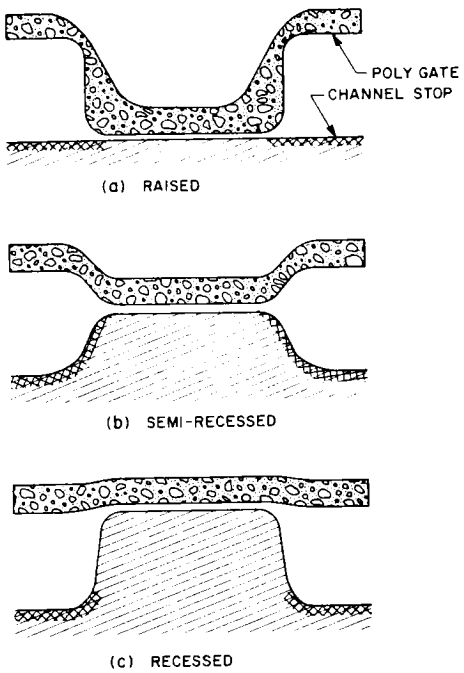


Fig. 6-40 In (a) *grow-oxide-and-cut-window* isolation, and in (b) LOCOS isolation structures, the field oxide rises above substrate. (c) In trench isolation structures, the field oxide is flat with respect to the substrate surface (unless, if the trench oxide is overetched, a downward step exists at the edge of the active regions, see, for example Fig. 6-41).

subthreshold curves (Fig. 6-41). Thus, the parasitic device increases the subthreshold leakage current of the active device.

This effect becomes even stronger if the isolation region is partly etched below the silicon mesa top surface. Then, the gate wraps around the corner of the mesa as shown in Fig. 6-40c. Now the regions of the channel near the edges of the device will behave much differently than the region at the center of the mesa. Consequently, 3-D modeling is required to accurately simulate the device characteristics. Such simulations verify that the corner regions have lower threshold voltage than do the center regions of the device.

5. Thinning of the gate oxide at the trench edge is another concern. Thinned gate oxide not only degrades the gate oxide reliability but also worsens the parasitic device problem, since a thinner gate oxide increases the electric field strength at the trench corner.

6. Problems arose as result of the etchback step. That is, to ensure complete removal of the CVD SiO_2 from the active areas, the field oxide must be

overetched by about 100 nm. The field oxide surface is then lower than active-area surface, causing a portion of the active-area sidewall to become exposed (Fig. 6-40c). This exacerbates the parasitic sidewall conduction described above, as well as to an enhanced inverse narrow-width effect (see chap. 5).

The combination of the above problems (especially the complex processing required to produce adequately planar surfaces), kept BOX isolation from displacing LOCOS-type isolation structures. However, the benefits offered by shallow-trench isolation are so attractive that work has proceeded to overcome the above problems. It appears highly likely that shallow-trench and refill isolation will become more widely adopted as dimensions approach 0.5 μm and smaller. Indeed, an isolation approach for NMOS that combined the best features of both BOX and LOCOS was reported by Fuse et al,[71] as is discussed in more detail in the following section.

6.5.2. Improved Shallow Trench and Refill Isolation Processes

Progress in overcoming the four major problems of early BOX isolation is summarized as follows:

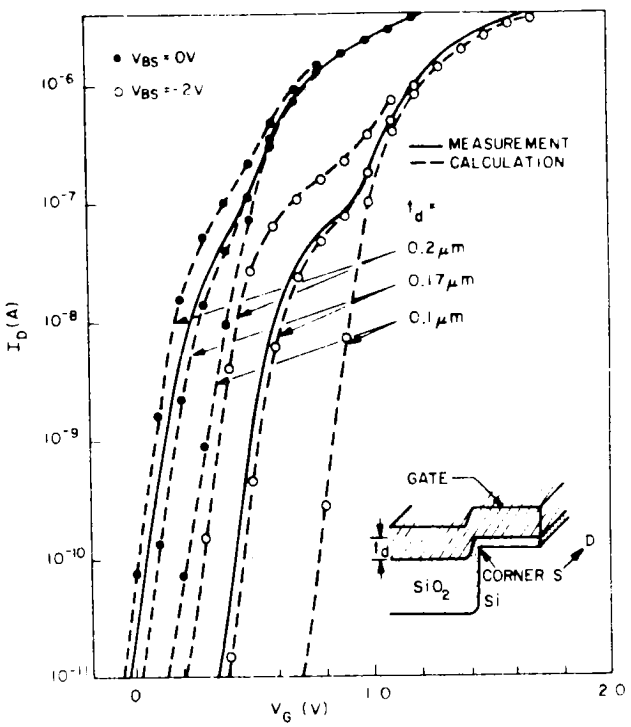


Fig. 6-41 Measured and simulated drain current versus gate potential for devices having a downward step in the field oxide (t_d).[66] (© 1981 IEEE).

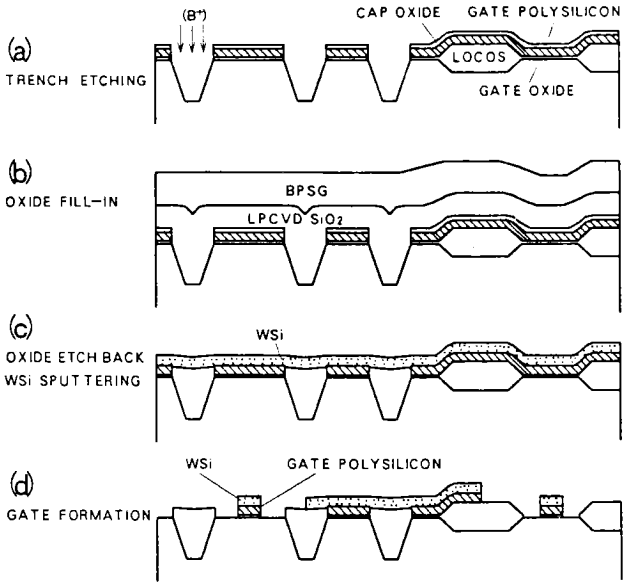


Fig. 6-42 Shallow trench process with tapered trenches.[70] (© IEEE 1992).

1. The problem of void formation when refilling narrow trenches with CVD SiO_2 has been ameliorated through the development of various advanced CVD techniques. Such CVD processes which exhibit less tendency to form void during trench filling include LPCVD TEOS, atmospheric-pressure ozone-TEOS, and various deposition-etchback approaches. A new shallow-trench process (Fig. 6-42) that produces tapered trenches by dry-etching with Cl_2 and N_2 gases (and which uses a cap oxide as a mask) attacks this problem from another direction.[70] Such tapered trenches can be more easily filled without void formation and the deposited dielectric can be planarized with less difficulty. However, the wider trenches have hollow tops (i.e., when the trench width is is greater than twice the depth), resulting in poor isolation characteristics in the resulting parasitic field transistors). Hence, the shallow tapered trench isolation structures are used only for small isolation spaces (<0.6 μm), with wider spaces being isolated using LOCOS structures. Film stress of the trench-fill dielectrics must also be smaller (<200 MPa) than in intermetal dielectric films, due to the use of thicker films.

2. Several approaches have been developed to overcome the inability to adequately planarize the surface through etchback. One involves the use of a block mask, whose function is to allow LOCOS oxide to be grown in wide isolation regions (see chap. 4, Vol. 2 for a discussion of complete global planarization through the use of a block mask). After employing the block mask, a variety of methods involving CVD deposition and etchback are used to planarize the remaining narrow isolation spaces. Many of these methods

followed the early BOX methods of conventional resist etchback, but one novel technique used a thin SiO_2 layer (sputter deposited at low temperature) between double photoresists (the SiO_2 layer acts as an etchback layer) and a polysilicon film above active regions (as an etchback buffer). This process is called *planarization with resist/oxide/resist and polysilicon* (PRORPS).[71] Another technique used LOCOS to fill the wide spaces in the trench and spin-on-glass SOG to fill the narrow spaces. Three or four layers of SOG were spun on to get good planarization.[72] All of the above, however, are complex processes. A relatively new planarization method called chemical-mechanical polishing has been applied to this problem, as is discussed in subsequent paragraphs.

3. As described above, the subthreshold characteristics of narrow MOSFETs (e.g., 0.5 μm wide) isolated by refilled shallow- trench structures are sensitive to the shape of trench upper corner shape.[73] The edges of the silicon mesa at the upper corners are turned on by the portion of the gate that overlaps such corner regions (see also section 5.2.2). This edge effect becomes even stronger if the isolation region is partly eroded away, in which case the gate wraps around the corner of the device as shown in Fig. 6-40c. The optimum condition for the shallow trench isolation structure is thus a gate that runs flat across the isolation, requiring a planar surface prior to poly deposition. A process utilizing RIE etchback followed by a chemical-mechanical polishing step offers the ability to produce a highly planar surface after etchback. The top shape exhibits virtually no step. However, single wafer processing with *in situ* control of the CMP process is likely to be needed to obtain the exact surface required.

4. Inversion of trench sidewalls. This problem and techniques to overcome it are described in section 6.6.

A shallow trench isolation (STI) process incorporating the above improvements is described by Fazan and Mathews.[121] They project that this isolation approach will be feasible for DRAM generations as dense as 4 Gbits (minimum feature size of 0.12 μm). Another trench process that uses trenches with sloped sidewalls and CMP for etchback is described by Deleonibus.[127] Finally, Inowaka describes how a shallow-trench isolation process (with a minimum trench width of 0.28 μm) is applied to a 0.2 μm CMOS technology.[120]

6.6 CMOS ISOLATION TECHNOLOGY

In CMOS ICs isolation must exist not only between *like* kinds of devices (e.g., between two *n*-channel transistors or two *p*-channel transistors within a given well or within the substrate region) but also between devices of *opposite* polarity (e.g., between *p*- and *n*-channel devices separated by at least one well). The isolation of like kind of devices involves the same techniques used to isolate the devices in either NMOS or PMOS circuits (i.e., a combination of a thick field oxide and channel-stop doping). However, the isolation of *p*-channel from *n*-channel devices involves additional considerations,

insofar as two opposite-type FETs must be separated physically by at least one well. That is, in CMOS isolation two additional requirements that arise in neither PMOS or NMOS technology must be simultaneously satisfied: (1) any possible leakage currents that could flow between a *p*-channel device adjacent to an *n*-channel device must be suppressed; and (2) the susceptibility of CMOS to latchup must be minimized. In this section, we describe isolation techniques to suppress the leakage currents between *n*- and *p*-channel devices. Latchup suppression is described in Vol. 2, chap. 6.

Interest in isolating *n*- from *p*-channel devices is very keen because CMOS isolation structures generally consume much more chip area than the isolation structures needed between like types of devices. That is, earlier in this chapter we showed that isolation between *like* kinds of MOS devices can be realized with isolation spacings ≤ 1.0 μm. But in single-well CMOS technologies the minimum required isolation spacing is about three times the depth of the well.[76] Thus, for a 4-μm-deep well, the minimum n^+ to p^+ spacing is ~12 μm. Even in twin-tub CMOS technologies with dual channel-stops, a minimum of 4-9 μm of lateral space is necessary for effective isolation. In advanced submicron CMOS technologies, it has been estimated that the minimum interwell (i.e., n^+-to-p^+) spacing is about three times the minimum intrawell (i.e., n^+-to-n^+ or p^+-to-p^+) spacing.

The large area penalty of *p*-channel-to-*n*-channel device isolation is the most important reason why CMOS technologies using conventional isolation methods cannot achieve as high a packing density as NMOS. Furthermore, while new techniques such as epitaxy greatly reduce latchup susceptibility as CMOS is scaled down, they generally do not suppress leakage currents in the parasitic MOS structures. Hence, the layout spacing between an *n*-channel and a *p*-channel device may be limited by isolation failure rather than by latchup.

In this section we describe why greater isolation spacing is needed in CMOS than in NMOS or PMOS, and how the minimum spacing for a given CMOS technology is determined. We begin by considering CMOS isolation from a qualitative perspective. Next, we give a description of how CMOS isolation is quantitatively modeled. Then, test structures for empirically assessing the adequacy of isolation among opposite-type FETs in a CMOS technology are described. Procedures for establishing the design rules for minimum CMOS-isolation-spacing from the simulation results and measured data are then presented. Finally, technological approaches for fabricating isolation structures in CMOS are discussed.

6.6.1 Qualitative Description of Isolation in CMOS

In CMOS structures, the isolation spacing between the *n*- and *p*-channel devices is defined as the sum of: (1) the distance between the edge of the n^+ region of the *n*-channel device and the edge of the well; and (2) the distance between the edge of the p^+ region of the *p*-channel device and the edge of the well (or, in other words, the n^+ to p^+ spacing, as shown in Fig. 6-43).[74] As noted in the introduction, the minimum isolation spacing in CMOS is much larger than that required in NMOS or PMOS technologies. Here we explore the reasons why this is the case.

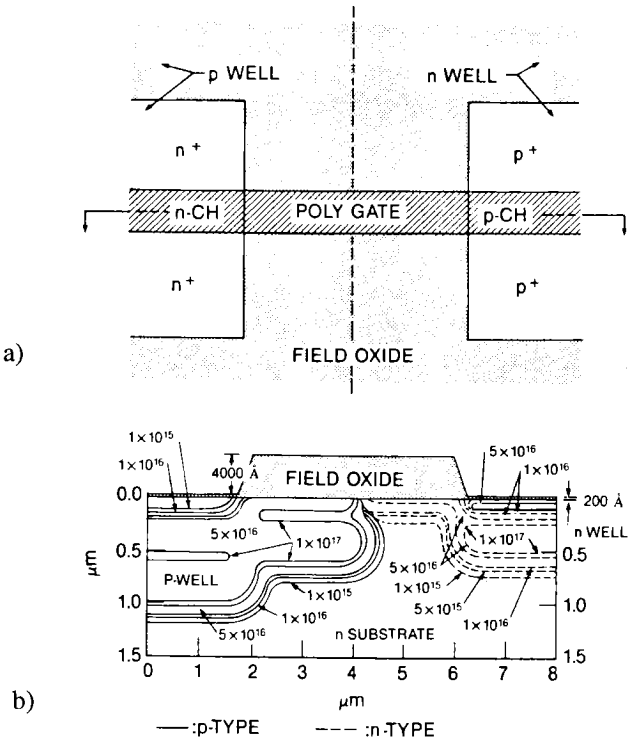


Fig. 6-43 (a) Layout top view of the isolation region between *n*-type and *p*-type transistors in CMOS. (b) Cross sectional view of a CMOS inverter showing the channel stops that are designed to prevent the surface under the field oxide from inverting.[74] (© 1986 IEEE).

An overall grasp of the issues of isolation in CMOS can be obtained by analyzing the various causes of leakage current in a CMOS inverter. The circuit schematic, layout, and a device cross-section of an *n*-well CMOS inverter is shown in Figs. 6-44a, b, and c, respectively. The layout actually shows two CMOS inverters connected in series, allowing the figure to be used to illustrate intra-well as well as inter-well isolation issues. In addition, the cross-section depicts an inverter with a different layout than the one shown in part (b) to enable us to show the source, drain, substrate and well contacts. Finally, note that *n*-well CMOS is chosen as the example CMOS architecture for this discussion, but the concepts are also equally valid for *p*-well and twin-well CMOS.

If a normally operating CMOS inverter is in one of its two logic states (i.e., it is not in transition between states), only a negligibly small leakage current I_{DD} should exist between V_{DD} and V_{SS}. To ensure that the total I_{DD} is within acceptably small limits, all of the sources of leakage between V_{DD} and V_{SS} need to be identified and characterized. Having such information in hand permits appropriate measures to be taken to assure that none of the leakage current components will exceed the maximum allowable I_{DD} value. From this perspective, we can identify the sources of such leakage by dividing them up into two groups: (1) those which arise as a result of leakage

phenomena within active devices or from isolation breakdown between *like* types of devices, and; (2) those which occur as a result of isolation failure in CMOS structures (e.g., loss of isolation between devices in the well and in the substrate).

6.6.1.1 Leakage Currents in CMOS Gates Due to Active Device Phenomena and Intra-Well Isolation Loss. The leakage-current phenomena associated with active devices and intra-well isolation structures have already been discussed in depth in chaps. 4 and 5 and in earlier sections of chap. 6. Therefore, they are considered here only in the context of their contribution to leakage in CMOS structures.

Recall first that if a CMOS inverter is in one of its two logic states, one of the two complimentary MOS devices of the inverter will be *OFF* (see Fig. 6-44a). In the *OFF* state there are two components of leakage current that the device itself can contribute to I_{DD}, namely the subthreshold drain current I_{Dst} (arising from either conventional subthreshold surface current or subsurface lateral punchthrough current) and the reverse-biased drain-substrate junction leakage current. By proper circuit and device design these leakage components can be kept adequately small under normal operating conditions.

In addition, leakage between the drain of one device and the source of neighboring devices within the same well (or substrate) constitutes another possible component of I_{DD} (see Fig. 6-44b). Such leakage can be suppressed by the techniques used to isolate two *n*-channel (or *p*-channel) devices from one another, as described earlier in the chapter. In most cases, leakage current due to inadequate isolation among intra-well devices is due to surface inversion effects which permit excessive subthreshold drain current to flow in the intra-well parasitic field transistors, as described in section 6.1.

6.6.1.2 Leakage Currents in CMOS Gates Due to Loss of Isolation Among Opposite-Type FETs. Basic isolation among opposite-type FETs in CMOS involves several considerations. The first and most obvious involves the suppression of leakage across the *pn* junction which isolates the well from the substrate. To keep the component of I_{DD} due to this effect negligible, the well-substrate junction must always be kept under reverse bias. To ensure that this requirement is met (under normal operating conditions) the *n*-well (or *n*-substrate) is connected to the most positive circuit voltage (i.e., V_{DD}), and the *p*-well (or *p* substrate) to the most negative circuit voltage (i.e., V_{SS}), as shown in Fig. 6-44c for *p*-well CMOS.

Other components of I_{DD} in CMOS, however, can also arise from the loss of isolation involving leakage in the parasitic FETs created by the proximity of the active device source/drain regions and the well edges. There are two of these parasitic FETs formed at the edge of the *n*-well - one being a parasitic NMOSFET and the other a parasitic PMOSFET (as shown in Fig. 6-45). Notice that the n^+ and *n*-well (and the p^+ and *p*-substrate), form the source and drain for such parasitic NMOSFETs (and PMOSFETs), respectively. The *p*-substrate and *n*-well are the channels of these parasitic NMOS and PMOS FETs respectively, and the field oxide and poly (or Al) runners over the field oxide form the gate oxide and gates of these parasitic structures. However, it is important to note that unlike in active MOSFETs, the source and drain in these parasitic FETs are not symmetrical. That is, the substrate and well are deep junctions

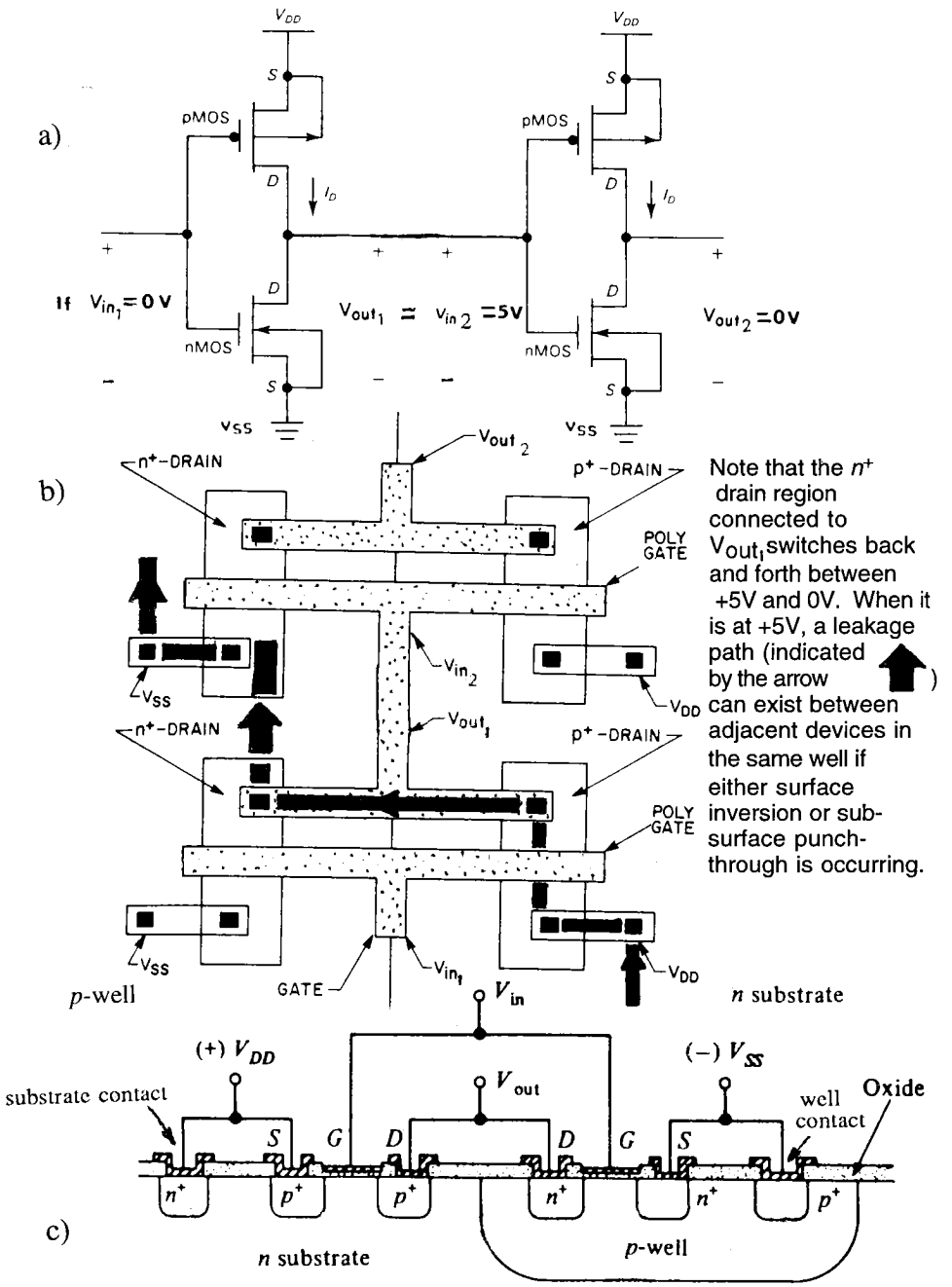


Fig. 6-44 (a) Circuit schematic of two CMOS inverters connected in series. (b) Layout of two *p*-well CMOS inverters connected in series. (c) Cross-section of a CMOS inverter, showing also the well and substrate contacts.

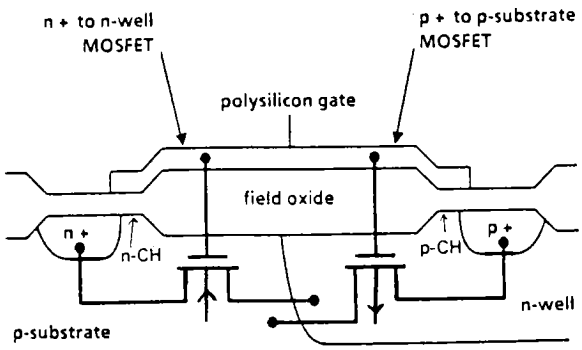


Fig. 6-45 Parasitic MOS transistors formed at the well edge in *n*-well CMOS.[91] (© IEEE 1987).

and the n^+ and p^+ regions are shallow junctions. Because a deeper drain junction and a highly conductive source always produce more subthreshold current (thereby representing a worst case condition for isolation), the *n*-well and *p*-substrate are designated as the drains when characterizing the behavior in these parasitic *asymmetric* FET structures.

Here we identify the phenomena in these parasitic FETs that can give rise to leakage current components in CMOS (again using *n*-well CMOS as the example CMOS architecture). Later we will describe techniques for mitigating these leakage currents. First, we point out that while the *n*-well (or *n*-substrate) is connected to V_{DD}, the n^+ source or drain regions can be biased at V_{SS}. (As an example, the n^+ source of the NMOS device in the CMOS inverter is connected to V_{SS}.) Thus, if a leakage path exists between the *n*-well and the n^+ source, the potential difference between the *n*-substrate and the n^+ regions can cause a component of I_{DD} to flow between V_{DD} and V_{SS}. The same problem arises between the p^+ source regions in the well and the *p*-substrate. It is also important to note that each of the parasitic FETs in which such leakage paths can arise act independently. Thus, when a CMOS technology is being developed, all of the parasitic FETs must be must be designed correctly to prevent any one of them from drawing a significant I_{DD}.

The leakage paths which can be established in the parasitic FETs of an *n*-well CMOS structure are the following:

1. Inversion of the channel region of the n^+-to-*n*-well parasitic FET.
2. Inversion of the channel region of the p^+-to-*p*-substrate parasitic FET.
3. Lateral subsurface punchthrough of the n^+-to-*n*-well parasitic FET.
4. Lateral subsurface punchthrough of the p^+-to-*p*-substrate parasitic FET.
5. Vertical punchthrough of the p^+-to-*p*-substrate parasitic FET.

Let us consider the *n*-well CMOS structure shown in Fig. 6-46. If the polysilicon gate is biased at $V_{DD} = 5V$, the channel region of the active NMOS device under the

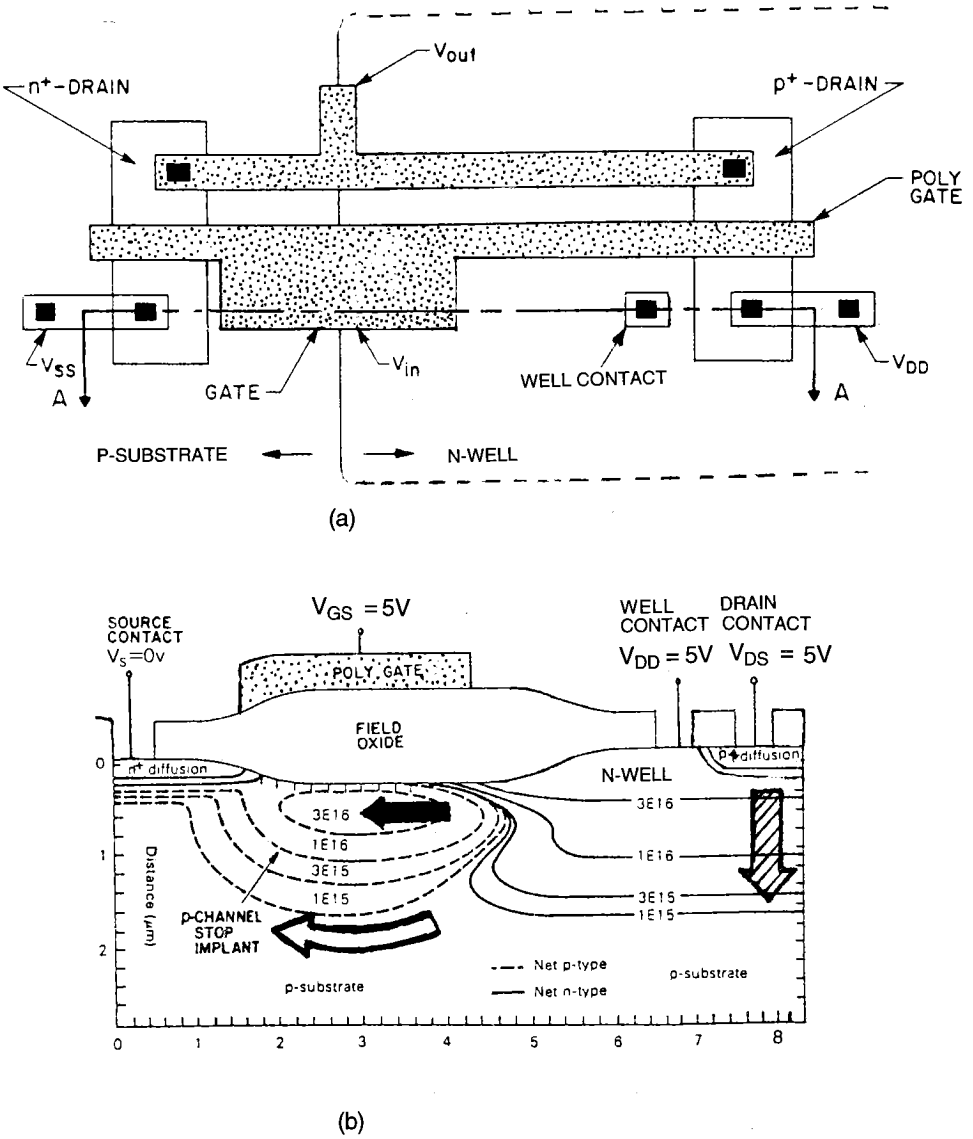


Fig. 6-46 (a) Layout of an *n*-well CMOS inverter; (b) Cross-section of this inverter taken through section A-A showing various possible leakage paths, including: ⬅ - lateral surface leakage path from n^+-to-*n*-well; ⇦ - lateral subsurface (punchthrough) leakage path from n^+-to-*n*-well; and ⇩ - vertical punchthrough path from p^+ (source)-to-*p*-substrate.

polysilicon gate is inverted to n-type. However, as long as the field region under the polysilicon in the p-substrate is not inverted, the active NMOS device remains isolated from the n-well by the p-substrate. However, if V_{TFiso} of the parasitic NMOSFET is not high enough, the surface of the p-substrate can invert under the field oxide, creating a path for current flow from the n^+ source to the n-well, causing the isolation provided by the well to be lost (path #1 above, and shown as in Fig. 6-46b). The subsurface lateral punchthrough path in this device is also shown in Fig. 6-46b as . This path can arise if the isolation spacing is decreased to a point where the punchthrough voltage of this parasitic FET becomes smaller than V_{DD} (path # 3 above). Vertical punchthrough from the p^+ source through the n-well to the p-substrate could also cause leakage in the parasitic PMOSFET (see also 6-46b). Finally, in n-well CMOS, leakage can also be due to surface-inversion and lateral-subsurface (punchthrough) phenomena in the p^+-to-p-substrate parasitic FET (paths #2 and #4 above).

6.6.1.3 Why n^+-to-p^+ Isolation Spacing is Greater than n^+-to-n^+ Spacing in NMOS (or p^+-to-p^+ Spacing in PMOS). Having identified the possible leakage mechanisms arising from the two parasitic FETs formed at the edge of the well, we can go on to explain why larger n^+-to-p^+ spacing must be used to provide adequate isolation in CMOS. There are several reasons which underlie the need for larger n^+-to-p^+ spacing:

1. The technology of conventional well formation leads to lower concentrations of well doping at the Si surface near the well borders (i.e., due to lateral redistribution and interdiffusion of well dopants during the drive-in step that forms the well). Thus, if the p^+-to-p-well spacing gets too small, the maximum n-substrate concentration in the channel of the parasitic FET will eventually become too low to sustain an adequately large value of V_{TFiso}. Hence, leakage will occur at the surface of the channel of the parasitic FET.

2. Even if the doping concentration at the surface of the channels of the parasitic FETs is large enough to sustain an adequately large long-channel V_{TFiso}, subthreshold-leakage-currents between opposite type FETs in submicron CMOS may still become excessive as the isolation spacing is reduced below some minimum dimension. Three additional effects that can give rise to such increased subthreshold leakage components of I_{DD} in submicron CMOS are the following:

a. Short-channel V_T rolloff and DIBL in the parasitic FETs;

b. Short-channel and channel doping-profile effects that can increase the susceptibility to lateral subsurface punchthrough in the parasitic FETs;

c. Decreased lateral n^+-to-p^+ spacing (which leads to shallower well depths) can give rise to vertical punchthrough (between the source/drain junction in the well and the substrate). CMOS structures fabricated on silicon substrates with thin epitaxial layers are also more prone to vertical punchthrough.

Here we discuss dopant redistribution at the well borders and short-channel and lateral subsurface punchthrough effects in the parasitic FETs. Vertical punchthrough is discussed in section 6.6.5.

6.6.1.4 Effects on n^+-to-p^+ Spacing of Dopant Redistribution at the Well Borders. The need for large isolation-spacing dimensions in CMOS is due in part to the processes used to fabricate CMOS devices. Wells are typically driven in quite deeply to: (1) ensure that enough charge exists below the transistor to prevent vertical punchthrough to the substrate, and; (2) to reduce the gain of the vertical parasitic bipolar transistor (decreasing the gain of parasitic bipolars reduces latchup susceptibility). The deep well drive-in, however, also results in both substantial lateral diffusion of the well dopants (which increases the minimum distance from the p^+-region and the well border) and a reduction in the surface concentration near the border of the n-well. The interdiffusion (i.e., compensation) of the tub and substrate impurities in single-well CMOS (or the two tub impurities in twin-well CMOS) further reduces the net surface concentration near the border, making the threshold voltage of the parasitic MOS field transistors even smaller.

The effects of dopant redistribution and interdiffusion at the well border in single p-well and twin-well CMOS processes are illustrated in Fig. 6-47. A 2-D simulation of

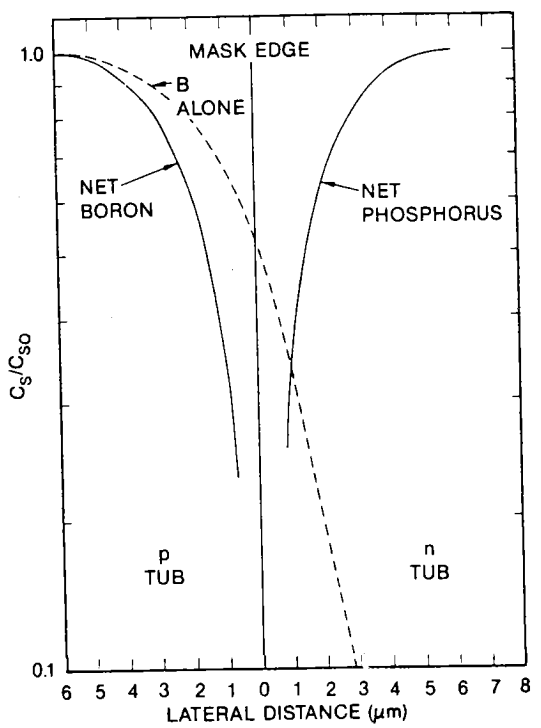


Fig. 6-47 Impurity surface concentration near the border of the two tubs in twin-tub CMOS.[77] (© 1980 IEEE).

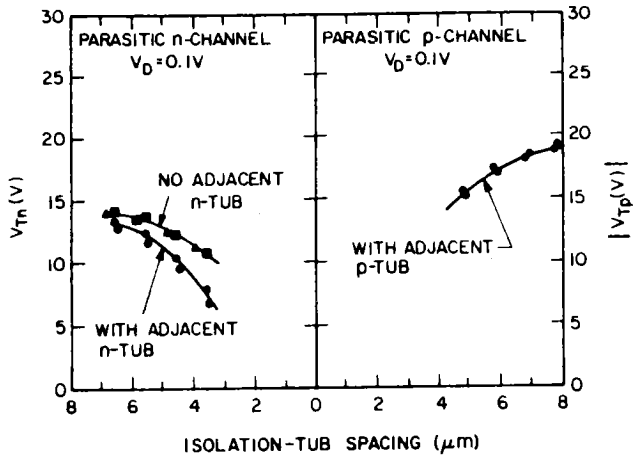


Fig. 6-48 Threshold voltage of n- and p-channel parasitic device as a function of the separation between the transistor edge and the tub border.[77] The upper curve in the left-hand graph shows the parasitic n-channel threshold-voltage reduction near the tub border when no adjacent n-tub is present. When an n-well is present, the threshold voltage is reduced (lower curve), since the interdiffusion of the two types of impurities reduces the net surface concentration of each dopant near the well border.[77] (© 1980 IEEE).

the surface doping concentration near the well border as a result of the well drive in is shown. (Note that this calculation assumes no extra channel-stop implants are used.) We see that near the border the net impurity concentrations at the surface are significantly decreased. The dashed line in Fig. 6-47 shows the reduction in net impurity concentration at the well boundary for the case in which only a boron well implant is used (single p-well). The solid curves represent the case when both boron and phosphorus well implants are used (twin well). We see that the slope of the curves in twin-well CMOS (Fig. 6-47) is greater than in conventional single-well CMOS. But, since the doping change occurs on both sides of the junction compared to a single side for the single-well case, these two effects offset each other. As a result, both technologies require comparable minimum n^+-to-p^+ spacings across the well. However, an advantage of the twin-well is that the well border is located close to its drawn position, while in the single-well case the well border is displaced by lateral diffusion of the well dopants. Thus, in single-well CMOS technologies, the circuit layout must be compensated for this displacement, either by design rules or by mask biasing. In twin-well CMOS processes like those of Fig. 6-47, this problem does not occur.

Figure 6-48 illustrates the effect on V_{TFiso} of the parasitic field transistors when the surface doping concentration near the well borders is reduced. Here, the experimental V_{TFiso} values of such structures (fabricated with the well processes that yield the well-border surface doping profiles of Fig. 6-47), are plotted as a function of the separation

between the transistor edge and the well border.[77] The upper curve in the left-hand graph shows the parasitic n-channel threshold-voltage reduction near the tub border when no adjacent n-tub is present. Note that when an n-well is also present (twin-well CMOS), the threshold voltage is reduced even more than in the single-well case (lower curve), since the interdiffusion of the two types of impurities decreases the net surface concentration of each dopant near the well border. Unless the net dopant-reduction effects are somehow counteracted, the spacing between adjacent n^+ and p^+ devices must be kept quite large (e.g.,>10 μm) to prevent inversion beneath the field oxide.

If smaller spacings between n- and p-channel devices are to be possible, the channel-stop doping concentration must somehow be increased, particularly in the p-substrate regions of n-well CMOS. During field oxidation, boron segregates into the oxide, while phosphorus piles up at the silicon surface. As a result, in an n-well process a separate p-type channel-stop implant must be added to increase the surface concentration of the lightly doped p-substrate. Without such an implant, inversion between the n-channel devices in the substrate is likely to occur. In p-well technologies the well itself is an adequate n-type channel stop because heavier boron doping exists in the p-well. In the n-substrate of p-well CMOS the concentration of phosphorus is increased at the surface of the n-substrate during field oxide growth so that a separate substrate channel-stop implant may not be necessary.

Because of increased interdiffusion, two channel stops are needed in twin-tub processes to reduce the isolation distance as much as possible. In one recently described approach, an increase in the abruptness of the doping transition between the two wells is realized with two channel stops as follows. A double implant is used to simultaneously introduce the n-well dopant and n-well channel stop dopants. That is, a shallow arsenic implant is performed together with a phosphorus implant into the n-well regions. This increases the net n doping near the well edge because the smaller diffusivity of arsenic keeps it close to its original location during the drive in that diffuses the phosphorus deeply to form the n-well. In addition, a second shallow boron implant is done into the p-well following the well drive in. This increases the boron surface concentration near the well edge and thus compensates for the loss of boron due to lateral diffusion and for the diffusion of phosphorus into the p-well near the well border. This process is described in more detail in section 6.6.6.

6.6.1.5 Threshold Voltage Rolloff, DIBL, and Lateral Punchthrough in Submicron CMOS Parasitic Field FETs. As n^+-to-n-well or p^+-to-p-substrate spacing is reduced, the magnitude of V_{TFiso} for either NMOS or PMOS parasitic FETs decreases for the same reasons that V_T decreases in active short-channel MOSFETs (or that V_{TFiso} decreases in parasitic field MOSFETs between like devices) as the effective channel length is reduced (see chap. 5 and section 6.1). V_{TFiso} also decreases as the drain bias is increased (DIBL), just as is observed in short-channel active MOSFETs. Lateral punchthrough from drain (n-well) to source (n^+ region) is more likely to occur in the inter-well parasitic FETs than in the intra-well parasitic FETs because of the deep drain in the former devices. Moreover, the two failure modes (field inversion and punchthrough), are coupled through 2-D effects.[74] For worst-case

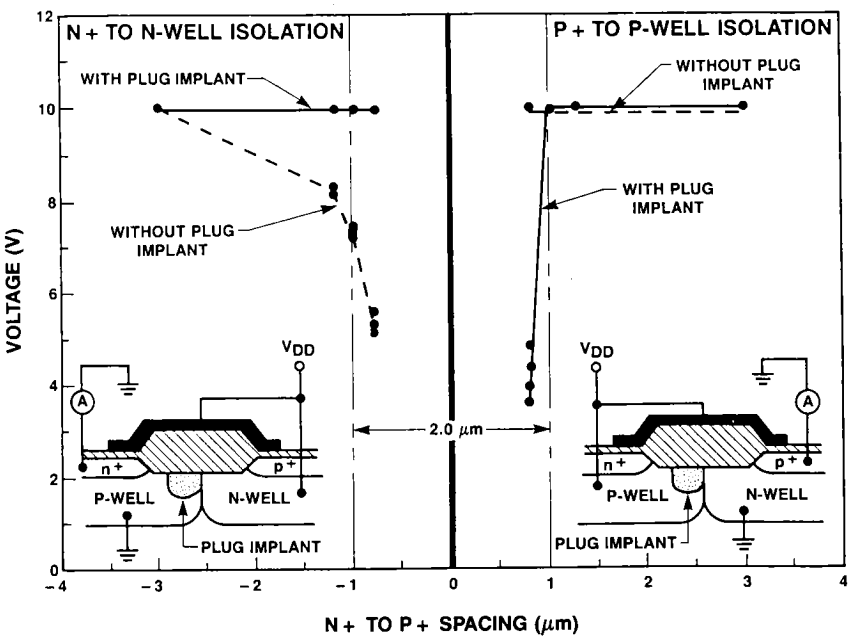


Fig. 6-49 Inter-well isolation voltage as a function of n^+ to p^+ spacing, demonstrating that proper implant selection can improve n-$^+$ to n-well leakage without significantly degrading p^+ to p-well isolation. Interwell isolation voltage is defined as the voltage required for 1 pA/μm of interwell leakage. The n^+ to n-well isolation is preserved on the left and p^+ to p-well on the right. The test setup is depicted in the insets at the bottom of the figure.[84] (© 1991 IEEE).

considerations, field inversion must be experimentally determined with the drain bias set at its maximum value (e.g., V_{DD} + 10%). Punchthrough must be experimentally determined with the gate voltage set at its maximum operating value. Fig. 6-49, shows the results using these test structures to monitor the "plug"-implant CMOS isolation structure discussed in section 6.6.7.[84] Here the interwell isolation voltage V_{TFiso} is defined as the gate voltage at 1 pA/μm of interwell leakage.

6.6.2 Quantitative Modeling of CMOS Isolation

A qualitative description of the issues pertinent to CMOS isolation, as given in section 6.6.1, is useful for gaining an intuitive understanding of the overall problem. However, in the process of developing a new CMOS generation (or modifying an older one), a *quantitative* characterization of the isolation behavior of the parasitic FETs near the well borders is needed. As will be described in section 6.6.3, such characterization can be experimentally obtained from appropriate test structures and test procedures. That is, if a set of test structures that represent the parasitic well-edge FETs in a particular CMOS technology is available, their experimental V_{TFiso} and punchthrough voltage values can

be obtained for each channel length dimension of interest. From plots of such experimental values it is easy to derive the minimum distance for safe operation at the supply voltage, usually V_{DD} plus 10% safety margin.

On the other hand, when a new CMOS process is being developed, a minimum n^+-to-p^+ spacing may be chosen as one of the design rules by the circuit designers. In this case, one would like to be certain early in the design cycle (namely, even before any circuits or device structures have actually been fabricated), that the parasitic well-edge FET structures of the new process will provide adequate isolation, thus allowing this isolation spacing design rule to be adopted for circuit design layout purposes. The experimental test data, in this case, would not yet be available. In fact, if the parasitic FETs could be properly designed prior to fabrication, the data from the test structures would merely serve to confirm the adequacy of the isolation characteristics as predicted from the simulation.

However, the isolation behavior of a well-edge parasitic FET with a given channel length can vary, depending on the device structure and fabrication techniques. In fact, the interdependence of doping profiles, thermal cycles, and device dimensions makes optimization of these parameters during process design a challenging task. To avoid many experimental runs for finding the optimal well-edge FET device configuration, quantitatively modeling of the isolation behavior of these structures is necessary. Such modeling early in the design cycle can also be used to determine whether or not adequate isolation is possible at the chosen minimum n^+-to-p^+ spacing, and if not, what the minimum achievable isolation spacing with the available process capabilities would be.

Such modeling of isolation behavior involves simulation of the subthreshold device characteristics of the parasitic FETs of interest. However, since the device structures of these parasitic FETs are quite different from those of conventional active MOSFETs, the simple analytical models of subthreshold behavior derived in chaps. 4 and 5 for such simulations are inadequate. That is, in submicron CMOS the parasitic devices are asymmetric FETs with deep drains, thick gate oxides, and short channel lengths. Furthermore, the doping in the channel regions is non-uniform in both the lateral and vertical dimensions. Hence, to quantitatively simulate the electrostatic potential and carrier concentrations as a function of position in the channel regions when the device is operated under subthreshold dc bias conditions, the Poisson equation and the current equations must be solved in two-dimensions. This can only be accomplished with two-dimensional process and device simulators, as described, in chaps. 1 and 2. Specific examples of simulating the parasitic well-edge FETs in CMOS are described in refs. 74 and 91. Here we summarize their results.

Basically, the first step in such quantitative modeling of CMOS isolation is to use a two-dimensional process simulator (such as SUPREM IV) to calculate the isoconcentration impurity contours of both well-edge FET isolation structures. Figure 6-50a shows the configuration of the test structure used to simulate n^+-to-n-well well-edge FET of an n-well CMOS technology, in which the doping concentration as a function of position has been plotted using a 2-D process simulator. By using the doping concentration results from this simulation procedure as an input to a two-dimensional device analyzer (such as GEMINI, MINIMOS, PISCES, or MEDICI), together with the

appropriate bias conditions, the potential contours in the device can be obtained. An example of this type of simulation for the same device structure is shown in Fig. 6-50b.[91]

We noted earlier that the two mechanisms of isolation loss (leakage) in the parasitic well-edge FETs are inversion of the channel in the field region and punchthrough of the source and drain of the parasitic device. For analyzing the *onset of inversion* in the parasitic FETs it is useful to plot the surface potential versus lateral distance under the isolation field-oxide at various V_{GS} values for a given n^+-to-p^+ spacing. An example of this type of plot is shown in Fig. 6-52c. We show how this plot is derived with the aid of Fig. 6-51. That is, in Fig. 6-51a1 we show the structure of the n^+-to-n-well parasitic FET. Beneath it (a2 to a5), the surface potential $\varphi_{surf}(y)$ is plotted for various V_{GS} values (e.g., 0, 15, 20 and 25V, respectively). We see that the minimum surface potential exceeds 0.51 mV when V_{GS} is between 20 and 25V (~21.4V). In Fig. 6-51b we show the same $\varphi_{surf}(y)$ plots for the p^+-to-p-well FET (b1). Here the maximum surface potential drops below -51 mV when V_{GS} is between -5 (b4) and -10V (b5), i.e., at ~ -7V.

In Fig. 6-52a we show the layout and in Fig. 6-52b the cross-section and 2-D net doping contours of these two parasitic FETs together. This particular structure represents a configuration that would be found in a *quadruple-well CMOS technology* (see chap. 8). In Fig. 6-52c the variation of the surface potential versus lateral distance is plotted for various V_{GS} values for the structure of Fig. 6-52b. (Note that this structure has a 2 μm n^+-to-p^+ spacing, and that the field oxide overlaps the p-well on the left-hand side, and the n-well on the right-hand side.) The minimum surface potential φ_{surf} in the p-well becomes 0V when V_{GS} = 20V, and reaches 0.22V at V_{GS} = 25V, indicating that the surface is becoming inverted as V_{GS} exceeds about 20V. The minimum value of φ_{surf} becomes 0V in the n-well when V_{GS} reaches about -7V.

The following procedure can be used to specify the value of V_{TFiso} more precisely

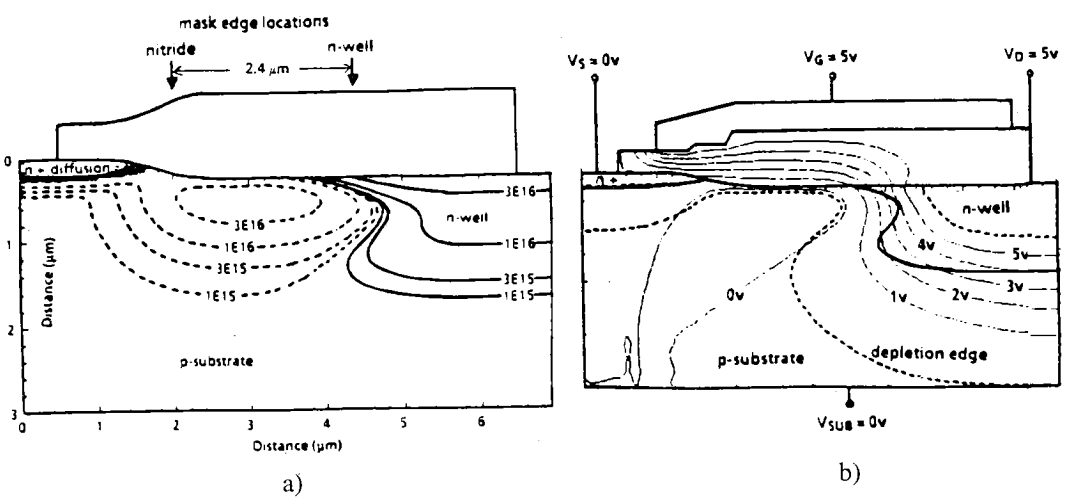


Fig. 6-50 2-D simulated: (a) doping contours; and (b) potential contours in the n^+-to-n-well parasitic MOSFET.[91] (© 1987 IEEE).

First, a specific carrier density is selected as an indicator of the onset of field inversion (e.g., $1x10^{11}$ cm^{-3}). Then, the value of φ_{surf} needed to establish this carrier concentration is found. Next, the value of V_{GS} necessary to produce this value of φ_{surf} is calculated, and this V_{GS} value represents the V_{TFiso} for the parasitic FET. The value of φ_{surf} can also be used to calculate the carrier density in the surface channel (and thus also at the V_{GS} necessary to establish this φ_{surf} value).

As an example, in Fig. 6-53 the minimum minority carrier concentration values in the parasitic channel versus V_{GS} are plotted for the structure depicted in Fig. 6-52b. If field inversion is defined as occurring when the minority carrier density exceeds ~$1x10^{11}$ cm^{-3} (corresponding to a surface potential of 51 mV at 25°C), then the simulated value of V_{TFiso} of a 1 μm n^+-to-n-well structure as shown on the left hand side of Fig. 6-52b is 21.4V. Likewise, a 1-μm p^+-to-p-well parasitic FET will have a corresponding V_{TFiso} = -7.0V. Thus, a safe V_{GS} operational range for this CMOS technology with a 1 μm n^+-to-n-well spacing and a 1 μm p^+-to-p-well spacing (i.e., 2-μm n^+-to-p^+ spacing) is $-7.0V<V_{GS}<21.4V$.

The V_{TFiso} calculated for, say the n^+-to-n-well FET in the manner described above, can be found for various channel lengths (i.e., n^+-to-n-well spacings), and these simulated V_{TFiso} values can then be plotted versus this distance (as is done in Fig. 6-57a). The same can be done for the p^+-to-p-well FETs, as shown in Fig. 6-59. From these plots one can find the minimum n^+-to-p^+ spacing that can be used and still have adequate protection against loss of isolation due to field inversion in the parasitic FETs.

For simulating *lateral punchthrough*, the voltage applied to the well junction and to the neighboring source/drain regions must be considered in the simulations. Furthermore, the effect of the gate voltage applied to the field parasitic MOSFETs must also be included, since it can indirectly affect punchthrough by controlling the underlying surface potential and carrier density. An example of these simulations is given in Fig. 6-54, which illustrates 2-D equipotential contours for 2 μm isolation with 1.5 μm p^+-to-p-well spacing and 0.5 μm n^+-to-n-well spacing.[74] The n-well is tied to the supply voltage (5V), while the p-well and the substrate are tied to ground. The graph at the top shows good electrical isolation with a gate voltage of 10V due to the undepleted portion of the p-well surface, which is characterized by zero surface potential. However, if the gate voltage is raised to 20V, then punchthrough occurs from n^+ to n-well, because the p-well surface becomes fully depleted or inverted, as demonstrated by the fact that it is at positive potential. By analyzing a matrix of bias conditions and spacings, the design of the parasitic well-edge FETs can be optimized even before a single silicon wafer lot is processed, providing a great savings of development costs and time.

6.6.3 Experimental Characterization of Isolation in CMOS

The cross-sections of two test structures for measuring the degree of leakage in the diffusion-region-to-well-edge parasitic field MOSFETs in CMOS are shown in Fig. 6-55.[91] (Note that an n-well CMOS technology is again used here, but the same kind of

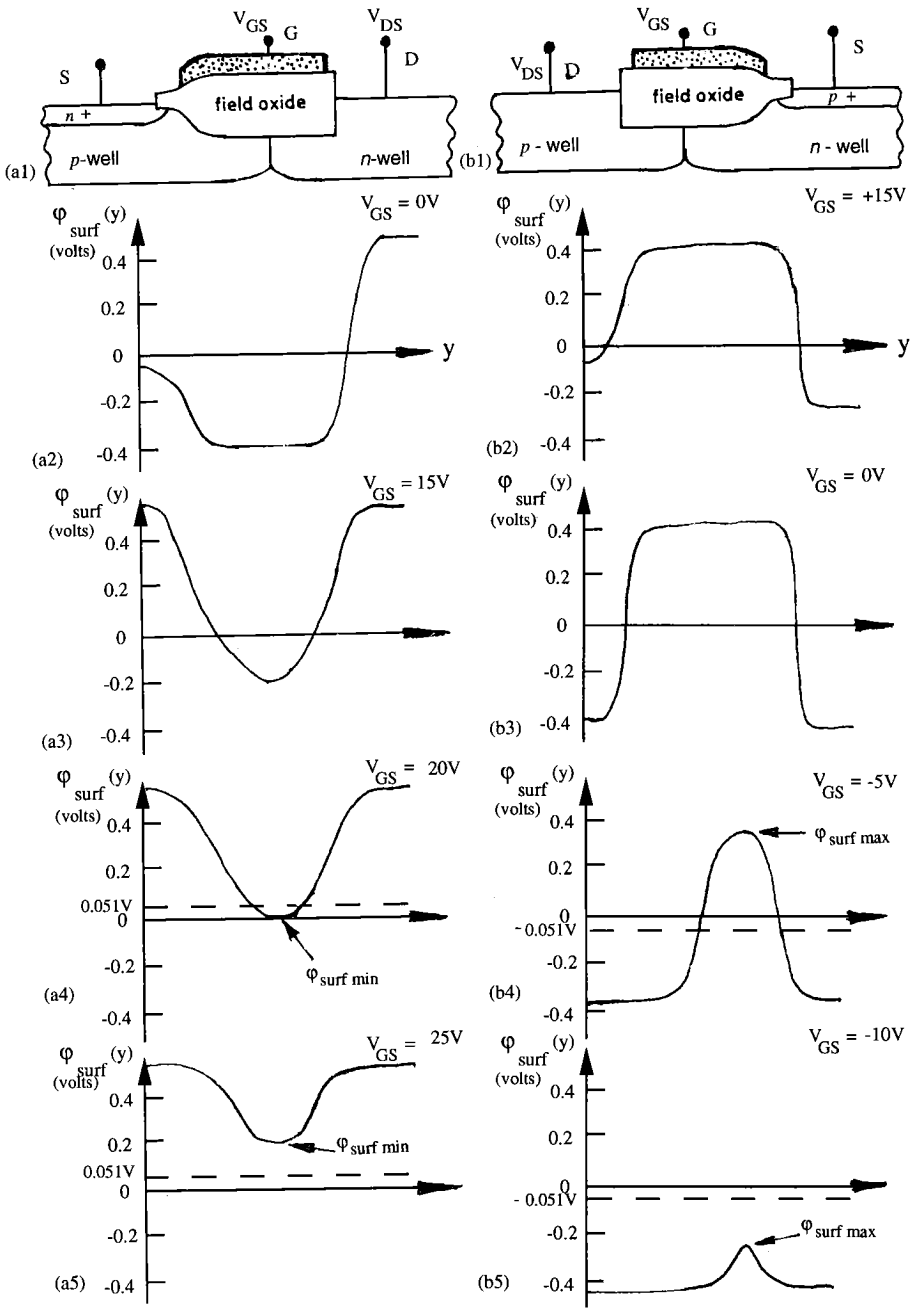


Fig. 6-51 (a1) Cross-section of an n^+-to-n-well parasitic FET with $L = 1$ μm; (a2 - a5) Plots of φ_{surf} in this structure for V_{GS} values of 0, 15, 20 and 25V. (b1) Cross-section of an p^+-to-p-well parasitic FET; (b2 - b5)) Plots of φ_{surf} in this structure for V_{GS} values of +15, 0, -5, and -10V.

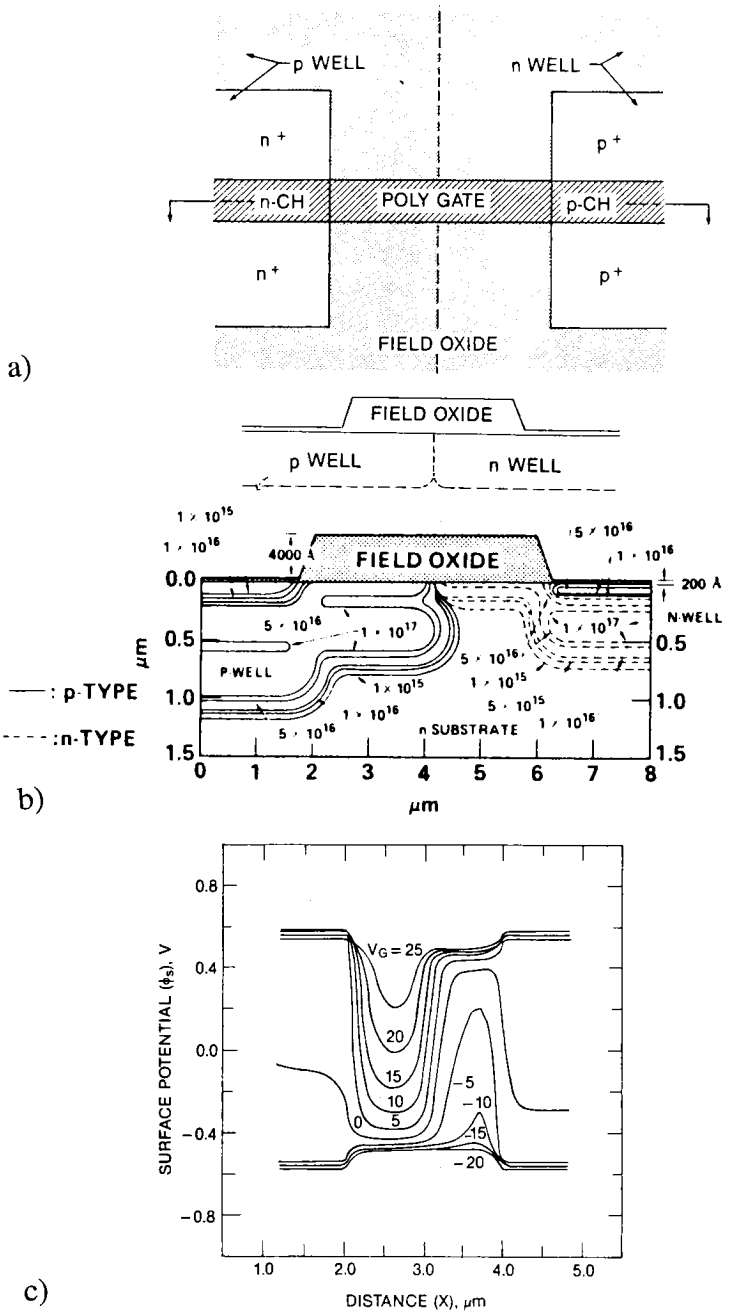


Fig. 6-52 (a) Layout and (b) cross-section and simulated 2-D net impurity contours in the CMOS isolation region of a quadruple-well CMOS technology; (c) Simulated surface potential for 2-μm isolation spacing at various gates voltage V_{GS}. (© IEEE 1986).

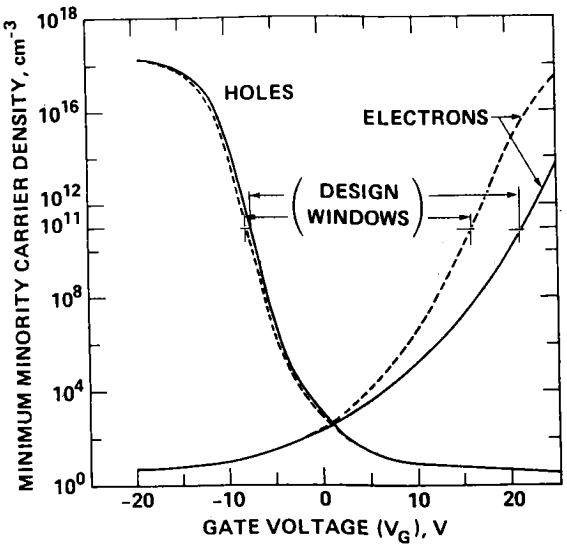


Fig. 6-53 Minimum minority carrier density under the 2-μm field oxide. For a total of 2-μm n^+-to-p^+ spacing, solid lines represent 1-μm p^+-to-p-well spacing and dashed lines present 1.5 μm p^+-to-p-well spacing.[74] (© 1986 IEEE).

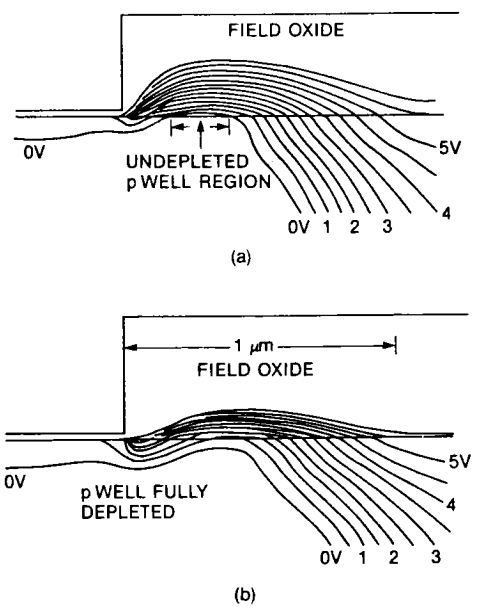


Fig. 6-54 2-D equipotential contours for 2-μm isolation with 1.5 μm p^+-to-p well and 0.5 μm n^+-to-n-well spacing. V_n = 5V, V_p = V_{sub} = 0V. (a) No punchthrough at 10V. (b) Punchthrough at 20V.[74] (© 1986 IEEE).

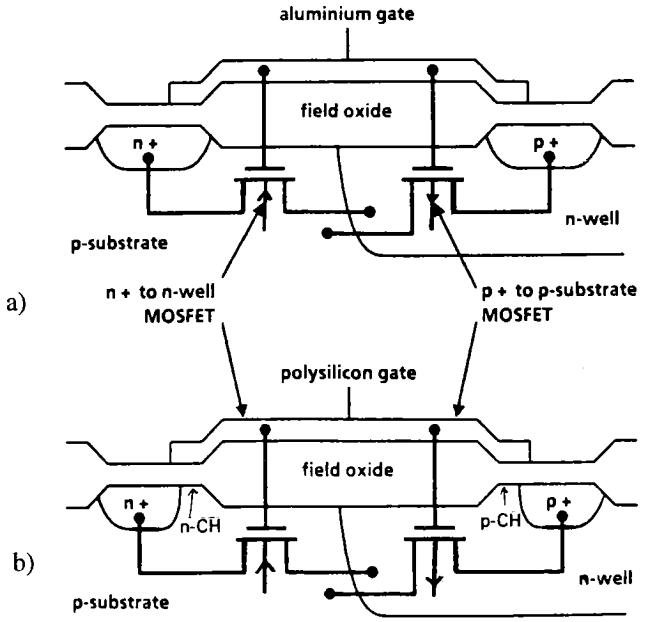


Fig. 6-55 Asymmetric parasitic FETs formed at *n*-well edge: (a) Aluminum gate structure; (b) Polysilicon gate structure.[91] (© 1987 IEEE).

test structures can be designed for *p*-well or twin-well CMOS.) These cross-sections indicate that both aluminum (Fig. 6-55a) and polysilicon (Fig. 6-55b) gate parasitic transistors exist in CMOS, and hence it is important to characterize each of these types to determine if one or the other might limit the minimum isolation spacing.

As noted earlier, in these test structures, the n^+ regions and *n*-well form the source and drain for the parasitic NMOSFETs while the p^+ and *p*-substrate serve as source and drain for the parasitic PMOSFETs. As in the NMOS isolation test structures described in section 6.1, two important differences exist between the two structures shown here: (1) the field oxide is thicker in the structures with aluminum rather than polysilicon gates (due to the addition of a dielectric layer to isolate aluminum from polysilicon where they overlap); and (2) the n^+ and p^+ regions in the polysilicon gate structures are self-aligned to the polysilicon gate edge instead abutting the LOCOS edges (see Fig. 6-55b), as occurs in the aluminum gate structures. Thus, a thin gate oxide region exists under the polysilicon between the n^+ and p^+ diffusion and the LOCOS edge. This region inverts at a very much lower voltage than the field, and the resultant inversion layers act as very shallow source and drain regions. Nevertheless, the true source and drain regions are kept further apart than in the aluminum gate case, and this tends to inhibit lateral punchthrough.

Note this implies that the polysilicon-gate parasitic-FET test structure is a good vehicle for characterizing V_{TFiso} under the polysilicon lines, but is not as good for characterizing the punchthrough voltage of actual parasitic CMOS FETs. That is, because lateral punchthrough is inhibited in the polysilicon gate test structure, this test

structure does not test for the worst case conditions under which lateral punchthrough will occur in such parasitic FETs.

The isolation characteristics of these parasitic transistor structures are typically determined separately, as was done in the section on modeling, where one well-edge FET was modeled at a time (see Fig. 6-50b). Usually, both V_{TFiso} and the lateral punchthrough voltages are also determined using separate test procedures. A gate voltage needed to establish a predefined maximum drain/source leakage current in these transistors (e.g., 1 pA/μm width) in each case is measured as a function of the n^+-*to*-*n*-well or p^+-to-*p*-well distance. As noted earlier, for worst case considerations, V_{TFiso} must be measured at maximum drain bias, while punchthrough must be measured at highest gate voltage. The minimum n^+-to-p^+ spacing needed to maintain adequately small leakage currents between opposite device types in a CMOS technology for the specified power supply voltage (plus, usually, a safety margin of 10%) can be derived from the data obtained from such measurements.

Specifically, a minimum V_{TFiso} is selected for the CMOS technology being developed. The minimum distance for the n^+-to-*n*-well and p^+-to-*p*-well needed to establish this value for both the aluminum and polysilicon gate structures is determined from the experimental data. The larger of the sum of the two minimum spacings for each parasitic device is the minimum acceptable n^+-to-p^+ spacing to maintain adequate field inversion. The same procedure is used to choose the minimum n^+-to-p^+ spacing to maintain a sufficiently high punchthrough value. The larger of these two spacings (i.e., the minimum n^+-to-p^+ spacing to maintain an adequate V_{TFiso} value in each device, and the minimum n^+-to-p^+ spacing to allow an adequate punchthrough voltage in each device) is the minimum overall n^+-to-p^+ spacing value needed to provide adequate CMOS isolation. Note that since this distance also affects latchup, the final design rule must exceed both the value found from this experimental data and the one necessary for avoiding latchup.

It should also be noted that if both the leakage due to field inversion and punchthrough are measured simultaneously, the same values of V_{GS} and V_{DS} are applied simultaneously as their values are ramped up, as shown in Fig. 6-56.[78] The voltage at which some predetermined drain current is observed is then plotted, and these quantities are referred to as V_C and *d* in Fig. 6-56.

6.6.3.1 Experimental Characterization of Isolation in n^+-to-*n*-well Parasitic FETs. Fig. 6-56a shows the measured values of field inversion voltages V_{TFiso} for n^+-to-*n*-well parasitic MOS structures as a function of n^+-to-*n*-well separation. V_{TFiso} is defined at parasitic FET drain current values of 10 pA, 1 nA, and 0.1 μA, and it assumes that the drain voltage is set at 5V as the gate voltage is ramped, for both the aluminum and polysilicon gate test structures.[91] Results from simulation data are also plotted, and good agreement with the measured data is found.* Below about

* It should be pointed out that the authors indicated that the fixed charge in the threshold equation used to calculate V_T was treated as a fitting parameter to match the experimental and

2 μm nominal n^+-to-n-well separations V_{TFiso} is seen to decrease due to short-channel V_T rolloff effects. That is, at n^+-to-n-well spacings smaller than about 2 μm, the depletion region of the n-well extends far enough toward the n^+ source that the short-channel V_T lowering and DIBL effects observed in active MOSFETs at smaller L values now also begin to occur in the parasitic field device.

The values of V_{TFiso} of the polysilicon-gate parasitic MOS transistors are lower than those of the aluminum gate devices because the field (gate) oxide is thinner in the former. But, the curves corresponding to the different subthreshold drain currents are closer together in the poly-gate devices, due to the smaller subthreshold swing S_t. (As discussed in chap. 4, S_t decreases as the thickness of the gate oxide is reduced.) Furthermore, the threshold lowering is smaller for polysilicon-gate FETs, again due to the thinner parasitic gate oxide. At the smallest n^+-to-n-well separations (1-1.5 μm), the faster V_T rolloff in the aluminum-gate FETs causes their V_{TFiso}s to become smaller than those of polysilicon-gate devices. As a result, it is the aluminum gate structures which ultimately limit the minimum n^+-to-n-well separation. Furthermore, the oxide quality and thickness control of the oxide between the poly and the aluminum is worse than that between the poly and substrate. Hence, the V_{TFiso}s of the aluminum gate

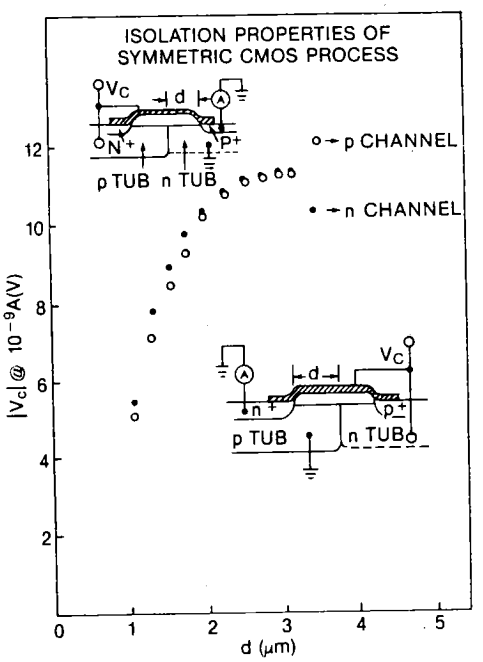


Fig. 6-56 Isolation characteristics of well-edge parasitic as a function of the distance from well edge to active area edge.[78] (© 1986 IEEE).

(* cont.) simulated long-channel V_T values (at values of $Q_{tot}<2x10^{12}q$ cm^{-2}). Thus, the significance of these simulations lies primarily in their predictions of the short-channel roll-off rather than their absolute values.

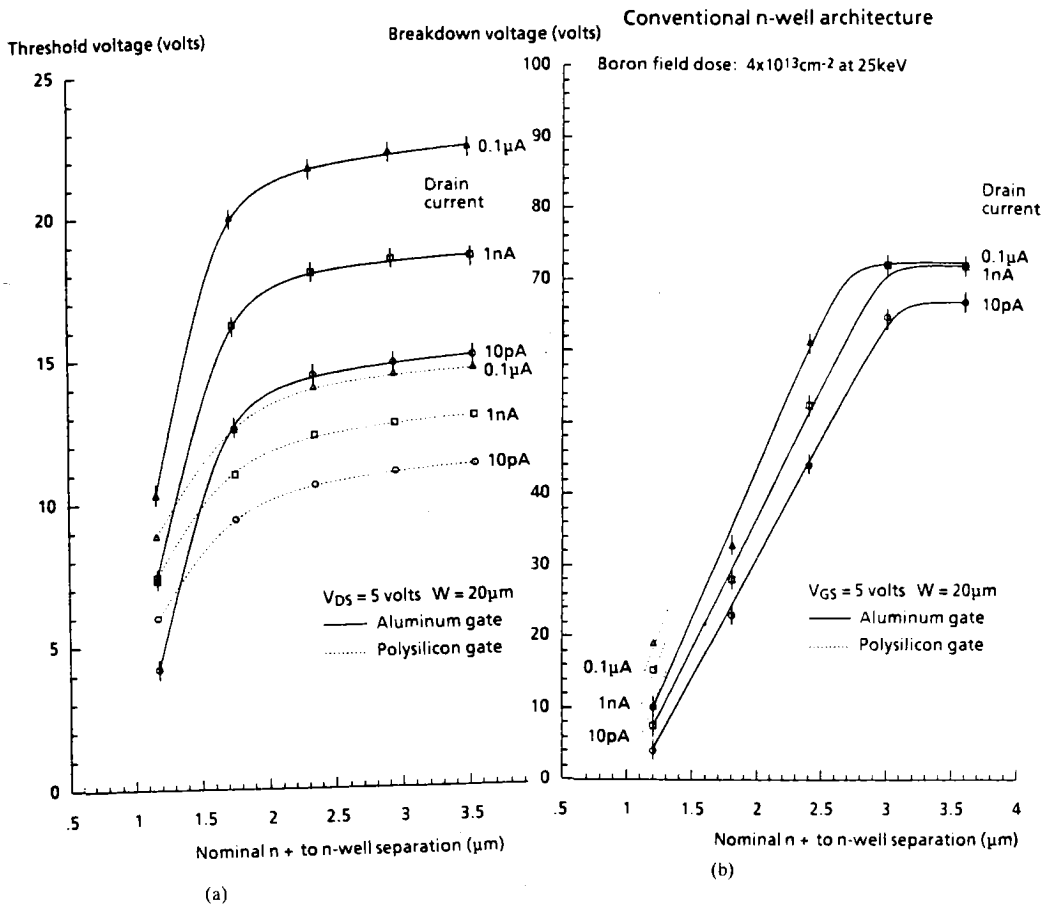


Fig. 6-57 (a) Threshold voltages V_{TFiso}; (b) breakdown voltages for the asymmetric n^+-to-n-well parasitic FETs.[91] (© 1986 IEEE).

structures are less reproducible from run to run. Thus, a greater margin of safety must be allowed (i.e., larger nominal values of V_{TFiso} for the aluminum devices are required).

Figure 6-57b shows the breakdown voltages (i.e., either due to punchthrough or avalanche breakdown, whichever occurs first) of the n^+-to-n-well parasitic test structures as a function of n^+-to-n-well spacing. They are measured by ramping the drain voltage until a voltage is reached which produces drain currents of 10 pA, 1 nA, and 0.1 μA, with $V_{GS} = 5V$. Also shown are the simulated values, and very good agreement between the two is again found. The breakdown voltages of n^+-to-n-well parasitic FETs indicate once again that the aluminum gate FETs ultimately determine minimum separation. Fig. 6-57b shows that at 1.2 μm n^+-to-n-well spacing, the polysilicon gate punchthrough voltages are all about twice the corresponding values for aluminum gate devices. At larger separations, breakdown in the polysilicon gate devices is limited by the dielectric strength of the thin gate oxide in the region where the polysilicon overlaps

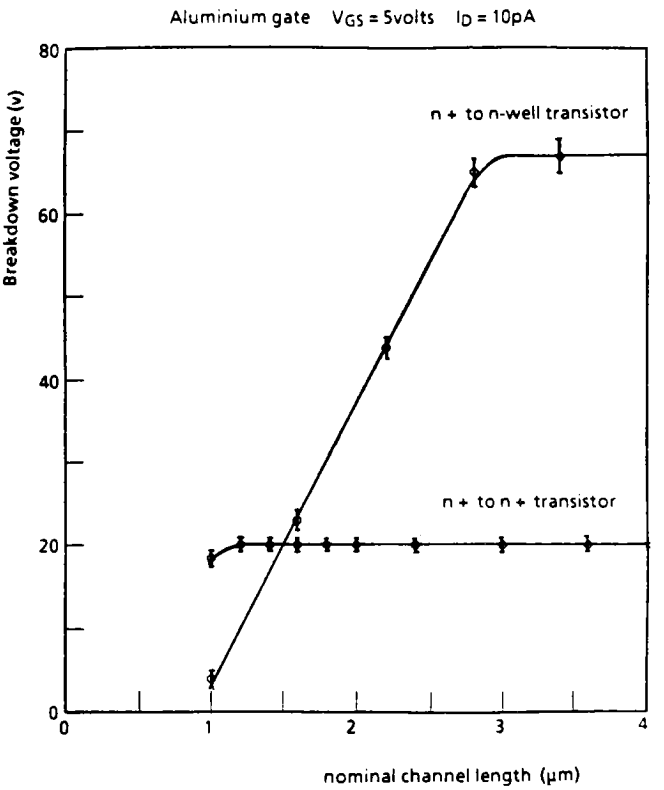


Fig. 6-58 Avalanche and punchthrough voltages for the symmetric (n^+-to-n^+) and symmetric (n^+-to-n-well) parasitic NMOSFETs.[92] (© 1987 IEEE).

the active area in the n-well (indicated as p-CH in Fig. 6-51a). Thus, punchthrough or avalanche breakdown cannot be measured in these structures.

As compared with the symmetric parasitic FETs previously described in section 6.1, punchthrough in asymmetric parasitic short-channel FETs is much more severe because of the deeper drains in these devices. In Fig. 6-58 the breakdown voltages of n^+-to-n^+ parasitic FETs and n^+-to-n-well devices are compared.[92] We observe that when the nominal channel length is reduced below 2 μm (i.e., n^+-to-n-well spacing), this parasitic device exhibits much smaller punchthrough voltages than does the n^+-to-n^+ parasitic FET.

6.6.3.2 Experimental Characterization of Isolation in p^+-to-p-substrate Parasitic FETs. Measured and simulated field inversion voltages V_{TFiso} for p^+-to-p-substrate parasitic FETs are shown in Fig. 6-59 for both aluminum and polysilicon gate devices.[91] Again, agreement between the two is good. We

observe, however, that the structure of this parasitic FET differs considerably from that of a conventional FET, and even that of the n^+-to-n-well FET. That is, the p-substrate (which acts as the drain) is not only deep, but wraps around under the p^+ source. Thus, there is the possibility of lateral and vertical leakage paths between the p^+ and p-substrate. But the vertical leakage path is not affected by the parasitic gate voltage, and hence may be neglected in the present discussion. We discuss vertical punchthrough in such devices in section 6.6.5.

We see in Fig. 6-59 that the V_{TFiso} values are much higher than those of n^+-to-n-well FETs, especially for the case of the aluminum gate devices. This is due to several effects, including the fact that higher doping is needed in the n-well to ensure adequate compensation of the boron, and that any fixed charge at the Si-SiO_2 interface is likely to be positive and therefore increase the magnitude of V_{TFiso}. Short-channel effects are also

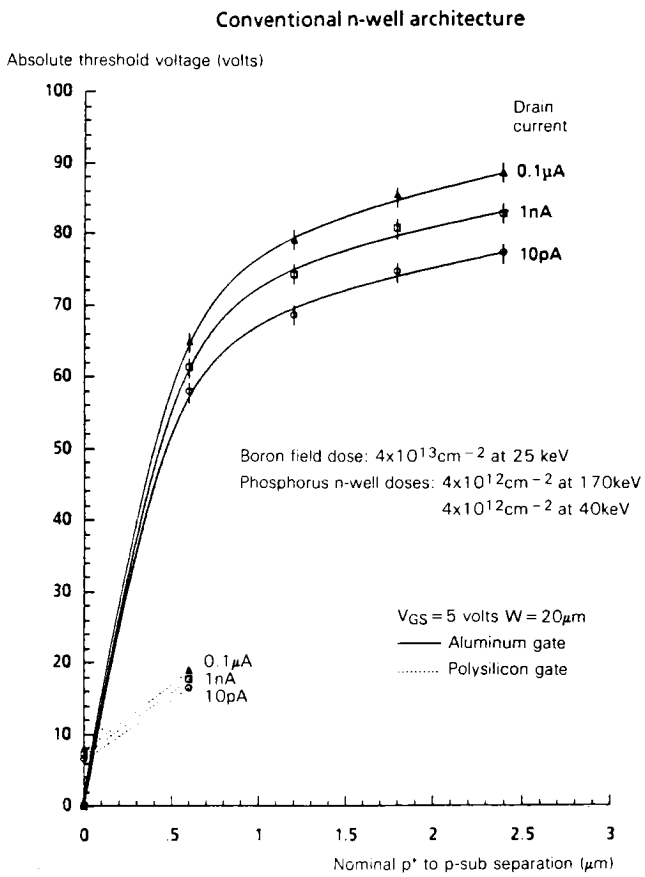


Fig. 6-59 Threshold voltages for the symmetric p^+-to-p-substrate parasitic FETs.[91] (© 1987 IEEE).

less severe, partly due to the larger doping in the n-channel, and partly because the drain (substrate) doping is lower than that of the channel. Thus, the drain-channel (substrate-well) depletion region width extends into the true substrate rather than encroaching through the well toward the p^+ source.

Since nominal p^+-to-p-substrate separation represents the separation of the nitride LOCOS mask and n-well resist mask, and the actual p^+ in the polysilicon device isabutted to the polysilicon edge rather than the nitride edge, the actual p^+ to p-substrate separation is larger than its nominal value by the length of the poly-overlapping thin oxide. The values of V_T can only be found at 0.6 μm or less, because of the dielectric breakdown in the thin oxide region. At nominal zero separation in the polysilicon devices, the actual p^+-to-p-substrate separation is about the same as the length of the poly-overlapping thin oxide, that is, 1.2 μm. In the case of an aluminum gate in which the p^+ is abutted to the LOCOS edge, the devices fail to operate.

Lateral punchthrough between the p^+-to-p-substrate is not seen until the nominal separation becomes smaller than 0.6 μm. As described earlier, this minimum punchthrough occurs mostly because the majority of the drain depletion region is extended into the lightly doped p-substrate, which serves as the drain in this structure. Vertical punchthrough between p^+ to p-substrate can be a more worrisome problem, particularly when the n-well depth is scaled down in submicron CMOS. This issue, as noted earlier, is discussed in section 6.6.5 which deals with vertical punchthrough.

6.6.4 Establishing the Minimum n^+-to-p^+ Isolation Spacing in CMOS from Modeling and Experimental Data

The data presented in Fig. 6-60a indicate that using conventional wells permits adequate isolation to be established between n^+ diffusions and n-wells in n-well CMOS at about 1.5 μm nominal separation (i.e., for this structure V_{TFiSO} is ~10V at 1.5 μm, and V_{PT} is ~20V). Figure 6-60b also shows that this separation can be reduced to about 1.0 μm with retrograde wells. Likewise, it has been shown that p^+-to-p-substrate separations down to 0.6 μm can be achieved in n-well CMOS with either conventional or retrograde wells. Thus, it should be possible to realize full CMOS structures with about 2 μm n^+-to-p^+ separation using conventional n-wells, and about 1.6 μm separation with a retrograde-well technology. However, these results are obtained with isolated n^+-to-n-well and p^+-to-p-substrate parasitic transistors. To establish isolation design rules, it is necessary to show that adequate isolation exists at such small separations in full CMOS devices.

Figure 6-61 shows supply currents drawn by full CMOS structures with 1.8 μm n^+-to-p^+ separations (1.2 μm n^+-to-n-well edge and 0.6 μm p^+-to-n-well edge).[91] Fig. 6-61a shows the test structures used for this arrangement. The output node, connecting n^+ and p^+ diffusions closest to the well edge, is either shorted to ground ($V_{out} = 0$) or to the supply line ($V_{out} = V_{DD}$). These cases simulate an inverter with its output in either a Low or High state. The leakage current drawn by an ideal CMOS device between V_{DD} and V_{SS} should be zero in both cases, but any leakage due to

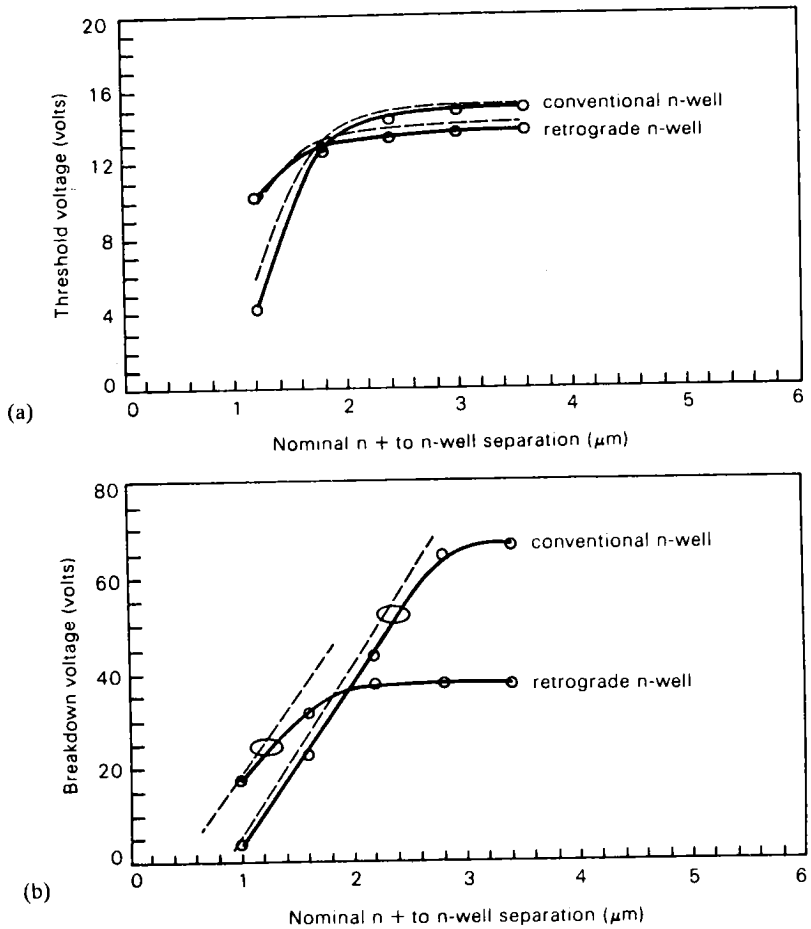


Fig. 6-60 Comparison of retrograde and conventional well structures for n^+ to n-well parasitic FETs: (a) field threshold voltage; and (b) n^+-to-n-well breakdown voltage. Solid lines: experimental data, dashed lines: modeled results, breakdown voltage is modeled by punchthrough only.[91] (© 1987 IEEE).

insufficient isolation between the n- and p-channel devices will show up as a component of the supply current.

Fig. 6-61b shows results for conventional-well devices. When $V_{out} = V_{DD}$, no potential difference exists between the n^+ diffusion and the n-well, and no leakage current should flow between them. The limiting phenomenon here is thus p^+-to-p-substrate breakdown, and consequently no breakdown should be observed below 20V. This indicates that the worst case for leakage is when the output node is shorted to ground. In such cases, the full supply voltage appears between the n^+ diffusion and the n-well, and the mechanism limiting the maximum operating voltage is instead n^+-to-n-well punchthrough leakage. The results shown in Fig. 6-61b for $V_{out} = 0$ are consistent

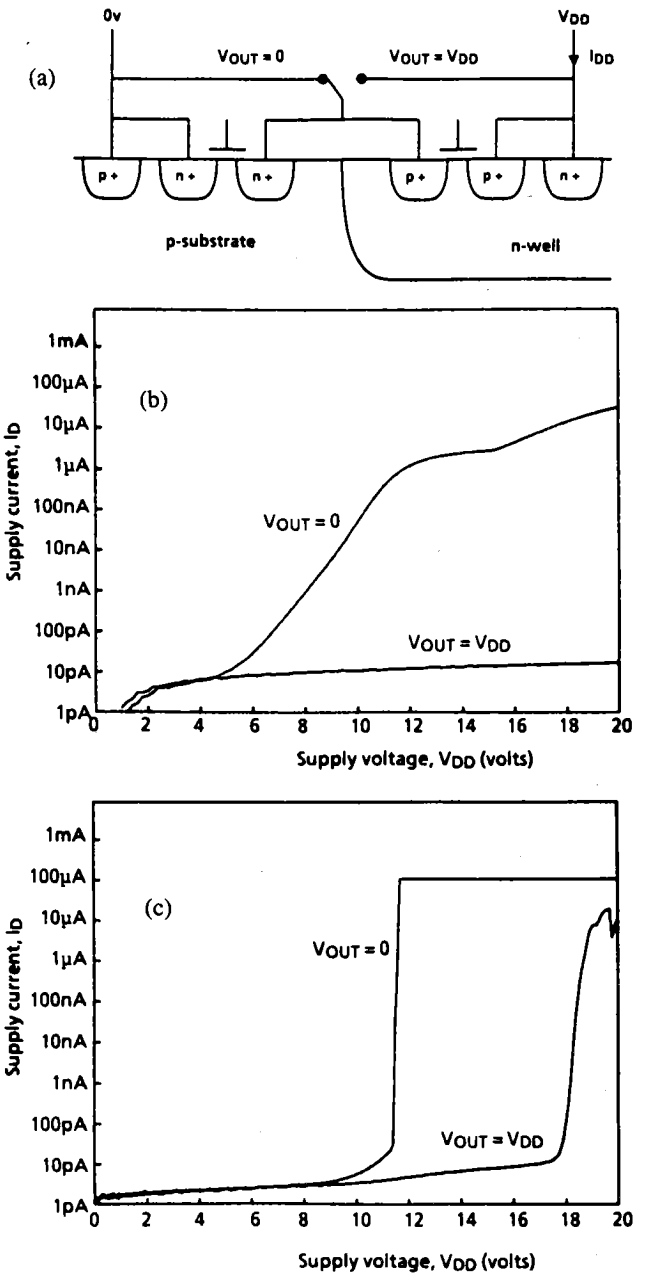


Fig. 6-61 Test structures and I-V curves for a full CMOS isolation structure: (a) test structure and measurement arrangement; (b) leakage current for conventional-well CMOS. (c) leakage current for retrograde well devices.[91] (© 1986 IEEE).

with the punchthrough breakdown voltages exhibited in Fig. 6-57b where at nominal n^+-to-n-well separation, the breakdown (punchthrough) voltage is ~10V. Figure 6-61b shows that at V_{DD} = 5V the leakage current is very close to the maximum acceptable value of 10 pA, and increases steeply as V_{DD} is increased above 5V. This demonstrates that the CMOS device under test provides barely adequate isolation at V_{DD} = 5V, but also that this does not provide adequate safety margin for 5V operation. However, by increasing the n^+-to-n-well spacing to 1.8 μm (n^+-to-p^+ separation = 2.4 μm), the maximum operating voltage becomes limited instead by junction breakdown, which in this case is just over 20V. Thus, the minimum n^+-to-p^+ separation acceptable for 5-V operation with conventional wells lies between 1.8 and 2.4 μm.

Figure 6-61c shows results for retrograde well devices with 1.8 μm n^+-to-p^+ separation. In these instances, the leakage current remains below 10 pA for supply voltages as high as 10V, providing a good safety margin for 5V operation. Again, the condition with V_{out} = 0 sets the lowest breakdown voltage, although n^+-to-n-well junction avalanche breakdown occurs at only a few volts higher than the onset of n^+-to-n-well punchthrough leakage.

In conventional LOCOS-isolated CMOS ICs, isolation among similar type transistors can be achieved with 1.0 μm isolation spacing. Isolation among opposite-type transistors is more involved, however, and modeling and measurement results show that it is possible to achieve isolation between an n- and p-channel transistor with approximately 2.4 μm isolation spacing. The use of retrograde wells can further reduce this isolation dimension to below 2 μm. At such small isolation dimensions, latchup becomes a major problem and ultimately may instead become the factor limiting the isolation spacing between NMOS and PMOS transistors.

The results presented above were obtained from n-well technologies, and thus, one might ask how well such might apply to p-well and twin-well processes. Although a p-well structure could be made by merely exchanging the n- and p-type dopants, (implying that the isolation characteristics would be comparable), data indicates that in one major respect the isolation behavior is somewhat different than for n-well CMOS. That is, the most important isolation leakage mechanism in submicron CMOS appears to be due to lateral punchthrough when the lateral spacing between the well border and source/drain regions outside the well is small. In the case of n-well CMOS, the boron channel-stop implant in the substrate must be fairly heavy due to boron segregation into the field oxide during field oxide growth. As a result, the boron profile in the substrate penetrates relatively deeply. This is exactly the type of substrate doping profile needed to prevent subsurface lateral punchthrough. In p-well CMOS, on the other hand, the phosphorus channel-stop dopant tends to pile up at the surface, yielding a shallower final phosphorus channel-stop doping profile in the substrate. This type of profile is less effective in suppressing lateral subsurface punchthrough. Thus, it may be more difficult to suppress p^+-to-p-well punchthrough in p-well CMOS.

In twin-well CMOS, both n-type and p-type devices lie inside a well. Thus, it might appear that the superior p^+-to-p-well isolation performance observed in the n-well CMOS results would be duplicated between the n^+-to-n-well in twin-well CMOS. However, the well profiles produced in fabricating the n-well CMOS for the above-cited

study appear quite like twin wells, with the boron field implant extending as deep or deeper than the shallow *n*-well. Consequently, it is not expected that significantly better isolation is achievable using a twin-well technology.

6.6.4.1 Isolation-Spacing Design Rules for 0.35 and 0.25 μm CMOS. Recent reports summarizing the work required to establish the minimum isolation spacings in submicron CMOS (as described above) indicate the following design rules will be needed in 0.35 μm and 0.25 μm CMOS technologies:[84,133]

1. For 0.35 μm CMOS, the minimum active pitch at which adequate *intrawell isolation* (n^+-to-n^+, or p^+-to-p^+) can be maintained appears to be about 1 μm (i.e., ~0.5 μm active region width plus 0.5 μm isolation spacing width). For LOCOS-based approaches, the minimum dimension for *interwell isolation* (i.e., n^+-to-p^+ spacing) is about three times that required for intrawell isolation. Thus, the minimum n^+-to-p^+ spacing for 0.35 μm CMOS is about 1.5 μm.

2. For 0.25 μm CMOS the above design rules can be scaled by about 30-40%. Using this criterion, the active pitch for 0.25 μm CMOS would be 0.6-0.7 μm. This means the minimum intrawell isolation spacing would be 0.3-0.35 μm, and that for interwell isolation spacing, would be 0.9-1.05 μm.

These design rules represent the nominal requirements. Obviously, a manufacturable isolation process will need to specify allowable tolerances to these dimensions to accommodate the inevitable variations in patterning technology.

6.6.5 Vertical Isolation in CMOS

By reducing the depth of the wells in CMOS, smaller n^+-to-p^+ spacing can be used. However, when shallow wells are implemented the problem of maintaining adequate vertical isolation must be addressed. That is, shallow wells are more prone to vertical punchthrough between the source/drain diffusions in the well and the underlying substrate. Retrograde wells (described in chap. 8), with their high doping levels at the well bottom provide excellent vertical isolation, even at well depths of the order of 1 μm,[115, 116] but adequate vertical isolation at such shallow well depths is much harder to achieve with conventional-well structures. Shallow wells fabricated with a significantly reduced thermal budget also allows the use of thinner epitaxial substrate material, and hence better latchup protection. At the same time, however, the shallow wells introduce a new constraint on the epitaxial layer thickness due to the degradation in vertical isolation caused by the underlying heavily doped substrate. As we will see, this is especially true in the case of *n*-well CMOS, where out-diffusion of boron from the heavily doped substrate imposes a major limitation on the minimum epitaxial layer starting thickness.

Example 6-1: An *n*-well CMOS process is to be developed for operation with a power-supply voltage of V_{DD} = 5 V. The substrate doping of the *p*-type wafers is $1x10^{15}/cm^3$. Assume the *n*-well has a uniform doping profile with an average dopant concentration of $1x10^{16}/cm^3$. The *p*-channel MOSFET sources

and drains are to have junction depths of 0.4 μm and an average dopant density of $10^{18}/cm^3$. What is the minimum *n*-well depth needed to avoid vertical punchthrough to the substrate?

Solution: Vertical punchthrough will occur if the depletion region of the source/drain-to-well junction were to merge with the depletion region of the well-substrate junction when V_{DD} = 5 V (see section 5.5.2). The source-to-well junction is essentially a one-sided *pn* junction with a built-in voltage of V_{bi} = 0.82 V. In this case, the depletion-region width is calculated to penetrate into the *n*-well to a distance of ~0.35 μm, if the voltage applied across the junction is zero (i.e., both the source and the well are connected to V_{DD}). The *np*-junction to the substrate has a built-in voltage of 0.63 V, and the total depletion width of this junction at a 5 V bias is ~1.9 μm. We calculate that about 0.19 μm of this depletion region is in the *n*-well. The *n*-well must therefore be deep enough to accommodate the depth of the source junction (0.4 μm) as well as the sum of the depletion region widths in the well in order for vertical punchthrough to be avoided. While the total of these dimensions is 0.94 μm, it is good engineering practice to increase the depth of the well by about 50% to account for process variations. A reasonable well depth might therefore be 1.5 μm.

It is more feasible to fabricate CMOS having adequate vertical isolation with shallow conventional *p*-wells than with conventional *n*-wells. It has been demonstrated that successful *p*-well CMOS circuit operation can be achieved at V_{DD} = 5V with conventional well depths of approximately 1 μm.[117] Consider first the case when the n^+ source/drain diffusion in such shallow *p*-wells is biased more positively than the *n*-substrate. In this instance, the vertical punchthrough voltage is only 3-4 V. However, since the *n*-substrate is tied to V_{DD} and the n^+ region is biased between V_{DD} and V_{SS} this bias arrangement does not occur in normal circuit operation (except during transient overshoot of the voltage at the circuit node). Thus, a punchthrough voltage value of 3-4V is adequate in this case.

Let us also consider the case of vertical punchthrough current flowing in the opposite direction in a *p*-well CMOS structure. That is, when the substrate is biased at a higher potential than the n^+ region. In this case, in a *p*-well circuit under normal operating conditions, the *n*-substrate must always be at the most positive potential. In such circumstances (i.e., with the *n*-substrate at the higher potential and the n^+ region at V_{SS}), the *n*-substrate-to-n^+ punchthrough voltage is measured to be well above 10V. Since the substrate is usually tied to V_{DD} = 5V, under normal operation this punchthrough voltage value is also adequate for *p*-well CMOS operation.

Under some biasing conditions, however, shallow wells in *n*-well CMOS are much more prone to vertical punchthrough, and the problem becomes worse if the *n*-well CMOS is fabricated in a *p*-epi layer on a p^+ substrate. Note first that in *n*-well CMOS vertical punchthrough current flows between the p^+ region and the *p*-substrate through an *n*-well. In normal CMOS operation, the *n*-well is at V_{DD}, the *p*-substrate is at V_{SS} and the p^+ is biased between V_{DD} and V_{SS}. However, during transient overshoot, the p^+

or p-substrate can become negatively biased. Since this condition represents the worst case for vertical punchthrough, it is useful to analyze it in more detail.

The problem can be illustrated by considering an example of a conventional but shallow n-well with its doping profile (obtained using the process simulator SUPREM III) given as in Fig. 6-62a.[93] The well depth is 1.1 μm and the p^+ junction depth is about 0.3 μm. The simulated potential distribution for this case is depicted in Fig. 6-62b. In this figure, the p^+ is biased between +5V and -5V while the n-well and p-substrate are biased at normal operating voltages (i.e., the n-well is biased at V_{DD} = 5V and the p-substrate at V_{SS} = 0V).[93] As the p^+ is varied from +5V to 0V, the potential energy barrier between the p^+ and p-substrate regions is always present, even when the p^+-to-n-well and n-well-to-p-substrate depletion regions meet. Consequently, under normal operating conditions, vertical punchthrough current is not expected to flow. When the p^+ potential reaches -5V, however, the barrier disappears, permitting punchthrough current to flow from the substrate up to the p^+. p-channel devices formed in such an n-well would thus be limited to operation at drain biases smaller than -10V.

Punchthrough current can also flow from the p^+ down to the substrate if the substrate is negatively biased (a condition also examined in Fig. 6-62b). However, because most of the well-to-substrate depletion region is in the lightly doped substrate rather than into the well, the p^+-to-substrate potential barrier cannot be lowered significantly until the substrate is at a large negative bias (<-70V).

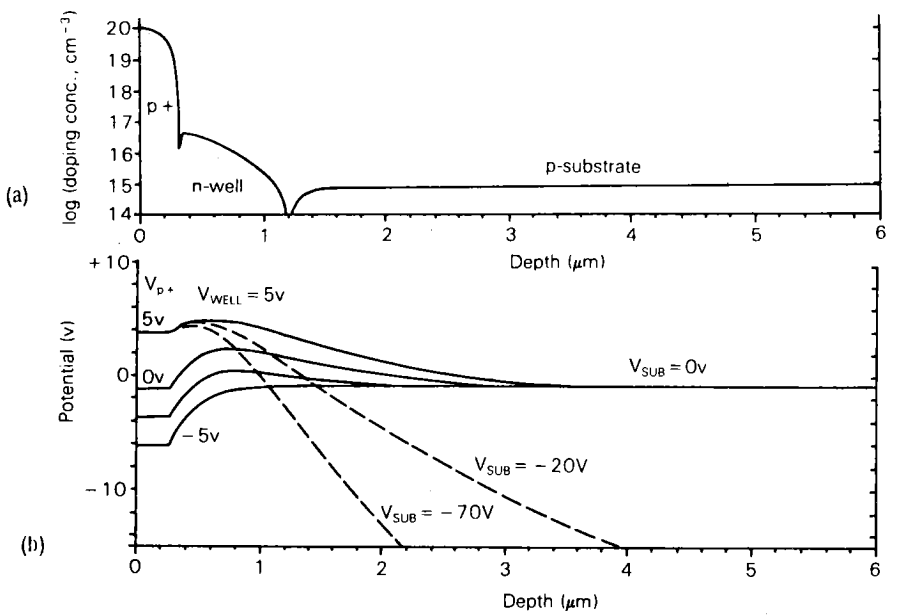


Fig. 6-62 Vertical doping and potential profiles for a 1.2 μm CMOS structure. (a) Net doping, and (b) potential profiles, solid lines: p^+ voltage varies with p-substrate grounded, dashed lines: substrate bias changes with p^+ fixed at 5 V. The n-well is always at 5 V. (c) Potential profile showing vertical punchthrough with p^+ more positive than p substrate.[93] (© 1987 IEEE).

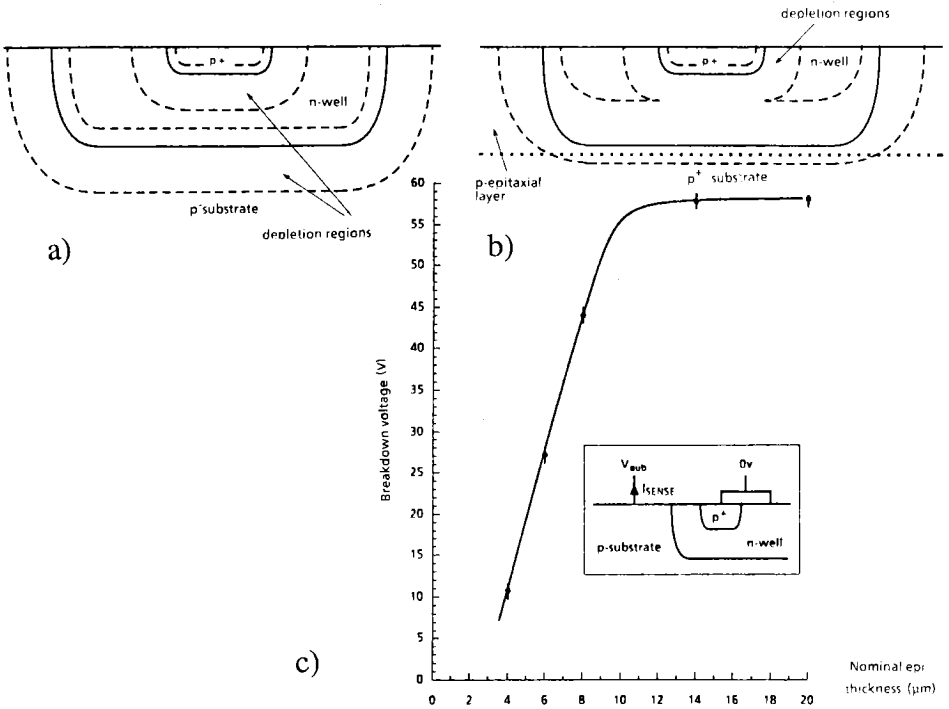


Fig. 6-63 (a) Influence of thin epitaxial layer on depletion regions in an n-well. (b) Thin epitaxial material. (c) Vertical punchthrough voltage versus nominal (i.e., as-grown) epi thickness in a CMOS structure with a 1.1 μm well depth. Sense current = 1 nA.[93] (© 1987 IEEE).

The situation is altered, however, if the p-substrate is replaced by a p-epi layer on a p^+ substrate. Epitaxial substrates are often required for latchup protection, especially as dimensions are reduced (see Vol. 2, chap. 6). As the minimum n^+-to-p^+ spacing is shrunk, the epitaxial layer thickness must also be reduced.[118] In the case of n-well technology, this means that the heavily doped p^+ substrate starts to approach the base of the n-well. Because the heavily doped substrate prevents the extension of the well-substrate depletion region down to the substrate, the depletion region is forced up into the well (see Figs. 6-63a and b). As a result, the V_{SUB} needed to lower the potential energy barrier is reduced, thus making it easier to cause vertical punchthrough. Figure 6-63c shows vertical punchthrough voltages measured in 1.2 μm deep wells as a function of nominal p-on-p^+ epitaxial thickness.[93] It indicates that the magnitude of V_{SUB} at the breakdown of vertical isolation decreases as the epi layer is thinned. At 4 μm nominal (as-grown) epi thickness, isolation breaks down at V_{SUB} of about -5V (i.e., if the p^+ diffusion is biased at +5V, the breakdown voltage is 10V). Therefore, vertical isolation may limit the use of the thinnest epi layer. Where thin epi layers are required for high-packing-density CMOS circuits in latch-up sensitive applications, the as-grown epitaxial layer thickness will need to be tightly controlled to avoid unacceptable degradation of vertical isolation.

6.6.6 LOCOS-Based Isolation in CMOS

The AT&T Twin-Tub V technology is an example of an advanced CMOS process that uses a LOCOS-based approach to isolate like devices.[79] The starting material is a p-epi layer on a p^+ substrate. The technology is implemented with a single-mask, self-aligned twin-well process that uses two channel-stop implants (as well as two separate well implants).

First, phosphorus and arsenic are sequentially implanted into the n-well areas, forming a high/low doping profile (channel-stop and well; Fig. 6-64). Since arsenic diffuses much more slowly than phosphorus, it will remain near the surface of the well during the drive-in step. This high arsenic concentration provides an effective channel stop for p-channel devices and also protects against punchthrough.

Next, a relatively thick masking oxide is selectively grown over the n-well region, and the p-well implant (boron) is then performed. Because the implant is blocked by the oxide that covers the n-wells, it enters the silicon only in the p-well regions (a self-aligned implant step). Both well regions are then driven in (with the thick oxide over the n-wells being retained), and a second boron implant (which will serve as the p-well channel stop) is carried out. This implant is kept shallow because a high pressure LOCOS process is subsequently used to grow the field oxide, minimizing the lateral and vertical impurity-profile spreading. Once again, the implant is self-aligned to the p-well regions by the presence of the oxide on the n-wells. The masking oxide is then removed, and a nitride masking layer is deposited and patterned to cover the active areas. Finally, the field oxide is formed.

This process makes it possible to achieve adequate isolation and a reduced tendency toward latchup with smaller device separations than are possible with an atmospheric-pressure oxide-growth process. Nevertheless, a 7-μm spacing must be used between the n^+ and p^+ regions in the Twin-Tub V structure to provide adequate isolation. Since this spacing still represents a significant layout-density limitation, alternatives to the LOCOS-only process are being vigorously pursued.

In another study using n-well technology, a high dose p-type channel-stop implant for the substrate (yielding a peak concentration of $5\text{x}10^{16}$ boron atoms/cm^3) allowed n^+ to p^+ spacings of 2.4 μm with conventional wells, and 1.8 μm spacings with retrograde wells.[82]

A novel LOCOS-based submicron-CMOS isolation approach employs a FMPBL isolation structure and a high energy boron "plug" implant.[84] The FMPBL with a conventional channel-stop implant provides nominal V_{TFiso}s in excess of 10V between NMOS transistors (intra-well isolation). The plug implant allows inter-well p^+ to n^+ spacings of less than 2.0 μm without the use of epi layers, trench structures, or retrograde wells. Instead, a low-dose of high-energy boron atoms are implanted through the field oxide (5500Å-thick) to form a p^+ plug just under the oxide in the p-well adjacent to the boundary of the n-well (Fig. 6-65). The lateral extent of the implant is controlled by implanting through a narrow slit in the implant mask. This plug effectively controls NMOS field leakage without degrading PMOS isolation. With the plug, lower implant channel-stop doses can be used, which helps curtail the narrow-

width effect due to boron redistribution. Although the implant could be done before field oxidation, by doing it afterwards the problem of boron diffusion during the oxide growth step is circumvented.

6.6.7 Trench Isolation for CMOS

Trench isolation has also been explored for twin-tub CMOS as a means of implementing n^+-to-p^+ isolation structures, primarily as a means for achieving smaller interwell (n^+-to-p^+) isolation spacing. (Other advantages of trench isolation over LOCOS are described in section 6.5.) A trench at the n-well to p-well interface prevents dopant interdiffusion where the two wells join. Thus, dopant compensation (which lowers the surface dopant concentration at the well borders) is eliminated. This permits larger parasitic V_{TFiso} values to be maintained at smaller isolation spacings. Shallow and deep trenches of micron and submicron widths have been used to reduce p^+ to n^+ spacings to 2-2.5 μm.[85]

When trenches deeper than the well depth are used, they replace the reverse-biased pn junction as the isolation structure at the well sidewalls (Fig. 6-66). However, since fabrication of deep-trench isolation for CMOS is substantially more complex than shallow-trench isolation, and latch-up measurements have not fulfilled expectations, this approach has not yet been widely used for such applications. Nevertheless, some recent reports have described deep trench isolation structures as potential candidates for 0.5 μm (and smaller) CMOS technologies.[119,120] In this section we limit our discussion to shallow-trench isolation structures for CMOS. Readers interested in more details on deep-trench isolation are referred to chap. 2, Vol. 2 and the references.

While their use appears to permit smaller n^+-to-p^+ spacing, the presence of trench structures also gives rise to additional leakage paths not present in conventional CMOS isolation structures. These new paths can degrade the isolation between the active devices and the well. In some cases these paths can also increase the subthreshold leakage current within *active* devices. Generally, these new leakage phenomena involve the formation of an inversion channel along the vertical trench sidewall (see Fig. 6-67a).

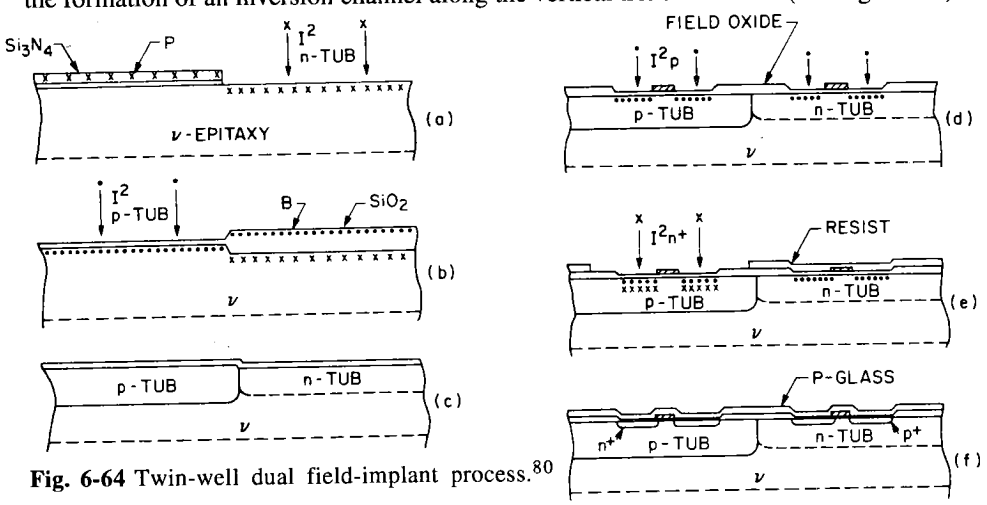


Fig. 6-64 Twin-well dual field-implant process.[80]

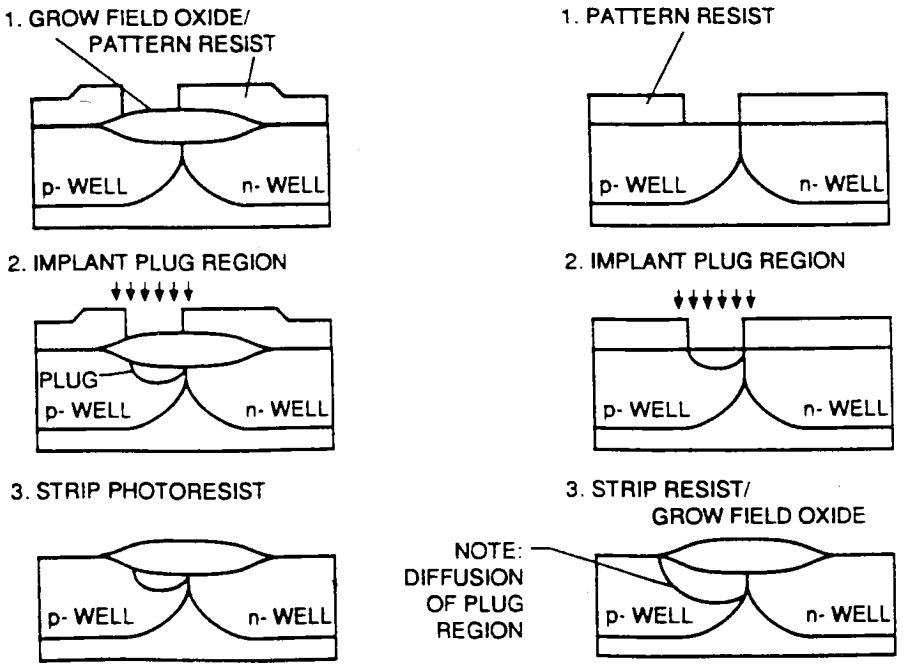


Fig. 6-65 Process flow for the plug implantation process. The figure shows that the implant can be performed either before (left) or after (right) field oxide growth.[84] (© 1991 IEEE).

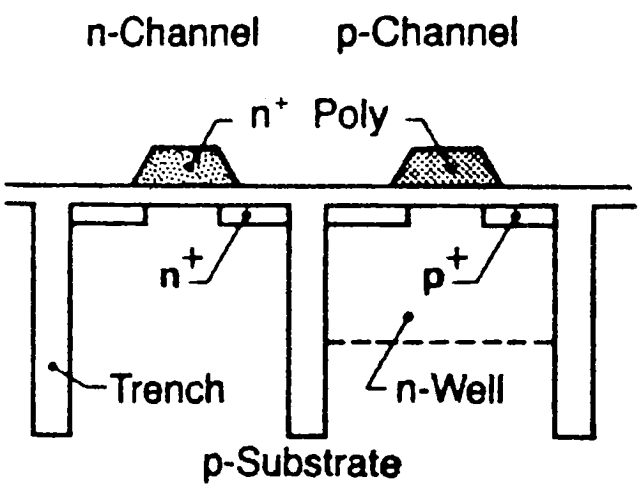


Fig. 6-66 Schematic of an *n*-well CMOS with deep trench isolation.[83] (© 1986 IEEE).

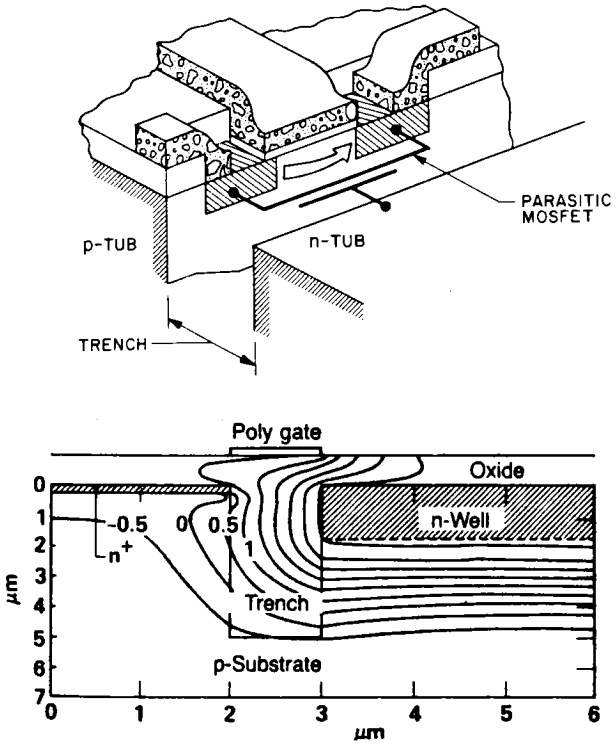


Fig. 6-67 (a) A parasitic sidewall MOSFET with the *n*-well as the gate and the trench serving as the gate oxide.[75] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission. (b) Cross-sectional view of a simulation of the 2-D equipotential contours in an *n*-well CMOS with trench isolation. The simulation is done at $Q_{fc} = 5x10^{10}$ cm^{-2}, $N_A = 6x10^{14}$ cm^{-3}, $V_{well} = 3V$, $V_{GS} = 0V$, $V_{sub} = -1V$.[137] (© 1983 IEEE).

The sidewall inversion is caused by the *horizontal* parasitic MOS device, in which the positively biased *n*-well acts as the gate electrode, and the trench dielectric serves as the MOS gate oxide (with a thickness equal to the trench width). If the *n*-well is biased at +5V, and the trench is narrow (e.g., ≤ 1 μm), inversion along the sidewall of the trench between the source and drain of the MOSFET can easily occur. (As shown in Fig. 4-67a, this is the trench sidewall that is outside of, but facing, the *n*-well.) In Fig. 6-67b, the simulated 2-D potential contours in a trench isolated CMOS structure are shown. Note that the potential on the sidewall of the trench on the *p*-substrate side is >0.5V, indicating that it is inverted.

It would appear that to take maximum advantage of such trench structures for reducing isolation spacing the active devices should be placed right at the tub edge (as shown in Fig. 6-67a). However, if an active NMOS device abuts an inverted trench sidewall, a conducting path will exist between the source and drain of the active NMOS device. This path will exist regardless of the bias on the "real" gate of the active

NMOS device, giving rise to an increase in the subthreshold leakage current of the active device. In practice, to avoid this problem, active devices are usually moved back somewhat from the tub edge.

Even if the NMOS device is moved back from the trench edge, the trench sidewall can still become inverted by a positively-biased *n*-well. If this happens and the poly over the field region also inverts the top surface of the *p*-well, it is possible for a leakage path from the NMOS source to the *n*-well to be established, as shown in Fig. 6-68.

Assuming that the leakage path of Fig. 6-68 is somehow suppressed (e.g., by implanting channel-stop dopants into the bottom of the trench), the inverted trench sidewall can still produce other, unexpected leakage paths. One such dc leakage path between V_{DD} (e.g., 5V) and V_{SS} (e.g., 0V) is illustrated in Fig. 6-69 for the case of two series-connected CMOS inverters which suffer from inadequate CMOS isolation.[87] This path arises from leakage along both the inverted surface of the field regions and the trench sidewall (which, in this example, is offset from the active device edge). Here V_{in1} to the first inverter is at logic 1 (i.e., $V_{in1} = V_{DD}$), causing the NMOS transistor of the first inverter to be turned *ON*. The output of the first inverter V_{out1} (and thus the input to the second inverter V_{in2}) is at logic 0. This causes the output of the second inverter V_{out2} to be at logic 1, and causes the PMOS device of the second inverter to be turned *ON*. The black highly conductive paths in Fig. 6-69 from V_{DD} to the edge of the n^+ region in the second inverter and from V_{SS} to the p^+ region in the first inverter are connected by a less-well-conducting parasitic conductive path. This path consists of an inverted region that runs under the poly of gate 1 in the *p*-tub, then along the inverted trench sidewall, then under the poly of gate 2 (note that for clarity, in Fig. 6-69 the path under the poly of gates 1 and 2 is drawn as running adjacent to the poly, whereas the path actually exists beneath the poly).

More severe trench-sidewall inversion problems are exhibited in *p*-wells than in *n*-wells. This is due to two interrelated phenomena. The first involves the dielectric (insulator) material used to fill the trench (typically SiO_2), and the resultant Si-SiO_2

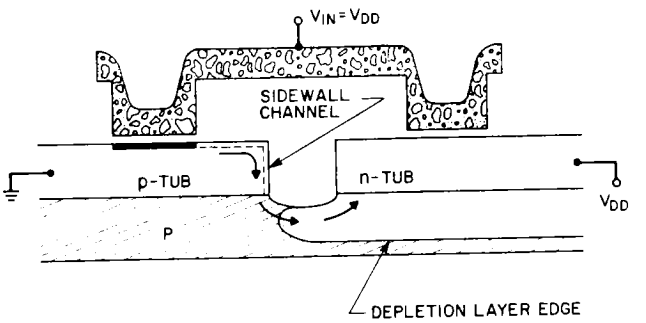


Fig. 6-68 A trench isolated *p*-well and *n*-well with a parasitic vertical sidewall MOSFET (with *n*-well as the gate) and a parasitic horizontal MOSFET (with the V_{in} poly line as the gate).[75] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission.

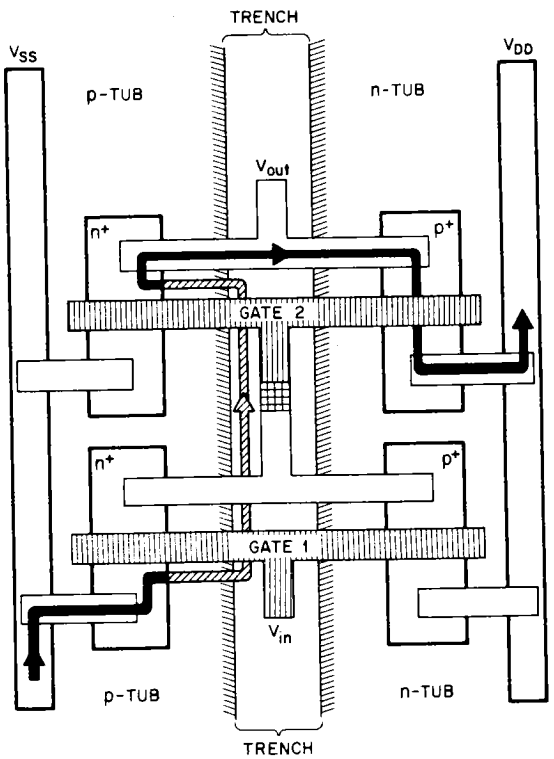


Fig. 6-69 A dc leakage path through two inverter stages in series. The solid black path is a good conducting route and the shaded path is a parasitic route. The parasitic consists of a field oxide corner under gate 1 (cf. Fig. 6-68), a sidewall path along the trench (cf. Fig. 6-67a), and a corner path under the V_{out} strap.[87] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission.

interface at the trench sidewall. The interface fixed or trapped charge that exists at an Si-SiO_2 interface is positive, which tends to lower the inversion threshold of *p*-channel regions (in this case the *p*-tub). To minimize this interface charge, a thin, good-quality thermal-oxide layer must be grown inside the trench. The oxide growth process, however, gives rise to the second problem. That is, during thermal oxidation some of the boron doping in the *p*-well segregates into the oxide. Thus, the boron surface concentration at the interface is reduced, further decreasing the inversion threshold. As a result, inversion of *p*-well trench sidewalls occur at lower threshold voltages than do those of an *n*-well for comparable well-doping concentrations. Finally, note that trench-etch processes often damage the Si surface in a manner that leads to increased interface trapped charge. This tends to worsen the problem of trench sidewall inversion in *p*-wells even more.

To counteract intra-well and inter-well leakage in trench-isolated CMOS several steps are necessary. First, the surface isolation path must be suppressed, by use of sufficient

channel stop-doping between the active device and tub edge (as is also done in LOCOS-isolated CMOS). Second, to avoid subthreshold leakage in single active devices, the devices are moved back somewhat from trench edge. Third, inversion of the trench sidewall should be prevented. One way to do this is to increase the trench width. However, this will increase the isolation spacing (as will moving the active devices away from the active device to the trench edge). Because of these limitations, the

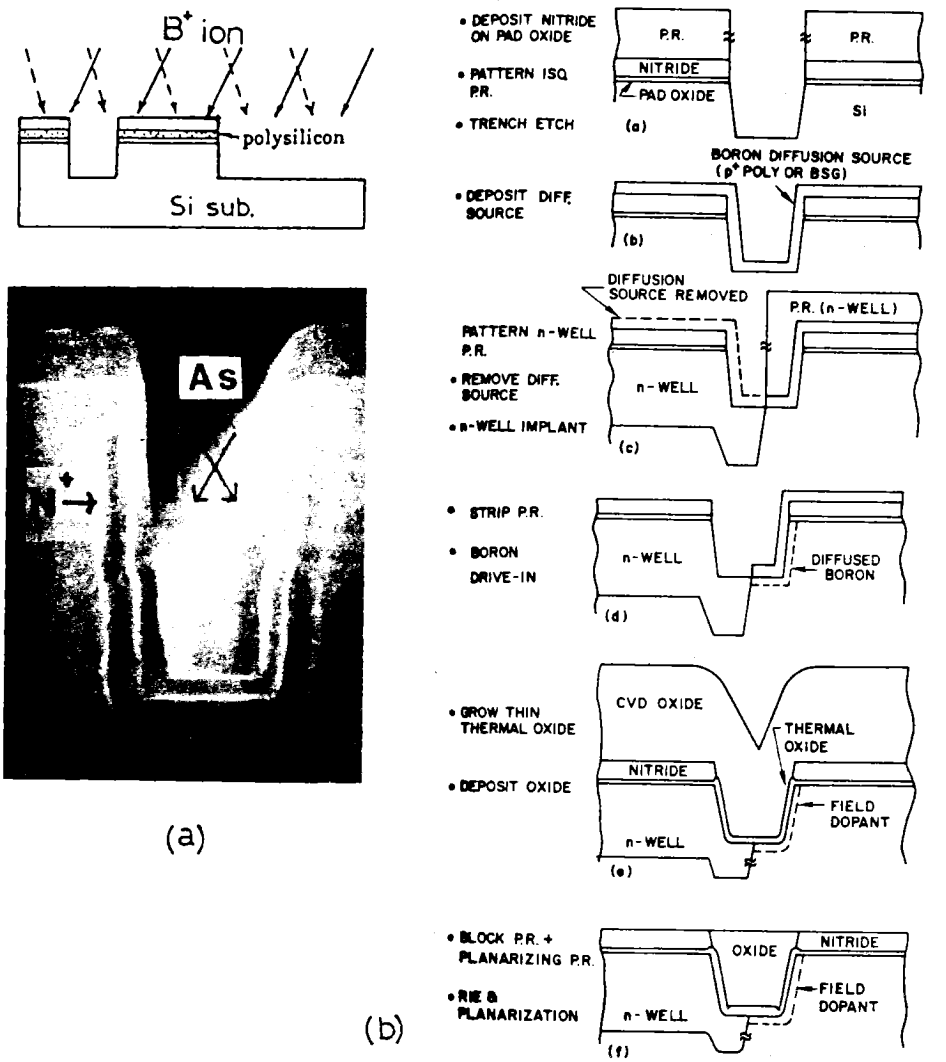


Fig. 6-70 (a) Schematic of oblique ion implantation and SEM cross section result of this procedure.[89] (© 1987 IEEE). (b) Shallow trench isolation process flow using a doped oxide as diffusion source of trench sidewall doping.[90] (© 1988 IEEE).

minimum n^+ to p^+ isolation distance in trench isolation in 1992 was still approximately 2-2.5 μm.

Another way to prevent inversion of the trench sidewalls is to increase the doping at the sidewall surface. Use of a direct, low-tilt-angle implant (e.g., boron, at an angle of 8° off the perpendicular to the top wafer surface, or at an angle of 82° off the perpendicular to the trench walls) is one reported way to do this (Fig. 6-70a).[89] However, since only one of the walls is doped by one implantation, the implant must be done four times, with the wafer rotated 90° after each step to ensure that all four walls of the trench are doped. In addition, this procedure is only effective for intra-well isolation (i.e., to isolate like-types of devices).

An alternative method for doping the sidewall of the p-well without doping the n-well with Si (thus making it suitable for CMOS inter-well isolation), was proposed by Davari et al.[90] In this approach, the p-well sidewall dopant is provided by boron diffusion from p^+ poly or borosilicate glass. The n-well does not require additional sidewall dopant as phosphorus piles up during the brief thermal oxidation of the trench prior to CVD oxide refill.

First a nitride/pad oxide mask stack is formed and patterned. Next, after shallow trenches are etched by RIE to a depth of 0.3-0.5 μm, a boron diffusion source (*in situ* doped p^+ poly or borosilicate glass, BSG) is deposited. A lithography step is then used to etch the poly of BSG away from the n-well sidewall and trench bottom (see Fig. 6-70b). A thermal cycle is then carried out to drive the boron into the sidewall and bottom of the p-well trench. No boron is diffused into the active device areas since they are covered with nitride. The boron source is then etched away, and the remainder of the trench refill process is conducted. The boron concentration at the sidewall surface is controlled by the segregation during the brief thermal 50-nm-thick oxide growth step. This approach has been used for trench widths as small as 0.8 μm. For 0.5 μm widths, control of sidewall doping and geometry is difficult, and hence for such dimensions control of sidewall parasitic MOSFETs represents a major obstacle to miniaturizing isolation.

These doping techniques have been used for ~0.8 μm wide trenches. For 0.5 μm widths, control of the sidewall doping and geometry effects is difficult.[73] No matter what technique is used, however, sidewall doping impacts the active devices through increased junction capacitance and source-substrate bias dependence. Control of sidewall parasitic MOSFETs thus represents a significant challenge for scaling CMOS isolation.

Trench-isolated 0.5-μm CMOS using shallow trenches filled with void- and seam-free TEOS/Ozone films has been implemented.[126] The 0.5 n- and p-channel MOSFETs showed no subthreshold degradation compared to the same device fabricated with LOCOS isolation. Field punchthrough voltages on intra-well and inter-well isolation structures were significantly higher than the control devices, which used PBL isolation (see Fig. 6-71). That is, punchthrough voltages were well above 10V and independent of trench isolation width.

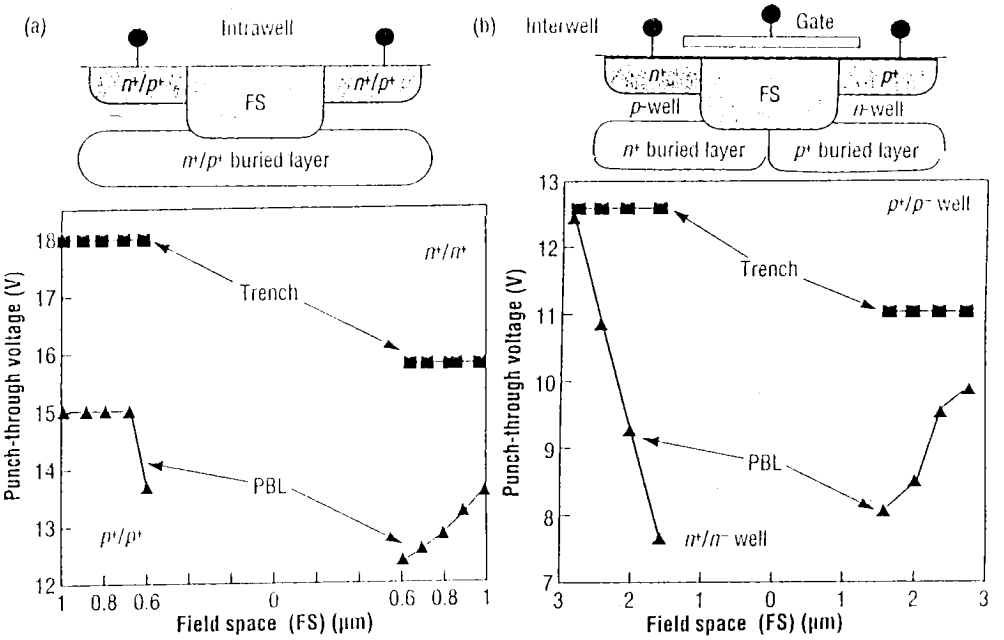


Fig. 6-71 Field punch-through voltages for trench isolation and PBL isolation. Consistently higher punchthrough voltage is observed for trench isolation with no degradation of subthreshold 0.5 μm MOSFET characteristics.[126] Reprinted with permission of Solid State Technology.

REFERENCES

1. J. Y. Chen, CMOS Science and Technology, Prentice-Hall, Englewood Cliffs, N. J., 1990, p. 237.
2. K. L. Wang et al., IEEE Trans. Electron Dev. ED-29, p. 541, 1982.
3. A. G. Lewis, R. A. Martin, and J. Y. Chen, "Retrograde and Conventional n-well CMOS technologies: A comparison," Internal Report, EIL-87-17, Xerox Palo Alto Research Center, October, 1986.
4. J.W. Lutze and J.P. Krusius, "Electrical Limitations of Advanced LOCOS Isolation of Deep Submicron CMOS," *IEEE Trans. Electron Dev.*, **ED-38**(2), p. 242, Feb. 1991.
5. J.A. Appels et al., *Phillips Research Rep.*, **25**, 118, (1970).
6. E. Kooi, *The Invention of LOCOS* IEEE Press, 1991, New York.
7. R. C. Jaeger, *Introduction to Microelectronic Fabrication*, 1988, Addison-Wesley, Reading, MA.
8. S. M. Sze, Ed., *VLSI Technology*, 2nd Ed., 1988 McGraw-Hill, New York.
9. D.R. Craven and J.B. Stimmel, *Semiconductor International,* June 1981, p. 59.
10. Cham et al., *Computer-Aided Design and VLSI Device Development* Kluwer Academic Publishers, Boston 1986.

11. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.
12. B. Cochran, "Tradeoffs in Submicron CMOS Process Design," *Semiconductor International,* May 1991, p. 146.
13. J.R. Pfiester and J.R. Alvis, "A Novel CMOS VLSI Isolation Using Ge/B Ion Implantation," *IEEE Electron Dev. Letts.*, August 1988, p. 391.
14. J.R. Pfiester and J.R. Alvis, "Improved CMOS Field Isolation Technology Using Selective Cl Implantation," *IEEE Electron Dev. Letts.*, November 1988, p. 561.
15. P. A. Van den Plas et al., *Field Isolation Process for Submicron CMOS*, Digest Symposium on VLSI Technology, Karuizawa, Japan, p. 19-20, 1987.
16. J. Manoliu, *Semiconductor Internat.*, April 1988, p. 90.
17. J. Hui et al., "Scaling Limitations of Submicron LOCOS Technology," *Tech. Dig. IEDM,* 1985, p. 392.
18. B. Coulman et al., *Ext. Abs. Electrochem. Soc. Meeting,* Spring 1989, p. 293.
19. J. W. Lutze, H. Perera, and J.P. Krusius, J. ECS, Vol. 137, No. 6, June 1990, p. 1867.
20. J. Hui et al., *IEEE Trans. Electron Dev.* **ED-29**, p. 554, 1982.
21. E. Kooi and J.A. Appels, "Selective Oxidation of Silicon and its Device Applications," *Semiconductor Silicon,* Electrochemical Society, 1973., p. 860.
22. A.E.T. Kuiper et al., "Oxidation Behavior of LPCVD Oxynitride Films," *Applied Surface Science* 33/34, p. 757, 1988.
23. P.B. Griffin and C.S. Rafferty, "A Viscous Nitride Model for Nitride/Oxide Isolation Structures," *IEDM Tech. Dig.*, p. 741, 1990.
24. R. A. Chapman et al., "High Performance Sub-Half Micron CMOS Using RTP," *IEDM Tech. Dig. 1991,* p. 101.
25. L. M. Loewenstein and C.M. Tipton, "Chemical Etching of Thermally oxidized Silicon Nitride: Comparison of Wet and Dry Etching Methods," *J. Electrochem. Soc.*, **138** (5) p. 1389, May 1991.
26. E. Kooi et al., *J. Electrochem. Soc.*, **123,** 1117, (1976).
27. T.A. Shankoff, *J. Electrochem. Soc.*, **129,** 1066, (1982).
28. C.A. Goodwin and J.W. Brossman, *J. Electrochem. Soc.*, **123,** 1117, (1976).
29. H.H. Tseng and P.J. Tobin, "A Double Sacrificial Oxide Process for Smoother Gate Oxide Interfaces," *Extended Abs. Meeting Electro. Chem. Soc.*, Spring, 1992, p. 392.
30. N. Guillemot et al., *IEEE Trans. Electron Dev.*, May 1987, p. 1033.
31. D.J. Chin et al., *IEEE Trans. Electron Dev.*, **ED-30**, July 1983, p. 993.
32. E.P. EerNisse, "Stress in Thermal SiO_2 during growth," *Appl. Phys. Lett.*, , **35** (1) 1 July, 1979.
33. D. Kao et al., *Tech. Dig. IEDM,* 1985, p. 388.
34. D.-B. Kao et al., "Two-Dimensional Thermal Oxidation of Silicon-II. Modeling Stress Effects in Wet Oxides," *IEEE Trans. Electron Dev.*, **ED-35**, January 1988, p. 25.
35. P. Sutardja, W.G. Oldham and D.B. Kao, *Tech. Dig. IEDM*, 1987, p. 264.
36. P. Sutardja, Y. Shacham-Diamond, and W.G. Oldham, *Tech. Dig. IEDM*, 1986, p. 526.
37. S. Isomae et al., "Oxidation-Induced Stress in a LOOCS structure," *IEEE Electron Dev. Letts.*, **EDL-7**, June 1986 p. 368.

38. S. Odanaka et al., "SMART-P: Rigorous three-dimensional process simulator on a supercomputer," *IEEE Tans. Computer-Aided Design,* p. 675, June 1988.

39. H. Unimoto and S. Odanaka, "Three-Dimensional Process Simulation of Local Oxidation of Silicon," *IEEE Tans. Electron Dev.,* **ED-38**, p. 505, March 1991.

40. R. L. Hodges, et al., *Ext. Abs. Meeting Electrochem. Soc.,* Spring 1991, p. 625.

41. R. Chapman et al., "0.8 μm CMOS Technology" *IEDM Tech. Dig.* 1987, p. 362.

42. Y. Han and B. Ma "Poly Buffer Layer for Scaled MOS", *VLSI Science &Technology* Electrochem. Soc., 1984, p. 334.

43. N. Hoshi et al., *Tech. Dig. IEDM*, 1986, p. 300.

44. N.A.H. Wils, P. A. Van den Plas, and A. H. Montree, "Dimensional Characterization of Poly Buffer LOCOS in comparison with Suppressed LOCOS," *Digest of ESSDERC,* Nottingham, England, p. 535, 1990.

45. J.M. Sung et al., "Crab-Eye Formation in Poly-Buffer LOCOS Isolation Technology," *J. Electrochem. Soc.,* **137,** p. 2350, 1990.

46. M. Ghezzo et al., *J. Electrochemical Soc.*, July 1989, p. 1992.

47. R. L. Guldi, et al., *J. Electrochemical Soc.*, December 1989, p. 3815.

48. J. M. Sung, et al., *IEEE Trans. Electron Dev.,* Vol. 38, p. 1970, Aug. 1991.

49. T. Lin, N. Tsai, and C. Yoo, "Twin-White-Ribbon Effect and Pit Formation Mechanism in PBLOCOS," *J. Electrochem. Soc.,* **138**(7), p. 2145, July, 1991.

50. H.S. Yang et al., "Poly-void formation in PBL Process," *Ext. Abs. Electrochem. Soc. Mtg.* Spring 1992, p. 442.

51. T. Nishihara et al., "A 0.5 μm Isolation Technology Using Advanced Poly Silicon Pad LOCOS (APPL)," *Tech. Dig. IEDM,* 1988, p. 100.

52. B.-Y. Nguyen, et al., *Ext. Abs. Electrochem. Soc. Conf.*, Spring, 1991, p. 614.

53. J. D. Hayden et al., *IEEE Trans. Electron Dev.*, **ED-38**, p. 876, April 1991.

54. J. M. Sung, et al., *Electron Dev. Letts.*, November 1990, p. 549.

55. N. Shimizu et al., "A Poly-Buffer Recessed LOCOS Process for 256 Mbit DRAM Cells," *Tech. Dig. IEDM,* 1992, p. 279.

56. S.S. Roth et al., "Polysilicon Encapsulated Local Oxidation," *IEEE Electron Dev. Letts.*, p. 92, March 1991.

57. S.S. Roth et al., *IEEE Trans. Electron Dev.,* **ED-39**, May 1992, p. 1085.

58. K.J. Cooper et al., "Recessed Polysilicon Encapsulated Local Oxidation," *IEEE Electron Dev. Letts.*, p. 515, October 1991.

59. S.W. Sun and K.C. Weng, "Effects of field-oxide volume ratio and interface-charge density on submicron modified LOCOS isolations," *Solid-State Electronics.,* vol. 32, April 1989, p. 333.

60. S.S. Roth et al., "PELOX Integrated PBL," *IEEE Trans. Semi. Manf.* Vol. 6, August, 1993, p. 246.

61. S.-P. Tay, S. Grant, and I. Wylie, "Submicron MOS Device Isolation Using HiPOX-PBL Technology," *Ext. Abs. Electrochem. Soc. Mtg.,* Ext. Abs. #287, Fall 1992, p. 410.

62. R.L. Hodges, F.T. Liou, and S.S. Kim, "Reduction of Oxide Thinning and Bird's Beak for PBL Isolation Using HIPOX," *Ext. Abs. Electrochem. Soc. Mtg.* Spring 1992, p. 421.

63. S. Deleonibus and S.S. Kim, "High Pressure Oxidation: The Key for Poly-Buffer LOCOS Scalability towards the Gigabit Density," *Ext. Abs. Electrochem. Soc. Mtg.* Fall 1993, p. 281.

64. E. DuBois et al., "A Study of the Electrical Performances of Isolation Structures," *IEEE Trans. Electron Dev.* **ED-37**(6), p. 1477, June 1990.

65. T. Shibata et al, *IEDM Tech. Dig.* 1983, p. 27.

66. T. Iizuka et al., "Double Threshold MOSFETs in Bird's-Beak Free Structures," *Tech. Dig. IEDM* 1981, p. 384.

67. D. C. Cole et al., "The Use of Simulation in Semiconductor Technology Development," *Solid-State Electronics*, Vol.33, No. 6, p. 591, 199.

68. D. Foty., J. Mandelman and T. Furukawa, "Behavior of an NMOS Trench-Isolated Corner Parasitic Device at Low Temperature," *Electrochemical Soc. Mtg. Fall, 1989,* p. 295.

69. T. Furukawa and J. Mandelman, "Process and Device Simulation of Trench Isolation Corner Parasitic Device," *Electrochemical Soc. Mtg. Fall, 1988,* p. 329.

70. K. Shibahara et al., "Trench Isolation with ∇-Shaped Buried Oxide for 256-Mbit DRAMs," *Tech Dig. IEDM*, 1992, p. 275.

71. G. Fuse et al., *Tech. Dig. IEDM* 1987, p. 732.

72. G. Smolinsky et al., "Material Properties of Spin-on Silicon Oxide (SOX) for Fully Recessed NMOS Field Isolation," *J. Electrochem. Soc.*, **137**(1) p. 229, January 1990.

73. K. Ohe et al., "Narrow-Width Effects of Shallow-Trench-Isolated CMOS with n^+ Polysilicon Gate," *IEEE Trans. Electron Dev.* **ED-36** June 1989, p. 1110.

74. J.Y. Chen and D.E. Snyder, "Modeling Device Isolation in High-Density CMOS," *IEEE Electron Dev. Letts.*, February 1986, p. 64.

75. J.R. Brews, "The Submicron MOSFET," in *High-Speed Semiconductor Devices,* Ed. S.M. Sze, John Wiley & Sons, New York, 1990, p. 183.

76. R. Chwang and K. Yu, "CHMOS - An *n*-Well Bulk CMOS Technology for VLSI," *VLSI Design,* Fourth Quarter, 1981, p. 42.

77. L.C. Parrillo et al., "Twin-Tub CMOS," *IEDM Tech Dig.,* 1980, p. 752.

78. S.J. Hillenius et al., "A Symmetric Submicron CMOS Technology," *Tech. Dig. IEDM,* 1986, p. 252.

79. M.-L. Chen et al., *Tech. Dig. IEDM,* 1986, p. 256.

80. L.C. Parrillo, G.W. Reutlinger, and L.K. Wang, U.S. Patent No. 4,435,895, March 31, 1984.

81. S.J. Hillenius and L.C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.

82. A.G. Lewis et al., *IEEE Trans. Electron Dev.,* June 1989, p. 1777.

83. J.Y. Chen, "CMOS-The Emerging Technology," *IEEE Circuits and Devices Magazine,* March 1986, p. 16.

84. J.D. Hayden et al., "A High-Performance Half-Micron Generation CMOS Technology for Fast SRAMS," *IEEE Trans. Electron Dev.*, **ED-38**(4), p. 876, April 1991.

85. Y. Nitsu et al., "Latchup Free CMOS Structure Using Trench Isolated CMOS Process," *Tech. Dig. IEDM 1985, p. 509.*

86. K.M. Cham and S.-Y. Chiang, "Study of Trench Surface Inversion Problem for the Trench Isolation Technology," *IEEE Electron Dev. Letts.,* Sept. 1983, p. 303.

87. *ibid.* ref. 75, p. 185.

88. K. Cham et al., *Tech. Dig. IEDM 1986, p. 296.*

89. G. Fuse, "A New Isolation Method with Boron-Implanted Sidewalls for Controlling Narrow-Width Effect," *IEEE Trans. Electron Dev.*, **ED-34**(2) p. 356, Feb. 1987.

90. B. Davari et al., "A Variable-Size Shallow trench Isolation Technology with Diffused Sidewall Doping for Submicron CMOS," *IEDM Tech. Dig.*, 1988 p. 92.

91. A. Lewis, et al., "Device Isolation in High Density LOCOS-Isolated CMOS,"*IEEE Trans. Electron Dev.*, Vol. ED-34, p. 1337 June 1987.

92. J. Y. Chen and A. G. Lewis, "Parasitic transistor effects in CMOS VLSI," *IEEE Circuits and Devices Magazine,* Vol. 4, p. 8, 1988.

93. A. G. Lewis et al., "Vertical Isolation in Shallow *n*-wWell CMOS Circuits," *IEEE Electron Dev. Lett.*, EDL-8, p. 107, 1987.

94. A.T. Wu et al., *IEDM Tech. Dig.*, 1989, p. 271.

95. J. Manoliu and J.O. Borland, "A Submicron Buried Layer Twin-Well for High Density CMOS," *Tech. Dig. IEDM,* 1987, p. 20.

96. L.C. Parrillo et al., "Twin-Tub CMOS-II," *Tech. Dig. IEDM,* 1982, p. 706.

97. M.-L. Chen et al. *Tech. Dig. IEDM* 1986, p. 256.

98. S.J. Hillenius and L.C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.

99. Ch. Zeller, C. Mazure, and M. Kerber, "Field-Implant-Free Isolation by Double-Well Split Drive-In," *IEEE Electron. Dev. Letts.*, p. 215, May 1990.

100. J.Y. Chen,"CMOS - The Emerging Technology," *IEEE Circuits and Devices Magazine,* March 1986, p. 16.

101. D.M. Brown et al., "Trends in Advanced CMOS PRocess Technology," *Proceedings of the IEEE,* December 1986, p. 1646.

102. R.D. Rung, C.J. Dell'Oca, and L.G. Walker, "A Retrograde *p*-well for High Density CMOS," *IEEE Trans. Electron Dev.*, ED-28, p. 1115, 1981.

103. R. Martin and J.Y. Chen, "Optimized Retrograde *n*-well for 1 μm CMOS, *Proc. Custom Integ. Circuits Conf.,* 1985, p. 199.

104. S.R. Coombs, "Scalable Retrograde *p*-well CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-28**, p. 346, 1981.

105. Y. Taur et al., *IEEE J. Solid-State Circuits,* **SC-20,** p. 123, 1985.

106. G.J. Hu et al., *IEDM Tech. Dig.* 1982, p. 710.

107. Y. Mochizuki, "Special Report: High Energy Ion Implantation Technology," Nikkei Microelectronics, (Dec. 1991), p. 92.

108. M. Moslehi and K. Saraswat, "Thermal Nitridation of Si and SiO_2 for VLSI," *IEEE Trans Electron Dev.* **ED-32**, p. 116, January 1985.

109. P. Fahey, R.W. Dutton, and M. Moslehi, "Effects of Thermal Nitridation Processes on Boron and Phosphorus Diffusion in <100> Silicon," *Appl. Phys. Letts.*, **43**(7), p. 683 July, 1983.)

110. R. de Werdt et al., *Tech. Dig. IEDM,* 1987, p. 532.

111. K. Tsukamoto et al., *Nucl. Instrumentation & Methods,* B59/60, p. 584 (1991).

112. K. Tsukamoto et al., "High Energy Ion Implantation for ULSI: Well Engineering and Gettering," *Solid State Technology* P. 49, June 1992.

113. K. Kurosawa, T. Shibata, and H. Iizuka, "A new bird's beak free field isolation technology for VLSI devices," *IEDM Tech. Dig.* 1981, p. 384.

114. A. Stolmeijer, "A twin-well CMOS process employing high-energy ion implantation," *IEEE Trans. Electron Dev.*, **ED-33**, April, 1986, p. 450.

115. F.-S. Lai et al., *IEEE Trans. Electron Dev.*, **ED-33**, p. 1308, 1986.

116. R.A. Martin et al., *Tech. Dig. IEDM,* 1985, p. 403.

117. A.G. Lewis and S.L. Partridge, *IEEE Trans. Electron Dev.*, **ED-30**, p. 1680, 1983.

118. J.O. Borland and R. Koelsch, "MeV implantation technology: Next generation manufacturing with c rrent generation equipment," *Solid-State Technology*, December 1993, p. 28.

119. P.V. Gilbert et al., "Latch-Up Performance of a Sub-0.5 Micron Inter-well Deep Trench Technology," *Tech Dig. IEDM,* 1993, p. 731.

120. H. Inowaka et al., "Ultranarrow Trench-isolated 0.2 μm CMOS and its Application to Ultralow-power Frequency Dividers," *Tech Dig. IEDM,* 1993, p. 887.

121. P.C. Fazan and V.K. Mathews, "A Highly Manufacturable Trench Isolation Process for Deep Submicron DRAMs," *Tech Dig. IEDM,* 1993, p. 57.

122. P. U. Kenkare et al., "Sensitivity of Field Isolation Profiles to Active Pattern," *Tech. Dig. IEDM,* 1993, p. 479.

123. T. Uchida et al., "A Simulation of Plastic Deformation of Silicon During Thermal Oxidation," *Tech. Dig. IEDM,* 1993, p. 315.

124. P.E. Riley et al., "Reactor Characterization for a Process to Etch Si_3N_4 Formed on Thin SiO_2," *IEEE Trans. Semi. Manuf.*, Vol. 6, August 1993, p. 290.

125. P.K. Kenkare et al., "Scaling of Poly-Encapsulated LOCOS for 0.35 μm CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-41**, January 1994, p. 56.

126. H. Wallace et al., "Applications of APCVD TEOS/O_3 Thin Films in ULSI IC Fabrication," *Solid-State Technology,* March 1994, p. 31.

127. S. Deleonibus et al., "Optimization of a shallow trench isolation refill process for high density non-volatile memmories using 100% chemical mechanical polishing," *Ext. Abs. of the Electrochemical Soc. Mtg.*, Spring 1994, p. 267.

128. M.A. Chonko, "Gate Dielectric Issues in 0.5μm Processing," *Ext. Abs. of the Electrochemical Soc. Mtg.*, Spring 1994, p. 256.

129. K.C. Brown et al., "Poly-Buffered LOCOS Isolation Pattern Definition," *Ext. Abs. of the Electrochemical Soc. Mtg.*, Spring 1994, p. 291.

130. S. Deleonibus, "The thickness and temperature dependence of field oxidation three-dimensional behavior: Example on Poly-Buffer LOCOS Isolation," *Ext. Abs. of the Electrochemical Soc. Mtg.*, Spring 1994, p. 273.

131. D. M. Rossman et al., Physical Characterization of Poly-Buffered LOCOS," *Ext. Abs. of the Electrochemical Soc. Mtg.*, Spring 1994, p. 295.

132. P.U. Kenkare et al., "Scaling of Poly-Encapsulated LOCOS for 0.35 μm CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-41**, January 1994, p. 56.

133. P.U. Kenkare and J.R. Pfiester, "Extending LOCOS-based Approaches for Sub-0.5 μm CMOS," *Semiconductor International,* May 1994, p. 64.

134. J.R. Pfiester et al., "Nitride-Clad LOCOS Isolation for 0.25 μm CMOS," *Digest of 1993 Symp. of VLSI Tech.*, 1993, p. 139.

135. E. Bassous, H.N. Yu, and V. Maniscalco, "Topology of Silicon Structures with Recessed SiO_2," *J. Electrochem. Soc.* **123**, p. 1729, 1976.

136. M.E. Law, C. Rafferty, and R.W. Dutton, *SUPREM IV User's Manuel,* Technical Report, Integrated Circuits Laboratory, Department of Electrical Engineering, Stanford University, July, 1986.

137. K.M. Cham et al., "Characterization and modelling of the trench surface inversion problem for the trench isolated CMOS technology," *Tech. Dig., IEDM* 1983, p. 23.

138. H.-H. Tseng and P.J. Tobin, "Thin CVD Stacked Gate Dielectric for ULSI Technology," *Tech. Dig. IEDM,* 1993, p. 321.

139. T.T. Sheng et al., "From "White Ribbon' to 'Black Belt': A Direct Observation of the Kooi Effect Masking Film by Transmission Electron Microscopy," *J. Electrochem. Soc.,* **140**, p. L163, November 1993.

140. A. Bryant et al., "The Current-Carrying Corner Inherent to Trench Isolation," *IEEE Electron Dev. Letts.,* **EDL-14**, August 1993, p. 412.

141. N. Shigyo et al., "Three-Dimensional Analysis of Subthreshold Swing and Transconductance for Fully Recessed Oxide (Trench) Isolated 1/4-μ m Width MOSFETs," *IEEE Transs. Electron Dev.*, **ED-35**, July 1988, p. 944.

PROBLEMS

1. Two n^+ diffused lines are running parallel in a substrate doped with 10^{15} boron atoms/cm^3. The substrate is biased to -5V, and both lines are connected to +5V. Using one-dimensional junction theory, calculate the minimum spacing needed between the lines to prevent their depletion regions from merging.

2. Explain why bird's beak and boron encroachment limit the maximum dimension in which conventional LOCOS isolation can be used.

3. Discuss each of the following drawbacks of LOCOS isolation technology:
a) bird's beak encroachment
b) boron encroachment;
c) oxide thinning in narrow field oxide spacings
d) non-planarity
e) process complexity
f) sensitivity of bird's beak encroachment on pattern geometry.

4. Explain how the frame-masked PBL (FMPBL) and the polysilicon-encapsulated LOCOS (PELOX) processes achieve a smaller bird's beak penetration length the conventional PBL.

5. Assume the field oxide beneath an n+ polysilicon line is 600 nm thick, and the p-type channel stop beneath this oxide is doped to $1x10^{17}$/cm^3. Calculate V_{TF} and S_t if Q_{tot} = $1x10^{11}$ cm^{-2}.

6. If the nominal V_{TF} (i.e., measured when I_{DSt} = 1 μA) of an aluminum gate field oxide test device is 8V, what is the leakage current of this test structure when the aluminum line at 5V (assuming that S_t = 0.5V/dec at T = 300K)? If V_{TF} decreases to 6.5V when T is increased to 125°C, what is the leakage current in this case when the aluminum line is biased to 5V.

7. Based on Figs. 6-4 and 6-5 obtain the subthreshold swing S_t for the parasitic n- and p -channel FETs (both Al- and polysilicon gate devices)with a 2 μm separation between two n^+ and two p^+ diffusion regions.

8. If the circuit application requires that no more than 10 pA leakage flows between the $n+$ and n-well in a CMOS technology, and we have obtained the process and device data from Fig. 6-57a. What design rule should be specified to provide for the minimum n^+-to-n-well separation?

CHAPTER 7

THIN GATE OXIDES:

GROWTH and RELIABILITY

Gate oxide thickness has grown smaller with each generation of MOS ICs. Table 7-1 shows the evolution of MOSFET scaling from 1977 to 1991.[1] As the gate length shrank from 3.0 μm in 1977 to 0.6 μm in 1991, the gate oxide thickness went from 70 nm to 13.5-12.5 nm. During the shrinkage from 3.0 μm to 0.6 μm technologies, the power-supply voltage remained at 5V. At gate lengths below 0.5 μm, however, it is projected that the power supply voltage will have to be reduced. In fact, a reduction to 3.3V in some 0.5 μm CMOS technologies has already been implemented. The decrease in power-supply voltage is needed because the gate oxide thickness in such deep-submicron MOSFETs is 10 nm or less. As will be discussed, the oxide electric fields produced by 5V cause severe reliability problems in oxide films thinner than 10 nm.

The thickness of gate oxides is made as small as possible because the drive current in MOSFETs increases with decreasing gate oxide thickness. Since this trend extends to gate oxide thicknesses below 5 nm, it seems apparent that even thinner oxides will be specified as MOSFETs are scaled below 0.5 μm (see Table 5-3 in Section 5.6.1).

Table 7-1 Evolution of MOSFET Technology Since 1977[1]

Year of Introduction	1977	1979	1982	1985	1988	1991
Gate Length (μm)	3	2	1.5	1.1	0.9	0.6
Minimum Feature Size (μm)	3	2	1.5	1	0.7	0.5
V_{CC} (V)	5	5	5	5	5	5
MOSFET Gate Oxide (nm)	70	40	25	25	20	13.5
Junction Depth (μm)	0.6	0.4	0.3	0.25	0.2	0.15
NMOS I_{Dsat} @ V_{GS} = 5V (mA/μm)	0.1	0.14	0.23	0.27	0.36	0.64
PMOS I_{Dsat} @ V_{GS} = 5V (mA/μm)		0.06	0.11	0.14	0.19	0.31
Gate Delay @ FO = 1 (ps)	800	350	250	200	160	90

Furthermore, thin oxides (and shallow junction depths) are needed to control short-chan -nel effects; i.e., thin gate oxides permit the gate to retain strong control of the channel charge, and shallow junctions keep the drain field from extending far into the channel.

Table 5-4 in chap. 5 gives an estimate of MOSFET technology scaling made by Hu in 1993.[1] As the gate length is decreased and approaches 0.1 μm it is predicted that gate oxides of only 4 nm thickness will be used (albeit at a power supply voltage of 2.2V). An example of a deep submicron CMOS technology (0.25 μm) is given by Chang et al. of IBM, in which t_{ox} = 7 nm and V_{DD} = 2.5V.[2] Other studies predict that even thinner oxides will be needed in memory technology than in logic CMOS technologies. For example, in 64 MB DRAMs the capacitor oxide thickness is estimated to be below 10 nm, and in EEPROMs tunneling oxides are even thinner than those used in DRAMs.

While the benefits of using thin oxides are apparent, there are two key issues which must be satisfied to permit them to be utilized: (1) it must be possible to grow such thin oxide layers precisely and uniformly across a wafer; and (2) MOSFETs with such thin oxides must exhibit adequate reliability characteristics under normal circuit operating conditions. These two issues are the subject of this chapter.

7.1 GATE OXIDE CHARACTERISTICS NEEDED FOR SUBMICRON MOSFETS

The desired characteristics of MOSFET gate oxides are the following:

1. The grown gate oxide thickness must closely match the thickness specification of the MOSFET device design.

2. The specified oxide thickness must also be sufficiently uniform across the entire wafer, and from wafer to wafer, and from run-to-run.

3. The gate oxide film and the Si/SiO_2 interface must exhibit adequately small values of charge in the oxide and at the $Si\text{-}SiO_2$ interface (i.e., low Q_f, D_{it}, Q_{ot} and Q_m values - see chap. 3).

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

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

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

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

From the list of desired gate oxide characteristics given above, we see that thin oxide reliability issues fall into two categories:

a. Dielectric breakdown (which causes catastrophic failure of device).

b. Hot-carrier-injection degradation (as a result of which device characteristics are degraded).

In this chapter we cover gate oxide breakdown, and in chapter 9 we describe device degradation due to hot-carrier injection. However, here we will also cover all the mechanisms by which charge is injected into the gate oxide (section 7.3), including injection by "hot" as well as by "cold" carriers.

7.2 PHYSICAL PICTURE of SiO_2 and the Si/SiO_2 INTERFACE

Before setting out to describe the technology of growing thin gate oxides and their reliability, it is useful to discuss the physical and chemical properties of thermally grown SiO_2 and the physical picture of the Si/SiO_2 interface. We cover these topics in this section.

7.2.1 Physical and Chemical Properties of SiO_2

SiO_2 (silica) can be found in crystalline, vitreous, or amorphous forms. When the atomic structure of silica exhibits long-range order it is said to be in a *crystalline* form. Many varieties of crystalline silica exist, including quartz, crystobalite, coesite, etc. However, the type encountered in silicon ICs (i.e., thermal silicon dioxide films grown under normal conditions) is silica in its *vitreous* (or *glassy*) form. Vitreous solids do not exhibit long range order. Instead, their structure is highly ordered only over short ranges. This glassy state of SiO_2 is also often referred to as *fused silica.* If no short-range order is exhibited by the silica its structure is termed *amorphous*. Amorphous SiO_2 films may be encountered in oxides deposited under poorly controlled conditions. Since we are interested in thermally grown SiO_2, most of our discussion will focus on the glassy state of SiO_2 (fused silica).

The basic structural unit of silica is centered around the *structural formula* (SiO_4). The spatial arrangement of Si and O atoms is due to their respective valence (+4 and -2), their relative sizes, and their electrostatic interactions. These factors give rise to elementary SiO_2 cells of tetrahedric configurations. That is, a silicon atom (with a valence of +4) is located at the center of the tetrahedron, with oxygen atoms (O^{2-}) at each of the corners (Fig. 7-1a).[3] In crystalline SiO_2 (quartz), each oxygen atom belongs to two tetrahedra and is thus bonded to 2 silicon atoms. Such oxygen atoms are then known as *bridging oxygen atoms*. (Fig. 7-1b).[3] In vitreous or amorphous SiO_2 some of the vertices of the tetrahedra have *nonbridging oxygen atoms*, meaning they are not shared between two tetrahedra (Fig. 7-1d). The greater the ratio of bridging to non-bridging oxygens, the greater the cohesiveness of the SiO_2 structure.

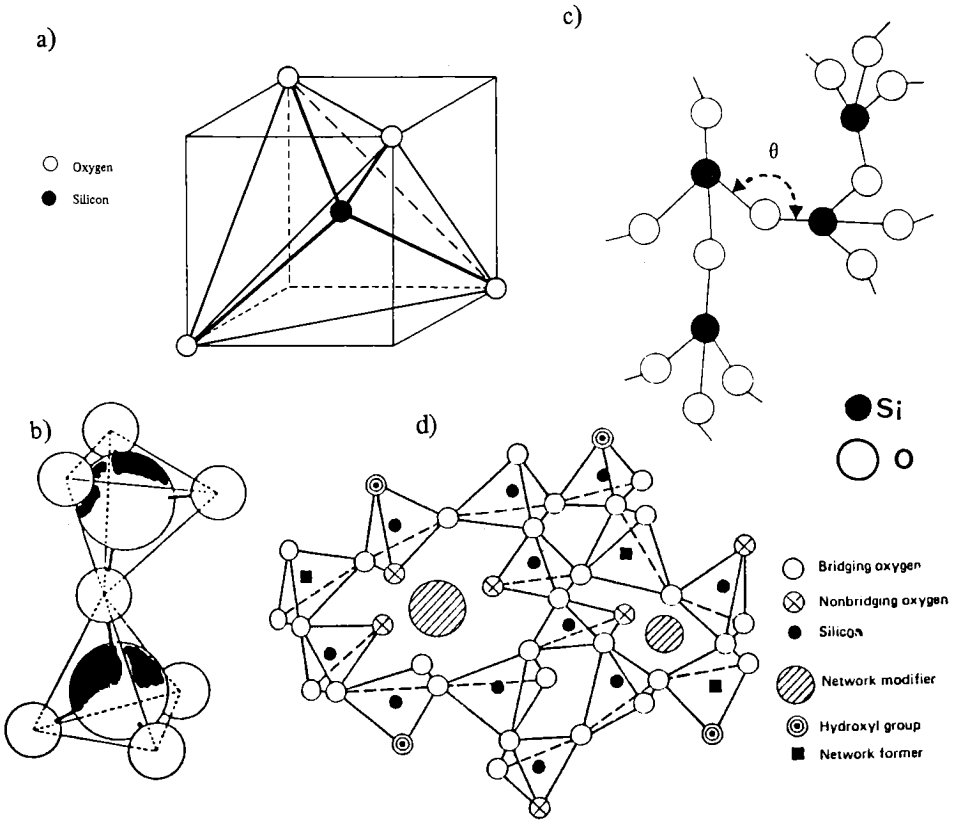


Fig. 7-1 (a) The basic structure of SiO_2.[18] (b) Three-dimensional representation of two neighboring SiO_4 cells, bridged by an oxygen atom. (c) Two-dimensional lattice representing vitreous SiO_2.[18] (d) The structure of thermally grown SiO_2 showing bridging and non-bridging oxygen atoms and dopant (i.e., network modifier) atoms.[132] (© IEEE 1965).

The interatomic distances (from the center of one atom to the center of its nearest neighbor) have been measured and found to have the following mean values: 1.62Å for the Si-O bond; 2.27Å for the O-O distance; and 3.12Å for the Si-Si distance. Note this implies that a 10 nm gate oxide is only 40-50 atomic layers thick!

The structure of fused silica is a continuous random network of tetrahedra, where the Si-O-Si bond angle varies from 110° to 180°, with the most probable value being 144°. A two-dimensional representation of the vitreous SiO_2 lattice is shown in Fig. 7-1c).[3]

The density of thermal SiO_2 (fused silica) is 2.20 g/cm^3 which is smaller than that of quartz (2.65 g/cm^3). In fused silica only 43% of the lattice space is occupied, making its structure much more open than that of quartz. Consequently, a large variety of impurity atoms can easily enter this oxide network and diffuse through it. Since it is

difficult to directly assess the density of thin films, the refractive index of SiO_2 is usually measured instead, and the density is then inferred from it. A refractive index of 1.460 corresponds to a density of 2.20 g/cm^3. Generally, CVD oxides exhibit smaller densities than thermally grown SiO_2, and worse electrical and material properties correspond to less dense films. Hence, measurements of the index of refraction also provide a method for rapidly comparing the characteristics of deposited oxides.

7.2.2 The Si/SiO_2 Interface

The characteristics of the Si/SiO_2 interface, as well as the SiO_2 bulk characteristics, play an important role in the functioning of the gate dielectric. Hence, a discussion of the structure of this interface is also useful. In chap. 3 we introduced the concept of surface states and interface trapped charge, and their impact on the threshold voltage of MOSFETs. Here we consider these topics in further detail. We recall that surface states are extra allowed energy states for electrons that are present at the semiconductor surface, but not within the bulk.

The term *surface states* was coined by Tamm in 1932.[4] Using quantum mechanical calculations he showed that new electronic energy states arise in a crystal if the lattice is terminated at a surface. Such new states are confined to the region very close to the surface. He also calculated that these states will arise at interfaces (such as the Si/SiO_2 interface), as well as at free surfaces. Each of the states is associated with a single atom at the surface. Hence, an electron occupying one of these states is localized (i.e., forced to remain in a restricted region of space centered on that atom). Since such states thus effectively trap free carriers at the surface, they are also referred to as *interface traps.* The charge per unit area stored in the traps is symbolized by Q_{it}.

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

From a qualitative perspective, the dangling bond model can be visualized by with the aid of Fig. 7-2a. Here it can be seen that if a Si lattice is abruptly terminated along a given plane to form a surface, one of the four bonds of each Si atom at the surface is left dangling. There will thus be extra states on this surface because the energy field of the crystal is one-sided (i.e., the electrons in the surface region are bonded only from the side directed toward the bulk). It is plausible that the thermal formation of SiO_2 can tie up these surface Si bonds, and that along a perfect Si/SiO_2 interface (Fig. 7-2b) all such bonds could be tied up. In this case, surface states at the interface would be suppressed.[5] However, it is also plausible that such oxidation might not tie up all the bonds (Fig. 7-2c). If even a very small fraction of the number of dangling bonds was left unsatisfied, a significant number of surface states could exist. For example, on a (100) Si surface there are 6.8×10^{14} Si atoms per cm^2. If oxidation left 1/1000 of these bonds dangling, and each of them gave rise to a surface state, the density of interface trapped charges would be 6.8×10^{11} cm^{-2} (assuming one electronic charge is associated with each energy state). If a gate oxide thickness of 20 nm was being used, this would

cause a threshold voltage shift of 0.63V. This indicates that if the dangling bond model correctly describes the origin of interface states, then only a relatively small number of residual dangling bonds can significantly perturb the device characteristics. Crystal defects near the surface or foreign atoms bonded at the surface are other interface perturbations that have been identified as possible sources of surface states.

As proposed by Deal,[6] surface states have the following properties:

1. A surface state can trap or emit a carrier.

2. The energy levels of these states are located within the forbidden gap.

3. The surface state can be of either type: donor or acceptor. The classification is made in a manner analogous to bulk dopant atoms. That is:

a. A monovalent *donor* trap possesses two charge states +1 and 0. Its charge is positive (+1) when the trap is empty (i.e., its energy level E_T is above E_F), but is neutral (0) when full (when E_T is below E_F).

b. A monovalent *acceptor* trap possesses two charge states 0 and -1. When the trap is empty its charge is neutral (0) (i.e., its

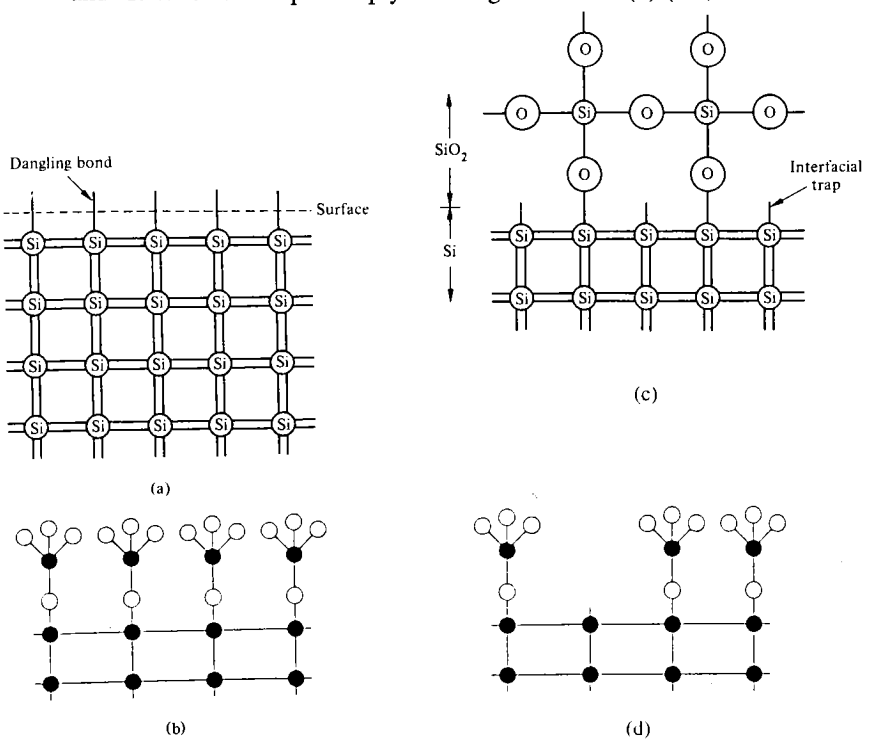


Fig. 7-2 Physical model for the interfacial traps. "Dangling bonds" which occur when the Si lattice is abruptly terminated along a given plane to form a surface. (b) Post oxidation perfect interface. (c & d) Post oxidation dangling bonds that become interfacial traps.[118] Reprinted with permission of the publisher, the Electrochemical Society.

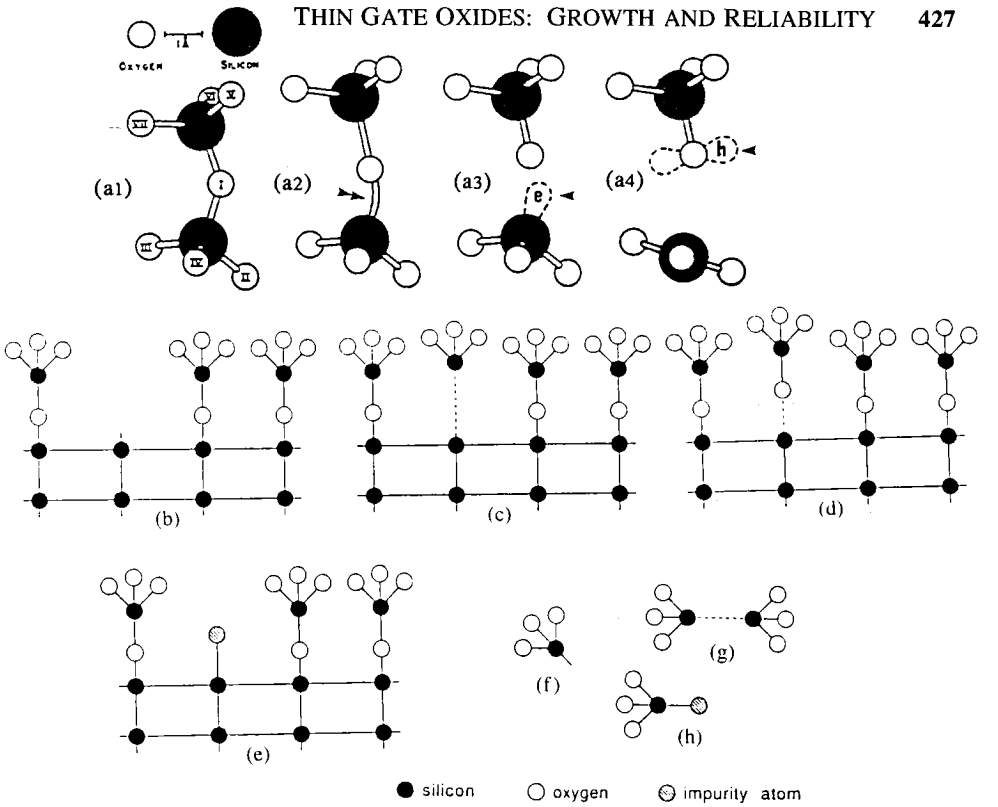


Fig. 7-3 (a) Strained and broken bonds: (a1) normal Si-O bonds; (a2) strained bond in vitreous silica ; (a3) broken bond represented with a traped electron; (a4) broken bond represented with a trapped hole. Possible interface defects. (b) Si dangling bonds. (c) Si-Si stretched bond (or oxygen vacancy). (d) Si-O stretched bond. (e) impurity at interface. Possible oxide bulk defects: (f) Si dangling bond; (g) Si-Si stretched bond (or oxygen vacancy). (h) impurity in SiO_2.[5] (© IEEE 1988).

energy level E_T is above E_F), but when the trap is full its charge is negative (-1) (i.e., when E_T is below E_F).

4. Because they are mostly due to a disruption of the periodicity of the crystal lattice (and only less frequently caused by the presence of foreign impurities), interface states are not associated with well-defined energy levels. Instead, interface states are distributed throughout the entire band gap at closely spaced levels.

Figure 7-3 represents the bond stretching and broken bond model of the Si/SiO_2 bulk of Verwey[18] and of the Si/SiO_2 interface of Sakurai and Sugano.[5] That is, Fig. 7-3a depicts the stretched and broken Si-O bonds in the bulk, and Figs. 7-3b-d at the interface and Figs. 7-3e-h the possible defects in the oxide bulk. In Fig. 7-3b, the dangling Si bond at the surface is depicted as an *interfacial trivalent Si atom,* while in Fig. 7-3f, a dangling Si bond in the bulk is portrayed as a *trivalent Si atom.* Trivalent Si atoms introduce a deep trap level near the Si midgap when located at the interface (see Fig. 7-4) and a deep trap level in the SiO_2 bandgap when located in the

oxide bulk (1 in Fig. 7-4). The weak Si-Si* and Si-O stretched bonds (Figs. 7-3c and 7-3d) at the interface introduce a continuum of deep trap levels, or if such stretched bonds exist in the bulk (Fig. 7-3g), they produce shallow trap levels. If a dangling Si bond is tied up with an impurity (most commonly H or OH) this is thought to produce an electron trap, while the oxygen vacancy (or weak Si-Si stretched bond) is considered to be the precursor of the hole trap. Thus, as we shall see, processes that decrease the electron trap densities tend to increase the hole trap densities.

An anneal in a hydrogen ambient (100% H_2 or 4%-H_2 forming gas) at approximately 450°C is normally the final step prior to assembly and packaging of the ICs. This step is thought to allow hydrogen to penetrate the gate oxide and tie up the remaining dangling bonds at the Si/SiO_2 interface not tied up by thermal oxidation (as suggested in Fig. 7-5). However, the Si-H or Si-OH bond can be easily broken by injected hot electrons, giving rise to interface traps. Thus, hydrogen introduced in the post-metal anneal or other process steps (i.e., during the steam reflow of an interlevel BPSG layer, or during the deposition of silicon nitride), can increase hot-electron degradation in a MOSFET. Consequently, it has been suggested that the ideal gate dielectric and all subsequent dielectric layers should contain as little hydrogen as possible to make them hot-electron resistant.

The defects in the oxide bulk and at the interface are believed to be responsible for:

1. Oxide integrity degradation (which may cause catastrophic failure).

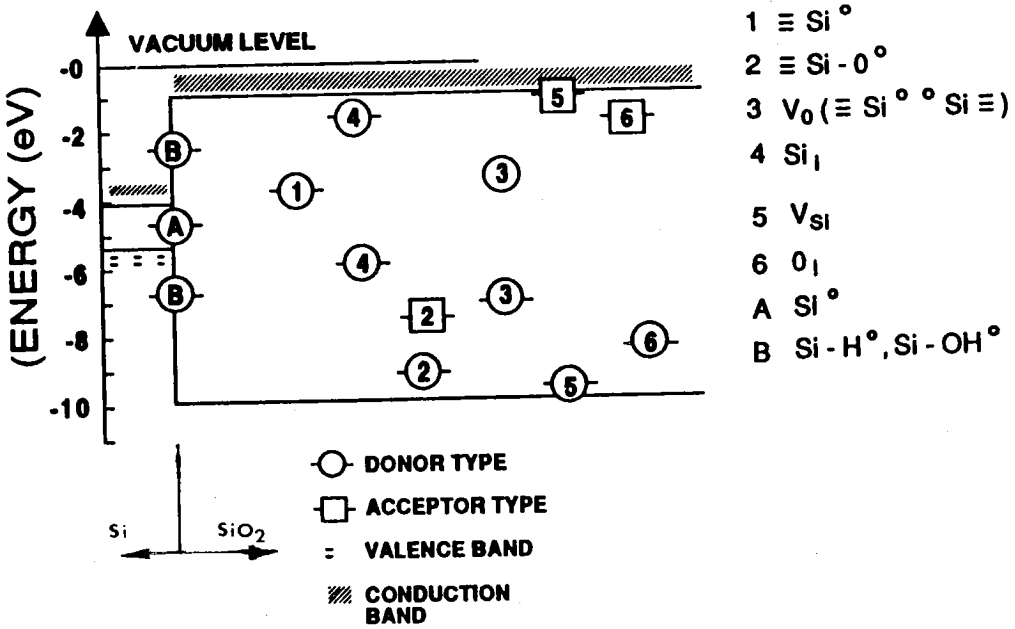


Fig. 7-4 Energy band diagram showing the energy levels of the various defects in the oxide-silicon interface and in the gate oxide.

* The weak Si-Si bond is more frequently referred to as an *oxygen vacancy*.

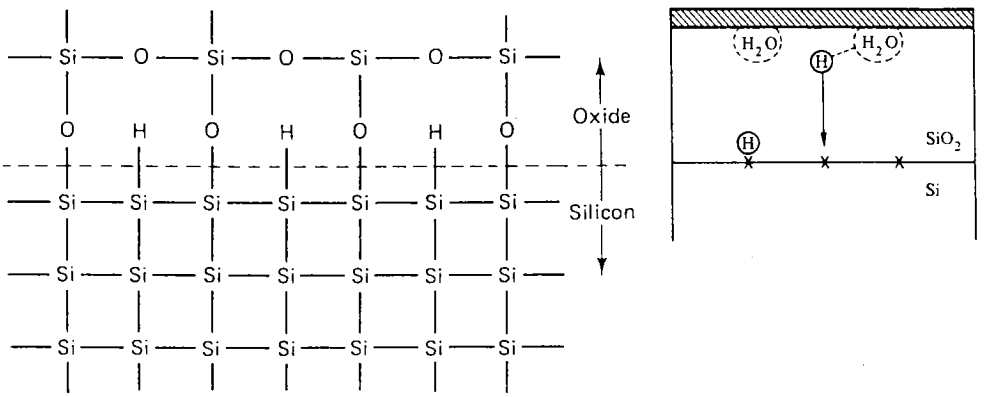


Fig. 7-5 How hydrogen can passivate the dangling bonds at the silicon surface shown in Fig. 7-2c and reduce the interface trap density. From D.L. Pulfrey and N. G. Tarr *Introduction to Microelectronic Devices*, p. 222. Copyright 1989. Prentice-Hall. Reprinted with permission.

2. Hot-carrier induced degradation, which causes some of the device operating characteristics to be degraded (e.g., V_T and g_m), but the device may still work.

7.3 CARRIER INJECTION IN THE Si/SiO_2 SYSTEM

Failures in MOSFETs associated with the gate oxide (both catastrophic failures and device-characteristic-degradation failures) usually involve carriers injected into the gate oxide. A variety of carrier-injection mechanisms have been proposed for the Si/SiO_2 system. Here we review the basic physics of the carrier injection phenomena in MOSFETs as useful background material for the discussion of gate oxide breakdown (this chapter), and hot-carrier degradation (chap. 9).

Carriers can be injected either over an energy barrier (hot-carrier injection), or through a barrier (tunneling by cold carriers). The most important injection phenomena are:

1. Channel hot electron (CHE) injection — Hot carrier injection phenomena
2. Drain avalanche hot carriers (DAHC) injection — "

and

3. Injection by Fowler-Nordheim tunneling — Cold carrier injection phenomena
4. Injection by direct tunneling. — "

Hot carrier injection implies that the injected carriers are no longer in equilibrium with the lattice at the point of injection (i.e., they possess far more kinetic energy than they would normally acquire from the ambient lattice temperature). They gain such energy

upon being accelerated by the high channel fields in the device. This is in contrast to tunneling injection phenomena, in which the carriers being injected are in equilibrium with the Si lattice (i.e., they are "cold") and do not become hot until after they are injected into the oxide and are accelerated by the oxide field to the gate electrode.

The energy-band diagram of the MOS structure, including the barrier heights for electrons and holes, is shown in Fig. 7-6. The oxide band gap is ~9 eV. The barrier that suppresses electron injection into the oxide (or *modified electron affinity*) is 3.1 eV, and the barrier for holes (*modified hole affinity*) is ~4.9 eV. Thus, the barrier to hole injection into the oxide is higher than that for electron injection.

7.3.1 Channel Hot Electrons (CHE)

Injection of channel hot electrons (CHE) occurs in MOSFETs biased to strong inversion. The electric field responsible for accelerating the carriers in this case is the lateral electric field. Ideally carriers flow only in the channel of the MOSFET since an energy barrier exists between the channel and the gate oxide (3.1 eV for electrons and 4.9 eV for holes). This barrier is important, since the electric field along the channel also has a vertical component. The vertical component pulls the carriers toward the gate, but the presence of the barrier suppresses the injection of the carriers into the oxide. Thus, the vertical field component merely acts to constrain the electrons to the silicon surface. However, if the carriers acquire sufficient energy from the lateral electric field and become 'hot', they can surmount this barrier and jump into the oxide conduction band. Since the barrier for electron injection is smaller, the probability of hot-electron injection is far greater than for hot-hole injection. Hence, the problem of hot-carrier injection is much worse in *n*-channel MOSFETs (in which electrons are the channel current charge carriers) than in PMOSFETs. Consequently, only *channel hot electron* (CHE) injection is generally considered to be important, and *channel hot hole* injection (which would be associated with PMOSFETs) is generally ignored.

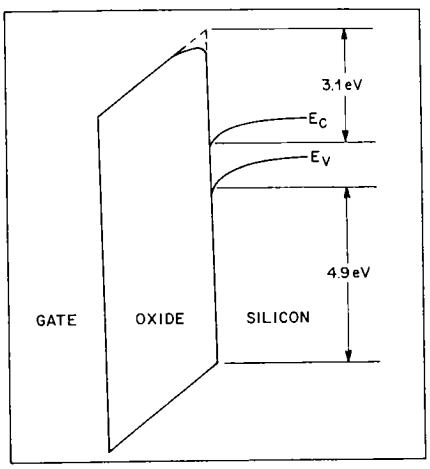


Fig. 7-6 Energy band diagram of a MOS device in the direction perpendicular to the surface. The barriers at the Si-SiO_2 interface are ~3.1 eV for electrons and ~4.9 eV for holes.

The energy-band diagram at the drain end of a MOSFET for the condition when hot electron injection can occur is shown in Fig. 7-7. Since the mean electron energy is a reasonably strong function of the electric field, and the fraction of hot electrons with an energy greater than the oxide barrier height is an exponential function of this mean energy, one expects most of the hot electron emission into the dielectric to occur where the channel field is the greatest. Figure 7-8a illustrates the field distribution along the channel for the saturation regime, which indicates that the channel electric field is very large near the drain in the pinchoff region. Thus, hot electrons are most likely to be generated in the pinchoff region of MOSFETs. The source of the electrons is the drain current, which is controlled both by the gate and drain voltages. More details on the channel electric field in short-channel MOSFETs and its effects on hot-carrier degradation are given in chap. 9. Nevertheless, once the channel hot electrons are injected into the oxide conduction band they can be transported to the gate electrode (in which case they represent a gate current), or they can be returned to the substrate, depending on the oxide field direction.

Figure 7-9 illustrates a plot of the gate current due to CHE versus gate voltage V_{GS}, with the drain voltage as the parameter. The CHE gate current initially increases as

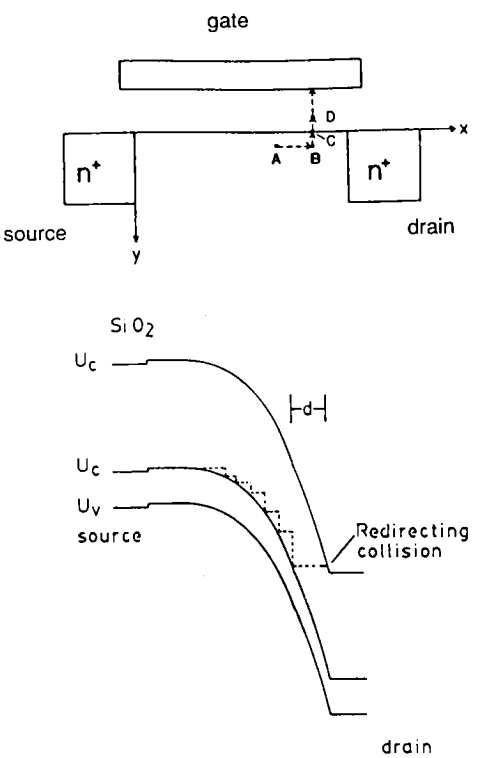


Fig. 7-7 The energy band diagram at the drain end of a MOSFET for the condition when hot electron injection can occur.

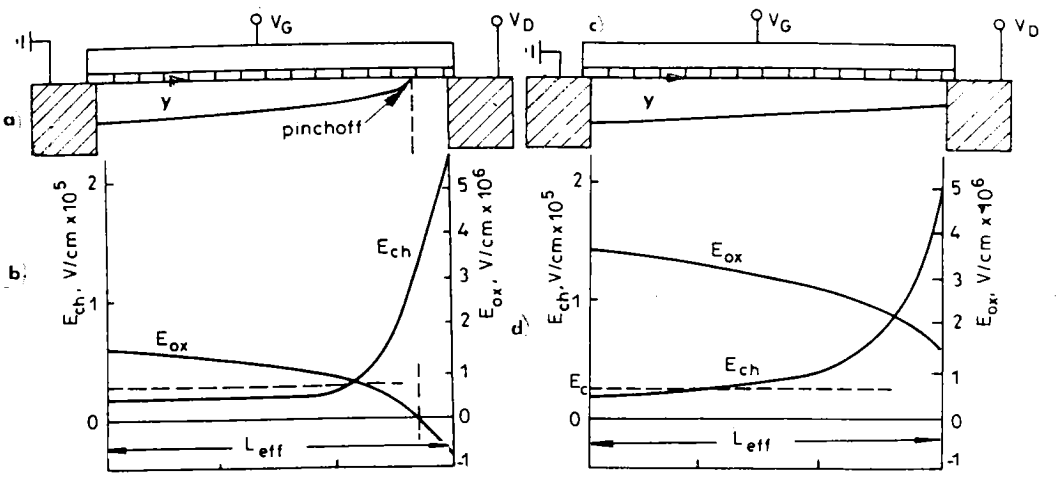


Fig. 7-8 (a) Cross section of MOS transistor operating in saturation. (b) The oxide and channel fields in the *y*-direction of the channel when the MOSFET is in saturation (i.e., $V_{GS}<V_{DS}$, here V_{GS} = 6V and V_{DS} = 8V). (c) Cross section of a MOS transistor operating in linear regime. (d) The oxide and channel fields in the *y*-direction of the channel when the MOSFET is in the linear regime (i.e., $V_{GS}>V_{DS}$, here V_{GS} = 15 V and V_{DS} = 8V).[119] (© IEEE 1991).

V_{GS} is increased - due to the increase in the number of carriers in the channel. The gate current peaks at the point when $V_{GS} \cong V_{DS}$ because when V_{GS} is increased beyond V_{DS}, the MOSFET is driven toward the linear region of operation and the channel electric field is rapidly reduced. As the channel electric field decreases below its maximum value, fewer energetic electrons are produced. Thus, hot electron emission also declines. (Fig. 7-8b shows the electric field along the channel in the linear regime. It is seen that the maximum channel field strength is smaller than in the saturation regime case). Note, however, that as drain bias increases, hot-electron injection into the oxide increases because of increased channel field.

It should also be pointed out that when V_{GS} is smaller than V_{DS} the field in the oxide near the drain end of the channel is directed so that electrons in the oxide conduction band at that location are drifted back toward the substrate. In this case, many electrons that jump over the energy barrier into the oxide simply return to the MOSFET channel. Since only the electrons that reach the gate are detected as gate current, the gate current due to CHE will decrease as gate bias decreases even when electron emission at the Si/SiO_2 interface is increasing (because of higher electric field).

The two common models for gate current due to CHE are the *lucky electron model* and the *electron-temperature model.* In the lucky-electron model,[7] the probability of an electron suffering a collision within a certain distance is determined (as the electron is being accelerated along the channel by the drain bias). In MOSFETs with small channel lengths, it is calculated that a few "lucky" electrons may simply escape such energy loss collisions and only gain energy from being accelerated. These lucky electrons, become "hot", and acquire sufficient energy to be able to jump into the conduction band. (Collisions resulting in a carrier *gaining* energy from a collision are

ignored). Using this model, the experimentally measured ratio of gate current to drain current has been successfully fitted at large drain voltages. The main discrepancy between the lucky-electron model and experimental data is that the model fails to describe the gate current observed under small drain bias.

The *electron-temperature model* assumes that the electron temperature T_e is higher than that of the lattice, and hence the electrons are not in equilibrium with the lattice.[8] The higher temperature of the channel electrons near the drain increases the number of electrons in the high-energy tail of the distribution. Some of these higher-energy electrons may be injected into the oxide. However, like the lucky-electron model this model predicts a much smaller gate current at low values of drain bias than is found experimentally.

Frey has presented an explanation of why gate current can occur at such low voltages. He noted that electrons can acquire (as well as lose) energy by collisions with the lattice.[9] The highest-energy electrons are also the ones most likely to undergo such energy-gaining collisions - i.e., with optical (high-energy) phonons. Calculations show that enough energy-gaining collisions are possible even at drain voltages as small as 1.5V, to produce a significant number of high-energy electrons.

However, as channel lengths shrink below about 0.15 μm, Frey postulates that electrons in the channel may not be able to undergo enough energy-gaining collisions

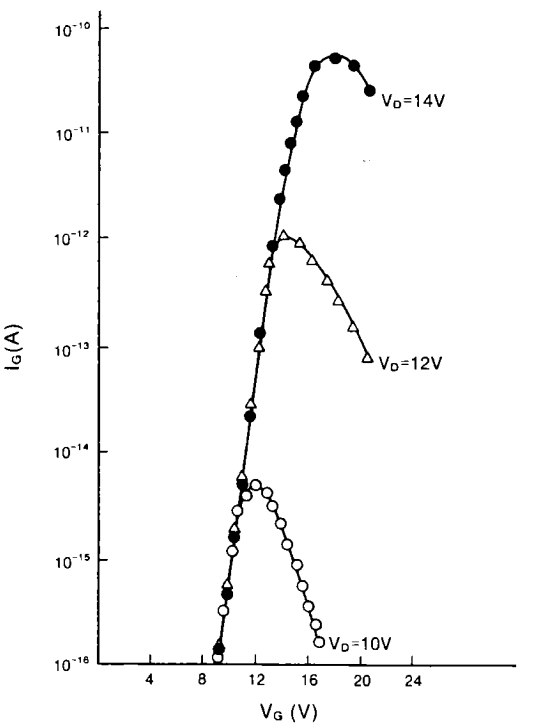


Fig. 7-9 Gate current due to channel hot electrons versus gate voltage, with V_{DS} as the parameter.[120] (© IEEE 1981).

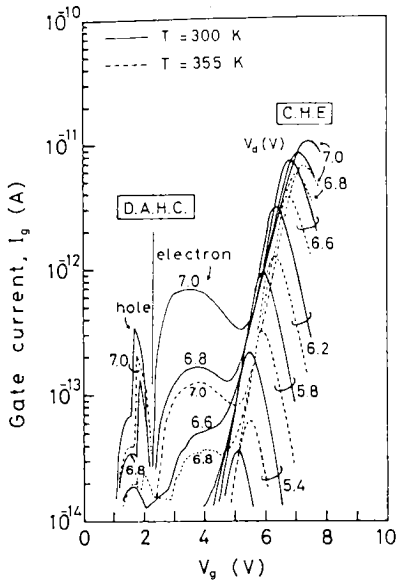


Fig. 7-10 At very low V_{GS}, I_G changes sign, indicating that hot holes are emitted into the gate oxide. Their presence at low V_{GS} is shown in this figure.[8] (© IEEE 1982).

before being collected at the drain. Consequently there will be fewer high-energy electrons available to cause hot-electron degradation reliability problems in such ultra-short-channel devices.

7.3.2 Injection of Drain Avalanche Hot Carriers (DAHC)

Energetic channel electrons accelerated by the lateral electric field can produce additional channel carriers (i.e., electron-hole pairs) by impact ionization in the channel region. The electrons created by impact ionization are either collected by the drain or injected into the oxide. The holes formed by this process give rise to substrate current, providing a measurable quantity for assessing the degree of impact ionization. Under conditions of gate bias much smaller than drain bias some of these holes can also cause a hot-hole gate current.

If the level of impact ionization is quite high a substantial substrate current can flow. Excessive hole flow into the substrate can lead to such undesirable behavior as snap-back or latchup. Hence, it is essential to monitor substrate current in a production environment and ensure that it does not become excessive in normal device operation.

As mentioned above, when V_{GS} is much smaller than V_{DS} (but V_{GS} is still greater than V_T), a gate current can flow due to injection of impact-ionization-generated hot holes (see Fig. 7-10). Since the magnitude of such hot-hole currents is very small (typically $<10^{-12}$ A) they are difficult to measure directly, especially in oxides thicker than 10 nm. However, indirect measurements using EPROM structures permit currents as small as 10^{-19} A to be detected. With these techniques, hole gate currents in oxides as

thick as 25 nm have been observed.[10,11] Figure 7-10 shows that hole gate currents arise when the gate voltage is much smaller than the drain voltage (e.g., V_{GS} = 1.2V and V_{DS} = 6V). As V_{GS} is increased above 2V (while V_{DS} is kept fixed at 6V), the hot-hole gate current decreases to vanishingly small values.

As the gate voltage is increased beyond the hot-hole regime, however, the bias regime associated with drain-avalanche hot electrons (DAHE) is encountered (Fig. 7-10). Such DAHE gate current arises because as the voltage between drain and gate decreases (but the drain voltage remains at a constant, high value), the probability increases that hot electrons injected into the oxide conduction band can make it to the anode before being forced back into the substrate by the oxide field. This results in larger gate currents due to DAHE as the gate voltage initially rises. Further increases in gate voltage, however, lower the lateral electric field to such an extent that hot electrons generated by impact ionization events will not contribute significantly to the gate current. Thus, for sufficiently high gate voltages I_g is dominated by CHE.

While the shoulder region of gate current attributed to DAHE is not observed for thick oxides, it is detected in thinner oxides. Some workers have argued that this shoulder region current is not due to DAHE, but instead is due to the increased collection efficiency of the emitted electrons for thinner oxides. More work is required to resolve this issue.

Some models of hot-electron degradation in MOSFETs also attribute this degradation to avalanche-generated hot carriers. Others, however, dispute this view, arguing that there exists no physical basis for the expectation that secondary (impact-ionization, or avalanche generated) carriers could cause more degradation than the primary hot carriers (i.e., CHE). They note that CHE are, on average, more energetic and numerous by at least an order of magnitude than those secondary carriers generated by impact ionization.[12]

7.3.3 Fowler-Nordheim (FN) Tunneling

The phenomenon of carrier injection into the oxide by tunneling is due to the quantum-mechanical nature of the electron. In the classical sense, the oxide represents an impenetrable barrier to injection of electrons into the conduction-band of the silicon if they possess kinetic energies smaller than 3.1 eV. However, the wave nature of the electron (or hole) allows a finite probability of crossing the barrier even if the electron does not possess sufficient kinetic energy. This probability increases with larger gate electric fields and/or thinner barriers. As mentioned in the introduction, electrons injected by tunneling are generally considered to be in equilibrium with the lattice, and thus are characterized as being "cold," i.e., not "hot."

The first type of tunneling is called *Fowler-Nordheim tunneling.* Here electrons are injected by tunneling into the conduction band of the oxide through the triangular energy barrier. Once injected into the oxide conduction band, electrons are accelerated by the oxide field toward the anode (gate), causing a gate current. Figure 7-11a sketches the energy band diagram for this case assuming a positive bias is applied to the gate.

To determine the probability of an electron exhibiting FN tunneling through a barrier, Schroedinger's equation is solved for a triangular barrier. The WKB approx-

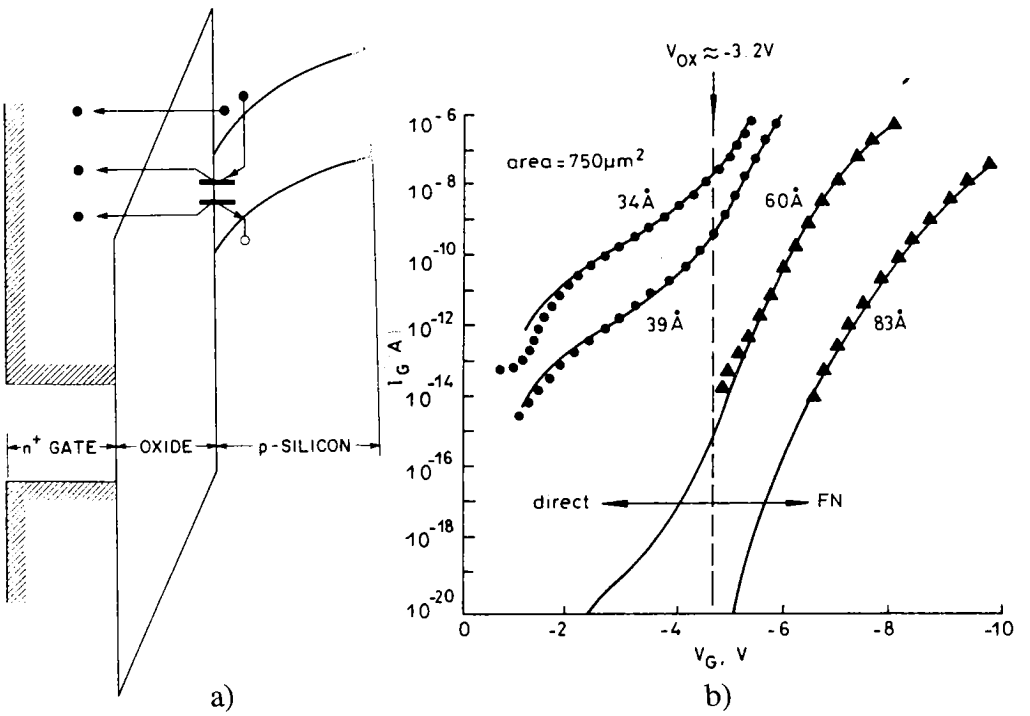


Fig. 7-11 (a) Energy-band diagram for the phenomenon of Fowler-Nordheim tunneling in the MOSFET gate oxide. Also shown are some possible interface-trap assisted leakage paths, with trap levels indicated by short, solid bars. From J.R. Brews, "The Submicron MOSFET," Chap. 3, in *High-Speed Semiconductor Devices*, Ed. Sze, p. 153. Copyright 1990, John Wiley & Sons. Reprinted with permission. (b) Theoretical and experimental tunneling I-V curves of Al-gate *n*-channel MOS structures under negative gate bias, illustrating the I-V characteristics of Fowler-Nordheim tunneling in 6 and 8.3 nm thick oxides and the calculated direct tunneling currents in 3.9 and 3.4 nm thick oxides.[15] (© IEEE 1983).

imation (discussed in most quantum mechanical textbooks) is invoked, and the following expression for the tunneling current expression is derived:

$$J = A_F \mathcal{E}_{ox}^2 \exp(-B/\mathcal{E}_{ox}) \qquad (7-1)$$

where $A_F = 1.25 \times 10^{-6}$ A/V^2, B ~240 MV/cm, J is the current density in A/cm^2, and $\mathcal{E}_{ox}$ is the oxide field in V/cm.[13] According to this equation, the oxide current is exponentially dependent on the electric field and experimental studies have found that the equation accurately describes the oxide current. This can be seen in Fig. 7-11b which shows the measured gate current and the gate current calculated using Eq. 7-1 in oxides 6 nm and 8.3 nm thick. The tunneling probability rapidly rises as $\mathcal{E}_{ox}$ increases

because the tunneling distance (i.e., the distance from the Si surface, through the triangular barrier, to the oxide conduction band) becomes smaller as $\mathcal{E}_{ox}$ increases.*

It should also be noted that it may be possible for electrons to be hot but not possess sufficient energy to surmount the barrier. In this case, carrier injection may occur by allowing these hot carriers to tunnel through a smaller triangular barrier distance. Thus, a synergistic combination of processes which individually would not have caused injection may yet enable carriers to be injected. This effect was modelled by Ning and Yu.[14]

7.3.4 Direct Tunneling

For gate oxides thinner than 6 nm, the tunneling current is observed to exhibit a weaker dependence on oxide fields than that described by Eq. 7-1. This has been explained by attributing the gate current in such oxides to direct tunneling through the forbidden gap of the SiO_2 to the gate electrode Such direct tunneling is known to occur in thin oxides at low voltages. In direct tunneling, electrons tunnel through a trapezoid-shaped barrier as shown in Fig. 7-12. An expression for the tunneling probability can be derived, but

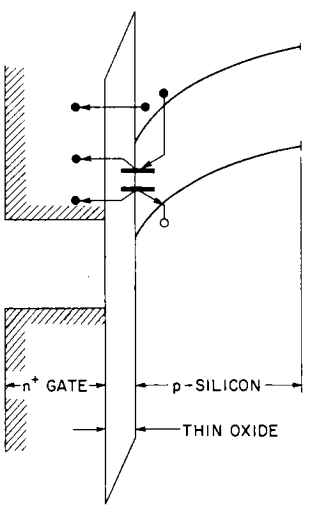


Fig. 7-12 Energy-band diagram for the phenomenon of direct tunneling through the gate oxide for thin oxides. Also shown are both unassisted tunneling and some possible interface-trap assisted leakage paths, with trap levels indicated by short, solid bars. From J.R. Brews, "The Submicron MOSFET," Chap. 3, in *High-Speed Semiconductor Devices*, Ed. Sze, p. 160. Copyright 1990, John Wiley & Sons. Reprinted with permission.

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

no analytical expression is available as in the case of FN tunneling. Figure 7-11a also shows the direct tunneling current calculated from numerical analysis.[15]

For oxide thicknesses smaller than 3 nm the direct tunneling current becomes large enough that it removes carriers faster than they can be supplied by thermal generation. In such cases an inversion layer is kept from forming in an MOS capacitor. This appears to set a lower limit of ~3 nm on gate oxide thickness.

7.4 TESTS FOR GATE OXIDE RELIABILITY

Various types of tests are used to characterize the breakdown characteristics of gate oxides. Depending on the type of test employed, the breakdown phenomenon may bear a different name.

7.4.1 Measuring Dielectric Strength

The *dielectric strength* of an oxide - i.e., the maximum electric-field strength that can be applied before it breaks down (typically expressed in units of V/cm) is measured with a

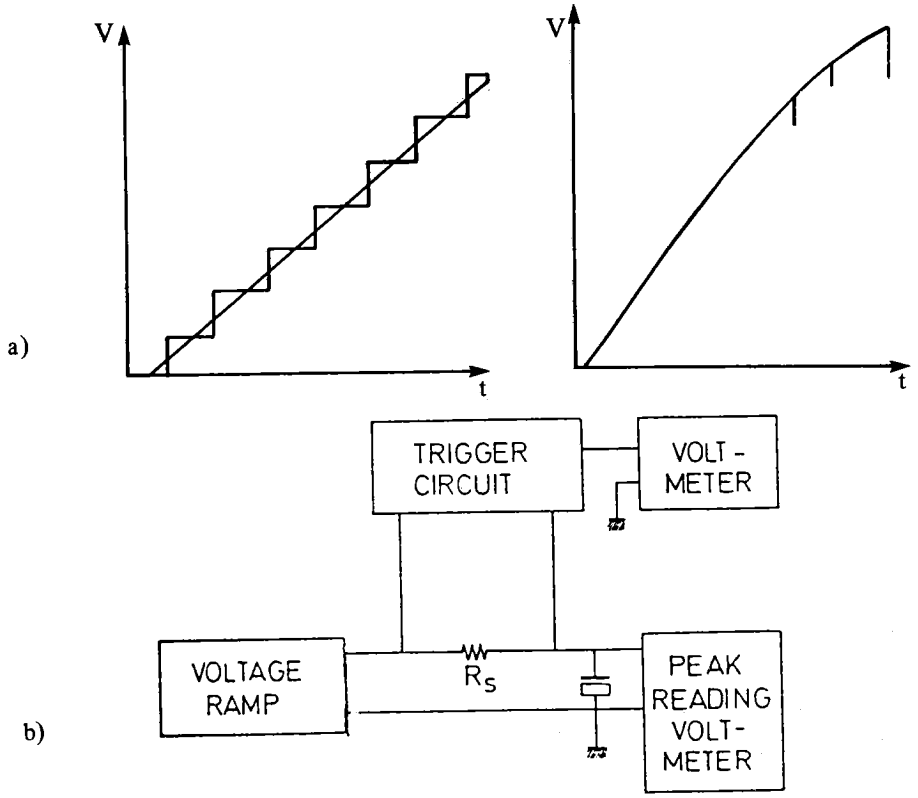


Fig. 7-13 (a) Breakdown voltage value of the gate oxide can be obtained by ramping or stepping the applied voltage and detecting sudden voltage drops at the electrode using a sensing resistor. (b) Diagram of the ramp-voltage oxide breakdown test setup.[16] Reprinted with the permission of the publisher, the Electrochemical Society.

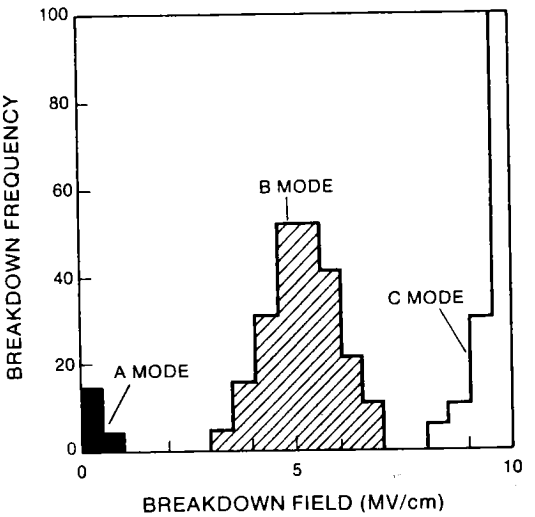


Fig. 7-14 Ramp voltage oxide breakdown data of Yamabe and Taniguchi.[17] (© IEEE 1985).

ramped-voltage test. That is, as shown in Fig. 7-13a a voltage that increases linearly with time (ramp voltage) is applied between the gate and substrate of the MOS-C test structure until the oxide breaks down. Breakdown is defined at the voltage value V_{BD} at which a sudden voltage drop across the oxide is observed. When measuring the dielectric strength of a set of oxide samples with the same thickness, the ramp rate is ($\Delta V/\Delta t$) is held constant (e.g., at a value of 0.1 or 1V/sec).* Figure 7-13b depicts the test setup for the ramped-voltage breakdown test.[16]

A histogram is usually used to plot the results of the ramped-voltage tests for a group of oxide samples, with the number (or fraction) of the group that failed within each voltage interval (e.g., within each 1 MV/cm, see Fig. 7-14).[17] The ramp breakdown tests are attractive because they require very little time. However, the V_{BD} data that is obtained is not directly relevant to practical concerns, since what is really needed is how long a given capacitor will survive at lower field strengths. A procedure for correlating V_{BD} to the time-to-breakdown at lower bias will be presented later.

On the other hand, a JEDEC-recommended current-ramp test applies a gate current that is ramped at a constant logarithmic rate until the oxide is broken down. In this test the charge-to-breakdown, Q_{BD} (see section 7.4.2.3), occurs in a very short time (i.e., approximately two to three seconds per capacitor), making it one of the most frequently used methodologies. Since it offers very poor reliability resolution, however, it can only be used to flag major oxide problems during the process.

* A variation of the ramped-voltage test is the *stepped-voltage test,* in which the voltage is stepped to various values, and held at those values for a constant time period.

7.4.2 Measuring Time-Dependent Breakdown Behavior

7.4.2.1 Time-to-Breakdown under Constant-Voltage Stressing. If the electric field in the oxide $\mathcal{E}_{ox}$ is kept constant during the stress test (implying also that the applied voltage is held constant), the biasing time (instead of the oxide voltage) becomes the variable. The length of time t_{BD} elapsed until breakdown occurs is then measured. The time-to-breakdown behavior of a group of oxide samples under such test conditions is referred to as *time-dependent dielectric breakdown* (TDDB). Breakdown is said to occur when the voltage across the oxide suddenly drops (Fig. 7-15.)[18] This test is frequently referred to as the *constant-voltage time-to-breakdown* test.

7.4.2.2 Time-to-Breakdown under Constant-Current Stressing. An alternative to the constant-voltage test is the *constant-current time-to-breakdown* test. Here, a preset current is injected into the insulator (by Fowler-Nordheim injection), and this value I_{inj} is held constant as the test proceeds. Voltage and time are recorded until the voltage suddenly drops.[18] Results from such tests indicate that the time-to-

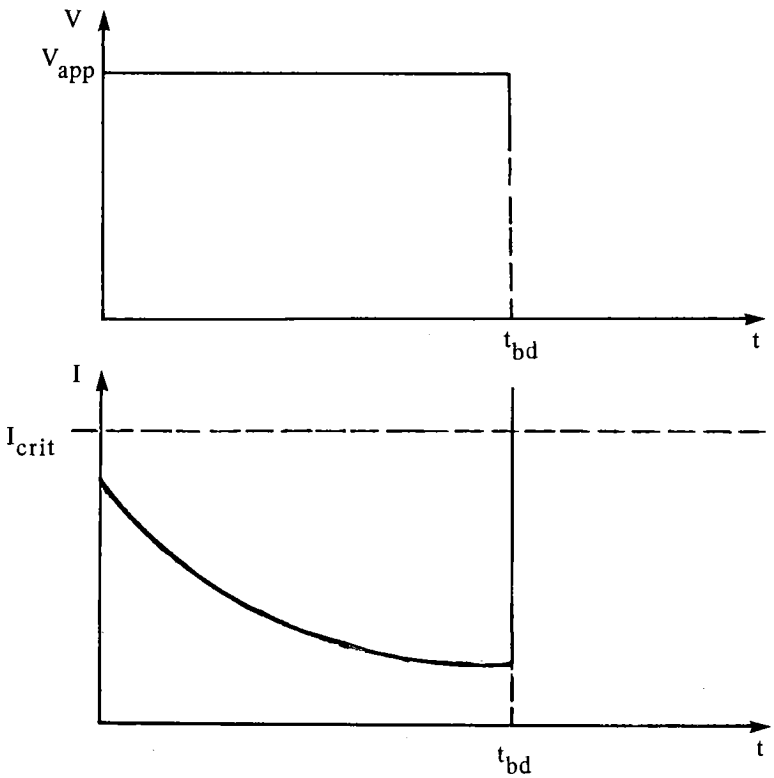


Fig. 7-15 (a) One technique for obtaining the time to breakdown t_{BD} is by applying a constant voltage and detecting a sudden voltage drop.

breakdown is dependent on the value of the current density $J_{inj} = I_{inj}/A$ - the larger the current density, the shorter the time t_{BD} until breakdown occurs.

7.4.2.3 Charge to Breakdown (Q_{BD}). If the constant-current test is used to determine t_{BD}, the product of the current density and t_{BD} provides another measure of the oxide breakdown behavior. This quantity is referred to as the *charge-to-breakdown* Q_{BD}. When comparing oxide breakdown behavior, those oxides exhibiting larger Q_{BD} are deemed superior. In a constant-current test the value of Q_{BD} is simply calculated from

$$Q_{BD} = J_{inj} \cdot t_{BD} \qquad (7 - 2)$$

If a constant-voltage test is used Q_{BD} can still be determined, but in this case the more general definition of Q_{BD} must be used (i.e., Q_{BD} is the time integral of J_{inj} from the time the bias is applied, to the time of breakdown)

$$Q_{BD} = \int_0^{t_{BD}} J_{inj}\, dt \qquad (7 - 3)$$

It has been found that for small values of J_{inj} the value of Q_{BD} for defect-free gate oxide films (i.e., intrinsic oxides) is constant as a function of the magnitude of J_{inj} (with a value of ~10 C/cm^2). For larger values of J_{inj} the value of Q_{BD} tends to decrease. We discuss this dependence of Q_{BD} on J_{inj} in section 7.5.1.

7.4.3 Mathematics of Reliability Characteristics

The dielectric strength ($\mathcal{E}_{ox}$ at breakdown observed in the ramp-voltage test), the time-to-breakdown t_{BD}, or charge-to-breakdown Q_{BD} are all important parameters that describe the breakdown of a single oxide sample. However, in a group of oxide samples of the same area and thickness, the above values will vary from sample to sample. Likewise, when a large number of MOS ICs (each containing an oxide of the same thickness and total gate oxide area) are operated under the same conditions, they will exhibit the same sort of variation in their breakdown parameters.

The *reliability* of an oxide is defined as:

> *the probability that it will not suffer breakdown under the stated operating conditions for the defined lifetime of the system.*

To predict the reliability of circuits and systems in which such MOS ICs are used, the oxide reliability must be expressed in quantitative form. Here we present the terms and the mathematical functions with which oxide reliability is quantitatively characterized. In later sections, these terms and functions will be applied to specific examples. A more rigorous treatment of this subject is given in reference 19.

7.4.3.1 The Failure Unit (FIT). If system contains a large number of electronic parts (e.g., MOS ICs) operated under the same conditions, a unit of failure rate can be

used to characterize the reliability of the system. This unit of failure rate is the ***Failure Unit*** or FIT. One FIT is defined as 1 failure per 10^9 hours of device operation. (Note that other failure rate units, e.g., 1 failure per 1000 hours, have also been defined, but for electronic systems containing a large number of parts, the FIT is most useful.)

For example, 10 years of operation equals ~88,000 operating hours. If a system contains 1000 MOS ICs this corresponds to 8.8×10^7 device hours of operation in 10 years. Assume the failures are known to be caused by oxide breakdown and the system is designed to exhibit a failure rate of 1000 FIT. At this failure rate (1000 FIT), 1000 failures are expected to occur in 10^9 hours, so 88 failures should occur in 8.8×10^7 hours. Hence, 88 of the 1000 parts (~10%) are expected to fail during 10 years of operation. Likewise, it can be shown that if a system is designed for 100 FIT, the percentage of parts that are expected to fail in 10 years of operation is ~1%, and at 10 FIT, only 0.1%. Note that 100 FIT also corresponds to 0.01% failure per 1000 device hours of operation.

It should also be noted, however, that 1 FIT represents approximately 125,000 years of failure-free operation for one device. Thus, it is impossible to wait that long to determine whether all of a large number of similar devices have failed by then and to record the time of individual failure. As a consequence, some sort of *accelerated life testing* is needed to determine reliability. This type of approach is described in section 7.5.3.

7.4.3.2. The Cumulative Distribution Function. Let there be N devices in a population, whose "reliability" one wishes to determine, and let these devices be placed under test at time $t = 0$. Assume that at time t it is observed $N_f(t)$ of the devices have failed and that $N_S(t) = N - N_f(t)$ devices have survived. The ratio of surviving devices to the number of original devices placed under test (N_S/N) is defined as the *reliability.* Similarly, the ratio of $N_f(t)/N$ is a measure of the unreliability of the devices and is usually termed the *cumulative distribution (or failure) function* $F(t)$. $F(t)$ describes the probability that a device will fail at a time before time t. In general, $F(t)$ has the following properties:

$$F(t) = 0 \qquad t < 0 \tag{7 - 4a}$$

$$0 \le F(t) \le F(t') \qquad 0 \le t \le t' \tag{7 - 4b}$$

$$F(t) \text{ --> } 1 \qquad t \text{ --> } \infty \tag{7 - 4c}$$

The *cumulative failure percentage* (i.e., the fraction of parts from a group of samples that will fail prior to a specific time t) can be obtained from $F(t)$. The time-to-failure data from the constant-voltage or constant-current tests are frequently plotted on a cumulative percentage failure graph, with the time axis scale plotted in terms of ln t.

7.4.3.3 Probability Density Function. The probability that a device will fail in the interval from t to $t + dt$ is determined by a function $f(t)$ referred to as the *probability density function.* It can be shown that $f(t)$ is the derivative of $F(t)$ with respect to time, or

$$f(t) = dF(t)/dt \tag{7 - 5}$$

Furthermore, because all devices fail eventually,

$$\int_0^\infty f(t)\,dt = 1 \qquad (7-6)$$

7.4.3.4 The Instantaneous Failure Rate. In many applications, the two most important reliability quantities are the *cumulative percentage of failure* [determined using $F(t)$], and the *instantaneous failure rate*, $\lambda(t)$. The instantaneous failure rate is the failure rate during an interval of time divided by the population remaining at that time. In fact, $\lambda(t)$ may be the more critical of these two reliability functions. It can shown that $\lambda(t)$ can be calculated from $F(t)$ according to

$$\lambda(t) = f(t)/[1 - F(t)] = \{1/[1 - F(t)]\}\ dF(t)/dt \qquad (7-7)$$

The distinctive shape of the $\lambda(t)$ curve makes it the more useful measure of the reliability of a part. This 'bathtub'-shaped curve can be divided into three regions, as shown in Fig. 7-16. A high infant mortality rate is initially observed, followed by a constant random (and generally low) failure rate (midlife), and ultimately a wearout region. The infant failures are most often due to a manufacturing process defect that was not evident at wafer-level or package-level testing.

7.4.3.5 The Mean Time to Failure. Another commonly used measure of reliability is the mean time to failure (MTTF) of the device or system, which can be expressed as

$$\text{MTTF} = \int_0^\infty t\,f(t)\,dt \qquad (7-8)$$

MTTF is the device's average age at failure for a population whose reliability is represented by the function $(1 - [F\{t\}])$ with a probability density function of $f(t)$.

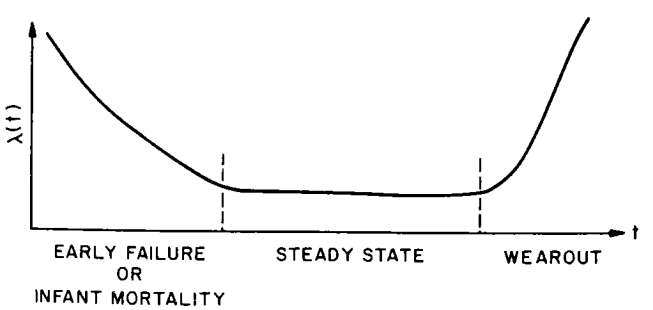


Fig. 7-16 Failure rate versus time for typical ICs.

7.4.3.6 The Weibull Cumulative Distribution Function. If the failure rate of a group of parts varies as the power of the age of the device (as is observed to be the case of oxide breakdown failures), the time-to-failure versus cumulative percentage failure is likely to be well fitted with a *Weibull* (or *extreme*) *cumulative distribution function.* The early failure of gate oxides can be represented by a Weibull distribution, and Weibull plotting paper is often used in presenting oxide TDDB data. The failure rate of parts whose failure statistics can be represented with the Weibull function is expressed as

$$\lambda(t) = \frac{\beta}{\alpha} t^{\beta - 1} \tag{7 - 9}$$

where α and β are constants. From Eq. 7-9 we can calculate that

$$F(t) = 1 - \exp\left[-\frac{1}{\alpha} t^{\beta}\right] \tag{7 - 10}$$

and

$$f(t) = \frac{\beta}{\alpha} t^{\beta - 1} \exp\left[-\frac{1}{\alpha} t^{\beta}\right] \tag{7 - 11}$$

When experimental data is fitted to an assumed cumulative failure function to determine the function parameters as well as the goodness of fit, appropriate plotting paper must be used. That is, the scales of the paper must be such that the experimental data will lie on a straight line if the data is well represented by the assumed distribution. Such plotting paper is available for use with the Weibull distribution function. For the Weibull distribution function

$$1 - F(t) = \exp\left[-\frac{1}{\alpha} t^{\beta}\right] \tag{7 - 12}$$

$$\ln\left\{\ln\left[\frac{1}{1 - F(t)}\right]\right\} = \beta \ln t - \ln \alpha \tag{7 - 13}$$

which is linear in the form $y = mx + b$. Figure 7-17 shows an example of Weibull plotting paper; the ordinate is marked in cumulative percent failed devices $F(t)$, and the abscissa is $\ln t$. If the experimental data are represented by a Weibull distribution, a straight line is obtained when the data are plotted, as shown in Fig. 7-17. The slope of the line is the parameter β. Note that if $\log t$ is the horizontal axis, the vertical axis can be scaled either as $\ln \{\ln (1 - F[t])^{-1}\}$ or as F (see Fig. 7-17).

7.4.3.7. The Log-Normal Cumulative Distribution Function. The *log-normal distribution. function* is another form of $F(t)$ that been empirically shown to successfully describe the failure statistics of oxide breakdown. Many reports which describe oxide TDDB experiments report their data using log-normal plotting paper. The function $f(t)$ of the log-normal distribution is given by

$$f(t) = \frac{1}{\sigma t \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right] \tag{7 - 14}$$

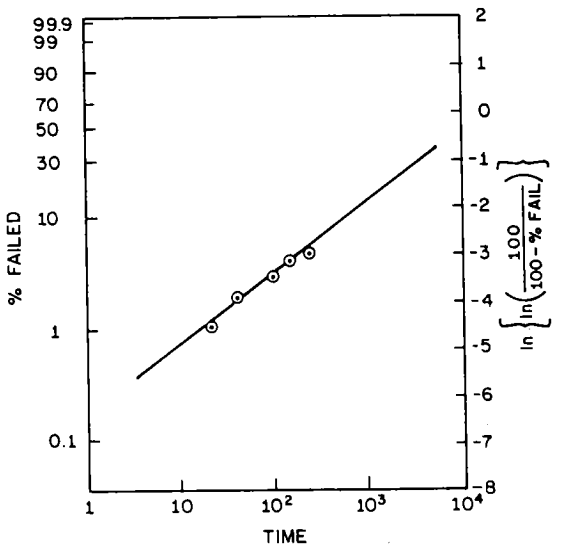


Fig. 7-17 An example of Weibull plotting paper showing the failure distribution for a typical device. From W.J. Bertram, "Yield and Reliability," Chap. 14 p. 632, in *VLSI Technology,* 2nd. Ed., S.M. Sze Ed., Copyright 1988 Bell Telephone Laboratories. Reprinted with permission.

$$= \frac{1}{\sigma t \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma} \ln \frac{t}{t_{50}}\right)^2\right]$$

where μ is the *mean* and σ is the *dispersion* of the distribution. The median time to failure (the time when 50% of the devices have failed) is given by

$$t_{50} = e^{\mu} \tag{7 - 15a}$$

or

$$\ln t_{50} = \mu \tag{7 - 15b}$$

and the *dispersion* (or *scale parameter*) σ is approximately

$$\sigma \cong \ln (t_{50} / t_{16}) \tag{7 - 16}$$

where t_{16} is the time when 16% of the devices have failed (more precisely, 15.866%).

The form of $F(t)$ is given by

$$F(t) = \int_0^t f(\tau)\, d\tau = \frac{1}{\sigma\sqrt{2\pi}} \int_0^t \frac{d\tau}{\tau} \exp\left[-\frac{1}{2}\left(\frac{\ln \tau - \mu}{\sigma}\right)^2\right] \tag{7 - 17}$$

Experimental data can be fitted to the log-normal distribution function by plotting the data on standard *log-normal plotting paper*, on which the ordinates have a normal probability scale and the abscissae have a logarithmic time scale (or vice-versa). A straight line representing the distribution is obtained if the probability distribution is log-normal. The value of the σ of the distribution can be estimated from the plot by the intersection of a line parallel to the experimental data and passing through the index point $\dot{\Phi}$, with the sigma scale on the right hand side of the plot (see Fig. 7-18). The MTTF is the time shown on the log-normal plot of the TDDB data when the cumulative percent failure is 50 percent.

7.4.3.8 Characterizing the Reliability of Oxides in Terms of the Mathematical Reliability Functions. As noted earlier, the quantities of most interest with respect to the reliability of an electronic system containing a large number of parts are: (1) the *cumulative failure percentage* of these parts during the expected lifetime of the system under the specified normal operating conditions; and (2) the *instantaneous failure rate* as a function of time over the system lifetime. Note that $\lambda(t)$ will generally vary during this time, but typically, a maximum value is specified. However, it may be possible to use a screening technique to prevent defective parts from being used in the system (by burn-in, as described in section 7-6). This can not only reduce the value of $\lambda(t)$, but can also make $\lambda(t)$ become essentially constant over the

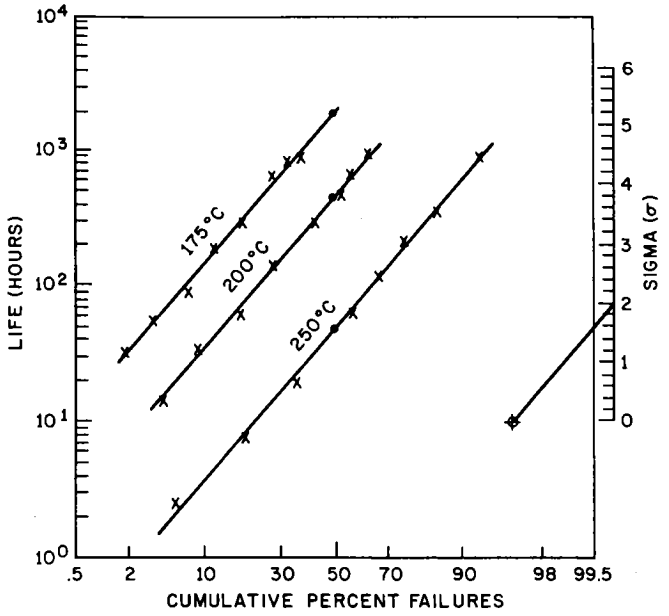


Fig. 7-18 An example of log-normal plotting paper showing device failure distribution at three temperatures. From W.J. Bertram, "Yield and Reliability," Chap. 14 p. 636, in *VLSI Technology,* 2nd. Ed., S.M. Sze Ed., Copyright 1988 Bell Telephone Laboratories. Reprinted with permission.

expected lifetime (i.e., the surviving parts behave like those in the constant (or random -failure) portion of the instantaneous failure rate curve shown in Fig. 7-16). In such cases, a single λ (t) value can be used to characterize the failure rate of the system.

The cumulative percentage of failure is usually determined from experimental values of TDDB data under normal or accelerated stress conditions (but, by using a relationship that correlates V_{BD} to t_{BD}, the cumulative percentage failure can also be determined as a function of experimental V_{BD} values, as will be discussed in section 7.5.3).

To determine values of λ (t) however, we must calculate them from Eq. 7-7. This implies that an analytical expression for F (t) must be available. The following two approaches are used to establish the analytical form of F (t) from the empirical cumulative failure data.

1. In the first approach, F (t) is assumed to exhibit either a Weibull or log-normal form. In this case, the data of cumulative percentage failure versus t_{BD} is plotted on the graph paper appropriate for the assumed form of F (t) (i.e., if we assume $F(t)$ has a log-normal form, the data is graphed on log-normal plotting paper). If a straight-line dependence is exhibited by this graphed data, the parameters needed to obtain the analytical form of F (t) can be extracted from the plot [e.g., μ and σ for the log-normal form of the F (t) function, or α and β for the Weibull form of the F (t) function]. Once these parameters are extracted, the analytical form of $F(t)$ can be determined, and this function can then be used in Eq. 7-7 to determine λ (t).

2. In the second approach (as is used in the oxide breakdown reliability model discussed in section 7.5.4), F (t) is derived indirectly. Here, oxide breakdown is assumed to be hastened by defects present in the oxide film.[20] For instance, when such a defect exists, the oxide breaks down at a smaller voltage than expected for a film of that nominal oxide thickness. Since the distribution of such defects is also used in estimating the yield of ICs, F (t) is likewise assumed to have a form (1 - Y), where Y is the function that estimates the yield (ratio of good chips to the number of chip sites per wafer). The experimental cumulative failure percentage data is then used to calculate Y, which in turn is used to establish the analytical form of F (t).

Specifically, if the defects are assumed to be uniformly distributed Y is assumed to follow the Poisson distribution

$$Y = \exp[-D(\Delta t_{ox})A]. \qquad (7 - 18a)$$

If they are assumed to be nonuniformly distributed, Y has a Gamma function form

$$Y = \frac{1}{(1 + A D (\Delta t_{ox}) S)^{1/S}} \qquad (7 - 18b)$$

The data of cumulative percent failure versus t_{BD} is then used to establish a set of empirical data points representing the cumulative oxide defect density [$D(\Delta t_{ox})$] versus effective oxide thinning. (The procedure for doing this is described in detail in section 7.5.4.) Finally, an analytical function of $D(\Delta t_{ox})$ is fitted to this set of points (see for example, Eq. 7-40). The analytical form of this function is used to calculate the defect density value required in the function used to calculate Y. That is, once A and $D(\Delta t_{ox})$ are known, Y can found, and this allows an analytical form of $F(t)$ to be established (see Eq. 7-41). From this expression, $\lambda(t)$ can be found using Eq. 7-7.

7.5 THE PHENOMENON OF OXIDE BREAKDOWN

Thin oxide films undergo catastrophic breakdown when stressed by high electric fields. The maximum electric field that SiO_2 films (fabricated in normal IC processes) can withstand without instantaneously breaking down is approximately 12 MV/cm. However, not all oxide films of the same thickness undergo breakdown at the same electric field strength. Results of experiments on MOS-C structures in which an increasing electric field is applied until breakdown occurs indicate that such oxide failures fall into three groups (as shown in Fig. 7-14).

Some fraction of the sample population can withstand the highest electric fields (8-12 MV/cm). These constitute the C-mode group shown in Fig. 7-14. The high-field failure mechanism exhibited by this group is referred to as "intrinsic" failure. It is generally assumed that such oxides are defect free. Intrinsic oxide films are observed to fail at about the same time if they are stressed for long periods at lower electric fields.

Another fraction of the stressed oxides fail instantly upon the application of a relatively small gate bias, e.g., <1 MV/cm (the A-mode group in Fig. 7-14). It is believed that such MOS-Cs may already be shorted before the application of the low-strength field. Oxides that cannot withstand even such relatively low electric fields are thought to possess gross defects (such as pinholes).*

The remaining fraction of oxide films (the B-mode group in Fig. 7-14) suffers failure at electric fields of intermediate strength (2-6 MV/cm). The oxides of the B-mode group are thought to contain weak spots that do not produce instant shorting but may give rise to early failures of ICs under normal operating conditions (infant-mortality failures). Thus, B-mode failures are believed to be responsible for the early breakdown of oxides. Since such premature failures occur at various times, they are also referred to as *time-dependent dielectric breakdown* failures. Note that if 5V is applied across a 15 nm gate oxide, the oxide field falls in the range of the B-mode failures (~3 MV/cm). Hence, B-mode failures represent a serious problem for MOSFETs with such thin oxides. In this section we discuss the physical models of intrinsic failure (C-mode) and early failure (B-mode).

* A technique for gate oxide pinhole detection prior to metalizing the oxide has recently been described.[21] It works by carbonization of the Si surface by propane infiltration. After surface carbonization at the pinhole sites, the oxide is stripped, leaving behind SiC spots that are easily visible with an optical microscope.

Before beginning the discussion of breakdown models, one other general observation regarding breakdown should be mentioned. It appears that oxide degradation occurs in two stages. At first, the oxide is gradually weakened by the passage of gate current. At some point this weakening permits a continuous conductive path to be locally established in the oxide. The energy stored in the MOS-C is then abruptly discharged by the current flowing along this path. As a result, the oxide film (and also possibly the gate material above the oxide at that location), is vaporized at that spot.

Examination confirms that the breakdown is highly localized. That is, a conductive short of very small area is found at a specific location within the MOS-C (Fig. 7-19).[18] However, if the gate material over this spot is also completely vaporized, the capacitor itself may not be shorted. Such blowing of the gate material over the breakdown spot acts a self-healing mechanism. During oxide stress testing many such self-healing breakdown events may be observed on one capacitor.[22]

7.5.1 Qualitative Physical Models of Intrinsic Oxide Breakdown

The physical origin of intrinsic oxide breakdown is still not entirely understood. A

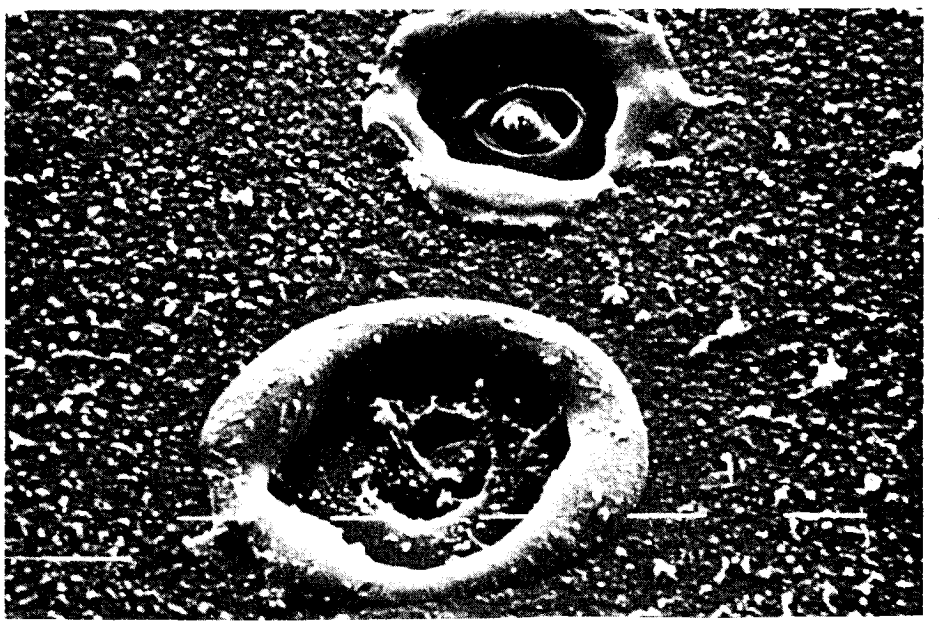


Fig. 7-19 A SEM photograph of two breakdown spots. The polysilicon electrode has melted away over 20 μm^2. The polysilicon is spread over a large area in the form of dust and little globules. The size of a dash is 1 μm. From D.R. Wolters and J.F. Verwey, "Breakdown and Wear-Out Phenomena in SiO_2 Films," Chap. 6 p. 329, in *Instabilities in Silicon Devices.* G.M. Barbottin and A. Vapaille Eds., Copyright 1986 Elsevier Science Publishers. Reprinted with permission.

number of models have been proposed to explain the phenomenon, but experimental evidence does not convincingly support one over the others. Here we describe the two leading models, the *hole-generation and trapping model* and the *lattice damage model.* Several other models have also been proposed including the electron-trapping model, the resonant tunneling model, and the electron trap-generation model. References are listed that give the details of the latter three.[23,24,25]

7.5.1.1 Hole Generation and Trapping Model. The most widely accepted theory of oxide breakdown is the *hole generation and trapping model.*[26,27,28] This model proposes that at sufficiently large fields electrons are injected into the conduction band of the oxide by Fowler-Nordheim tunneling. In the oxide conduction band these injected electrons are accelerated by the oxide field toward the anode, thereby gaining kinetic energy.* Some of these high-energy electrons generate electron-hole pairs within the oxide. A fraction of these generated holes are trapped at localized areas of the oxide, which, for some reason, have a greater probability to trap positive oxide charge than the "robust" oxide areas. (Some sort of 'super-fine' defects such as trivalent Si atoms in the bulk [resulting in a dangling bond], impurities, or very small micropores at such locations are believed to act as a 'seed' for the generation of 'larger' defects during high current injection.)** Locations with such a propensity to trap positive charge are postulated to make up only $\sim 10^{-6}$ of the total oxide area.[28] The locally trapped holes increase the positive charge (and thus the cathode field) at such points in the oxide, causing the tunneling current density to increase there (a positive feedback effect). When the density of trapped-oxide hole charge Q_{ot}^{+} at one of such locations reaches some critical value, the local tunneling current is increased to the point that breakdown occurs.

Several mechanisms for hole generation within the oxide by the energetic electrons have been suggested, including band-to-band impact ionization, trap-assisted impact ionization, and transitions between the conduction-band and valence-band tails of the SiO_2 (Fig. 7-20). The latter two effects could permit significant hole generation at even relatively low voltages (albeit at reduced generation rates).

The argument that trapped holes are the responsible species for oxide damage gains more plausibility once it is recognized that hole trapping in oxides is far more likely to

* Note that the *anode* and *cathode* are the terminals at which the polarity of the applied bias is, respectively, positive and negative. Thus, either the gate or the substrate in a MOS-C can be the cathode, depending on the polarity of the applied bias. In most cases, the gate of a MOS-C with a *p*-type substrate is biased positively in oxide stress tests (making the substrate the cathode). Occasionally, however, results are reported for the case of negative gate bias, in which case the gate becomes the cathode.

** It has been speculated by Sah that if the oxide film was microscopically uniform and free of such superfine defects, that the 'intrinsic' breakdown fields observed in normally fabricated thin oxide films (10-15 MV/cm) would be much higher (possibly by as much as by a factor of 5).[29] Under such conditions, if current runaway was not initiated by very-high Fowler-Nordheim injected gate currents, the field-strength might eventually become so strong that the electrons would be directly ripped from their silicon-oxide bonds, and the lattice would disintegrate or explode!

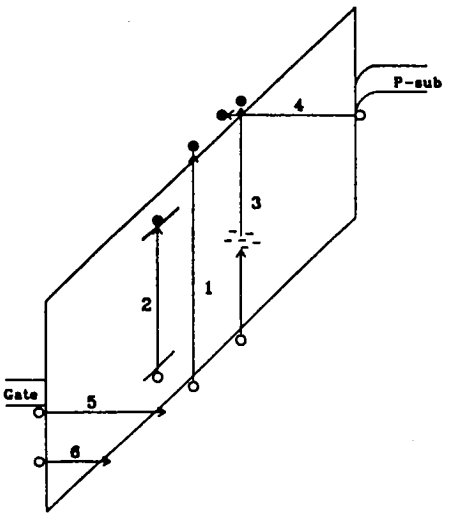


Fig. 7-20 Several possible mechanisms of the substrate hole current.[31] (© IEEE 1986).

occur than electron trapping. That is, electrons in the oxide conduction band are quite mobile (20-40 $cm^2/V{\cdot}s$), and hence most are apparently swept rapidly from the oxide. However, the relatively low mobility of holes in the oxide (~$2x10^{-5}$ $cm^2/V{\cdot}s$) makes it far more probable that they will be trapped. The greater likelihood of hole trapping has been demonstrated by independent experiments, in which MOSFETs were exposed to ionizing radiation capable of generating electron/hole pairs and lattice damage. Such irradiated MOSFETs exhibited negative threshold voltage shifts, implying that positive charge accumulated in the oxide after being irradiated. From this it is inferred that if the lattice is damaged and both electrons and holes are available to be trapped on the damaged sites, positive charge (hole) trapping appears to be more probable than negative charge (electron) trapping.

It should be noted, however, that some electron trapping is also possible, and in fact it evidently occurs simultaneously with hole trapping as a result of the electron transport in oxides. This is evident from Fig. 7-21 which shows the measured and calculated gate current (using the quantitative model described in section 7.5.3) through a 13 nm-thick gate oxide as a function of V_{GB}.* The saturation of the gate current as V_{GB} increases implies that negative charge is being trapped throughout the oxide, as depicted in Fig. 7-22. (Note, however, that typical oxide films have been shown to trap only about one percent of the total injected electrons.[30])

Nevertheless, while trapped electrons may be accumulating throughout most of the oxide volume as gate current flows, the hole-trapping model assumes that such electron trapping does not contribute to the damage that leads to breakdown. Instead, only at

* Note that in this case, V_{GB} is negative, implying that the gate is the cathode, and electrons are injected into the oxide from the gate.

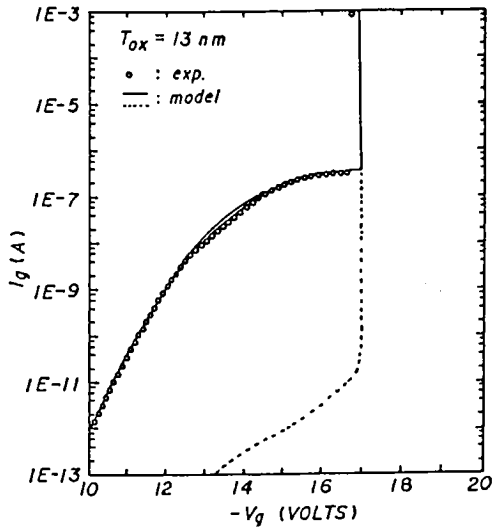


Fig. 7-21 I-V curve of a ramp-voltage test measured and simulated by the model. The dotted curve is the current carried by the weak area, which is responsible for breakdown.[121] (© IEEE 1985).

those local spots where hole trapping is dominant does such damage occur. The current density carried in these 'mini-defective' regions increases as the positive charge builds up. While this current component initially makes up only ~10^{-6} of the total tunneling current (since the defective areas consist of only ~10^{-6} of the total oxide area), the

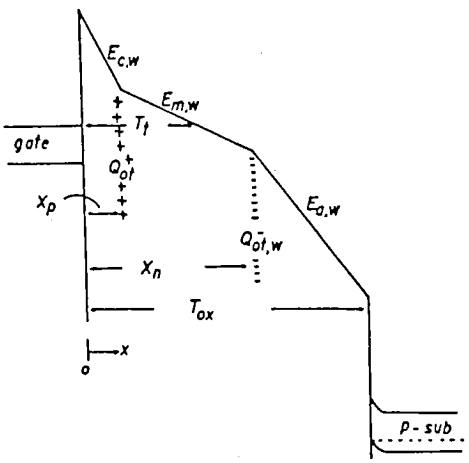


Fig. 7-22 SiO_2 energy diagram showing trapped holes Q_{ot}^+ and trapped electrons Q_{ot}^-.[121] (© IEEE 1985).

positive feedback effect causes such defect-region current to increase until it dominates the gate current flow and produces breakdown. (Figure 7-21 also shows the calculated current flowing in the mini-defective region, as the dashed curve. Notice that it rises until it eventually dominates the gate current, finally leading to breakdown).

The presence of hole flow in the gate oxide during oxide stressing (resulting from the generation of holes in the oxide as discussed above) has also been directly observed using two different approaches. In the first, the gate of an *n*-channel MOSFET is positively biased to cause Fowler-Nordheim tunneling of electrons into the oxide. This gives rise to a gate current I_g (consisting of electron flow) shown in Fig. 7-23.[31] At the same time, however, a small hole current is measured flowing from the substrate I_p (originating from holes generated in the oxide, as discussed above, which are then drifted to the substrate by the oxide field). The charge to breakdown due to I_g (Q_{BD}) is monitored versus Fowler-Nordheim (electron injection) fluence, as is the hole-charge-to-breakdown (Q_p) versus hole fluence. For low densities of gate current, the value of Q_{BD} in intrinsic oxides is found to be relatively constant at ~10 C/cm^2. The value of Q_p is also constant, but is much smaller, at a value of ~0.1 C/cm^2 (Fig. 7-24a).

In the second approach of directly examining the effect of hole flow on the breakdown characteristics of oxide, hot holes originating in the channel are injected into the oxide. That is, hot electrons are produced in the channel of an NMOSFET by the application of a high drain electric field (e.g., V_{DS} = 10V).[32] These hot electrons generate additional channel electron/hole pairs by impact ionization, as described in section 7.3.2. By keeping the gate voltage much lower than the drain voltage (e.g., V_{DS}

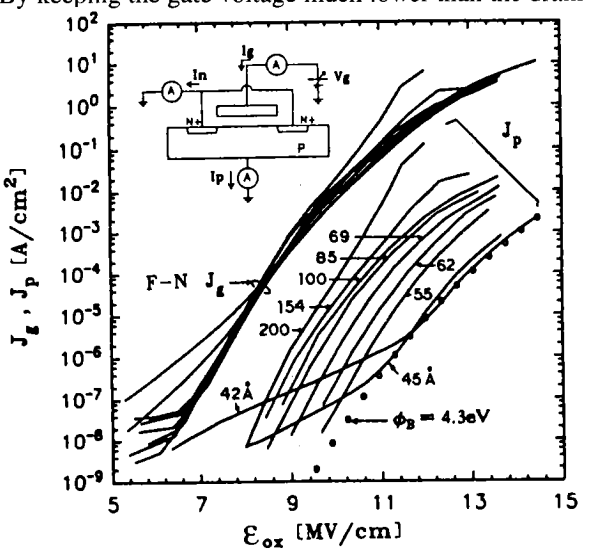


Fig. 7-23 Gate and hole current density versus oxide field for different oxide thicknesses. The gate currents follow the same Fowler-Nordheim characteristics as expected. The dependence of hole current on oxide thickness eliminates cold carrier tunneling as the origin of holes. The dotted curve labeled φ_B = 4.3 eV represents the theoretical J-$\mathcal{E}$ for valence band electron tunneling.[31] (© IEEE 1986).

= 10V and V_{GS} = 1.5V), any channel hot holes injected into the oxide conduction band will be drifted by the oxide field toward the gate. While hot electrons can also be injected and transported through the oxide (by making the gate voltage *larger* than the drain voltage), it was found that MOSFETs stressed with enough hot electrons to produce threshold voltage and transconductance degradation did not exhibit reduced Q_{BD} values if subjected to Fowler-Nordheim stressing after such hot electron stressing. On the other hand, stressing by hot hole injection prior to Q_{BD} testing, degraded the Q_{BD} of the oxides. That is for every 0.01 C/cm^2 of hole fluence, Q_{BD} was reduced by ~1 C/cm^2. This also appears to support the model assumption that hole transport through the oxide is somehow related to oxide breakdown.

One inconsistency between this model and experimental data is that Q_{BD} appears to *decrease* with increasing temperature. Since impact ionization is predicted to decrease with increasing temperature, the model would imply that Q_{BD} should also *increase* with temperature. Another question concerns the fact that oxide breakdown occurs at electric fields smaller than those that could cause band-to-band impact ionization in the oxides. This concern was addressed by postulating that trap-assisted ionization or transitions between the conduction-band and valence-band tails of the SiO_2 could be responsible for producing holes in the oxide, as mentioned above.

Nevertheless, if the mechanism of hole trapping is indeed chiefly responsible for oxide breakdown, it would be useful to characterize the origin, growth, and shrinkage of hole trap-center densities. As noted earlier, hole traps are thought to be caused by oxygen vacancies in bulk and/or dangling bond at the surface.

A recent study by Satake and Toriumi sheds some light on why Q_p is independent of (Fowler-Nordheim induced) gate current density, but that Q_p (and Q_{BD}) decreases with increasing temperature.[102] Their model postulates that hot holes in thin oxides are generated in the gate electrode, and that these holes are captured at traps (usually weak Si-O bonds randomly distributed throughout the oxide) having a small but finite repulsive barrier (as shown in Fig. 7-24b). The hole trapping induces local structural changes in the SiO_2 network (weak Si-O bond becomes a dangling bond), and the oxide band gap at these sites becomes smaller than that of SiO_2. Eventually a breakdown path whose conductivity is much larger than that in SiO_2 is created, leading to oxide breakdown. At room temperature, holes can easily be trapped at these sites by surmounting the repulsive barrier merely with the aid of thermal energy. Thus, oxide breakdown brought about by such trapping can occur without the help of the oxide field. At low temperatures, holes do not have enough thermal energy to overcome the barrier, and the traps remain unfilled unless the oxide field is increased such that the barrier is lowered (Fig. 7-24c). Thus, Q_p increases for smaller gate current densities at lower temperatures (Fig. 7-24d). This data supports the model that holes captured in the oxide causes local structural changes in the oxide lattice that eventually leads to oxide breakdown.

It has also been shown that the number of *electron* trap sites is related to the water diffused into the SiO_2 film.[33] Furthermore, the negative charge build up in ultra-dry SiO_2 is 1/10 that in wet oxides. These results suggest that the hydrogen contained in the oxide strongly affects the generation of electron traps in SiO_2. However, some of

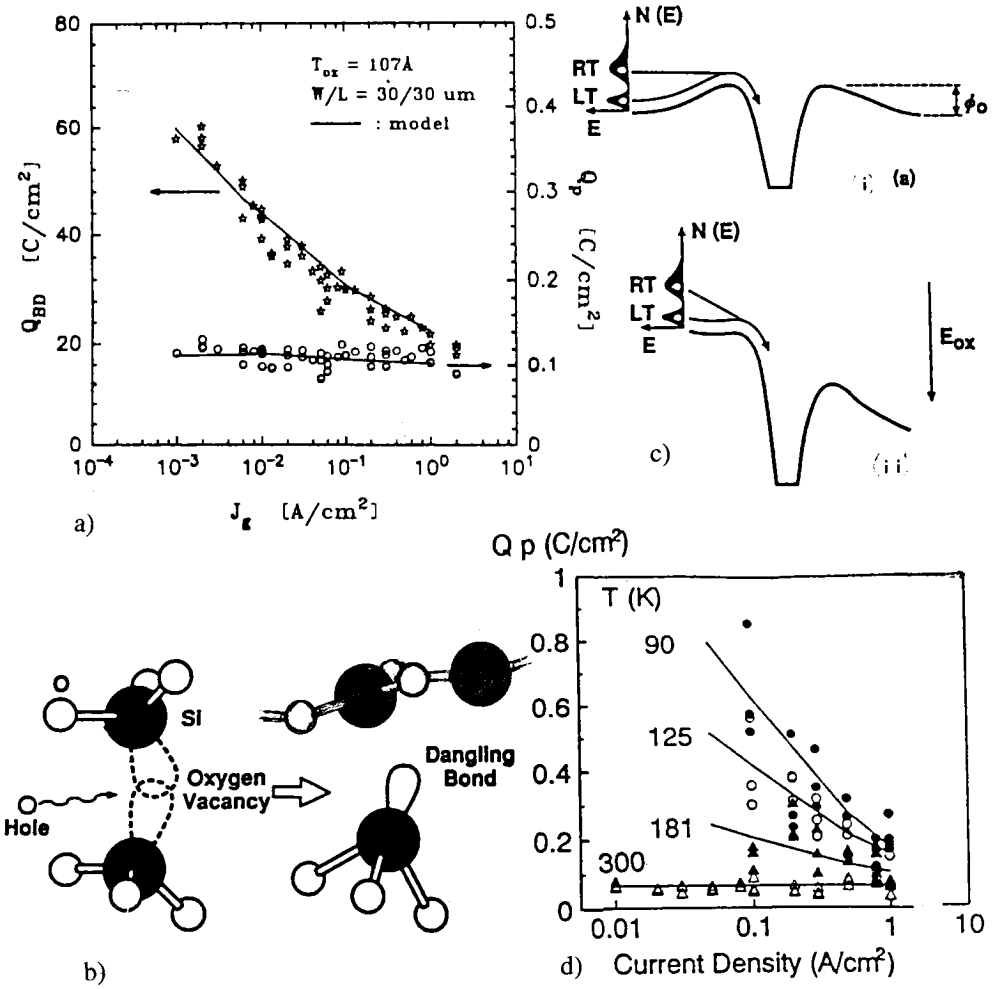


Fig. 7-24 (a) The charge-to-breakdown, Q_{BD}, and the hole-fluence-to-breakdown, Q_p, for a 107Å oxide under constant current stress. Q_{BD} increases with decreasing current density, while Q_p remains constant as predicted by the hole-trapping oxide breakdown model.[31] (© IEEE 1986). (b) Schematic picture of structural change in Si-O bonds caused by a trapped hole. Oxide vacancy acts as a hole trap. The hole trapping locally induces a structural change in the SiO_2 network. The conductivity in the changed structure becomes quite large compared to that of the initial SiO_2 due to its small band-gap. (c) Schematic picture for the repulsive hole trap with a finite energy barrier φ_o. Figures (i) and (ii) correspond to the trap without and with oxide field $\mathcal{E}_{ox}$, respectively. At room temperatures, holes are easily trapped with only the help of thermal energy. At low temperatures, barrier lowering by the electric field is needed for holes to be captured. (d) Hole-fluence-to-breakdown Q_p as a function of current density. At 300K, the oxide-field dependence of Q_p agrees with previous work [31]. However, note that Q_p has a strong oxide field dependence at lower temperatures.[102] (© IEEE 1993).

the hydrogen contained in the oxide is eliminated by high temperature annealing in dry N_2. Since the Si-OH structure is thought to be an electron trap, and oxygen vacancies in the bulk and dangling bonds at the surface are thought to be hole trap precursors, such N_2 anneals are expected to reduce electron traps, but also to increase hole traps. Since it does not appear to be possible to simultaneously achieve a reduction in electron and hole trap densities by annealing alone, other kinds of oxide quality improvement by adding small amounts of impurities to the oxide have been investigated, including fluorine and nitrogen. These approaches are described in chap. 9.

7.5.1.2 Wolters' Electron Lattice-Damage Model. In the lattice-damage model of Wolters, it is agreed that electrons are injected into the conduction band of the oxide by Fowler-Nordheim tunneling where they undergo acceleration by the oxide field. Upon reaching the anode they have the highest energy. This energy is dissipated by creating atomic defects at the anode interface (e.g., breaking Si-O bonds). These defects are charged positively and thus they attract newly injected electrons. This mechanism directs the growth of tracks of defects into the dielectric. Such bond breaking proceeds from anode to cathode forming a conductive path. Figure 7-25 depicts the simulation of these tracks which evolve during random charge injection. Initially a large number of small tracks grow. The largest track will grow faster than the others. Tracks also tend to branch and closely resemble "trees" in the pre-breakdown stage. Eventually, a conductive path that has grown from the anode to cathode is produced.[34] Once the conductive path is established across the entire film thickness, the capacitor can be discharged through this path. Since this discharge occurs very rapidly, the energy stored in the capacitor must be quickly dissipated by the small volume of oxide material that

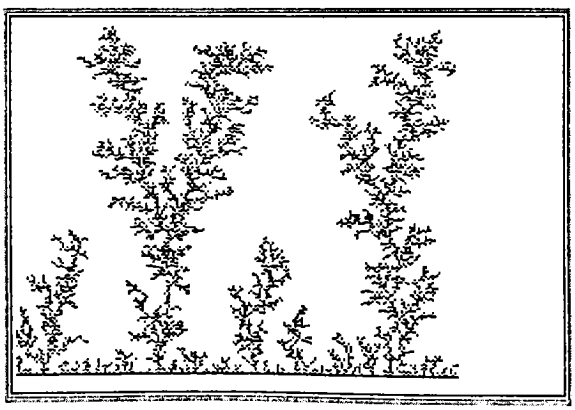


Fig. 7-25 Modeling of growing "trees" at the draining electrode of a capacitor under stress. Under charge injection the "trees" grow until they reach the opposite electrode and make a short.[34] This paper was originally presented at the Spring 1990 Meeting of the Electrochemical Society, Inc. held in San Francisco, CA.

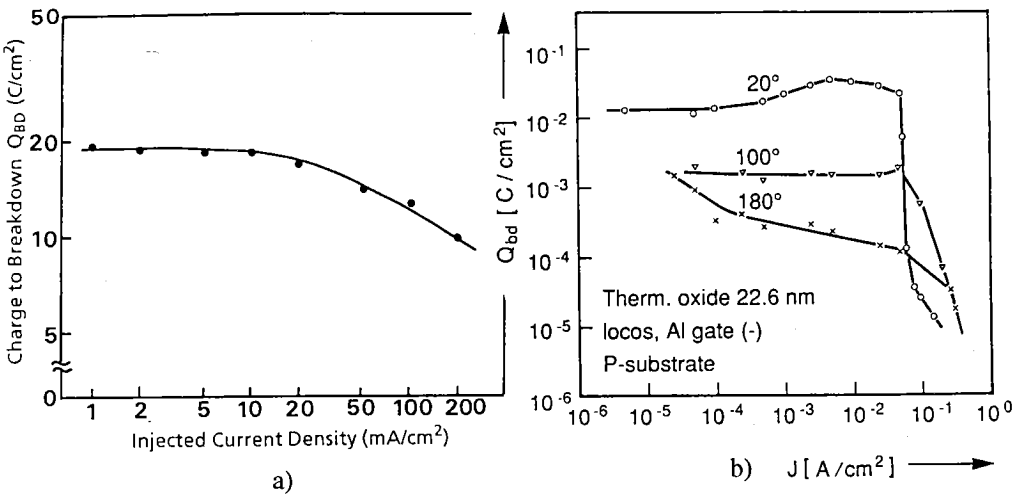


Fig. 7-26 (a) The dependence of the breakdown charge on the F-N injection current density. Q_{BD} is constant below certain current density and decreases gradually beyond it.[123] (© IEEE 1991). (b) Temperature effect on Q_{BD} values. Devices have often higher operation temperatures and hence, the lowering of Q_{BD} with T is an important aspect of end-of-life prediction.[34] This paper was originally presented at the Spring 1990 Meeting of the Electrochemical Society, Inc. held in San Francisco, CA.

constitutes the current path. This vaporizes the oxide along the path of current flow, and the capacitor is broken down. (The area of melted polysilicon above the breakdown spot, as shown in Fig. 7-19, is ~20 μm^2. For a breakdown spot of a few square microns in even a very small-area MOS-C [e.g., 0.02 mm^2], the energy density arising from the dissipation of the charge stored on the capacitor can exceed 10^5 J/cm^3. Such energy densities can indeed melt and evaporate polysilicon spots of 20 μm^2.)

Wolters has also observed that the average value of Q_{BD} is essentially constant as a function of Fowler-Nordheim injection current density (i.e., electron fluence) for low current densities. According to him, this precludes avalanche multiplication as a stepping stone in the oxide breakdown mechanism because the multiplication is strongly field dependent. This constancy of Q_{BD} has also been observed by Uraoka et al[35, 123] and Takeda et al (Fig. 7-26a).[122] Wolter's model also correctly predicts that Q_{BD} significantly decreases with increasing temperature (Fig. 7-26b).

In addition, Wolters argues that the critical parameter determining intrinsic oxide breakdown is Q_{BD}. He notes that while Q_{BD} is constant versus injected current density (J_{inj}) at low values of J_{inj}, its value decreases somewhat as J_{inj} rises (as shown by the curve fitted to the squares in Fig. 7-27a).[34] Thus, failure data from the four tests used to characterize oxide breakdown (ramp-voltage and ramp-current breakdown tests, and constant-voltage and constant-current TDDB tests) can only be meaningfully compared if the above facts are taken into account. That is, when any one of the four tests is

performed, the total charge transported through the oxide Q is related to J_{inj} in a different manner (as is indicated by the lines given in Fig. 7-27a). From Fig. 7-27a we see that, in general, the Q versus J_{inj} curve for each test can intersect the Q_{BD} curve at a different point (i.e., when $Q = Q_{BD}$ breakdown occurs). Thus, it is not surprising that the tests may appear to give conflicting results with respect to oxide failure. However, if all four tests are designed so that the charge transported through the oxide intersects the Q_{BD} versus J_{inj} curve at the same point, all of the tests should yield identical failure data. Figure 7-27b shows that this indeed the case. That is, apart from the early (B-mode) failures, the predicted failure rates of the intrinsic oxide are identical, regardless of the type test used to produce such data.

7.5.2 Qualitative Models for the Physical Origin of B-Mode Failures in Oxides

While intrinsic failures are easier to study, the B-mode failures shown in Fig. 7-14 are more important. This is because oxide electric fields in submicron CMOS fall in the range at which B-mode failures occur (e.g., when $t_{ox} = 12.5$ nm and $V_{GS} = 5.0$, $\mathcal{E}_{ox}$ is ~4.0 MV/cm). The B-mode failures of oxide films shown in Fig. 7-14 are believed to be due to defects in the oxide film that reduce its ability to withstand electrical stress. Such "walking wounded" oxide films will cause parts to fail prior to their designed lifetimes during normal operation. To build reliable systems, parts with such weaknesses must be screened out and excluded from use. Various types of defects have been proposed as being responsible for B-mode failures, including:

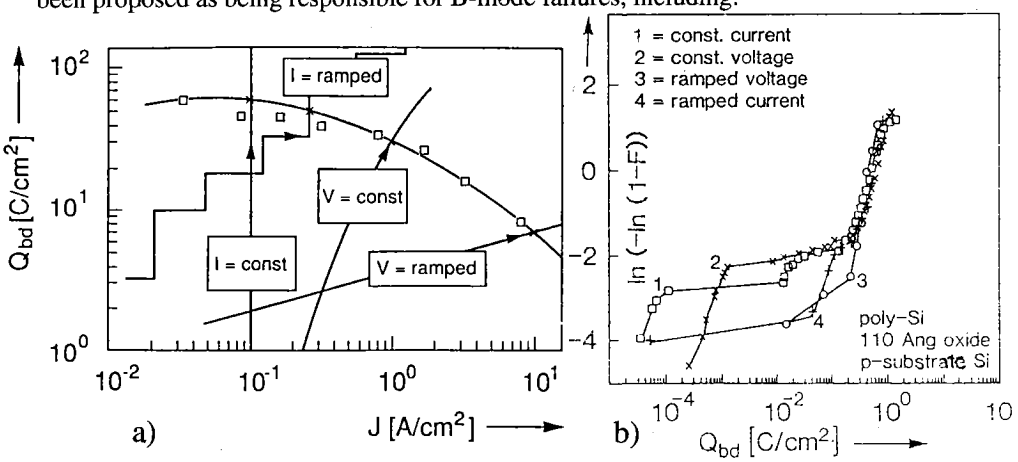


Fig. 7-27 (a) An example of the variation of Q_{BD} with injection current density. Lines in the figures depict the applied stress for different measuring techniques. In order to get comparable results, the lines should intersect the Q_{BD} curves at the same point. (b) Cumulative failure percentage *F* versus Q_{BD} for the various measuring techniques showing their correspondence. The differences in the tails are not only due to stochastical variations on the wafer but are caused by the different types of stress as well.[34] This paper was originally presented at the Spring 1990 Meeting of the Electrochemical Society, Inc. held in San Francisco, CA.

1. *Sodium contamination in the oxide film.* When a field is applied to the oxide, any positive ionic charges present in the oxide (e.g., Na^+) eventually drift toward the cathode. The proximity of these charges to the cathode increases the cathode field and enhances electron injection. This field increase continues as more charge builds up until the oxide ruptures.

It is thought that the origin of such Na^+ is the W-furnace filament. The sodium evaporates from the filament and then diffuses through the quartz furnace tube into the oxidation ambient. Periodic cleaning of the oxidation furnace tube by flowing a mixture of O_2 and a chlorine-containing gas (e.g., Cl, HCl, or TCA) through it for several hours (followed by an O_2 purge), is a standard technique used to reduce Na contamination from this source in the thermal oxide. Use of double-walled tubes (in which gas flowing in the outer tube carries away metal ions which have diffused through the outer wall), has also been reported to reduce Na (and other metallic) contamination.[36]

2. *Substrate metal contamination.* The presence of metals (especially Al, Fe, and Ca) on the Si surface prior or during oxide growth can significantly degrade the gate oxide properties. Metal contamination may exist in the starting material (in the form of heavy metal precipitates in the wafer),[37,38] or it may be introduced during processing. Examples of the latter include sputtering of stainless-steel fixtures and wafer holders by the beamline of ion implanters, diffusion of heavy metals from the heater coils through the quartz walls of diffusion and oxidation tubes (see above and Vol. 1, chap. 2), and metal-contaminated chemicals or DI water used in the cleaning process.[39] The latter reference also suggests adding complexing agents for Ca or Al to the cleaning solutions. An example is EDTA (ethylene diamine tetra acetic acid) which is an agent that binds metal ions (e.g., Ca and Al) in basic solutions through the unshared pair of electrons on the nitrogen atoms.

3. *Surface roughness.* With gate oxide films approaching 10 nm in thickness, uncontrolled surface roughness can significantly reduce the effective oxide thickness. A total Si-surface roughness of less than 5Å is now being specified as a requirement before gate oxide growth. This can be monitored using atomic-force microscopy, which enables 3-D subnanometer roughness measurements.

In the conventional RCA cleaning process used prior to gate oxidation, silicon surface roughening has been found to occur as a result of using $NH_4OH{:}H_2O_2{:}H_2O$ in the ratio of 1:1:5. If the process is modified, using instead a solution with less NH_4OH (e.g., ratios of ($NH_4OH{:}H_2O_2{:}H_2O$ = 0.05:1:5,[40] or 0.25:1:5[41]), the roughening is reported to become negligible.

Increased silicon surface roughness near the gate edge has also reportedly been measured with atomic force microscopy.[107, 108] This roughness is attributed to the high compressive stress caused by the field oxidation. The roughness degrades the oxide breakdown voltage. An annealing step after the field oxidation growth at 1000-1050°C (prior to the sacrificial gate oxide growth step) reportedly improves the breakdown voltage strength of the oxide films.

4. *Localized regions where the oxide is thinner.* Such thinning may be due to such causes as particulates embedded in the oxide during growth, the Kooi effect, or stress effects that locally reduce the oxide growth rate. The gate thinning problem is especially severe near the field edge due to the latter two effects (see Fig. 7-28a).[112, 130]

Note that stripping the sacrificial gate oxide in HF can result in impurities being left behind on the Si surface after the strip. These impurities cause a rough Si/SiO_2 interface that may thus result in a thinner SiO_2 film in some locations than others, and hence degrade the gate oxide reliability. Using ultra-high purity HF for this procedure appears to significantly improve the electrical breakdown yield.[42] The use of a stacked gate oxide (i.e., a two layer oxide, one layer thermally grown, and the deposited by CVD, see section 7.9.4.1) has also been reported to reduce this problem, since only the part of the gate oxide that is thermally grown is subjected to the thinning effects (see Fig. 7-28c).

5. *Crystalline defects in the substrate.* An example is the presence of SiO_2 precipitates near the surface of the Si.[43,129] These occur in Czochralski-grown silicon wafers containing oxygen dissolved in the lattice as result of the crystal growth process (see Vol. 1, chaps. 1 and 2). Unless the surface is made free of oxygen by a "denuding" step that out-diffuses the oxygen from the near surface regions, such precipitates can form near the surface during wafer processing steps. "Knock-on" oxygen from the implantation of heavy atoms (such as As) through screen oxides can also be responsible for the formation of such SiO_2 precipitates. Stacking faults are also possible crystalline-defect culprits. It has

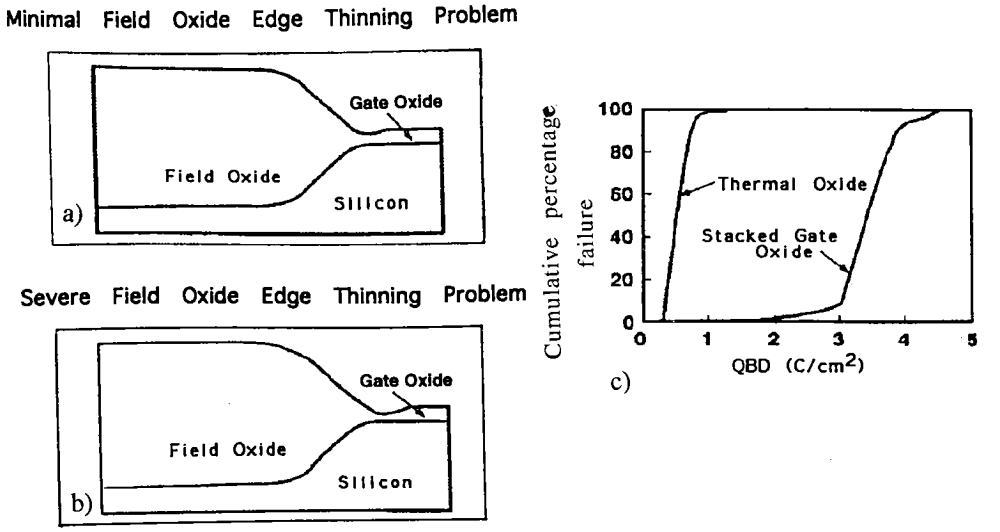


Fig. 7-28 Reduction of gate oxide thinning problem at field edge using stacked gate oxide. Nitride mask for PBL isolation is 2000Å thick.[112] (© IEEE 1993).

been reported that the presence of bulk defects which degrade gate oxide integrity can depend on the specific conditions of the ingot growth process (e.g., pull rate and thermal history of the ingot during the pulling process).

One measure that is reportedly effective in minimizing many of the defects listed above is the use of epitaxial- rather than bulk-Czochralski (CZ) wafers.[97,98] That is, epitaxial wafers show superior gate oxide integrity. Recently, a new high grade silicon bulk CZ wafer (which is termed a *Hi Wafer*) has been announced by Toshiba Semiconductor and Toshiba Ceramic.[99,100,124] In Hi-wafers a preprocessing high temperature anneal in high-purity hydrogen is used to form a zone denuded of oxygen near the wafer surface. Annealing in an inert ambient appears to result in lower surface oxygen levels than other ambients. Hi Wafers are reported to give gate oxide integrity results that are superior to conventional bulk CZ wafers and equal to those of epitaxial wafers (but at half the price of the latter).

7.5.2.1 Test Structure for Intrinsic Oxide Breakdown Measurements. It should be mentioned that to measure the true intrinsic properties of an oxide, an appropriate test structure is needed (i.e., one that does not degrade the measured intrinsic properties). Most conventional MOS-C test structures used for characterizing oxide breakdown follow the design rule of not connecting a metal line to the poly gate inside the active MOS area. Instead the metal/poly contact is made as shown in Fig. 7-29a. In this case, however, the poly gate, overlaps the field oxide around the periphery of the MOS-C (field-oxide surrounded MOS-C). It has been shown that when an advanced isolation technology, such as poly-buffered-LOCOS (see Chap. 6), or SWAMI[44] (see Vol. 2, chap. 2), a slight amount of gate oxide thinning may occur at the gate edge (i.e., up to 30% of the nominal gate oxide thickness, due to stresses involved with 2-D oxidation growth effects[45,46] (see also chap. 6, section 6.3.2). Since the poly passes over a rather long section of such a thinner oxide, the measured breakdown characteristics of the oxide film will appear degraded, compared to those in which no such thinner regions occurs. To circumvent this problem, the test structure can be modified so that the poly gate is not allowed to overlap the field oxide, and the contact between the metal and poly is as shown in Fig. 7-29b.[125] In addition, following the poly gate definition, a short reoxidation step is carried out to produce a gate bird's beak (or *graded-gate oxide* - see chap. 5). This smoothly increases the gate oxide thickness at the edge, preventing the gate edge from becoming a weak spot, in terms of dielectric strength. Thus, with this improved test structure, the intrinsic oxide breakdown characteristics can be measured independent of edge effects.

7.5.2.2 Damage at the Gate Periphery Resulting from the Polysilicon Etch Process. When the polysilicon gate is patterned by dry etching it is possible that during the overetch time the gate oxide may be removed from a small area adjacent to the gate periphery, even when the selectivity of the poly-to-SiO_2 of the etch process appears to be sufficiently high that the SiO_2 should not be completely removed. In such cases, the etch nevertheless "punches through" to the substrate and microtrenches are etched into the substrate near the gate periphery (Fig.

7-29c. This effect has been analyzed and modeled.[115] The cause of this microtrenching has been identified as ion scattering from the feature sidewalls, with a flux that is sufficient to sputter away the thin SiO_2 near the base of the sidewall and expose the substrate to the silicon etchant species.

The process engineer must minimize this damage to ensure adequate device reliability. Although the damage can be visually monitored with SEM/TEM, this is cumbersome. A special test structure has been suggested for electrical monitoring of such damage. It consists of an MOS-C with an additional polysilicon layer deposited and etched after the gate has been patterned (see Fig. 7-29d). This creates a conductive spacer over the region at the gate periphery and permits stress testing of the gate oxide at the periphery. If damage has occurred during the gate etch, such test-MOS-Cs will breakdown, or exhibit worse cumulative failure behavior than MOS-Cs without such conductive spacers over the damaged areas.[116] With the aid of such measurements the gate etch process can be adjusted to minimize such damage.

7.5.3 Empirical Quantitative Model of Intrinsic Oxide Breakdown

The exact physical origin of intrinsic oxide breakdown is still being debated. Nevertheless, the experimental evidence surrounding the breakdown event has led to the development of a statistical empirical model for oxide breakdown that allows the reliability of gate oxide films to be quite well estimated. The model can be applied to both "intrinsic" breakdown events (mode-C failures), and wearout events (mode-B failures).

The theoretical basis of the statistical reliability model for intrinsic failures is the following:[20] (1) Q_{BD} of intrinsic oxides (i.e., those containing no defects) is approximately constant (at ~10 C/cm^2) for relatively small Fowler-Nordheim tunneling current densities (i.e., less than 0.1 A/cm^2); and (2) Q_p is also constant at ~0.1 C/cm^2. Since $Q_p = J_g \alpha t_{BD}$ = constant (0.1 C/cm^2), where J_g is the Fowler-Nordheim gate current density, and α is the hole-generation coefficient, this implies that the breakdown time t_{BD} can be modeled as

$$Q_p \propto J_g \, \alpha \, t_{BD} \qquad (7 - 19)$$

where the gate current density due to Fowler-Nordheim tunneling is given by Eq. 7-1. Thus

$$J_g \propto \exp(-B/\mathcal{E}_{ox}) \quad \text{with } B = 240 \text{ MV/cm} \qquad (7 - 20)$$

and the hole-generation coefficient α is given by:

$$\alpha \propto \exp(-H/\mathcal{E}_{ox}) \quad \text{with } H \sim 110 \text{ MV/cm} \qquad (7 - 21)$$

where the oxide electric field, in terms of the voltage across the oxide V_{ox} and its thickness t_{ox}, is expressed as

$$\mathcal{E}_{ox} = V_{ox}/t_{ox} \qquad (7 - 22)$$

Therefore

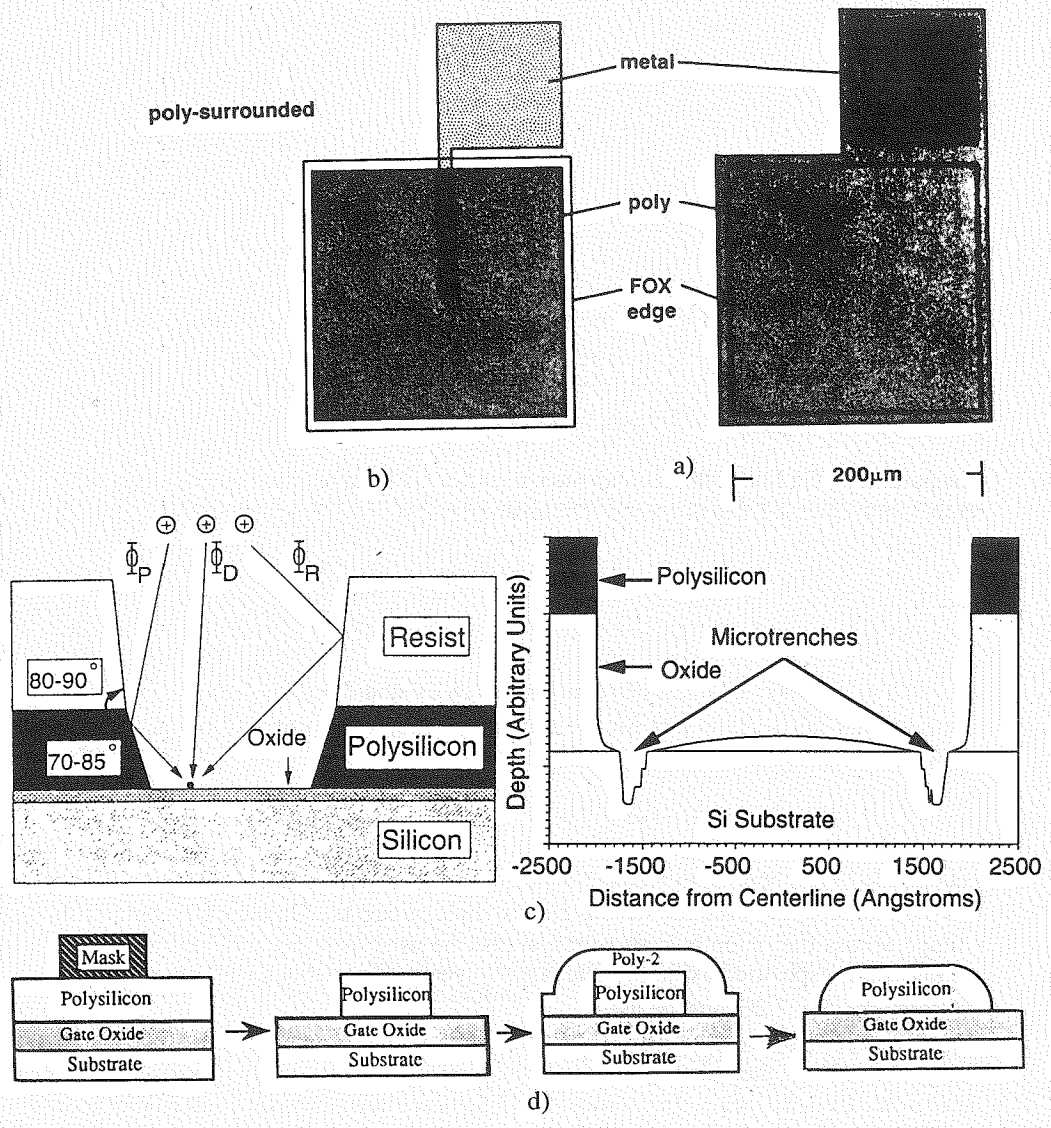


Fig. 7-29 (a) Layout of the conventional field-oxide-surrounded test structure with the gate poly expanded to the field-oxide area. (b) Layout of the gate-surrounded MOS capacitor.[125] (© IEEE 1992). (c1) Cross section of the film structure at polysilicon etching end point showing typical values of polysilicon and photoresist sidewall angles. (c2) Surface topography after etching. Note: microtrench locations correspond to the maxima in the flux profile.[115] Reprinted with the permission of the publisher, the Electrochemical Society. (d) Special test structure for electrical monitoring of such damage, consisting of an MOS-C with an additional polysilicon layer deposited and etched after the gate has been patterned.[116] This paper was originally presented at the Spring 1993 Meeting the Electrochemical Society.

$$Q_p \propto \exp(-G/\mathcal{E}_{ox})\, t_{BD} \qquad (7\text{-}23)$$

where

$$G = B + H \cong 350 \text{ MV/cm}. \qquad (7\text{-}24)$$

Thus, the time required to accumulate a certain amount of Q_p to cause breakdown t_{BD} is proportional to:

$$t_{BD} \propto \exp(G/\mathcal{E}_{ox}). \qquad (7\text{-}25)$$

The relation given by 7-25 can be made into an equation by finding the appropriate pre-exponential factor, which we will call $\tau_o(T)$. Then, in general, we can write

$$t_{BD} = \tau_o(T) \exp[G(T)/\mathcal{E}_{ox}] \qquad (7\text{-}26)$$

The parameter G (T) and the pre-exponential factor $\tau_o(T)$ have been shown to depend on temperature (over a limited range of temperatures: 25°C-150°C) in the following way[47]

$$G(T) = G[1 + \{(\delta/k)[(1/T) - (1/300)]\}] \qquad (7\text{-}27)$$

and

$$\tau_o(T) = \tau_o \exp[(-E_b/k)(\{1/T\} - \{1/300\}] \qquad (7\text{-}28)$$

where $\tau_o = 10^{-11}$ sec, G = 350 MV/cm, E_b = 0.28 eV and δ = 0.0167 eV.

Equation 7-26 is the model for the time-to-breakdown in intrinsic oxide films. At 300K, the equation is simply

$$t_{BD}(300K) = 10^{-11} \text{ (sec)} \exp[350 \text{ (MV/cm)}\, t_{ox} \text{ (cm)}/ V_{ox}(V)]. \qquad (7\text{-}29)$$

Equation 7-29 can be used to predict the time to breakdown of a defect-free oxide film of a given thickness at 300K for various electric field strengths. On the other hand, if the voltage across the oxide is known, it can also be used to select the minimum oxide thickness for a specified oxide lifetime, e.g., 10 years. To predict t_{BD} at other operating temperatures, Eq. 7-26 must be used.

The theoretically derived Eq. 7-29 has since been experimentally verified, as shown in Fig. 7-30. Here, the value of log t_{BD} for several thin oxides (ranging from 52 to 103 Å) is plotted versus $1/\mathcal{E}_{ox}$. In addition, Eq. 7-29 using a value of G = 350 MV/cm, and $\tau_o = 10^{-11}$ sec is also plotted. It can be seen that the agreement between the experimental data and Eq. 7-29 is very good. The value of τ_o is obtained by extrapolating the curve of the experimental points until they intercept the t_{BD} axis (see Fig. 7-31).

Equation 7-26 is also the basis for extrapolating the lifetime under normal operating conditions from the lifetime measured at higher-than-normal stress voltages (*accelerated testing*). That is,

$$t_{BD2} = t_{BD1}(\tau_o/ t_{BD1})^{(1 - V_{ox1}/V_{ox2})} \qquad (7\text{-}30)$$

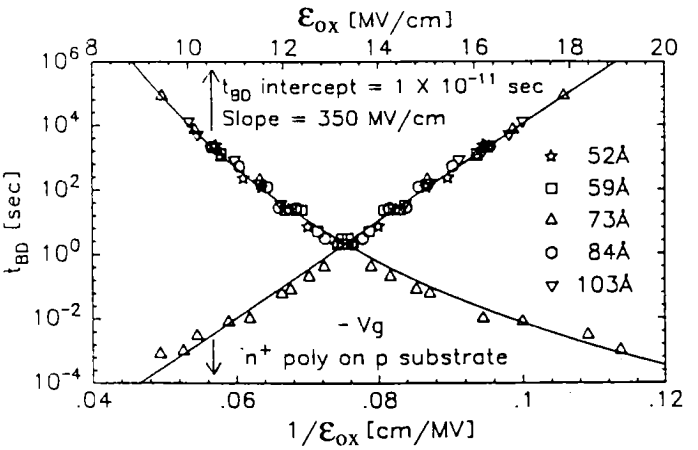


Fig. 7-30 Log (t_{BD}) versus $1/\mathcal{E}_{OX}$ for different gate oxide thicknesses. Log (t_{BD}) follows a linear relationship with $1/\mathcal{E}_{OX}$, not $\mathcal{E}_{OX}$. Small-area oxide samples (80 μm^2) were used such that intrinsic breakdown dominates and little statistical variation of t_{BD} at any one applied electric field was observed.[20] (© IEEE 1988).

Example 7-1: Use Eq. 7-29 to calculate the time to breakdown of 8 nm thick, defect-free oxide film if V_{ox} is 5V at 300K.

Solution: We use Eq. 7-29 to calculate t_{BD} with $t_{ox} = 8 \times 10^{-7}$ cm

$$t_{BD}(300K) = 10^{-11}(\text{sec})\, e^{350\,(\text{MV/cm})\, t_{ox}\,(\text{cm})/V_{ox}(V)} \qquad (7 - 31)$$

$$t_{BD} = 5.18 \times 10^{10} \text{ sec (i.e., in excess of 1000 years!).} \qquad (7 - 32)$$

Example 7-2: Calculate the minimum thickness of a defect-free oxide that could be used in a MOSFET that is to operate at 5.5V for 10 years at 150°C (425K) without suffering oxide breakdown.

Solution: Using Eqs. 7-27 and 7-28, the value of G(425K) = 283 MV/cm, and τ_0(425K) = 0.75 x 10^{-11} sec. With these values we solve for t_{ox} assuming t_{BD} = 10 years (3 x 10^8 sec), to get

$$t_{ox\,min} = 5.5V/(283\ \text{MV/cm})\ \ln(3\text{x}10^8\ \text{sec}/(.75\text{x}10^{-11}\ \text{sec}) = 88Å$$

Example 7-3: What is the maximum electric field that can be applied across a gate oxide so that is will operate for 10 years without suffering breakdown at 125°C.

Solution: At T = 125°C ≅ 400K, and t_{BD} = 10 years = 3.1 x 10^8 sec.

$$G(400K) = 297\ \text{MV/cm and}\ \tau_0(425K) = 0.75\ \text{x}10^{-11}\ \text{sec}$$

Thus, using Eq. 7-26 and solving for $\mathcal{E}_{ox}$ we get

$$\mathcal{E}_{ox} = V_{ox}/t_{ox} < 6.7 \text{ (or } \cong 7) \text{ MV/cm.}$$

It should also be noted that V_{ox} is the voltage across the oxide, which is generally different than the applied gate voltage V_{GB}. The difference is due to band bending and the work-function difference. The worst case occurs for a MOS-C with an n^+ polysilicon gate and a p-substrate (which also corresponds to the structure of the NMOSFET), when the MOS-C is biased into inversion. Here, $V_{ox} \sim (V_{GB} - 0.2V)$. For a MOS-C with an n^+ polysilicon gate and an n-substrate (or for a MOS-C with a p-substrate biased into accumulation), the oxide field for the same V_{GB} is smaller [$V_{ox} \sim (V_{GB} - 1.2V)$]. This explains why n-channel devices exhibit higher oxide failure rates than p-channel devices in CMOS ICs.[48]

Another point that should be emphasized concerns the upper limit of oxide field strength for accelerated tests. Since the log (t_{BD}) versus $1/\mathcal{E}_{ox}$ plot is linear, one might be tempted to use very high fields to obtain t_{BD} data in a very short time, and then use Eq. 7-26 to extrapolate the t_{BD} value at smaller fields. However, if $\mathcal{E}_{ox}$ becomes too large the Fowler-Nordheim current density begins to deviate from the theoretical J-$\mathcal{E}_{ox}$ curve (see Fig. 7-32, inset). The upshot is that testing at such excessive fields will lead to overly optimistic extrapolated lifetimes. To avoid this undesirable outcome a maximum oxide field $\mathcal{E}_{max}$ as a function of oxide thickness (for accelerated tests of "intrinsic" oxide breakdown), is suggested in Fig. 7-32.[49] By keeping $\mathcal{E}_{ox}$ below $\mathcal{E}_{max}$, a higher confidence in the accuracy of the extrapolated lifetime is ensured. Note that the

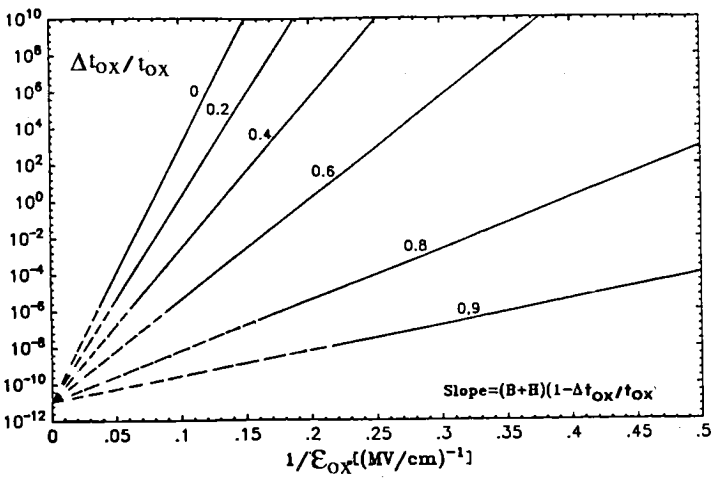


Fig. 7-31 An illustration of the dependence t_{BD} on the severity of defects Δt_{ox} as a function of the oxide field $\mathcal{E}_{ox}$, using Eq. 7-26. The value of τ_0 is obtained by extrapolating the curve of the experimental points of Fig. 7-30 until they intercept the t_{BD} axis.[20] (© IEEE 1988).

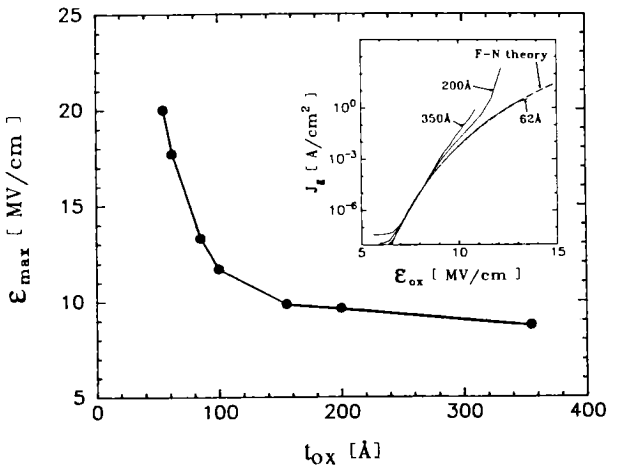


Fig. 7-32 Maximum recommended oxide field for extrapolating lifetime from log (t_{BD}) versus $1/\mathcal{E}_{ox}$ plots as a function of oxide thickness. At fields higher than $\mathcal{E}_{max}$, the J-$\mathcal{E}$ curve deviates from the theoretical Fowler-Nordheim values as shown in the inset. Fields should be kept to 20 percent below the $\mathcal{E}_{ymax}$ shown for testing of early failures.[49] (© IEEE 1987).

value of $\mathcal{E}_{max}$ given in this figure should be reduced by ~20% when accelerated field tests are used to study B-mode failures. This is necessary because the local oxide field at weak spots is higher than the nominal $\mathcal{E}_{ox}$.

On the other hand, Q_{BD} is independent of J and is therefore not affected by the distortion of the J-$\mathcal{E}_{ox}$ curve at high field. Thus, one can perform Q_{BD} measurements at high fields and extrapolate high-field Q_{BD} against $1/\mathcal{E}_{ox}$ to a low field. This low-field Q_{BD} can then be divided by the low-field current density to find the low-field lifetime $t_{BD} = Q_{BD}/J$. Using this approach it appears that highly accelerated testing (i.e., test times of only a few seconds) appears to be feasible, especially for very thin oxides.

7.5.4 Quantitative Models for B-Mode Failures

7.5.4.1 Reduced Barrier Height Model. In an early quantitative model of B-mode failures, Shatzkes et al. proposed that oxide defects may effectively lower the barrier for Fowler-Nordheim tunnel injection of electrons into the oxide conduction band. The breakdown mechanism, is then identical to intrinsic oxide breakdown at high fields. The breakdown distribution is merely shifted to lower fields because of the defect barrier lowering. Figure 7-33 shows a curve of oxide current versus electric field in such a defective oxide.[50] At low fields, more leakage current flows than would be expected in a defect-free oxide film (J_O). At low fields the leakage current characteristic is the same as if it were caused by Fowler-Nordheim tunneling if the oxide barrier was 2.4 eV instead of the normal 3.1 eV. Shatzkes found that capacitors with this excessive

leakage also suffered the intermediate field breakdown with far greater frequency than the less leaky MOS capacitors.

7.5.4.2 Oxide Thinning Model. The effect on the lifetime of oxides containing the B-mode failure defects listed in section 7.5.2 can be modeled by assuming that each such defect (as well as any others not listed) has the same effect on an "intrinsic" oxide film of a given thickness. That is, the presence of any one of these defects will cause an otherwise defect-free oxide to fail at a shorter time - corresponding to the time-to-failure of a thinner "intrinsic" oxide. Hence, any B-mode defect (weak spot) causes the oxide to behaves as a film having an effective thickness t_{oxeff} that is thinner than the nominal thickness t_{ox} by an amount Δt_{ox} (i.e., $t_{oxeff} = t_{ox} - \Delta t_{ox}$, see Fig. 7-34a).

Quantitatively, an oxide film of nominal thickness t_{ox} having no B-mode defects would fail at time t_{BD} given by Eq. 7-26. However, one containing a B-mode defect (that makes it appear to be thinner by Δt_{ox}) would fail at a time t_{BD} given by[20]

$$t_{BD} = \tau_o(T) \exp [G(T) (t_{ox} - \Delta t_{ox})/V_{ox}] = \tau_o(T) \exp [G(T) t_{oxeff}/V_{ox}] \quad (7 - 33)$$

The statistical variation of breakdown in the B-mode group is thus incorporated into the statistical distribution of Δt_{ox}. *That is, if we can determine the statistical variation of*

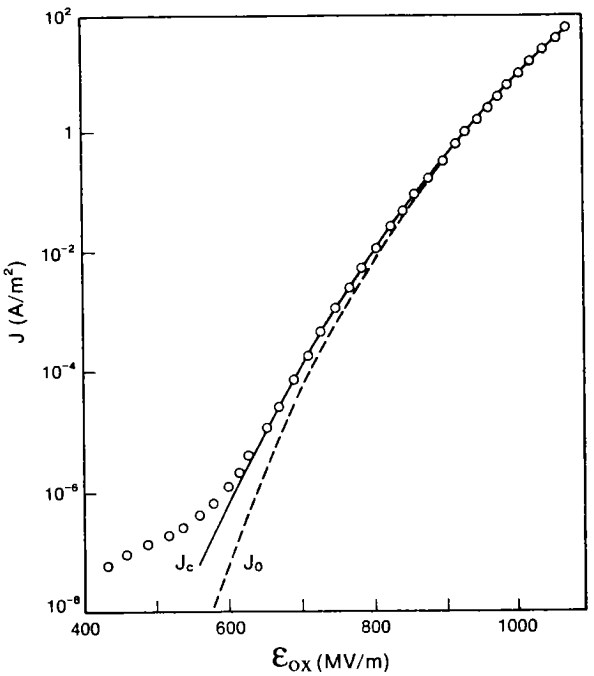


Fig. 7-33 J-V characteristics of defective oxide (J_C) compared to the theoretical J-V characteristics of an ideal oxide.[50] Reprinted with permission of IBM J. Research and Devel.

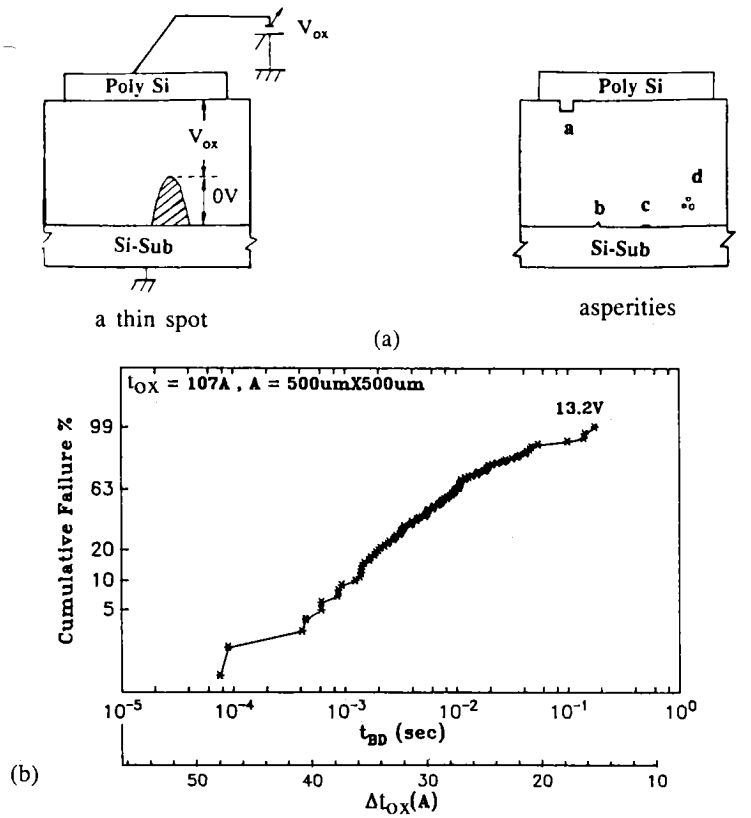


Fig. 7-34 (a) Oxide defects are modeled as effective oxide thinning by Δt_{OX}. b) Cumulative percentage failures versus time for a group of 0.25 mm^2 MOS-Cs with t_{OX} = 10.7 nm under a stress voltage of 13.2 V.[20] (© IEEE 1988).

B-mode defects in an oxide film we can estimate the failure rate and cumulative percentage of failures of such films under normal operating conditions. Note that the larger Δt_{OX} becomes, the sooner the oxide film will fail.

The statistical variation of Δt_{OX} in oxide films produced by a particular process in a specific fab environment can be inferred from time-dependent dielectric breakdown (TDDB) data of cumulative percentage failure versus time obtained from performing a constant-voltage stress test. For example, Fig. 7-34b shows the cumulative percentage failures versus time for a group of 0.25 mm^2 MOS-Cs (with an *n*-type substrate) with t_{OX} = 10.7 nm under a stress voltage of 13.2V. Using Eq. 7-34 we can find Δt_{OX} for each t_{BD} value Hence, from experimental data we can obtain a relation between cumulative percentage failure and Δt_{OX}. This relationship is also shown in Fig. 7-34. The data supports the notion that the short time to breakdown of the early failure samples corresponds to oxide samples containing spots with large Δt_{OX}.

7.5.4.3. The Cumulative Defect Density Function $D(\Delta t_{ox})$. To be able to project the lifetime of an oxide for arbitrary oxide areas, we need to obtain the *cumulative defect density as a function of* Δt_{ox} $D(\Delta t_{ox})$ [i.e., for a specific Δt_{ox}, $D(\Delta t_{ox})$ is the density of *all* the B-mode defects/cm^2 greater than Δt_{ox}]. Figure 7-35 is an example of a plot of $D(\Delta t_{ox})$ versus Δt_{ox}.

Basically, $D(\Delta t_{ox})$ is derived from the TDDB data of Fig. 7-34, but since obtaining $D(\Delta t_{ox})$ is the key to being able to predict the oxide reliability from experimental data, we will first outline the conceptual basis for how it is derived, and then give an example of how to actually calculate the values of $D(\Delta t_{ox})$ from a plot of cumulative percentage failure versus time.

Let us assume that oxide defects with varying degrees of severity are distributed over the wafer. Hence, the t_{BD} value for a particular oxide sample having a nominal thickness t_{ox} and area A is determined by the spot on that sample that has the largest Δt_{ox} value. For a group of samples each with the same oxide area, the percentage of samples that will fail below a specific t_{BD}, $P(t'_{BD}<t_{BD})$ is equal to the probability of finding a defect with an effective oxide thinning larger than Δt_{ox} in such an area ($P(\Delta t'_{ox}>\Delta t_{ox})$, or

$$P\,(t'_{BD}<t_{BD}) \;=\; P\,(\Delta t'_{ox}>\Delta t_{ox}) \qquad (7\text{ - }34)$$

where Δt_{ox} and t_{BD} are related through Eq. 7-33. The complement of this is the probability of not finding a defect with an effective thinning greater than Δt_{ox}, i.e.,

$$P\,(\Delta t'_{ox}>\Delta t_{ox}) \;=\; 1 \;-\; P\,(\text{no defect with effective thinning} >\Delta t_{ox}). \qquad (7\text{ - }35)$$

If the defects across a wafer are randomly distributed, implying a Poisson distribution can be used to estimate the probability, then

$$P\,(\text{no defect with effective thinning}>\Delta t_{ox}) \;=\; \exp\,(-\,A\,D[\Delta t_{ox}]) \qquad (7\text{ - }36)$$

where A is the oxide area, and $D(\Delta t_{ox})$ is the cumulative defect density function. Combining Eqs. 7-34, 7-35, and 7-36 we get

$$P(t'_{BD}<t_{BD}) = 1 - P(\text{no defect with effective thinning}>\Delta t_{ox}) = 1 - \exp[-AD(\Delta t_{ox})] \qquad (7\text{ - }37)$$

On the other hand, if the defects are nonuniformly distributed across a wafer, a Gamma distribution function can be used, and Eq. 7-37 becomes

$$P\,(\,t_{BD}'< t_{BD}) \;=\; 1 \;-\; \frac{1}{(1 \;+\; A\,D\,(\Delta t_{ox})\,S)^{1/S}} \qquad (7\text{ - }38)$$

where S is a measure of the degree of clustering. In the limiting case of S --> 0, Eq. 7-37 reduces to Eq. 7-36.

The example of Fig. 7-35 shows that the defect density with increasing Δt_{ox} gets smaller. This trend explains why MOS-Cs of larger area are more likely to incorporate

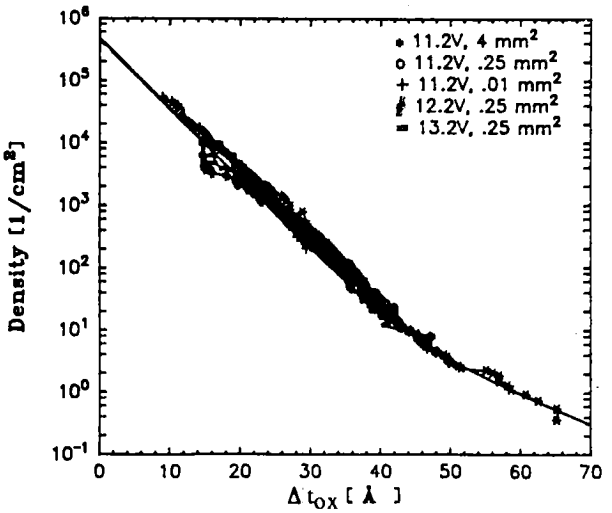


Fig. 7-35 $D(\Delta t_{OX})$ [i.e., Defect density as a function of Δt_{OX}], using Gamma distribution for the data in Fig. 7-34. Once $D(\Delta t_{OX})$ is derived for a given process, many reliability parameters such as yield, failure time, and optimum screen conditions can be determined.[20] (© IEEE 1988).

defects of larger Δt_{OX}'s and exhibit shorter t_{BD}s. This distribution of defect sizes is similar to the distribution of airborne particulates as a function of particle diameter.

Example 7-4: Calculate the value of $D(\Delta t_{OX})$ if: (a) $\Delta t_{OX} = 30$Å; and (b) $\Delta t_{OX} = 20$Å for the oxide technology that yields the TDDB data of Fig. 7-34, assuming the defects are nonuniformly distributed and S = 0.6. The area of the oxide samples is 25 mm^2.

Solution: Since the defects are assumed to be distributed nonuniformly, we use the Gamma distribution and Eq. 7-38 to extract $D(\Delta t_{OX})$. That is, we solve Eq. 7-38 for $D(\Delta t_{OX})$, using S = 0.6 and A = 0.0025 cm^2 = 25×10^{-4} cm^2, or

$$D(\Delta t_{ox}) = \frac{1}{A\,S}\left\{\left[\frac{1}{1 - P(t_{BD}' < t_{BD})}\right]^S\right\} \qquad (7\text{-}39)$$

Then, using Fig. 7-34 we see that

$$P(t_{BD}' < t_{BD}) \cong 0.5 \text{ for } \Delta t_{OX} = 30\text{Å},$$

and

$$P(t_{BD}' < t_{BD}) \cong 0.9 \text{ for } \Delta t_{OX} = 20\text{Å}.$$

{i.e., from Fig. 7-34, $t_{BD} \sim 5x10^{-3}$ sec for Δt_{ox} = 30Å. The probability that a failure will occur sooner is found from the cumulative failure percentage at $t_{BD} = 5x10^{-3}$ sec, which in Fig. 7-34 is ≅ 50%. Thus, P ($t'_{BD} < 5x10^{-3}$ sec) ≅ 0.5.}
Using these values in Eq. 7-38 we find that

$$D(\Delta t_{ox}) \cong 330/cm^2 \quad \text{for } \Delta t_{ox} = 30Å$$

$$D(\Delta t_{ox}) \cong 2000/cm^2 \quad \text{for } \Delta t_{ox} = 20Å.$$

Note that these values correspond to the values shown in Fig. 7-35, on which the calculated $D(\Delta t_{ox})$ values for many more Δt_{ox} points are also plotted for this oxide process.

7.5.4.4. Using $D(\Delta t_{ox})$ to Predict Oxide Breakdown Reliability.

Once the calculated $D(\Delta t_{ox})$ points are plotted as shown in Fig. 7-35, a closed-form equation describing $D(\Delta t_{ox})$ as a function of Δt_{ox} can also be obtained from this experimental data by curve fitting a mathematical function to the set of the plotted points. This fitted curve can be used to predict the cumulative percentage of failure curves and the failure rates for any oxide area and stress voltage.

For example, the data in Fig. 7-35 can be curve fitted by the following equation:

$$D(\Delta t_{ox}) = a_1 \exp(b_1 \Delta t_{ox}) + a_2 \exp(b_2 \Delta t_{ox}) \qquad (7 - 40)$$

where $a_1 = 13.1$, $a_2 = 6.3$, $b_1 = -0.26$, $b_2 = -0.11$ are fitting parameters. This equation is shown plotted in Fig. 7-35 as the solid line. Assuming a nonuniform distribution of defects, Eq. 7-38 can be used to formulate the cumulative distribution function F (expressed in terms of the percentage of failure) as a function of the oxide area, the oxide field, and time

$$F(A, V_{ox}, t_{BD}) = 1 - \frac{1}{\left[1 + A\left(a_1 e^{b_1 \Delta t_{ox}} + a_2 e^{b_2 \Delta t_{ox}}\right) S\right]^{1/S}} \qquad (7 - 41)$$

where Δt_{ox} is related to V_{ox} and t_{BD} through Eq. 7-33. In other words, it is possible to derive F for a specific oxide thickness from a set of measurements at a single oxide stress voltage V_{ox} and oxide area A. This function F can then be used to estimate the cumulative percentage failure versus time for any area or stress voltage (for a specific oxide thickness and set of process conditions). In Fig. 7-36a, the theoretical F for various areas is plotted as the solid lines, and the experimental data as the points (with t_{ox} = 107Å). In Fig. 7-36b the theoretical F is plotted for various stress voltages for this same oxide thickness. Agreement between the estimated and the experimental values of F is reasonably good.

The instantaneous failure rate λ (t) can also be predicted with the aid of Eq. 7-41. For example, a common specification of reliability of an electronic system is to have a failure rate of less than 0.01 percent per 1000 device hours (i.e., 100 FITs) before shipping the product to customers. From Eqs. 7-7, 7-26, and 7-41, one obtains

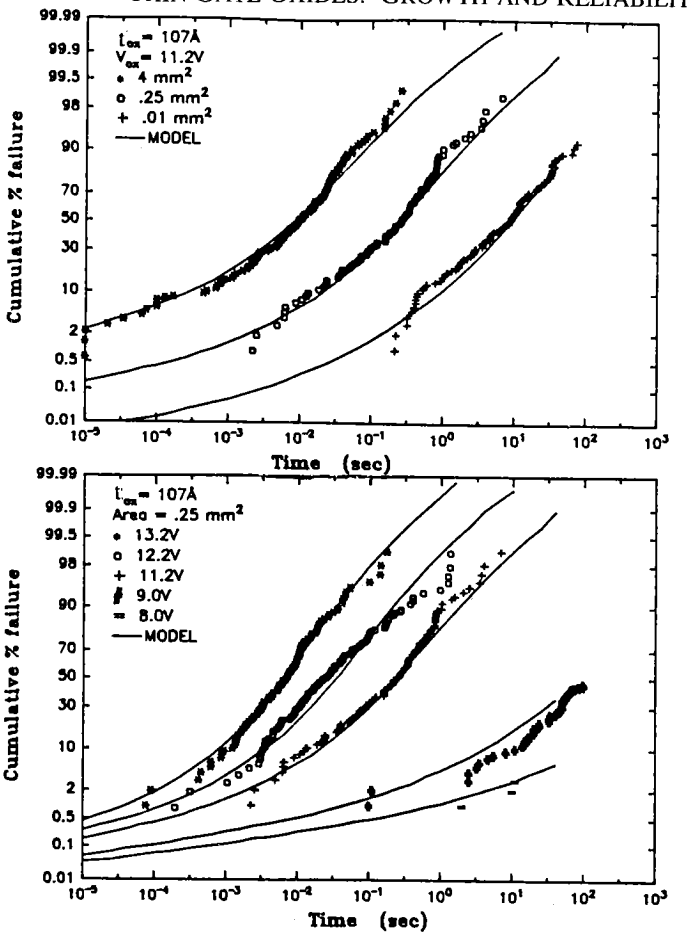


Fig. 7-36 Cumulative percentage failure versus time for (a) different areas and (b) different stress voltages of 10.7 nm oxide. Solid curves are derived from the defect density and Δt_{ox} models. Reasonable agreement between the data and model prediction is obtained.[20] (© IEEE 1988).

$$\lambda(t) = \frac{A}{1 + ADS}\left(a_1 b_1 e^{b_1 \Delta t_{ox}} + a_2 b_2 e^{b_2 \Delta t_{ox}}\right)\frac{V_{ox}}{t\,G} \qquad (7 - 42)$$

for the Gamma distribution, and with S = 0, for the Poisson distribution. Figs. 7-37a and 7-37b show the experimental and calculated failure rates. Reasonable agreement is again observed.

Equations 7-26, 7-41 and 7-42 constitute a quantitative model for the breakdown failure of oxides. That is they allow the following failure parameters of an oxide in a particular IC to be calculated: (a) Eq. 7-26 is used to find the oxide lifetime (t_{BD}) of a film with an effective oxide thickness t_{oxeff}, for a specific set of operating conditions (V_{ox} and T); (b) the cumulative failure of this oxide F (A, V_{ox}, t_{oxeff}) as a function of

time for a given oxide area and stress voltage is estimated from Eq. 7-41; and (c) the failure rate λ (t) is calculated using Eq. 7-42.

It should also be noted that the $D(\Delta t_{ox})$ curves can be used to monitor the quality of the oxide growth process. As discussed earlier, it is expected that the quality of the oxide should depend on such factors the wafer quality (i.e., defect and metal precipitate densities), the gettering conditions, the cleanliness of the process, and the oxide growth and annealing conditions. By comparing the $D(\Delta t_{ox})$ curves of two different processes or wafer lots, one can obtain a quantitative measure of the oxide breakdown characteristics. Examples of applying the reliability model of oxide breakdown to such cases are given in section 7.5.5.

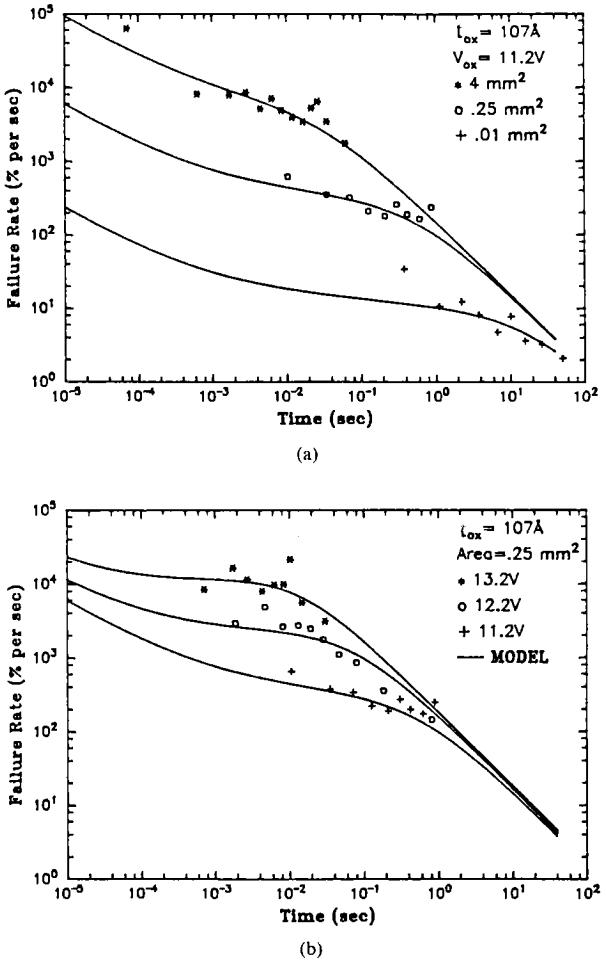


Fig. 7-37 Failure rate vs. time for (a) different areas and (b) different stress voltages of 10.7 nm oxide. Solid curves are derived from the defect density and Δt_{ox} models.[20] (© IEEE 1988).

7.5.4.5. Ramp-Voltage Breakdown Test for Obtaining $D(\Delta t_{ox})$ Data. Data for extracting $D(\Delta t_{ox})$ curves can be obtained using the ramp-voltage breakdown test instead of the constant-voltage TDDB test. Since the ramp-voltage test is much shorter than the constant-voltage test, this approach is very attractive. Note that a slow-ramped voltage (1-2V/sec) is used rather than the fast-ramped technique (100-200 V/sec) used to obtain dielectric strength (as described in section 7.4.1). A slow ramp allows TDDB failures to occur due to traps in the gate oxide, and yet each MOS-C can be tested in a matter of seconds.

The following equation gives the breakdown criteria in terms of the ramp-voltage test parameters.[51]

$$\{V_{BD}^2/RG\ (t_{ox} - \Delta t_{ox})\} \exp\ (-\ G\ [t_{ox} - \Delta t_{ox}]/V_{BD}) = \tau_0 \qquad (7-43)$$

where R is the ramp rate in volts per second, and V_{BD} is the ramp-breakdown voltage, and τ_0 is the pre-exponential term in Eq. 7-26 (i.e., the value of τ_0 is 10^{-11} sec).

In general, Eq. 7-43 has to be solved iteratively to find the value of $[t_{ox} - \Delta t_{ox}]$ corresponding to a measured V_{BD}. However, since the $[t_{ox} - \Delta t_{ox}]$ dependence of the exponential term in Eq. 7-43 dominates over the $[t_{ox} - \Delta t_{ox}]$ dependence of the pre-exponential coefficient, it is possible to obtain a reasonably accurate closed-form solution for $[t_{ox} - \Delta t_{ox}]$ by setting $[t_{ox} - \Delta t_{ox}]$ in the pre-exponential term to 80Å. Then Eq. 7-43 can be solved approximately for $[t_{ox} - \Delta t_{ox}]$ in the following form:

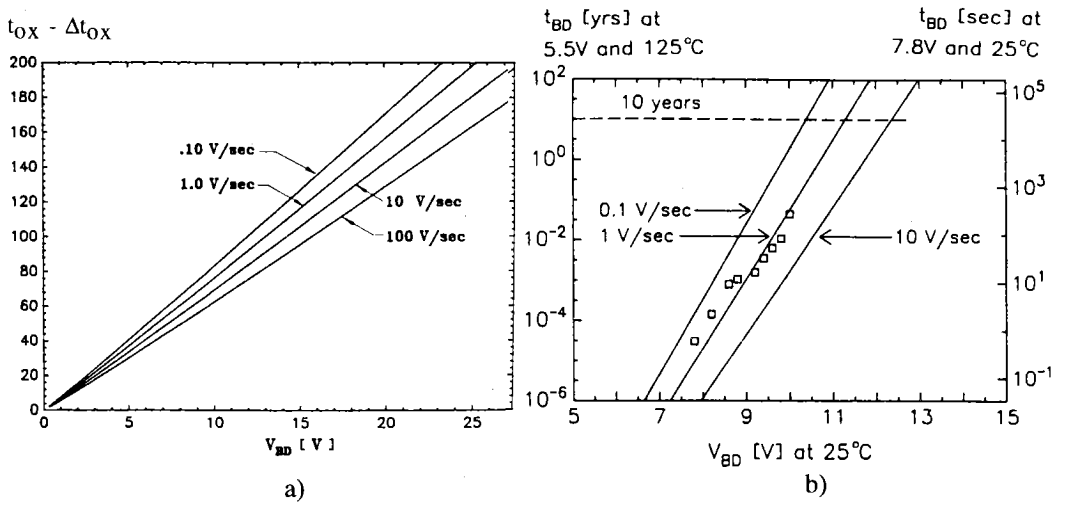


Fig. 7-38 (a) $t_{ox} - \Delta t_{ox}$ versus ramp breakdown voltage V_{BD} as a function of ramp rate. Using this figure, the D(Δtox) curve can be derived from a ramp-voltage test.[20] (© IEEE 1988). (b) Correlation between t_{BD} and V_{BD} as predicted by Eq. 7-46. The experimental data were taken at 7.8V (for a constant voltage stress) and 1.0 V/s (for the ramp voltage stress) for a 125Å oxide [n^+ poly on (100) p substrate, biased in accumulation]. Each data point represents the breakdown voltage and time-to-breakdown for a specific cumulative failure percentage.[47] (© IEEE 1990).

$$(t_{ox} - \Delta t_{ox}) = \frac{V_{BD}}{G\,(T_{Vbd})} \ln\left(\frac{V_{BD}^2}{R\,G\,(T_{Vbd})\,\tau\,(T_{Vbd})\,80\text{Å}}\right) \qquad (7 - 44)$$

Figure 7-38a plots Eq. 7-44 (i.e., $[t_{OX} - \Delta t_{OX}]$ versus V_{BD} with ramp rate as the parameter). From such a plot, one can then relate V_{BD} to Δt_{OX}. Then, using a cumulative failure versus V_{BD} plot, one can extract $D(\Delta t_{OX})$. Once $D(\Delta t_{OX})$ is obtained, all the reliability predictions discussed earlier can be performed.

7.5.4.6. Correlating Ramp-Voltage Test Data and Constant-Voltage TDDB Data. It has been proposed that the correlation between V_{BD} of the ramp breakdown test and the t_{BD} of a constant voltage test can be established.[47] That is, by substituting Eq. 7-44 into Eq. 7-33, an approximate relationship between V_{BD} of the ramp voltage test (at temperature T_{Vbd}) and the t_{BD} of the constant voltage test (performed at temperature T_{tbd}) is found

$$t_{BD} \cong \tau_o\,(T_{tbd}) \left| \frac{V_{BD}^2}{R\,G(T_{Vbd})\,\tau_o\,(T_{Vbd})\,80\text{Å}} \right|^{G(T_{tbd})V_{BD}/\,G(T_{Vbd})\,V_{ox}} \qquad (7 - 45)$$

For the typical case where the ramp voltage test is performed at room temperature (T_{Vbd} = 25°C and T_{tbd} = T), Eq. 7-45 reduces to

$$t_{BD} \cong \tau_o\,(T) \left| 3.6 \times 10^8 \left(\frac{1}{V\,sec}\right) \frac{V_{BD}^2}{R} \right|^{[G(T)\,/350]\,(V_{BD}\,/V_{ox})} \qquad (7 - 46)$$

where G(T) (in units of MV/cm) and $\tau_o(T)$ are given by Eqs. 7-27 and 7-28. This equation predicts that oxides with a lifetime of 10 years at 5.5V and 125°C will exhibit failure in a ramp-voltage test at V_{BD} ~ 11.5 V at 1 V/sec and a lifetime of 100 years at 125°C corresponds to $V_{BD} \cong$ 12 V at 1 V/sec.

The theoretical correlation between V_{BD} and t_{BD} using Eq. 7-46 at T_{Vbd} = 25°C and T_{tbd} of 125°C and V_{OX} = 5.5V is plotted in Fig. 7-38b for various ramp rates (the left-hand axis represents these theoretical t_{BD} values). The experimental data for 125Å oxides correlating V_{BD} to the t_{BD} when the oxides are stressed at 7.8 V is also plotted, using the right-hand axis for representing the experimental t_{BD} values, with each point representing V_{BD} and t_{BD} for a specific cumulative percentage failure. (For example, the lowest experimental point corresponds to the V_{BD} and t_{BD} values at 10 percent cumulative failure in Figs. 7-39a and b.) Very good agreement between the theoretical and experimental data is seen. This implies that one can theoretically predict the lifetime of an oxide operating at different voltages and temperatures just by performing a simple ramp test at room temperature and carrying out the calculations needed to produce the curves as shown in Fig 7-38b. For example, the curves in Fig. 7-38b predict that an oxide sample (with any thickness and area) having any combination of V_{BD}s and ramp rates above the dashed line, will have a lifetime of more than 10 years at 5.5V.

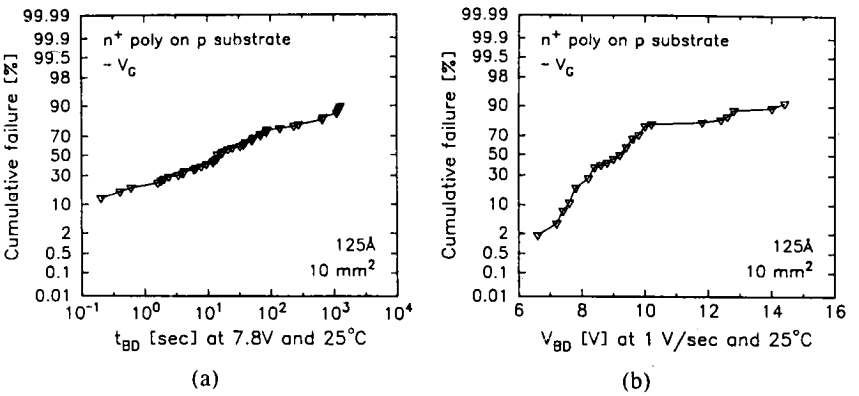


Fig. 7-39 (a) TDDB and (b) ramp breakdown data for 10 mm^2 125Å samples [n^+ polysilicon gate on (100) p substrate, biased in accumulation].[47] (© IEEE 1990).

7.5.4.7 Interface Trap Generation as a Measure of Oxide Damage Due to Charge Injected into the Oxide. It has been observed that the passage of tunneling current through a thin gate oxide also introduces interface traps at the Si/SiO_2 interface. Furthermore, while breakdown may not occur until about 10 C/cm^2 of charge has passed through the oxide, interface traps are detectable even after 0.001 C/cm^2 of charge passage. Thus, measuring interface traps has been proposed as a more sensitive means of quantitatively measuring oxide damage prior to breakdown due to Fowler-Nordheim stressing of oxides.[52,53] Good agreement between TDDB values predicted from trap densities in the oxide caused by voltage stressing and measured TDDB values was found.[103] A physical model that describes how the traps are generated inside the oxide by high voltage stressing and how their presence in the oxide leads to wearout and breakdown in thin silicon oxides has also been developed.[104] The physical trap generation model was quantified to describe the measured breakdown distributions using a statistical model, the mathematical details of which are described in ref. 105.

Quasi-static CV curves of oxides before and after passage of charge through the oxide show that the CV curves after such stressing are degraded compared to the CV curves prior to stressing (Fig. 7-40), and the degree of degradation is correlated to the increase in trap density in the oxide. The more charge passed through the oxide, the more severe is the degradation. In section 7.10.4 we will describe how this method has been applied to measure oxide damage due to plasma etching.

7.5.5 Applications of the Oxide Breakdown Reliability Model

7.5.5.1 Projecting the Reliability of Gate Oxides with Various Thicknesses. The oxide breakdown reliability model described in this chapter can

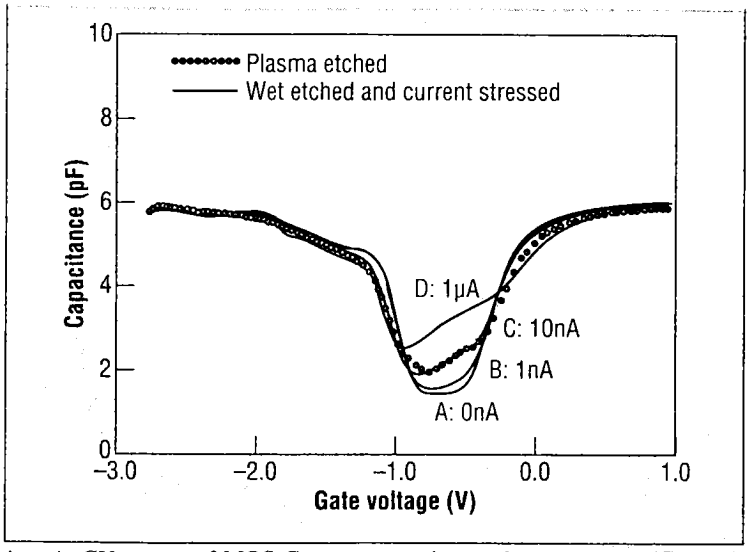


Fig. 7-40 Quasi-static CV curves of MOS-C structures prior to charge passage (Curve A = 0 nA), and after charge passage (Curves B - D). The creation of interface traps by current passage through the gate oxide degrades the CV curves.

be used to estimate and compare the reliability of different oxide technologies or thicknesses. Here we give an example of this procedure.

Example 7-5: Establish a set of curves which will allow the lifetime of various thin gate oxides of different thicknesses (e.g., 50-200Å thick) and arbitrary areas to be projected from accelerated stress data (assuming all are produced in the same process environment).

Solution: A group of MOS-Cs is fabricated for each oxide thickness of interest, with the gate oxide areas in all groups being equal. Then, the following procedure is used to estimate, for arbitrary area, the projected lifetimes of each of the oxide thicknesses:

1. Each group of MOS-Cs with a specific oxide thickness is subjected to ramp-voltage breakdown tests. The results, i.e., the cumulative percentage of failure versus V_{BD} for each oxide thickness are plotted (Fig. 7-41).

2. The experimental data from the previous step is used to calculate $D(\Delta t_{ox})$ versus t_{oxeff} for each oxide thickness. The procedure for obtaining $D(\Delta t_{ox})$ is:

a. For each value of V_{BD} for a given oxide thickness, we solve either Eq. 7-43, or Eq. 7-44 (the approximate form of Eq. 7-43), to get t_{oxeff}.

b. Using the calculations from part 2(a) we plot the cumulative percentage of failures versus t_{oxeff}.

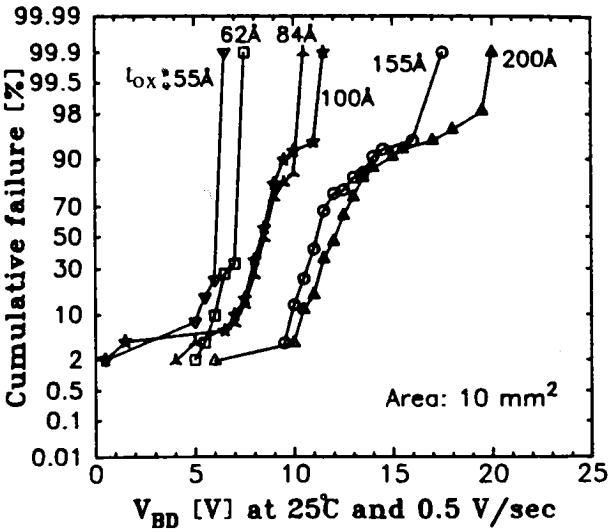


Fig. 7-41 Cumulative failure ramp breakdown data for 10 mm^2 oxide samples [n^+ polysilicon gate (100) p substrate, biased in accumulation].[47] (© IEEE 1990).

c. We then find $D(\Delta t_{ox})$ versus t_{oxeff} for each oxide thickness. This is done by first finding the percentage of failures for a given t_{oxeff} from the plots generated in part 2(b), and then using Eqs. 7-37 or 7-38 to calculate $D(\Delta t_{ox})$ for that value of t_{oxeff} (using the area of the test MOS-Cs).

d. The values of $D(\Delta t_{ox})$ are then plotted versus t_{oxeff} for each oxide thickness (see Fig. 7-42). The experimental plot of $D(\Delta t_{ox})$ has now been obtained.

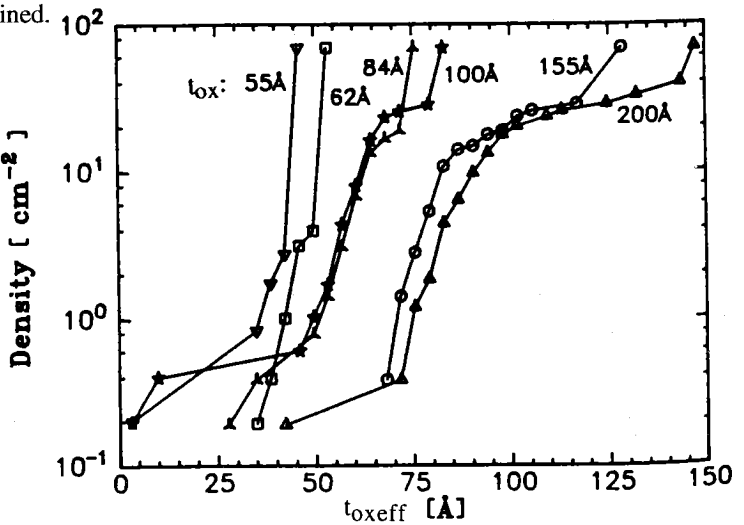


Fig. 7-42 Defect distribution for the oxide samples of Fig. 7-41.[47] (© IEEE 1990).

3. Now that $D(\Delta t_{ox})$ is available for each t_{oxeff}:

a. For a given operating temperature, we can calculate t_{BD} for each t_{oxeff} from Eq. 7-33.

b. From part 2, for a given t_{oxeff} we know $D(\Delta t_{ox})$. So from Eq. 7-37 or 7-38 we can determine the cumulative percentage of failures for any other specific oxide area.

c. Following the procedure of part 3b, the cumulative percentage of failures vs t_{BD} is now plotted (i.e., a point can be calculated for each t_{oxeff}).

4. From these curves we can find the cumulative percentage of failure for the various oxide thicknesses at that area for various times (Fig. 7-43). As an example, note that the cumulative failure for the 155Å oxide after 10 years is projected to be approximately 5%.

7.5.5.2 If a MOS IC is to have a Specific Lifetime (e.g.,10-Years), what is the Minimum Oxide Thickness that can be Used? (Assume the Maximum Cumulative Failure Percentage, Operating Temperature, and Oxide Area are Specified.)

Here we outline the procedure for selecting the minimum oxide thickness that will ensure a specific maximum cumulative failure rate over a 10-year lifetime at a specific operating temperature in a part that has a known gate oxide area.

1. Assume the $D(\Delta t_{ox})$ versus t_{oxeff} for each oxide thickness of interest has been established, as described in the previous section.

2. The maximum electric-field strength that allows defect-free (intrinsic) oxides to operate for a 10-years at 125°C is ≅ 7 MV/cm (see section 7.5.3). If an oxide has a defect such that its effective oxide thickness $t_{oxeff} = t_{ox} - \Delta t_{ox}$ is reduced from the nominal thickness of t_{ox}, we must derate the maximum allowed electric

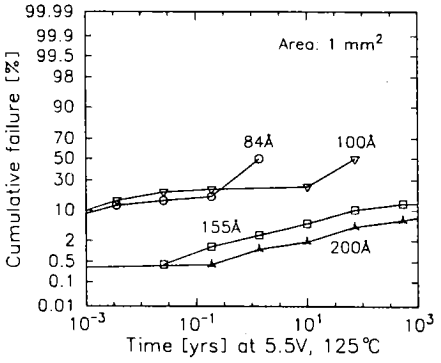


Fig. 7-43 Projected reliability during 5.5V operation is calculated from the defect density curves shown in Fig. 7-42.[47] (© IEEE 1990).

field and oxide voltage to obtain a 10-year lifetime at 125°C operation for this weakened oxide. That is, the maximum electric-field strength (as measured by V_{ox}/t_{ox}) must be derated to $\mathcal{E}_{max}$ = 7 MV/cm (t_{oxeff}/t_{ox}), and the maximum voltage that can be applied to such a defective film having a nominal oxide thickness t_{ox} must be reduced to V_{max} = (7 MV/cm) t_{oxeff} (cm).

3. For a specified cumulative failure percentage over 10 years (e.g., 3% cumulative failure) at some operating temperature (e.g., 125°C) and specific oxide area (e.g., 1 mm^2), Eq. 7-37 is used to calculate the value of $D(\Delta t_{ox})$ needed to ensure this cumulative failure percentage for that specific oxide area. (Eq. 7-37 is used because we also assume the defects are distributed uniformly). As an example, solving Eq. 7-37 for $D(\Delta t_{ox})$ indicates that a 3% cumulative percent failure in 10 years for a 1 mm^2 oxide corresponds to a $D(\Delta t_{ox})$ of 3 defects/cm^2 on the wafer.

4. Next, we use the plot of $D(\Delta t_{ox})$ versus t_{oxeff} to determine the value of t_{oxeff} corresponding to this value of $D(\Delta t_{ox})$. In this case, from Fig. 7-35 we see that a value of $D(\Delta t_{ox})$ = 3 defects/cm^2 corresponds to a t_{oxeff} of 55Å, and since the oxide under study was 125Å thick, Δt_{ox} is found to be 70Å.

5. Once the value of t_{oxeff} is known, we can calculate the $\mathcal{E}_{max}$ or V_{max} that can be applied to oxide films of this area at various thicknesses, using the relations given in step 2. On the other hand, if it is known that a specific voltage will be applied to the oxide (e.g., 5.5V), we can use the value of Δt_{ox} to calculate t_{oxeff} and then derate $\mathcal{E}_{ox}$ (intrinsic) or V_{ox} (intrinsic) to $\mathcal{E}_{max}$ or V_{max} accordingly. For example, if we know that at 125°C, a 10-year life can be had for an $\mathcal{E}_{ox}$ (intrinsic) of 7 MV/cm, then the thinnest intrinsic oxide that can live for 10 years at 125°C under a 5.5V stress is t_{ox} ~ 80Å. Since Δt_{ox} = 70Å, the thinnest oxides we can use (with an area of 1 mm^2) that will give a 10 year lifetime at 125°C with a maximum 3% cumulative failure are 150Å thick.

6. The calculation performed in part 5 can be repeated for other voltages to yield other minimum oxide thicknesses. In addition, other defect density values can be used when performing this calculation. A set of points can be plotted using this procedure, as is done in Figs. 7-44 to yield the minimum oxide thickness for a given oxide voltage, with the defect density (cm^{-2}) as a parameter. Note that if a higher defect density (corresponding to a higher cumulative failure rate) can be tolerated, these figures imply that a higher operating voltage could be used.

A final remark about oxide integrity models should be interjected. The structures for studying gate oxide integrity are MOS-Cs, in which the electric field is uniform across the entire device area. In MOSFETs, however, the oxide field can varies spatially along the channel due to the increase in the channel voltage from source to drain. Thus, one can speculate whether nonuniform injection would lead to a spatially non-uniform Q_{BD}

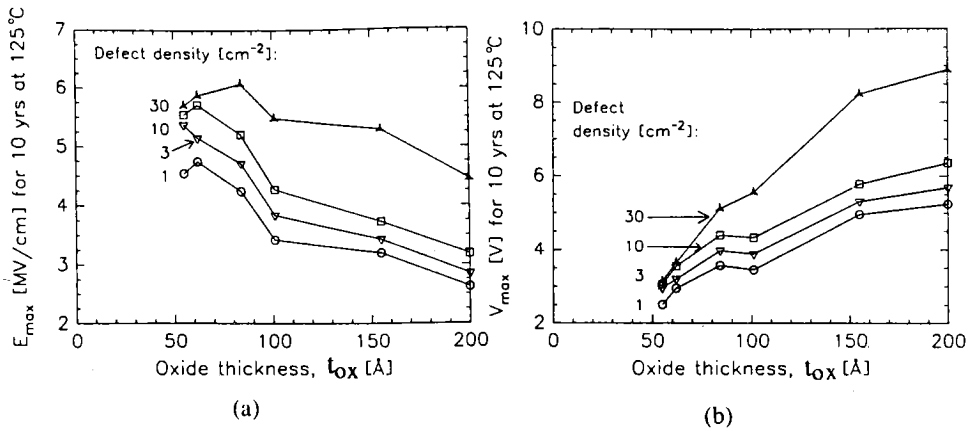


Fig. 7-44 Given a desired average failure rate and the oxide area, the largest acceptable defect density is calculated from Eq. 7-37. Using this defect density, it is possible to determine (a) the maximum allowable oxide field and (b) voltage for different oxide thicknesses or technologies.[47] (© IEEE 1990).

in MOSFETs. This might also yield different gate oxide integrity results in MOSFETs as compared to those obtained from tests using MOS-Cs.

7.6 SCREENING TESTS FOR OXIDE BREAKDOWN (BURN-IN)

As noted above, oxides in the B-mode group will fail at times sooner than those in the C-mode group (intrinsic oxides). *Burn-in* is a common technique used to screen out B-mode group parts. In this procedure, the intent is to operate the devices for some period of time, during which most of those parts that would undergo early failure actually do fail. Failing while undergoing burn-in is preferable to having the devices fail after they are installed in a system and delivered to the customer. By excluding such weak parts the reliability of the installed system can be significantly improved.

In general, burn-in conditions are selected so that the failure mechanisms that contribute to the early failures are presumably accelerated. In the case of oxide breakdown, this is achieved by burning-in under higher than normal voltages and temperatures. That is, application of high field and elevated temperature yields a much more efficient burn-in procedure than straightforward stress under normal operating conditions.

For example, Fig. 7-45 shows the theoretical failure rate curves at 5V predicted from the model for the parts that exhibit the cumulative defect density function of Fig. 7-35. These curves indicate that the failure rate of such products would stay higher than the specification of 0.01% per 1000 device hours (100 FIT) for more than 100 years. That is, only after 100 years of operation would a sufficient number of the weaker devices have expired so that the failure rate after that time would be less than 100 FIT.

To decrease the failure rate to an acceptably low rate more quickly, the failure of these weakest devices must be hastened by applying above-normal stress, i.e. screening.

The instantaneous failure rate of parts that fail by a mechanism that exhibits a log-normal distribution decreases monotonically in time. Early failure (B-mode) oxides have been shown to exhibit such log-normal behavior. Thus, the device failure rate (measured in this case as the fraction of the instantaneous population of oxides breaking down per unit time), decreases as the stress time increases. Figures 7-45 and 7-46 both depict typical experimental data expressing the instantaneous failure rate of a group of oxide capacitors as a function of time (that have not been subject to a burn-in screen).[54] The fact that the instantaneous failure rate decreases with time is the logical basis of the burn-in screen tests. That is, if a reliability engineer specifies a maximum allowable rate of failure of constituent MOS capacitors of, say, 100 failure units (100 FIT), the population may be burned-in until the instantaneous failure rate drops below this level.

Figure 7-47 presents an example of how such a burn-in procedure reduces the failure rate of parts under normal operating conditions.[55] The instantaneous failure rate of two groups of oxides is shown. In one group (black dots), a continuous stress of 2 MV/cm (which we assume to be the normal operating electric-field stress) is applied, while in the other group (white circles), a 2.5 MV/cm stress is applied for the first second of the test. Then the stress is reduced to 2 MV/cm. During this first second, the failure rate in the second group is increased. Upon dropping the stress to its normal operating value (2 MV/cm), however, the failure rate drops to a value well below that of the first group. That is, it falls to a value that is not attained by the first group until a much later time.

Unfortunately, while a longer burn-in time lowers the failure rate, it also increases the yield loss (i.e., the fraction of the parts that fail during burn-in). Thus, it is necessary to determine the optimal set of burn-in conditions that minimizes the burn-in time, yet

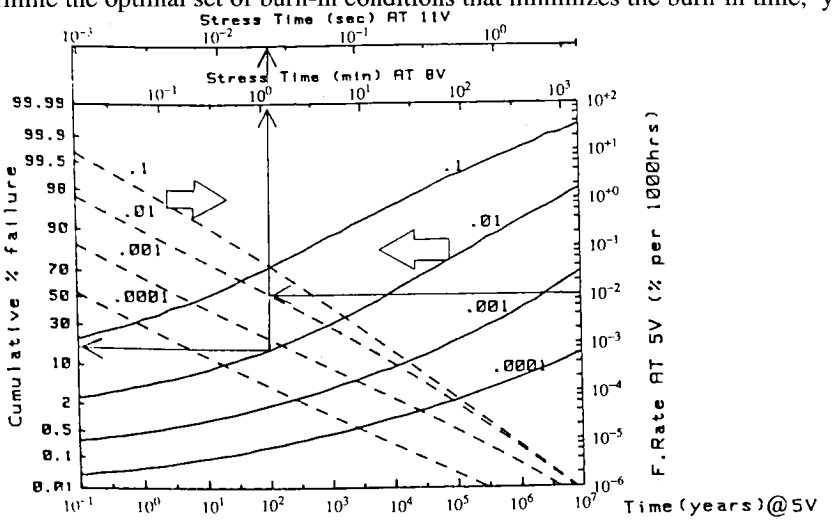


Fig. 7-45 Theoretical curves of failure rate and cumulative percent failures for different areas at 5-V operation. These curves are used for screening and screen yield optimization.[20] (© IEEE 1988).

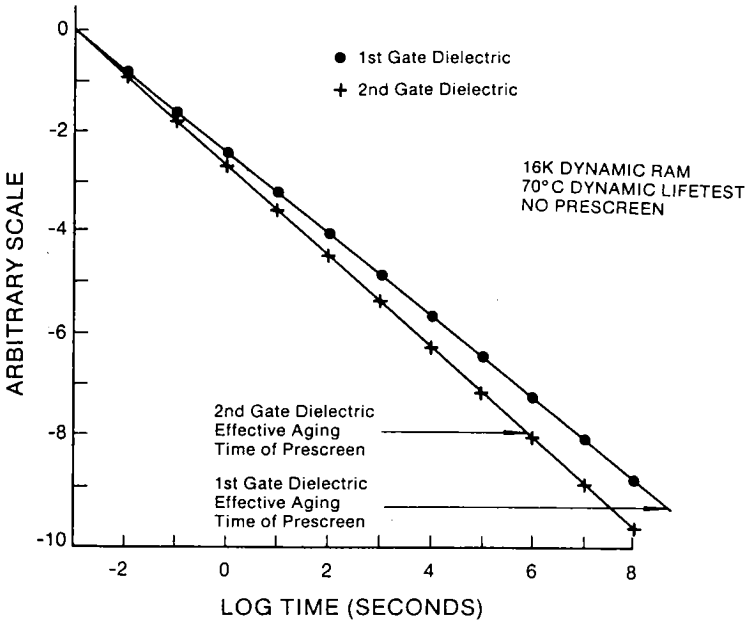


Fig. 7-46 Instantaneous failure rate data.[54] (© IEEE 1979).

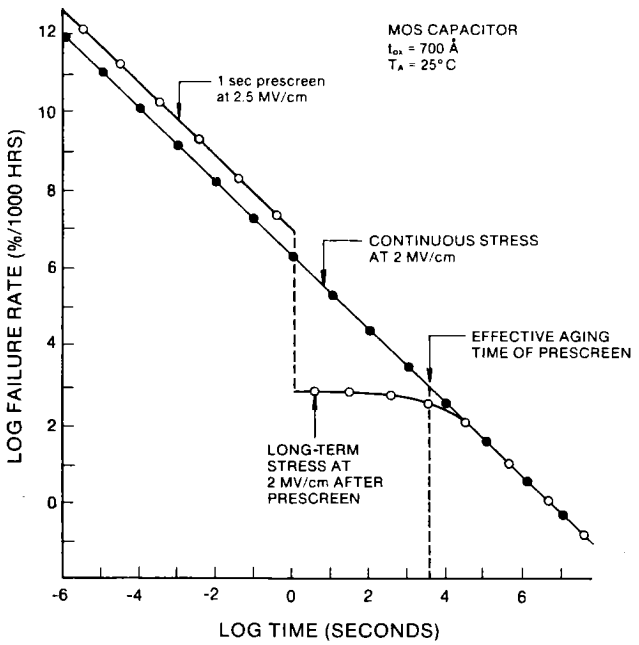


Fig. 7-47 Instantaneous failure rate data showing the effect of aging stress at elevated field.[54] (© IEEE 1979).

adequately lowers the failure rate with minimal yield loss. However, burn-in duration has often been set arbitrarily or determined by experimental results which are applicable to only certain lots of devices. It would be better to use a model that simulates the burn-in effects of various sets of burn-in conditions. Here we describe such an approach.

In order to determine the appropriate stress time, stress field, and stress temperature for optimally screening oxides of a given area, operating voltage, and operating temperature, an accurate quantitative model of predicting the failure rate and TDDB behavior is required. In fact, the model for t_{BD} due to B-mode failure described above has been applied to burn-in condition selection. Using this model, the reliability of oxides [characterized by a cumulative defect distribution $D(\Delta t_{ox})$] following a given burn-in condition can be projected for any oxide area, operating voltage, and temperature.[47]

7.6.1 Modeling the Effect of the Burn-In Stress on the Post-Burn-In Reliability of the Oxides.

With the addition of a burn-in step, oxide breakdown during post-burn-in operation occurs as a result of damage incurred during both the burn-in and the normal operation of the device. The subsequent lifetime of a device after a specific burn-in has been carried out t_{op} is calculated from

$$t_{op} = \tau_o(T_{op})\, e^{G(T_{op})\, t_{ox\,eff}/V_{op}} \left[1 - t_{BI}\, e^{-G(T_{BI})\, t_{ox\,eff}/V_{BI}}\right] \qquad (7\text{-}47)$$

where t_{BI} is the burn-in time, V_{BI} is the burn-in voltage, V_{op} is the operating voltage, and T_{BI} and T_{op} are the burn-in and operating temperatures, respectively. Comparing Eq. 7-47 with Eq. 7-33 we see that the first two product terms of Eq. 7-47 give the original oxide lifetime, i.e., the lifetime without burn-in t_{BD}. Thus, the quantity in square brackets represents the fraction of the original oxide lifetime remaining after burn-in, L_{BI}

$$L_{BI} = \left[1 - t_{BI}\, e^{-G(T_{BI})\, t_{ox\,eff}/V_{BI}}\right] \qquad (7\text{-}48)$$

This quantity (fraction of the lifetime remaining after burn-in, L_{BI}) is plotted in Fig. 7-48 as a function of the ramp breakdown voltage (using Eq. 7-43 to find the value of V_{BD} corresponding to the value of t_{oxeff} in Eq. 7-48), for the case of a burn-in test performed for 10 hours at 150°C at various burn-in temperatures. The expected lifetime at 5.5V and 125°C before burn-in is found using Eq. 7-33.

Plots such as those in Fig. 7-48 are used to determine the fraction of oxides that will be damaged by the burn-in to a degree that they will exhibit a significantly shorter lifetime after the burn-in (e.g., oxides whose life after burn-in is less than 90% of their original life). As an example, Fig. 7-48 indicates that after a burn-in of 7V for 10 hours at 150°C, only those oxides which originally would have exhibited a V_{BD} between 11.5 and 12V are predicted to have had their lifetimes reduced to less than 90% of their pre-burn-in value (note that those oxides which would have had $V_{BD}<11.5$V would not have survived this burn-in). The same conclusion is drawn when comparing the projected lifetime after burn-in versus the original lifetime before burn-in as

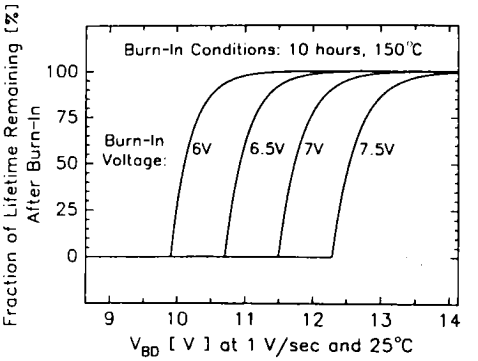


Fig. 7-48 The effect of a given burn-in condition on an oxide technology can be estimated based on a simple ramp voltage test. For example, following a 10-h burn-in at 7V and 150°C, all oxides with breakdown voltages below 11.5V are expected to fail. Of the oxides surviving burn-in, only those with $11.5V<V_{BD}<12V$ are expected to exhibit a reduction of more than 10% from their original lifetime (the lifetime without burn-in).[47] (© IEEE 1990).

shown in Fig. 7-49. Of the oxides that survive the burn-in, only the ones having lifetimes within one half of one decade have significantly reduced lifetimes (i.e., a reduction of more than 10% from the original lifetime) after burn-in. With the aid of Eq. 7-48, the optimal burn-in conditions can now be determined.

7.6.2 Procedure for Selecting an Optimal Set of Burn-In Conditions for Oxide Breakdown Screening

Here we outline the general procedure for selecting optimal burn-in conditions. Then we give an example which illustrates how the steps of this procedure are carried out.

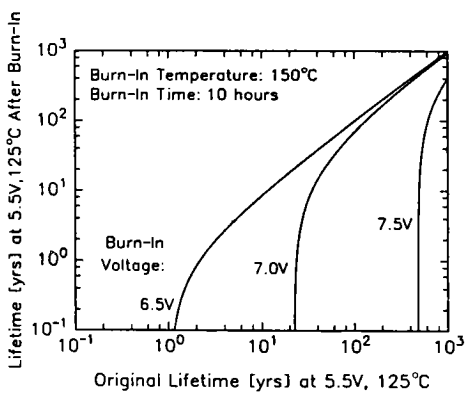


Fig. 7-49 Replot of Fig. 7-48 in terms of lifetime under operating conditions.[47] (© IEEE 1990).

We recall that the goal of burn-in is to *lower the failure rate* so that: (1) it becomes constant over the subsequent normal operating lifetime; and (2) its value becomes small enough to meet the reliability specification of the system. In addition, the yield loss due to burn-in should be minimized. The general procedure for establishing the design curves from which optimal burn-in conditions can be selected is the following:

1. When an optimal set of burn-in conditions is to be selected, it is assumed that the following parameters are given: the nominal oxide thickness t_{ox}, the oxide area A in the circuit being fabricated, the operating voltage V_{ox}, the operating temperature T_{op}, the expected lifetime after burn-in, and the allowable cumulative percent failure after burn-in over the expected lifetime of the oxides.

2. A plot of cumulative percent failure versus V_{BD} for oxide structures with the given t_{ox} and area A is first established from a ramp-voltage breakdown test.

3. The lifetimes t_{BD} (at the specified T_{op} and V_{op}) versus V_{BD} from Eq. 7-46 (using ramp rates of 0.1V/sec, 1V/sec, and 10V/sec as the parameter) are then calculated and plotted as shown in Fig. 7-38b.

4. A burn-in temperature T_{BI} is now selected (e.g., 150°C).

5. Equation 7-46 is then used to calculate the V_{BD}'s that correspond to oxide burn-in lifetimes t_{BI} for various burn-in voltages V_{BI} and the selected T_{BI}. That is, V_{BD} values for oxide lifetimes ranging from t_{BI} = 1 sec to 10^6 sec (in decade intervals), and V_{BI} values ranging from 5-10V, might be calculated.* If this group of calculated V_{BD} values is plotted versus V_{BI} (with t_{BI} as the parameter), a set of design curves is produced. An example of such a set of curves is shown in Fig. 7-50 for the case of T_{BI} = 150°C, T_{op} = 125°C, V_{op} = 5.5V, and R = 1V/sec. (Note that at this point readers might not yet be able to interpret this figure, as it also contains other relevant information that still needs to be described).

6. A set of plots of L_{BI} versus V_{BD} {as shown for example in Fig. 7-48 for the case of T_{BI} = 150°C, $T(V_{bd})$ = 25°C, R = 1V/sec, and various burn-in voltages} are next computed. The values for such plots are calculated using Eq. 7-48, with a separate set of curves calculated for each t_{BI} used in step 5. A set of operating-lifetime-after-burn-in versus original-lifetime plots for each t_{BI} (with V_{BI} as a parameter) should also be developed (see Fig. 7-49).

7. Next, on the same set of design curves developed in step 5, we also plot the value of V_{BD} that pre-burn-in oxides would have to exhibit in order to be able to survive both the burn-in and the expected lifetime after burn-in. This is done as follows:

* One might also wish to compute the V_{BD}s for $t_{BI} = 3.6 \times 10^4$ sec, and plot these points on the design curve, as this corresponds to a burn-in time of 10 hours (the curve for this time is not plotted in Fig. 7-50).

a. First we calculate the value of V_{BD} that an oxide would have to exhibit if it only had to survive the stress of normal operation (i.e., no burn-in). This is done using the curves generated in step 3. As an example, if an oxide was operated at 125°C and 5.5V, Fig. 7-38b indicates it would need to exhibit a V_{BD} = 11.3V to ensure an operating lifetime of 10 years (with no burn-in).

b. Second, from the plots developed in step 6 (see for example, Fig. 7-49), we note that for low values of V_{BI} (e.g., 6V or less in Fig. 7-49), the lifetime after the burn-in remains essentially the same as for the case of no burn-in. Thus, for the case of an oxide operated at 125°C and at a voltage of 5.5V, any oxide exhibiting a $V_{BD} \geq 11.3V$ will not have its original lifetime altered by burn-ins at 150°C for 10 hours and burn-in voltages smaller than approximately 6V.

c. We next superimpose another set of curves on the design curves of step 6. These are the curves that indicate the minimum original V_{BD} value that an oxide would have to exhibit if it was to survive both the burn-in and the expected operating lifetime (i.e., oxides exhibiting original V_{BD} values above this curve would be such survivors). The value of V_{BD} at the intercept of the vertical axis is the V_{BD} value needed by an oxide to survive merely the normal operating stress (in Fig. 7-38b equal to 11.3V). However, as V_{BI} is increased, the burn-in eventually starts to damage some oxides to the degree that their post-burn-in lifetime is significantly reduced. The value at which this begins to occur is determined from the curves of step 5 as follows: We see in Fig. 7-49 that at V_{BI} = 6.5V for a 10 hour burn-in at 150°C, the original lifetime of a 11.3V oxide is reduced by ~10%, and at V_{BI} = 7V it is reduced by 100% (i.e., all the oxides exhibiting original V_{BD} values smaller than 11.4V would be eliminated by the burn-in). Thus, in this case, the minimum original V_{BD} an oxide would have to exhibit in order to survive both the burn-in and a 10-year lifetime is greater than 11.4V. To find this value (which would become one of the points on these curves), we use Figs. 7-38b and 7-49. That is, from Fig. 7-49 we see that for an oxide to have a 10-year life after a 7V, 10-hour, 150°C burn-in, its original (i.e., pre-burn-in) projected lifetime would have to be 30-years. From Fig. 7-38b we see that this corresponds to an original V_{BD} of 11.6V. This value would become one of the points we would plot on the set of design curves of Fig. 7-50.

8. From the now-completed set of design curves we can determine a number of things about a burn-in test. First, from the design curves (such as those shown in Fig. 7-50) we can see that samples with original breakdown voltages between the dashed and solid lines in this figure are projected to fail during normal operating conditions after burn-in. Those with breakdown voltages above the solid lines are projected to satisfy the specified operating lifetime after burn-in. Thus, it is apparent that in order to achieve a low post-burn-in failure rate, a set of burn-in conditions (voltage, time, and T) should be chosen such that there is a small difference between the V_{BD} corresponding to the burn-in and the V_{BD}

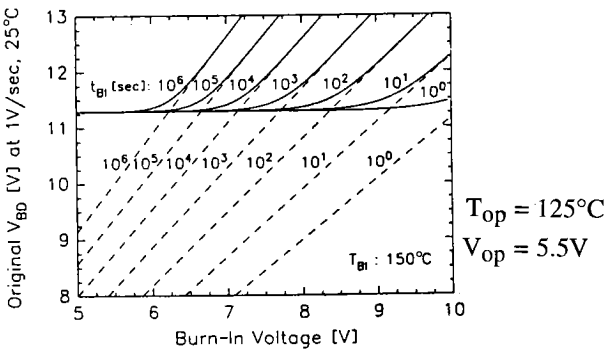


Fig. 7-50 Burn-in conditions can be optimized using the design curves shown here. The optimal burn-in condition at T_{BI} = 150°C can be chosen using these curves and a breakdown voltage distribution characterizing the oxide technology.[47] (© IEEE 1990).

corresponding to the normal operating condition plus burn-in. Second, there are many combinations of voltage and time that can produce an equivalent difference between these values. In addition, the values of V_{BI} and t_{BI} to achieve such goals can be selected from the burn-in design curves described above.

9. After a set of burn-in conditions is chosen, we now have sufficient information to characterize the selection. That is, from the set of selected burn-in conditions and the design curves generated for the particular oxide technology and reliability specifications, we can determine: (1) the range of original V_{BD} values that will be screened out by the burn-in test; (2) the range of V_{BD} values that will correspond to early failures among the survivors of the burn-in.

10. Using the values from step 10 and the cumulative percent failure versus V_{BD} plot, one can determine: (a) the yield loss due to burn-in; and (b) the fraction of oxides that, having survived the burn-in, would still fail prior to the expected lifetime (cumulative percent failure during post-burn-in lifetime).

7.6.3 Example of Applying the Procedure for Determining an Optimal Set of Burn-In Conditions

Assume we are asked to choose a set of conditions for burning-in a set of ICs having a nominal gate oxide thickness of 125Å and a gate oxide area 10 mm^2. The expected operating lifetime is 10 years at an operating temperature of 125°C. The specified cumulative percentage failure after burn-in during the operating lifetime is 2%. How do we select a set of burn-in conditions to meet this goal?

First, we carry out a ramp-voltage breakdown test (with R = 1V/sec) on a set of MOS-Cs having the specified t_{OX} and oxide area at 25°C. The resulting cumulative percentage failure versus V_{BD} data is plotted as shown in Fig. 7-38a. The t_{BD} versus V_{BD} for the normal operating conditions (and T_{VBD} = 25°C) is computed and plotted as

shown in Fig. 7-38b. If we choose a T_{BI} of 150°C, and a t_{BI} of 10 hours ($3.6x10^4$ sec), the V_{BD} as a function of V_{BI}, T_{BI}, and t_{BI} can computed using Eq. 7-46. The computed points of V_{BD} for these chosen burn-in parameters are plotted as the dashed line in Fig. 7-51.

Next, the L_{BI} versus V_{BD} curve, and the curve of lifetime after burn-in versus lifetime prior to burn-in (for several burn-in voltages and t_{BI} = 10 hours) are plotted from Eq. 7-48, as shown in Figs. 7-48 and 7-49. From Fig. 7-48, we see that Eq. 7-48 predicts that if oxides are subjected to a 10-hour burn-in at 150°C at 7V, all those with $V_{BD} \leq 11.4V$ will fail during burn-in, and hence will be screened out.

The values of V_{BD} (as a function of V_{BI}) that would have to be exhibited by the pre-burn-in oxides to be able to survive the burn-in *and* a 10-year lifetime at 5.5V at 125°C are then computed (with the aid of Figs. 7-48 and 7-49). These points are then plotted on Fig. 7-51 as shown by the solid line. (Note that for an oxide to have a projected 10-year lifetime at 5.5V and 125°C, it must exhibit an original V_{BD} = 11.3V). We see that at V_{BI} = 7V, the V_{BD} that would have to be exhibited by the oxides is 11.6V.

Based on the selected values of the burn-in and the information graphed in Fig. 7-51, we can draw the following conclusions about the burn-in procedure. First, since yield loss from burn-in is determined by the fraction of the total oxide population with a breakdown voltage of ≤11.4V, we estimate from Fig. 7-38a that the burn-in yield loss will be 75%. Second, after this particular burn-in, Fig. 7-51 indicates that only oxides with $11.4V < V_{BD} < 11.6V$ are projected to fail during the subsequent 10-years of operation at 5.5V at 125°C. From Fig. 7-38a, we see that the percentage of oxides that exhibit V_{BD} between 11.0 and 12.V is very small, implying that the burn-in will probably achieve the goal of a cumulative percent failure of 2% after burn-in.

The corresponding average failure rate is determined by the fraction of the surviving population with $11.4V < V_{BD} < 11.6V$. Therefore, it is clear, that the burn-in condition

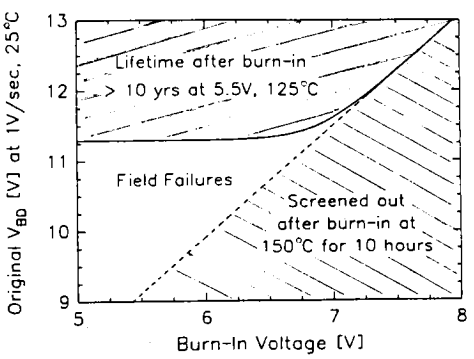


Fig. 7-51 Samples with breakdown voltages below the dashed line are screened out after burn-in; samples with breakdown voltages between the dashed line and the solid line are projected to fail during operation at 5.5V and 125°C and those with breakdown voltages above the solid line are projected to satisfy a 10-year lifetime specification.[47] (© IEEE 1990).

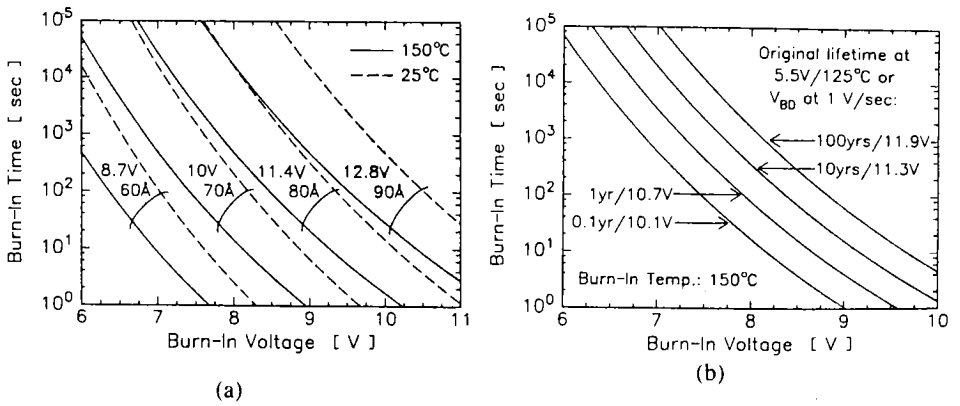


Fig. 7-52 Once an optimal burn-in condition is chosen using Fig. 7-50, equivalent burn-in conditions can be determined using the curves shown here. For example, burn-in conditions which screen out all oxides with breakdown voltages below 10V, i.e., t_{oxeff}<70Å (see the 10V/70Å) curve for 150°C burn-in (a) or the 0.1-year lifetime/10.1V curve in (b) will have the same burn-in yield and projected reliability.[47] (© IEEE 1990).

has to be chosen such that the cumulative percentage that will fail during the operating lifetime (in this case that fraction of the population that exhibits failures with 11.4V<V_{BD}<11.6V) is within the failure specification while keeping the yield loss from the burn-in as low as possible. If, for a given technology, a large fraction of the oxides have 11.4V≤V_{BD}≤11.6V (80.5Å<t_{oxeff}<81.5Å), as determined by taking the difference between the values of F (V_{BD} = 11.4V) and F (V_{BD} = 11.6V) on the cumulative percent failure versus V_{BD} plot, further screening is necessary to reduce the failure rate. Conversely, an oxide technology may have such a low fraction of oxides with V_{BD}≤11.6V (i.e., the defect density may be very low) that no screening is necessary.

Also note that since three variables must be selected (burn-in time, temperature, and voltage), it is possible to obtain the same burn-in yield and projected failure rate for many combinations of burn-in time, temperature, and voltage. As an example, if the optimal burn-in condition is determined from Fig. 7-50 to be 7V, 150°C, and $3x10^4$ sec (corresponding to a breakdown voltage of 11.4V), all equivalent burn-in conditions are given by the 11.4V/80Å line in Fig. 7-52a or the 11.3V/10-year line in Fig. 7-52b. In other words, any burn-in voltage and time combination on the 11.4V/80Å line in Fig. 7-52a, or the 11.3V/10-year line in Fig. 7-52b can be used to screen out those oxides characterized by V_{BD}≤11.4V or t_{oxeff} <80Å (i.e., to screen out those oxides that are projected to fail within 10 years at 5.5V at 125°C). Since it is desirable to select burn-in conditions which minimize testing costs, the shortest burn-in time (i.e., the maximum burn-in voltage) is usually picked. This maximum allowable voltage is usually set by such circuit or technology constraints as the junction breakdown voltage.

7.7 OXIDE BREAKDOWN MODULE IN BERKELEY RELIABILITY TOOL (BERT)

The oxide reliability model described above has been implemented in the *Circuit Oxide Reliability Simulator* (CORS) program,[56] which is part of a larger reliability simulator software package BERT.[57] CORS helps determine if the oxide quality for a particular process is high enough to meet the specified circuit lifetime under expected operating conditions with a sufficiently low failure rate. Thus, it should help circuit designers determine if a particular design is feasible for a given oxide process. The user is required to provide ramp-voltage breakdown statistics or time-to-breakdown statistics of test MOS-Cs. The simulator uses this data to derive $D(\Delta t_{ox})$ for the particular technology.

The CORS program generates statistics about circuit failures due to MOS oxide breakdown. It predicts the probability of circuit failure as a function of operation time, temperature, power-supply voltage, and input waveforms. In addition to supplying TDDB data, the user creates an input deck similar to an ordinary SPICE input deck. CORS then determines ε_{ox} for each oxide element in the circuit from the SPICE node voltages and a consideration of the work function difference between the electrodes. Figure 7-53 shows a block diagram of this oxide reliability simulator. Circuit failure statistics are calculated using the rule that if any MOSFET experiences breakdown, then the entire circuit fails. CORS is also a tool for optimizing burn-in, in that it can simulate the effect of burn-in in circuit reliability. The results of such simulations are useful for balancing conflicting goals such as cost, avoidance of hot electron degradation, and reduction of field failures due to oxide breakdown in choosing the burn-in condition for a product. HSPICE-compatible versions of the BERT simulator with software support are also commercially available (e.g., from BTA, Santa Clara, CA[127]).

7.8 MODELS OF THIN OXIDE GROWTH

The Deal-Grove model described in Vol. 1 for the growth of SiO_2 provides excellent agreement with experimental observations for thick oxides.[58] That is, the rate constants

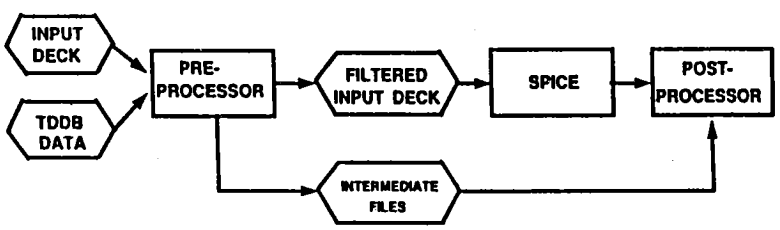


Fig. 7-53 Block diagram of oxide reliability simulator.[56] (© IEEE 1989).

derived in the Deal-Grove model allow the thickness of oxide films >350Å to be well predicted as a function of temperature, furnace ambient, silicon doping concentration, and silicon crystal orientation. However, the model does not give a detailed understanding of the mechanisms that produce these dependencies nor is it relevant when the oxide is thinner than 350Å. This is particularly troublesome because gate oxide thicknesses of ≤100Å are needed for submicron MOSFETs.

Several physical mechanisms have been proposed as models for such enhanced oxidation including:[59]

1. Models based on *space-charge effects* where the oxidation enhancement in the thin regime is due to field-assisted oxidant diffusion.

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

3. Models based on *stress effects* that can lead to a larger value of the diffusivity of the oxidizing species.

4. Models which postulate that the increased solubility of the oxidant in the thin oxide could affect the growth rate by its influence on the oxidant concentration at the silicon surface.

5. A model by Massoud et al., based on the presence of a *thin surface layer in the silicon where additional reaction sites are available.* The concentration of these sites decays exponentially with a characteristic length L_2 in Eq. 7-50.

Despite the large number of models proposed to explain oxide growth in this very thin regime, none has yet been convincingly accepted as more valid than the others. Nevertheless, several empirical models have been developed that yield reasonable agreement with measured values of oxide thickness in the thin regime.

Massoud et al. demonstrated that a much better fit can be accomplished that with the Deal-Grove model alone by adding an additional term to the oxidation rate that decays exponentially with thickness. The expression for the growth rate in the classical Deal-Grove model (see Vol. 1, chap. 7)

$$\frac{d\,t_{ox}}{d\,t} = \frac{B}{(2t_{ox} + A)} \qquad (7 - 49)$$

then becomes

$$\frac{d\,t_{ox}}{d\,t} = \frac{B}{(2t_{ox} + A)} + C_2\, e^{-t_{ox}/L_2} \qquad (7 - 50)$$

where C_2 is a pre-exponential constant with a value of 3.2 eV, and L_2 is the characteristic decay length described in paragraph **5** above. For dry-oxidation of lightly doped substrates in the 800-1000°C range, L_2 was found to be ≅ 70Å, independent of surface orientation. Heavy doping was found to only slightly affect the oxidation rate enhancement in the thin regime. Equation 7-50 has been implemented in SUPREM III as the model for thin oxide growth in dry-oxygen.

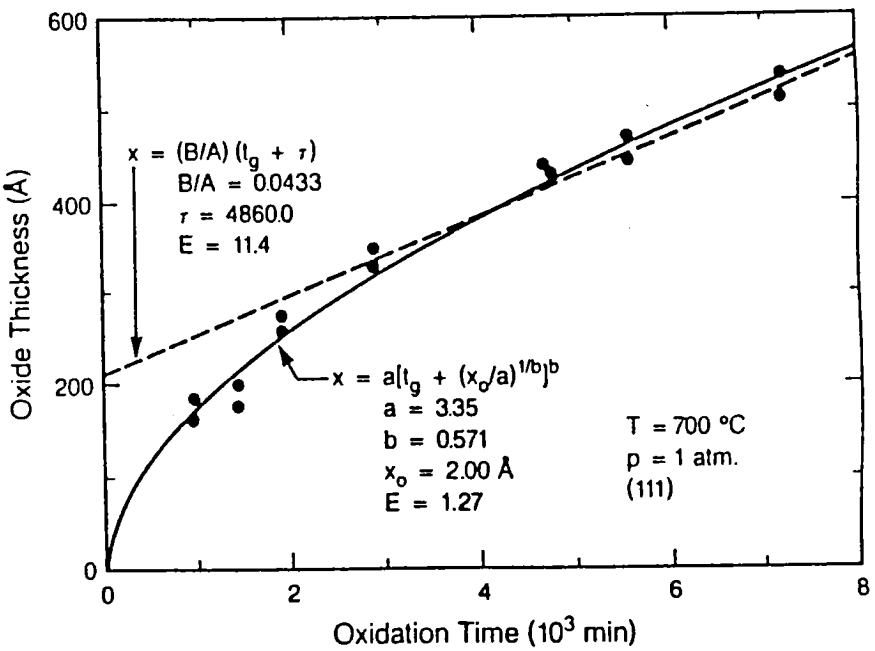


Fig. 7-54 Fits to oxide thickness versus oxidation time using the Deal-Grove model and the power-law model expression.[128]

Another empirical model for thin oxide growth has been developed by Reisman and Nicollian. By analyzing a vast amount of data, they were able to derive an empirical model that calculates the oxide thickness versus time using a general power law of the form

$$t_{ox} = a\,(t_g + \tau)^b \qquad (7 - 51)$$

where a and b are constants, and t_g is the growth time measured in a given experiment, and τ is the time to grow an oxide of thickness t_{oxi} already present on the surface.

A comparison between the power law description of oxide thickness and the Deal-Grove model is shown in Fig. 7-54. Data from Deal and Grove at 700°C are plotted on linear scales and Eq. 7-49 is plotted with measured values of A and B. Shown for comparison is Eq. 7-51 with the appropriate constants obtained by a fractional weighted least squares fit to the data.[60] Kouvatsos has extracted the a and b parameters of Eq. 7-51 at 950°C and 1100°C for oxide growth in O_2 diluted in He, and found very good fit to the Reisman-Nicollian power-law model.[61] This model has been implemented in the process simulator PREDICT-1 (see chap. 1).

7.9 TECHNOLOGY OF THIN OXIDE GROWTH

In this section we will describe the process technology of growing 10-20 nm thick gate oxides. As mentioned earlier, the goal is to grow uniform, reliable oxides of precisely the specified thickness - across each wafer, from wafer-to-wafer, and from run-to-run.

7.9.1 Oxidation Furnaces

Gate oxide growth is still carried out in production as a batch process. The typical load size is 150 wafers for 150 mm wafers and 100 wafers for 200 mm wafers (at 0.125" wafer-to-wafer spacing). Vertical furnaces are used almost exclusively for growing such gate oxides on 150-200 mm wafers (see Fig. 7-55).[62] In such furnaces neither the wafers nor the wafer holders make contact with the tube during loading or during the growth process. Thus, far fewer particles are generated than when loading was carried out by sliding a quartz boat into the horizontal furnace tube. (This approach was improved by cantilevering the load during insertion into the tube, then lowering the boat onto the tube, and withdrawing the cantilever, see Vol. 1, chap. 7). However, to achieve this same level of particle reduction in a horizontal furnace, a cantilever must be used to suspend the wafer load during the growth process. This necessitates use of a cantilever sufficiently rigid to hold the wafer load at high temperatures without sagging. As the wafers get larger, the load becomes heavier, making this approach too cumbersome. (Variable deflections of the cantilever rod may also cause problems in centering the load in the process tube, and the massive construction of cantilevers increases the amount of thermal energy transported from the wafers, leading to temperature nonuniformities in the wafers during oxide growth).

In the vertical furnace, lighter boat materials can be used, permitting the return to all-quartz boat construction. The lighter, more open boat design also enables virtually 360 degrees of line-of-sight radiant heat transfer between the wafers and furnace walls, thus improving the thermal uniformity characteristics. Additionally, gas-phase mass transport of the process gases is distributed uniformly along virtually the entire wafer circumference. Other advantages are: (1) precise control of wafer-to-wafer spacing in vertical furnace boats (which promotes better thickness uniformity); (2) ease of implementing automated wafer loading and unloading onto the boat (as well as boat loading into the furnace (in fact, loading wafers from plastic cassettes to the quartz boat is performed by Class-1 robots); and (3) contamination of the furnace by ambient atmosphere during load cycles is reduced. Thus, vertical furnaces have virtually replaced horizontal furnaces for 150 mm or larger wafers.

7.9.2 Control of Gate Oxide Growth Rates

Relatively slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. A number of approaches to achieving such slower growth rates have been reported, including: growth in dry O_2 at atmospheric pressure at lower temperatures (800-900°C); growth at reduced total pressures (i.e., lower than atmospheric pressure);[63] growth a reduced partial pressures of O_2 (i.e., use of a diluent inert gas -

such as N_2,[64] Ar,[65] or He,[66] together with the gas containing the oxidizing species); and use of composite oxide films (i.e., using gate-oxide films consisting of a layer of thermally grown SiO_2 and an overlying layer of CVD SiO_2). By and large, however, the mainstream approach for 10-15 nm thick gate oxides is apparently to grow the oxide film at atmospheric pressure and lower temperatures (800-900°C). With this approach, processes using modern vertical oxide furnaces can grow 10 nm oxides to within 0.1 nm (100Å ± 1-2Å) across the wafer, from wafer-to-wafer, and from run to run.

Note that particle densities resulting from gate oxide growth processes are very low in modern vertical furnaces for a number of reasons. Besides avoiding contact between the boat and the furnace walls, particle generation by vibration of wafers standing vertically in the boat (as can occur in horizontal furnaces) is suppressed. In addition, particles present in the vertical furnace are likely to fall in the annular space between the wafers and furnace walls. Each wafer is protected from these particles by the wafer just above it. Particle densities of less than $0.05/cm^2$ (at sizes $<0.2\ \mu m$) are now being specified for oxide furnaces.

The oxidation process is also inherently a particle reducer (as opposed to a supplier). That is, many particles are carbonaceous, and are thus consumed by the oxidation process itself (being turned into gaseous CO_2 or CO). However, other types of particles, such as silicon particles (from scribe marks, laser marking tools, or scratches), or metallic precipitates, etc., are of course not removed, and can represent problems. Particulates are more likely to be introduced from other fabrication processes than from the oxide furnace. In fact, the cleanest wafers (e.g., ≤10 particles per 200 mm wafer), and the most reproducible gate oxide films, are routinely achieved on wafers as received from the silicon wafer suppliers.

7.9.3 Factors Impacting Gate-Oxide Thickness Uniformity

Some of the factors that can adversely impact the control of oxide thickness uniformity include: (1) the presence of native oxide films on the Si surface prior to gate oxide growth; (2) growth of the oxide during ramp up and ramp down; (3) pre-oxidation cleaning effects; and (4) the residual water content of the O_2 gas in dry-oxidation processes. (Although the doping concentration in the Si substrate can also impact the oxide growth rate, this factor is not significant in gate oxidation processes since the substrate doping is relatively light.) Let us discuss the important factors in more detail.

A thin "native oxide" grows on bare silicon at room temperature and usually achieves a maximum thickness of about 1 nm (10Å).[67] Although there is some controversy concerning the exact nature of this oxide, (i.e., organic contamination, adsorbed H_2O, or true oxide), its presence undoubtedly affects subsequent oxidation. Nevertheless, the native oxide growth is self-limiting, and seems to attain most of its final thickness (10Å) within 24 hours after a bare Si surface has exposed to atmosphere. Thus, by either rigidly controlling the time between cleaning (HF dip) and oxide growth to within an hour, or by allowing at least 24 hours as the minimum time between cleaning and oxide growth, the native oxide layer thickness from run-to-run can be kept sufficiently constant so that the final oxide layer thickness will not be impacted by this

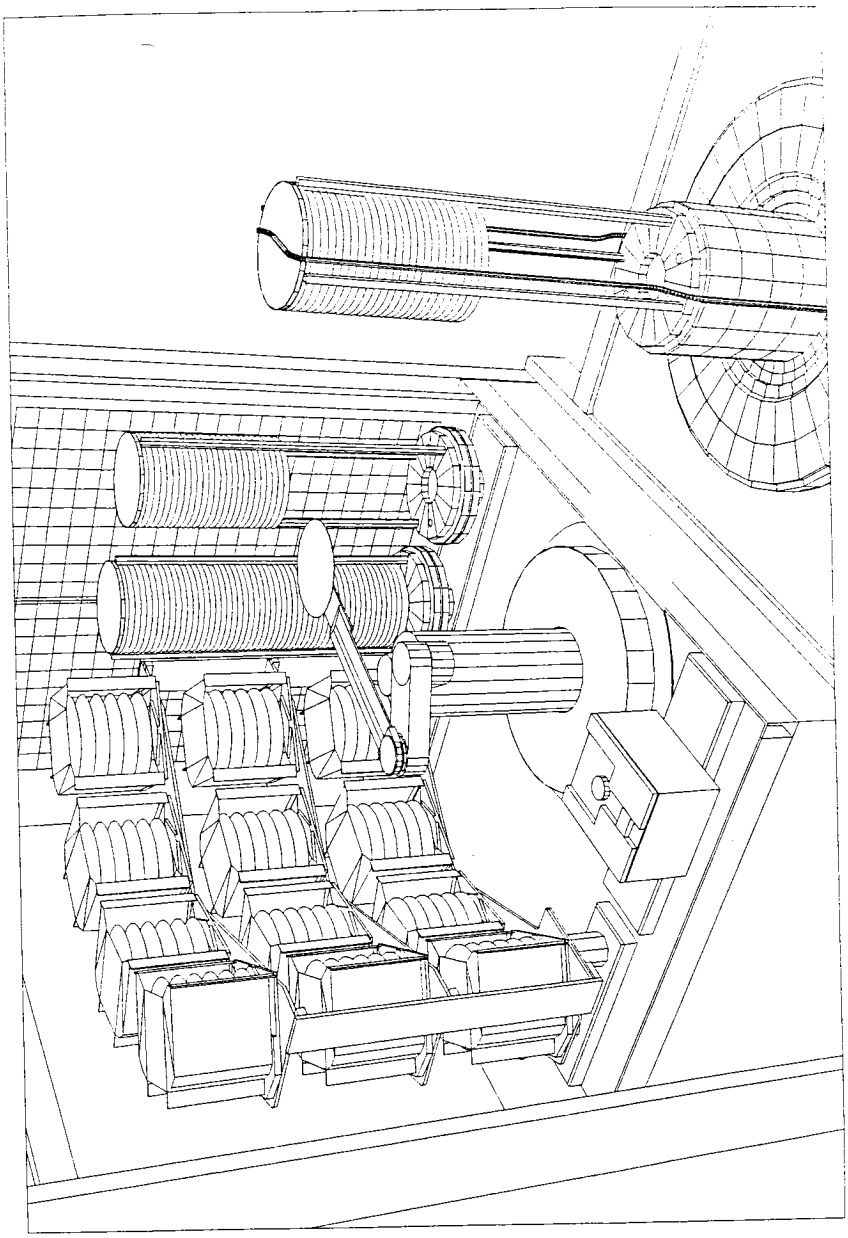


Fig. 7-55 Vertical furnace for growing thin gate oxides. (a) Robotic loading of vertical quartz boat. Courtesy of Semitherm.

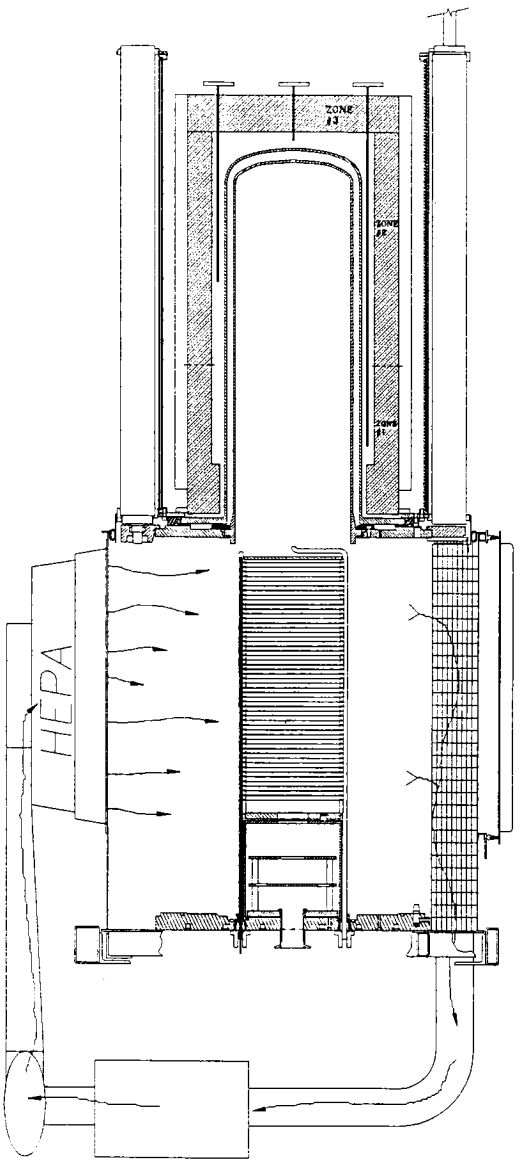


Fig. 7-55 Vertical furnace for growing thin gate oxides. (b) Furnace tube and boat. Courtesy of Semitherm.

factor. (The time wafers are permitted to remain in the DI rinse water after the HF dip may also need to be specified and controlled.)

Ohmi has also reported that native oxide can be prevented from forming by storing the wafers after cleaning in a gas ambient containing very little residual oxygen and moisture.[68] He also reported that native oxide growth can occur as a result of rinsing wafers with exposed Si in DI water containing high concentrations of dissolved oxygen (>600 ppm). To suppress such growth, he suggests rinsing such wafers in DI water containing very little (i.e., <20 ppm) dissolved oxygen, and then drying the wafers by blowing with N_2 gas.[69]

The furnace temperature during insertion is typically 700°C. The temperature is then ramped up to the growth temperature (~15°C/min). To prevent (uncontrolled) oxide growth during the temperature ramp up and ramp down of the furnace, the procedure is normally carried out in an ambient of N_2 and 2-4%O_2 (i.e., the oxygen is present to prevent either nitridation or micropitting of the Si surface exposed to pure N_2 at elevated temperatures). In another approach to control oxide growth and control surface microroughness during ramp up, Morita describes a technique of growing a "pre-oxide" of 4Å thickness in ultraclean O_2 at 300°C, and then ramping the wafers up to temperature in ultraclean Ar.[70] Yet another technique was reported by Wu et al., in which wafers were loaded at a lower temperature (600°C) into an N_2 ambient to reduce thermal stress and minimize native oxide growth. This is followed by a ramp-up to 900°C. The method includes a 60 minute pre-annealing step at 900°C prior to the oxide growth process, and a 15 minute N_2 anneal after the oxide is grown.[114]

Pre-oxidation cleaning procedures can have significant impact on the growth rate. With respect to cleaning, common procedures involve the use of acidic or basic hydrogen peroxide solution, sputter cleaning, *in situ* gas cleans using HCl, or flash heating to high temperatures. All of these cleans leave the surface in different states with varying surface films. For example, cleans composed of as base-acid-HF sequence produce a silicon surface that can oxidize up to 20% faster than a base-only clean. Apparently, the different cleans, on fundamentally different surface structures, can produce a change in the interfacial reaction rate which affects the linear oxidation rate constant.[71]

For dry O_2 oxidations there is also the requirement that trace amounts of water be controlled. A relatively small amount of water (125 ppm) in O_2 is sufficient to increase the oxidation rate at 980°C enough to throw the oxidation process out of spec for a sub-0.5 μm technology. *In situ* gas analysis with an RGA can monitor trace moisture in dry-oxidzing ambients to improve gate oxide thickness uniformity.[133]

7.9.4 Gate Oxide Growth Processes for 10-20 nm thick Gate Oxides on 150-200 mm Wafers

Gate oxide processes for submicron MOSFETs down to ~0.5 μm generally vary from company-to-company, and the exact details are usually proprietary. Despite this fact, some general statements can be made about thin gate oxide growth in this regime. As noted earlier, the processing is carried out in vertical furnaces, with wafer loads of 100-

150 wafers. Dummy wafers (e.g., 5 at the top and bottom of the wafer holder) are used to prevent thickness nonuniformity on wafers at the ends of the boat.

Prior to the gate oxide growth step, various procedures, as described in section 7.5.2 are carried out to optimize the gate oxide breakdown characteristics after it is grown. These include the growth and stripping of a sacrificial gate oxide (to counteract the Kooi effect), and careful cleaning procedures (see section 7.9.4). The wafers are then loaded onto quartz carriers by robotic wafer handling, and the carrier is moved into the hot zone of the furnace at insertion rates of ~15 cm/min for 150 mm wafers and ~10 cm/min for 200 mm wafers, with the furnace temperature during such insertion at ~650-700°C.

After being loaded, the furnace temperature is ramped to the oxide growth temperature. During this time, the wafers are held in an N_2 ambient (or maybe an Ar ambient) to prevent uncontrolled oxide growth during ramp up. As noted earlier, a small percentage of O_2 is also included in the inert ambient to slightly oxidize the wafer. This prevents thermal nitride from forming, which could inhibit subsequent oxidation. In addition, extended exposure of bare Si wafers to inert high temperature gases can result in the formation of micropits.[72] When the furnace reaches the growth temperature a short time for temperature stabilization of the wafers is allowed (e.g., 5 min) before the oxidizing gases are introduced. (In some furnaces, the furnace is pumped out prior to introducing pre-heated process gases.)

The growth process is usually carried out at atmospheric pressure, either in dry O_2, or in some cases in a 3-step process (i.e., the growth time is divided into parts: 1/3 dry O_2/1/3 wet O_2/1/3 dry O_2).* Typical process temperatures range from 800-900°C with the total growth time in the range of 15-30 minutes. Many processes use a chlorine-containing gas as well as O_2. HCl or trichloroethane (TCA - $C_2H_2Cl_3$) have been widely used for this purpose (2-4% concentration in the O_2 is typical).[73] Some of these may also be 3-step processes (e.g., 1/3 dry O_2/ 1/3 dry O_2 + TCA/ 1/3 dry O_2).[109]

The drawbacks of using HCl stem from its highly corrosive nature. That is, small leaks can cause damage to the gas cabinet electronic circuitry. Also, to prevent corrosion of the gas lines careful measures must be employed to exclude any moisture. TCA is much less corrosive. It is also the safest organic chloride available (with a relatively high toxic limit value [tlv] of 350 ppm), and generally only a catastrophic container failure may result in harmful release. However, TCA has been identified as an ozone depleter, and thus its use is also being phased out.** New gases, including 1,2,-trans dichloroethylene (t-DCE) have been introduced as substitutes for TCA, and are commercially available from Schumacher, Olin Hunt, and others.

* In a few processes, purely wet oxidation may be carried out, usually where it is desired to keep as low a thermal budget as possible (i.e., where very shallow junctions are needed and have been formed prior to the oxide growth step).

** Note that both the Montreal Protocol and the U.S. Clean Air Act will not regulate TCA that is used as a 'feedstock' material. So TCA is not restricted for use in applications like silicon oxidation, where it is a feedstock for chlorine atoms. Despite this fact, many U.S. semiconductor manufacturers are replacing TCA with t-DCE, which has been dubbed "the more responsible ecological choice." Its environmental half-life is four days, compared to six years for TCA.[74]

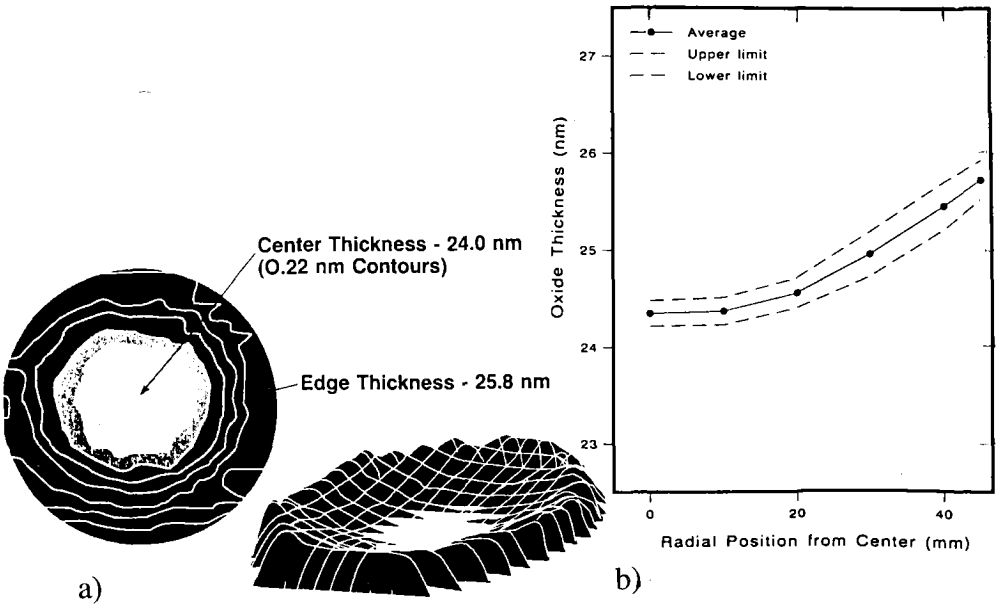


Fig. 7-56 Radial oxide thickness variations in oxides grown in O_2 + HCl: (a) thickness contour map for a single wafer, (b) oxide thickness as averages over an entire boat of wafers, as a function of radial position. These graphs illustrate the oxide thinning towards the center of the wafer.[73] Reprinted with the permission of the publisher, the Electrochemical Society.

However, it has been reported that use of Cl-containing gases in the oxide growth process can cause oxide thickness non-uniformity (see Fig 7-56).[73]

Some gate oxide growth processes, however, reportedly do not use chlorine-containing gases. The philosophy behind such approaches is that the Cl is primarily used to enhance local uniformity of the thin oxide film. This same goal is believed to be achievable by using meticulous and frequent cleaning of the furnace quartzware (which also prevents inadvertent Na contamination). In such cases, the use of chlorine is considered to be unnecessary.

Upon concluding the oxide growth process, the wafers are usually subjected to a post-oxidation anneal (POA), most commonly in N_2 at either the oxide growth temperature, or a somewhat higher temperature (50-100°C higher). A report comparing a 900°C/60 min POA in N_2 + 5%O_2 with a 1000°C/20 min POA in N_2 + 5%O_2 for 20 minutes claims that the higher temperature POA improves the oxide Q_{BD} and t_{BD} values.[75] (The small percentage of oxygen in the N_2 annealing ambient is thought to bridge non-bridging positively-charged oxygen atoms, and to increase the probability of oxidizing free Si trivalent atoms at the Si-SiO_2 interface.) Thus, surface states and fixed oxide charge are reduced. (Some reports also indicate that POA in Ar may result in incrementally lower oxide defect densities and reduced electron traps. Argon, however, is 3 to 5 times as expensive as N_2, so the increased performance may not justify the added production cost.[76]) Note that another approach, involving a two-step oxidation has also

been reported to improve the quality of the oxide. Here an initial 30-40Å oxide is grown, followed by an anneal at 1050°C for 60 min in Ar. Then the second oxidation step is performed to obtain the desired oxide thickness. The anneal in argon is postulated to cause viscous flow of the oxide allowing relaxation and release of built-up strains.[77]

Following the POA the furnace temperature is ramped back down, and the wafers are unloaded. The thickness of the grown oxide is measured with either a reflection spectrophotometer or an ellipsometer.[78] (UV-reflection spectrophotometers can measure thinner films than visible-light spectrophotometers, i.e., repeatable to <1 Å on 50Å oxides).

7.9.5 Alternative Gate Oxide Growth Processes

7.9.5.1 Stacked Gate Oxide Films (Thermally Grown SiO_2 Layer and CVD SiO_2 Layer). An alternative technique for forming the gate dielectric film by using a stack of thermal and CVD oxide layers has been investigated. This approach attempts to decrease the density of low-voltage defects (the A-mode defects shown in Fig. 7-14) in thin (<15 nm) gate oxide films. In one report, a three-step process sequence was used to fabricate a gate oxide film that had low defect density and interface state density.[79] A 5-nm-thick thermal oxide was first grown at 850-900°C in a mixture of O_2:TCA. Next, a 5-nm-thick LPCVD (0.2-0.3 torr) SiO_2 layer was deposited by TEOS pyrolysis at 625-650°C. Finally, a third layer of oxide was thermally grown underneath the first grown layer by performing an oxidation during the densifying anneal at 850-900°C. During this third step three events occur: the CVD oxide is densified, the traps get annealed out and a planar stress-free Si/SiO_2 interface is formed by the newly grown SiO_2 layer. The decrease in low-voltage defects is attributed to a mismatch of the micropores (10-25Å in diameter) that occur in both the thermal oxide and CVD layer. That is, there is high probability that these weak spots will not occur in both layers at the same location at the thermal/LPCVD oxide interface (Fig. 7-57a). In addition, since less silicon is consumed than in a conventional thermal oxide fewer weak spots originating from substrate defects (such as metal-contamination-related defects in the bulk) are incorporated into the bottom thermal oxide.

In a second report, a 11-nm-thick LPCVD TEOS layer was deposited at 650°C on a 4-nm-thick thermal oxide layer (grown at 900°C in O_2 and HCl) followed by a densification step to form a 14-nm-thick gate oxide. The densification was performed at 950°C in a 1% oxygen ambient diluted with argon, so that the shrinkage of the CVD layer during the anneal was offset by simultaneous thermal SiO_2 growth, resulting in a negligible effect on the final thickness of the stacked oxide. The temperature of the anneal is also important, as a lower one (i.e., 900°C) resulted in a less densified CVD layer, which allowed enhanced diffusion of oxidizing species during the later poly reox step, resulting in a larger gate bird's beak (see section 5.2.3.1). This can impact the electrical channel length in small MOSFETs. That is, whereas the 950°C anneal produced devices with electrical channel lengths comparable to those with thermal gate oxides, those with 900°C anneals had significantly longer electrical channel lengths.

Like the oxide in the report above, a significantly smaller number of low-voltage defects are also observed.[80,81] In later work, the results of hot-carrier and Fowler-Nordheim stressing on these films was reported. That is, the stacked films exhibited less degradation due to such stressing (i.e., less V_T-shift and degradation of the subthreshold slope S_t were observed in MOSFETs with these stacked gate oxides than with those made with thermal gate oxides).[113]

Moazzami and Hu investigated the impact of the dielectric deposition and growth sequence and the ratio of thermal to LPCVD oxide thickness on gate dielectric integrity.[82] They suggest that the decrease in low-voltage defects in such stacked gate dielectrics is not due to mismatch of the micropores, but to the filling of the micropores by the LPCVD oxide film (see Fig. 7-57b). If this assertion is correct (implying that only a very thin LPCVD oxide layer is necessary to fill micropores in the thermal oxide layer), it is projected that the stacked oxide technology can be scaled further by reducing the thickness of the thermal oxide layer. For example, with a 5-nm-thick thermal oxide layer, a total stacked oxide thickness as thin as 7 nm (suitable for 0.25-μm CMOS technology) appears feasible. They also note that for oxide technologies that are not

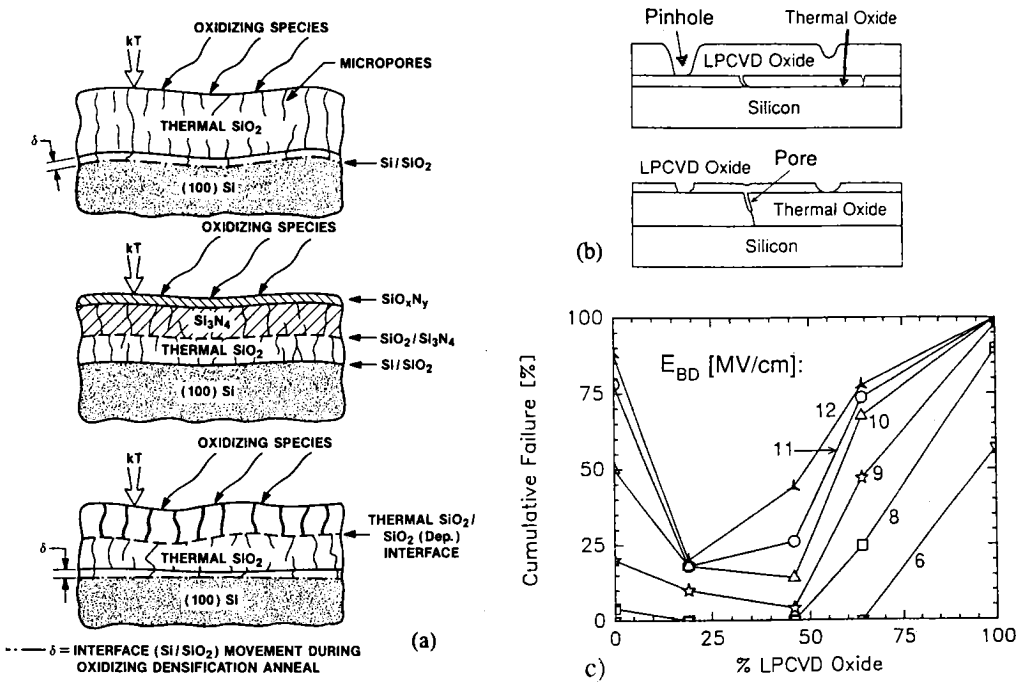


Fig. 7-57 (a) Effects of oxidizing densification anneal: (*i*) SiO_2 films, (*ii*) SiO_2 -Si_3N_4 dual dielectric: (*iii*) Thermal SiO_2 -CVD Dep SiO_2 multilayered stacked film.[79] (© IEEE 1988). (b) Thermal reoxidation does not eliminate a pinhole in the LPCVD oxide layer. (c) By introducing a thermal oxidation step prior to the LPCVD oxide deposition, the defect density is improved appreciably.[82] (© IEEE 1993).

believed to contain micropores (such as wet thermal oxide), stacked oxide technology may not lower A-mode defect density significantly. In addition, CVD SiO_2 processes are more subject to particulate formation than are thermal SiO_2 processes. Thus, gate oxide processes involving a CVD step will need to demonstrate that this problem can be overcome in order to be considered for mainstream production processes.

In another recent report, the advantages of the stacked gate dielectric was demonstrated for overcoming the following problems: gate thinning at the field edges; decreased oxide reliability following removal of the poly-buffer LOCOS nitride/poly mask by dry etching, and; damage of the gate oxide by MERIE plasma etch processes. The merits of using an N_2O oxynitride layer together with a CVD TEOS layer to form the stacked gate dielectric were also discussed.[112]

7.10 GATE OXIDE DAMAGE FROM PLASMA PROCESSING

Plasma processes use glow discharges to lower the process temperature and obtain directional ion bombardment (see vol. 1, chaps. 9 and 16). By their nature, glow discharges result in some wafer-surface charging, which is normally not significant. However, if this charging occurs on conductor films formed on the wafer surface, under some circumstances a thin (i.e., gate) oxide under these conductors can be severely degraded by such charging. That is, the oxides can be either completely shorted or weakened to the degree that they will fail early during normal device operation. Since such oxide damage may represent even further IC yield loss by enhancing the oxide's vulnerability to hot-carrier-induced degradation, the issue of plasma-induced degradation is a critical one for VLSI processing.

Evidence indicates that non-uniformities in the plasma across the wafer surface can give rise to significant local charge build-up. Here we describe models of how such charge build-up occurs and how it may damage the gate oxide, as well as suggested methods to avoid or mitigate the problem. A second plasma damage mechanism, radiation induced damage, is also briefly discussed at the conclusion of section 7.10.

7.10.1 How Isolated Conductor Regions on a Wafer Surface Can be Charged Up by an RF Plasma

When a plasma is generated with an rf power supply, the reactor electrodes are subjected to an alternating polarity voltage (i.e., at 13.56 MHz frequency in many RIE etchers). During the first half of the rf cycle (i.e., when a positive bias is applied) the electrode surface attracts electrons from the plasma. When the bias is negative (in the second half of the rf cycle of the voltage waveform) the electrode surface attracts positive ions from the plasma. Thus, the electrode is impinged upon by both positive and negative charges. These charged particles will therefore also bombard the surface of wafers mounted on such electrodes (see also Vol. 1, chap. 9).

In a uniform plasma a condition of balance is reached in which the flux of electrons during the positive part of the rf signal is equal to the flux of ions during the negative part of the signal. Charging is not a problem because precise recombination of opposing charges occurs and the surface potential stays close to that of the substrate.

However, if the plasma is non-uniform, ion and electron currents do not have to balance locally over the rf cycle (although there must be a net balance over the electrode as a whole).

If the effective electrode surface is a conductor, surface currents flow to balance the local nonuniformity in conduction currents from the plasma. If the surface is an insulator, this imbalance can cause the surface to charge locally. Specifically, if an island of conductor material is sitting on an oxide (and this island is elsewhere not connected to the substrate (e.g., a poly line that serves as a gate electrode), the locally unbalanced ion and electron currents can charge the conductor. The surface area of such a conductor exposed to impingement by the charged plasma species is referred to as the *antenna* surface (see Fig. 7-59). Since the antenna can sit on both thick oxide (field regions) and thin oxide (gate regions) the antenna/thin oxide area-ratio can exceed unity. The charging will result in an increased voltage across the gate oxide and a decreased voltage across the plasma sheath. This charge buildup will continue until the currents balance or the oxide begins to conduct. The positive feedback mechanism that drives the charge buildup is caused by the exponential dependence of electron current on sheath voltage. If the oxide begins to conduct (i.e., by Fowler-Nordheim tunneling), charge passage through the oxide may cause damage that can eventually lead to oxide breakdown. In particular, if the time integrated oxide current approaches Q_{BD} for the oxide, significant damage is incurred.

Let us consider the degree of charging from a quantitative perspective.[83] If the plasma is uniform, and positive and negative charges balance over a single rf cycle, the amount of charging of a gate *during* the cycle is very small. For instance, if we have a gate oxide thickness t_{ox} of 10 nm, an impinging ion current density J_i of 10 pA/μm^2 during the second half of the cycle, a 13.56 MHz excitation frequency, and an antenna/thin oxide area-ratio (antenna ratio) of 1000, the maximum gate voltage excursion in 1 rf cycle ΔV_g will be less than

$$\Delta V_g = \frac{J_i \times 1000}{2\,f \times C_{ox}} = \frac{10\ \text{pA}/\mu\text{m}^2 \times 1000}{2 \times 13.56\ \text{MHz} \times 345\ \text{nF/cm}^2} = 0.11\text{V} \qquad (7\text{-}52)$$

where C_{ox} is the oxide capacitance per unit area.

However, when the ion and electron currents are not in local balance, net charging over each rf cycle occurs. It continues until a steady-state oxide voltage is reached. At this point the Fowler-Nordheim tunneling current through the oxide balances the net current collected by the antenna. To get some idea of how long it takes for such a steady-condition is reached, let us assume that the ion current density J_i is larger than the electron current density J_e (e.g., J_i = 10 pA/μm^2 and the time-average electron current density $\langle J_e \rangle$ = - 8 pA/μm^2), the time needed to reach a steady-state gate voltage of 10V ($t_{charging}$) is approximately

$$t_{charging} = \frac{C_{ox} \times V_g}{(J_i - \langle J_e \rangle) \times 1000} = \frac{345\ \text{nF}/\text{cm}^2 \times 10\ \text{V}}{2\ \text{pA}/\mu\text{m}^2 \times 1000} = 17.3\ \mu\text{sec} \qquad (7\text{-}53)$$

Thus, after this short charging time a steady-state bias of 10V exists across the oxide in this example. At this steady-state condition, the voltage on the gate remains nearly constant, with a small-amplitude, superimposed 13.56 MHz ripple, similar to the effect calculated in Eq. 7-52. Plasma charging can therefore be approximated as a dc stress to the oxide. Note that both positive and negative charging can occur with voltages over 50V possible for thick dielectrics in very nonuniform plasmas. The actual voltage seen by a gate is reduced because the tunneling currents flowing the rough the gate oxide limit the peak charging voltage.

7.10.2 Causes of Non-Uniformities in Plasmas

The non-uniformities in a plasma which can give rise to local imbalances in collected charge can result from a variety of factors, including the following: (1) poor electrode design;[84] (2) non-uniform and/or time-varying magnetic fields;[83] (3) gas composition; e.g., use of highly electronegative gases (such as SF_6);[85] and (4) gas flow conditions inside the plasma reactor that cause unstable plasmas. Besides plasma nonuniformity there is concern that the drive toward higher plasma densities will increase charge damage since the charged particle currents are higher.

An example of how a non-uniform magnetic field causes plasma non-uniformities is shown in Fig. 7-58. Here a magnetic field in a MERIE (magnetically enhanced RIE)

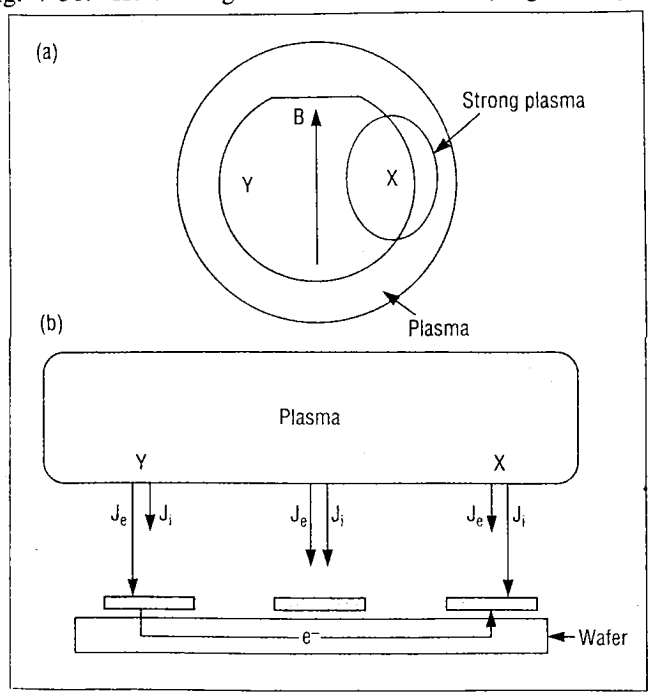


Fig. 7-58 (a) A strong plasma is present in region X during static MERIE etching. (b) A possible plasma-charging current path.[83] Reprinted with permission of Solid State Technology.

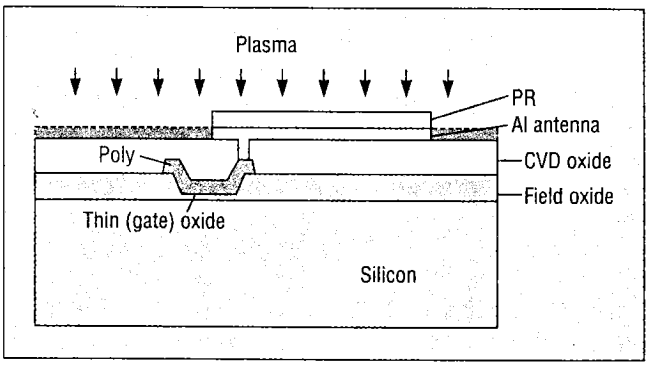


Fig. 7-59 Device cross-section of "antenna structure."[83] Reprinted with permission of Solid State Technology.

reactor with a static magnetic field is depicted. (Note that this is not necessarily the exact magnetic-field configuration of any particular MERIE reactor.) The effect of this magnetic field non-uniformity is to cause the electron flux to be larger than the ion flux near the wafer periphery, and the (positively charged) ion flux to be greater near the wafer center. Such magnetic field non-uniformities can also arise in ECR as well as MERIE plasma etchers.

7.10.3 Test Structures Used to Characterize Plasma-Induced Oxide Damage

Special test structures can be used to maximize the charging sensitivity of oxides to plasma damage. Measurements on these test structures can then be compared to results on product wafers. The two types of test structures most widely used to enhance the effects of plasma damage on thin oxides are: (1) large area capacitors; and (2) so-called "antenna" structures (structures with a large conductor area on thick oxide connected to a relatively small-area gate electrode on thin oxide - see Fig. 7-59. Note that both types of test structures should also be layed out in area-intensive and edge-intensive configurations (Fig. 7-60). This is because damage may be proportional to gate periphery or area (as will be discussed in section 7.10.5.

Oxide degeneration tends to increase with capacitor size because defects in the grown oxide tend to exhibit the same density distribution as particulates that cause chip defects (see section 7.5.4.3). That is, there is a greater likelihood of a defect being present in a capacitor as it's area is increased.

Antenna structures increase the density of Fowler-Nordheim tunneling current in thin oxides during plasma exposure by increasing the collected dc or rf currents which drive the Fowler-Nordheim tunneling current. That is, collected charge is proportional to the antenna electrode area or edge length (depending on whether the entire surface or just

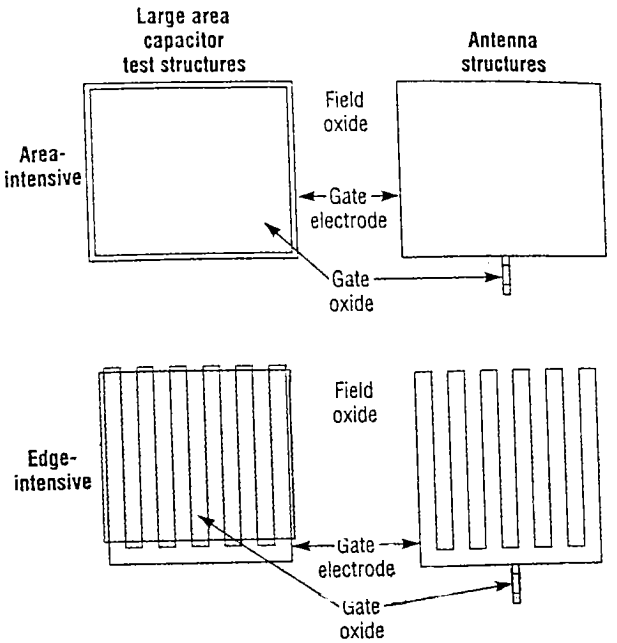


Fig. 7-60 Top view comparison of large area capacitor and antenna damage test structures with area-intensive and edge-intensive configurations. Field oxide is grown wherever gate oxide is not.[86] Reprinted with permission of Solid State Technology.

the edges are exposed to the plasma).[86] After the small area of thin oxide of the antenna structure is stressed by the collected charge, it is subjected to one of the electrical tests which measures oxide integrity (e.g., ramped-voltage V_{BD} or charge-to-breakdown [Q_{BD}] testing). Note that the antenna structure is less sensitive to damage but is probably more representative of real device structures than the large area capacitor. The antenna:gate-area ratio is usually varied (e.g., from 16:1 to 1000:1 for gate oxide areas of 20 μm x 20 μm (400 μm^2).

Large capacitors with area- and edge-intensive configurations are recommended for studying oxide damage during gate etching and ashing. Such structures are useful as monitors of both gate-oxide quality and overall damage because they are very sensitive to oxide defects. Antenna structures are effective for measuring damage from charge buildup after gate formation (i.e., interconnect etching) as well as damage that occurs during gate formation and ashing.

The threshold voltage of a MOSFET structure subjected to charging current damage will change due to the generation of interface states (as described in the next section). Although the shift in V_T may be small, a technique for monitoring such small V_T shifts has been suggested. It involves the use of a differential pair amplifier containing two MOSFETs. One MOSFET of the pair has a gate connected to an antenna with small

area, and the other to an antenna of larger area. The input offset voltage of the pair will vary if a change in the V_T of the two MOSFETs is not equal. The MOSFET whose gate is connected to the larger antenna will suffer more V_T shift than the other in the differential pair. Thus, the input-offset voltage of the circuit will increase proportionally to the increase in antenna area after a plasma processing step that causes oxide damage.[111]

7.10.4 Electrical Tests for Measuring Oxide Damage

The damage to gate oxides by oxide charging during plasma etching is measured after the oxide has been exposed to the plasma-etching process suspected of causing the damage. The same test techniques as those used to characterize the reliability of gate oxides (as described in section 7.5) can be used to monitor the damage from the plasma charging. That is, the oxide can be tested for: (1) increased Fowler-Nordheim leakage, as described in section 7.5.4.1; (2) the time-to-failure, t_{BD}, or charge-to-breakdown, Q_{BD}, using either a constant-voltage or constant current oxide breakdown test, as described in section 7.4.2.2; (3) V_{BD} using a slow-ramp voltage breakdown test, as described in section 7.5.4.5; (4) degradation of the CV characteristics (which is correlated to the degree of interface state generation by the injected charge), as described in sect. 7.5.4.7.

From the latter test, the stress charge that is passed through the oxide as a result of the plasma charge buildup can also be extracted from the CV measurements. That is, a group of MOS-Cs are fabricated with gates defined by wet etching (i.e., they are not exposed to any plasma charging process). These control samples are then subjected to known values of injected charge using a constant-current test setup. The injected charge causes these control MOS-Cs to exhibit degraded CV curves. Such degraded CV curves become the reference curves for extracting the charge injected during plasma etching. As an example, the CV curve of a plasma etched large-area MOS-C (16,000 μm^2) is superimposed on a set of CV reference curves of the same size control MOS-Cs that were respectively subjected to various known values of injected stress charge (as shown in Fig. 7-40). The plasma-etched CV curve matches the 10 nA reference curve quite well, indicating that the plasma-etched sample collected an average of 10 nA of charging current during the time of etch.

A non-plasma bias-temperature stress has also been reported to be a predictor of how susceptible an oxide will be to plasma charging. In the study, 12 nm-thick oxides were subjected to a -7V stress at 200°C for 15 min. The threshold-voltage shift after such bias-temperature stressing was correlated to the percentage failure in a slow-ramped voltage breakdown test after plasma-charging stress. Good correlation was found, suggesting that this bias-temperature stressing may be used to predict oxide susceptibility to plasma charging.[87]

Annealing in forming gas (e.g., at 400°C for 30 min) has been suggested as a technique to reverse plasma etch damage. Such an anneal is effective in passivating the interface traps caused by plasma etching (i.e., by the formation of Si-H bonds). This is evidenced by the disappearance after annealing of the CV shifts caused by the plasma process. Thus, it may appear that the annealing step heals the plasma damage.

However, the bonds formed between hydrogen and silicon during annealing are very weak (0.3 eV), and thus are more likely to be easily broken when the oxide is electrically stressed. This indeed occurs if the gate oxide is subjected to Fowler-Nordheim stressing after the annealing step (i.e., about 60% of the traps reappear subsequent to such electrical stressing). The oxides also continue to exhibit premature failure in time-dependent breakdown tests. The conclusion is that the damage caused by oxide charging is latent and remains even after annealing.

7.10.4.1 Wafer Mapping to Determine Positional Dependence of Damage on the Wafer Surface. Wafer mapping is also commonly used to locate the positions on the wafer where damage is most severe. From this information clues are obtained as to the possible cause of the damage. Figure 7-61 shows a wafer map of failed test MOS-Cs after being exposed to an Ar magnetron plasma with a nonuniform magnetic field. The position of the failures was correlated to this non-uniformity and the magnet was redesigned, reducing the damage problem.[88]

As another example, Fig. 7-62 shows a wafer map of the oxide-charging current across a wafer extracted from CV measurements as described in the previous section. This demonstrates the value of the CV measurement technique for monitoring plasma-oxide damage. That is, such spatial patterns would be very difficult to extract from oxide breakdown data because the latter become confounded with the random distribution of oxide defects.

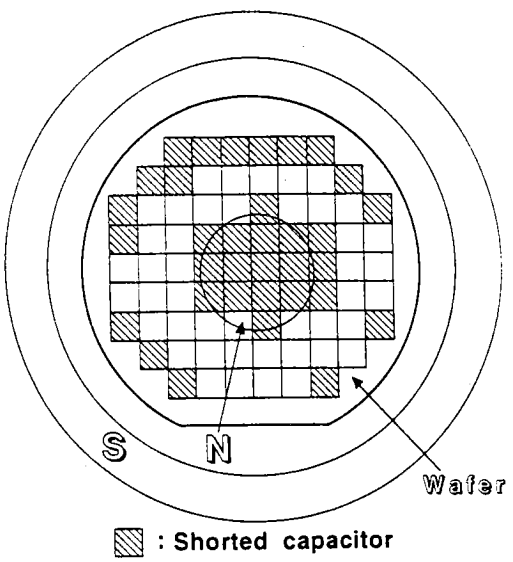


Fig. 7-61 (a) Gate oxide breakdown mapping in wafer exposed to an Ar magnetron plasma with a nonuniform magnetic field as shown in (b). (b) Magnetic field distributions as a function of distance from wafer center and magnet.[88] This paper was originally presented at the Spring 1992 Meeting of the Electrochemical Society, Inc. held in San Francisco, CA.

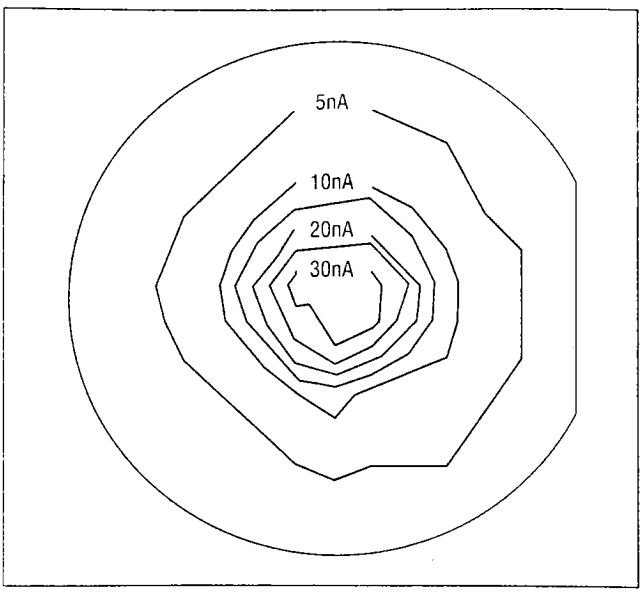


Fig. 7-62 Wafer map of the oxide-charging current across a wafer in a RIE etcher as extracted from CV measurements.[83] Reprinted with permission of Solid State Technology.

7.10.5 Oxide Damage Models in Various Plasma Etching Processes

7.10.5.1 Oxide Damage During Polysilicon or Metal Etch Processes. The first report of oxide damage during plasma etching was made by Watanabe and Yoshida in 1984.[89] They observed that the gate oxide could be destroyed by dielectric breakdown during reactive ion etching (RIE) of the polysilicon gate layer under some etching conditions. The mechanism of breakdown is not obvious, since the conductor is protected by a resist mask after the gates become isolated. A model has been proposed by Fang and McVittie to explain the mechanism of this damage.[90]

That is, as the doped polysilicon gate film is being plasma etched, non-uniformities in the plasma cause local unbalanced ion and electron currents. During most of the actual etching process the polysilicon film completely covers the wafer. Thus, the exposed poly-Si provides a low resistance path across the wafer so that surface currents can flow and balance the local nonuniformity in conduction currents from the plasma. No charge builds up during this phase, and no oxide damage occurs.

As the etching of polysilicon film is nearly completed, the long surface conduction paths across the wafer become too resistive or eventually discontinuous, and thus surface current paths are broken. Islands of gate electrode start to appear, and local charge buildup may begin. This results in a brief period (estimated to be a few seconds) during which the gate can be sufficiently charged to cause a significant Fowler-Nordheim tunneling current.

As noted earlier, the charging rate is determined by the net local current imbalance and the amount of exposed gate area in the islands. In the model by Fang and McVittie, the charge collector area during the transitory charging period is represented as a 'halo' of residual poly film around the resist edge (Fig. 7-63). That is, the charging current is collected by such halo regions around the mask. This model implies that the gate periphery may be important in determining the charging current. That is, the stress-producing current may depend on the perimeter of the polysilicon. This hypothesis must be tested for each plasma etch process using area-intensive and edge intensive test structures.[86]

During the overetch period, this 'halo' eventually disappears, and only the gate electrode edges are exposed to the plasma. Thus, current collection decreases significantly because it is limited by the small sidewall area. Further damage to the gate oxide from current tunneling is unlikely after the etch reaches endpoint. This model is supported by plasma measurements, SPICE simulations, and etching damage results.

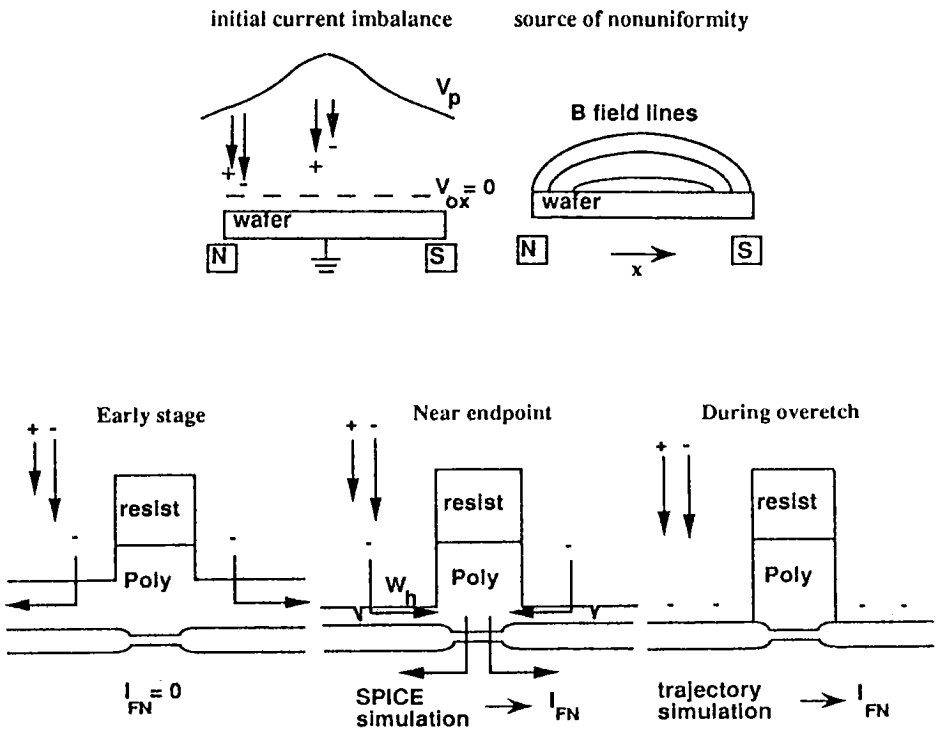


Fig. 7-63 The sequence leading to charge damage during poly-Si etching. Initially, surface currents prevent charging. Near endpoint, the long surface conduction paths across the wafer become too resistive, and excess current charges the gate and flows through the substrate. During overetch, current collection is limited by the small sidewall area.[86] (© IEEE 1992).

A report from the Microelectronics Center of North Carolina (MCNC) in 1992 studied cumulative breakdown failure of 11 nm-thick gate oxides in MOS-C in which the polysilicon gate was dry etched. They found that oxide breakdown failures after dryetch increase as the polysilicon gate perimeter-to-area ratio is made larger. Since the trend is to use ever larger polysilicon perimeter-to-area ratios, there is concern that the polysilicon gate etching process may become a serious yield limiter in submicron CMOS processes.[101]

The type of oxide damage that occurs during polysilicon etching described by the above model is also likely to occur during the etching of other conductors, including aluminum thin films, and during the oxide-spacer etch used to form the spacers in LDD MOSFETs (if the etching is continued past end point).[110]

7.10.5.2 Oxide Damage During Resist Stripping. In contrast to conductor pattern etching, when a plasma photoresist stripping process is carried out, the entire wafer surface is exposed to the reactive environment during the overetching step. In this case, any charging current that is present is thus proportional to both the pad area and the overetch time. Consequently in-plasma ashing is potentially a more damaging process. However, since there is no need for resist stripping to be anisotropic, the process can be carried out in a charge-free afterglow.

Solvent and low molecular weight residues from the resist can also contribute to plasma nonuniformity. In high rate strippers, the wafer is typically heated to ~200°C at the start of the cycle. At this temperature, light materials volatilize rapidly and the transient gas load can change the plasma's composition and electrical properties. Remote strippers generally avoid this difficulty because the flow and pressure of the feed gas are usually enough to prevent downstream products from reaching the plasma source.

It should be noted that other process steps besides plasma etching may also cause oxide damage by charging, including ion implantation[131] and pre-metal sputter-etch cleaning.[91] Generally speaking, however, charge build-up in these processes does not seem to represent as severe a problem as charging due to plasma etching.

7.10.5.3 Radiation Induced Oxide Damage. Another mechanism besides that due to plasma non-uniformities has been identified as a cause of gate oxide damage by the plasma, namely UV induced generation of traps and trapped charge at the exposed gate/thin oxide edge. That is, O_2 plasmas are known to emit light from the visible range down to 130 nm (i.e., down to the UV range). Although such UV radiation is unable to penetrate through the poly-Si gate layer to damage the oxide below, it can strike the exposed edge of the oxide. If the oxide is charged so that the field in the oxide is not sufficient to cause damage on its own (e.g., 1 MV/cm), the presence of this field may enhance UV degradation. Thus, edge-intensive structures can suffer this type of damage during the overetch of the poly-Si gate.

7.10.6 Procedures for Reducing Plasma-Induced Oxide Damage

Various measures can be taken to reduce oxide damage during plasma processing. These include: (1) reducing the excitation frequency;[92] (2) modifying the electrode design of the plasma etcher (e.g., by altering the magnet design in a MERIE etcher);[88] (3) eliminating highly electronegative gases such as SF_6 from the etch process;[85] (or, as another example, add O_2 to HBr during the polysilicon etch);[93] (4) modifying the pre-oxide cleaning process;[91] (5) selecting an oxide growth process that yields an oxide less susceptible to charging damage;[94,95]* and (6) utilizing device design techniques, such as limiting conductor area or edge length on field oxide attached to a gate electrode to reduce charge collection.

Since it seems that resist ashing in any process that subjects the wafer to charged particle bombardment can potentially become a source of oxide charge damage, resist should be stripped with processes that do not cause such damage (such as wet stripping, ozone ashing, or ashing in the charge-free afterglow of a plasma [downstream etching]).

It has also been reported that it may be possible to anneal out oxide damage incurred during plasma etching with an anneal at 650°C prior to metallization. This step is believed to detrap the holes thought to be responsible for the oxide damage.[96] However, another report indicates that while it appears that a post-plasma-etch heat treatment does seem to completely passivate the interface traps formed by plasma charging, these traps reappear after subsequent stressing. (The anneal is thought to form Si-H bonds that tie up the traps, but such bonds are weak and can be easily broken when electrically stressed.) Thus, this report concludes that the damage caused by oxide charging is latent and remains, even after annealing.

Another approach minimizing damage gate oxide during the LDD spacer etch involves depositing a protective dielectric layer over the gate electrode (e.g., 1500Å of thermal TEOS deposited) prior to gate electrode etching. After the gate is patterned, the spacer oxide layer is deposited by CVD, and then the spacers are created with an anisotropic dry etch step. The thicker oxide atop the gate remains intact during spacer oxide overetching. This prevents charge buildup across the gate oxide because the conductive surface of the gate is not exposed to charging currents, even during overetch time.[110]

As a final comment, it has been observed that those oxides which survive plasma processes that cause charge-up damage show longer lifetimes (t_{BD}) than those oxides not subjected to such processes. This is not surprising since the effect of plasma stressing is, in a sense, similar to oxide burn-in. However, the yield from such quasi-burn-in stressing may be intolerably low since the conditions of stress are inadvertent and certainly not optimized (as is possible in a genuine burn-in test). Hence, plasma oxide charge-up is still considered to be detrimental, and is not a substitute for burn-in.

* Interestingly, some evidence indicates that use of TCA during the dry-oxide growth process makes the oxide more susceptible to plasma charging damage than an oxide grown in dry-O_2 without a chlorine-containing gas.[94]

REFERENCES

1. C. Hu, "Future CMOS Scaling and Reliability," *Proceedings IEEE* May 1993, p. 682.
2. W.S. Chang et al., "A High-Performance 0.25 μm CMOS Technology: II - Technology," *IEEE Trans. Electron Dev.*, April 1992, p. 959.
3. S. Rigo, "Silica Films on Silicon," in *Instabilities in Silicon Devices - Vol. 1,* Ch. 1 Eds. G. Barbottin and A. Vapaille, (Elsevier) North-Holland, Amsterdam, 1986.
4. I.E. Tamm, *Phys. Z. Sowjetunion* **1**, 733, (1932).
5. P. Roblin, S. Samman and S. Bibyk, "Simulation of Hot-Electron Trapping and Aging in *n*MOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2229.
6. R.R. Razouk and B.E. Deal, *J. Electrochem. Soc.* **126,** p. 1573, (1979).
7. S. Tam, P.K Ko, and C. Hu, "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFETs," *IEEE Trans. Electron Dev.*, **ED-31**, (1984), p. 1116.
8. E. Takeda et al, "Submicrometer MOSFET Structures for Minimizing Hot-Carrier Generation," *IEEE Trans Electron Dev.*, **ED-29**, (1982), p. 611.
9. J. Frey, "Where Do Hot Electrons Come From?" *IEEE Circuits and Devices Mag.*, Nov. 1991, p. 31.
10. Y. Nissan-Cohen, "A novel floating-gate method for measurement of ultra-low hole and electron gate currents in MOS transistors," *IEEE Trans. Electron Dev. Lett.*, **EDL-7**, 1986, p. 561.
11. N.L. Saks et al., "Observation of Hot-Hole Injection in NMOS Transistors Using a Modified Floating-Gate Technique," *IEEE Trans. Electron Dev.*, **ED-33**, October 1986, p. 1529.
12. J.M. Pimbley et al., *Advanced CMOS Process Technology*, VLSI Electronics Microstructure Science, Vol. 19, Academic Press, Inc., San Diego CA, 1989, p. 186.
13. Y. Nissan-Cohen, J. Shappir, and D. Frohman-Bentchkowsky, "Measurements of Fowler-Nordheim Tunneling Currents in MOS Structures under Charge Trapping Conditions," *Solid-State Electron.* **28** (7) p. 717, (1985).
14. T.H. Ning and H.N. Yu, "Optimally induced injection of hot electrons into SiO_2," *J. Appl. Phys.*, vol. 45, p. 5373, (1974).
15. C. Chang et al., "Carrier tunneling related phenomena in thin oxide MOSFETs," *Tech. Dig. IEDM*, 1983, p. 194.
16. C.M. Osburn and D.W. Ormond, *J. Electrochem. Soc.*, **121**, (1972) p. 526.
17. K. Yamabe and K. Taniguchi, "Time-dependent dielectric breakdown of thin thermally grown SiO_2 films," *IEEE Trans. Electron Dev.*, **ED-32**, p. 423, (1985).
18. D. Wolters and J.F. Verwey, "Breakdown and Wearout Phenomena in SiO_2 Films," chap. 6, in *Instabilities in Silicon Devices - Vol. 1,* Ch. 1 Eds. G. Barbottin and A. Vapaille, (Elsevier) North-Holland, Amsterdam, 1986.
19. W.W. Hines and D.C. Montgomery, *Probability and Statistics in Engineering and Management Science,* Wiley, New York, 1972.
20. J.C. Lee, I.-C. Chen, and C. Hu, "Modeling and Characterization of Gate Oxide Reliability," *IEEE Trans. Electron Dev.*, **ED-35** December 1988, p. 2268.
21. J.P. Lee and A.J. Steckl, *J. Electrochem Soc.*, June 1993, p. L89.

22. P. Solomon, "Breakdown in Silicon Dioxide - A Review," *J. Vac. Sci. Technol.,* **14**, 1122, (1977).

23. E. Harari, "Dielectric breakdown in electrically stressed thin films of SiO_2," *J. Appl. Phys.*, **vol. 49**, no. 4, p. 2478, 1978.

24. B. Ricco et al., "Novel mechanism for tunneling and breakdown of thin SiO_2 films," *Phys. Rev. Lett.,* **51**, 1795 (1983).

25. Y. Nissan-Cohen and T. Gorczyca, "The Effect of Hydrogen on Trap Generation, Positive Charge Trapping, and Time-Dependent Dielectric Breakdown of Gate Oxides," *IEEE Electron Dev. Letts.,* vol. 9, June 1988, p. 287.

26. T.H. DiStefano and M. Shatzkes, "Impact ionization model for dielectric instability and breakdown," *Appl. Phys. Lett.* vol. 25, no. 12, Dec. 1974, p. 685.

27. N. Klein and P. Solomon, "Current runaway in insulators affected by impact ionization and recombination," *J. Appl. Phys.,* vol. 47, no. 10, 1976, p. 4364.

28. I.C. Chen, S. Holland, and C. Hu, "Electrical breakdown of thin gate and tunneling oxides," *IEEE Trans. Electron Dev.*, p. 413, Feb., 1985.

29. C.T. Sah, in *Properties of Silicon*, INSPEC, London, 1988, p. 587.

30. T.H. Ning, C.M. Osburn, and H.-N. Yu, "Threshold instability of IGFETs due to emission of leakage electrons from silicon substrate into silicon dioxide," *Appl. Phys. Lett.,* **29**, 198 (1976).

31. I.C. Chen, S. Holland, and C. Hu, "Oxide breakdown dependence on thickness and hole current-enhanced reliability of ultra-thin oxides." *Tech. Dig. IEDM,* 1986, p. 660.

32. I.C. Chin, et al., "The Effect of Channel Hot-Carrier Stressing on Gate-Oxide Integrity in MOSFETs," *IEEE Trans. Electron Dev.,* **ED-35**, Dec. 1988, p. 2253.

33. D.L. Griscom, "Diffusion of radiolytic molecular hydrogen as a mechanism for the post irradiation buildup of interface states in SiO_2 on Si structure," *J. Appl. Phys.,* vol. 58, p. 2524, 1985.

34. D.R. Wolters and A.T.A. Zeegers-van Duijnhoven, "Breakdown of thin dielectrics," *Ext. Abs. Mtg. of Electrochem. Soc.,* Spring 1990, p. 272.

35. Y. Uraoka et al., "Evaluation technique of gate oxide reliability with electrical and optical measurements," *Proc IEEE 1989 Conf. Microelectronic Test Structures,* March 1989, p. 97.

36. M.W. Hillen et al., *Insulating Films on Semiconductors,* J.F. Verwey and D.R. Wolters, Eds., North-Holland, New York, (1983), p. 274.

37. K. Honda , A. Ohsawa, and N. Toyokura, "Breakdown in Silicon Oxides - Correlation with Cu Precipitates," *Appl. Phys. Letts.,* vol. 45, no. 3, p. 270, Aug. 1, 1984.

38. K. Honda, A. Ohsawa, and N. Toyokura, "Breakdown in Silicon Oxides - Correlation with Fe Precipitates," *Appl. Phys. Letts.,* vol. 46, no. 6, p. 582, Mar. 15, 1985.

39. S. Verhaverbeke et al., "The Effect of Metallic Impurities on the Dielectric Breakdown of Oxides and Some Ways of Avoiding Them," *Tech. Dig. IEDM,* 1991, p. 71.

40. M. Miyashita et al., "Dependence of Surface Microroughness of CZ, FZ, and Epi Wafers on Wet Chemical Processing," *J. Electrochem. Soc.,* August 1992, p. 2133.

41. M. Meuris et al., "Investigating techniques to improve gate-oxide integrity," *Microcontamination,*, May 1991, p. 31.

42. D. Dimetrius et al., "Effects of HF-Last Clean Process Sequence on Gate-Oxide Quality," *Abs. Mtg. Electrochem. Soc.*, Fall 1993, p. 473.

43. M. Liehr, G.B. Bronner, and J.E. Lewis, "Stacking-Fault-Induced Defect Creation in SiO_2 on Si (100)," *Appl. Phys. Lett.*, **vol. 52**, no. 22, p. 1892, May 30, 1988.

44. W.M. Greene, P.J. Marcoux, and C.K.Lau, "Time Dependent Dielectric Breakdown: Lateral Areal Defects and Isolation Edge Defects," *Abs. Mtg. Electrochem. Soc.*, Fall 1991, p. 437.

45. R. Burmester, M. Kerber, and C. Zeller, "Optimisation of polysilicon oxidation for 0.8 μm technology," in *Proc. ESSDERC 88, J. de Physique*, coll. C4, suppl. 9, vol. 49, Sept. 1988, p. C4-545.

46. K. Shiozaki et al., "Influence of Field Isolation Stress on Gate Oxide Reliability," *Abs. Mtg. Electrochem. Soc.*, Spring 1992, p. 444.

47. R. Moazzami and C. Hu, "Projecting Gate Oxide Reliability and Optimizing Reliability Screens," *IEEE Trans. Electron Dev.*, **ED-37**, July 1990, p. 1643.

48. J.M. Soden and C.F. Hawkins, "Reliability of CMOS ICs with Gate Oxide Shorts," *Semiconductor International*, May 1987, p. 240.

49. I.C. Chen and C. Hu, "Accelerated Testing of Time-Dependent Breakdown of SiO_2," *IEEE Electron Dev. Letts.*, **EDL-8**, April 1987, p. 140.

50. M. Shatzkes, M. Av-Ron, and R.A. Gdula, "Defect-related breakdown and conduction in SiO_2," *IBM J. Res. Dev.* **24,** 469 (1980).

51. I.C. Chen, S. Holland, and C. Hu, "A quantitative physical model for time-dependent breakdown in SiO_2," in *Proc. Intl. Rel. Physics Symp.*, p. 24, 1985.

52. D.J. Dumin et al., "Correlation of Wearout and Breakdown in Sub-10 nm Silicon Oxide," *Tech Dig. IEDM*, 1988, p. 718.

53. D.J. Dumin and J.R. Maddux, "Correlation of Stress-Induced Leakage Currents in Thin Oxides with Trap Generation Inside the Oxides," *IEEE Trans. Electron Dev.*, **vol. 40**, May 1993, p. 986.

54. D.L Crook, "Method of Determining Reliability Screens for Time Dependent Dielectric Breakdown," *Proc. 17th Int. Rel. Phys. Symp.* 1, San Francisco CA, April 1979.

55. *ibid.*

56. E. Rosenbaum et al., "Circuit Reliability Simulator - Oxide Breakdown Module," *Tech. Dig. IEDM*, 1989, p. 331.

57. C. Hu, "IC Reliability Simulation," *IEEE J. Solid State Circuits*, p. 241, March, 1992.

58. B.E. Deal and A.S. Grove, "General relationship for the thermal oxidation of silicon," *J. of Appl. Phys.* vol. 36, p. 3770, 1965.

59. H.Z. Massoud , J.D. Plummer & E.A. Irene, *J. Electrochem. Soc.*, **132,** 1985, p. 2693.

60. A. Reisman et al., *J. Electron. Mat.* **16**, p. 45, (1987).

61. D.N. Kouvatsos and M.K. Hatalis, "Thin silicon dioxide films grown at very low oxygen partial pressure in oxygen diluted by helium," *Abs. Mtg. Electrochem. Soc.*, Fall 1992, p. 399.

62. R. Pong "LPCVD - Are Vertical Reactors the Answer?", *Microelectronics Mfg. and Test.*, July, 1990, p. 1.

63. A.C. Adams, T.E. Smith, and C.C. Chang, *J. Electrochem. Soc.*, **127**, p. 1787 (1980).

64. Y.J. van der Meulin, *J. Electrochem. Soc.*, **119**, p. 530 (1972)

64. Y.J. van der Meulin, *J. Electrochem. Soc.,* **119**, p. 530 (1972)

65. H.Z. Massoud, J.D. Plummer and E.A. Irene, *J. Electrochem. Soc.,* **132**, p. 2685 (1985).

66. D.N. Kouvatsos and M.K. Hatalis, "Thin silicon dioxide films grown at very low oxygen partial pressure in oxygen diluted by helium," *Abs. Mtg. Electrochem. Soc.,* Fall 1992, p. 399.

67. A. Licciardello, O. Puglisi, and S. Pignataro, "Effect of Organic Contamination on the Oxidation Kinetics of Silicon at Room Temperature," *Appl. Phys. Lett.,* vol. 48 (1), p. 41, Jan. 6, 1986.

68. M. Morita et al., *J. Appl. Phys.,* July 15, 1990, p. 1272.

69. M. Miyawaki, S. Yoshitake, and T. Ohmi, "Improvement of Al-Si Contact Performance in Native-Oxide-Free Processing," *IEEE Electron Dev. Lett.,* , **11**, October 1990, p. 448.

70. M. Morita and T. Ohmi, "Pre-Gate Oxide Si Surface Control," *Abs. Mtg. Electrochem. Soc.,* Spring 1992, p. 380.

71. G. Gould and E.A. Irene, "The Influence of Silicon Surface Cleaning Procedures on Silicon Oxidation," *J. Electrochem. Soc.*, vol. 134 (4) , p. 1031, April 1987.

72. A. Reisman, S. Edwards, and P. Smith, *J. Electrochem. Soc.,* **vol. 132**, p. 284, (1990).

73. C.M. Osburn et al., "Silicon Gate Oxide Thickness Uniformity during HCl Oxidation," *J. Electrochem. Soc.,* **vol. 138**, January 1991, p. 268.

74. *Semiconductor International*, March 1993, p. 21.

75. M.K. Khan and L. Schmidt, "A Simple Way to Improve Gate Oxide Quality," *Semiconductor Internat.,* September 1989, p. 136.

76. *Semiconductor International,* October 1988, p. 42.

77. A. Battacharyya et al., *J. Electrochem. Soc.* **Vol. 132**, August 1985, p. 1900.

78. T.S. Choo, C.L. Lee, and T.F. Lei, "Measurement of Ultrathin (≤100Å) Oxide Films by Multiple-Angle Incident Ellipsometry," *J. Electrochem. Soc.,* June 1991, p. 1756.

79. P.K. Roy et al., "Synthesis and Characterization of High Quality Ultrathin Gate Oxides for VLSI/ULSI Circuits," *Tech. Dig. IEDM* 1988, p. 714.

80. H.H. Tseng, P.J. Tobin, J.D. Hayden, and K.-M. Chang "Advantages of CVD Stacked Gate Oxide for Robust 0.5 μm Transistors," *Tech. Dig. IEDM*, 1991, p. 75.

81. H.H. Tseng and P.J. Tobin, "A Robust Gate Dielectric for Submicron Technology," *Semicond. Internat.*, July 1992, p. 68.

82. R. Moazzami and C. Hu, "A High Quality Stacked Thermal/LPCVD Gate Oxide Technology for ULSI," *IEEE Electron Dev. Lett.,* **14**, February 1993 p. 72.

83. H. Shin et al., "Plasma Etching Charge-Up Damage to Thin Oxides," *Solid-State Technology*, August 1993, p. 29.

84. H. Shin et al., "Spatial Distributions of Thin Oxide Charging in Reactive Ion Etched and MERIE Etcher," *IEEE Elect. Dev. Lett.,* vol. 14 (2), p. 88, February 1993.

85. C.T. Gabriel, "Gate Oxide Damage from Polysilicon Etching," *J. Vac. Sci. Tech. B* vol. 9 (2), p. 370 (1991).

86. C.T. Gabriel and J.P. McVittie, "How Plasma Etching Damages Thin Gate Oxides," *Solid-State Technology,* June 1992, p. 81.

87. S. Fang and J.P. McVittie "Thin-Oxide Damage from Gate Charging During Plasma Processing," *IEEE Electron Dev. Letts.,* May 92 p. 288.

88. M. Sekine et al., "Gate Oxide Breakdown Phenomena in Magnetron Plasma," *Ext. Abs. Mtg. Electrochem. Soc.,* Spring 1992, p. 126.

89. T. Watanabe and Y. Yoshida, "Dielectric Breakdown of Gate Insulator Due to RIE," *Solid State Technology,* April 1984, p. 213.

90. S. Fang, S. Murakawa, and J.P. McVittie, "A New Model for Thin Oxide Degradation from Wafer Charging in Plasma Etching," *Tech. Dig. IEDM* 1992, p. 61.

91. I.-W. Wu et al., "Breakdown Yield and Lifetime of Thin Gate Oxides in CMOS Processing,"*J. Electrochem. Soc.,* June 1989, p. 1638.

92. K. Tsunokuni, K. Nojiri, M. Nawata, "Dependence of RF Frequency on Gate Oxide Breakdown During Microwave Plasma Etching," *Ext. Abstr. 38th Mtg. Jap. Soc. Appl. Phys.,* p. 499, (1991).

93. W.M. Greene, J.B. Kruger, G. Kooi, "Magnetron Etching of Polysilicon: Electrical Damage," *J. Vac. Sci. Tech. B* vol. 9 (2), p. 1638 (1989).

94. S. Fang and J.P. McVittie, "Thin Oxide Damage from Charging During Plasma Processing," *IEEE Elect. Dev. Lett.,* vol. 13 (5) p. 288, May 1992.

95. W. Ting, et al., "MOS Characteristics of Ultrathin SiO_2 Prepared by Oxidizing Si in N_2O," *IEEE Elect. Dev. Lett.,* vol. 12 (4) p. 416, April 1991.

96. S. Fang and J.P. McVittie, "Thin Oxide Damage from Charging During Plasma Processing," *IEEE Elect. Dev. Lett.,* vol. 13 (5) p. 288, May 1992.

97. J. Borland, "Borland's Overview of the Latest in Intrinsic Gettering," *Semiconductor Internat.* Part 1, April 1989 p. 144, & part 2, May 1989, p. 154.

98. T. Ohmi et al., *IEEE Symposium on VLSI Technology,* Section 3-4, p. 24, June 1992.

99. "Special Report on Flash Memory and Hi Wafers," *Nikkei Microdevices,* p. 34, June 1993.

100. S. Samata et al., *Japan Society of Applied Physics Spring Meeting,* sections 30a-ZS-1 -4, p. 676, April 1993.

101. P.H. Singer, "Evaluating Plasma Etch Damage," *Semiconductor International,* May 1992, p. 78.

102. H. Satake and A. Toriumi, "Substrate Hole Current Generation and Oxide Breakdown in Si MOSFETs under Fowler-Nordheim Electron Tunneling Injection," *Tech. Dig. IEDM,* 1993, p. 337.

103. D.J. Dumin et al., "Use of Wearout Measurements to Analyze TDDB Distributions in Thin Silicon Oxides," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 204, Abs. No. 133.

104. D.J. Dumin, "Wearout and Breakdown in Thin Silicon Oxide," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 200, Abs. No.131.

105. J. Sune et al., *Thin Solid Films,* **185**, 347, 1990.

106. R. Winkler and G. Behnke, "Gate Oxide Quality Related to Bulk Properties and Its Influence on DRAM Device Performance," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 778, Abs. No.488.

107. K. Shiozaki et al., "Improvement of Stress-Induced Surface Microroughness for Highly Reliable Gate Oxide," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 207 , Abs. No.135.

108. M. Gardener, J. Seaton and J. Fulford, "The Role Silicon Surface Has Upon Producing Highly Reliable Tunnel/Gate Dielectrics," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 211, Abs. No.137.

109. A. Philiposian and K. Van Wormer,"Temperature Measurement Studies in Atmospheric Thermal Silicon Oxidation Reactors," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 223, Abs. No.145.

110. C. Gabriel and M. Weiling, "Gate Oxide Damage Reduction during LDD Oxide Etch," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 322, Abs. No. 205.

111. H. Shin, Z.-J. Ma, and C. Hu, "Impact of Plasma Charging Damage and Diode Protection on Scaled Thin Oxide," *Tech. Dig. IEDM* 1993, p. 467.

112. H.-H. Tseng and P.J. Tobin, "Thin CVD Stacked Gate Dielectric for ULSI Technology," *Tech. Dig. IEDM,* 1993, p. 321.

113. H.-H. Tseng et al., "A Comparison of CVD Stacked Gate Oxide and Thermal Gate Oxide for 0.5 μm Transistors Subjected to Process-Induced Damage," *IEEE Trans. Electron Dev.,* **ED-40**, March 1993, p. 613.

114. S.L. Wu, C.L. Lee, and T.F. Lei, "Characterization of Ultra-Thin Oxide Prepared by Low-Temperature Wafer Loading and Nitrogen Pre-Annealing before Oxidation," *J. Appl. Phys.,* vol. 72, p. 1378, 1992.

115. T.J. Dalton et al., "Microtrench Formation in Polysilicon Plasma Etching over Thin Gate Oxide," *J. Electrochemical Soc.,* August 1993, p. 2395.

116. S. Krishnan and J. McKee, "A Monitor for Polysilicon Etch Damage at the Gate Periphery," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 328, Abs. No.208.

117. S. Ma and J.P. McVittie, "Effect of Exposed Gate Oxide Edge on the Separation of Plasma Charging and Radiation Induced Damage," *Ext. Abs. Electrochem. Soc. Mtg.,* Spring 1994, p. 326, Abs. No. 207.

118. B.E. Deal, *J. Electrochemical Society.,* **121:** 198C (June 1974).

119. J.J. Sanchez and T.A. DeMassa, "Review of Carrier Injection in the Silicon/Silicon-Dioxide System," *IEE Proceedings-G* **138**, No. 2, June 1991.

120. B. Eitan and D. Frohman-Bentchkowsky, "Hot Electron Injection into the Oxide in *n*-channel MOS Devices," *IEEE Trans. Electron Dev.*, **ED-28**, p. 328, (1981).

121. C. Hu, "Thin Oxide Reliability," *Tech. Dig. IEDM,* 1985 p. 368.

122. E. Takeda et al., "VLSI Reliability Challenges," From Device Physics to Wafer Scale Systems," *Proceedings of the IEEE,* **81**, May 1993, p. 653.

123. Y. Uraoka et al., "Evaluation Technology of VLSI Reliability Using Hot Carrier Luminescence," *IEEE Trans. on Semiconductor Mfg.,* **4**, August 1991, p. 183.

124. S. Samata et al., "Hydrogen Annealing of Silicon Wafer," *Ext. Abs. Electrochem. Soc. Mtg.*, Fall 1993, p. 426, Abs. No. 264.

125. M. Kerber, "Entirely Gate-Surrounded MOS Capacitor to Study the Intrinsic Oxide Quality," *IEEE Trans. Electron Dev.,* **ED-39**, December 1992, p. 2814.

126. R. Moazzami et al., "Projecting the Minimum Acceptable Oxide Thickness for Time-Dependent Dielectric Breakdown," *Tech. Dig. IEDM,* 1988 , p. 710.

127. BTA, P.O. Box 4868, Santa Clara CA 95054.

128. A. Reisman et al., *J. Electron. Mat.* **16**, p. 45 (1987).

129. H. Abe et al., "Analysis of Defects in Thin SiO_2 Thermally Grown on Si Substrate," *Tech. Dig. IEDM*, 1985, p. 372.

130. H. Uchida et al., "The Effect of Oxide Charges at LOCOS Isolation Edges on Oxide Breakdown," *IEEE Trans. Electron Dev.*, **ED-40**, October 1993, p. 1818.

131. H. Muto et al., "A Mechanism of Gate Oxide Deterioration During As^+ Ion Implantation," *IEEE Trans. Electron Dev.*, **ED-38**, June 1991, p. 1296.

132. A.G. Revesz, "The Defect Structure of Grown Silicon Dioxide Fims," *IEEE Trans. Electron Dev.*, **ED-12**, p. 97, (1965).

133. M. Moinpour, "In Situ Reactor Analysis for Silicon Nitride and Silicon Dioxide Thin Films," *Ext. Abs. Electrochem. Soc. Mtg.*, Spring 1994, p. 223, Abs. No. 146.

PROBLEMS

1. Describe the difference between hot and cold carrier injection in MOSFETs.

2. Explain why the peak value of substrate current I_{sub} and gate current I_G occur at different values of V_{GS} for a fixed value of V_{DS}.

3. Explain why it appears that the minimum gate oxide thickness in MOSFETs may be 30Å.

4. Using the data plotted in Fig. 7-18, determine the value of the dispersion (σ) of the log-normal cumulative distribution function plotted therein, as well as the mean time to failure at 200°C and 250°C.

5. Calculate the value of $D(\Delta t_{OX})$ if: (a) Δt_{OX} = 40Å; and (b) Δt_{OX} = 15Å for the oxide technology that yields the TDDB data of Fig. 7-34, assuming the defects are nonuniformly distributed and S = 0.8. The area of the oxide samples is 75 mm^2.

6. Assume that a chip having a total gate oxide area per chip of 75 mm^2 is to be manufactured and the defect density due to weaknesses in the gate oxide of the specific chosen thickness is found to be $D(t_{OX}) = 40/cm^2$. Calculate the manufacturing yield of this chip if: (a) the defects are assumed to be uniformly distributed; and (b) the defects are assumed to be nonuniformly distributed, and S = 0.5.

7. Use Eq. 7-29 to calculate the time to breakdown of 10 nm thick, defect-free oxide film if V_{OX} is 5V at 300K. If a defect in the film reduces the thickness to 6 nm, calculate the new time to breakdown at the same gate voltage and operating temperature.

8. Calculate the minimum thickness of a defect-free oxide that could be used in a MOSFET that is to operate at 3.5 V for 10 years at 150°C (425K) without suffering oxide breakdown.

8. What is the maximum electric field that can be applied across a gate oxide so that is will operate for 2 years without suffering breakdown at 125°C.

9. Using Eq. 7-51 calculate the time to grow a 10 nm oxide film at 700°C on bare (111) silicon, assuming that the constants a and b at 700°C equal $a = 3.35$ and $b = 0.571$.

10. List some of the advantages of vertical furnaces that have caused them to replace horizontal furnaces fro the growth of thermal oxides in submicron technologies.

CHAPTER 8

WELL FORMATION IN CMOS

Both *n*- and *p*-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite to that present in the starting material must be formed. These regions are called *wells* (or sometimes *tubs*) and are usually (but not always) the first features to be defined on a starting wafer. The well-formation process starts with a uniform, lightly doped *p*- or *n*-type substrate. Excess dopants of the opposite type are then selectively implanted or diffused to attain the proper well depth and doping profile in the well regions. The doping type in the wells becomes the identifying characteristic of the CMOS technology (e.g., *p*-well or *n*-well CMOS). Conventional well-implant doses range from $2x10^{12}$-$8x10^{12}/cm^2$ at energies from 100-200 keV. The term "conventional" in this sense refers to wells formed by low-energy (<200 keV) ion implantation and a thermal drive-in. Alternative well-formation techniques that do not require a drive-in will also be discussed separately. Epitaxial *n*-on-n^+ or *p*-on-p^+ starting substrates have also come into widespread use, since they help reduce the susceptibility of CMOS circuits to latchup.

As well as being of the correct doping type, the well and substrate must be electrically isolated from each other under every operating condition. As noted in chap. 6, to ensure that *pn* junctions are not forward-biased during circuit operation, if a *p*-type substrate is used as the starting material of an IC, the substrate must be connected to the most negative circuit voltage. If an *n*-type substrate is used, it must be connected to the most positive circuit voltage. When wells are present, however, the well regions must also be connected to the appropriate circuit voltages to prevent forward-biasing of the junctions within the well (and between the well and substrate). Because the wells are totally junction-isolated from the rest of the wafer, it is especially important that provisions for well connections be made at the top surface of the wafer. That is, in non-well technologies, it may still be possible to contact the substrate from the backside of the wafer even if no provision is made for a substrate contact from the top surface. Obviously, such backside connections cannot be established to the well regions.

The dopant concentration near the surface of the well is one of the factors that impact V_T, carrier mobility, source/drain capacitance, and field isolation. The well concentration deeper beneath the surface affects other characteristics of MOSFET behavior such as latch-up and vertical punchthrough (i.e., between drain and well-substrate junctions).*

* To ensure that vertical punchthrough will not occur, the depletion regions of the drain/well and well/substrate junctions must be prevented from merging. That is, the depth of the well/substrate junction must be deep enough (typically >1.5 μm) so that it is greater than the sum of the widths of the two depletion regions in the well, even when a reverse bias of V_{DD} is

8.1 *p*-WELL, *n*-WELL, and TWIN-WELL TECHNOLOGIES

There are many trade-offs involved in the optimization of a CMOS process. The choices revolve around the highly interrelated parameters of circuit performance, layout density, fabrication cost, and susceptibility to latchup. As described in previous chapters, obtaining the best circuit performance from an MOS device involves maximizing the drive current and minimizing junction capacitances and body effect - all characteristics which favor lower doping concentrations in the device body. Optimizing density, however, requires raising these same doping concentrations (to avoid punchthrough and to achieve high values of threshold voltage in the parasitic field-transistors). Higher density is thus achieved by allowing closer packing of adjacent *n*- and *p*-channel transistors. Conversely, latchup tolerance can also be improved by keeping *n*- and *p*-channel transistors farther apart (see chap. 6, Vol. 2), but this in turn lowers circuit density.

These complex interacting tradeoffs converge on several processing configurations that are determined at the outset of the processing sequence through the selection of the type of well doping.[1,2] It is therefore useful to consider the ramifications of the well-type choice with respect to its impact on circuit applications. Because at least one of the device types must be located in the well, the selection of well architecture influences device and circuit behavior. In particular, the performance of devices in the well for the most part suffer as a consequence of the higher well doping. That is, they exhibit higher junction capacitance, increased sensitivity to body effect, and decreased transconductance (due to reduced carrier mobility). Furthermore, substrate current (e.g., due to hot-carrier effects) is harder to collect from the well regions. The issue thus becomes one of deciding which device type (if any) should be subjected to such performance degradation.

Figure 8-1 shows the five basic well structures employed in CMOS technologies. Figure 8-1a illustrates the *single p-well structure*, in which *p*-wells are fabricated in an *n*-type starting substrate. NMOS devices are made in the well, and PMOS transistors in the substrate. If the dopant types of the *p*-well technology are reversed, a *single n-well structure* is obtained (Fig. 8-1b). The structures in Figs. 8-1c to 8-1e involve approaches in which *both p*- and *n*-wells are formed. CMOS technologies employing such structures are called *twin-well, dual-well,* or *twin-tub CMOS.* The first two of these (Figs. 8-1c and 8-1d) are designed to be fabricated on a particular substrate type, like the single *p*- and *n*-well structures. The third of these twin-well technologies (Fig. 8-1e), however, is independent of starting material type.

In the next sections we will outline the pros and cons of both the single *p*-well and *n*-well configurations, as well as those of the more complex well configurations that have been developed (e.g., twin-well and retrograde-well CMOS).

* (cont.) applied to the drain (see example 8-1 in this chapter.). In conventional-well CMOS structures, this depth is achieved by a long, high-temperature drive-in (up to 1200°C) following the well implants. For example, in one twin-well process the well-implants are followed by a drive-in at 1100°C for 500 minutes (8 1/3 hours).[13]

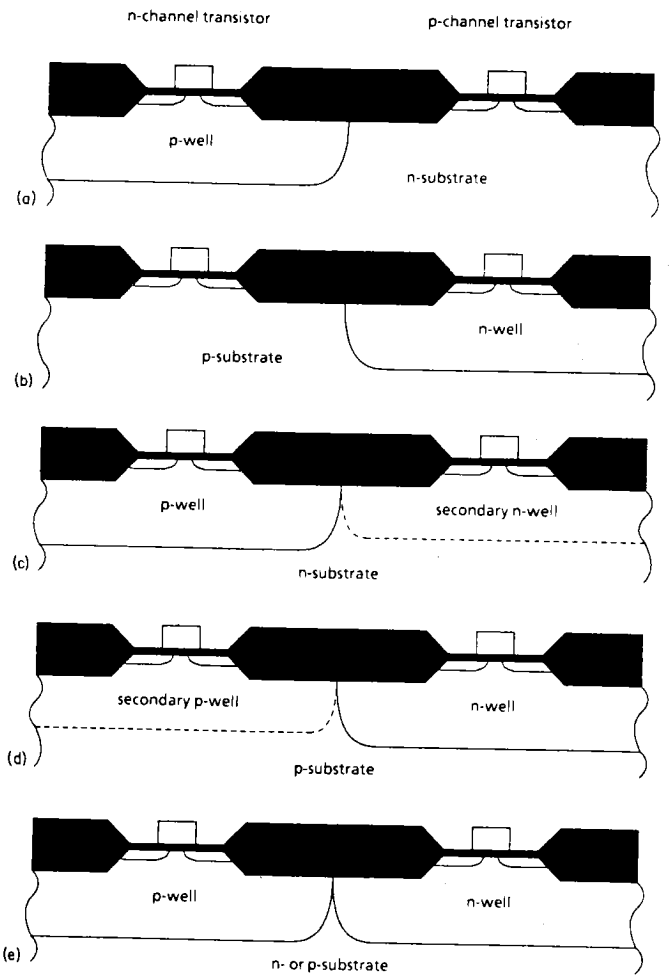


Fig. 8-1 Well structures available for CMOS circuit fabrication: (a) *p*-well CMOS; (b) *n*-well CMOS; (c) *p*-well with secondary *n*-well (twin-well constrained to *n*-substrate); (d) *n*-well with secondary *p*-well (twin-well constrained to *p*-substrate); (e) twin-well CMOS on *n* or *p*-substrate. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics,* Academic Press, San Diego CA, 1989, Reprinted with Permission.

8.1.1 *p*-Well CMOS

The *p*-well process, illustrated in Fig. 8-1a, involves the creation of *p*-regions in an *n*-type substrate to permit the fabrication of NMOS devices. The *p*-wells are formed by implanting a *p*-type dopant into an *n*-substrate at a high enough concentration to overcompensate for the substrate doping and to give adequate control over the *p*-type

doping in the well. The starting *n*-type substrate, however, must also have sufficient doping to ensure that the characteristics of devices fabricated in the substrate regions are adequate (a minimum doping concentration of $3x10^{14}$-$1x10^{15}/cm^3$ is required). If the *p*-well is doped too highly, however, the performance of the *n*-channel devices will be degraded through lower carrier mobility, increased source/drain to *p*-well capacitance, and increased sensitivity to body-biasing effects. The *p*-well doping is therefore usually selected to be about five to ten times higher than the doping in the *n*-substrate.

p-well CMOS was the first type of CMOS that could be practically manufactured, and such circuits were introduced in the late 1960s. The first companies to commercially offer CMOS components produced many designs in *p*-well technology. As this experience was spread throughout the industry, *p*-well CMOS became widely established.

There were several reasons for choosing *p*-well CMOS at that time. First, the earliest MOS ICs were based on a metal-gate PMOS process which utilizes an *n*-substrate. The work-function of aluminum is such that enhancement-mode PMOS devices can be made in a lightly doped *n*-type substrate without the need for additional channel doping. Thus, when CMOS circuits were invented (to take advantage of the reduced power consumption possible with CMOS), it was simplest to manufacture them merely by adding a *p*-well to the basic PMOS process. Second, higher channel doping levels were needed to permit sufficiently large V_T values to be established in the enhancement-mode NMOS transistors. Such higher doping was also needed in the field regions of the NMOS sections of the circuit to ensure adequate isolation (see chap. 6). These requirements are compatible with the need to have well surface concentrations five to ten times larger than the substrate doping level. Finally, it was relatively simple to manufacture *p*-wells with the existing process fabrication technology. That is, in the late 1960s ion implantation was not yet available and it was necessary to introduce the well dopants by chemical diffusion. At the doping levels required in the wells it was much easier to achieve adequate doping profile control using boron rather than phosphorus diffusion processes.

There are also several circuit and system advantages of *p*-well over *n*-well CMOS. First, *p*-well technology may be the better choice for pure-static logic, in which a good balance between the performance of both MOS device types is beneficial. (The performance of the PMOS and NMOS devices are "more balanced" because the carrier mobility reduction associated with higher well doping is compensated by forming the NMOS devices in the *p*-well; i.e., the electrons have higher mobility.) Second, it is attractive for applications that require an isolated *p*-region (such as those using an *npn* bipolar transistor as an on-chip driver or an *n*-channel FETs for an analog input). Third, *p*-well CMOS is less susceptible to field-inversion problems than *n*-well CMOS (which uses a lightly doped *p*-substrate), and can thus be slightly easier to fabricate. (We will describe later how *p*-well CMOS can use the inherent substrate and well doping as the channel stops, whereas *n*-well CMOS must use a separate substrate channel-stop process.) Fourth, if the so-called retrograde-well process (rather than a diffusion of a shallow implant) is used to form the wells, *p*-well technology is more feasible. It is easier to form a *p*-retrograde well than an *n*-retrograde well, since boron ions penetrate deeper than arsenic or phosphorus ions at the same implant energy. Finally, *p*-well

CMOS may be better for fabricating SRAMs. Since the alpha-particle-induced soft-error rate (SER) becomes significant even in SRAMs if feature sizes are scaled to submicron dimensions, the cells should be made inside a well. The sensing of the state of an SRAM cell depends on the current provided by the cell. As a result, high-gain NMOS devices are more desirable for the pass gates and drivers in the cell, and must be built in a *p*-well. Furthermore, it facilitates the use of polysilicon load resistors in 4-transistor memory cells and buried contacts.

Some of the disadvantages of *p*-well CMOS are the following: (1) Since NMOS transistors are fabricated in the *p*-well, their characteristics will be degraded compared to those made in a lightly doped substrate. In CMOS circuits that rely heavily on optimum NMOSFET behavior, this can lead to reduced circuit performance; (2) Substrate currents, the result of impact-ionization generated carriers in short-channel devices (see chaps. 7 and 9), are much larger in NMOS than PMOS devices. It is more difficult to collect these carriers from a well than from the substrate. Since the NMOSFETs in *p*-well CMOS are in the well, this represents a disadvantage of submicron CMOS *p*-well technology.

8.1.2 *n*-Well CMOS

By the late 1970s, the technological barriers that prevented enhancement-mode NMOS transistors from being manufactured were overcome. Primarily, the introduction of ion implantation made V_T control in NMOSFETs a simple, straightforward procedure. As a result, NMOS technologies ousted PMOS in high-speed, high-packing-density ICs. By the mid-1980s, however, the power dissipation in VLSI NMOS ICs became intolerably high (see chap. 6, Vol. 2), and CMOS development was pursued as a mainstream replacement technology for NMOS. For those companies that had been producing NMOS as their main IC product, *n*-well CMOS appeared to be an attractive approach for making the switch to CMOS.

In the *n*-well process, shown in Fig. 8-1b, the *p*-channel devices are formed in the more heavily doped *n*-well, while the NMOS devices are made in the substrate. Since it was now feasible to build NMOS devices in lightly doped substrates, this made it possible to make *n*-channel devices in an *n*-well CMOS process that were identical to those made by NMOS technology. Thus, virtually all of the experience that had been amassed in fabricating high-performance NMOS could be transferred to an *n*-well CMOS process. Furthermore, CMOS circuits could be made that exploited existing NMOS device design. The availability of ion implantation for introducing well dopants into the substrate (instead of having to rely on diffusion), now made dopant control of *n*-wells much more feasible. This lead to the successful development of *n*-well CMOS technology.

As a result, virtually all EPROMs, microprocessors, and dynamic RAM designs in the generations of technology built with 1.25-2.0 μm dimensions were implemented with *n*-well CMOS. Mixed CMOS-NMOS designs were also made, in which the high noise immunity and negligible stand-by-power dissipation of CMOS was exploited in the input/output sections, and faster NMOS devices were used internally.

However, it should be noted that *n*-well technology also has some disadvantages. First, as mentioned earlier, it is more sensitive to field-inversion problems than *p*-well CMOS. Second, it may be more difficult to build pure-static, high-performance logic circuits with *n*-well CMOS.

8.1.3 CMOS on Epitaxial Substrates

As described in vol. 2, chap. 6, heavily doped substrates with a more lightly doped surface epitaxial layer have been utilized to suppress latchup in CMOS.[4] When such starting material is used with single-well CMOS, the epitaxial layer is doped to a concentration equal to that of the substrate in a nonepitaxial wafer used for that process. If a twin-well CMOS process is used, the epitaxial layer is doped to a level significantly lower than that required for building either the *p*- or *n*-channel MOSFETs (see section 8.1.4).

The epitaxial layer is made thicker than the well depth, since the dopants in the heavily doped substrate under the epi layer diffuse toward the surface as the well dopants are diffused toward the bulk (Fig. 8-2). Thus, some of the epitaxial layer becomes more heavily doped during the CMOS process flow. The process is designed so that the bottom of the well is eventually adjacent to the heavily doped substrate region.

Either *n*-epi on n^+ substrates or *p*-epi on p^+ substrates can be used, with each method having advantages and drawbacks. Because the problems with *n*-epi on n^+ are more serious, *p*-epi on p^+ is more widely used. The major limitation of the latter

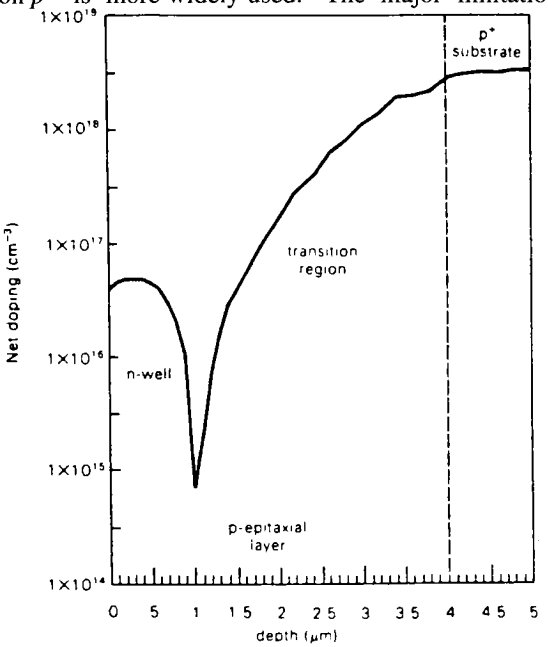


Fig. 8-2 A *p*-on-p^+-epi doping profile and diffused *n*-well profile measured after the entire CMOS process has been completed. The as grown epi thickness is 4 μm and the final lightly doped epi thickness is about 2 μm.[3] (© 1987 IEEE).

approach is that the outdiffusion of boron from the p^+ is much more severe than is the outdiffusion of As or Sb in n-epi on n^+. (The reason for this is that boron diffuses much more rapidly than antimony, which is the most widely-used n^+ dopant material.) Thus, a thicker p^-- epitaxial layer must be used.

In addition, the transition region between the p^+ substrate and the p^--epi layer is thicker, producing a larger series resistance (R_{sub}), which in turn reduces latchup immunity. On the other hand, the p-on-p^+ material is less sensitive to process-induced defects, and the p-type substrate provides higher conductivity under NMOS devices. Such additional conductivity is desirable, since it reduces the voltage drops caused by the substrate currents (generated as a result of the hot-carrier effects in short-channel devices). It is especially important in the regions containing NMOS devices, since the hot-electron substrate current is much higher in such devices than in PMOS devices.

The n-on-n^+ epitaxial substrates also offer some advantages. First, SRAMs are often built on n-type substrates, because the p-well-to-n-substrate junction provides protection from radiation-induced discharging of the memory's n-type storage nodes in the p-well.[5] Second, retrograde p-wells are easier to implement because their implantation energy requirements are much lower. Third, the n-to-n^+ transition region is smaller than that in p-on-p^+ substrates, and the smaller value of R_{sub} provides improved latchup protection.

The limitations of n-epi on n^+ involve the process by which the heavily doped substrate is grown. Antimony (Sb) is the n-type dopant used, both because its diffusion coefficient is so low and because it exhibits much less lateral autodoping than arsenic (the other slow-diffusing n-type dopant). The problem with Sb is that its segregation coefficient, k_O, is very small (i.e., $k_{O,Sb} = 0.023$; see, Vol. 1, chap. 1), and a large quantity of Sb must therefore be put into the Si melt to ensure that a sufficiently heavily doped ingot will be produced. Furthermore, in even the most highly refined Sb there are high concentrations of unwanted heavy metals, which become incorporated into the growing silicon crystal. Finally, the oxygen content of the Sb-doped crystal is relatively small, due in part to the special growing conditions used when the Sb-doped ingot is pulled.[6,7] As a result, intrinsic gettering techniques that would immobilize the metals in the substrate are not as effective as they are in p-on-p^+ epi. For these reasons, n-on-n^+ substrates are less frequently chosen when epi-CMOS is implemented.[4]

A problem that exists with both types of epitaxial substrates for CMOS is that the wells cannot be made too deep, since the lateral diffusion would then take up too much area. On the other hand, if the wells are too shallow, vertical punchthrough (between the substrate and the source/drain regions of devices in the well) will ensue.[8] In this case, appreciable leakage current can flow along this vertical path. As noted, punchthrough can occur if the depletion regions of the p^+/n-well junction and the n-well/p^+-substrate junction merge. The problem is even more severe in the case of the heavily doped p^+ substrate, due to the high degree of boron outdiffusion. A major limitation is thus imposed on the minimum epitaxial-layer thickness.

Another problem of epi-CMOS is back-surface autodoping. For example p-type 20 Ω-cm (7×10^{14} boron/cm^3) epi on a 0.005 Ω-cm (2×10^{19} boron/cm^3) substrate is representative of epi for CMOS devices. If the back surface of the substrate is not

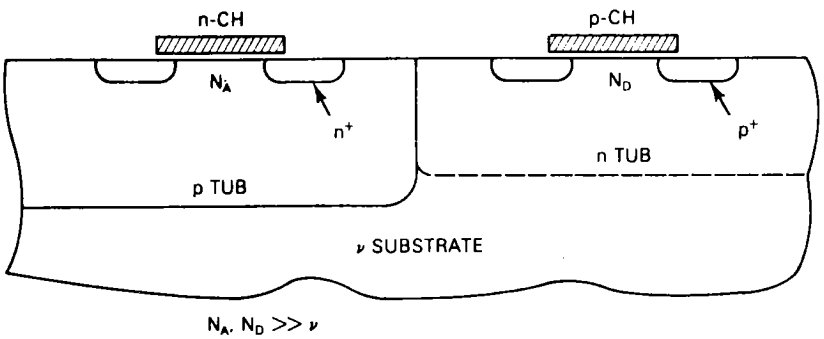


Fig. 8-3 Cross section of a twin-well 1.2-μm CMOS structure.[5] Reprinted with permission of Semiconductor International.

sealed (e.g., by using a sealing layer such as undoped silicon dioxide or silicon nitride), boron evaporation from the substrate can contribute to autodoping on the front surface during the entire epi deposition cycle. This widens the epi/substrate interface and may even prevent the epi from reaching the 20 Ω-cm specified resistivity. Note that if silicon nitride is used as a sealing layer it should only be used in thin layers (e.g., less than 100 nm thick) since its high intrinsic stress causes it to bow the silicon.[9]

8.1.4 Twin-Well CMOS

With the twin-tub approach, two separate wells are formed for *n*- and *p*-channel transistors on a lightly doped substrate (Figs. 8-1c and 8-3). The substrate may be either a lightly doped wafer of *n* or *p* material, or a thin, lightly doped epitaxial layer on a heavily doped substrate. In either case, the level of the well surface doping is significantly lower than that required for building either the *p*- or *n*-channel MOSFETs. Each of the well dopants is implanted separately into the lightly doped surface region and is then driven in to the desired depth.

In twin-well CMOS, the doping profiles of each of the device types can be set independently, since the constraint of single-well CMOS does not exist (i.e., that the well doping must always be higher than the doping of the substrate in which one type of device is made). This was originally cited as an advantage of twin-well CMOS over single-well CMOS, with the argument made that both device types could thus be optimized.[12] This benefit for 1-2 μm CMOS has been questioned by Chen,[2] who points out that in modern single-well CMOS processes, an additional implant is used to prevent punchthrough without the need to raise the entire substrate-doping concentration. By incorporating this additional implant, it is possible to build higher-performance devices than can be achieved with the twin-well approach, in terms of junction capacitance and sensitivity to body effect.

Twin-well CMOS does offer some significant benefits for devices with submicron-channel lengths (although these advantages are gained at the cost of greater process

complexity). The first, and most important advantage arises when devices with submicron channel lengths are fabricated. Since it is recognized that the two device types perform similarly as channel lengths approach 0.5 μm, it is useful to provide symmetrical n- and p-channel devices. Furthermore, at submicron dimensions the body doping of both transistor types must be raised significantly to prevent punchthrough and to maintain adequate threshold-voltage levels. Thus, the advantage of having one type of MOS transistor in a lightly doped region (to optimize its performance at the expense of the other) disappears. It is instead more beneficial to produce two types of active device wells, each with its own optimized doping profile (i.e., formed by separate implants into a lightly doped substrate).*

The second advantage of the twin-well process is that it is compatible with the technologies of either isolation by selective epitaxial growth (SEG) or trench isolation.[1] Both approaches restrict the lateral diffusion of the dopants in each of the wells. In addition, sidewall inversion along the trench is less likely when both device sidewalls are butted against a trench that has been formed in a more highly doped well.[10] Finally, when deep trenches are used with a thin epitaxial layer on a heavily doped substrate, latchup can be eliminated. The combined use of the twin-well process and advanced isolation techniques allows n^+ to p^+ spacing to be dramatically reduced in comparison to single-well technologies.

A third benefit is that the combination of epitaxial substrates and the twin-well process allows either substrate type to be chosen with no effects on transistor performance and essentially no change in process flow. Such flexibility is useful since some applications are best met with n^+ substrates and others with p^+ substrates. This advantage may be important if a single process is needed for implementing designs with different applications.

Finally, self-aligned channel stops can be easily implemented in the twin-well approach, allowing the spacing between n- and p-channel devices to be reduced. Although this spacing reduction is not as great as it would be if trenches or SEG were used, the process is much less complex with the twin-well method.

A twin-retrograde-well 0.7-μm-CMOS process for fabricating 1-Mbit SRAMs has been reported.[11] The high energy implants also allow a restricted thermal budget to be used, thus reducing the up-diffusion from the p^+ substrate. This permits the use of a thinner epi layer (which also helps prevent latch-up), and allows implantation of the channel-stop dopants after the field oxide has been grown and planarized (consequently, minimizing lateral diffusion of the boron channel-stop dopants). We discuss retrograde-well CMOS in more detail in section 8.3.

* It is not useful to start with a substrate doped to the optimum level needed for just one of the submicron devices. If this were done, a single well of much higher doping would have to be established for the other type of submicron device, which would unnecessarily degrade the device performance. It is possible, with the twin-well approach, to have both types placed in a well of optimum doping profile.

8.1.5 Summary of CMOS Well-Technology Issues

The selection of a particular CMOS process depends largely upon circuit applications, and to a lesser degree on technology evolution. For VLSI circuits with 1-2 μm design rules n-well CMOS has been the more widely used single-well technology, both because n-channel devices can be optimized in the p-substrate and because most circuits use more NMOS than PMOS devices. In circuits that use transistor channel lengths smaller than 1 μm, twin-well and retrograde-well technology are more attractive. As noted, the price paid for their benefits is increased process complexity.

The optimum process-integration design strategy is to first select the well technology (and to decide whether or not to use epitaxial substrates), and then to select the isolation method. Once these decisions have been made, the well depth and doping profile can be determined. The well depth impacts the lateral-diffusion distance of the well and the vertical-punchthrough voltage. The doping profile affects transconductance, threshold voltage, source/drain punchthrough, junction capacitance, carrier mobility, source/drain-to-substrate breakdown, sensitivity to body effect, and hot-electron effects.

As previously discussed, early CMOS used lightly doped n-type wafers and diffused p-wells (in which the NMOSFETs were fabricated). The p-well approach was initially adopted because NMOSFETs could only be reliably produced in more heavily doped regions until ion implantation was developed in the mid 1970's (see chaps. 5 and 6, Vol. 2). Eventually, however, n-well as well as p-well single-well CMOS processes were implemented. Later twin-well processes were introduced, and these now dominate the newest CMOS fabrication lines (although older n- and p-well single-well technologies are still used to fabricate a variety of CMOS products).

The well-formation process was also one that quickly exploited the technology of ion implantation. The doping precision required in the well gave implantation a significant edge over diffusion as a method for introducing the dopants into the substrate. Advances in implantation technology are allowing development of processes involving multiple implants and higher energies to avoid the long, high-temperature drive-in and to form retrograde well profiles (as will be discussed in sections 8.3 and 8.4).

8.2 CONVENTIONAL FORMATION OF WELLS AND CHANNEL STOPS

There are several issues associated with conventional-well and channel-stop formation that are covered in this section, including: (1) alignment of subsequent patterns to the well locations; (2) process sequences for forming twin-well and channel-stop regions, and; (3) novel well-formation schemes that reduce mask count and process complexity.

8.2.1 Aligning Subsequent Patterns to the Wells

The wells are often the first features to be formed in the CMOS process flow. However, the implant process that introduces the well dopants does not produce visible features on the wafer surface. Thus, it is necessary to provide some indirect indication of the well locations so that subsequent layers can be aligned to them. Here we illustrate some

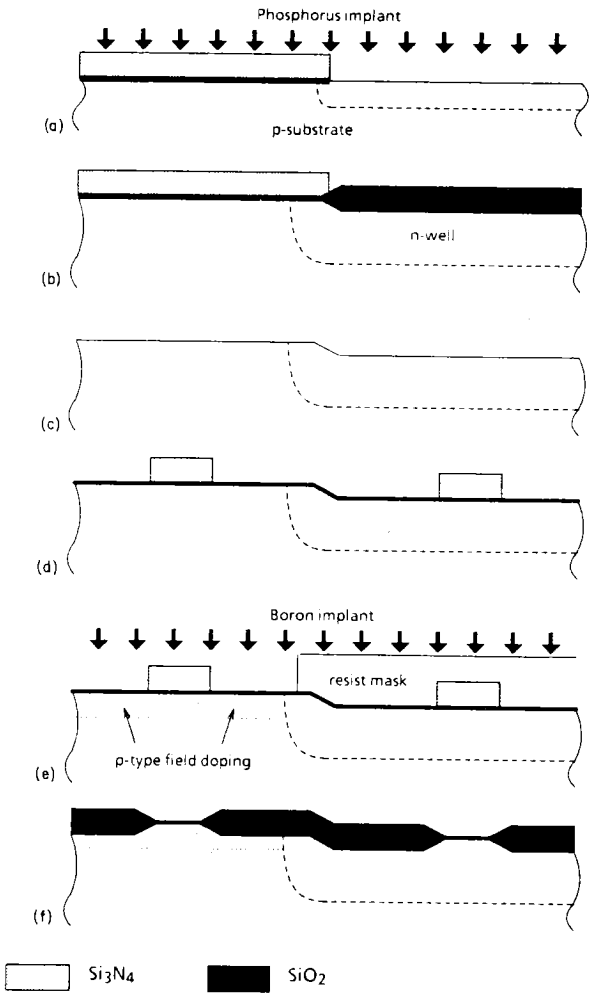


Fig. 8-4 Conventional *n*-well architecture with *n*-well as first mask. (a) Pad oxide/nitride well mask and well implant. (b) Well oxidation and drive-in. (c) Nitride/oxide strip. (d) LOCOS pad oxide/nitride mask and active area definition. (e) NMOS field implant. (f) LOCOS field oxidation. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics,* Academic Press, San Diego CA, 1989, Reprinted with Permission.

methods that have been developed to provide such alignment marks. For simplicity we utilize an *n*-well CMOS structure in depicting the process flows, but later we show how these methods can be incorporated into twin-well CMOS processes.

In the first approach the wells are the initial features created on the wafer. Here a nitride mask is employed to both define the well locations and to serve as an oxidation mask (see Fig. 8-4a and 8-4b). That is, the patterned nitride/resist layer blocks the phosphorus *n*-well-implant from penetrating to the *p*-substrate (or *p*-well) regions. Part of the drive-in step after the implant is carried out in an oxidizing ambient with the nitride mask layer left in place. This causes a thermal oxide layer to be grown selectively over the *n*-well regions. At the conclusion of the drive-in process, this oxide is stripped. Since the oxide growth consumes some silicon in the *n*-regions and since no such silicon loss occurs in the regions covered by the nitride mask, a step is produced on wafer surface when the oxide is stripped (Fig. 8-4c). This step is also is self-aligned to the well edge, therefore providing a visible indication of the well boundaries that allows the active area features to be aligned to them. (Subsequent layers, however, are usually aligned to the active area.) The disadvantage of this approach is the step that is created on the wafer surface. In submicron processing such nonplanar topography is undesirable.

The second approach is a simpler process and it also avoids the creation of the step produced in the previous method (see Fig. 8-5). Here, however, the first features to be defined are the active areas (Mask #1). That is, a nitride mask stack layer is patterned first and all subsequent layers are aligned to the active area pattern. Then the well mask (Mask #2) is aligned to the active area pattern, and the well implant is performed. Note that the implanted dopants must penetrate the nitride mask stack. This does not present a serious difficulty, but it does mean that the *n*-well tends to be shallower and slightly less-heavily doped beneath active regions than under the field areas. Both the first and second approaches require three masking steps (including the boron channel-stop implant that is performed after the well implant - but prior to the LOCOS field oxidation process). However, the second approach offers the benefit of needing fewer oxidations and depositions.

The first method can be applied to a twin-well process by growing an oxide after the *n*-well implant (Mask #1), but not doing much of the drive-in step during this oxidation. Then the nitride mask is stripped and a second *p*-well mask is aligned to the *n*-well oxide (Mask #2). After doing the *p*-well implant the well drive-in process is performed. An even easier (i.e., one mask) procedure would be to use the *n*-well oxide to mask the *p*-well implant (instead of using the second mask). The second method can also be applied to a twin-well process but two masks must be used to create the twin wells. However, if a shallow *p*-well is acceptable, the same resist mask that is used to form the channel-stop region in Fig. 8-4e (Mask #3) can be used to form the shallow *p*-well. In this case the *p*-well implant must be performed at a higher energy than the field implant in order to penetrate the thin nitride mask layer.

The third method utilizes an initial alignment mark etched into the silicon surface prior to forming any device features. This mark is also used as an automatic alignment target by the wafer stepper. Figure 8-6 shows an example of such an alignment pattern.[16] The use of such initial alignment marks is likely to increase with the proliferation of automatic stepper alignment.

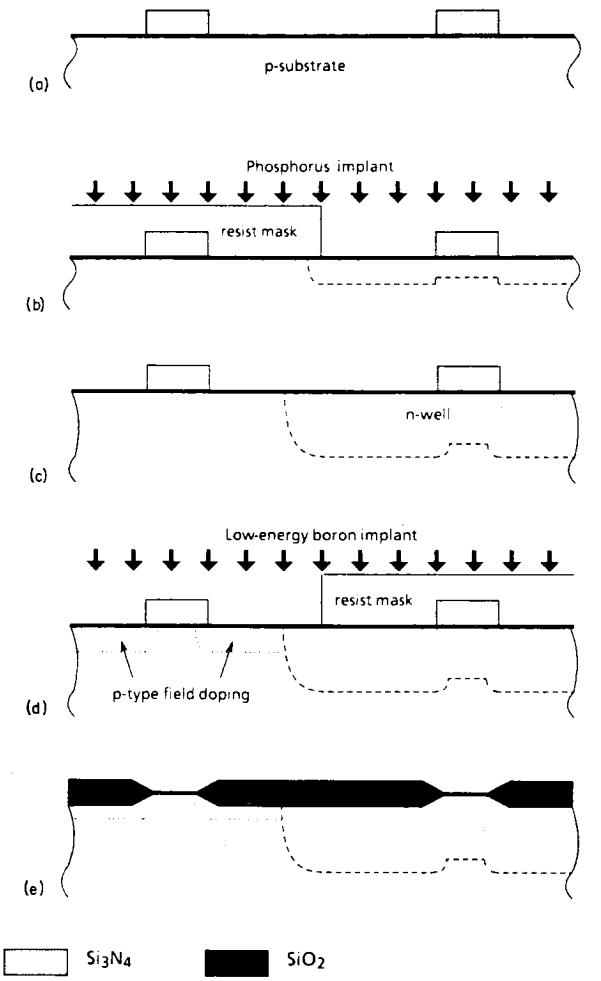


Fig. 8-5 Conventional *n*-well architecture with active area as first mask. (a) Active area definition using LOCOS nitride mask. (b) *n*-well mask and implant. (c) *n*-well drive-in. (d) boron field implant, and (e) LOCOS field oxidation. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics,* Academic Press, San Diego CA, 1989, Reprinted with Permission.

8.2.2 Twin-Well and Channel-Stop Formation.

The most obvious method of forming twin wells is to use two masking steps, each of which blocks one of the well implants. A procedure for forming twin-wells with a single masking step, however, has also been suggested (see Fig. 8-7).[12] In this method,

a single mask is used to pattern a nitride/oxide film formed on the bare silicon surface (*Mask #1*). The openings in the film become the *n*-well regions, and phosphorus is implanted into them (e.g., at 80 keV). Next, a thermal oxide is selectively grown on the *n*-well regions, to a thickness that is sufficient to block the boron implant used to form the *p*-wells. (The nitride mask layer is used to prevent the oxide from growing in the *p*-well regions.) The pad-oxide/nitride mask layer is then stripped, exposing the silicon in the *p*-well regions. The *n*-well regions are however, still covered with the oxide grown earlier. Thus, during the *p*-well implant (e.g., at 50 keV), this oxide masks the *n*-well regions, permitting the boron to enter the silicon only in the *p*-well areas.

Next the wells are driven in (e.g., at 1100°C for 500 min). At the conclusion of this drive-in step the concentrations in the wells for an 0.8-μm-CMOS process are ~$1x10^{16}/cm^3$ for the *p*-well, and ~$3x10^{16}/cm^3$ for the *n*-well. The *n*-well is selected to have a higher doping concentration (i.e., to improve the punchthrough performance of the PMOS devices and to eliminate the need for a separate channel-stop implant for the *n*-well). A higher concentration in both wells would still produce devices with relatively low capacitances at the bottoms of the source/drain-to-well junctions.

The channel-stop implant procedure is also usually included in the part of the processing sequence in which the wells are formed. In one reported twin-well process,[13] only a *p*-well channel-stop implant (boron) is used, because the doping in the *n*-well is high enough ($3x10^{16}$ cm^{-3}) that a separate channel stop implant is not required. In this case, an unmasked boron implant is performed following both well implants, but prior to field-oxide growth.

In a second approach,[14] an additional mask can be used to provide both *n*-well and *p*-well channel stops (Fig. 8-8a). Note that in this procedure, a mask is saved by

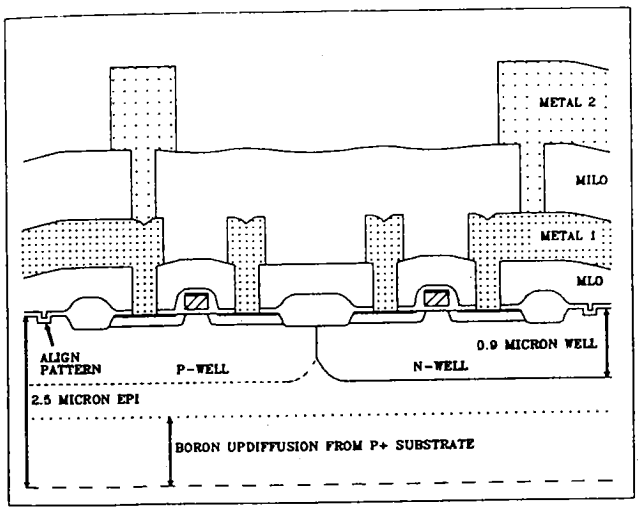


Fig. 8-6 Twin-well CMOS showing initial alignment mark.[16] (© 1991 IEEE).

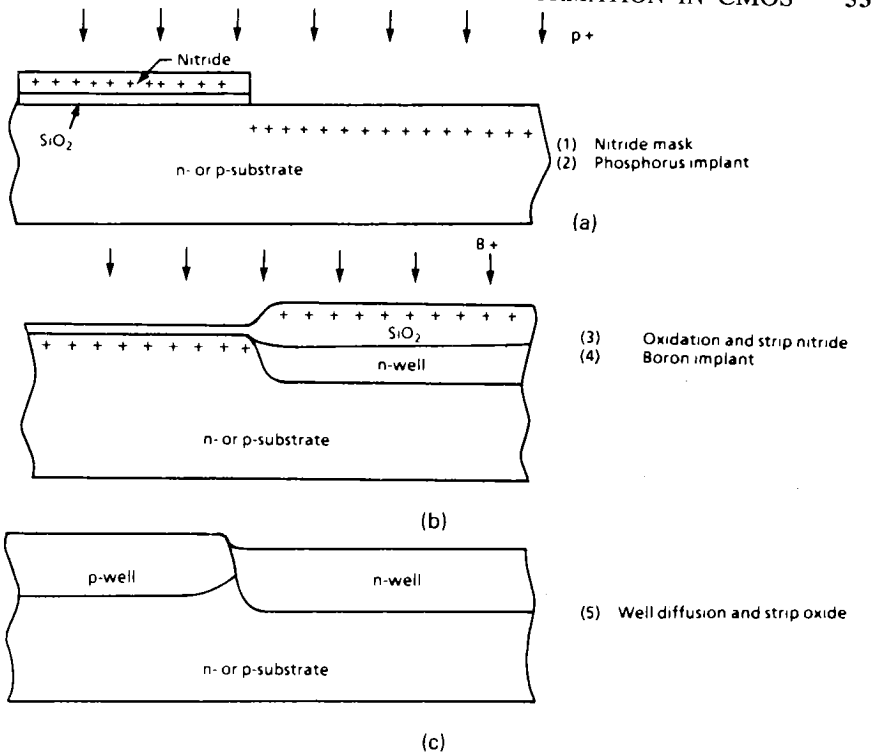


Fig. 8-7 Twin-well CMOS process using a single mask to form both wells. (a) P^+ implant for the *n*-well. (b) B^- implant to form the *p*-well (blocked by the *n*-well oxide). (c) well drive-in for both wells.[12] (© 1982 IEEE).

performing a blanket boron implant. As a result, the boron channel stop dopant ends up in the *n*-well as well as in the *p*-well. Hence, the phosphorus channel stop dose must be increased to compensate for the boron's presence in the *n*-well field regions. The disadvantages of this method include the additional alignment step between the channel stops and the well masks, the interdiffusion that occurs during the oxidation step, and the asymmetry of the doping profiles. (The latter is due to the increased phosphorus concentration that must overcompensate the nonselective boron channel stop).

A variation of this method that requires no masks at all to create the channel stops has been developed to overcome the above drawbacks (Fig. 8-8b).[14,15] In this sequence, both arsenic and phosphorus are co-implanted into the *n*-well regions. This places both the dopants that form the well and the channel stop into the *n*-well regions prior to implanting of the *p*-wells. An oxide is then selectively grown on the *n*-well regions and the boron dopant for the *p*-wells is implanted. After the wells have been driven in, the oxide over the *n*-wells is retained, and a second boron implant is carried out. This implant serves as both a channel stop in the *p*-well *field* regions and a punchthrough-prevention implant in the *active* regions of the *p*-well.

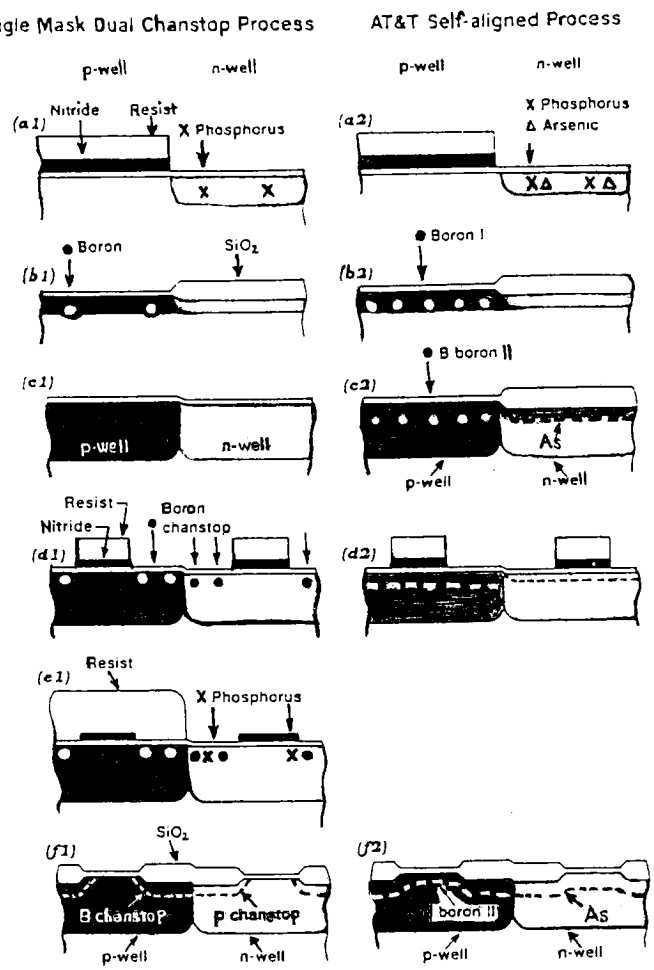


Fig. 8-8 (a) Twin-well CMOS with a single-mask dual channel stop process. (b) ATT twin-well CMOS process which requires no masks to form the dual channel stops (self-aligned process.)[15] (© IEEE 1981).

It should also be noted that the minimum separation between two wells of the doping type at two different potentials is limited by electrical isolation. In single-well processes, to prevent punchthrough in the lightly-doped substrate between two such wells, adequate spacing between them must be allowed.

8.2.3 Split Well-Drive-In Process

In another report, the long, high-temperature drive-in step is performed in a novel manner, allowing the *p*-well transistors to be fabricated without use of a channel-stop

implant.[17] After the *p*-well implant is carried out, the field oxide is first grown without driving the implanted well dopants. During this step, some of the implanted boron impurities segregate into the field oxide, leaving a surface concentration as low as would exist after any field oxide step. Next, the remainder of the well drive-in step is performed in a non-oxidizing ambient. The well dopant deep in the bulk now back-diffuses toward the surface, raising the Si-SiO_2 interface concentration by a factor of 2-3 (to 2-3x10^{16} cm^{-3} in this case, see Fig. 8-9). The well concentration at surface of the field regions is only about 1.5x lower than at the surface in the active regions. V_{TFiso}s greater than 10 V (at I_D at 0.5 pA/μm at V_{DS} = 8V) can be produced for isolation lengths of 1.2 μm. It has been observed that for very narrow NMOSFETs (e.g., 0.35 μm) fabricated with the split-well drive-in scheme, V_T is decreased by ~100 mV compared to wider channel devices (e.g., 5 μm). This occurs because boron also diffuses from the channel regions to the field regions during the post-oxidation drive-in step. Such diffusion occurs because, as described earlier, the boron concentration at the silicon surface in the field regions has been reduced during the field oxide step by boron segregation effects.

8.3 RETROGRADE-WELL CMOS

As previously discussed, conventional wells are formed in single- and twin-well CMOS technology by implanting the well dopants and then diffusing them to the desired depth. However, lateral as well as vertical diffusion occurs, which has the effect of reducing packing density. For example, since the lateral diffusion is ~0.7 times the junction

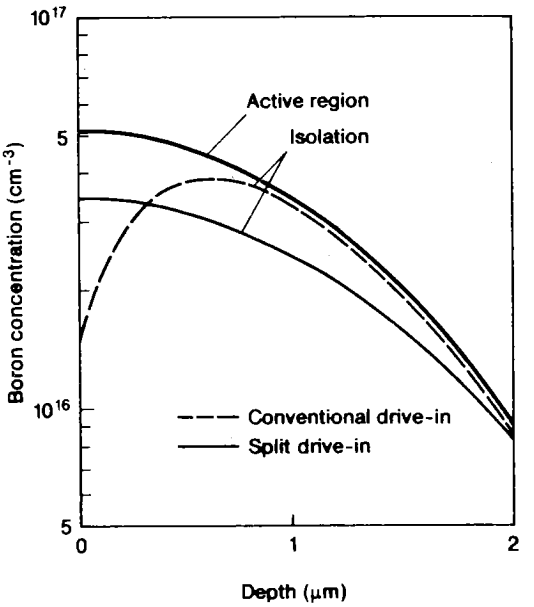


Fig. 8-9 SIMS profiles of boron well underneath field oxide and in the active area for split and non-split drive-in. Implanted boron dose: 8x10^{12} cm^{-2}. The active area profile is the same for both drive-in cases.[17] (© 1990 IEEE).

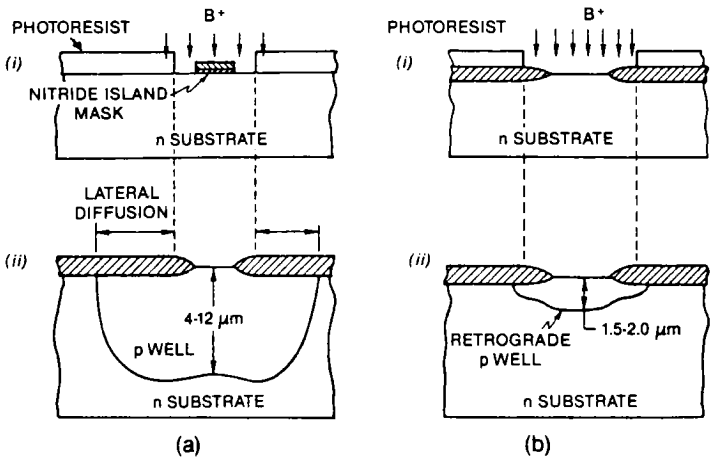


Fig. 8-10 Formation of (a) conventional and (b) retrograde p-well CMOS structures.[19] (© 1981 IEEE).

depth, a typical diffused well of 3 to 4 μm results in a lateral spread of well dopant of ~2 to 3 μm. If a higher energy implant is used to place the dopants at the desired depth without significant additional diffusion (i.e., the high-energy implant is performed after the active area isolation structures [field oxide] is already in place, and only a short anneal at a relatively low temperature is used to activate the implant), much less lateral dopant spread occurs. This is because ion implantation is a more anisotropic dopant distribution process than diffusion. That is, the lateral implant straggle is ~0.7 μm and barely exceeds one micron after all thermal processing.[18] In addition, since the high-energy implanted wells are formed after the LOCOS isolation step, lateral spread of the well dopants is further reduced. Hence, the 7-9 μm n^+-to-p^+ spacing required for a twin-well 1.2 μm CMOS process which uses conventional wells, can be reduced to ~5 μm if retrograde wells are used instead.

Figure 8-10 depicts the cross-section of a high-energy implanted p-well in comparison with a conventional diffused well. In this example, the implanted p-well is formed by boron implantation at 400-600 keV, followed by a brief 30 min anneal at 1000°C (instead of the 6-20 hour drive-ins at 1100-1150°C used for conventional-well formation). Notice the lateral spread associated with the p-well is greatly reduced.

Such high-energy implants also cause the peak of the implant to be buried at a certain depth within the silicon substrate (depending on the implantation energy, but at roughly 1 μm for the energies most frequently used), and the impurity concentration decreases as the doping profile approaches the wafer surface (see Fig. 8-11). Since the well profile in this case is different from that of conventionally formed wells (in which the doping concentration after drive-in is highest at the surface, and decreases monotonically with depth), such deeply implanted wells are known as *retrograde wells*. It is estimated that, in order to avoid excessive spread of the active device electrical

parameters, the surface concentration after the retrograde implant should be at least five times smaller than its final value, which is set by shallower low-dose implants specifically designed for threshold and punchthrough control. Comparison of the doping profiles in conventional and implanted p-well profiles (obtained by simulation) is shown in Fig. 8-11.[19] The retrograde-doping profile is retained by minimizing the temperature cycles of later process steps (e.g., by using RTP). Retrograde wells can be implemented on both bulk and epitaxial wafers.[20]

Besides the potential benefit of increased packing density, retrograde wells offer the following advantages:[1]

1. A retarded electric field is created in the parasitic vertical bipolar transistor, thereby reducing its current gain. This affords some protection against latchup.

2. Susceptibility to vertical punchthrough is reduced without having to significantly raise the surface doping concentration.

3. The conductivity in the bottom of the well is increased, which also provides additional latchup protection.

4. A higher threshold voltage can be achieved in the field regions of the p-wells if the boron implant is done following field oxidation (i.e., the boron does not segregate out into the oxide during the field oxide growth step).[19]

5. Lateral diffusion of the channel-stop boron is also eliminated, thereby reducing encroachment of the boron into the active regions. This lessens narrow width effects in the active devices.

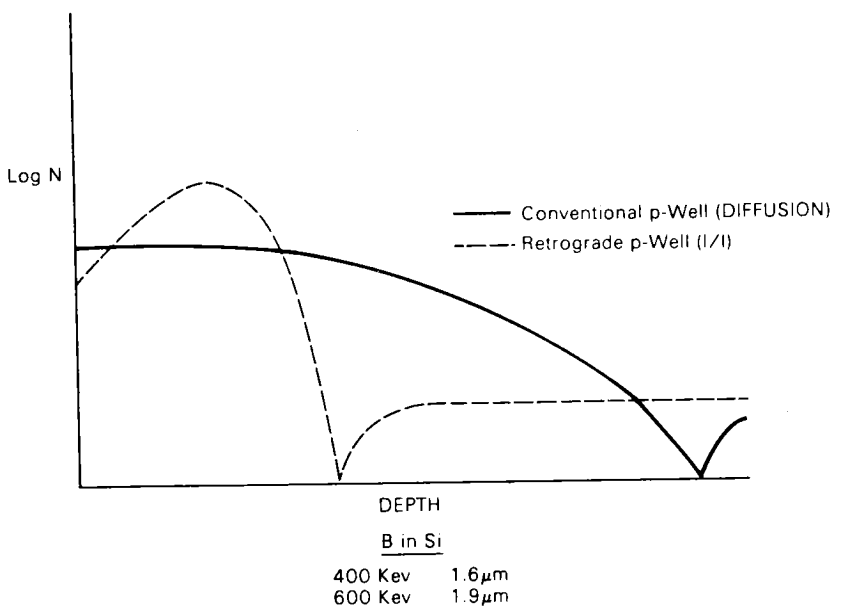


Fig. 8-11 Comparison of conventional and retrograde well p-well profiles.[19] (© 1981 IEEE).

6. Greater flexibility exists in retrograde structures to optimize channel doping profiles for active device performance. That is, in conventional wells (especially shallow ones), the surface doping is primarily determined by the need to maintain sufficient vertical isolation. In retrograde structures, however, vertical isolation is secured by the high doping level towards the well bottom. Thus, it is possible to adjust the surface doping profile to a larger extent. This helps to achieve adequate short-channel device characteristics with retrograde wells.

7. The shallower well depth should also make the drain of the parasitic n^+-to-n-well (or p^+-to-p-well) parasitic FETs shallower. This should improve both the short-channel threshold roll-off and the punchthrough characteristics of these parasitic structures, as is discussed further in chap. 6.

However, several concerns associated with the retrograde process should also be mentioned. First, the process requires a thick photoresist as a masking material for the high-energy implant. Typical resist thicknesses are about 5 μm, and special exposure and development techniques are required for this process. Second, there is concern about the annealability of the damage caused by the high-energy implant. Third, the retrograde implant may cause both the junction capacitance and the body factor to be significantly increased. The p^+-to-n-well junction capacitances measured in retrograde-well structures formed with a 700 keV implant have been shown to be about twice the values obtained in conventional wells. While this can adversely impact the propagation delay, circuit performance is also impacted by a number of other parasitic capacitances, and the net result on typical CMOS circuit speed and power dissipation is often relatively small. For example, in a simulation of a 32-bit CMOS arithmetic logic unit (ALU), it was found that the circuit delay increases by about 7% as the p-channel junction capacitance increases by 30 percent.[1]

When the retrograde well is formed by means of very high-energy implants, however, the doping concentration under the bottom of the source and drain regions is decreased, which reduces the junction capacitance. This implies that if such high-energy ion implanters (~1 MeV) become available for production, some of the disadvantages of retrograde wells can be overcome.[18] In addition, the higher doping that will be required in the wells for fabrication of submicron CMOS devices may mean a less severe junction capacitance penalty.

Although both p-type[21] and n-type[22] retrograde wells have been demonstrated (see Figs. 8-12a and 8-12b,[20] and Fig. 8-12c,[23] respectively), the p-type technology has been more widely implemented. This is because at least 600-keV beam energies are required for the formation of n-type (phosphorus) retrograde wells suitable for 0.5-0.8 μm CMOS. As of 1994 such machines were commercially available, but had only begun to be introduced into mainstream production environments. Retrograde p-type wells adequate for 0.5 μm devices require implants with energies of ~300 keV at a dose of ~1×10^{13} cm^{-2}. Such p-well implants can be achieved by implanting singly ionized boron at 300 keV or doubly ionized boron at 150 keV. (The required p-well energy is smaller because boron has a larger projected range than either As or P; see Vol. 1, chap. 9). Although the doubly ionized boron approach is achievable with conventional

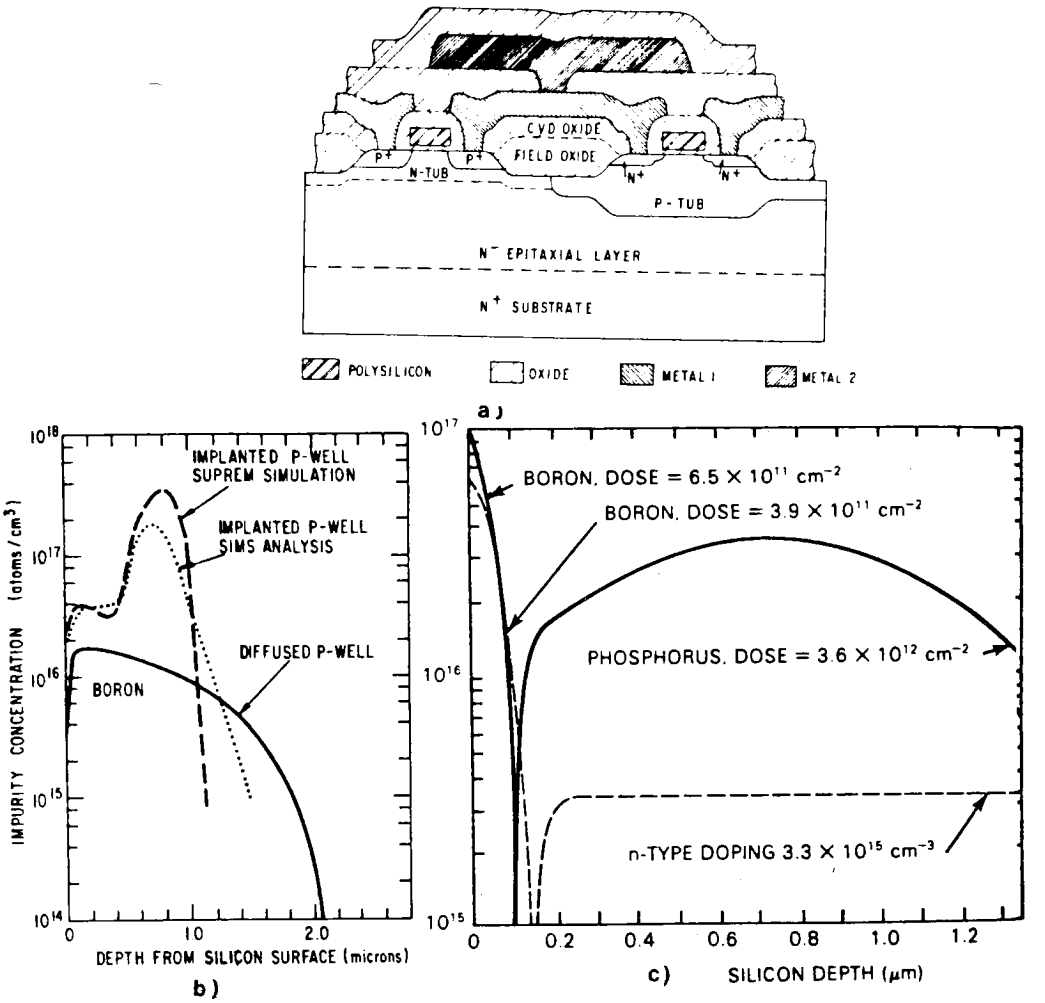


Fig. 8-12 (a) Twin-tub device cross section in which a retrograde p-well is used. General Electric AVLSI Process. (b) Retrograded p-well implanted impurity concentration profile. General Electric AVLSI process. Also shown is a conventionally thermally diffused well.[20] (© 1986 IEEE). (c) Simulated doping profiles in PMOS channel regions: dashed lines, lightly doped substrate; solid lines, retrograde n-well.[23] (© 1982 IEEE).

production implanters it is not a routine process to implement.* Subject to the precautions mentioned in the footnote, retrograde well CMOS processes have been

* One problem is that a large fraction of double-ionized ions, as high as 20%, can interact with gas molecules at the entrance of the acceleration column and become single-charged ions because of partial neutralization. These ions would only penetrate into silicon about half the desired range, distorting the profile and causing abnormalities of the threshold

the successfully implemented in a production environment with good results. However, these have primarily been limited to custom and advanced system applications, where higher costs of retrograde well processing are justified by the higher density and performance benefits offered by the retrograde-well device structures.

Two examples in which a 300 keV boron implant is used to form a retrograde p-well CMOS structure have recently been published. In the first, this p-well structure (together with a 1.2 μm epi layer) allowed an NMOS device-to-device spacing of less than 1 μm to be used when an optimized field implant was employed.[25] In the second, the p-well was fabricated with a 320 keV implant (using, in fact, doubly-charged boron accelerated at 160 keV), and a 5 min 1100°C RTP anneal. Together with the use of a 2-2.5 μm epi layer, a sub-half micron CMOS structure was fabricated (which permitted an n^+ to p^+ spacing of less than 2.5 μm).[16]

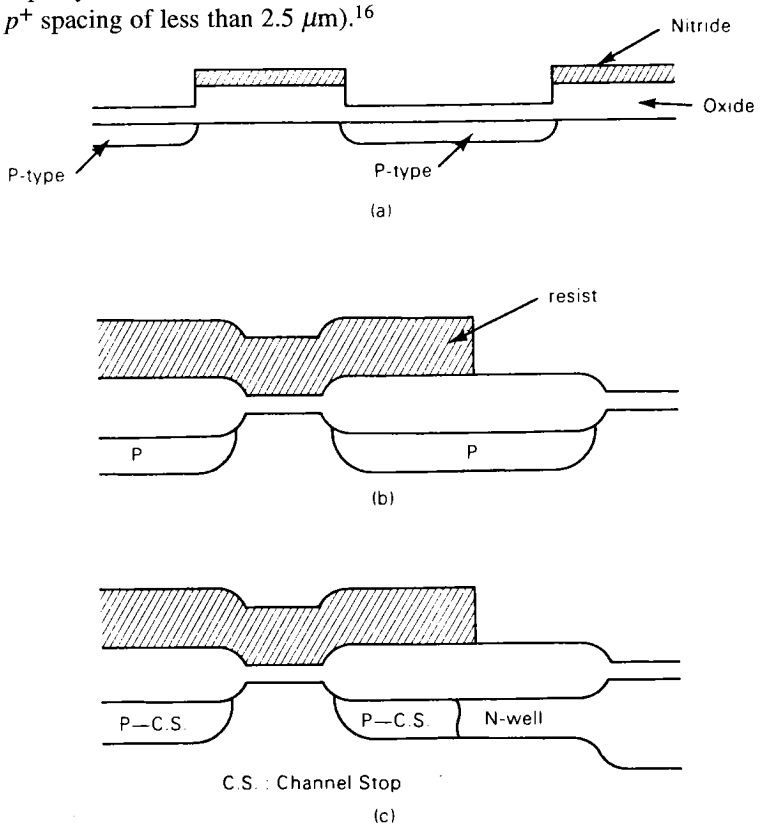


Fig. 8-13 A simplified retrograde n-well process. (a) Nitride mask for active area and B^+ implant for p-type channel stop.; (b) photoresist mask for n-well; and (c) high energy P^+ implant and any additional shallow implants in the n-well.[18] (© 1986 IEEE).

* (cont.) voltage and other electrical parameters. Thus, if a high-current implanter is not available, a medium-current implanter of the post-analysis type should be used to eliminate single-charge ions before they reach the wafers.[24]

However, an n-well retrograde CMOS process using a 700-keV phosphorus implant has also been demonstrated.[18] The process sequence used to form this retrograde n-well CMOS structure (with the well implant performed after field oxidation) is shown in Fig. 8-13. The active areas are first defined using nitride masking. Next, a blanket low-energy boron channel-stop implant is performed. The field oxidation is then carried out and the retrograde n-well implant step is performed. A 30 minute, 1000°C anneal follows the implant. Note that the blanket boron channel-stop dopant is overcompensated in the n-well regions by the phosphorus implant. This retrograde-well process permits very small (~2 μm) n^+-to-p^+ spacings.

Some advanced twin-well CMOS processes with both retrograde n- and p-wells, (using short [~5 min] RTP well-drive-in steps), have also been reported for submicron applications.[26] Scaling CMOS to 0.5 μm and below reduces the well depths required to prevent vertical punchthrough and latchup. In addition, the well depth and well-dopant concentration can be reduced if additional, lower energy implants are precisely located with respect to the transistor structure to suppress punchthrough of the MOSFET channels and to enhance the field isolation characteristics.

One novel method of achieving such short well-drive-in times is to enhance the effective diffusion rate. It has been demonstrated that the diffusion of boron and phosphorus is substantially enhanced by oxynitridation in an NH_3 ambient with a thin oxide layer over the diffusion region.[27]* These factors combine to produce a trend of lower implant energy in retrograde wells for scaled devices at and below 0.5 μm.

A twin-retrograde-well 0.7-μm-CMOS process for fabricating 1-Mbit SRAMs on p^--on-p^+ epi substrates has been reported.[11] The high energy implants also allow a restricted thermal budget to be used, thus reducing the up-diffusion from the p^+ substrate. This permits the use of a thinner epi layer (which also helps prevent latch-up), and also allows implantation of the channel-stop dopants after the field oxide has been grown and planarized (thus minimizing lateral diffusion of the boron channel-stop dopants).

The use of retrograde wells for both the n- and p-wells of a twin-tub CMOS process has also been reported.[29] The approach was called *quadruple-well CMOS* to emphasize the fact that for each type (n- or p-well), two wells were needed, one in the active area, and another, shallower one in the field region. The deep and shallow wells, however, are created simultaneously using the same higher energy implant steps. That is, since the implant is carried out after the field oxidation (and after the LOCOS nitride mask has been stripped), in regions of the wafer where no field oxide exists the implant penetrates deeply, but in regions where the implant must penetrate the field oxide, it enters the Si substrate less deeply (creating a shallower well there).

Another approach for fabricating retrograde wells without the need for very high ion implantation is to use epitaxy. Such a method for forming an n-well CMOS structure is shown in Fig. 8-14. Shallow, heavily doped wells are formed in the p-substrate by ion implantation of a slow diffusing dopant (typically arsenic) in order to prevent significant redistribution during subsequent processing. A p-type epitaxial layer is then grown to a thickness of a few microns, creating an n^+ buried layer. A low-energy phosphorus implant is then used to complete the n-wells. No drive-in is needed if it is

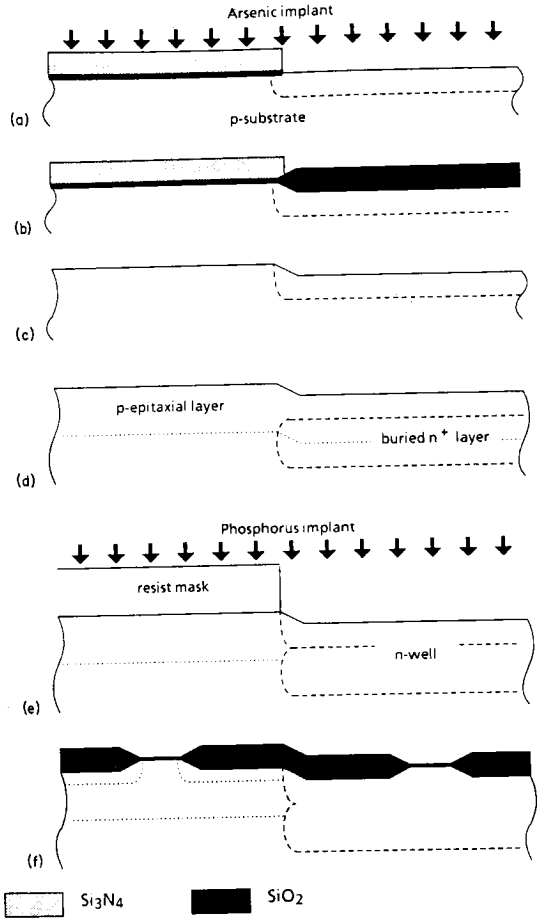


Fig. 8-14 Retrograde *n*-well architecture based on epitaxy. (a) Oxide/nitride well mask and buried layer implant. (b) Well oxidation. (c) Nitride/oxide strip. (d) Epitaxial growth. (e) Surface *n*-well implant. (f) Boron field implant and standard LOCOS field oxidation. From A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics,* Academic Press, San Diego CA, 1989, Reprinted with Permission.

arranged that the surface phosphorus implant merges with the buried layer to complete the *n*-well. The formation of buried layers in this way is especially well-suited to BiCMOS technologies.

8.3.1 Retrograde Well Technologies to Improve Isolation

Since retrograde wells are formed by a high-energy implantation step following the LOCOS oxidation growth process, lateral diffusion of the well dopants is greatly reduced compared to a conventional well process. This results in significant saving of layout space. Isolation among opposite type devices is also expected to be improved.

Here we present data that supports this latter claim, as a supplement to the material on isolation presented in chap. 6.

The comparison of isolation performance between conventional-well and retrograde-well CMOS structures was performed using an *n*-well process and aluminum-gate MOSFETs.[18] To establish a fair comparison, equal well depths were used in the active area of both the conventional-well and retrograde-well structures.

The process sequences used to form both the conventional-wells and retrograde-wells in an *n*-well CMOS process are shown in Figs. 8-15a and 8-15b, respectively.[30] The active areas are first defined in both structures using nitride masking. In the retrograde structures, a blanket low-energy boron channel-stop implant is performed next. In the case of the conventional well, the *n*-well resist mask is aligned to the patterned nitride, and the *n*-well implant (phosphorus) is carried out, followed by a blanket low-energy boron channel-stop implant. The field-oxide is now grown on both structures. In the retrograde-well structure, the *n*-well implant step is carried out after the field oxidation, followed by a short (~30 min) anneal at 1000°C. In both structures, the boron channel stop dopant is overcompensated in the *n*-well regions by the phosphorus implant.

Because the retrograde well is formed by high-energy ions penetrating the field oxide in the field regions the well depth there is only half of the well depth in the active regions. The shallower well depth in the field regions of the retrograde-well structure (compared to the well field-region depth in the conventional-well structure) is advantageous. That is, it creates a shallower drain for the n^+-to-*n*-well parasitic FET,

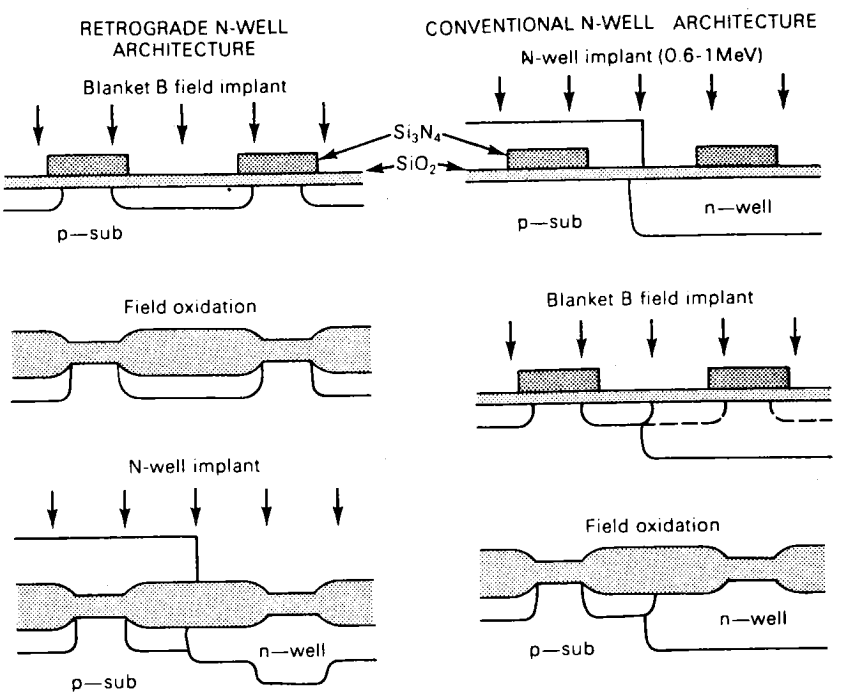


Fig. 8-15 Process architectures for retrograde and conventional *n*-well CMOS structures.[30]

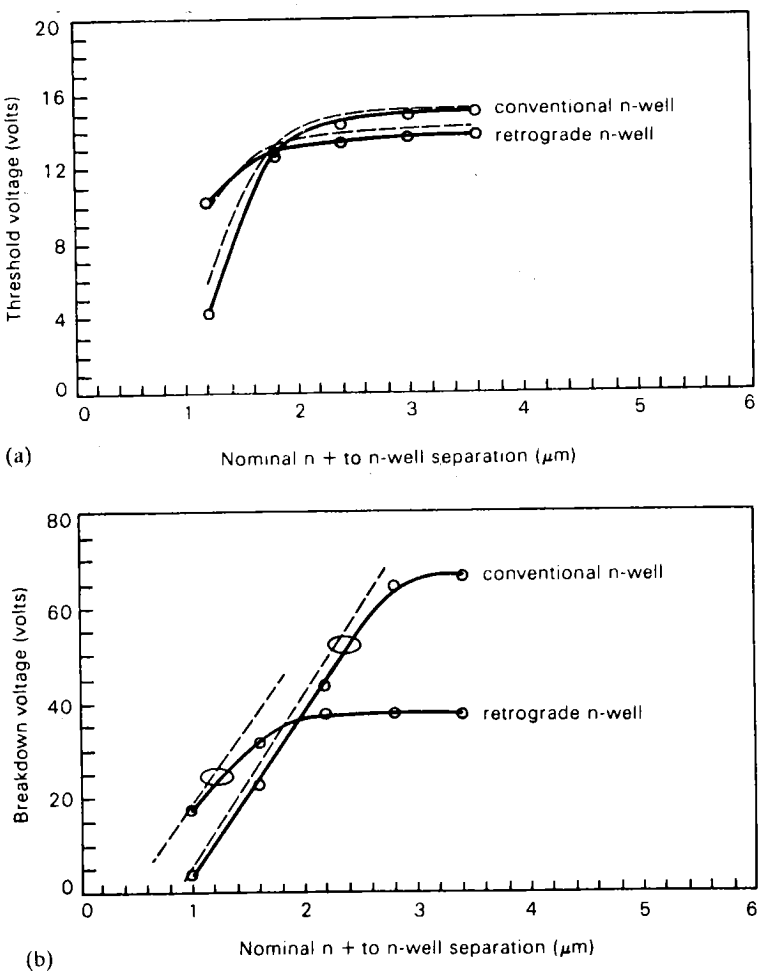


Fig. 8-16 Comparison of retrograde and conventional well structures for n^+ to n-well parasitic FETs: (a) field threshold voltage; and (b) n^+-to-n-well breakdown voltage. Solid lines: experimental data, dashed lines: modeled results, breakdown voltage is modeled by punchthrough only.[37] (© 1987 IEEE).

thereby improving isolation. At the same time, the depth of the well in the active region is greater, so that the performance of the devices in the well are not significantly degraded because the n-well doping concentration near the surface remains sufficiently low. In addition, in the active regions the peak of implanted profile is near the bottom of the well, which is beneficial for vertical isolation. Meanwhile, the implant profile peaks near the Si surface in the field areas, which is again better for producing a channel stop and for preventing p^+-to-p-substrate lateral punchthrough.

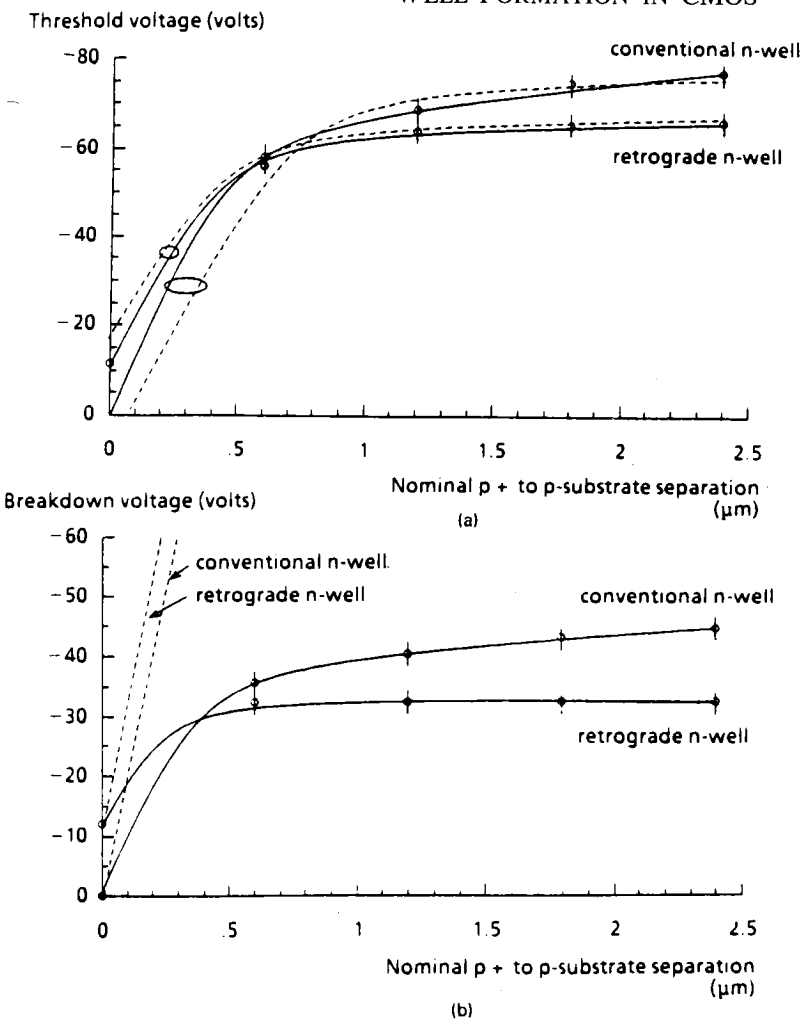


Fig. 8-17 Comparison of retrograde and conventional well structures for p^+ to p-well parasitic FETs: (a) field threshold voltage; and (b) p^+-to-p substrate breakdown voltage. Solid lines: experimental data, dashed lines: modeled results, breakdown voltage is modeled by punchthrough only.[37] (© 1987 IEEE).

The measured and simulated V_{TFiso}s and breakdown voltages (all defined at a drain current in the parasitic FET test structure of 10 pA) for n^+-to-n-well parasitic transistors in both the conventional and retrograde aluminum-gate CMOS test structures are shown in Figs. 8-16a and 8-16b. At large n^+-to-n-well separations, avalanche breakdown at the well-substrate junction occurs before lateral punchthrough occurs. The lower value of avalanche breakdown in the retrograde structures is attributed to higher well doping. As described earlier, shallower retrograde wells should exhibit less propensity to degradation

caused by short-channel effects (e.g., to threshold lowering and DIBL). These predictions are confirmed from the experimental V_{TFiso} and punchthrough data.

Figure 8-17a gives the V_{TFiso} values for lateral p^+-to-p-substrate parasitic devices. The absolute values are much larger than those for the n^+-to-n-well devices. These higher values are due in part to the greater field doping required in the n-well to guaranty sufficient compensation of the boron. The larger values are also partly caused by the fixed charge at the Si-SiO_2 interface, which is generally positive and thus increases the magnitude of the threshold voltage of the p-channel transistors. It is also evident that short-channel effects are much less severe in these p-channel parasitic devices, especially for the retrograde structures; the thresholds do not decrease until the channel length approaches 0.5 μm. This effect is caused by the higher channel (n-well) doping in these parasitic transistors, and to the drain (substrate) doping being lower than the channel (n-well) doping.

Figure 8-17b depicts the breakdown voltages exhibited by the same p^+-to-p-substrate devices, and comparison with Figs. 8-17a and b again shows the superior performance of these structures. In these types of parasitic FETs, the p-substrate, which is considered to be the drain, is not only deep but wraps around under the source (see Fig. 6-45). Thus, both lateral and vertical leakage paths may possibly exist between the p^+ source and the p-substrate. The limiting values of breakdown voltage for retrograde n-well structures is set by well to substrate avalanche breakdown, as is the case of the n^+-to-n-well transistors. In conventional wells, however, the limit is set by p^+-to-p-substrate vertical punchthrough breakdown. Thus, the maximum breakdown voltage shown on Fig. 8-17b for these devices is not the same as that shown in Fig. 8-16b, even though the drain-channel junction is in fact the same in both cases.

The application of retrograde wells for both the n- and p-wells to a twin-tub CMOS process has also been reported. The approach, as described in the previous section, was called *quadruple-well CMOS* to indicate that for each well type (i.e., n- or p-), two wells were needed, one in the active area, and another, shallower one in the field region (see Fig. 8-18). The two shallow wells actually function as self-aligned channel stops, in which the surface concentration of the well dopants is tightly restricted on each side of the well boundary. That is, as shown in Fig. 8-19 the reduction of the net impurity concentration is restricted to a small distance from the boundary of the well edge. Specifically, the surface doping concentration near the well boundary is decreased significantly only within a few tenths of a micron from the boundary, instead of a few microns, as occurs in a conventional twin-well structure (see Fig. 6-47). Experimental data shows that at a 4 μm n^+-to-p^+ separation, subthreshold leakage in the isolation region is <1 pA for gate voltages between -15V and +20V (Fig. 8-20). Simulation data indicate that it is possible to obtain adequate isolation with a 2 μm n^+-to-p^+ separation.

A CMOS process based on high energy implantations without epi-layers or trenches that allows n^+-to-p^+ spacings as small as 1.5 μm has been reported. To realize such small isolation spacing, the n-well was formed with 700 keV, $1x10^{13}$ cm^{-2} phosphorus and 1.5 MeV, $3x10^{13}$ cm^{-2} boron implantations. A unique feature of the well structure in this report is that these same two implantations simultaneously form not only the two retrograde wells and the n-well and p-well channel stops, but the boron implant also

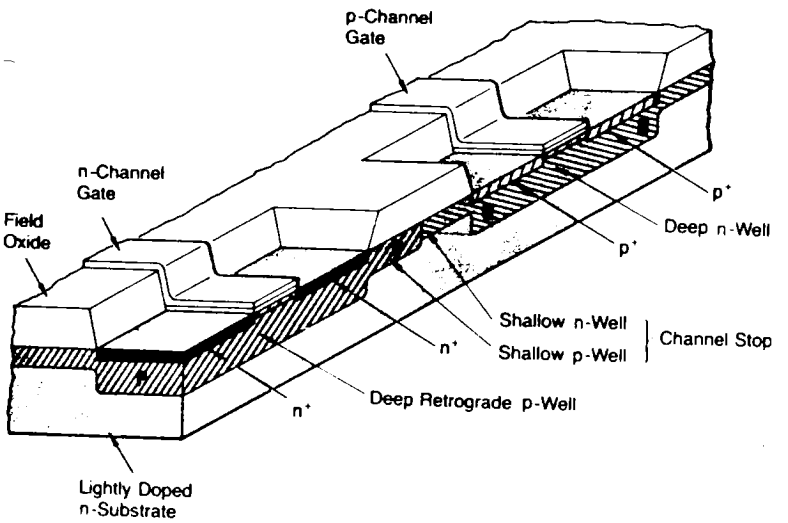


Fig. 8-18 Cross section of a quadruple-well CMOS device.[36] Reprinted with permission of VLSI Design.

formed a deep, buried p^+ layer under the phosphorus n-well profile. The p^+ layer eliminates the need for substrates with an epi layer, since the low resistivity substrate required to suppress latchup is created by the high energy implantation. It was also reported that the high energy implantation did not induce any excess leakage or measurably degrade MOS transistor characteristics.

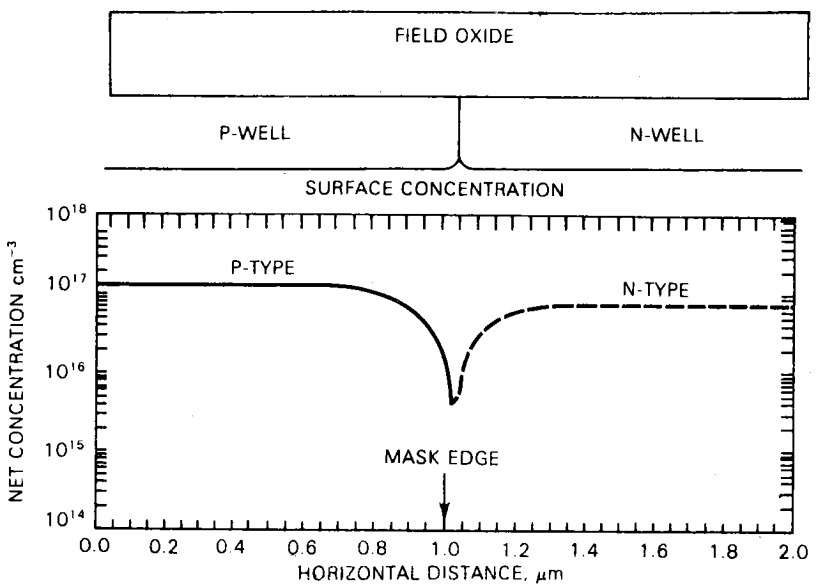


Fig. 8-19 Surface impurity concentration at the boundary of the p- and n-retrograde wells in quadruple-well CMOS.[29] (© 1986 IEEE).

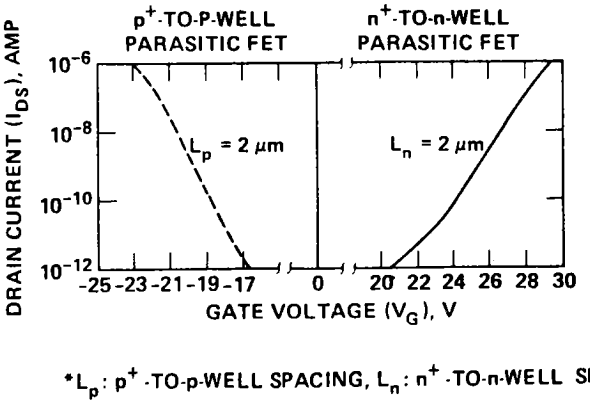


Fig. 8-20 Subthreshold characteristics of symmetric parasitic FETs in quadruple-well CMOS.[29] (© 1986 IEEE).

8.4 FUTURE TRENDS IN WELL FORMATION

It seems that high-energy ion implantation will find more application for generating profiled wells. Profiled wells are achieved by multiple implants at varying energies. They offer the ability to independently control several MOSFET characteristics, including threshold voltage, hot-carrier generation, punchthrough voltage, junction capacitance, latchup susceptibility, soft-error immunity, and the body-effect coefficient. The optimization of well profiles for improving device performance has been termed "well engineering."[31]

Unlike a conventional well, which has an almost flat impurity distribution, a profiled well can have an impurity profile tailored to independently optimize a variety of device characteristics. Figure 8-21 shows the impurity profile of a multiply-implanted well and the relation of the device characteristics to the profile.[32]

In profile-implanted wells formed with high-energy implants, the wells are implanted after LOCOS isolation. In the *n*-well, multiple high-energy implants of phosphorus ions form a profiled well for active as well as isolation regions (with a thick photoresist serving as the *p*-well mask). Subsequently, low-energy implants are performed to adjust the PMOS V_T. A profiled *p*-well is next produced in the same way.

The deepest implant of the well in this example is fabricated with a 1 MeV implant. These implanted dopants form a heavily doped buried layer (10^{18}-10^{19} cm^{-3}) at a depth of 2-4 μm that serves to reduce latchup susceptibility. Such a p^+ buried layer is more effective than an epitaxial layer in reducing substrate resistance and suppressing latchup, thus eliminating the need for an epi layer. The ability to use non-epi substrates when wafer sizes reach 150 mm and 200 mm represents a major cost savings, as the cost of 200 mm epitaxial wafers are ~$100 US higher than bulk-CZ wafers. The replacement of epitaxial wafers with MeV implantation processing also alleviates problems with autodoping control and backside sealing.[35]

The use of MeV implantation avoids the long (3-21 hour), high-temperature (1150-1200°C) well-formation drive-in step of the conventional process. This minimizes

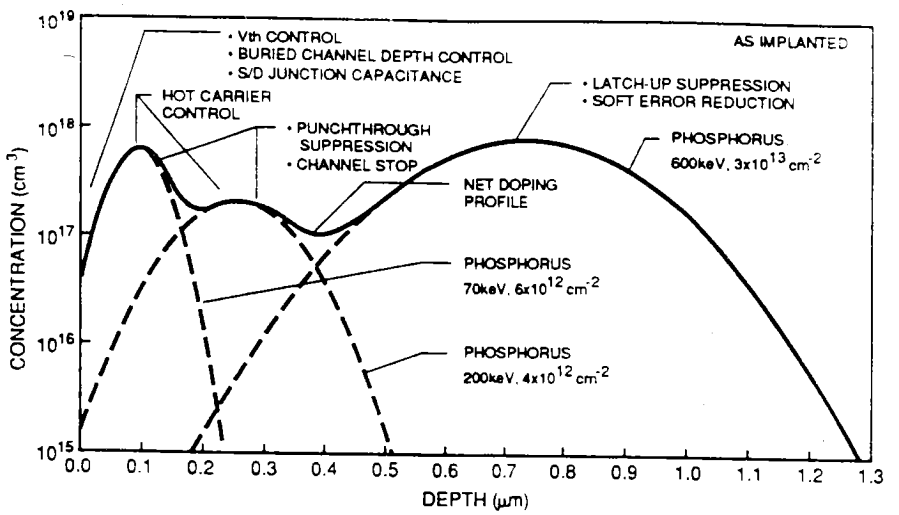


Fig. 8-21 The relation between device characteristics and the impurity concentration at various depths in multiple implanted well.[32] Reprinted with permission of Solid State Technology.

thermally induced stress in large wafers that can cause wafer warpage, bow, and flatness degradation, as well as offering simplified processing with shorter fabrication times. Process simplification also results from the fact that the retrograde-well and field-isolation regions are formed by the same implant (after field oxidation). As noted in section 8.3.1, implantation through field oxide can also improve active device isolation characteristics. Employing MeV implantation to dope the bottom of the well more heavily leaves more room to tailor the doping profile at the top of the well in the channel area with lower energy implants.

Concern has existed that the high-energy implants may cause damage that will degrade device and circuit performance. Increased junction leakage is observed if the dose or energy are too low (i.e., $<3\text{x}10^{14}$ cm^{-2} and ~1-1.5 MeV, respectively). Recent studies, however, indicate that if the dose and energy exceed $3\text{x}10^{14}$ cm^{-2} and 1.5 MeV, respectively, leakage due to implant damage is dramatically reduced with both boron and phosphorus implantation. In fact, when an optimum combination of dose, energy, and annealing process is used, *less* leakage is observed than in control devices having no buried layer. This behavior is believed to be caused by gettering of electrically active damage or impurity defects away from the depletion regions of active transistors by secondary defects near the concentration peak of the buried layer. The effect is called "self-gettering." For all of the above reasons it appears that high-energy implantation will become more widely used to form wells in device generations at and below 0.5 μm.

8.5 AFTER-GATE IMPLANTATION (AGI)

An alternative technique which exploits high-energy implantation to reduce the complexity of CMOS process called *after-gate implantation* (AGI) has been suggested. That is, by performing all implantations after the gate electrodes are patterned (wells, channel-stop doping, and source/drain), the number of mask steps can be reduced.

Figure 8-22 illustrates this approach, in which only seven masks are required to form a single p-well CMOS structure.[33] The LOCOS-isolation field-oxide regions are formed first, then the polysilicon gates are deposited and patterned. Using a mask that obscures the PMOS regions, the NMOS p-well implant, multiple-channel implants, and the n^+ source/drain implants are carried out. Then a mask is used to cover the NMOS areas, and the PMOS-channel and p^+ source/drain implants are performed. Since the well and channel implantations must penetrate the field oxide and the polysilicon gate, they must be carried out at higher energies. The p-well implant energy is 600 keV (at a dose of $1x10^{13}$ cm^{-2}), and the NMOS devices receive low-dose channel implants of 135 keV, 175 keV, and 215 keV. The channel of the PMOS devices receives two low-dose, higher-energy phosphorus implants (at 220 keV and 440 keV). Such multiple-channel implants are used in both NMOS and PMOS devices to avoid unacceptable V_T variations caused by polysilicon gate thickness variations encountered in processing (and to improve the punchthrough resistance of the device structures). Studies also indicate

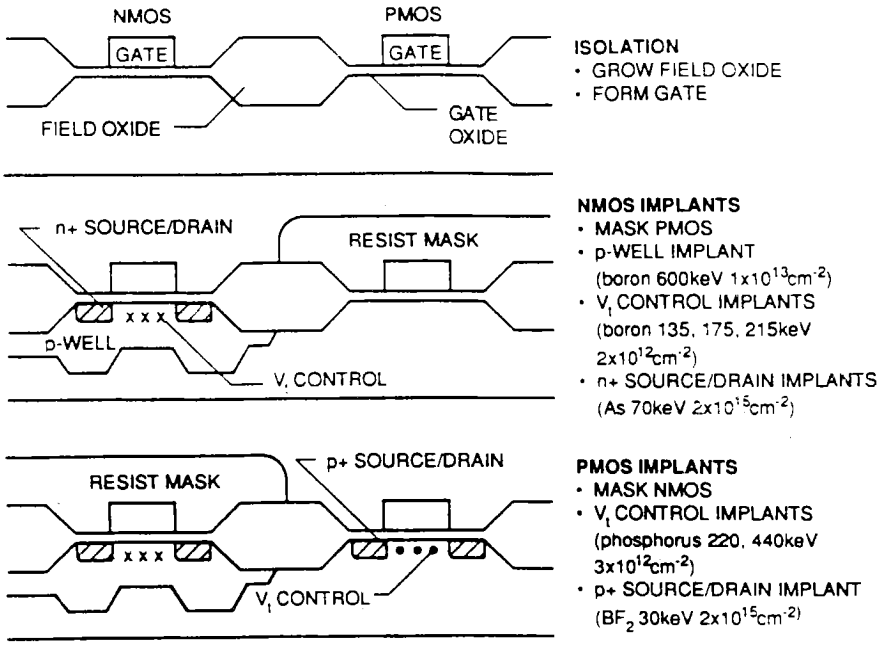


Fig. 8-22 After-Gate Implant (AGI) process sequence.[33] (© 1986 IEEE).

that the high energy AGI does not degrade the gate oxide integrity[32] nor the MOS transistor characteristics.[34]

The AGI process thus offers the benefits of implantation processing freedom and flexibility, as well as reduced CMOS process complexity. The low doses required for the well and channel implantations can be performed with a medium-current implanter which may also have a higher energy implantation capability.

REFERENCES

1. J.Y. Chen, "CMOS - The Emerging Technology," *IEEE Circuits and Devices Magazine,* March 1986, p. 16.
2. J.Y. Chen, *CMOS Devices and Technology for VLSI*, Prentice-Hall, Englewood Cliffs, N.J., 1990.
3. A.G. Lewis et al., *IEEE Trans. Electron Dev.*, **ED-34**, 2156, (1987).
4. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, **ED-31**, p. 205 (1984).
5. L.C. Parrillo "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.
6. W. Wijaranakula et al., *J. Electrochem. Soc.,* December 1988, p. 3113.
7. H. Tsuya et al., *Japanese J. Appl. Phys.*, **22,** L16 (1983).
8. A. G. Lewis et al., "Vertical Isolation in Shallow *n*-Well CMOS Circuits," *IEEE Electron Dev. Lett.*, EDL-8, p. 107, 1987.
9. K. Tanno, F. Shimura, and T. Kawamura, *J. Electrochem. Soc.,* **128**, p. 395 (1981).
10. J. Manoliu and J.O. Borland, "A Submicron Buried Layer Twin-Well for High Density CMOS," *Tech. Dig. IEDM,* 1987, p. 20.
11. R. de Werdt et al., *Tech. Dig. IEDM,* 1987, p. 532.
12. L.C. Parrillo et al., "Twin-Tub CMOS-II," *Tech. Dig. IEDM,* 1982, p. 706.
13. R.A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.
14. M.-L. Chen et al. *Tech. Dig. IEDM* 1986, p. 256.
15. S.J. Hillenius and L.C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.
16. R.A. Chapman et al., "High Performance Sub-Half Micron CMOS Using Rapid Thermal Processing," *Tech. Dig. IEDM* 1991, p. 101.
17. Ch. Zeller, C. Mazure, & M. Kerber, "Field-Implant-Free Isolation by Double-Well Split Drive-In," *IEEE Electron. Dev. Letts.*, p. 215, May 1990.
18. R. Martin and J.Y. Chen, "Optimized Retrograde *n*-well for 1 μm CMOS, *Proc. Custom Integ. Circuits Conf.,* 1985, p. 199.
19. R.D. Rung, C.J. Dell'Oca, and L.G. Walker, "A Retrograde *p*-well for High Density CMOS," *IEEE Trans. Electron Dev.*, ED-28, p. 1115, 1981.
20. D.M. Brown et al., "Trends in Advanced CMOS Process Technology," *Proceedings of the IEEE,* December 1986, p. 1646.
21. S.R. Coombs, "Scalable Retrograde *p*-well CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-28**, p. 346, 1981.
22. Y. Taur et al., *IEEE J. Solid-State Circuits,* **SC-20,** p. 123, 1985.

23. G.J. Hu et al., "Design and Fabrication of *p*-channel FET for 1-μm CMOS technology," *IEDM Tech. Dig.* 1982, p. 710.

24. A. Stolmeijer, "A twin-well CMOS process employing high-energy ion implantation," *IEEE Trans. Electron Dev.*, **ED-33**, April, 1986, p. 450.

25. A. Shida et al., 3.3V BiNMOS Technology Using NPN Transistors without Buried Layers," *Tech. Dig. IEDM,* 1991, p. 93.

26. Y. Mochizuki, "Special Report: High Energy Ion Implantation Technology," Nikkei Microelectronics, (Dec. 1991), p. 92.

27. M. Moslehi and K. Saraswat, "Thermal Nitridation of Si and SiO_2 for VLSI," *IEEE Trans Electron Dev.* **ED-32**, p. 116, January 1985.

28. P. Fahey, R.W. Dutton, and M. Moslehi, "Effects of Thermal Nitridation Processes on Boron and Phosphorus Diffusion in <100> Silicon," *Appl. Phys. Letts.*, **43**(7), p. 683 July, 1983.

29. J.Y. Chen, "Quadruple-well CMOS for VLSI Technology," *IEEE Trans. on Electron Dev.*, **ED-31**, p. 910, 1984.

30. A.G. Lewis, R.A. Martin, and J.Y. Chen, "Retrograde and conventional *n*-well CMOS technologies: A Comparison," Internal Report, EIL-87-17, Xerox Palo Alto Research Center, Oct. 1986.

31. K. Tsukamoto et al., *Nucl. Instrumentation & Methods,* B59/60, p. 584 (1991).

32. K. Tsukamoto et al., "High Energy Ion Implantation for ULSI: Well Engineering and Gettering," *Solid State Technology* P. 49, June 1992.

33. H. Mikoshiba et al., "A Novel CMOS Process Utilizing After-Gate Implantation Process," *Dig. of Tech. Papers, 1986 Symp. VLSI Tech.*, p. 41.

34. J. Pfiester et al., "A Self-aligned Elevated Source/Drain MOSFET," *IEEE Trans. Electron Dev.*, September 1990, p. 365.

35. J.O. Borland and R. Koelsch, "MeV implantation technology: Next generation manufacturing with current generation equipment," *Solid-State Technology*, December 1993, p. 28.

36. J.Y. Chen, *VLSI Design* July 1984, p. 78.

37. A.G. Lewis et al., "Device Isolation in High Density LOCOS-Isolated CMOS," *IEEE Trans. Electron Devices,* **ED-34**, p. 1337, June 1987.

PROBLEMS

1. An *n*-well n^+ polysilicon-gate CMOS process starts with a substrate doping of $3x10^{15}/cm^3$. The well doping near the surface of is approximately constant to a level of $3x10^{16}/cm^3$. The gate oxide thicknesses are both 40 nm, and $Q_{tot} = 3x10^{10}/cm^2$.

(a) Calculate the threshold voltages of the *n*- and *p*-channel transistors.

(b) Calculate the boron doses needed to shift the NMOS threshold to +1V and the PMOS threshold to -1V. Assume that the threshold shifts are achieved through shallow ion implantations.

2. In a *p*-well CMOS technology, the *n*-substrate has a doping concentration of $1x10^{15}/cm^3$, and a *p*-well concentration neat the surface of $5x10^{16}/cm^3$. Calculate the source/drain junction capacitances (cm^{-2}) and body effect coefficients for both *n*- and *p*-channel MOSFETs assuming the gate oxide is 20 nm thick in both devices.

3. Describe the split well-drive process, and the advantages it offers compared to the conventional well-drive process.

4. Comparing retrograde well to conventional well, what is the most important difference in process technology and finished device structure? How does it affect device isolation?

5. In a (p^+-diffusion)-n-well-(p-on-p+ substrate)structure, the n-well is uniformly doped at $1x10^{16}$ cm^{-3} and is 1 μm deep, the p-epi thickness is 4 μm and is uniformly doped at $1x10^{15}/cm^3$, and the p^+ diffusion is 0.4 μm deep. If a substrate bias generator is needed and the p^+ diffusion and n-well are both biased at 5V, what is the maximum (in absolute value) substrate voltage that can be applied before punchthrough occurs?

CHAPTER 9

HOT-CARRIER RESISTANT PROCESSING AND DEVICE STRUCTURES

One of the more serious problems posed by the continued shrinking of MOSFETs into the submicron regime is the *hot-carrier effect.* If device dimensions are reduced and the supply voltage remains constant (or is not reduced as rapidly as the structures are scaled), the lateral electric field in the channel increases. This causes the inversion-layer charges to be accelerated (heated) to an extent that they can cause a number of harmful device phenomena (hot-carrier effects). Figure 9-1 depicts hot-carrier generation and the current components which result from such generation. The most important hot-carrier effect is the damage inflicted to the gate oxide and/or the Si-SiO_2 interface. This leads to a time-dependent degradation of various MOSFET characteristics, including threshold voltage (V_T), linear region transconductance (g_m), subthreshold slope (S_t), and

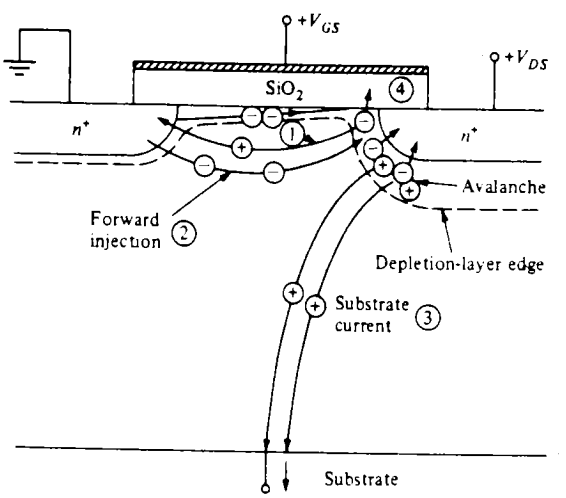


Fig. 9-1 Hot-carrier generation and current components. 1 Holes reaching the source. 2 Electron injection from the source. 3 Substrate hole current. 4 Electron injection into the oxide.

saturation current (I_{Dsat}). The lifetimes of conventional MOSFET structures when subject to such degradation may be reduced below acceptable limits (i.e., to less that 10 years of operation) in NMOS devices with channel lengths smaller than 1.5 μm, and in PMOS devices with submicron channel lengths. As a result, it has been necessary to develop techniques to combat this problem. The purpose of this chapter is to discuss hot-carrier effects in MOSFETs and the approaches developed to overcome them.

Techniques for Reducing Hot-Carrier Degradation. The degradation of device performance arising from hot-carrier effects can be suppressed by various methods. Most involve modifying the conventional MOSFET structure or strengthening the gate oxide and the Si-SiO_2 interface against hot-carrier damage. As an introduction to the material of this chapter, a list of the general categories of approaches to suppress hot-carrier degradation is presented here:

1. The common driving force of all hot-carrier phenomena is the channel electric field $\mathcal{E}_y$. Specifically, the greatest control over hot-electron effects is exerted by the *maximum channel electric field* $\mathcal{E}_{ymax}$, which occurs at the drain end of the channel. Thus, a key underlying goal in reducing hot-carrier degradation is to reduce the magnitude of $\mathcal{E}_{ymax}$. Toward this end, the factors that impact $\mathcal{E}_{ymax}$ in a MOSFET must be understood, and useful tools for gaining such understanding are models for $\mathcal{E}_{ymax}$. We thus begin the chapter by reviewing a simple model for $\mathcal{E}_{ymax}$ in non-graded-drain MOSFETs, and later we extend it to include graded-drain (i.e., LDD) MOSFETs.

2. One obvious way to decrease $\mathcal{E}_{ymax}$ is to reduce the voltages applied to the device (e.g., by lowering the power-supply voltage from 5V to 3.3V). The decision to implement such power supply reductions, however, is not in the hands of the device designer or process engineer.

3. The structure of the conventional MOSFET can be modified such that it is less sensitive to hot-carrier degradation. Such modifications attempt to reduce $\mathcal{E}_{ymax}$ as well as to separate the maximum current path in the channel from the maximum electric field location. This is accomplished by developing special drain field-reducing structures designed to accomplish these objectives. A discussion of the evolution of such structures constitutes the bulk of this chapter.

4. The resistance of the gate oxide and the Si-SiO_2 interface to hot-carrier degradation can be increased, thereby allowing a larger $\mathcal{E}_{ymax}$ to be applied. As an example, the density of trapping sites in the gate oxide can be reduced through the use of special processing techniques (which are described in section 9.14).

9.1 MODEL FOR THE MAXIMUM LATERAL ELECTRIC FIELD $\mathcal{E}_{ymax}$ IN NON-GRADED-DRAIN MOSFETS

The maximum lateral electric field in a MOSFET, $\mathcal{E}_{ymax}$, occurs near the drain. It has been determined that no significant hot-carrier effects occur at values of $\mathcal{E}_{ymax}$ smaller than $4x10^4$ V/cm (i.e., the field value at which electrons are velocity saturated). Hence, when hot-carrier effects are being considered, it is only necessary to be concerned with the MOSFET being operated in saturation, and then only with the lateral electric fields in the velocity-saturated region (VSR) of the MOSFET. Note that the VSR was introduced in section 5.4.3 of chap. 5.

A rigorous calculation of the field near the drain is a complex procedure, requiring a numerical analysis of the two-dimensional Poisson equation. An example of the results of one such analysis which shows $\mathcal{E}_y$ within an NMOSFET as a function of channel length is shown in Fig. 9-2. Nevertheless, in chap. 5 we showed how a reasonably accurate analytical model of $\mathcal{E}_{ymax}$ in the VSR of non-LDD MOSFETs could be derived using the pseudo-2D approach to solving the Poisson equation in the MOSFET channel. This model gives good agreement with the values of $\mathcal{E}_{ymax}$ obtained from numerical analysis. According to this model the value of $\mathcal{E}_{ymax}$ is given by Eq. 5-70, which for the case when $(V_{DS} - V_{DSsat})/l >> \mathcal{E}_{sat}$ becomes [1,2,8]

$$\mathcal{E}_{ymax} = (V_{DS} - V_{DSsat}) / l \qquad (9 - 1)$$

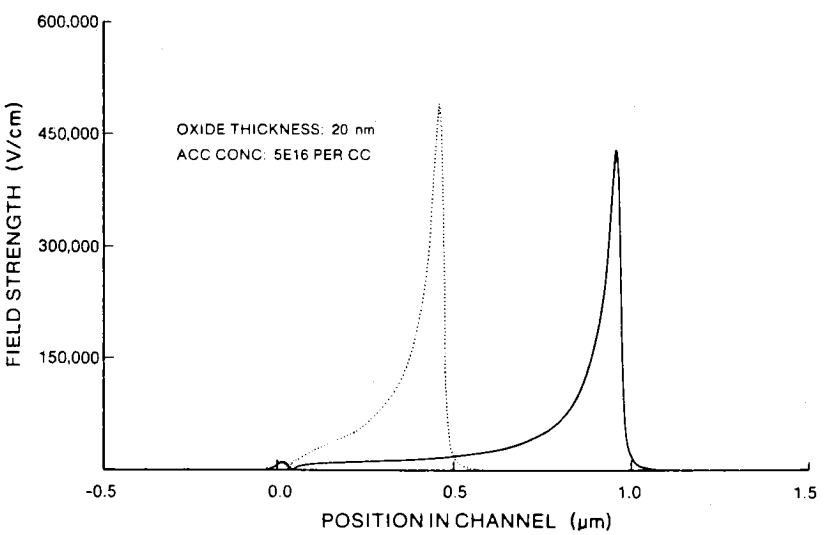


Fig. 9-2 $\mathcal{E}_y$ within an NMOSFET as a function of channel length for two devices with channel lengths of 1.0 μm and 0.5 μm, and the same value of t_{OX} = 20 nm.[146]

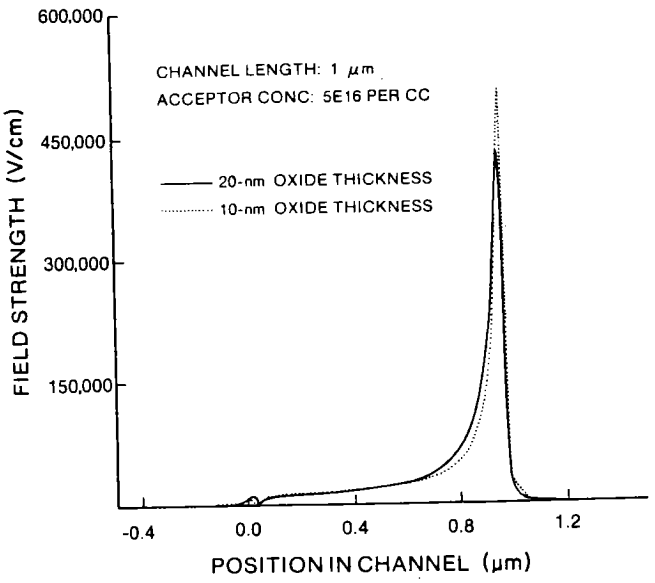


Fig. 9-3 $\mathcal{E}_y$ within an NMOSFET as a function of channel length and gate oxide thickness for two devices with a channel length (1.0 μm) and two values of t_{ox}, 10 and 20 nm.[146]

where[9]

$$l \approx 0.22\, t_{ox}^{1/3}\, r_j'^{1/3} \qquad \text{for } t_{ox} \geq 15 \text{ nm} \qquad (9\text{-}2a)$$

and

$$l \approx 1.7\text{x}10^{-2}\, t_{ox}^{1/8}\, r_j'^{1/3} L^{1/5} \qquad \text{for } t_{ox} < 15 \text{ nm, and } L < 0.5\ \mu\text{m} \qquad (9\text{-}2b)$$

and t_{ox} is the gate oxide thickness and r_j' approximately corresponds to the source/drain junction depth. As we calculated in section 5.4.3, when t_{ox} is ~15 nm and r_j ~0.35 μm, l ~ 0.15 μm. Equations 9-2a and 9-2b imply that l is independent of L until L becomes small enough that it approaches l. That is, for larger values of L than this, the channel length still has an influence on $\mathcal{E}_{ymax}$, but only through the V_{DSsat} term of Eq. 9-1, and this dependence is weaker than a linear one. A model for calculating V_{DSsat} was presented in chap. 5 (Eq. 5-37), which we also repeat here for convenience:[3]

$$V_{DSsat} = \frac{\mathcal{E}_{sat} L\,[V_{GS} - V_T]}{(V_{GS} - V_T) + \mathcal{E}_{sat} L} \qquad (9\text{-}3)$$

An example of the dependence of $\mathcal{E}_y$ on t_{ox} as calculated using numerical analysis for an NMOSFET with L = 1 μm is shown in Fig. 9-3. Here we see that the device with a thicker t_{ox} has a smaller $\mathcal{E}_{ymax}$, as is also predicted by the model of Eqs. 9-1 and 9-2.

9.2 SUBSTRATE CURRENTS DUE TO HOT CARRIERS

While the time-dependent degradation of MOSFET device parameters is the hot-carrier effect that receives the most attention, another important one involves the substrate current I_{sub} generated by such hot carriers. Since I_{sub} can be correlated to long-term device degradation, it can be used to predict the device lifetime. However, I_{sub} can also adversely impact MOSFET operation in other ways as well. Hence, it is useful to examine the phenomenon of I_{sub} prior to the discussion on how to characterize hot-carrier degradation.

If electrons in the channel of a MOSFET acquire more than about 1.5 eV of energy, impact ionization can result upon their collision with the lattice. Electron-hole pairs are generated from such collisions, with the total number being exponentially dependent on the reciprocal of the electric field, $-1/\mathcal{E}_{ymax}$. The electrons produced in this manner are either attracted to the drain (adding to the drain current) or if possessing sufficient energy, they may be injected into the oxide. The generated holes, on the other hand, enter the substrate and constitute a parasitic substrate current I_{sub} (Process 3 in Fig. 9-1). Such substrate current differs from the current which arises in *pn*-junction avalanche breakdown in that its magnitude is much smaller than the drain current. I_{sub} is thus classified as a *low-level-avalanche multiplication effect*. Nevertheless, I_{sub} can itself produce several problems, and thus must be kept below some maximum tolerable value. The problems which can be directly caused by I_{sub} include the following:

1. If some of the holes that should constitute I_{sub} are instead collected by the source (i.e., instead of by the body contact), and this collected hole current causes a voltage drop in the substrate material on the order of 0.6V, the substrate-source *pn* junction will begin to conduct significantly. Electrons will then be injected from the source to the substrate, just like electrons injected from emitter to base of an *npn* transistor (the forward injection shown in Fig. 9-1). These extra injected electrons can, in turn, gain sufficient energy as they travel toward the drain to cause additional impact ionization and create new electron-hole pairs. A positive-feedback mechanism thus exists, one that can sustain itself if the drain voltage exceeds a certain value. This is observed externally as a form of breakdown, referred to as a *snapback breakdown*. A particularly clear explanation of this effect, including the reason for the negative-resistance, or "snapback," portion of the curves, is given in reference 4.

2. Excessive I_{sub} may induce latchup in CMOS circuits.

3. As some of the holes are accelerated during their passage out of the drain-substrate depletion region, they may acquire enough energy to cause secondary impact ionization far from the drain region. Some the electrons generated in this manner may escape the drain field and, instead of being collected by the drain, may then travel to other nodes on the chip to be collected. This may lead to a reduction of the storage time of dynamic circuit nodes in DRAMs (i.e., manifested as a degradation of the refresh time).[5] This excess electron current is

reported to be around 10^{-4} times smaller than the substrate ionization impact current itself.

4. If a substrate bias-generator circuit is included on-chip, its output becomes less negative as I_{sub} is increased.

As noted earlier, I_{sub} can also be used as a monitor to correlate device degradation with predicted device lifetime. The reason that I_{sub} is chosen for this purpose is that both device degradation and I_{sub} are driven by a common driving force, namely the lateral electric field in the channel of the MOSFET. The magnitude of I_{sub} depends exponentially on the maximum value of this field $\mathcal{E}_{ymax}$.

An example of how I_{sub} is experimentally observed to depend on decreasing channel length is seen in Fig. 9-4.[6] This figure shows the maximum I_{sub} generated at a voltage of 5V versus the effective channel length for MOS transistors processed with the same technology (i.e., t_{ox} and r_j have identical values in all of the devices). Since I_{sub} also increases with $\mathcal{E}_{ymax}$, according to the model of Eqs. 9-1 and 9-2, the magnitude of I_{sub} would grow even more rapidly with shrinking L_{eff} if the oxide thickness and junction depth were scaled together with the channel length.

Two common methods are used to display the empirical data when I_{sub} in a MOSFET is being characterized through measurement methods. In the first (shown in Fig. 9-5), I_{sub} is plotted against the gate voltage V_{GS}, with V_{DS} as the parameter. The initial increase in I_{sub} with increasing V_{GS} observed in such plots is due to the rise in the population of electrons in the channel as V_{GS} is first increased. (We also recall that I_D increases as V_{GS} is made larger). Since this raises the number of hot electrons that can be heated, it implies that there are a larger number of electrons capable of participating in impact ionization. Eventually, however, when V_{GS} approaches V_{DS}, the lateral electric field begins to decrease. From that point on, I_{sub} decreases even as V_{GS}

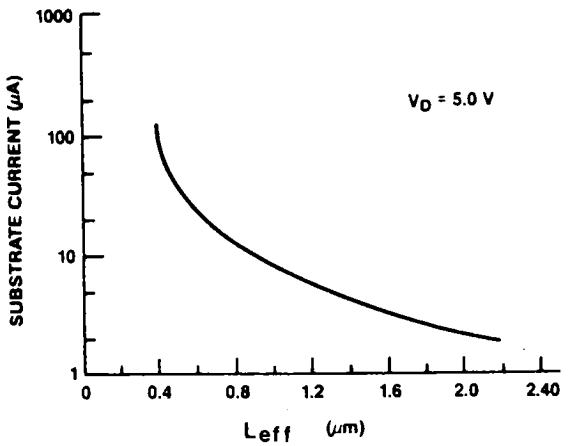


Fig. 9-4 The maximum substrate current due to impact ionization produced at a drain voltage of 5V vs. L_{eff} for MOSFETs with t_{ox} = 250Å.[6]

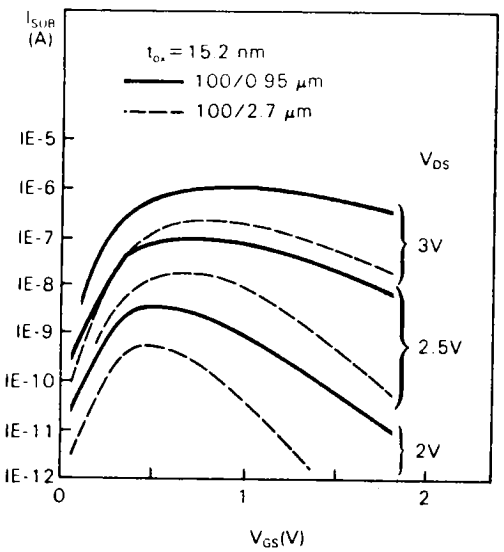


Fig. 9-5 Substrate current characterized as a function of V_{GS} with V_{DS} as a parameter.[7] (© IEEE 1984).

continues to rise. In the second approach used to display I_{sub} data, we plot (ln I_{sub}) versus $1/V_{DS}$, with the gate voltage as the parameter (as shown in Fig. 9-6). In this type of plot I_{sub} values are typically observed to follow a series of straight lines.

While both types of plots are useful for depicting the empirical behavior of I_{sub}, the resulting experimentally generated curves fail to reveal a simple relationship between I_{sub} and V_{GS}, V_{DS} and device parameters such as t_{ox}, r_j, and L. However, a rudimentary analytical model that allows I_{sub} to be calculated based on the these terminal voltages and device parameters was derived by Chan et al., as follows:[7]

$$I_{sub} = 1.2\,(V_{DS} - V_{DSsat})\,I_D \exp\,(-1.7\text{x}10^6/\mathcal{E}_{ymax}) \qquad (9-4)$$

$$= 1.2\,(V_{DS} - V_{DSsat})\,I_D \exp\,(-3.7\text{x}10^5\, t_{ox}^{1/3}\, r_j^{1/3}/\,(V_{DS} - V_{DSsat}) \qquad (9-5)$$

By plotting Eq. 9-5 (in terms of log I_{sub} versus V_{GS}, with V_{DS} as the parameter) a set of curves shown as the dashed lines in Fig. 9-7 is obtained. If the measured values of (log I_{sub}) are also plotted on the same figure (as is done in Fig. 9-7 using solid lines), good agreement between the model and the experimental data is observed. We also note that the I_{sub}-V_{GS} curves exhibit a bell-shape, with their maximum values occurring at low gate voltages. This shape results because, as noted earlier, $\mathcal{E}_{ymax}$ decreases as V_{GS} increases, but I_D increases monotonically. Thus, I_{sub} peaks where a

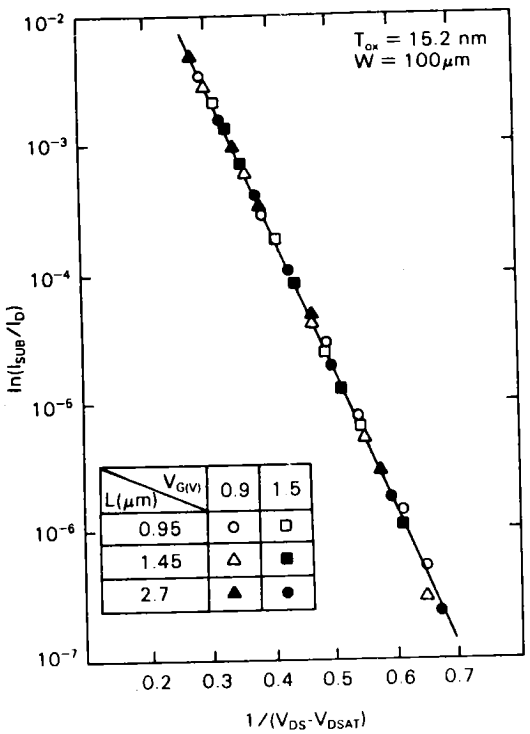


Fig. 9-6 Substrate current characterized as a function of $1/V_{DS}$ with V_{GS} as a parameter.[7] (© IEEE 1984).

significant number of carriers are available, but $\mathcal{E}_{ymax}$ has not yet been significantly reduced. These circumstances exist when $V_{GS} \approx 0.4V_{DS}$. The whole I_{sub} curves shift upward or downward depending on the drain bias.*

9.3 MODELS OF HOT-CARRIER DEGRADATION PHENOMENA

Generally speaking, a negative charge builds up in the gate oxide and/or at the Si-SiO_2 interface of MOSFETs as a result of hot-carrier injection from the silicon into the oxide. This charge contributes to the total oxide charge term Q_{tot} in the expression for the device threshold voltage (see chap. 3, Eq. 3-72). Furthermore, in NMOSFETs this charge continues to accumulate with time under conditions of operation that produce

* Note that Eq. 9-5 has been verified for MOSFETs with uniform channel doping. In order to be able to treat the complex channel and drain doping profiles found in short-channel MOSFETs (e.g., the doping profile of an LDD), some extracted parameters need to be added. Specifically, the bias and size dependencies of $\mathcal{E}_{sat}$ of Eq. 9-3 need to be taken into account.

hot-carrier stress. Due to the polarity of the trapped charge, the resulting V_T shift in n-channel devices is positive, which in turn causes both the drain current I_D and the transconductance g_m to be reduced. Reduced current, of course, will lead to reduction of circuit speed. Threshold voltage drift will cause the inverter transfer characteristics and linear circuit offset to drift. The concern is that such hot-carrier stressing may eventually cause a circuit to fail the speed specification or to experience racing conditions (or to otherwise become unacceptable).

Figure 9-8 shows typical I_D-V_{DS} characteristics of an NMOSFET before and after stressing. Notice that the decrease in I_D is larger if the source and drain are reversed after stressing. As we will describe in further detail in the next section, this indicates that the hot-carrier-induced negative charges are located near the drain (see Fig. 9-9a).

It is not clear exactly what sort of mechanism is responsible for such charge buildup, but most agree that it is caused either by interface trap generation[10] or by oxide trapped charge.[11] However, mounting evidence points to interface trap generation as the main degradation mechanism of NMOSFETs having normal quality oxides, operated at room temperature, and stressed under the V_{GS} condition that causes the largest I_{sub}.[17]

Some hypotheses blame hot electrons, others hot holes, and still others the combined effect of hot electrons and hot holes for the generation of such interface traps. However, hot holes alone have been effectively discounted as the culprits because it is observed that if I_{sub} is kept fixed (which implies that the hot-hole generation rate is not changing) and the degradation is monitored as a function of V_{GD}, degradation increases

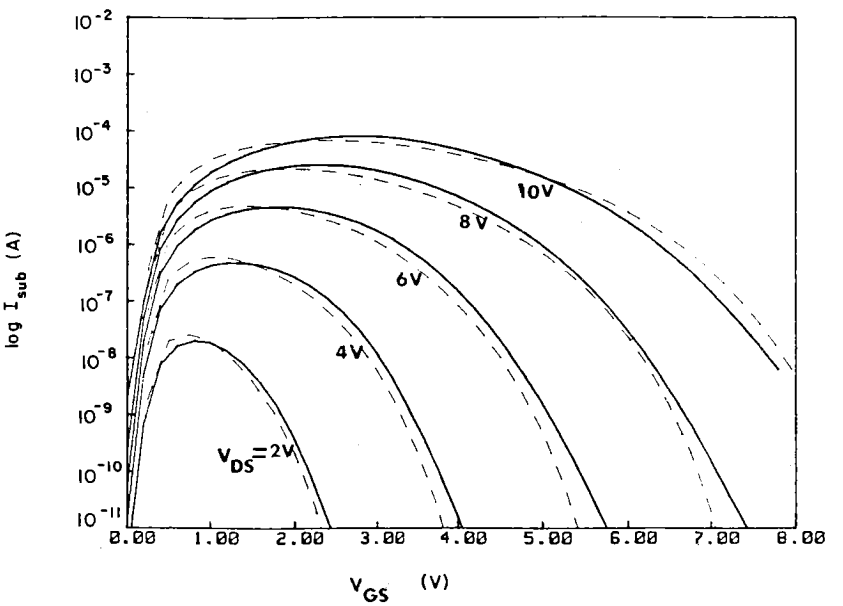


Fig. 9-7 Example of the calculated values of I_{sub} (dashed lines) using (Eq. 9-5) compared to the measured values of I_{sub}, depicted by the solid lines.[147]

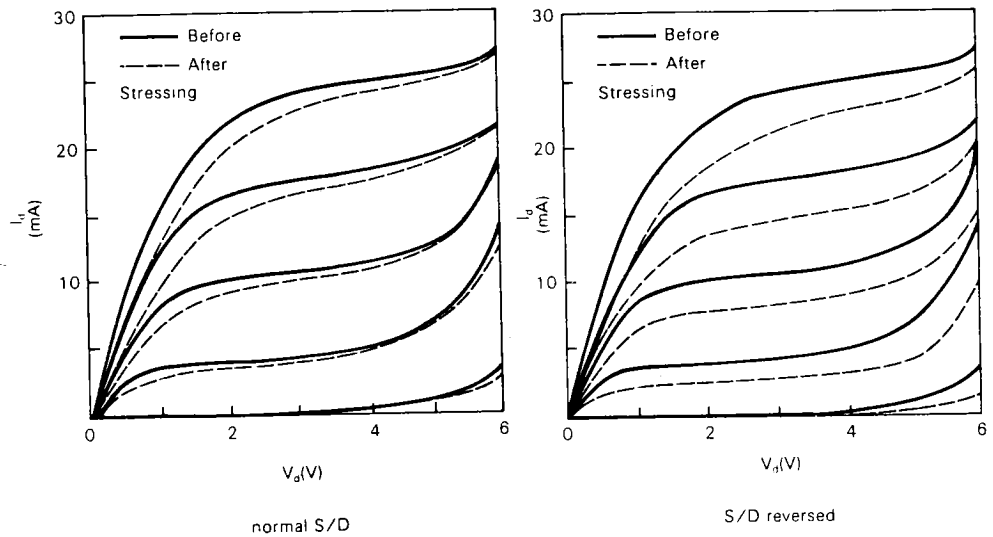


Fig. 9-8 Typical I_D-V_{DS} characteristics of an NMOSFET before and after hot-carrier stressing.[148] (© IEEE 1984).

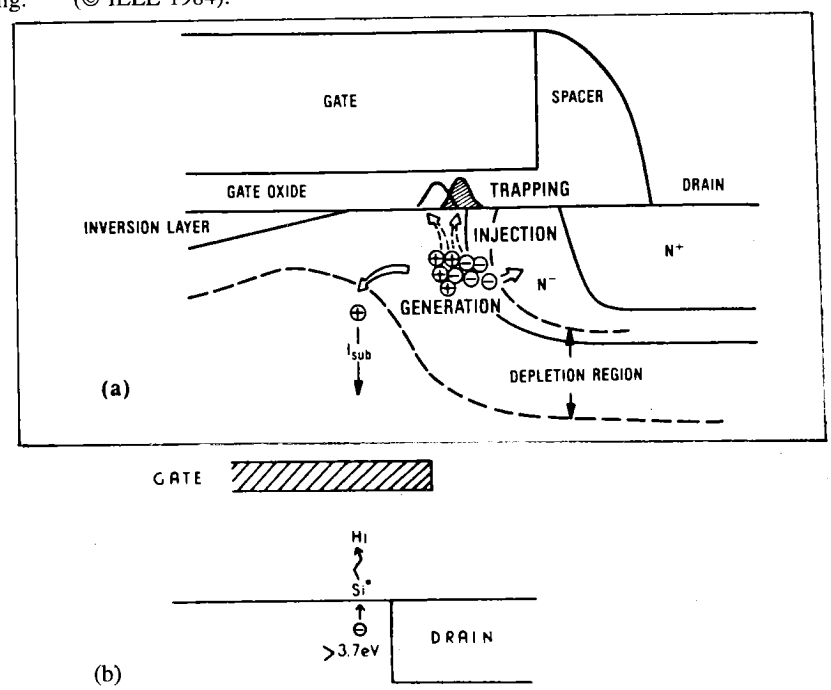


Fig. 9-9 (a) Illustration of hot-carrier effects, including hot-carrier generation, injection, and trapping.[92] (© 1988 IEEE) (b) Physical model for interface trap generation.[17] (© IEEE 1985).

monotonically with increasing V_{GD}. Increasing V_{GD} tends to attract electrons to the gate and to repel holes. This implies that hot-electron injection is at least partially responsible for the degradation.

9.3.1. Si-H Bond Breaking Model

One of the simplest microscopic degradation models is based on the assumption that damage arises from the creation of interface states. They are proposed to be generated as follows:[18] If $V_{GS}<V_{DS}$ in NMOSFETs, hot-electrons bombard the Si-SiO_2 interface against a field that tries to repel them. However, if the electrons possess sufficient energy to overcome the barrier of the Si-SiO_2 interface and the repelling field, they can break chemical bonds at the interface (Fig. 9-9b). Most likely the bonds that are broken are Si-H bonds. (The breaking of Si-H bonds by energetic electrons is a mechanism that has been confirmed by experiment.[12]) Such Si-H bonds at the interface can be formed during process steps such as BPSG reflow in steam, CVD of Si_3N_4, or the final post-metal anneal in H_2 or forming gas. The strength of Si-H bonds is about 0.3 eV. Hence, the barrier energy that would have to be overcome to break the bond would be the sum of the Si-SiO_2 energy barrier (~3.2 eV) plus this bond strength (0.3 eV) and the retarding potential difference between the current path and the interface (assuming $V_{GS}<V_{DS}$). Thus, a hot electron would have to possess ~4.0 eV of kinetic energy to cause such bond breaking and to create an interface trap. The resultant trivalent Si atom at the Si-SiO_2 interface becomes an electron trap. That is, if the bond created by separating the H atom from the Si atom is filled by trapping an electron (as opposed to when it was filled by the H atom), the site has a net negative charge. Such broken bonds accumulate with time, and their filling with electrons would account for the buildup of negative charge at the Si-SiO_2 interface under hot-carrier stressing. It should be noted that while this microscopic model is intuitively satisfying, its authors admit that there is no direct proof of its correctness.

9.3.2 Hot-Hole and Hot-Electron Trapping Model

Another microscopic model of interface trap generation proposes that both hot holes and hot electrons play a role in causing the device degradation.[13] That is, neutral trapping centers at a density N_t^0 are postulated to exist in the 'fresh' SiO_2 film (i.e., not yet having been subjected to hot-carrier injection). These sites are thought to exist within a distance of about 10 nm from the interface. Such neutral centers are initially able to trap only holes (i.e., but not any electrons) that are injected into the oxide. Such hole trapping converts them from neutral to positively charged sites. The density of positive charge centers initially grows during hot-carrier stressing, with their density at any time given by N_t^+. Due to their positive charge, these sites are attractive to electrons, and are also capable to trapping one of them. The latter electron trapping events give rise to two further effects. First, the charge of the trapping site is returned to neutral, with the symbol of the density of such centers being written as N_{the}^0. Second, an interface state is produced (by a mechanism that is still not understood) that becomes charged to a negative state during the transconductance measurement. Their density is denoted by N_{it}^-.

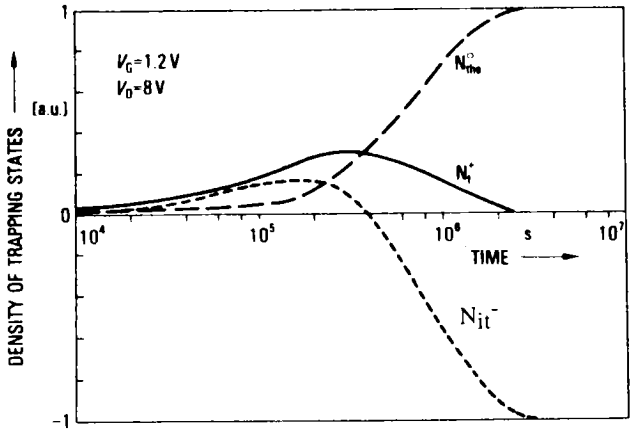


Fig. 9-10 The evolution (as a function of hot-carrier stressing time), of the densities of the different trapping states N_t^o, N_t^+, N_{the}^o, and N_{it}^-. (© IEEE 1984).[13]

Such negatively charged interface states are deemed to be responsible for the negative effective charge that builds up at the Si-SiO_2 interface as a result of the hot-carrier stressing. Eventually all of the original neutral sites (N_t^o) become converted to neutral sites (N_{the}^o) and corresponding interface states (N_{it}^-). The evolution of the densities of these respective charge centers (as a function of hot-carrier stressing time), is shown in Fig. 9-10.

Despite the fact that the degradation mechanisms in either model cannot be directly observed, the amount of degradation (both of V_T and g_m) is observed to follow the bias dependence of substrate current (see Fig. 9-11). Hence, it has been postulated that by monitoring I_{sub} we are, in effect, also monitoring the degradation in an actual device.

This becomes the rationale for using I_{sub} as a universal monitor of hot-carrier stressing in NMOSFETs, as will be described in the next section. However, others point out that attempting to infer the degree of degradation of two different devices by measuring the bias voltages at which the same I_{sub} values are produced may not be valid. This is because knowing the magnitude of I_{sub} does not provide any information as to how far beneath the Si-SiO_2 interface are the impact-ionization-generated carriers created. That is, if the carriers which constitute I_{sub} are being created deep within the substrate, not much damage should occur in the device (because the I_D current path, which is populated by the hot carriers that cause impact ionization, is too far from the surface to permit many of these hot carriers to be injected into the oxide). Conversely, for the same value of I_{sub} arising from impact-ionization near the interface, much more degradation would be expected. Furthermore, it is observed that different bias conditions can cause different degrees of degradation, even at the same value of I_{sub}. Perhaps this is due to the fact that the measured value of I_{sub} incorporates hot carriers of *all* energies, but only a small fraction of these may have sufficient energy to be injected into the oxide and cause degradation. Furthermore, the degradation is known to be localized along

a small region of the device channel (as will be discussed in the upcoming section). Hence, this damage must be caused by injection only in that region. This local component of injected current is not necessarily related to the average value of I_{sub} in a simple manner.

One final aspect of using I_{sub} to compare degradation in different MOSFETs should also be mentioned. That is, hot-carrier degradation is smaller in MOSFETs with thinner oxides, when compared to devices having the same channel length but a thicker oxide, if both are stressed to exhibit the same value of I_{sub}.[14] It is speculated that this happens because the gate electric field is stronger in the devices with the thinner oxides, which causes the electrons injected against this field to be repelled more strongly. Thus, fewer hot electrons with sufficient energy arrive to cause degradation. On the other hand, the devices with thinner t_{ox} exhibit larger I_{sub} for the same stressing conditions. As a result, the maximum power-supply voltage that can be applied to avoid hot-carrier degradation is essentially independent of t_{ox} as L is decreased (see Fig. 9-12).[15] Nevertheless, this is another example that using I_{sub} as a comparator of damage in devices with differing structures may not be valid.

9.3.3 Location of the Hot-Carrier Damaged Region

The location of the damaged region due to hot-carrier stress is found to be above or adjacent to the drain of the device. The lifetime of devices is impacted by the spatial non-uniformity of this damage in that a device which focuses the hot carriers in a small area behaves differently than one that spreads them over a wide area. The extent of the damaged region is a function of device geometry, of the duration and conditions of stress, and of spatial distribution of oxide and interfacial defects.

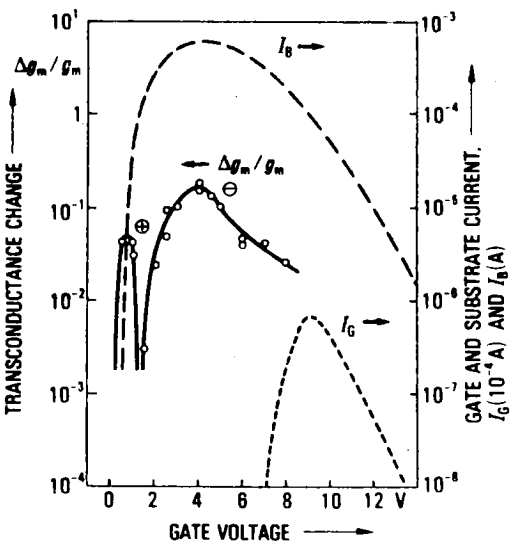


Fig. 9-11 The dependence of I_{sub}, and g_m on V_{GS} for a given value of V_{DS}. Note that both exhibit the same dependence on V_{GS}.[13,149] (© IEEE 1984).

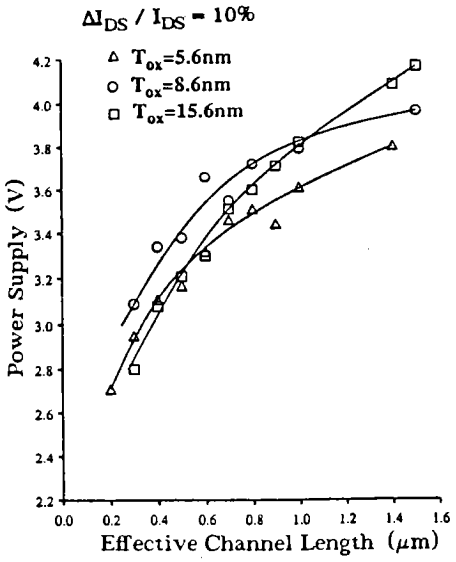


Fig. 9-12 Voltage required to insure 10 years of device lifetime for t_{OX} = 56Å, 86Å, and 156Å.[15] (© IEEE 1988).

Several reports indicate that damage occurs initially at the lateral position of $\mathcal{E}_{ymax}$. The damaged region grows toward the drain, and also at a slower rate toward the source. In very short channel devices the damaged region can grow to a significant fraction of the channel length.[16] However, it has also been reported that the length of the damaged region is independent of channel length; i.e., it appears to have the same length in devices with L_{eff} from 0.95 μm down to 0.15 μm. Thus, the damaged region becomes a larger fraction of the channel length as the devices shrink. This causes the percent degradation in I_D to increase more rapidly as L_{eff} is made smaller for the same stressing time and the same value of I_{sub} (Fig. 9-13).[42]

9.4 CHARACTERIZING MOSFET DEGRADATION DUE TO HOT-CARRIER EFFECTS

As we have been discussing, channel hot carriers in a MOSFET can cause device degradation. Earlier we described models for the driving force that produces such hot carriers ($\mathcal{E}_{ymax}$), as well as microscopic models for the mechanisms believed to be responsible for the damage that they cause. In this section we discuss how to characterize hot-carrier device degradation in a quantitative manner. That is, we identify the device parameters that are degraded, and examine how such damage is measured. Then we indicate how to determine if a device can be operated for a suitably long lifetime without exhibiting excessive hot-carrier degradation. This entails designing

stress tests to verify the device reliability, as well as the issue of accelerated stress testing. In this discussion we also consider the effects of dc versus ac stressing, and the effects of device degradation on circuit behavior. That is, the reduction in a *device* parameter due to hot-carrier stressing may in fact result in a smaller percentage degradation in a corresponding *circuit* parameter. For example, a 10% reduction in I_D may cause the circuit switching speed to drop by only 1%. If this is not taken into account, the device may end up being "overdesigned" in an attempt to mitigate what appears to be excessive hot-carrier *device* degradation. The price for such unnecessary protection may entail increased process complexity and/or the sacrifice in performance of some other device parameter (for example, current drive capability), through the introduction of a more complex drain structure (e.g., an LDD). This issue has been addressed in reliability simulators (e.g., CAS), which we will also describe in this section.

9.4.1 Device Parameters that are Impacted by Hot-Carrier Damage

We have indicated previously that hot-carrier damage can degrade the I_D-V_{DS} characteristics of a MOSFET during long-term operation under normal operating conditions, or during a stress test (see Fig. 9-8). At least four device parameters contribute to such degradation, namely: shift in threshold voltage ΔV_T; reduction in

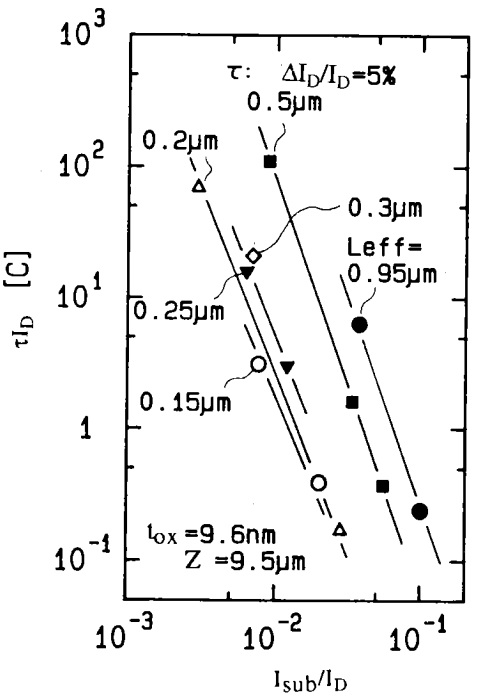


Fig. 9-13 Lifetime plot, τI_D versus I_{sub}/I_D. τ is defined as the time when $\Delta I_D/I_D$ reaches 5% at V_{GS} - V_T = 1.5V.[42] (© IEEE 1990).

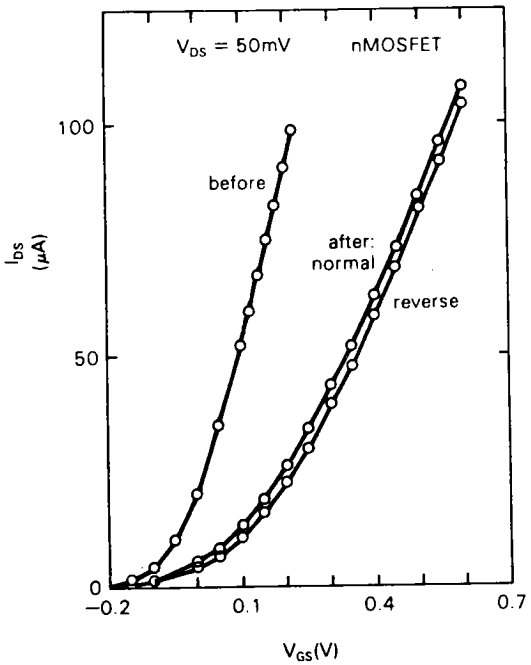


Fig. 9-14 I_D vs V_{GS} characteristics before and after hot-carrier stressing.[19] (© IEEE 1985.)

drain current ΔI_D; reduction in the linear transconductance Δg_m; and shift in subthreshold swing ΔS_t. Figure 9-14 shows an example of the changes in V_T and g_m after severe stressing.[19] The device under test in this example is 100 μm wide, with a 2 μm-wide gate length. Note that V_T here is defined as the gate voltage at which an I_D of 0.1 μA/μm occurs with V_{DS} = 50 mV. Thus, V_T in this figure is the gate voltage at I_D = 10 μA. The value of g_m is the linear slope of the curves in this figure. Note that the linear transconductance or drain current is usually measured since hot-carrier damage will lead to a larger change in g_m in the linear- rather than the saturation-region of operation.

9.4.2 How to Determine if a MOSFET Can be Operated Under Normal Operating Conditions without Exhibiting Failure due to Hot-Carrier Degradation?

We now consider the problem of how to determine if a MOSFET can be operated under normal conditions without exhibiting failure due to hot-carrier damage. To do this, we must first decide what actually constitutes device failure, and then choose an appropriate device failure criterion. Second, we must identify a monitor (e.g., I_{sub}) that is correlated to the stress mechanisms responsible for causing the failure-inducing damage. By tracking the monitor level during the test, the degree of stress applied to the device at any time can be determined. Third, the bias conditions of the stress test at the power-

supply voltage of normal operation (e.g., 5.5V, which is 5.0V nominal plus 10% margin) must be chosen. Most commonly, the combination of V_{GS} and V_{DS} that yields the maximum value of the monitor signal (e.g., maximum I_{sub} value, corresponding to $V_{GS} \approx 0.45V_{DS}$) is used, as this represents the worst-case stress conditions that the device will experience under normal operating conditions.* The device can then be operated under such bias conditions, with I_{sub} being continuously monitored. Periodically the test can be interrupted to measure the device parameter selected as the indicator of failure (e.g., ΔV_T or Δg_m). The stressing continues until the failure criterion is reached, with the time to reach such failure being the hot-carrier device lifetime τ. That is, τ is defined as the stress time at which the change in the device parameter value being monitored reaches the chosen failure criterion.

9.4.2.1 Selecting a Criterion that Identifies Device Failure Due to Hot-Carrier Damage. The first step in determining the hot-carrier reliability of a MOSFET is to select an appropriate criterion indicating that a device has "failed." Most such criteria are based on a specific degree of change in one of the four device parameters listed above. Quite a number of such criteria have been used, apparently somewhat arbitrarily. The ranges of change that have been picked to signal device failure are: shifts in V_T of 10 to 100 mV, and percent changes in I_D or linear g_m of between 3% and 10%.[20] In Fig. 9-15 we note that in non-LDD MOSFETs the shift in V_T is linearly related to the percentage change in g_m, with, for example, a 10 mV shift in V_T corresponding to a 3% change in g_m (with the degradations being caused by conditions of maximum I_{sub} for several V_{DS} values).[17]

While the value of the change in device parameter selected to indicate failure has often been selected somewhat arbitrarily, Ning et al.[21] who performed some of the earliest work on the lifetime of MOSFETs under hot-carrier stressing selected the value in a more logical manner. They described a process with a minimum gate length of 1.3 μm $\pm$ 0.3 μm. They reasoned that the shortest allowable gate length of 1.0 μm would exhibit the worst case hot-carrier degradation, but would also have the highest "fresh" (pre-stressed) value of g_m. They argued that such short devices could afford to relinquish some of their initial g_m advantage to hot-carrier degradation, and still perform as well as the nominal length (1.3 μm) devices. Thus, the device failure criterion was

* Note that not everyone agrees that selecting bias conditions which correspond to the worst-case value of the stress-signal monitor is an appropriate way to design a hot-carrier stress test for a device. That is, if device operation under normal conditions does not expose the device to such high levels of stress monitor signal, the test overstresses the device, and a pessimistic value of its lifetime compared to normal operating conditions will be obtained. This may again make it appear that hot-carrier protection structures need to be incorporated in a device design, when this may not be the case. On the other hand, if the stress time is derated to a time which more realistically mirrors the actual stress that a device will experience during normal operation, the test will reveal a more accurate estimate of the hot-carrier reliability of the device. As we will see, such considerations exist in the hot-carrier reliability simulators that are currently available.

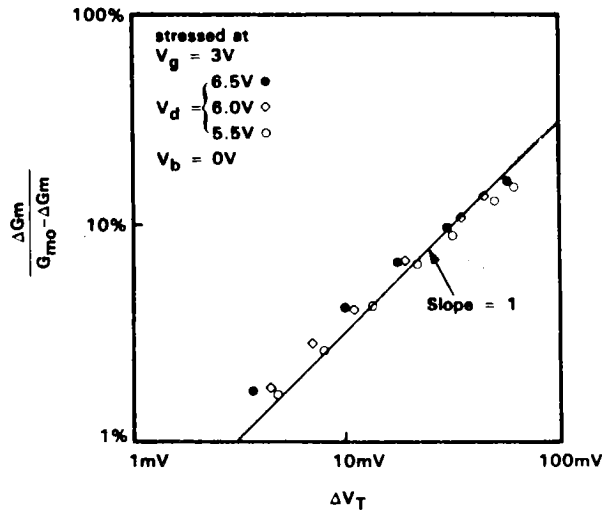


Fig. 9-15 The linear relationship between ΔV_T and Δg_m in non-LDD MOSFETs.[17] (© IEEE 1985).

chosen by Ning as follows: The 1.0 μm devices would be stressed until they lost 30% of their initial transconductance, at which point it would be deemed that they had failed. The time to failure would be the device hot-carrier degradation lifetime. In addition, Ning et al. recognized that the device would not face a continuous dc stress during normal operating circuit conditions. Thus, they estimated a duty cycle of 3% (the fraction of time in which the MOSFET is "ON" in their particular circuit application). This would allow a 10 year (10^5 hr) circuit operation requirement to be assessed by a continuous dc stress in only 3000 hours (~4 months).

9.4.3 Accelerated Testing of Hot-Carrier Device Lifetime

Devices are often designed with the goal of reliable operation (under normal operating conditions) of at least ten years (10^5 hours ~ 3.6×10^8 sec). That is, devices should be able to operate for ten years without exceeding the chosen hot-carrier failure criterion. However, having to wait for ten years to verify the reliability of a device is obviously not practical. Instead, it is necessary to establish some ways to accelerate the stress test to obtain accurate estimates of time-to-failure τ in less time. One way, as mentioned above, is to carry out a dc stress test and then extrapolate the results to normal (i.e., ac) operation. This can reduce the test time by an order of magnitude (or more).

Another means of accelerating the failure time is to increase the stress level. This can be done by increasing V_{DS} beyond the normal power-supply voltage and adjusting the V_{GS} value until a maximum I_{sub} is produced (to $V_{GS} \approx 0.45V_{DS}$). I_{sub} in this case will be higher than that observed under the supply voltage value, leading to shorter times to device failure. Since it has been observed that τ and I_{sub} are related by a power-

law function, by measuring τ at several higher values of V_{DS}, one can extrapolate from these high-stress lifetimes a value of τ at the lower power-supply voltage value.

The general relationship that has been observed to exist between τ and I_{sub} is*

$$\tau = B\,Z\,(I_{sub})^{-m}/(I_D)^{m-1} \qquad (9-6)$$

Since I_D is proportional to Z, but does not vary much otherwise, a simpler, approximate form of Eq. 9-6 that appears to give good agreement with experimental results for devices smaller than ~1.0 μm is[22]

$$\tau \approx B\,(I_{sub}/Z)^{-m} \qquad (9-7)$$

where Z is the device width. As a result, when characterizing device lifetime using accelerated stress testing, the relationship between the measured values of τ and I_{sub}/Z is often graphed (for several values of V_{DS}) on a log-log plot (as shown in Fig. 9-16). It should be noted that while the value of *m* in Eqs. 9-6 and 9-7 is roughly independent of technology (and typically is found to be around 3), B appears to vary quite widely. This variation in B is due to a number of factors, such as hydrogen concentration, mechanical stress, and drain-gate overlap. Nevertheless, it is generally advisable to empirically determine the values of both *m* and B for each device structure by constructing a plot such as the one shown in Fig. 9-16 and extracting the particular values from it. As illustrated in Fig. 9-16 such a plot yields straight lines with a slope of -3, which appears to support the validity of Eq. 9-7.[23] In any case, one can extrapolate the lifetime that would be exhibited by a device operated at the normal power-supply voltage from such a plot.**

Expressions that provide a reasonable fit to experimental data and hence yield useful first order estimates of τ in non-LDD MOSFETs are given as follows:[17]

$$\tau \approx 10^6\,[I_{sub}/Z]^{-3} \qquad (9-8a)$$

$$\approx 3\,(I_{sub}/I_D)^{-3} \qquad (9-8b)$$

Another method for plotting data in a form that is suitable for extrapolating τ at normal operating voltages from τ measured at higher V_{DS} stress voltages is to plot log τ versus $(1/V_{DS})$, with V_{GS} chosen to produce maximum I_{sub} for each particular value of V_{DS}.

* It appears that the more general relationship given by Eq. 9-6 must be used to accurately extrapolate τ in larger devices operated at higher power-supply voltages (such as MOSFETs designed for use in analog circuits or non-volatile memory applications). Since, Eq. 9-6 gives accurate results for the case of both large devices operated at high V_{DS} values and short-channel devices, it is used in the circuit hot-carrier device degradation simulator SCALE instead of Eq. 9-7 to calculate τ.

** Note that Eq. 9-6 can also be graphed on a log-log plot, and this often done in the literature by first converting Eq. 9-6 to the form $\tau I_D \propto (I_{sub}/I_D)^{-m}$ and then plotting τI_D versus I_{sub}/I_D on such a plot.

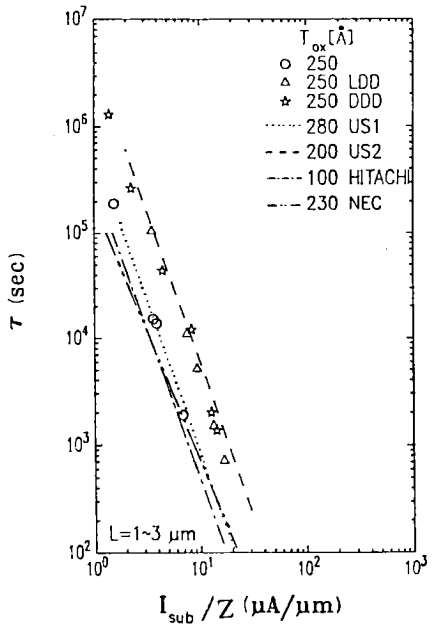


Fig. 9-16 Device lifetime τ versus I_{sub} /Z for various technologies.[23] (© IEEE 1987).

Since according to Eq. 9-5, ln I_{sub} is approximately linearly dependent on $(1/V_{DS})$, this plot should also yield straight lines. As shown in Fig. 9-17, this appears to be the case.[24] This approach is also a somewhat more straightforward way to visualize how τ depends on V_{DS}.

Despite the fact that the relationship given by Eq. 9-6 appears to permit the lifetime of a particular device structure to be accurately extrapolated from accelerated stress test data, this does not mean that we can create a universal curve of τ versus I_{sub} for all devices. Instead, when comparing τ of two or more device structures, it is customary to graph either the change in the selected device parameter (e.g., percentage change in I_D) versus stress time for each device on the same plot (Fig. 9-18), or τ versus $(1/V_{DS})$ for each device on the same plot (Fig. 9-19). Such plots permit us to compare the susceptibility of different device structures to hot-carrier degradation.

Note that the hot-carrier degradation of NMOSFETs by the I_{sub} model of Eqs. 9-6 and 9-7 assumes that the device is stressed at the condition of maximum I_{sub} (i.e., $V_{GS} \approx 045V_{DS}$). Since devices may not be subjected to such gate voltages during normal operation, a model for degradation for all gate voltages would be useful to characterize device degradation under conditions of other than maximum I_{sub}. Such a model has been reported by Woltjer and Paulzen in 1992.[25] They develop a model which predicts

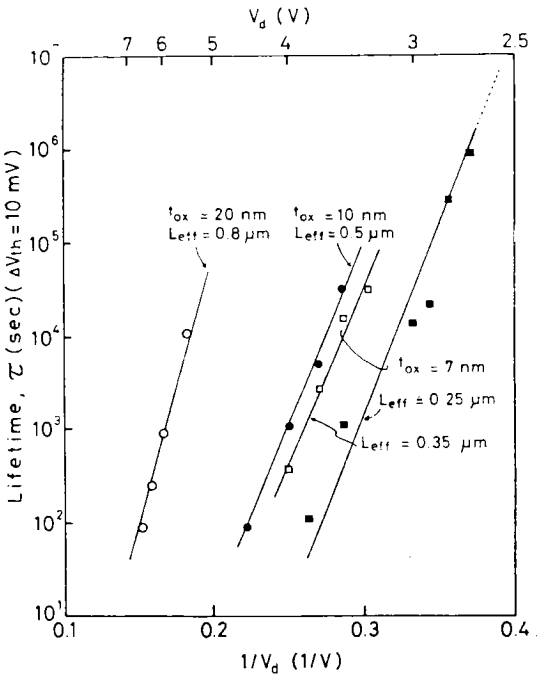


Fig. 9-17 Log τ versus $1/V_{DS}$ plot is an alternative way of plotting lifetime data to enable lifetime to be extrapolated.[24] (© IEEE 1983).

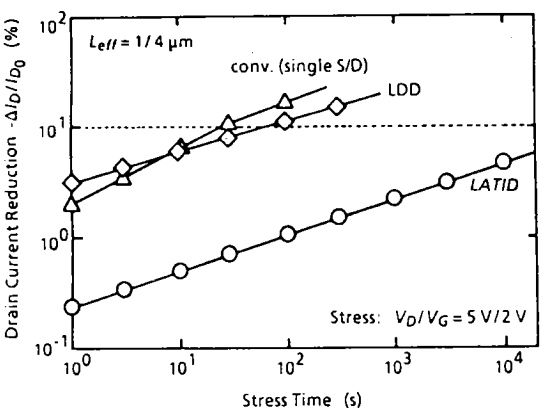


Fig. 9-18 Stress time dependence of hot-carrier-induced I_D degradation measured at V_{GS}/V_{DS} = 3V/3V (S/D reversed after stress) for the 1/4 μm LATID, conventional, and LDD devices.[117] (© IEEE 1989).

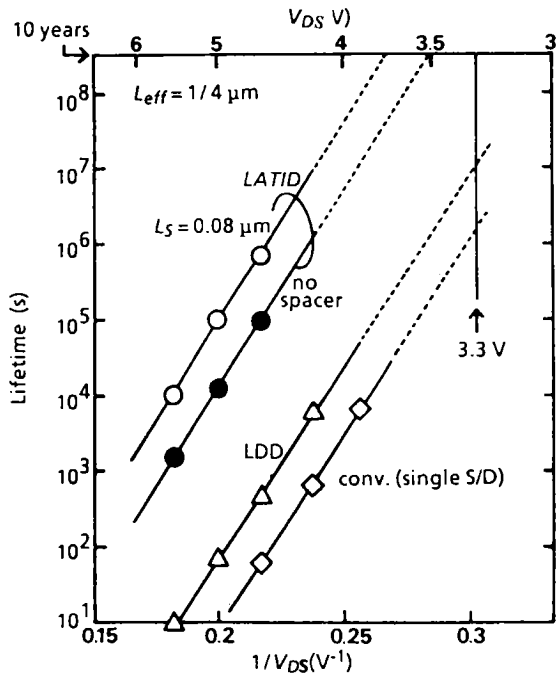


Fig. 9-19 Device lifetime reaching 10% I_D degradation (the more severe 10% Δg_{max} degradation for single S/D) versus the $1/V_{DS}$ for the 1/4 μm LATID, conventional, and LDD devices.[117] (© IEEE 1989).

the generation of interface states at all V_{GS} values for a given value of V_{DS}, and they select a maximum value of the generated interface states as the failure criterion, instead of ΔV_T, ΔI_D, etc.

9.4.4 Hot-Carrier Lifetime Under AC Stressing

Although most hot-carrier reliability stress testing is carried out under dc stress conditions, MOSFETs in an actual circuit are generally stressed in a dynamic (ac) manner. It is therefore important to understand device degradation under dynamic stressing. Modeling and characterization on simple circuits such as inverters, source followers, and transmission gates has been performed.[26] It was found that fast switching of an inverter's input voltage raises V_{GS} above V_T when the output voltage is still high. This corresponds to the condition of high substrate current generation because the driver transistor is biased at $V_{GS}>V_T$ and V_{DS} across the device is also high. Pulsed measurements made on an inverter chain has provided the following simple correlation between static and dynamic lifetime:

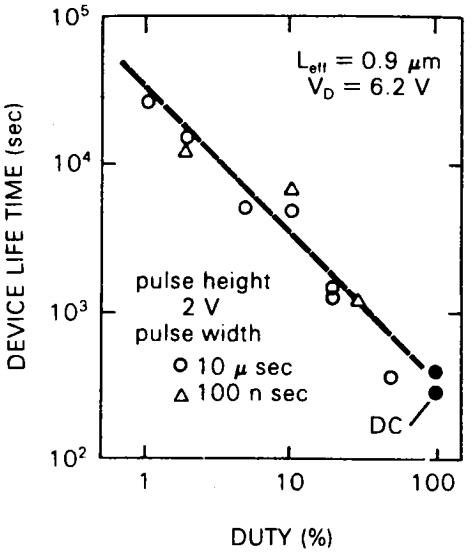


Fig. 9-20 Lifetime versus duty cycle measured after dynamic (ac) hot-carrier stressing.[150] (© IEEE 1986).

$$\tau_{dynamic} = \tau_{dc\ stress}/(\text{duty cycle}) \qquad (9\text{-}9)$$

The experimental confirmation of this relationship is shown in Fig. 9-20. If Eq. 9-8 is valid, this allows extrapolation of dc-stress device lifetime to the actual device lifetime under ac stress. Extrapolating $\tau_{dynamic}$ by use of a linear relationship is referred to as the quasi-static model of device lifetime.

Many initial reports claimed that enhanced degradation for ac stress was due to transient effects during the falling gate-voltage edges.[27-32] These effects have more recently been attributed to measurement difficulties, and thus it has been concluded that the quasi-static model given by Eq. 9-8 is valid for real digital circuits.[33, 34]

9.4.5 Effect of Device Degradation on Circuit Performance

There are two major components which determine the circuit level performance degradation. They are the amount of stress-induced degradation of the device and the sensitivity of circuit performance to the degradation of the individual devices. In a circuit environment, knowing the former is insufficient.[35] For example, a heavily degraded device may not cause significant circuit degradation if the circuit is insensitive to the effects of the degraded parameter. Conversely, a mildly stressed device may cause severe performance degradation due to circuit sensitivity. Traditionally, however, emphasis has been placed on estimating the stress and degradation sustained by individual devices. A test structure that permits hot-carrier degradation in CMOS circuits is described in ref. 36. More recently, several circuit level hot carrier simulators have been introduced, and these will be discussed next.

9.4.6 Hot-Carrier Reliability Simulation

9.4.6.1 SCALE. SCALE (Substrate Current and Lifetime Evaluator) calculates device degradation due to transient voltages (ac stress), which allows the degradation of individual devices in a circuit environment (i.e., where they are subjected to ac stress) to be simulated.[37] SCALE is linked to SPICE externally in a pre- and post-processor fashion so that it functions as an independent simulator. But by using SCALE and SPICE together as a simulator system, the circuit designer can easily identify those devices within the circuit most susceptible to adverse hot-electron degradation effects. Appropriate measures can then be taken to enhance the hot-carrier reliability of these devices.

SCALE carries out the simulation based on a quasi-static approach in the following way: *m* and B of Eqs. 9-6 and 9-7 are first supplied by the user. SCALE then calculates the lifetime τ under dc-stress that would cause the failure criterion to be met under normal operating bias voltages, using Eq. 9-5 to model I_{sub},* and Eq. 9-6 to model τ. (We will hereafter denote the failure criterion in symbolic form as the change in the parameter subscripted with the letter *f*, e.g., $[\Delta I_D/I_{DO}]_f$.) Next, the pre-processor takes the input deck and requests SPICE to calculate the transient voltages at the drain, source, gate, and substrate of all user-selected devices in the circuit. That is, it determines these voltages for each time interval of the period of the SPICE analysis (namely, the time of one switching cycle). With this information, and Eq. 9-5, the value of I_{sub} during each of these time intervals can be determined. Since the value of I_{sub} is known, the fraction of the total degradation of $[\Delta I_D/I_{DO}]_f$ during each of these time intervals can be calculated. By summing the fractional values of $[\Delta I_D/I_{DO}]_f$ for each time interval over one switching cycle the fractional degradation occurring during a single switching cycle is found. Finally, the lifetime due to ac stress for each device in the circuit can be computed by finding the time needed to reach $[\Delta I_D/I_{DO}]_f$. This time corresponds to the time per cycle multiplied by the number of cycles that must elapse to cause $[\Delta I_D/I_{DO}]_f$. As an example, in Fig. 9-21, a significant I_{sub} value is seen to occur only during the time interval around the 21.3 ns time of the switching cycle, as only then is V_{GS} within the range of $0.25V_{DS}$ to $0.5V_{DS}$ when V_{DS} is high. Thus, only during that fraction of the switching cycle is the device being subjected to hot-carrier stress sufficient to cause hot-carrier damage.

The system structure of SCALE is depicted in Fig. 9-22, where we see that SCALE is connected to SPICE in pre- and post-processor configurations. The pre-processor takes the input deck and BSIM parameters and requests a SPICE run to obtain the transient voltage waveforms at the drain, gate, source, and substrate of the user-selected device. The post-processor then calculates the transient substrate current using Eq. 9-5 and the transient drain voltages computed by SPICE. The value of I_D also needed in Eq. 9-5 is calculated by using the BSIM model for I_D.

* Note that Eq. 9-5 has been verified for MOSFETs with uniform channel doping. In order to be able to treat the complex channel and drain (e.g., LDD) doping profiles found in short-channel MOSFETs, some extracted parameters have been introduced.

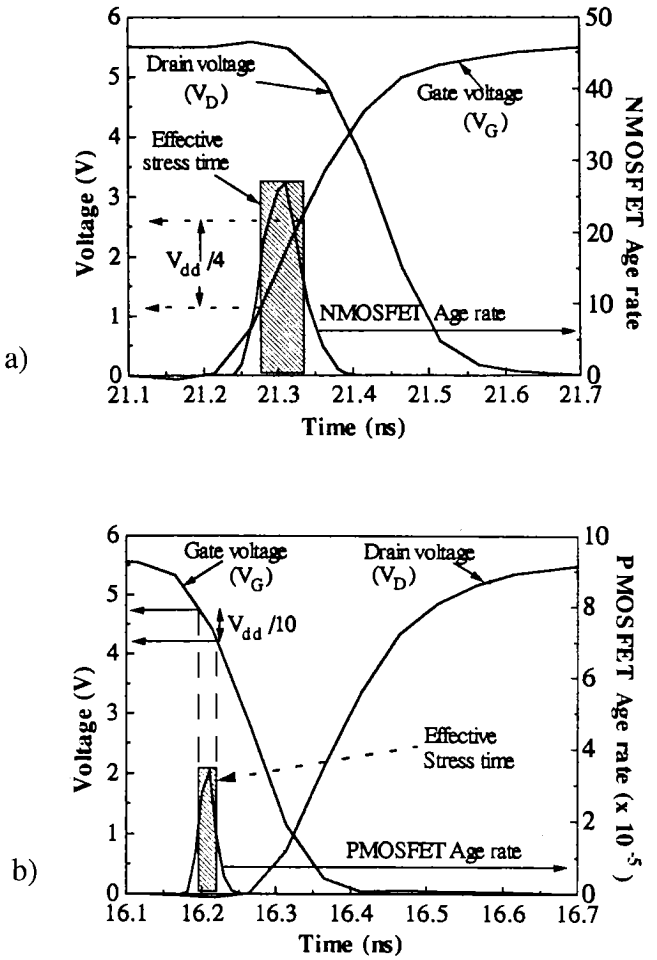


Fig. 9-21 (a) NMOSFET aging rate during circuit transient. The NMOSFET aging rate is a function of the substrate current. The overlap time is a quarter of the rise time. (b) PMOSFET aging rate during circuit transient. The aging rate is a function of the gate current and the effective overlap time is about a tenth of the fall time.[41] (© IEEE 1993).

9.4.6.2 Simulators that Relate the Device-Level Degradation Parameters to Circuit Performance (CAS and HOTRON). As we have mentioned earlier, it is not sufficient to characterize just the degree of degradation of device parameters. Instead, it is necessary to relate the device-level degradation parameters directly to circuit output behavior and performance. That is, it has been demonstrated that not all devices will affect the circuit output equally. Each device may have different performance sensitivities towards the output, and this relationship may

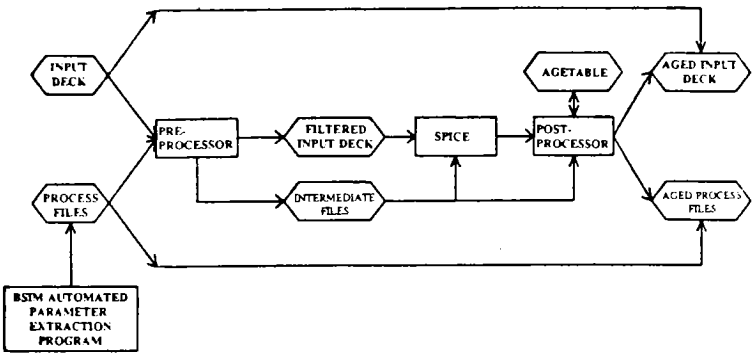


Fig. 9-22 Schematic configuration of how SCALE is incorporated as a pre-and post-processor to be used with SPICE.[151] (© IEEE 1988).

also change with the power-supply voltage. Several simulators have been developed to predict the effects of hot-carrier degradation on MOS circuit behavior, including the *circuit aging simulator* (CAS),[38] and HOTRON.[39] These simulators can directly simulate the entire circuit using degraded device parameters calculated from the hot-carrier degradation models Eqs. 9-6 and 9-7. The simulator CAS is thus a superset of SCALE in that transient substrate current and device lifetime can also be calculated using CAS.* However, CAS is also linked to SPICE so that at any time the device parameters that are degraded by hot-carrier effects can be substituted for the fresh device parameters, and the change in circuit behavior (i.e., aged output compared to fresh output) can be computed using SPICE (see Fig. 9-23).

A set of design guidelines for estimating hot-carrier lifetimes of devices under dynamic stress, as well the hot-carrier lifetimes of some classes of circuits in which such devices are being operated, have been developed using CAS.[41] The conversion from the lifetime under dc stress to the lifetime under ac stress can be accomplished by multiplying the dc stress lifetime by a dc-to-ac conversion factor, namely for NMOSFETs by NTF, and for PMOSFETs by PTF. These conversion factors have are termed *lifetime factors* and are found to be

$$\text{NTF} = 4 / (f\, t_r) ; \qquad \text{PTF} = 10 / (f\, t_f) \qquad (9 - 10)$$

where f is the operating frequency, and t_r and t_f are the rise and fall times of the V_{GS} signal. Figure 9-24 shows the calculated values of NTF (Fig. 9-24a) and PTF (Fig. 9-24b) as a function of f and t_r and t_f. As an example, for an operating frequency of 100 MHz and an input rise time of 0.35 ns, the NTF of NMOSFETs and the PTF of PMOSFETs, are 120 and 300 respectively. Thus, an NMOSFET operated under such

* A timing-simulator-based version of CAS, called IRSIM-CAS has also been developed. It is much faster than the standard CAS mentioned above, allowing it to be used to simulate hot-carrier aging of circuits containing thousands of transistors in a reasonable time.[40]

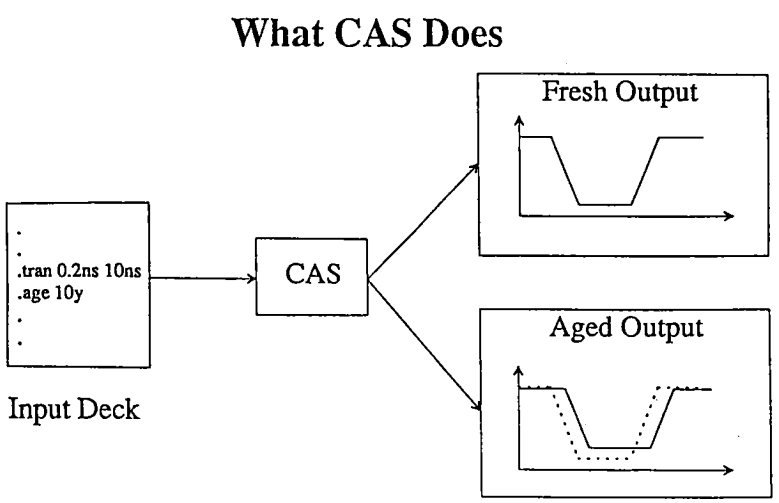


Fig. 9-23 CAS calculates the output of a circuit based on the device parameters as degraded by hot-carrier stressing.

an ac stress would exhibit a lifetime 120 times as long as the one predicted by a dc stress test.

Such guidelines have also been formulated for converting ac-stress lifetime of a device in a circuit to the circuit degradation lifetime with the aid of so-called *speed degradation factors*, NSF for NMOSFETs, and PSF for PMOSFETs. These factors take the following forms:

$$NSF = 1 / [2 k_n (1 + t_{ro} / t_{fo})] \quad (9 - 11a)$$

$$PSF = 1 / [2 k_p (1 + t_{fo} / t_{ro})] \quad (9 - 11b)$$

NSF and PSF appear to be about 1/4 and 1/2 for inverters.

Figure 9-25 shows how NTF and NSF translate dc-stress lifetime to circuit speed degradation. If a 10% $\Delta I_D/I_D$ of an NMOSFET is reached in 0.2 years under worst-case dc stress condition, it will take 16 years to produce a 10% $\Delta I_D/I_D$ if NTF = 120, and this $\Delta I_D/I_D$ degradation will cause a slow down of 2.5% of inverter speed under ac operation at 100 MHz if NSF = 1/4.

Although the circuit-level simulators discussed above can estimate hot-carrier lifetimes under ac stressing, they have generally not yet incorporated the latest understanding of ac stress degradation. A report that presents some guidelines that takes these effects into account, and which have been incorporated into another hot-carrier reliability simulator named ADHOC is given in ref. 54.

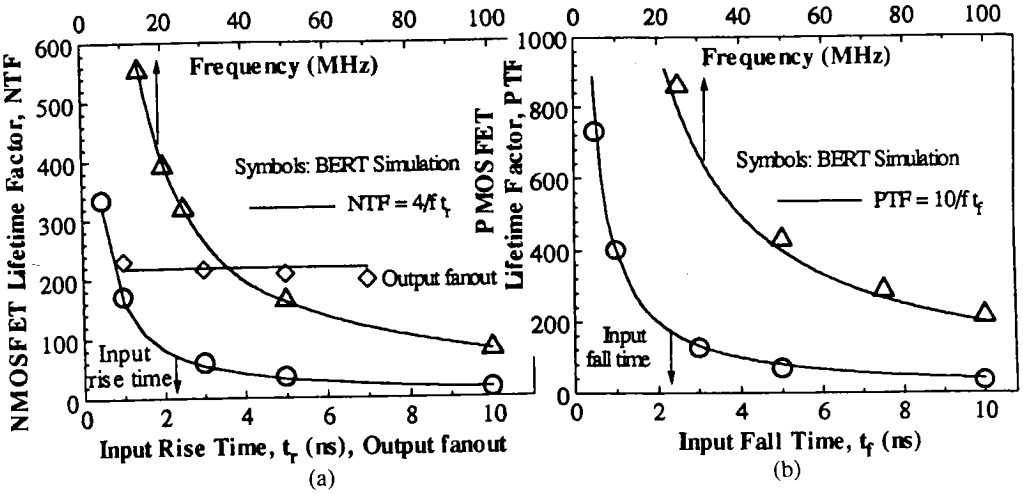


Fig. 9-24 (a) NMOSFET lifetime factor NTF as a function of input rise time, operating frequency, and output fanout factor. The frequency curve is for t_r = 0.5 ns and the input rise time curve is for f = 25 MHz, t_{ox} = 17.5 nm, L_{eff} = 0.5 μm. (b) PMOSFET lifetime factor PTF as a function of input rise time, operating frequency, and output fanout factor. The frequency curve is for t_f = 0.5 ns and the input fall time curve is for f = 20 MHz, t_{ox} = 17.5 nm, L_{eff} = 0.5 μm.[41] (© IEEE 1993).

9.5 DECREASING HOT CARRIER DEGRADATION BY MODIFYING THE MOSFET STRUCTURE

As was pointed out earlier, device lifetime (and other hot-carrier effects) can be improved in a number of ways, including, reducing the bias voltages, significantly strengthening the gate oxide against hot-carrier damage (i.e., in effect increasing B in Eq. 9-6), or

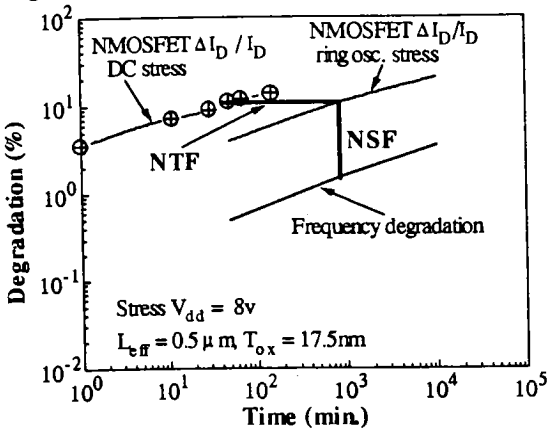


Figure 9-25 NMOSFET lifetime factor NTF, and NMOSFET speed degradation factor NSF, translate dc device degradation to circuit speed degradation.[41] (© IEEE 1993).

modifying the MOSFET structure. If the power-supply voltage cannot be reduced (e.g., it may need to be kept at 5V for system-compatibility reasons), and if the gate oxide cannot be sufficiently strengthened against hot-carrier degradation to provide adequate operating lifetimes, then the only remaining option is to modify the MOSFET structure.

A quantitative example will concretely demonstrate why a way must be found to reduce $\mathcal{E}_{ymax}$ in submicron MOSFETs. Assume that a 1 μm NMOSFET has the following structural and electrical device parameters: t_{ox} = 20 nm, r_j = 0.25 μm, V_{DS} = 5V, and V_{DSsat} = 1V. In this case, from Eq. 9-2a, $l \approx 138$ nm = 1.38×10^{-5} cm, and (V_{DS} - V_{DSsat}) = 4V. Then, from Eq. 9-1, $\mathcal{E}_{ymax} \approx 2.9 \times 10^5$ V/cm. Now we can roughly estimate the hot-carrier lifetime of the device using Eq. 9-8b. We first use Eq. 9-5 to model I_{sub}/I_D in this device. Then we insert this I_{sub}/I_D value into Eq. 9-8b to find τ. Equation 9-5 yields a value for I_{sub}/I_D ~0.015, and then Eq. 9-8b predicts that such a device will have a hot-carrier lifetime τ of only 1×10^6 sec or ~0.4 year. Obviously, this lifetime is too short (the design goal is typically 10 years). Clearly, a way to increase τ needs to be found.

Equation 9-1 suggests that one way to modify the MOSFET structure to reduce $\mathcal{E}_{ymax}$ is to increase the magnitude of l. According to Eq. 9-2, this can be accomplished by increasing t_{ox} or r_j, (or both). However, increasing either of these parameters will worsen the susceptibility of the MOSFET to short-channel effects (as was discussed in chap. 5). Thus, another tack must be taken to modify the device structure other than by using this approach.

An alternative approach in *n*-channel MOSFETs is to create a lightly doped (n^-) buffer region between the heavily As-doped (n^+) drain and the channel, *outside of the channel.* In other words, the idea is to produce a MOSFET device having a drain with a *graded,* or *lightly-doped* extension. Qualitatively, $\mathcal{E}_{ymax}$ is reduced by such a buffer region because the maximum electric field in a reverse-biased *pn* junction is highest when the junction is abrupt. By replacing the abrupt drain doping profile of a conventional MOSFET with a drain that has a more gradually decreasing lateral doping profile (i.e., a *graded drain*), the voltage drop becomes shared by the drain and the channel [in contrast to an abrupt n^+/p drain junction, in which almost the entire voltage drop occurs across the lightly doped (channel) side of the junction]. From a simple quantitative perspective using the model for $\mathcal{E}_{ymax}$ given by Eq. 9-1, $\mathcal{E}_{ymax}$ is significantly reduced by the presence of such a lightly doped region because this is another way to increase the denominator of Eq. 9-1.

Graded-drain regions can be created in MOSFETs in a number of ways, including: (1) using phosphorus in place of As as the dopant of the source/drain regions; (2) adding fast diffusing phosphorus to an As-doped drain region, and driving the phosphorus laterally ahead of the arsenic with a high temperature diffusion step (to create a so-called *double-diffused drain* [DDD] structure); and (c) pulling the highly doped (n^+) drain region away from the gate edge with an "oxide spacer" to create a so-called *lightly doped drain* (LDD) structure. Here we describe these approaches in more detail, focusing on the LDD, as it is the most important technique for forming graded-drain MOSFETs. Table 9-1 shows the evolution of the AT&T Twin-Tub CMOS technology,

Table 9-1 Evolution of Device Structures in AT&T's Twin-Tub CMOS Technology Development[43] (Twin-Tub VI announced in 1989)[44]

	Twin-Tub I	Twin-Tub II	Twin-Tub III	Twin-Tub IV	Twin-Tub V	Twin-Tub VI
Design Rule	3.5 μm	2.5 μm	1.75 μm	1.25 μm	0.9 μm	0.6 μm
L_{eff}	2 μm	1.5 μm	1.3 μm	1.0 μm	0.75 μm	0.4 μm
t_{ox}	600 Å	350 Å	250 Å	200 Å	150 Å	125Å
Device Structure	Conventional	Conventional	DDD	LDD	N&P LDD	N-DDD

which gives an indication of how graded-drain structures were incorporated into MOS devices as the devices were scaled.

9.5.1 Phosphorus-Drain Structure

The simplest method to incorporate a graded-drain region into a conventional (i.e., single As drain) MOSFET device, is to replace the arsenic with phosphorus as the source/drain donor impurity.[45] This structure is called the (single) *phosphorus-drain MOSFET*. The faster-diffusing phosphorus provides additional lateral length to the drain dopant profile, which widens the depletion region of the drain-channel junction, and thus reduces the maximum lateral electric-field strength $\mathcal{E}_{ymax}$.

To keep the sheet resistivity of the source/drain regions reasonably low, the dose of the phosphorus implantation must be greater than $1x10^{15}$ cm^{-2}. The resulting device exhibits values of I_{sub} about an order of magnitude smaller than those made with As-doped drains. The reduction in $\mathcal{E}_{ymax}$ is actually due to both the graded phosphorus diffusion profile and the much deeper source/drain junction depth. However, because of this deep source/drain junction, short channel effects are much more difficult to suppress in phosphorus-drain structures. Thus this method is not suitable for NMOSFETs with gate lengths smaller than about 1.3 μm.[46]

9.5.2 Double Diffused Drain (DDD)

While the single phosphorus-drain is not suitable for alleviating the hot-carrier reliability problem in short-channel MOSFETs, if arsenic and phosphorus are simultaneously used as drain dopants, additional improvement in hot-carrier reliability can be obtained with less excessive lateral encroachment into the MOSFET channel. The two dopants are co-implanted into the same region (using two separate implants, i.e, a medium phosphorus dose [e.g., $1x10^{14}$ to $1x10^{15}$ cm^{-2}] and a heavy arsenic dose [e.g., $5x10^{15}$ cm^{-2}] both at 0° tilt angle). This initially aligns both dopants to the same gate edge. Since these two implants are followed by a high temperature anneal step, in which both dopants diffuse simultaneously, the structure is called a *double-diffused drain* (DDD). However, the faster-diffusing phosphorus is driven farther under the gate edge (and elsewhere, see Fig. 9-26) than the arsenic, creating a less abrupt concentration gradient for the drain. Figure 9-27 shows a typical impurity distribution

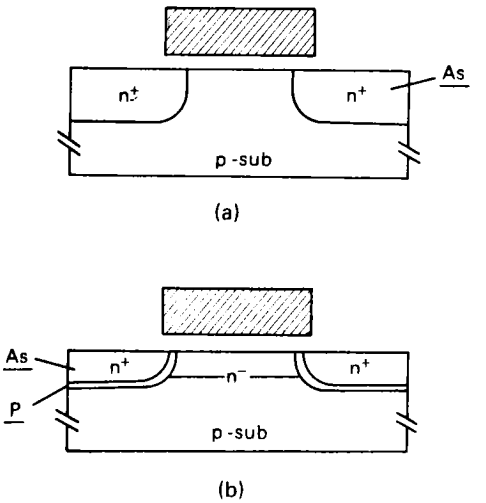


Fig. 9-26 An NMOSFET with (a) conventional single drain; and (b) double-diffused drain.

along the channel direction of a DDD MOSFET. The doping profile of a conventional single-As-drain MOSFET is also depicted for comparison.[47] A significant reduction in $\mathcal{E}_{ymax}$ occurs in the DDD, as is shown in Fig. 9-28. For devices with channel lengths greater than about 1.25 μm, I_{sub} reduction factors of 4 to 6 are typically obtained. It should be noted that some DDD structures are implemented using the oxide spacer technology generally associated with the LDD. In such structures, the implants are carried out after spacer formation.

The DDD structure is attractive because it adds little complexity to the process sequence. That is, it can be introduced into a CMOS process flow without any extra masking steps, whereas an NMOS LDD process may require an extra masking step to prevent the n^- implant from entering the PMOS source/drain regions. Furthermore, since the medium-dose phosphorus implant (n^-) and the high dose arsenic implant (n^+) are aligned to the same gate edge, the relative position between the n^- and n^+ edges (which is the basis of all graded-drain structures), is determined only by the high temperature anneal process. Thus, the DDD structure exhibits relatively consistent device characteristics and I_{sub} values.

On the other hand, this approach permits no freedom to optimize the length of the n^- region and simultaneously achieve a shallow junction. This limits the degree to which $\mathcal{E}_{ymax}$ can be reduced. That is, in order to get an n^- region of sufficient length, a long thermal drive needed. Unfortunately, this also causes a deep *vertical* phosphorus junction. Since shallow junctions are needed to prevent short-channel effects (see chap. 5), as the gate length approaches 1 μm in DDD structures, the deep vertical junction degrades the short-channel behavior to an unacceptable degree. The long drive-in also causes redistribution of dopants in the channel, which is also undesirable.

In addition, it should be mentioned that selecting the correct process and device parameters which yield an optimized DDD structure involves significant process and

device calibration. That is, too much phosphorus or excessive diffusion will yield punchthrough problems, while too little phosphorus will provide inadequate protection from hot-electron degradation. Furthermore, if an oxide spacer is used, the gate/drain overlap depends strongly on the width of the oxide sidewall spacer, which could make such devices very susceptible to hot-carrier degradation unless the spacer width is tightly controlled.

For 5V supply voltages, the DDD structure is therefore not suitable for devices with gate lengths smaller than 1.25 μm. As shown in Table 9-1 it was used in the 1.75 μm ATT CMOS-III technology, but not in the following 1.25 μm CMOS-IV process.[43] However, it is reported that the DDD structure has returned in the ATT 0.6-μm CMOS-VI technology, in which a smaller (3.3V) power-supply-voltage is used.[44]

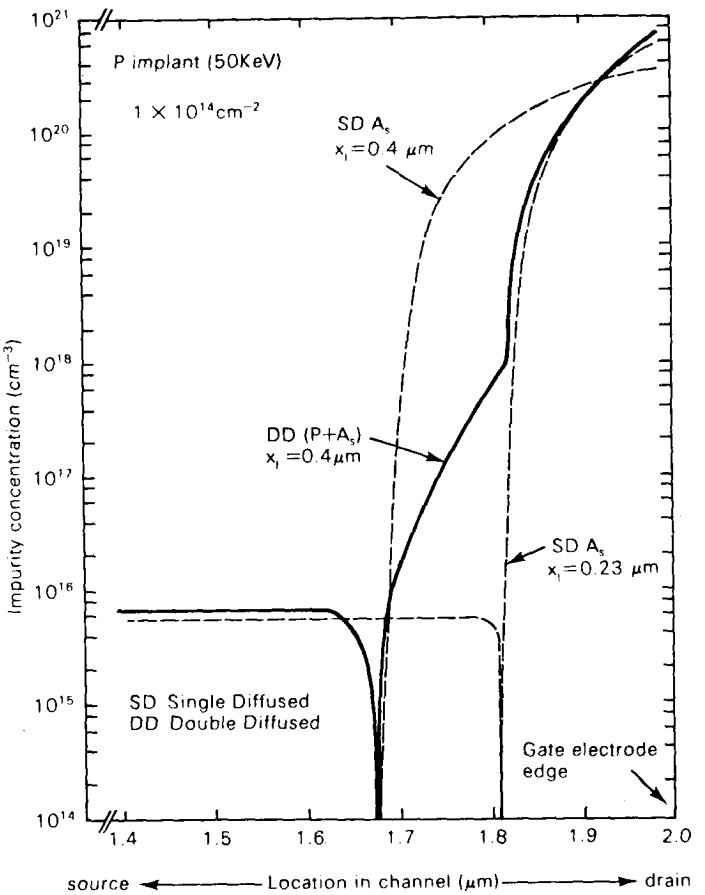


Fig. 9-27 Lateral impurity profiles for two single-diffused drain structures and one double-diffused drain structure.[47] (© IEEE 1985).

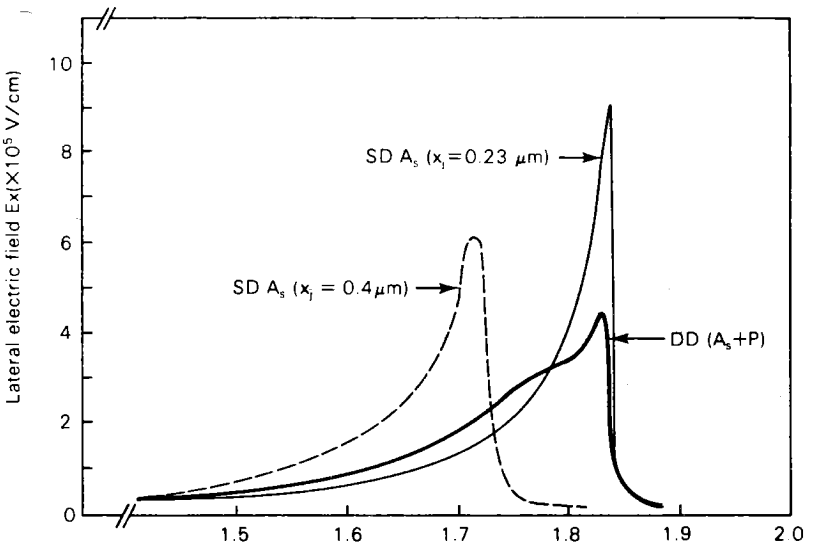


Fig. 9-28 Lateral electric field distributions associated with the three structures shown in Fig. 9-27.[47] (© IEEE 1985).

9.6 LIGHTLY DOPED DRAIN (LDD) STRUCTURES

The most widely used device structure for increasing the hot-carrier reliability in 5V-NMOSFETs with gate lengths smaller than 1.25 μm is the *lightly doped drain* (LDD), and in the upcoming sections we will describe it in detail. First we present models that allow the degree to which $\mathcal{E}_{ymax}$ is reduced in an LDD to be estimated in terms of the device parameters and bias voltages, as well as models which predict the performance penalty that must be paid by using an LDD in a MOSFET device. Next we discuss the process sequence used to fabricate a conventional LDD, and then we examine the limitations of such structures. Finally we describe how modifications have been made to the conventional LDD structure to overcome these limitations.

In the LDD structure, as first described in 1980 by Ogura et al.,[48,49] the drain is formed by two implants (Fig. 9-29). One of these is self-aligned to the gate electrode, and the other to the gate electrode on which two oxide sidewall spacers have been formed. The purpose of the first implant dose is to produce a lightly doped section of the drain at the gate edge near the channel.

Figure 9-30 shows the electric-field profile at the drain of a MOSFET, both with and without a conventional LDD structure (as obtained from 2-D numerical analysis of

the MOSFET device). In this case, the LDD provides about a 30-40% reduction in $\mathcal{E}_{ymax}$. Since we noted earlier that the average energy of the hot-electrons is exponentially dependent on $\mathcal{E}_{ymax}$ this reduction ought to significantly decrease the damage caused by such carriers in the MOSFET. Hence, the hot-carrier reliability of the LDD device should also be greatly increased. However, this benefit comes at a price, namely the lowering of the drive current of the MOSFET.

While 2-D numerical analysis techniques provide accurate simulations of the lateral electric field in both non-LDD and LDD MOSFETs, it would be useful if analytical models were available to estimate $\mathcal{E}_{ymax}$ and the resultant reduction in I_D in LDD devices. Here we describe such analytical models for estimating both phenomena in LDD MOSFETs.

9.6.1 Modeling $\mathcal{E}_{ymax}$ in LDD MOSFETs

9.6.1.1 Simple Model for Estimating $\mathcal{E}_{ymax}$ in LDD MOSFETs

The LDD reduces $\mathcal{E}_{ymax}$ because some of the drain voltage is dropped across the n^- region. This voltage drop does not occur in non-graded-drain devices because such an n^- region is absent. The simplest quantitative model for estimating the reduction of $\mathcal{E}_{ymax}$ in an LDD assumes that the voltage dropped across the n^- region is as high as possible.[18] For this to occur, the lateral electric field would have to be roughly constant across the entire length of the n^- region L_{n^-}, and be equal to its maximum possible value $\mathcal{E}_{ymax}$. In this case, a modified form of the equation used to calculate $\mathcal{E}_{ymax}$ in a non-LDD device in saturation (Eq. 9-1) would also be appropriate for estimating $\mathcal{E}_{ymax}$ in an LDD

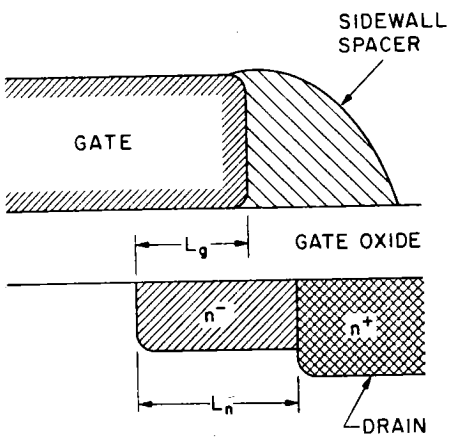


Fig. 9-29 Cross section of the drain end of an LDD device, showing the sidewall spacer, the gate overlap length, L_g, and the length of the lightly doped drain region L_{n^-}.

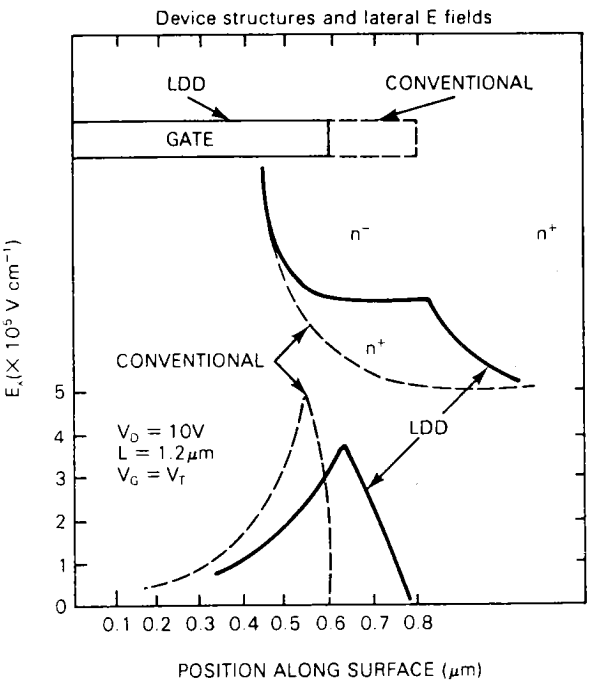


Fig. 9-30 Magnitude of the electric field at the Si-SiO_2 interface as a function of distance: $L = 1.2\ \mu m$, $V_{DS} = 8.5$ V, $V_{GS} = V_T$.[48] (© 1980 IEEE).

$$\mathcal{E}_{ymax}\ (\text{LDD}) = (V_{DS} - V_{DSsat} - \mathcal{E}_{ymax}L_{n^-})/\ 0.22\ t_{ox}^{1/3}r_j^{1/3} \qquad (9\text{-}12)$$

which after some algebraic manipulation becomes

$$\mathcal{E}_{ymax}\ (\text{LDD}) = (V_{DS} - V_{DSsat})/\ (0.22\ t_{ox}^{1/3}r_j^{1/3} + L_{n^-}) = (V_{DS} - V_{DSsat})/l' \qquad (9\text{-}13)$$

Equation 9-13 offers an alternative perspective of the phenomenon of $\mathcal{E}_{ymax}$ reduction in an LDD device. That is, Eq. 9-13 suggests that the potential difference $V_{DS} - V_{DSsat}$ is no longer dropped across the effective length of the velocity saturation region $l = 0.22\ t_{ox}^{1/3}r_j^{1/3}$, but instead is dropped across l', which is the sum of l and L_{n^-}, or $l = (0.2\ t_{ox}^{1/3}r_j^{1/3} + L_{n^-})$. It should be noted that in practice, the electric field will not be constant across L_{n^-} and therefore the term $\mathcal{E}_{ymax}L_{n^-}$ in Eq. 9-12 should in fact be reduced by some factor. Nevertheless, Eq. 9-13 suggests that the effect of an LDD

structure is to raise the maximum acceptable value of V_{DS} over that of a conventional MOSFET, regardless of L.

9.6.1.2 Analytical Model for Calculating $\mathcal{E}_{ymax}$ in LDD MOSFETs Using the Pseudo-2D Approach to Solving the Poisson Equation.

The above model is useful for providing some rough quantitative insight into the behavior of LDD MOSFETs but gives no help in designing an LDD structure for a given MOSFET with respect to selecting L_{n^-}, the n^- dose, or the degree of overlap of the n^- region under the gate. Furthermore, while this model helps to explain why an LDD reduces $\mathcal{E}_{ymax}$, it implies that this reduction will continue to increase as L_{n^-} grows longer. However, this is not the case. Instead it has been observed that there is some L_{n^-} length beyond which $\mathcal{E}_{ymax}$ no longer significantly decreases. Furthermore, the above model gives no clue as to how the implant dose is related to the reduction of $\mathcal{E}_{ymax}$, nor as to the location of $\mathcal{E}_{ymax}$ with respect to the gate edge.

A more elaborate analytical model for $\mathcal{E}_{ymax}$ in LDD MOSFETs which provides some of this information was described by Lee et al.[50] This model was based on the same pseudo-2-D solution approach of the Poisson equation of Ko et al., as was discussed in more detail in chap. 5, section 5.4.3.3. Here the doping is assumed to be constant throughout each device region (i.e., in the n^- and n^+ regions of the drain, and the channel region). The solution of the 2-D Poisson equation based on this pseudo-2D method provides the value of $\mathcal{E}_y$ versus y in the device. The maximum value found is $\mathcal{E}_{ymax}$. The location of $\mathcal{E}_{ymax}$ can therefore also be obtained from this model. The dependence of $\mathcal{E}_{ymax}$ on the doping concentration in the n^- region can also be determined with this model. Finally, a field reduction factor (FRF) is defined which gives the ratio of $\mathcal{E}_{ymax}$ in an LDD device to $\mathcal{E}_{ymax}$ of a non-LDD device with the same dimensions. This permits various LDD MOSFET structures to be quickly compared with respect to their effectiveness in decreasing $\mathcal{E}_{ymax}$.

Lee et al. used the model to calculate the FRF in LDD devices for three general device structures: (a) one in which the n^- region is completely overlapped by the gate; (b) one in which there is zero overlap between the n^- region and the gate; and (c) one in which a partial overlap exists between the n^- region and the gate. They drew several conclusions about the LDD device from their analysis: (1) The device with full n^--to-gate overlap can potentially provide the largest FRF of the three structures; (2) The maximum FRF in an LDD is about 0.5, and this occurs in full-overlapped structures having an optimum doping (and a maximum length of L_{n^-}). That is, if a longer L_{n^-} region is used, the FRF no longer increases. On the other hand, the resistance of the current path between the source and drain terminals does continue to rise as L_{n^-} grows longer, which will cause I_D to decrease; (3) In partially overlapped devices, the location of $\mathcal{E}_{ymax}$ depends on the doping level in the n^- region. That is, for light n^- doses the $\mathcal{E}_{ymax}$ location is at the n^-/n^+ junction (i.e., it exists beyond the edge of the gate), while for higher n^- doses the location of $\mathcal{E}_{ymax}$ moves to the p/n^- junction (i.e., under the gate). As we shall see, when $\mathcal{E}_{ymax}$ is located under the gate, devices exhibit improved hot-carrier reliability.

While this model is useful in providing insight into the general directions that LDD device-design should take, it does not provide enough information to be able to predict the actual hot-carrier lifetimes of LDD devices designed only from the data provided by the model. That is, the assumption is still made that an LDD MOSFET with the smallest value of $\mathcal{E}_{ymax}$ will yield a device structure with the highest hot-carrier reliability. Unfortunately, this is not always the case. Instead, other factors enter into the picture, as will be discussed in the section that examines the issues of hot-carrier reliability in LDD devices (section 9.7.2). We will also describe methods that are better able to predict hot-carrier hardness in LDD devices (section 9.10). These are based on more rigorous simulation procedures than those used in the above model.

9.6.1.3 Model for the Series Resistance Introduced by the n^- Region of the LDD. The value of the resistance R_{n^-} introduced by the n^- region of an LDD can be simply modeled if we assume that the doping concentration in the n^- region is uniform (with a value N_D) along the distance L_{n^-} and to the depth of the region r_j. In this case, we can model R_{n^-} in an NMOSFET according to:

$$R_{n^-} = Z L_{n^-} /(q \mu_n N_D r_j) \qquad (9 - 14)$$

where Z is the width of the device and μ_n is the mobility of the electrons in the n^- region (e.g., 800 cm^2/Vsec). The impact of R_{n^-} on I_D is described in section 9.7.1.1.

9.6.2 Overview of the Process Sequence to Form LDDs

The basic process sequence used to form early LDD MOSFETs in simple NMOS technologies is shown in Fig. 9-31. The poly gate is patterned with an RIE etch process designed to produce a poly feature with vertical sidewalls (Fig. 9-31b). In some processes the thermally grown gate oxide beyond the edges of the gate is kept in place after the poly etch, in an attempt to maintain a high quality interface under the sidewall oxide. This requires a polysilicon etch with extremely high selectivity to oxide, since the gate oxide is normally quite thin (e.g., 12.5-20 nm thick). Next, the n^- implant (also called the *graded-drain,* or *tip* implant) is carried out with phosphorus to form the lightly doped region of the drain. In early LDDs the doses ranged from $5x10^{12}$ to $1x10^{13}$ cm^{-2} at energies of 40-60 keV (Fig. 9-31c). Note that the n^- region at this point is self-aligned to the gate electrode. The n^- implant is carried out using a medium-current implanter with the ion beam tilted at 7° (to minimize implant channeling effects - see vol. 1, chap. 9).

Following the n^- implant step a thin oxide is thermally grown on the patterned poly gate using a short dry-oxidation process (this step is called *poly re-ox*). A conformal CVD oxide film is then deposited. A popular method is to decompose TEOS at 725°C, as this produces a film with excellent conformality, and if the poly profile is vertical, no re-entrant profile. A conformal deposition process produces a thicker film at the edge of the gate-to-active area step than on flat areas. Thus, an anisotropic dry-etch process will clear the oxide in the flat areas while leaving spacers at the sidewalls of the poly

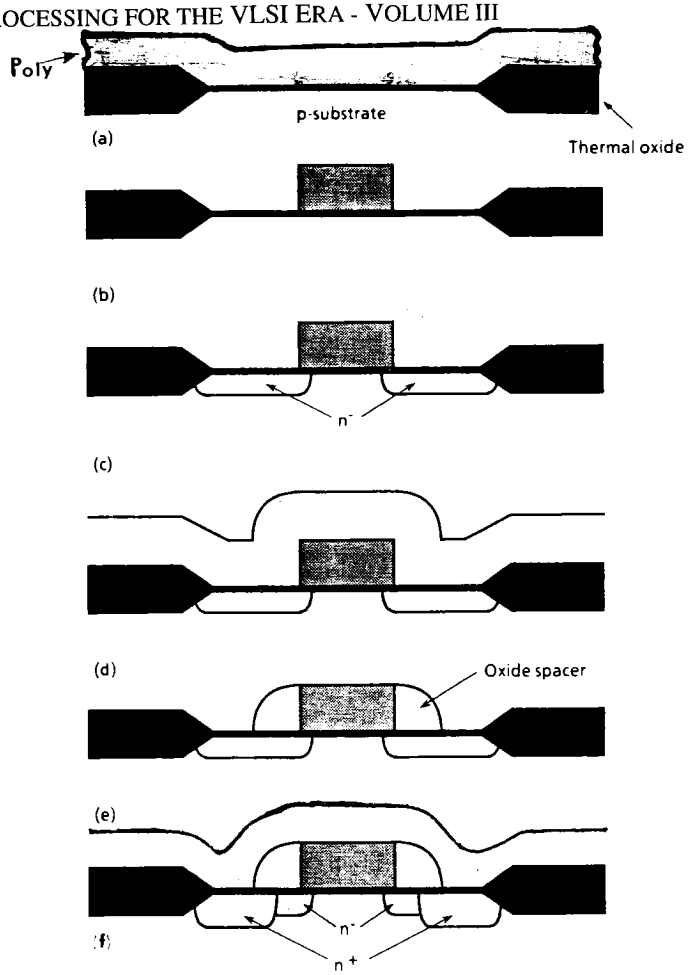


Fig. 9-31 Basic process sequence to form an LDD NMOSFET. (a) After poly deposition; (b) after poly etch; (c) LDD implant; (d) spacer deposition; (e) spacer etch; (f) n^+ implant and passivation.

gate (Fig. 9-31d). If the spacers are etched to endpoint with minimal overetch, the spacer height will be close to the same height as the poly. For poly gate thicknesses of 0.3 μm and a deposited oxide film thickness of ~0.3 μm, spacer widths of ~0.2 μm are typically obtained. A discussion of spacer thickness and shape dependence, with respect to variations in the deposited film thickness, conformality, overetch time etc. is given in ref. 51.

It should also be noted that to account for variations in the spacer oxide layer thickness, some overetch is needed in the etch step that creates the spacer. During the overetch time the spacer shape is altered (i.e., its thickness and height from the top of the poly gate will change). In addition, the field oxide and the silicon in the source/drain regions are also etched. Excessive overetching of these regions is undesirable from

device isolation and junction leakage viewpoints, respectively. Figure 9-32 shows a SEM photograph of an NMOS transistor with an oxide spacer at the conclusion of the spacer etch process, where the n^+ gate and drain regions have been decorated with a selective etch (the n^- region is not visible). Note that the silicon in the drain has been slightly etched during the spacer overetch as shown by the depression in the silicon surface to the right of the gate. Leakage of the source/drain junctions, localized in the junction directly beneath the spacer edge has been reported in oxide-spacer LDDs. This problem has been attributed to the silicon loss and etch damage in the substrate arising from the spacer etch,[52,53] and to defects formed by the subsequent n^+ implant and anneal. (The shape of the oxide spacer produced by the spacer etch process has also been reported to impact the formation of these defects.[54])

After the oxide spacers are formed, a high dose n^+ source/drain implant is carried out with a high-current implanter. Typically, arsenic is implanted at a dose of about $5x10^{15}$ cm^{-2} and at energies of 40-80 keV. This forms the low-resistivity drain region, which is merged with the lightly doped n^- region. Because the spacer serves as a mask for the As implant, the heavily doped n^+ region is self-aligned to the sidewall spacer edges and is thus offset from the gate edge. Since the edge of the n^+ region is further away from the channel than would be the case in a conventional drain structure, the depth of the heavily-doped region of the drain can be made somewhat greater without adversely impacting the device operation (e.g., 0.3 μm deep versus 0.18 μm deep). The increased junction depth lowers both the sheet resistance and the contact resistance of the drain. Deeper junctions also provide better protection against junction spiking. An anneal is then performed to cause the n^- dopant to sufficiently diffuse laterally under the gate edge and to activate the implanted n^+ dopants. (Actually, the entire thermal budget after n^- implant determines the degree of lateral diffusion, and this is often dominated by the interlevel dielectric densification and reflow step, e.g., 950°C for 20 min.)

By introducing the sidewall spacer after the n^- implant to laterally offset the n^+ implant, the LDD structure offers more freedom in optimizing the n^- length than the DDD structure. That is, the n^- length is controlled by the lateral spacer dimension and thermal drive cycle, and is made independent from the n^+ junction depth. If a salicide

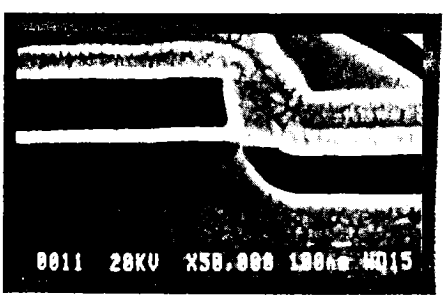


Fig. 9-32 SEM photograph of an NMOS transistor with an oxide spacer at the conclusion of the spacer etch process, where the n^+ gate and drain regions have been decorated with a selective etch (the n^- region is not visible).[67] Reprinted with permission of Semiconductor International.

process is being implemented, an oxide spacer is also needed to prevent shorting of the gate and the source/drain regions, and this structure is intrinsic to the LDD approach. However, as we shall describe in the next section, the conventional LDD has other problems and disadvantages, and improvement in hot carrier resistance in the basic LDD structure is obtained only by trading off device performance.

9.7 PERFORMANCE AND HOT-CARRIER LIMITATIONS OF THE CONVENTIONAL LDD

As emphasized earlier, the reason LDDs are incorporated into submicron MOSFETs is to increase their hot-carrier reliability. It would be ideal if this benefit could be obtained without degrading the other performance characteristics of the device. However, the introduction of an LDD generally does cause some performance degradation, primarily by causing a reduction in the drive current of the MOSFET. An example of this effect is illustrated in Fig. 9-33,[55] which compares the drain current in conventional and LDD MOSFETs as their gate length and n^- dose is varied. The LDD MOSFETs exhibit smaller I_D values. To make matters worse, if LDD MOSFETs are poorly designed they may even be more susceptible to hot-carrier degradation than comparable non-graded-drain MOSFETs.

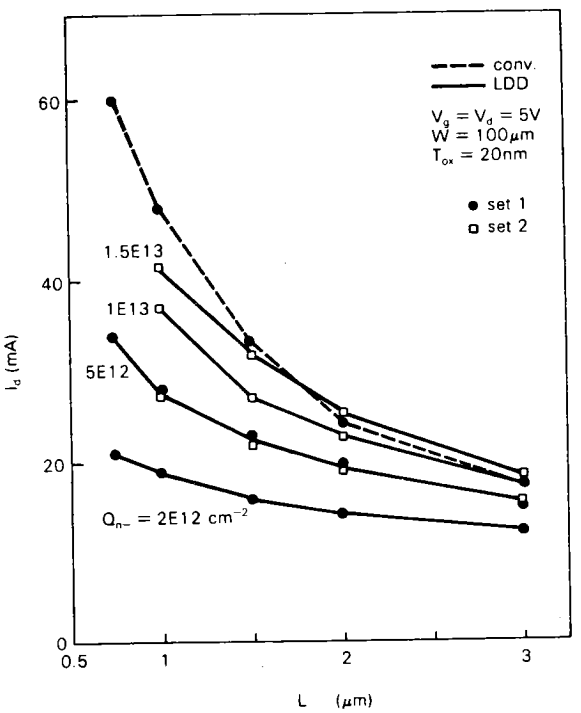


Fig. 9-33 Drain current in LDD MOSFETs with various gate lengths.[55] (© 1984 IEEE).

A report was published in 1985 that compared 1.0 μm non-graded-drain, DDD, and early LDD MOSFET structures.[56] The performance and hot-carrier reliability of the DDD at such gate lengths was somewhat better than that of the early LDD, but it appeared that the DDD would exhibit intolerable short-channel behavior if the supply voltage remained at 5V and the device dimensions were further reduced.

On the other hand, this study indicated that the LDD structure could be designed into smaller MOSFETs without excessively degrading their short-channel behavior. With this in mind, it became important to understand why early LDDs caused performance and hot-carrier reliability degradation. That is, if these drawbacks could be overcome, the LDD could be used to combat hot-carrier effects in MOSFETs with $L<1$ μm. Here we examine in detail the device structure and processing factors that are responsible for the limitations of the early LDD structure. An understanding of these issues also provides a background for the upcoming discussion of modifications to the basic LDD structure, as these were developed to overcome its limitations.

9.7.1 Performance Degradation in LDD MOSFETs

We have noted that the drive capability of conventional LDD MOSFETs, even when "fresh" (i.e., prior to hot-carrier stressing) is worse than that of non-graded-drain MOSFETs. That is, the introduction of an LDD is typically accompanied by a 10-20% reduction in I_D. In poorly designed LDD MOSFETs this performance degradation can be even larger. The reasons for such drive current decrease are discussed in this section. Note, however, that the gate/drain overlap capacitance is normally lower in LDD devices (because the LDD region is partially depleted), which recovers some of the reduced circuit performance. A study of the impact of the LDD on the gate-to-drain capacitance of the MOSFET is given in refs. 57 and 58. The overall loss in circuit performance of the combined effects is observed to be between 4-8%.

9.7.1.1 I_D Decrease Due to Added Resistance of n^- Regions. The most obvious reason that the drive current is reduced in an LDD MOSFET is that the n^- regions of the LDD add segments of lower doping along the current path between the source and drain. Hence, a higher total resistance exists along the current path between the source and drain terminals. As described in chap. 4, the series resistance of the MOSFET can be modeled using an equivalent circuit model as shown in Fig. 9-34.[59] Here we see that using the n^- regions is equivalent to adding two parasitic resistors each with a resistance R_{n^-} to the total source/drain resistance. A simple model for computing the value of R_{n^-} was given in section 9.6.1.3. As also noted earlier, this extra parasitic resistance in practice typically causes a 10-20% loss in both drain current and transconductance compared to non-LDD devices. Note that both the source and drain n^- parasitic resistors act to decrease I_D in the linear region, while only the source n^- resistor causes I_{Dsat} to decrease (because the I_{Dsat} is not impacted by a resistance change across the drain-channel junction). Thus, the effect of R_{n^-} on I_D is most pronounced in the linear (triode) region. (Note that an increase in the channel resistance as a result of hot-carrier damage will also cause the largest decrease in I_D in the linear region, and hence this quantity is often monitored when conducting studies of hot-carrier degradation.)

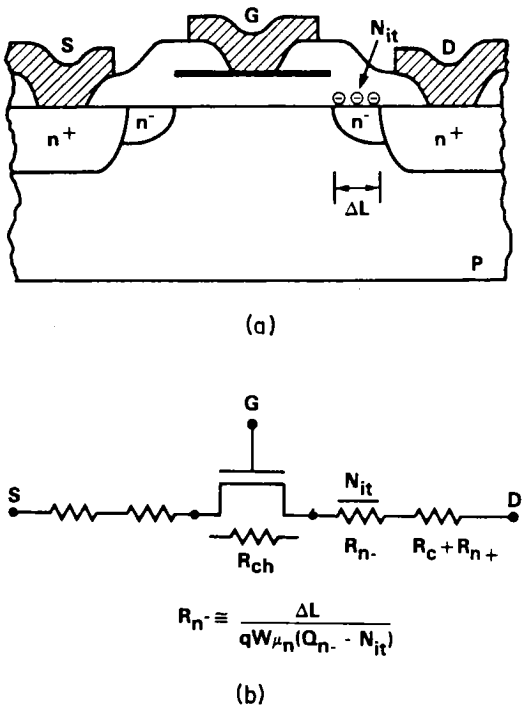


Fig. 9-34 Schematic diagrams showing (a) the two-section drain model and the external channel pinch-off effect, and (b) the equivalent circuit of the graded drain device structures.[59] (© 1984 IEEE).

However, in a device in which the source and drain parasitic resistance values are not identical (i.e., in asymmetric MOSFETs, which are discussed in section 9.7.1.5), the values of I_{Dsat} will not be identical if the source and drain terminals are reversed. As this can pose a problem in some circuits, we will discuss this issue in more detail later.

To minimize the value of R_{n^-} we could adjust several parameters. Primarily the doping concentration of the n^- region could be increased and its length L_{n^-} decreased. However, both of these variations impact the effectiveness of the LDD structure. That is, the optimum value of L_{n^-} has been found by two different studies to be either: (1) the same length as the effective length of the saturation region in a comparable non-graded -drain device, $L_{n^-} \approx l$;[50] or (2) equal to the depletion region width w_D on the n^- side of the p/n^- junction.[60] In practice both of these distances are comparable. This implies that if a longer value for L_{n^-} is chosen, no further reduction in $\mathcal{E}_{ymax}$ will be obtained, but a larger R_{n^-} will result (than would exist if L_{n^-} was shorter).* However,

* As L_{n^-} is increased, initially the depletion layer corresponding to doping in the n^- region and the drain voltage spreads out in order to sustain the drain electric field. This spreading continues until L_{n^-} becomes wide enough to sustain the drain field (i.e., which corresponds to a value of L_{n^-} equal to the depletion layer width in the n^- region). Hence, increasing L_{n^-} beyond l serves only to increase the parasitic series resistance with no further benefit in lateral field reduction.

if the n^+ region intrudes into the n^- depletion region (i.e., if L_{n^-} is shorter than w_D), the effective depletion-layer width is determined by L_{n^-}. Thus, L_{n^-} should be neither too short, nor too long.

Choosing the optimum doping concentration value is more complicated. Upper and lower bounds to the doping can be deduced from the limiting cases. That is, at one extreme, if the doping is too high the n^- region merely becomes an extension of the heavily doped n^+ region, and sufficient reduction in $\mathcal{E}_{ymax}$ is not achieved. At the other extreme, a very lightly doped n^- region hardly changes the structure from that of a conventional MOSFET with a high $\mathcal{E}_{ymax}$ at the heavily doped drain. Selecting the optimum doping between these two extremes involves various tradeoffs between performance and hot-carrier degradation. The relationship between these two characteristics is not straightforward, as will be discussed in section 9.10.

In practice, in early LDD MOSFETs the n^- dose was generally chosen to be in the range of $5x10^{12}$ cm^{-2} to $1x10^{13}$ cm^{-2} (which is roughly equal to a concentration of $2\text{-}4x10^{17}$ cm^{-3} in an n^- region of 0.25 μm in depth), as these values resulted in MOSFETs with the smallest I_{sub} values as well as I_D values that were about 20% below those of conventional 1 μm MOSFETs (see Fig. 9-33, for the conventional MOSFET and the LDD MOSFET at L = 1 μm and n^- dose = $1x10^{13}$ cm^{-2}). More recently, as is described in section 9.11, the n^- dose in modified-conventional LDDs has been increased to $4x10^{13}\text{-}1x10^{14}$ cm^{-2}. However, in such fully-overlapped LDD structures as the IT-LDD (proposed for 0.5 μm MOSFETs), the n^- dose has been decreased back to the lighter range. Some of the reasons for these choices will be considered in the upcoming section that deals with designing LDD MOSFETs for high hot-carrier reliability.

9.7.1.2 Doping Compensation Effects that can Increase R_{n^-} in Deep Submicron LDD MOSFETs. As device dimensions shrink, the doping concentration in the channel must be increased to achieve sufficiently high values of V_T and punchthrough voltage in submicron devices. At L = 0.5 μm, the peak channel doping concentration is estimated to be ~$2x10^{17}$ cm^{-2}. Thus, it is possible for this doping to compensate the n^- doping such that R_{n^-} is significantly increased (which in turn will decrease I_D).[61] It may be necessary to develop process sequences that somehow prevent the channel dopants from being introduced into the n^- regions. One such approach is described in section 9.12.1.3.

9.7.1.3 Increase of R_{sh} by Weak Overlap of the Gate and Drain. If the edges of the drain and gate are offset, a gap exists between the gate edge and the drain. A segment of the channel between source and drain thus extends over this gap. This condition (as well as the condition of only slight-edge-overlap between gate and drain), is referred to as *weak overlap*. The condition of weak overlap can lead to a degradation of both the performance (drive current) and hot-carrier reliability of a MOSFET. In this section, we former issue, and in section 9.7.2 the latter.

The decrease in drive-current performance arises because weak overlap results in an

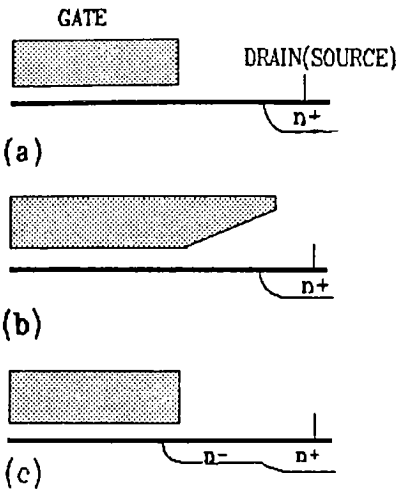


Fig. 9-35 Weak overlap occurs as a result of three conditions in a MOSFET: (a) offset of the drain edge from the gate edge in a single-drain MOSFET; (b) by a *gate bird's beak* (or *graded-gate oxide*, GGO; and (c) offset (or slight overlapping) of the gate and drain edges in an LDD. Note that gate offset and GGO may occur simultaneously, which can increase the severity of the weak overlap condition.[62] (© 1986 IEEE).

increase in R_{sh}. This effect occurs because the segment of the channel in the gap is under less control by the gate than is the channel region at the center of the gate. Consequently, this section is not inverted as strongly and the total channel resistance is increased. Basically, weak overlap[62] occurs as a result of three conditions in a MOSFET: (1) offset of the drain edge from the gate edge in a single-drain MOSFET (Fig. 9-35a). Such offset may occur for several reasons, as we shall discuss later; (2) weak overlap caused by a *gate bird's beak* (or *graded-gate oxide*, GGO, see Fig. 9-35b); and (3) offset (or slight overlapping) of the gate and drain edges in an LDD (Fig. 9-35c). Note that gate offset and GGO may occur simultaneously, which can increase the severity of the weak overlap condition.

9.7.1.4 The Impact of Poly Re-Ox on Weak Overlap in MOSFETs.

As described in chap. 5, the poly re-ox step grows a sidewall on the patterned poly gate and also gives rise to a graded-gate oxide, or GGO (9-35b). Weak overlap can be caused in MOSFETs as a result of both oxidation effects.[63,64] If the poly gate has been doped prior to re-ox, the poly sidewall oxide will grow much more rapidly than on the more lightly doped substrate regions (e.g., if 150Å of oxide is grown on the substrate about 500Å of oxide will grow on the doped poly). Thus, even if the gate and drain are overlapped prior to a poly re-ox step, the doped gate edge can retreat more rapidly (as a result of the rapid oxidation of the poly sidewall) than can the diffusing drain dopants which try to catch this retreating edge. As a result, the drain edge may end up being offset from the gate edge after re-ox even if alignment had existed prior to the process.

As noted in chap. 5, the GGO can provide several benefits to a MOS device, including improved gate-oxide edge integrity and lower gate-to-drain overlap capacitance. For example, the poly gate etch process may form microtrenches in the gate oxide at the periphery of the gate edge, causing all the oxide there to be removed (see Fig. 7-29d). The re-ox step can heal such damage. The heavy source/drain implant can also damage the gate oxide, even if it is covered by the CVD spacer.[106] A moderate re-ox process (which gives rise to a minor GGO) is effective in healing the former and preventing the latter damage. However, if a more extensive re-ox process is performed, i.e., in which the lateral penetration and increase in oxide thickness at the gate edge of the GGO is excessive, the performance of the device can be degraded by causing a weak gate/drain overlap effect (for different reason than is caused by the lateral offset of the gate and drain). That is, even if the drain edge ends up beneath the gate edge, the thicker oxide at the gate edges caused by GGO reduces the degree of inversion in the channel at the gate edge compared to that in the center of the gate. Since the lateral penetration of the GGO increases as the GGO thickens at the gate edge, these two effects can combine to worsen the reduction in the value of the drain current (or transconductance, as shown in Fig. 9-36). Here we see that the edge of the drain must be driven further under the gate edge (i.e., λ must be increased) to negate the effect of the GGO on transconductance.

A number of measures have been proposed to minimize the extent of poly oxide growth and GGO during poly re-ox. First, a short re-ox time and a relatively low oxide growth temperature should be chosen. Second, the re-ox step should be performed in a dry rather than a wet oxidation ambient. Finally, the poly can be doped after the re-ox step, namely during the n^+ implant step that is used to doped the source/drain regions (rather than before gate patterning, as is usually done).[65,66] The latter approach is effective because undoped poly oxidizes less rapidly than does heavily doped poly.

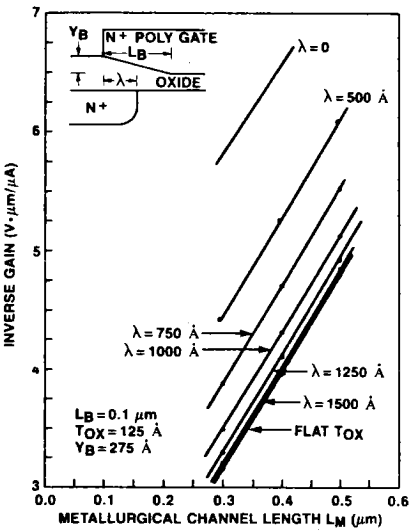


Fig. 9-36 Simulated results of inverse gain versus metallurgical channel length for different offset distances λ with L_B = 0.1 μm and Y_B = 275Å.[66] (© 1989 IEEE).

9.7.1.5 Off-Axis Implant Effects in MOSFETs, Leading to Weak Overlap Problems and Asymmetrical MOSFET Device Characteristics. Formation of the source/drain regions of a MOSFET is accomplished by means of ion implantation, with the polysilicon gate used as the implant mask. To produce shallow junctions the implant has traditionally been done off-axis (typically, at an angle of 7°), as this avoids channeling (see Vol. 1, chap. 9). However, an off-axis implant can also cause asymmetric doping of the source and drain regions.[68-72,76] That is, lateral shadowing (S in Fig. 9-37) of the substrate occurs on one side of the poly gate, and penetration of dopants through the leading corner of the poly on the other side. The impact of off-axis implants on the device characteristics can be examined by considering the n^- and n^+ implants separately.

If the implant that forms the lightly doped region of the drain is done off-axis, the as-implanted region in the shadow of the gate will have its edge displaced from the gate edge. As a result, there will be no overlap of the implanted region and this gate edge upon the conclusion of the implant. Even after a drive-in step, the overlap of this edge with the gate will be less than the overlap of the nonshadowed implant region edge and the gate. If the less-overlapped side is used as the source, the extra resistance introduced will reduce I_{Dsat} (I_{Dsno} in Fig. 9-37), thus degrading circuit speed. However, if the drain side has less overlap, the extra resistance will not affect I_{Dsat} (curve I_{Ddno} in Fig. 9-37). The asymmetry in drive current when the same device is connected in these two ways

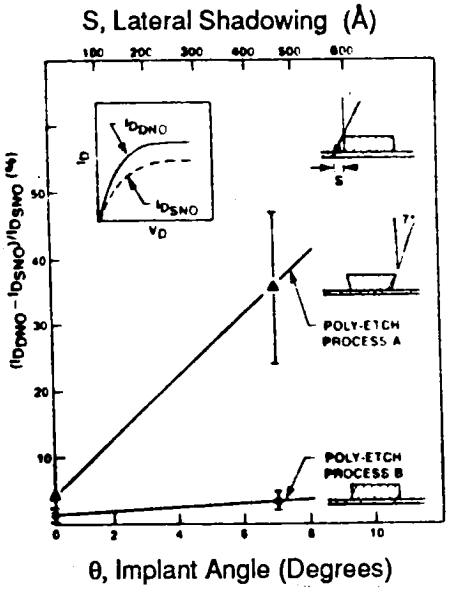


Fig. 9-37 Percent asymmetry in saturated drain currents ($V_{DS} = V_{GS} = 5V$) versus source/drain implant angle for LDD NMOS devices having reentrant- and vertical-wall polysilicon gates. Insert schematically shows source-non-overlap (S_{NO}) and drain-non-overlap (D_{NO}) I-V behavior.[67] Reprinted with permission of Semiconductor International.

can be as much as 40% and this can be catastrophic in circuits that require closely matched device characteristics.

The asymmetrical I_D-V_{DS} problem is severely worsened if the poly gate etch produces a structure with re-entrant angled sidewalls (rather than vertical sidewalls) as is shown in Fig. 9-37.[67] Thus, if an implant process with a 7° tilt angle is used, an etch process that produces vertical sidewalls on the poly (or polycide) is critical. This means that unless a 0° tilt angle is used in the implant processes, the etched profiles of the gate electrodes must also be monitored (and not just in the top views used to measure critical dimension, but in cross-section to expose re-entrant angles).

The problems arising from such n^- implant-shadowing can be alleviated by various measures, including the following:

1. A poly-etch process that produces vertical sidewalls in the poly gates should be used.
2. Excessive reoxidation of the polysilicon gate after it has been patterned should be avoided. (Note that minimizing re-ox also prevents a gap from being established between the drain region edge and the edge of the gate).
3. A vertical implant (0°) through a screen oxide can be used to reduce channeling, instead of using a 7° tilt implant.[27]
4. A poly-etch process that produces a slight positive bevel to the poly sidewall profile can be combined with a smaller angle (3° vs 7°) off-axis implant.[68]
5. The implant can be carried out at 7°, but in two segments, each at half the total dose. The wafer is repositioned by 180° after the first segment of the implant, so that each edge of the gate is the leading edge of the gate during one-half of the total implant, with respect to beam shadowing. In this way the implanted region on both sides of the gate will be symmetrical.

Let us next examine how the n^+ implantation can also impact the structure of the LDD.[73,75] If this implant is also performed at a 7° off-axis, additional asymmetric implant effects can occur. That is, on the side of the gate that is not shadowed, the arsenic atoms penetrate the leading corner of the spacer and thereby enter the substrate under the spacer to some extent. If the 7° tilt angle of the n^- region was from the opposite direction, the n^- region at the drain will have been offset from the gate. Therefore, if the spacer is too narrow, the tilted n^+ implant can wipe out the lightly doped drain-extension region. When this happens, the device can entirely lose the hot-carrier protection that was to be provided by the LDD structure, and it will become vulnerable to degradation by hot-carrier effects. To avoid this, a spacer with sufficient width must be employed, and a higher n^- dose will also offer more protection. Some guidelines for choosing the proper spacer width to deal with this problem are given in reference 74.

9.7.1.6 Implanter Issues that Impact the LDD Process Sequence. Medium current implanters are used to perform the n^- implants. In early machines the wafer holder could be inclined so that the beam struck the wafer surface at a fixed, non-0° angle. However, electrostatic scanning was typically the method used to deflect the beam over the wafer surface, and this causes the beam angle at the wafer surface to vary (e.g., from 3° to 9° when the nominal beam angle is 7°), as the beam is scanned. The n^+ implant is carried out in high-current implanters which use fully mechanically scanned beams. Mechanical scanning reduces the beam scan-angle variations that occur with electrostatic scanning. (In some cases the n^- implant dose was actually carried out at 0° on high current implanters to try to take advantage of this benefit.)

Later generations of medium current implanters have been introduced which permit more options. First, the scanning system has been modified so that reduced scan-angle variations are caused. Second, they offer the capability of angular repositioning of the wafer so that both edges of the gate of a MOSFET can be implanted with an off-axis implant to achieve symmetrical doping of both the source and drain sides of the device.

9.7.1.7 Process integration issues which arise when an LDD is integrated into a CMOS Technology. If an NMOS LDD structure is incorporated in a CMOS technology some issues must be addressed that do not arise in NMOS technology. Specifically, an extra mask is needed if it is desired to prevent the n^- implant dopants from being introduced into the source/drain regions of the PMOS devices. This is especially necessary if an LDD structure is also to be used in the PMOS devices. On the other hand, since the p^+ source/drain implant dose is heavy enough so that any n^- dopant implanted into the PMOS source/drain regions will be compensated by that dose, if no LDD is needed in the PMOS device, a blanket n^- implant can be carried out without the need for an extra mask step.

In the latter case, however, the p^+ source/drain implant step must be performed prior to the formation of the oxide spacer (to prevent the p^+ dopants from being offset from the gate edge by the spacer). This means that the p^+ dopants will be subjected to the anneal steps of the n implants. As a result, the depth of the PMOS source/drain junctions will be greater than if it was not necessary to subject them to this anneal cycle. In section 9-8 we will discuss how the use of disposable spacers is one possible way to circumvent this problem.

9.7.2 Hot Carrier Reliability Issues in Early LDDs

In early LDD MOSFETs, the criterion for selecting an optimum device structure for hot-carrier resistance was frequently based on monitoring I_{sub} for a given set of bias conditions and choosing the structure that exhibited the smallest I_{sub} value. This was done because the smallest I_{sub} value is correlated to the smallest value of $\mathcal{E}_{ymax}$. [If a tradeoff was made between low substrate current and high drain current, the devices which exhibited the lowest I_{sub}/I_D ratios were deemed best (see Eq. 9-5, which indicates that this ratio is an indicator of the $\mathcal{E}_{ymax}$ value).]

Although this assumption may be correct, unfortunately it is not always the case that devices with smaller $\mathcal{E}_{ymax}$ necessarily exhibit better hot-carrier reliability. In fact, as

discussed previously, some early LDDs which apparently had smaller $\mathcal{E}_{ymax}$ values than comparable non-graded-drain devices actually exhibited *worse* susceptibility to degradation after hot-carrier stressing. Hui and Moll reported that unless some minimum overlap between the gate and the n^- region of the drain was ensured, the LDD device would exhibit poor hot-carrier reliability.[77] They noted that LDD devices with extra thermal drive-in demonstrated significantly less hot-carrier degradation, and they surmised that the discrepancies in hot-carrier reliability among early LDD stressing reports was likely due to the different temperature cycles used in the various LDD devices.

It was eventually learned that a phenomenon dubbed the *structure-dependent degradation effect* could exacerbate hot-carrier degradation even as $\mathcal{E}_{ymax}$ was being reduced.[55,78] Actually, two effects contributed to the problem. First, the location of $\mathcal{E}_{ymax}$ is dependent on the doping concentration in the n^- region. That is, $\mathcal{E}_{ymax}$ moves closer to the edge of the n^+ region of the drain if the n^- implant dose is low (e.g., $5x10^{12}$-to-$1x10^{13}$ cm^{-2}, which was the dose typically chosen in early LDDs). Second, since the damage due to hot-carrier injection is localized in the gate oxide above the location of $\mathcal{E}_{ymax}$ this would have placed the damage in the spacer region of the channel (see Fig. 9-38b). Since this segment of the channel is outside of the edge of the gate, it will tend to already be less heavily inverted. The accumulation of damage-induced negative charge in the oxide above this location has the effect of reducing the level of inversion even more, thus causing the parasitic resistance to increase. In turn, the drain

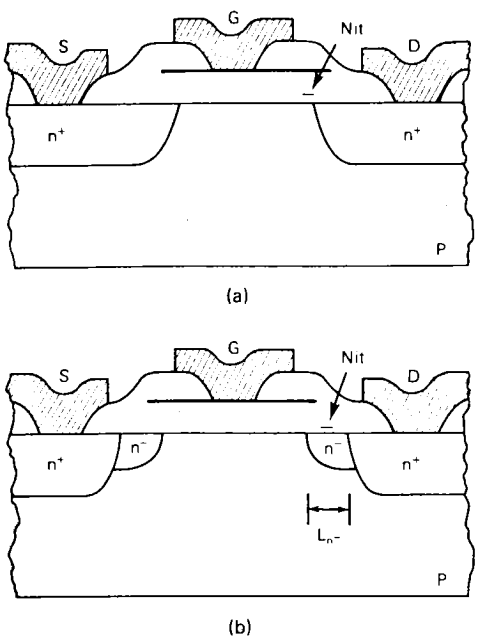


Fig. 9-38 Degradation in early LDDs (b) was above the drain current path beyond the edge of the gate. In conventional MOSFETs (a), the edge of the drain is under the gate.

current (and V_T and g_m) are decreased. (In extreme cases the charge buildup can cause the segment to be completely depleted, cutting off the device.)[55,78] This can cause the hot-carrier lifetime to decrease even more rapidly than in a non-graded-drain device, in which $\mathcal{E}_{ymax}$ is located beneath the gate edge. Another issue that compounds the problem is that the deposited oxide in the spacer region may not have as a good a quality as the thermal oxide beneath the gate, and thus may be more susceptible to damage by the injected hot carriers.

One simple cure for the above problem is to use a somewhat higher n^- dose. Although the price is a somewhat higher value of $\mathcal{E}_{ymax}$, this also has the effect of moving the location of $\mathcal{E}_{ymax}$ to the p/n^- junction, which is also beneath the gate edge. Here, the gate field counteracts the damage in the oxide, resulting in less channel depletion than before. Also, the heavier doping makes it harder to deplete the n^- region. These effects combine to make the device much more resistant to hot-carrier degradation. Devices with such higher n^- doses are referred to as moderately doped LDDs (M-LDDs), and are discussed further in section 9.11.2.1. Another way to guaranty that this problem does not occur is to fully overlap the n^- region with the gate. Devices have been designed to accomplish this, and we will also describe them later (e.g., fully overlapped structures: section 9.12.1, and LATID structures: section 9.12.2).

9.8 DISPOSABLE SPACERS

Instead of forming the oxide spacer after the n^- implant, and then doing the n^+ implant, an alternative process sequence can be used to form the LDD. That is, the spacer can be formed first, followed by an n^+ implant. Then the spacer is removed and the n^- implant is performed (hence the name *disposable spacer*). The benefits of disposable spacers apply mainly to CMOS technology, and they include the following: (1) the process becomes more flexible, allowing LDD formation in both or either NMOS and PMOS devices; (2) If both types of LDD structures are required they can be formed with only two masks (details later), whereas 3 or maybe even 4 masks are needed if the conventional sequence is used; (3) the *p*-implants for the PMOS device can be performed after the *n*-implants are annealed. Shallower source/drain junctions in PMOS devices can thus be fabricated since the *p*-implants do not have to be subjected to the *n*-implant anneal cycle; (4) As noted earlier, in the conventional LDD sequence the silicon substrate may be damaged during the overetch time of the oxide spacer etch formation process. Such damage has been blamed for junction leakage.[79] This problem is avoided in the disposable spacer process, since the damage is prevented during the disposable spacer etch (see below). Spacer etching is also simplified, because a sharp endpoint signal is easily detected as a large area is cleared at the end of etching, and; (5) The damage from the n^+/p^+ source/drain implant steps can be annealed prior to LDD implant step, thus allowing shallower n^- and p^- regions to be formed. Shallower junctions result in improved short-channel behavior.

The first disposable spacer process to be published used a polysilicon disposable spacer.[80] Here, a layer of polysilicon (instead of the conventional LTO oxide) is deposited on the reoxidized, patterned poly gate structure. The polysilicon is anisotropically etched back to form a poly spacer. This polysilicon etch, which has high

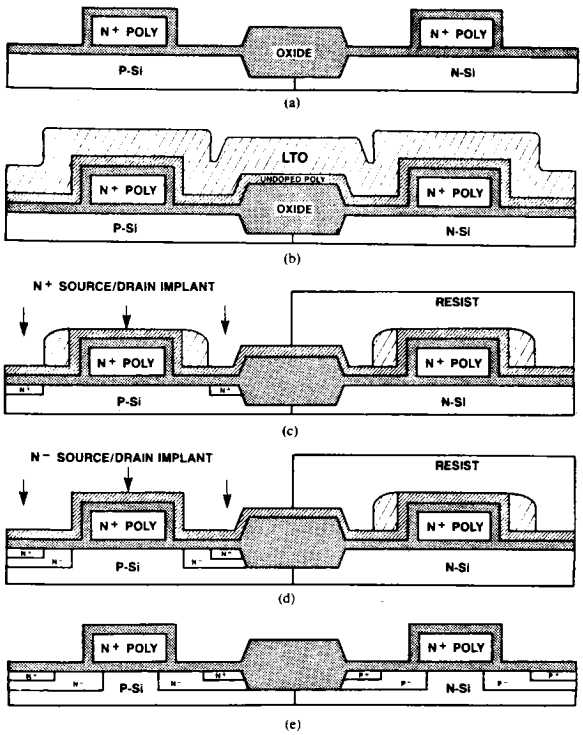


Fig. 9-39 Key processing steps for the NMOS device using the disposable sidewall spacer technology: (a) after poly etch and reoxidation: (b) after the 300-Å undoped poly and undoped LTO deposition; (c) after RIE of the LTO to form the sidewall spacers, which is followed by the n^+ source/drain implant with resist covering the PMOS areas; (d) with resist in place, the LTO sidewall spacers are stripped in 10:1 BOE and the n^- source/drain implant is performed; and (e) after the poly buffer is removed by etch or thermal oxidation.[81] (© 1988 IEEE).

selectivity to oxide, stops on the thin oxide over the source/drain regions without damaging the underlying silicon. The polysilicon spacers are removed after the n^+ implant, and the n^- implant is carried out. Note that in this process there is also virtually no loss of field oxide during the spacer formation etch step because of the high selectivity to the field oxide.

The next reported disposable spacer process was a *disposable oxide-sidewall-spacer process*,[81] in which a thin poly layer (300Å thick) was formed by LPCVD between the re-oxidized poly gate and the LTO spacer oxide. The thin poly under the spacer acts as an effective etchstop layer if a spacer-etch process with high etch selectivity of LTO to polysilicon is used. The oxide spacers are removed after the n^+ implant by a wet chemical etch step (10:1 BOE) which also does not attack the thin poly layer that covers the gate structure (see Fig. 9-39).[81] The thin poly layer can later be removed by wet or dry etching, or it can be thermally oxidized prior to the completion of the process. A

variation of this process that uses CVD silicon nitride in place of the LTO oxide as the disposable spacer material has also been described.[82] Since hot phosphoric acid is used to strip the nitride spacers, fewer defects occur in the re-ox oxide layer beneath the thin poly protective layer than when the BOE spacer removal etch process is used to remove the disposable LTO spacers. A thinner re-ox oxide can thus be used. This is advantageous as less gate-bird's beak (GGO) is formed in the gate oxide at the edges of the poly gate during the re-ox process.

If a salicide process is to be included in a CMOS technology, the disposable polysilicon spacer fabrication sequence described earlier must be modified in several ways. First, an additional permanent oxide spacer must be formed after all the LDD implant steps are completed (to permit the silicide to be formed on both the source/drain and poly gate regions without shorting between them). A sequence for fabricating a dual-gate CMOS technology (i.e., p^+ poly for PMOS and n^+ poly for NMOS) with disposable poly spacers is shown in Fig. 9-40.[84] To permit this technique to be used when devices are scaled to 0.5 μm an additional modification is made. That is, a thin CVD TEOS oxide is deposited on the re-oxidized polysilicon gate before the poly spacer layer is deposited. This is done because a sufficiently thick oxide layer must exist under the poly spacer film (i.e., at least 150Å) to prevent damage to the substrate during the

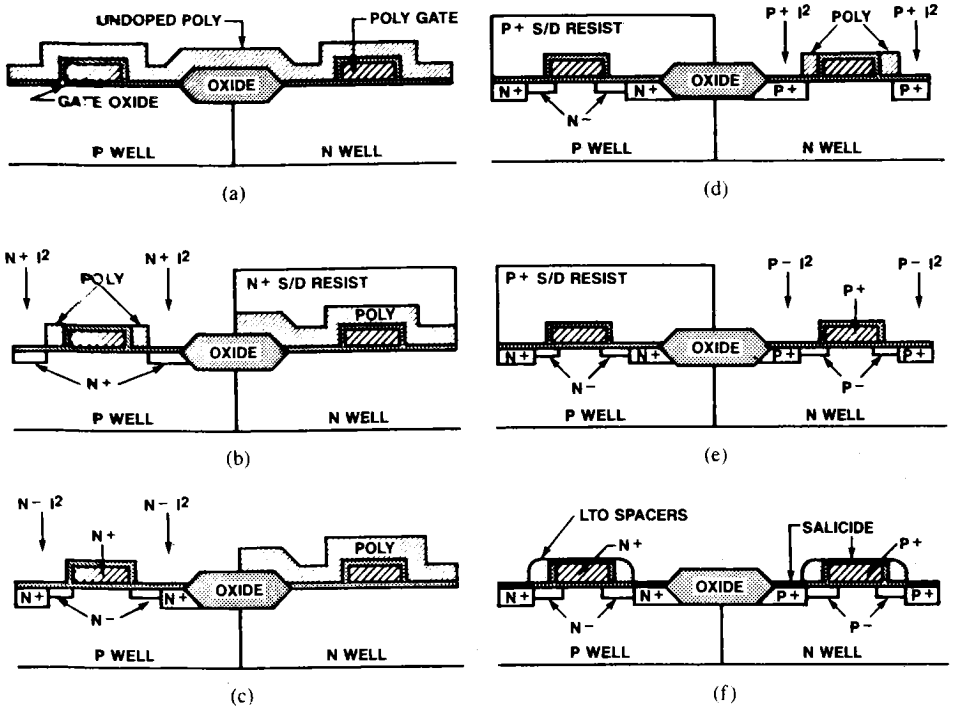


Fig. 9-40 Disposable polysilicon spacer process.[84] (© 1991 IEEE).

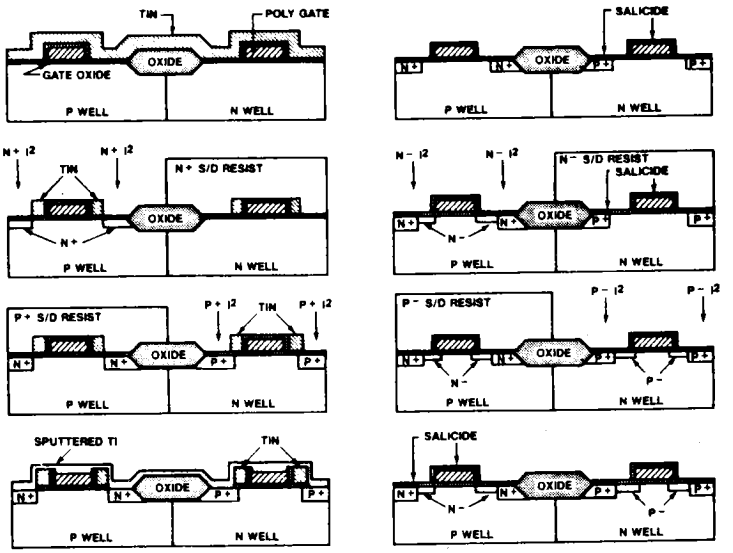


Fig. 9-41 Disposable TiN spacer process.[85] (© 1989 IEEE).

overetch time of the spacer formation process. If such a thick oxide film was grown during the poly re-ox step, the gate birds beak that would result would penetrate too far into the channel in devices (because of the heavy doping of the poly gate) whose gate lengths approached 0.5 μm. The use of the CVD TEOS film in combination with a thinner thermal oxide film formed during poly re-ox will sufficiently suppress the gate bird's beak and yet provide an adequately thick oxide under the poly spacer. Also note that to permit the dual gate technology to be implemented, the poly over the PMOS regions is not etched back to form spacers until both the n^+ and n^- implants are performed.

However, a technique has also been reported that allows a single disposable spacer to be used to form both the LDD and the salicide regions. But this approach requires the use of a different disposable spacer material than polysilicon, namely TiN. The disposable TiN spacer process sequence is shown in Fig. 9-41.[85] A technique that uses $CoSi_2$ disposable spacers has also been described.[86]

9.9 SUMMARY OF THE LIMITATIONS OF EARLY LDD DEVICES

In many cases, early LDD devices were able to provide improved hot-carrier reliability compared to conventional single-drain MOSFETs. However, these early LDDs also exhibited many drawbacks, including the distinct possibility that if they were improperly designed, or if the manufacturing processes were not tightly controlled, their

hot-carrier reliability could no better (or even worse) than that exhibited by conventional MOSFETs. The following reasons summarize why early LDD devices were prone to such variability.

First, the n^- implant dose in early LDDs was too low, which degraded the drain current more than necessary, even prior to hot-carrier stressing. Second, this low dose caused the location of $\mathcal{E}_{ymax}$ to be outside of the edge of the gate, and also made the n^- region under the spacer susceptible to depletion by the hot-carrier damage that was highest above the point of $\mathcal{E}_{ymax}$. It was possible for the latter two effects to combine and severely degrade the hot-carrier reliability of the device. Third, the above problems were exacerbated by the fact that the 7° tilt angle used in the LDD implants, together with non-optimized GGO, caused significant asymmetrical I_{Dsat} characteristics in 1 μm and smaller devices. Finally, the optimum field-reduction-factor (FRF) in the conventional LDD could not be achieved as this would have required a fully overlapped LDD structure. To achieve a fully overlapped n^- region, the vertical depth of the n^-/p junction would have ended up being too deep to control short-channel effects in submicron MOSFETs. For all these reasons, modifications to early LDD structures were needed to make them capable of providing hot-carrier protection in sub-micron devices. The design and fabrication effort undertaken to achieve such hot-carrier- resistant device structures in submicron devices has been termed *drain engineering*.

9.10 SUMMARY OF THE METHODS TO BE USED IN DESIGNING EFFECTIVE LDD STRUCTURES

Some of the weaknesses exhibited by early LDD structures arose because the guidelines used to design them were based on erroneous or incomplete understanding of the behavior of the structure. That is, it was believed that the most important issue with respect to hot-carrier reliability was to reduce $\mathcal{E}_{ymax}$ as much as possible. Simulations were carried out to calculate $\mathcal{E}_{ymax}$ for various values of L_{n^-} and n^- dose, and the drive currents of MOSFETs with the corresponding n^- regions were also computed. Device designs based on the values of n^- and L_{n^-} chosen from these analyses were fabricated, and I_{sub} was measured as a monitor of $\mathcal{E}_{ymax}$ to ensure that sufficient hot-carrier stability was indeed achieved from these choices. However, it was learned later that such factors as the location of the $\mathcal{E}_{ymax}$ point, and the "structural degradation effect" also played an important role in how much the LDD increased the lifetimes of the devices. Consequently, merely knowing $\mathcal{E}_{ymax}$ and I_{sub} in a device was not enough to determine whether an LDD was optimally designed.

Instead, optimization of an LDD MOSFET involves many device parameters besides the n^- dose, implant energy, and the value of L_{n^-}. These include the value of the offset of the n^- dose from the gate edge as a result of the re-ox step, the thermal budget after the n^- implant, and the channel doping (the latter two determine the gate-to-n^- overlap). Selecting the optimum set of these parameters to obtain the most effective LDD structure for the device at hand is a complex task.

The dilemma faced by process and device engineers is to how to devise an optimized drain structure in the face of all the possible choices. From a realistic perspective only a reduced number of the alternatives can possibly be investigated, partly due to the time-

consuming lifetime stress tests. These tests constitute an essential input and represent perhaps the ultimate criterion for selecting the best structure among the different possibilities. However, the necessity of conducting such lifetime tests runs up against a major obstacle. That is, the time pressure of shorter device-design cycle times conflicts with the need to investigate a broader spectrum of drain structures since the hot-carrier sensitivity increases as the devices continue to shrink. The only way to be able to carry out lifetime tests on a wide spectrum of drain structures in a short period of time when designing a next-generation MOSFET is to rely on simulation results to identify the most promising alternatives.

Such simulation, however, must be able to provide information beyond the values of $\mathcal{E}_{ymax}$ and I_{sub} that were the outputs of the early simulators. Such information permitted only the physics of degradation to be better understood and/or rough technological choices to be made. Unfortunately, this usually did not provide sufficient data to allow the degradation behavior of the actual devices to be correctly predicted. That is, an increase in $\mathcal{E}_{ymax}$ or I_{sub} does not necessarily translate into a lifetime reduction. Transistor degradation is also related to the damage of the gate oxide and Si-SiO_2 interface near the gate edge, both of which arise from hot-carrier injection. For an improved optimization of the drain structure, such effects require more evaluation than merely the value of $\mathcal{E}_{ymax}$. That is, the physical process which causes the degradation must also be considered in the simulation. A self-consistent calculation of the amount and position of the oxide damage is required in order to be able to carry out a direct evaluation of the device lifetime.

One such modeling tool for carrying out the above simulation task was developed by Hansch et al. in 1991.[87] This tool allows the time development of the device degradation under a dc stress condition to be directly monitored. It is able to predict the location and amount of gate oxide damage during the stress experiment by calculating the injection of carriers into the oxide region along the direction from the source to drain and by solving the continuity equation to obtain the carrier distribution in the oxide. This simulation routine was incorporated as a module of the device simulator MINIMOS IV. With such a simulation capability it becomes possible to direct and concentrate experimental efforts toward better devices while simultaneously helping to reduce development time and cost.

9.11 IMPROVED CONVENTIONAL (OXIDE-SPACER-BASED) LDD STRUCTURES

In section 9.6.1.2 we indicated that to optimize the hot-carrier resistance of the LDD structure, the n^+ region should be located at the edge (or just under the edge) of the gate, and the n^- region should be located entirely under the gate. Unfortunately, this optimum device structure is not easy to fabricate. That is, achieving full n^- overlap in an oxide-spacer-based LDD dictates the use of a very narrow spacer together with a substantial thermal drive-in. Unfortunately, in submicron CMOS technologies this approach would cause the S/D junctions in both the NMOS and PMOS transistors to be too deep to control short-channel effects. In fact, in the limit of very short spacer widths the LDD

would essentially revert back to being a DDD structure, making it subject to the same minimum channel-length constraints.

While alternative device structures with fully overlapped n^- regions have been developed, these require quite complex manufacturing sequences. Nevertheless, when gate lengths shrink below $L = 0.5$ μm their implementation may become necessary. Thus we will consider such alternative structures in later sections.

On the other hand, oxide-spacer-based LDDs with only *partial* n^--to-gate overlap have also been developed that provide adequate hot-carrier lifetimes for gate lengths that approach 0.5 μm. Such partial-overlap LDDs are less complex to fabricate than the full-overlap structures, but are no more susceptible to short-channel effects than early LDDs. Hence, some of them have been designed into 0.6-1.5 μm CMOS technologies. Here we describe such *modified oxide-spacer-based LDDs.*

The evolution of modified LDD devices has involved the following three approaches to improving characteristics of the conventional LDD:[90]

1. The doping concentration in the n^- region of the LDD structure is increased to improve the drive-current capability and to reduce the "structural degradation effect" that worsens the hot-carrier lifetime of the conventional LDD. The latter problem is alleviated by this approach because the location of $\mathcal{E}_{ymax}$ apparently moves beneath the edge of the gate when the doping concentration in the n^- region is sufficiently increased. The structures that exploit this approach are the *moderately-doped LDD* (M-LDD) and the *metal-coated LDD* (MLD).

2. The vertical profile of the n^- region is modified in an attempt to steer the path of current away from the surface of the MOSFET and to separate the location of $\mathcal{E}_{ymax}$ and the path of maximum current flow in the device. The structures that pursue this approach are the *buried LDD* (B-LDD), the *profiled LDD* (P-LDD), and the *sloped-junction LDD* (SJLDD).

3. Dopants are selectively implanted beneath the n^- region to reduce the susceptibility of submicron LDD structures to short-channel effects. The structure that pursues this approach is the *double-implanted LDD* (DI-LDD).

9.11.1 Increasing the Doping in the n^- Region of the LDD

9.11.1.1 Moderate Lightly Doped Drain (M-LDD). The first approach to improving the conventional LDD involves only an increase in the doping of the n^- region of the drain. This was first proposed by Kinegawa in 1985, who dubbed the resulting device the *moderate lightly doped drain* (or M-LDD) structure.[88] Another report detailing the characteristics of the M-LDD was published by Krieger et al. in 1991.[89] The spacer width and thermal budget after n^- implant are basically the same as in the conventional LDD, but the dose of the n^- implant is increased from $5x10^{12}$-$2x10^{13}$ cm^{-2} to $4x10^{13}$-$1x10^{14}$ cm^{-2}.

The increased n^- doping implies that $\mathcal{E}_{ymax}$ is not reduced as much as in the conventional LDD, but several problems of that structure are overcome such that the

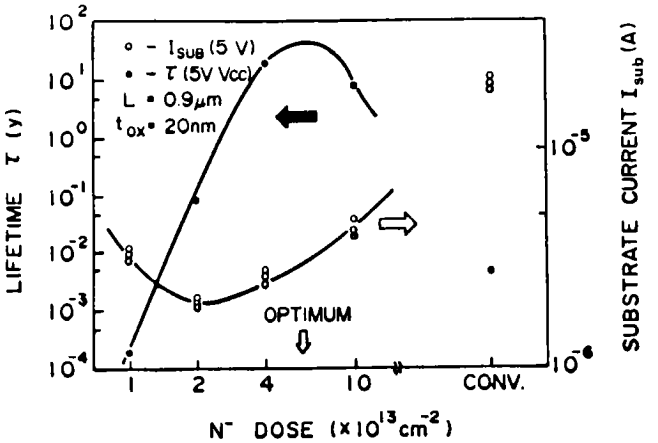


Fig. 9-42 (a) The dependence of device lifetime and substrate current level on the n^- region implant dose in the LDD device structure. The minimum substrate current point is at $2x10^{17}$ cm^{-2} and the maximum lifetime point is at $5\text{-}6x10^{13}$ cm^{-2}.[88] (© 1985 IEEE); (b) Device structure of the metal-coated LDD (MLD).[91] (© 1985 IEEE).

M-LDD exhibits increased hot-carrier resistance and improved drive current. Figure 9-42a shows the lifetime τ and I_{sub} of the LDD device as a function of the n^- implant dose. It is seen that τ in the M-LDD structure can be longer than that of a conventional LDD (in which $n^- = 1x10^{13}$ cm^{-2}), despite the fact that I_{sub} is larger. In fact, at an optimal n^- doping ($n^- = 5\text{-}6x10^{13}$ cm^{-2}), the M-LDD exhibits four to five orders of magnitude improvement in lifetimes over the conventional LDD with channel lengths down to 0.8 μm (for $V_{DD} = 5V$).[88] It is believed that this structure can be effectively used for devices with L_{eff} down to 0.6 μm.[89] Indeed, the advantages of the M-LDD are so great that to a large extent it has been incorporated into many devices (see for example ref. 81) - which are nevertheless generally still classified as LDD, but are actually M-LDD.

The benefits of increasing the n^- region doping in the MLDD arise for several reasons. First, the sheet resistance of the n^- region is lower than that of the light LDD, even prior to stressing. This improves the M-LDD current drive over that of lighter dosed LDDs. Second, the value of the current drive in the M-LDD is apparently much less sensitive to variations in the oxide spacer thickness than in the light-dose LDD. Third, the increased doping apparently moves the location of $\mathcal{E}_{ymax}$ under the gate edge, such that the damaged region in the oxide also lies beneath the gate edge, rather than outside the gate-overlap area as occurs in the light-dose LDD. Thus, the structural degradation effect that leads to more rapid hot-carrier degradation in light-dose LDDs is overcome. Finally, it should be noted that while gate-n^- overlap in the M-LDD is greater than in the light-dose LDD (in Krieger's report this overlap was found to be ~0.15 μm), it appears that the resultant increase in gate-to-drain capacitance does not significantly degrade the gate delays if an M-LDD is used in place of a light-dose LDD in such digital circuits.

9.11.1.2 Metal-coated LDD (MLD). The *metal-coated LDD* (MLD) is another oxide-spacer-based LDD structure that increases the doping in the n^- region above the levels used in light-doped LDDs.[91] Here, however, a deeper phosphorus implant dose is used than in the M-LDD and a silicide (e.g., $TiSi_2$) is also formed atop the n^+ source/drain regions (Fig. 9-42b). The deeper n^- implant steers the path of maximum current somewhat away from the position of maximum field at the surface. This reduces the impact ionization rate. As a result, a smaller I_{sub} occurs than is observed in light-dose LDDs with the same gate lengths. The silicide layer reduces the sheet resistance of the drain even further.

It has been shown that the MLD structure can sustain 5-V operation in MOSFETs with $L_{eff} = 0.6\ \mu m$, and the structure has reportedly been incorporated into the ATT Twin-Tub CMOS V technology.[92] In ref. 92 several issues were also raised relative to integrating the MLD structure into a CMOS process, including: ensuring that the gate-drain overlap is large enough to cover implant shadowing effects that can result in asymmetrical device characteristics; choosing a spacer width that is optimized to prevent bridging between the gate and source-drain while ensuring consistent silicide formation in a production environment, and not incurring excessive S/D resistance; and adjusting the *p*-LDD doping to overcome the blanket *n*-LDD implant under worst case implant condition, saving a masking step to form both *n*- and *p*-LDD devices (see chap. 8).

9.11.2 Tailoring the Vertical Doping Profile of the n^- Implant to Improve the Conventional LDD

Besides increasing the n^- doping concentration, exploiting the effect of variations in the vertical doping profile of the n^- region have also been explored. Several structures have been proposed that employ an n^- LDD region whose peak concentration is beneath (instead of at) the surface. Reportedly this produces devices with better hot carrier resistance.

9.11.2.1 Buried-LDD (B-LDD). In the first of such devices, termed the *buried-LDD* (B-LDD)[93] or the *sloped-junction LDD* (SJ-LDD)[94] a deep (e.g., 165 keV) n^- implant is used. (see Fig. 9-43a). The n^- implants in the two respective structures are As [BLDD] and P [SJLDD]. The deep implant produces a vertical retrograde n^- doping profile, which causes the contour of the channel/n^- junction to be inclined at a non-perpendicular angle to the Si surface (see Fig. 9-43b). Since the electric field of a *pn* junction is roughly normal to the junction contour, the channel/n^- junction field will thus assist the gate field $\mathcal{E}_x$ in steering the electron current (and the impact ionization region) away from the surface and deeper into the silicon bulk. As the location of $\mathcal{E}_{ymax}$ remains at the surface, the current path is steered away from this point and less heating of the channel carriers results. Not only are there fewer hot carriers (as indicated by a reduction in I_{sub}), but these also have a lower probability of being able to traverse the increased distance to the Si-SiO_2 interface without recombining. Hence, less injection (and thus less oxide damage) results, and longer hot-carrier lifetimes are obtained.

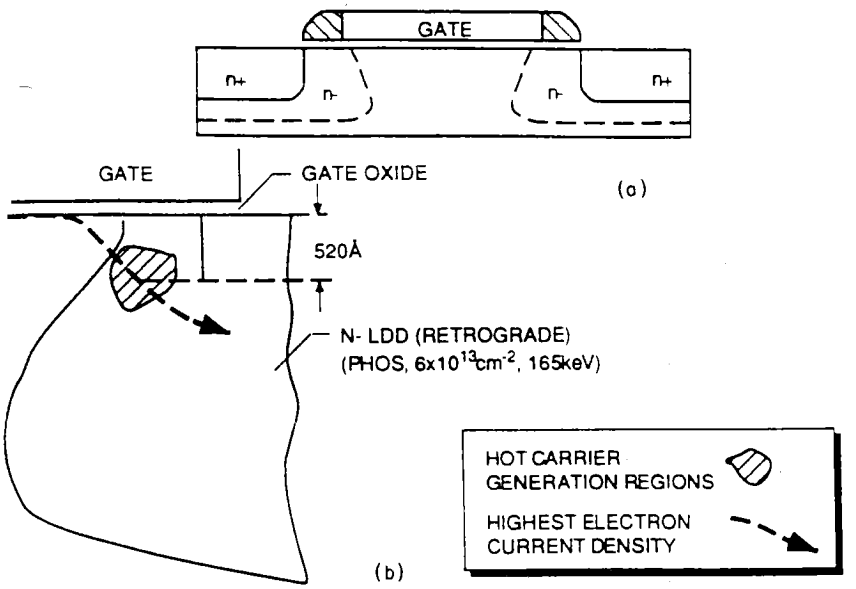


Fig. 9-43 (a) Schematic cross-section of the drain region of the *buried LDD* (B-LDD) structure.[93] (© 1985 IEEE); (b) Electric-field lines are normal to the buried *pn* junction, which steers the drain current away from the surface in this structure.[94] (© 1988 IEEE).

In the *graded-buried LDD* (GB-LDD)[95] the deep arsenic implant of the B-LDD is combined with the light n^- phosphorus dose of the conventional LDD (see Fig. 9-44). This results in a buried n^- profile that has a higher surface concentration than the B-LDD and which is also graded in the lateral direction when a thermal drive is carried out. This structure exhibits even better hot-carrier resistance than the BLDD, as seen in Fig. 9-45.

9.11.2.2 Profiled-LDD (P-LDD). If the implanted regions of the buried LDD structures just described are too deep, the concentration at the surface of the n^- regions just under the gate are reduced. This will cause degraded lifetime and increased parasitic drain resistance[96] (the latter effect is worsened by current crowding as described in ref. 145). This is one reason why the GB-LDD exhibits improved lifetimes compared to the B-LDD.

The GB-LDD structures can be even further improved by combining a very shallow moderately doped n^- region (e.g., As @ 50 keV) just under the surface, with a deeper retrograde n^- region (e.g., phosphorus, 1-5x10^{13} cm^{-2}) formed by a high-energy implant (≥150 keV). The shallow, abrupt As layer serves to increase the surface doping concentration and to reduce the impact of current crowding effects at the source end on the S/D resistance. After the two n^- implants are performed, a high temperature anneal is used to laterally diffuse the phosphorus more than the arsenic. After subsequent oxide spacer formation, the n^+ source/drain region is formed. Such a structure was first proposed by Toyoshima et al., in 1985, and the process sequence for forming it is

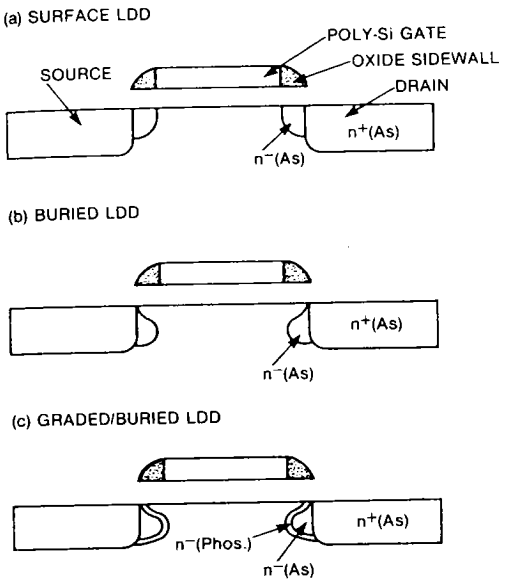


Fig. 9-44 Schematic cross-section of the *graded buried LDD* (GB-LDD).[95] (© 1986 IEEE).

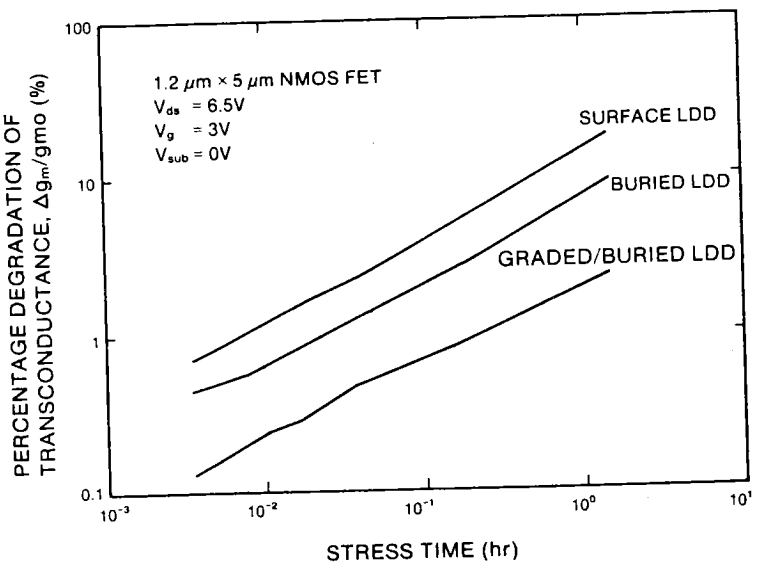


Fig. 9-45 Measured degradation in MOSFET transconductance under hot electron stress for the conventional LDD, B-LDD, and GB-LDD structures.[95] (© 1986 IEEE).

shown in Fig. 9-46.[97] A similar structure was reported by Chen et al., but they also added a $TiSi_2$ salicide layer.[92]

The P-LDD approach thus seeks to achieve all of the following objectives:

1. Reduction of the maximum electric field $\mathcal{E}_{ymax}$ in the silicon, as compared to the conventional LDD.
2. Ensuring that the injection position (i.e., the $\mathcal{E}_{ymax}$ point) is located under the gate edge.

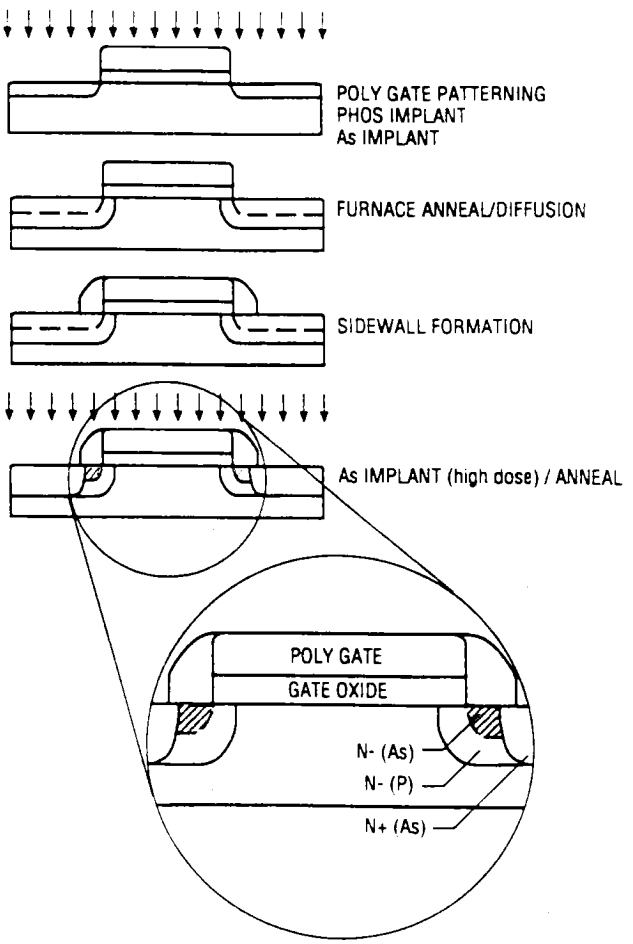


Fig. 9-46 Profiled LDD (PLDD) process sequence.[97] (© 1985 IEEE).

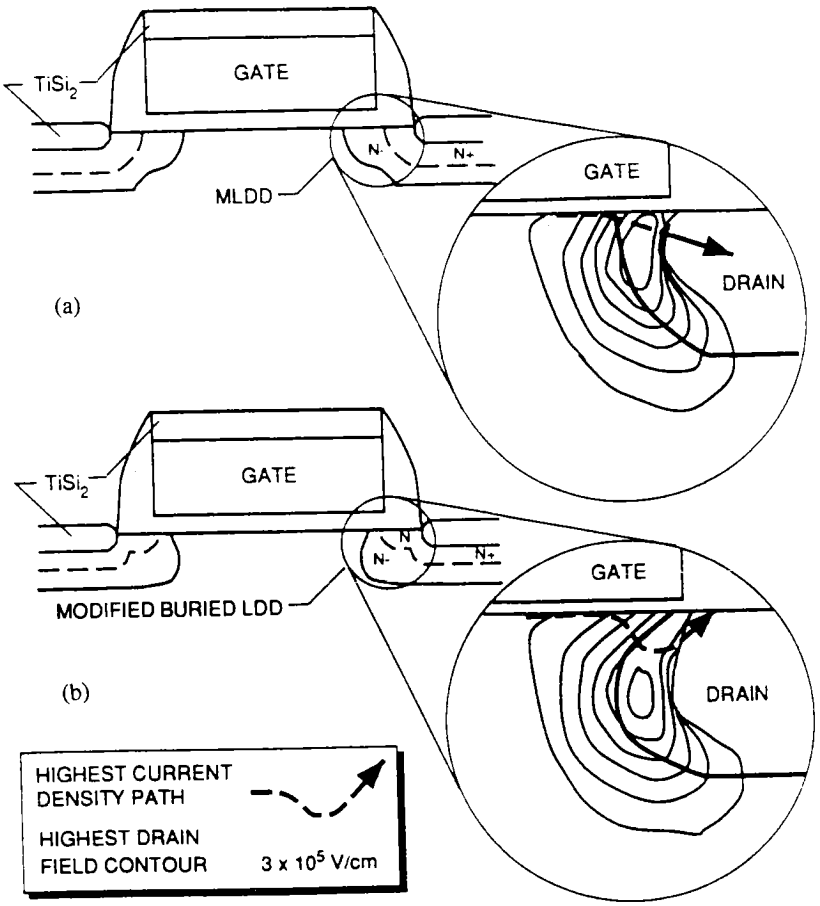


Fig. 9-47 Structure and lateral-field distribution for (a) MLD structure. (b) modified profiled LDD structure.[92] (© 1988 IEEE).

3. Separation of the point where the electric field in the silicon is at a maximum from the point of maximum current flow in the channel.
4. Ensuring that the impact-ionization region is pushed below the silicon surface to reduce the probability of hot carriers reaching the Si-SiO_2 interface.
5. Minimization of the increase in the parasitic resistance due to the LDD structure.

A simulation of the electric field and current path in this device is also shown in Fig. 9-47, where it is compared with those in an MLD-type structure. We see that in the

P-LDD structure there are two electric field peaks, but both are under the gate edge (Objective #2). We also observe that the current path is steered beneath the surface (Objective #4) and through the saddle point between these two peaks (Objective #3). A reduction in I_{sub} and the ratio of I_{sub}/I_D compared to MLD devices is also observed in P-LDDs, implying that $\mathcal{E}_{ymax}$ is reduced even more than in MLD-type devices (Objective #1). Objective #5 is met by the use of the shallow As n^- implant. Because the structure satisfies all of the above objectives it has been perceived as one of the most promising of the modified oxide-spacer-based LDD structures with respect to achieving high drive current and hot-electron resistance performance for MOSFETs down to 0.6 μm.[98] The lateral and vertical doping gradients in such structures, however, are quite complex, and this could cause manufacturing difficulties. Furthermore, it still requires a long high-temperature anneal to form the lightly graded region, and this may degrade the short-channel performance of such structures to a degree that they may not be suitable for devices with $L \leq 0.5$ μm.

9.11.3 Improving the Short-Channel Performance of the Conventional NMOS LDD

As noted in chap. 5, as MOSFET gate lengths are made shorter, various unwanted short-channel effects occur. One such effect in NMOSFETs is subsurface-DIBL (or punchthrough). While punchthrough stopper implants are used to suppress this problem, they become less effective as the gate length approaches 0.5 μm. An alternative technique to suppress punchthrough in such deep submicron LDD NMOSFETs is to locally implant p-type dopants (in NMOSFETs) under the lightly doped tip region of the LDD as shown in Fig. 9-48a. Such implants have been termed "halo" implants, and were first proposed for use in LDD NMOSFETs by Codella and Ogura in 1985. They called devices containing "halo" implants *double-implanted LDDs* (DI-LDDs).[99] Note that the use of halo-type implants as punchthrough stoppers in PMOSFETs is discussed in section 5.8.3.4.

Because the implanted dopant raises the doping concentration only on the inside walls of the source/drain junctions, the channel length can be decreased without needing to use a higher-doped substrate. That is, punchthrough does not set in until a shorter channel length because of the "halo." A penalty in increased sidewall junction capacitance is paid, but this cost is smaller than that incurred should the bottom area of the junctions lie above heavily doped substrate material. The maximum halo doping is determined by the onset of low-level avalanche breakdown of the pn^+ junction as is shown in Fig. 9-48b. Careful implant control is also essential to avoid compensation effects due to overlapping n- and p-dopants.

A halo NMOSFET with $L = 0.35$ μm was reported that operated at 3V with $t_{ox} = 15$ nm. The junction depth r_j was ~0.14 μm and the p-halo peak doping was ~3×10^{17} cm^{-3}. A more recent report applied a *large-angle tilt* implantation (LAT - see section 9.12.2) of boron ions in NMOS and phosphorus in PMOS (Fig. 9-48c) to form a halo-like punchthrough suppression region in each device, called SPI (*self-aligned pocket implantation*).[101,102] In NMOSFETs this approach first creates a conventional LDD

structure, using an n^- implant, a spacer oxide, and an n^+ implant. Note that the n^- dose will usually have to be increased to avoid compensation by the p^+ pocket implant. This is followed by the formation of $TiSi_2$ (salicide process). The SPI step is performed next, using a 40 keV, $8x10^{12}$ cm^{-2} phosphorus implant at a tilt angle of 25-30° to create the pockets of dopants. The gate/spacer structure and the $TiSi_2$ layer both act as self-aligned implant masks to define the regions into which the pocket dopant can penetrate.

Devices with channel lengths as small as 0.2 μm that still exhibited long-channel subthreshold V_T behavior were fabricated. Another report by Duvvury et al. indicated that unless the doping profile of the p^+ halo implant was correctly optimized, the DI-LDD would exhibit more susceptibility to electrostatic-discharge damage.[103]

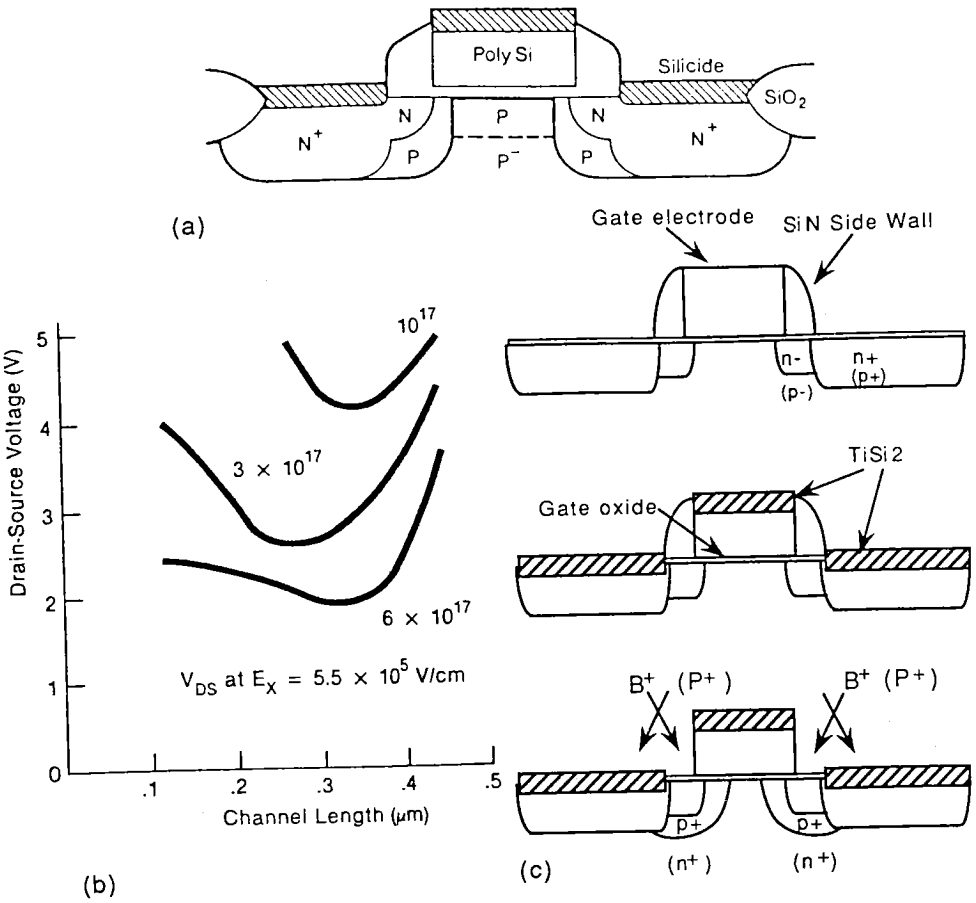


Fig. 9-48 (a) DI-LDD device cross-section. (b) Drain-Source voltage limit from low-level avalanche breakdown versus channel length for various peak halo doses.[99] (© 1985 IEEE). (c) Process sequence of SPI for NMOSFET and (PMOSFETs).[101] (© 1991 IEEE).

9.12 FULLY OVERLAPPED LDD STRUCTURES

As noted in section 9.6.1.2, an LDD structure with the gate fully overlapping the n^- region appears to be the optimum structure with respect to hot-carrier resistance and drive-current performance. Yet, this device structure is difficult to fabricate in submicron devices using the conventional oxide-spacer-based LDD approach. That is, the long thermal drive-in step required to drive the n^- region completely under the gate also produces a deep vertical junction, such that short-channel effects become excessive.

A variety of other device structures have been proposed to achieve the full-overlap condition, including the inverse-T structure (IT-LDD), the large-angle-tilt implanted drain (LATID) structure, the asymmetrical halo-source gate-overlapped LDD structure (HS-GOLD), and the high-dielectric-spacer structure. Here we describe these structures in more detail.

9.12.1 Inverse-T Gate LDD (ITLDD) and Gate-Drain Overlapped Device (GOLD) Structures

9.12.1.1 Inverse-T Gate LDD Structure. The first fully overlapped LDD structure was described by Huang et al. in 1986.[104,105] They named it the *inverse-T gate LDD* since the gate resembles that shape (see Fig. 9-49). Note that the conductive lateral poly extensions at the base of the gate completely overlap the n^- region of the LDD. The full overlap is produced by self-aligning the n^- region to the inside edge of

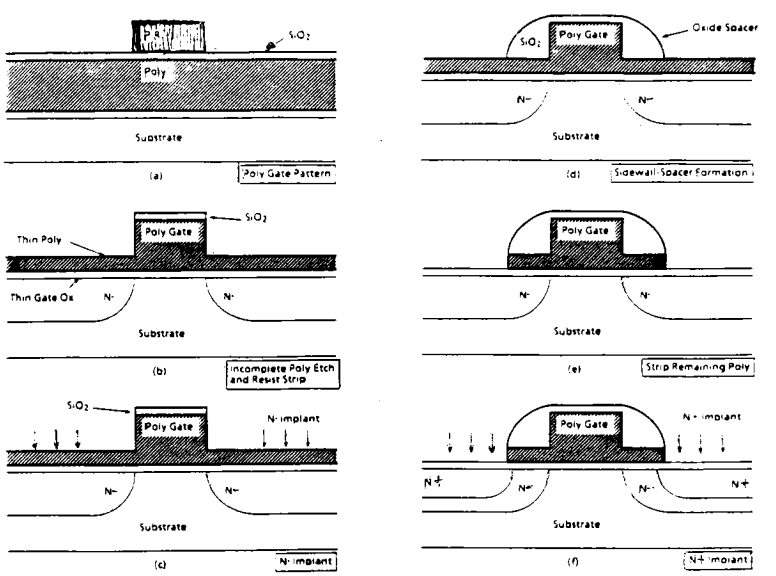


Fig. 9-49 Process sequence for ITLDD.[104] (© 1986 IEEE).

the inverse-T gate (by implanting phosphorus through the thin poly extensions), and by aligning the n^+ region to the outside edge of the gate. The latter alignment is achieved by forming an oxide spacer and then carrying out the n^+ implant. The n^+-to-gate offset is eliminated by this second self-alignment. (The remaining thin poly outside of the spacer regions is etched away prior to the n^+ implant.) No thermal step is needed to drive the n^+ dopants beneath the oxide spacer, and only a short RTP anneal is needed for implant dopant activation. Furthermore, since $\mathcal{E}_{ymax}$ is certain to be under the gate edge, a lower n^- dose can be used than in an M-LDD. Typically, n^- doses of $1x10^{13}$ cm^{-2} are used. The lower n^- doping helps reduce the gate-to-drain overlap capacitance.

IT-LDD structures with L as small as 0.6 μm have been found to exhibit excellent current drive, hot-carrier resistance, and short-channel behavior (punchthrough resistance). Hot-carrier lifetimes greater than 10 are years projected for 0.6 μm IT-LDD devices at 5V operation.

Despite the performance benefits that this structure provides, the original IT-LDD also has a number of serious limitations besides the obvious increase in process complexity, including the following:

1. *Difficulty in controlling the processes that form the poly extensions.* The key process in the sequence of fabricating the IT-LDD is the incomplete etch of the poly gate. An etch step creates the thin (50-100 nm thick) poly-gate extension regions. But it is difficult to control the poly thickness in a production environment because of the non-uniformity of the poly CVD and plasma etch processes, and because there is no natural etchstop or endpoint signal. Any nonuniformity in the thickness of these extensions results in changes to the vertical doping profile of the LDD n^- region.

2. *Damage of the gate oxide edge by the n^+ implant.* When the n^+ implant is carried out, the gate oxide at the outer edges of the poly gate is exposed to this implant. Arsenic implants through thin oxides at doses $\geq 1x10^{14}$ cm^{-2} have been identified as processes that degrade their dielectric breakdown voltage.[106] Hence, IT-LDD structures in which the gate oxide at the gate edge is not protected from the n^+ implant are prone to gate-oxide breakdown failures.

3. *Increase in gate-to-drain overlap capacitance.* The gate-to-drain overlap capacitance of the IT-LDD is higher than in non-fully overlapped LDD structures, despite the fact that a lower n^- dose can be used in the IT-LDD.

4. *Doping compensation problem in submicron LDDs is not corrected.* The problem of increased source/drain resistance due to compensation effects is not solved. This problem arises from the interaction of regions of high doping in the channels - needed in submicron devices - with those of comparable doping concentration, but opposite doping type, in the n^- regions.

We will see how modifications were made to the original IT-LDD structure to overcome these disadvantages.

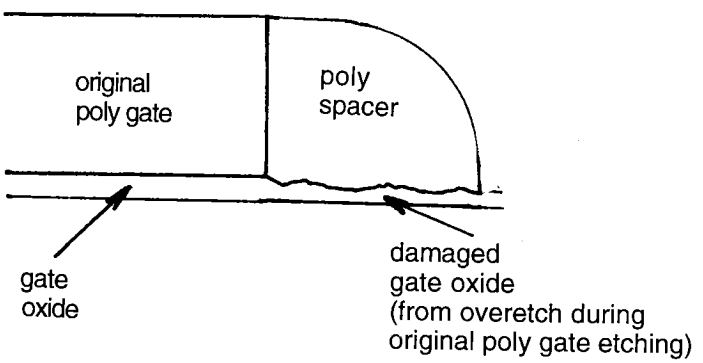


Fig. 9-50 Poly spacer acting as a poly gate extension can result in damaged oxide beneath the poly spacer that is subject to early breakdown.

Also note that poly extensions on a conventional poly gate could be created more simply, i.e., by forming poly spacers (instead of oxide spacers) after the n^- implant (see Fig. 9-50). The problem with this approach, however, is that when the original poly gate is etched (with an anisotropic etch to produce vertical sidewalls on the poly), some overetching must be employed. The exposed gate oxide is damaged during the overetch time and such damaged oxides exhibit early gate-oxide breakdown failures when subsequently covered by the poly spacers. Note that no re-ox can be used, because the poly spacer must make electrical contact to the main poly region to be able to act as an extension of the gate.

9.12.1.2 Gate-Drain Overlapped LDD (GOLD). A device quite similar to the IT-LDD structure was proposed by Izawa et al. in 1987,[107,108] which they termed the *gate-drain overlapped device* (GOLD). The GOLD device (Fig. 9-51) has two improvements designed to overcome two of the drawbacks of the IT-LDD structure described above. First, the poly gate film is deposited in two layers, with the first layer exposed to oxygen after deposition to allow a native oxide (5-10Å thick) to form on the surface at that time. When the two layer film is later etched to form the thin poly extensions, the native oxide layer serves as an etch stop - which should improve the control of this key poly etch process. Second, prior to implanting the n^+ region, the thin poly extensions are oxidized by a process that is called SELOCS (*selective oxide coating of silicon-gate*). This step serves two functions: (1) the length of the overlap of the gate and the n^- region overlap can be adjusted by this step; and (2) the gate oxide at the edge of the gate is now protected from exposure to the n^+ As implant, thus

preserving the gate oxide integrity of the MOSFET. (Note, however, that no information is provided as to how to ensure a good electrical contact between the two poly layers of the gate separated by the native oxide.)

9.12.1.3 Modified IT-LDD Structures. Several other versions of the IT-LDD structure have been reported which seek to improve the manufacturability and/or performance of the structure even further. The first is called an *inverse-T gate fully overlapped LDD* (FOLD),[109] which is an IT-LDD structure made with a more formidable etchstop layer between the two poly layers than the native oxide layer used in the GOLD structure. This etchstop layer is either a very thin (40Å) thermally grown oxide layer (Fig. 9-52a), or a PVD TiN layer (Fig. 9-52b), as both materials exhibit good selectivities to poly during reactive ion etching. The use of this etchstop further improves the control of the thickness of the poly extensions. In the process that uses the thermal oxide grown on the first poly layer, another thin layer of poly (300Å thick) must be deposited after the inverse-T poly gate is patterned and the thermal oxide etchstop layer is removed (with a short HF dip). The extra thin poly layer permits the upper and lower layers of the initial poly stack (which are separated by the thermal oxide etchstop) to be electrically connected. In the structure that uses the TiN etchstop, the extra poly layer is not needed. In both cases, As is used to form the n^- region as no additional drive-in is required to achieve n^- overlap. This structure exhibits adequate hot-

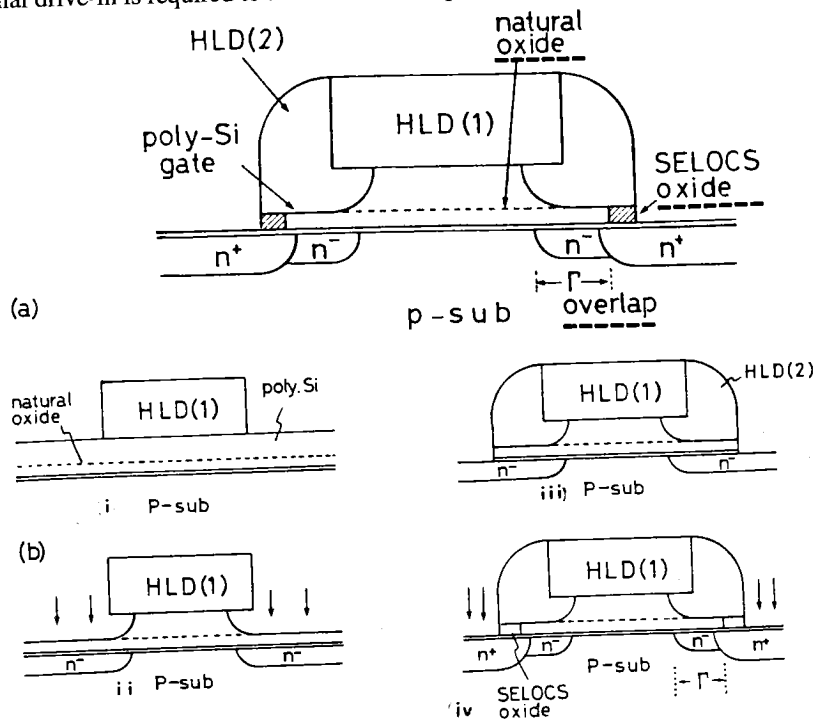


Fig. 9-51 (a) Schematic cross section; and (b) Process sequence of GOLD.[107] (© 1989 IEEE).

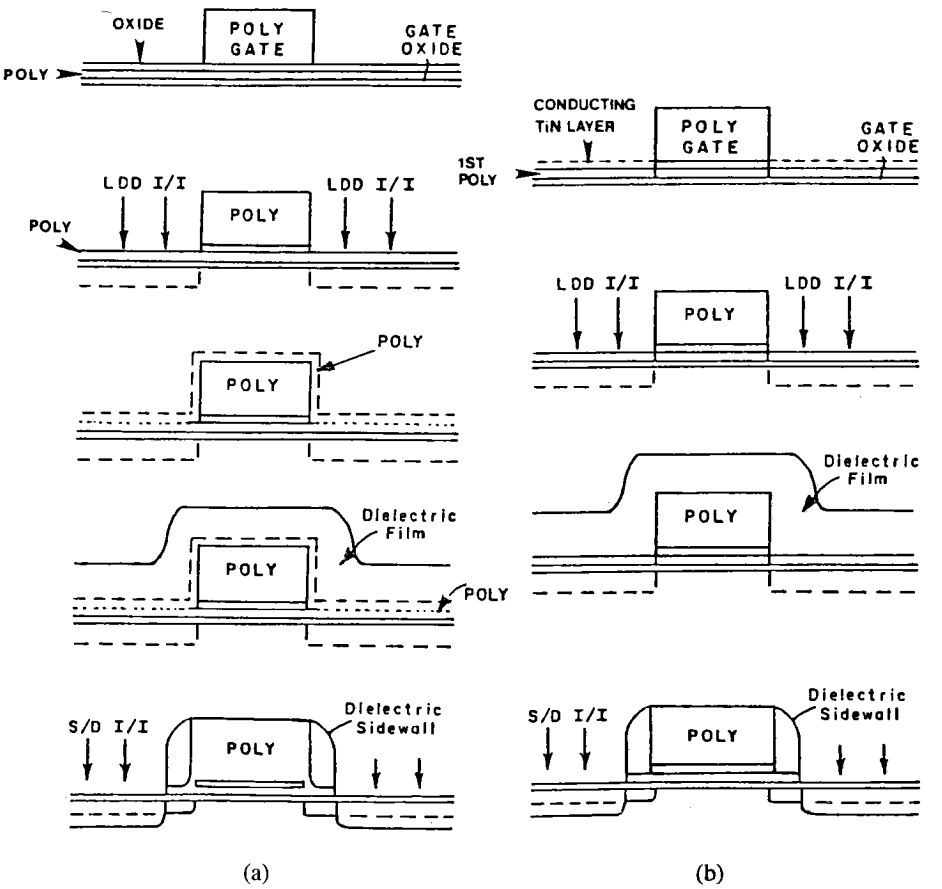


Fig. 9-52 (a) Process sequence for forming the FOLD structure; (b) FOLD structure with TiN etchstop layer.[109] (© 1989 IEEE).

carrier lifetimes for devices with $L = 0.35$ μm at a power supply of 3.5V.

A similar structure to the FOLD is the *total overlap with polysilicon spacer structure* (TOPS).[110] Instead of a thermal oxide etchstop layer in the poly gate stack, a 400Å CVD oxide layer is used (Fig. 9-53). The two poly gate layers are electrically connected with a poly spacer that is formed (in lieu of an oxide spacer) after the n^- implant is carried out. Note that it would be very difficult to form a salicide structure on the TOPS device as there is no oxide spacer to prevent the gate and source/drain from shorting during the silicidation step, and that neither the FOLD nor the TOPS structure provides protection of the oxide at the gate edges from the n^+ implant.

We noted earlier that it would be possible to create poly extensions on a conventional poly gate more simply by forming poly spacers (instead of oxide spacers) after the

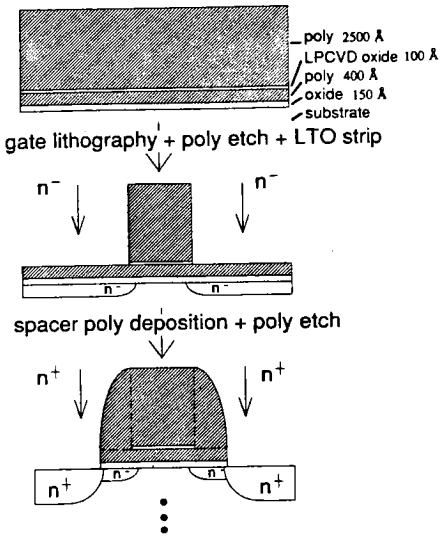


Fig. 9-53 TOPS structure layered gate and polysilicon spacer formation. Gate and spacer definition are accomplished by plasma etching with oxide endpoint detection. Typical CVD thicknesses are shown.[110] (© 1990 IEEE).

n^- implant. However, the gate oxide ending up beneath these poly spacers would be damaged by the poly etch. This would make such fully overlapped structures subject to early gate oxide breakdown failures. However, by re-growing gate oxide above the active regions exposed to the poly gate etch, as is done in the poly re-ox process, the resulting fresh oxide would exhibit adequate reliability. Unfortunately, if poly spacers were formed at this point, it appears that they could not serve as gate extensions, as they would be electrically disconnected from the gate.

However, in another IT-LDD-like structure proposed by I. Chen et al.,[111] poly extensions are formed after a re-ox process as described above, and the oxide grown on the poly sidewalls is apparently so heavily damaged by the n^+ implant that the gate and the poly spacer become electrically connected together. Hence, we can identify this structure as a *damaged re-ox layer fully overlapped structure*. Furthermore, by adding an additional oxide spacer to the edge of the poly spacer (Fig. 9-54),[112] the n^- overlap distance can be controlled (i.e., it becomes a partially overlapped device), and the oxide at the gate edge can be protected from the damage due to the n^+ implant in the same way that SELOCS protects the gate oxide in GOLD structures. In the partially overlapped device, the gate-to-drain capacitance is also significantly reduced, compared to a fully overlapped IT-LDD device.

Instead of using a dome-shaped poly spacer, as was done in the TOPS and in the damaged re-ox layer fully overlapped devices, M. Chen et al. suggest the use of *L-shaped spacers* (Fig. 9-55).[113] These are connected to the main gate poly by a $TiSi_2$ strap that is formed when a salicide is created on this structure. To prevent the gate and

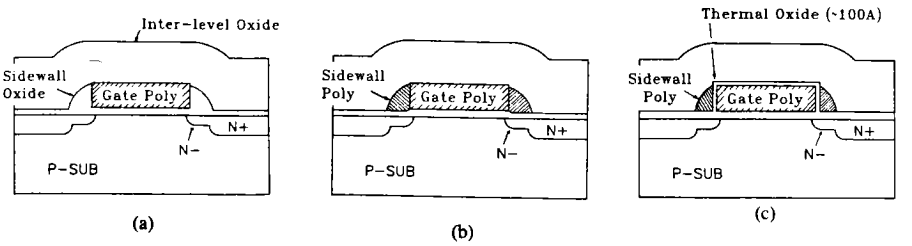


Fig. 9-54 Cross-sectional view of the (a) oxide spacer; (b) poly-spacer; and (c) oxide-padded poly spacer devices. The poly-spacer devices are permanent poly-sidewall to achieve the gate-to-drain overlap. For the oxide-padded poly-spacer device, due to the heavy source/drain implant, the thin thermal oxide between the gate and the poly spacer becomes conductive enough to keep the poly spacers at the same potential as the poly gate.[111] (© 1990 IEEE).

drain from shorting during the silicidation reaction, a nitride spacer is formed above the L-shaped spacer. It is much easier to integrate a salicide into this IT-LDD structure than in other IT-LDD structures.

The final IT-LDD structure we describe was proposed by Pfiester et al. in 1989. Here a novel approach is taken to forming the inverse-T gate.[114,115] That is, the main region of the gate is formed by selectively depositing poly, as shown in Fig. 9-56. After the main gate regions have been formed, a second layer of poly is blanket deposited to form the thin extension regions of the gate (see Fig. 9-56 again). The remainder of the process is similar to the conventional IT-LDD sequence. The advantages of this struc-

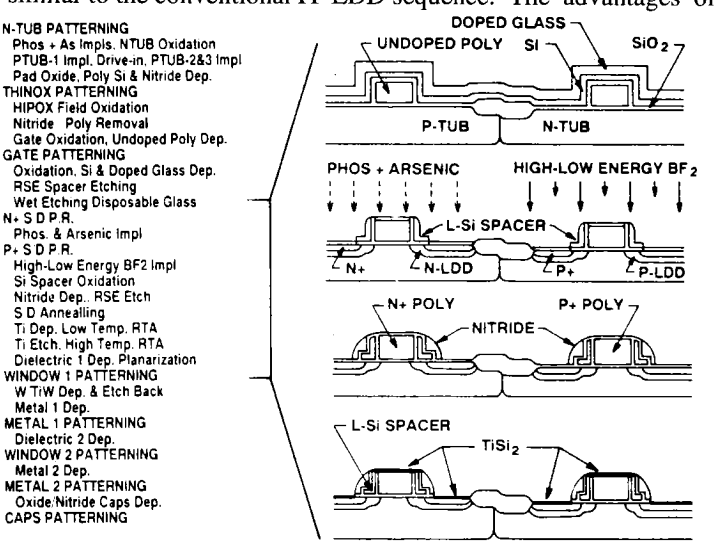


Fig. 9-55 Process sequence for self-aligned silicided L-shaped Si spacer LDD CMOS devices.[113] (© 1990 IEEE).

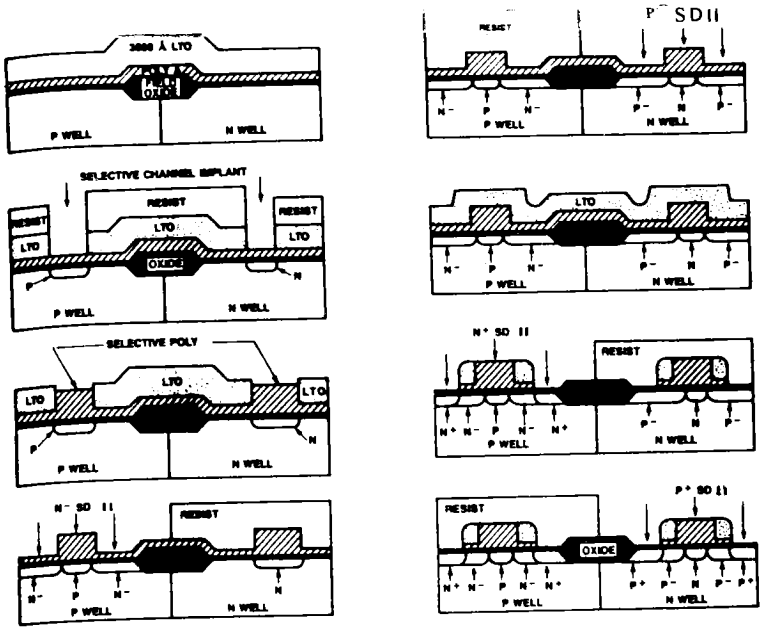


Fig. 9-56 Process sequence for forming self-aligned LDD/channel implanted ITLDD process with selectively-deposited poly gates for CMOS VLSI.[114] (© 1989 IEEE).

ture are: (1) a simpler, more manufacturable process sequence results, and (2) the channel dopants are only introduced into the region directly under the gate, not elsewhere. This avoids the compensation problem described in section 9.7.1.2 that occurs as heavier channel doping is encountered with shrinking channel length.

9.12.2 Large-Angle-Tilt Implanted Drain (LATID)

We have noted that fully overlapped drain structures can be fabricated with the IT-LDD approach, but that complex processing sequences are required that may be difficult to manufacture. A highly flexible technique for forming fully overlapped drains in submicron MOSFETs that is much simpler and also offers structure control and device performance improvements is based on large-angle-tilt implantation (LATID).[116-118,125] That is, this technique uses large tilt angles and target wafer rotational repositioning during implantation, without removing the wafer from the implant platen. The implant process sequence used to fabricate LATID is depicted in Fig. 9-57, together with the lateral doping profile that results (Fig. 9-58).

The first implant is performed after the poly gate is patterned, using an angle of 45-60° with a moderate phosphorus dose (1-2x10^{13} cm^{-2}) and energy (40-100 keV). This implant is carried out twice, with the wafer rotated 180° between these two implants, so that penetration of dopant under the gate is symmetrical. (In some circuit layouts in which some gates are oriented at right angles to other gates, four implants

90° apart would be needed.) The function of this implant is to introduce the n^- region dopant under the gate at the desired depth and doping concentration without having to use a diffusion step. After the n^- implant is performed, a gate sidewall spacer (0.08-0.10 μm) is formed, followed by a conventional high dose As implant that self-aligns the n^+ source/drain region to the gate, sets the gate/n^+ drain offset, and determines the source/drain junction depth. Note that to implement the LATID structure, an ion implanter that can vary the wafer tilt angle and also provide *in situ* repositioning of the wafer is needed. Recently, medium-current implanters that feature the capability of precise angle control (i.e., the beam angle remains constant across the wafer surface) and rotational repositioning of the wafer have become available. Note that the LATID structure requires no additional mask sets, and can be implemented as a "drop in" process module in current CMOS production processes, for example, as a reliability/ performance enhancement or as a part of a process "shrink."

The LATID technique permits the length of the gate/n^- overlap region, the horizontal and vertical concentration profiles of the n^- region, and the n^+-to-gate region overlap to be set independently. That is, the length of the lateral penetration of the n^- region can be increased by making the tilt angle larger (see Fig. 9-59). The length of the n^- region and its junction depth can be adjusted by the selection of the implant species and implant energy. The local peak concentration is set independently by the implant dose. These n^- parameters can be optimized independently of the n^+ drain-to-gate overlap, which is adjustable by the gate sidewall-spacer width.

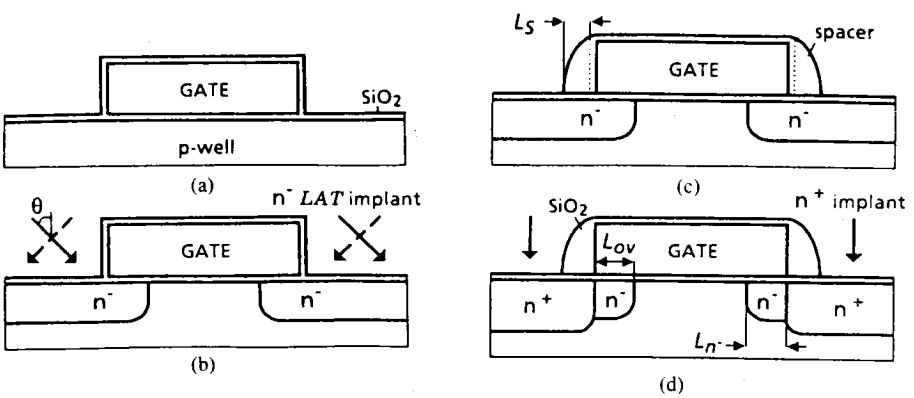


Fig. 9-57 Process sequences for 1/4 μm LATID FETs with ultrathin sidewall spacers. (a) 8.5-nm oxide formation and gate etching. (b) n^- LAT implant (tilt angle 45°). (c) formation of ultrathin ($L_S \leq 0.08$ μm) spacers. (d) n^+ implant and rapid thermal annealing (1000°C, 10 sec.)[117] (© 1989 IEEE).

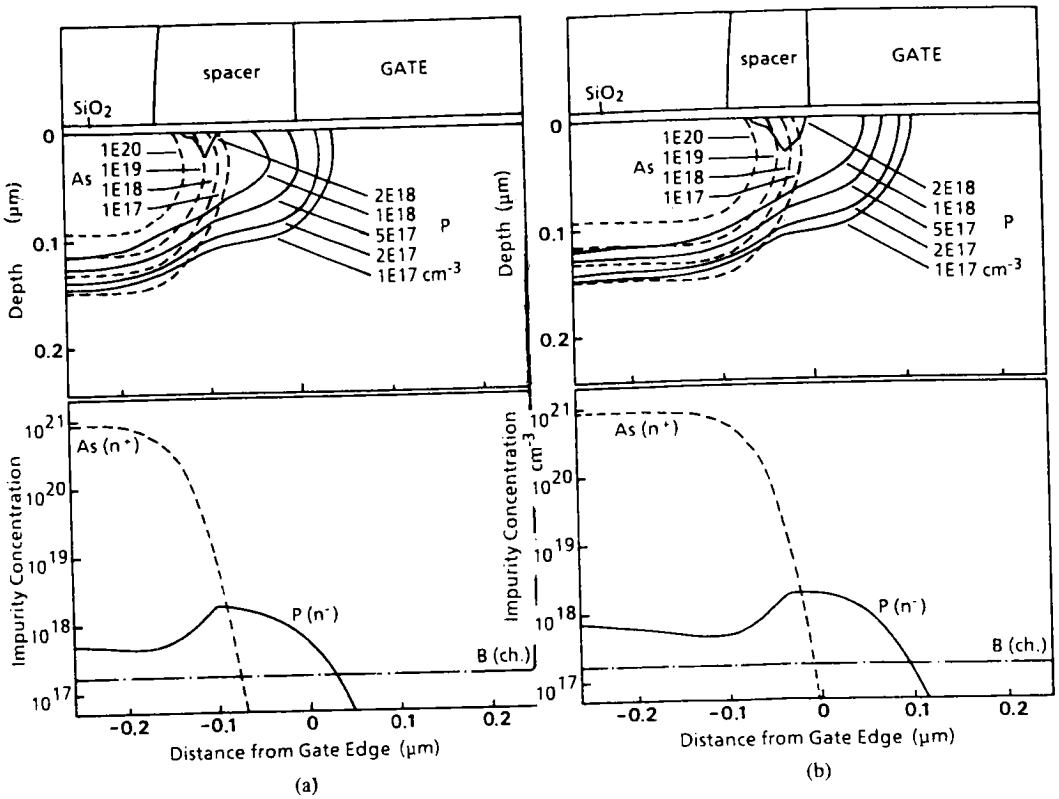


Fig. 9-58 Simulated 2D drain profiles and lateral ones near the surface (depth = 0.2 μm) for (a) conventional LDD; and (b) LATID with spacer devices after thermal annealing at 900°C for 10 min.[120] (© 1992 IEEE).

The LATID structures reportedly exhibit improved hot-carrier reliability and performance. NMOS LATID devices with a 0.5 μm gate length demonstrated extrapolated hot-carrier lifetimes of more than 10 years under 5V operation. The same device also exhibited high punchthrough immunity, i.e., similar to that of the control LDD device. The current drive capability was increased (i.e., by more than 50%) compared with the LDD control. The gate-to-drain capacitance is also reportedly significantly decreased below that of a conventional drain MOSFET.[119] It appears that the LATID device will be suitable for building MOSFETs to at least 0.25 μm, using power supplies of 3.3V.[120]

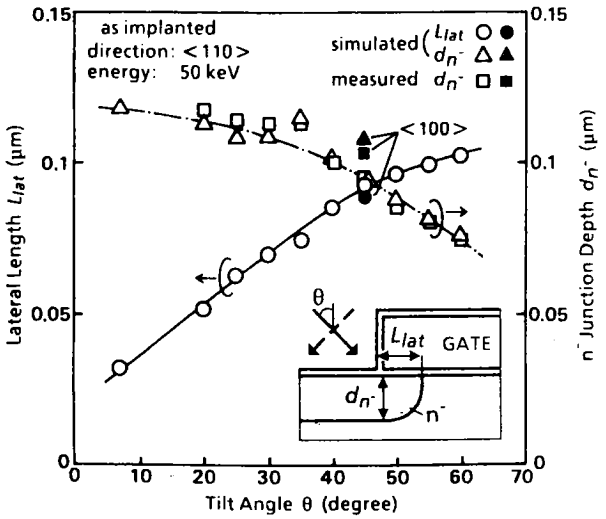


Fig. 9-59 Simulated lateral length L_{lat} from the implant-mask edge and n^- junction depth d_{n^-} at the gate edge (defined schematically in the inset figure) as a function of tilt angle θ for a constant ion flux N_{n^-} of $4x10^{13}$ cm^{-2}. Data of d_{n^-} estimated from SIMS measurements are also shown.[119] (© 1992 IEEE).

The gate-overlap capacitance of early (first-generation) LATID devices, however, was too high. This drawback was mitigated by incorporating a small (<0.1 μm) gate sidewall spacer in the second-generation LATID structures.[121] The resulting gate-overlap capacitance was reduced to a value comparable to that of a conventional LDD, even though the n^- region is well overlapped by the gate in the LATID. The small sidewall

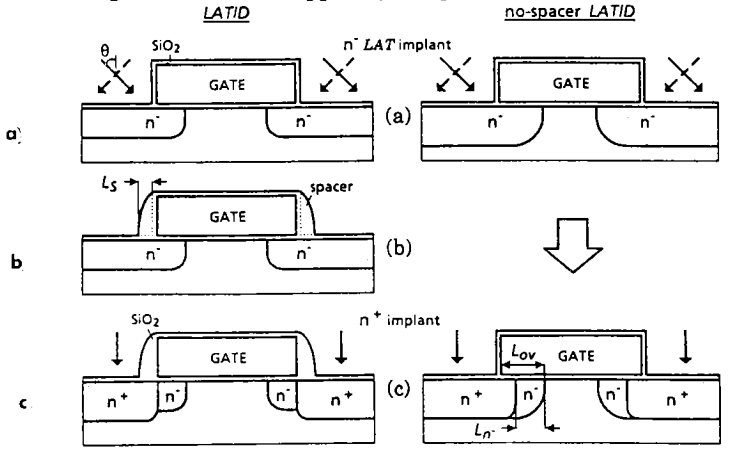


Fig. 9-60 Process sequence for LATID and *no-spacer* LATID. (a) n^- LAT implant after sidewall oxidation. (b) sidewall-spacer formation. (c) n^+ implant and annealing at 900°C for 10 min.[122] (© 1991 IEEE).

spacer is also needed to protect the gate oxide at the edge of the gate from being degraded by the n^+ implant, as described in the discussion on fabricating the IT-LDD.

An advanced LATID structure that also incorporates a self-aligned pocket implant that is also formed in combination with a LAT implant was described by Hori et al. in 1992.[122,123] Both NMOS and PMOS transistors can be built using this technique. The process sequence for fabricating this structure is illustrated in Fig. 9-60, where we also observe that a $TiSi_2$ film on the gate and source/drain regions and an epi substrate are also included. A dual-gate CMOS process is utilized, such that the PMOS transistor gates are doped *p*-type, and the NMOS gates *n*-type. Hence, both transistor types are surface-channel transistors. An inverter built with this CMOS process operates with a delay time of 30 ps and a power dissipation of 0.72 pJ.

Another LATID structure that deliberately incorporates a graded-gate oxide (GGO) region has also been described (see Fig. 9-61).[124] The GGO is formed by a poly re-ox step in a wet ambient at 850°C. This increases the 7-nm gate oxide thickness to about 25 nm at the edge of the gate. The presence of the GGO serves several functions in this structure. First, the vertical (gate) electric field is significantly smaller than that of the conventional LATID structure. This strongly suppresses the gate induced drain leakage (GIDL) current, which is high in LDDs (see chap. 4). Second, the gate-to-drain overlap capacitance is reduced as a result of the thicker gate oxide near the gate edges, where the gate-to-drain overlap occurs. This improves the propagation delay and power dissipation of the device compared with the conventional LATID fully overlapped MOSFET. These advantages in 0.6 μm T-gate LATID MOSFETs are obtained while maintaining comparable hot-carrier resistance as that of conventional 0.6 μm LATID structures.

9.12.3 High Dielectric Spacers

Fully overlapped LDDs have been created using IT-LDD and LATID structures. Another approach to obtaining the benefits of the fully overlapped LDD is to follow the conventional LDD process sequence, but to replace the oxide spacer with a spacer of higher dielectric constant (e.g., Si_3N_4, where $\varepsilon = 7.5$, or Ta_2O_5, where $\varepsilon = 30$), forming a so-called *high-dielectric spacer LDD* (H-LDD).[126-128] The effect of this replacement is to make a partially-overlapped LDD behave more like a fully overlapped LDD, especially as the gate lengths shrink to 0.5 μm or less. The use of high dielectric LDD spacers enhances the *gate fringing field* (GF) effects, which in turn increase the interaction of the gate and the n^- region beyond the gate edge. Up till now, the spacer material used for this purpose has been Si_3N_4. But since the GF effects are stronger if ε is larger, alternative spacer materials, such as Ta_2O_5 have also been considered. Furthermore, since the GF effects also become greater as devices are scaled down, the GF field becomes a more important factor in trying to exploit the benefits of scaled devices.

The H-LDD structure is shown in Fig. 9-62. Note that it could be created without the use of a poly sidewall and spacer-bottom oxide (Fig. 9-62b), but the Si/Si_3N_4 interface contains too high a density of interface states, and thus this structure exhibits poor hot-carrier resistance. The use of a thin poly sidewall and spacer-bottom oxide (Fig. 9-62c), reduces this problem, and the structure actually exhibits less degradation

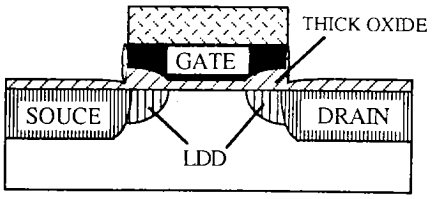


Fig. 9-61 T-gate overlapped LDD device structure. The device has the thick gate-oxide above the LDD region.[124] (© 1991 IEEE).

than an oxide-spacer LDD when $L = 0.5\ \mu m$ (see Fig. 9-63). The hot-carrier reliability of the H-LDD is enhanced by the GF effect for other reasons as well. First, the GF effect suppresses the structural degradation effect (negative charges built up in the spacer oxide deplete the n^- region and cause its resistance to rise).[128] This keeps the reduction of I_D during hot-carrier stress even smaller. Second, the GF effect causes the impact ionization rate to decrease because the high drain potential edge is pushed and moves toward the drain n^+ region. In essence, the GF effects enlarge the effective channel length, thereby reducing $\mathcal{E}_{ymax}$.

The drive-current capability of the H-LDD is also increased by the GF effect. That is, the GF increases the vertical field in the source n^- region, causing the electron density to increase there. Therefore, the LDD characteristic parasitic resistance of the n^- region under the LDD spacer decreases. Then R_S becomes smaller, resulting in a higher g_m.

A sensitivity study of the impact of varying the thickness of both the polyoxide sidewall thickness and spacer bottom oxide on the performance and reliability of the H-LDD structure was recently reported.[129] This was a theoretical study that used the process and device simulators SUPREM IV and PISCES-2B to examine how the lateral electric field, drive current, and subthreshold leakage were altered as these two oxide thicknesses were varied and a high-dielectric spacer was used for the LDD structure. The

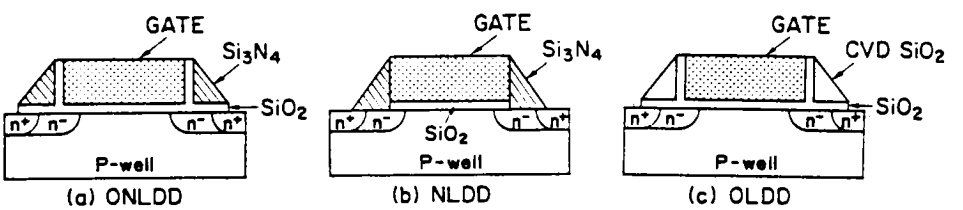


Fig. 9-62 Schematic cross section of three different LDD spacer structures. The oxide under the nitride film is fabricated by the re-ox step. (a) Si_3N_4/SiO_2 spacer (ONLDD); (b) Si_3N_4 spacer (NLDD); (c) oxide spacer (OLDD).[126] (© 1988 IEEE).

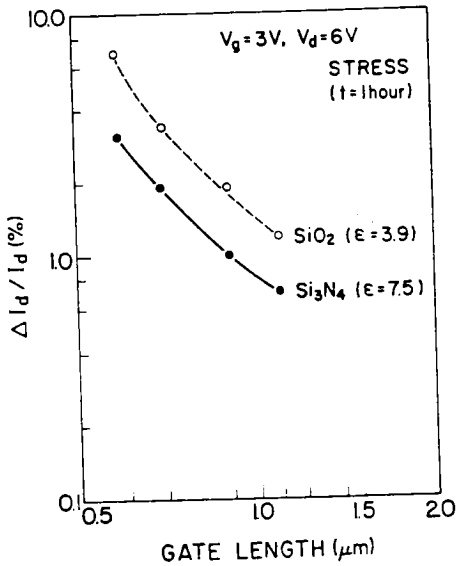


Fig. 9-63 Drain current degradation rate versus gate length for CVD-oxide spacer LDD and CVD-nitride spacer (with a 20 nm thick poly oxide).[127] (© 1989 IEEE).

results indicated that the structure can be optimized to give good performance and high hot-carrier reliability in devices with $L = 0.5$ μm.

9.12.4 Asymmetrical MOSFET Structure: Halo-Source-GOLD (HS-GOLD)

A novel approach to allow devices with gate lengths below 0.25 μm to possess adequate hot-carrier reliability is an asymmetrical MOSFET structure proposed in 1989 by Buti et al. They called this structure the *halo-source gate-overlapped drain* (HS-GOLD), and it is shown in Fig. 9-64.[130,131] Here the LAT implantation technique is used to form both a gate-overlapped structure only on the drain side of the gate (GOLD), as well as a halo implant punchthrough stopper (Halo Source - HS) only on the source side. This is accomplished by using two separate 45° implants, *p*- and *n*-type respectively. Superior hot carrier reliability and high punchthrough resistance were reported for this device, even at $L = 0.2$ μm. It is estimated that the structure would be adequate for gate lengths as small as 0.15 μm at a power supply voltage of 3.5V. The fabrication parameters used to build the structure are shown in Fig. 9-64b, including the LAT implant parameters. Figure 9-64c illustrates the doping profiles for a 0.5 μm device.

The key feature of this structure is that the n^- region and the *p*-type punchthrough stopper are completely separated. If the device was symmetrical, the *p*-type halo region would abut the n^- region doping. This would reduce the breakdown voltage and degrade

hot carrier reliability. By separating these two regions in the asymmetrical device, the punchthrough resistance and hot-carrier reliability can be independently optimized, something that is not possible in a symmetrical MOS structure.

Another important benefit of using the lightly doped region only on the drain side is that the parasitic series resistance and gate overlap capacitance are smaller than in a symmetric structure. For MOSFETs with L below 0.25 μm these two parameters become limiting factors in device and circuit performance, and hence their reduction is crucial. The main drawback of asymmetrical structures is that their use may restrict the flexibility of circuit layout, especially in some IC applications.

9.13 HOT-CARRIER EFFECTS IN PMOS TRANSISTORS

It is not necessary to use LDDs in PMOS devices whose gate lengths are longer than about 1.0 μm. The reason is that hot-carrier effects are not serious in such devices operated at 5V. That is, the impact-ionization rate due to holes (i.e., the number of

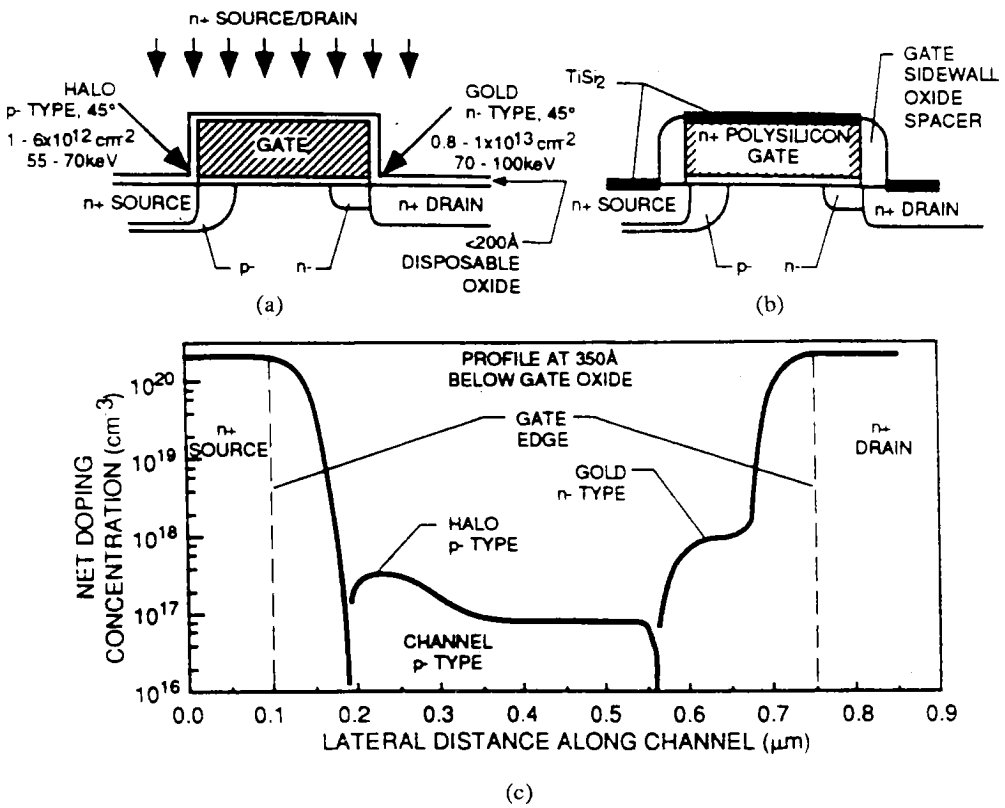


Fig. 9-64 The *halo-source gate-overlapped drain* (HS-GOLD) structure; (b) the fabrication parameters used to build this asymmetrical structure, including the LAT implant parameters; (c) the doping profiles for a 0.5 μm HS-GOLD device.[130] (© 1989 IEEE).

electron-hole pairs generated by a hot hole) is 1 or 2 orders of magnitude lower than that due to electrons at a given electric field. A secondary reason is that the buried-channel behavior associated with PMOS transistors with n^+ poly gates results in a deeper current path and possibly a lower electric field, both of which reduce hot carrier effects. At submicron channel lengths, however, hot-electron effects in PMOS begin to become significant.

It has been reported that two hot-carrier effects predominate in submicron PMOS devices. Both are apparently caused by hot electrons generated through impact ionization. These hot electrons are then injected into the oxide and become trapped. Hot-hole injection does not appear to cause significant degradation, despite the fact that holes are far more numerous in the channel of PMOSFETs than electrons. That is, hole injection into the oxide is far less probable than electron injection because the potential energy barrier for holes at the Si-SiO_2 interface is higher. In addition, when PMOS devices are operated in saturation (i.e., $V_{GS}<V_{DS}$), the gate-bias polarity favors electron drift through the oxide (which is opposite to that in NMOS), but retards the drift of holes. In fact, in PMOS devices only gate current due to hot-avalanched electrons has been measured, and no gate current resulting from channel hot holes has been reported in the literature.

9.13.1 Model for Hot-Carrier Damage in PMOSFETs

A qualitative model for the mechanism responsible for the PMOS device degradation mentioned above is as follows: At high channel fields, channel holes are sufficiently accelerated to cause impact ionization, thereby generating electron-hole pairs. If some of these generated electrons can gain sufficient energy from the channel field and then also happen to be scattered in the right direction, they can be injected into the gate oxide. Some of such injected electrons become trapped, and negative charge is thus built up in the oxide near the drain.[132,133]

It has been reported that the most severe degradation in PMOSFETs is observed under bias conditions corresponding to peak gate current I_{Gmax}. In general, I_{Gmax} in PMOSFETs occurs at low gate bias. That is, the worst stressing condition for PMOSFETs is at low V_{GS} (roughly V_{GS} = -1.25V when V_{DS} = -8V) because its corresponding oxide field results in a maximum gate current and a maximum stress-induced V_T-shift (although the maximum substrate current occurs at higher V_{GS}, i.e., $V_{GS} \sim 0.4V_{DS}$ or roughly -3V in this case). Again this condition is opposite to the results observed in NMOSFETs, in which the worst degradation is observed at the gate voltage corresponding to the largest substrate current. This is because in PMOSFETs when V_{GS} is increased (i.e., made more negative) the electron injection into the gate oxide is suppressed. A study of hot-electron damage in PMOSFETs with p^+ gates is given in Ref. 134.

9.13.2 Hot-Electron Induced Punchthrough (HEIP)

The chief way in which PMOS devices are impacted by hot-electron trapping in the oxide is that a negative charge buildup in the oxide occurs near the drain. This

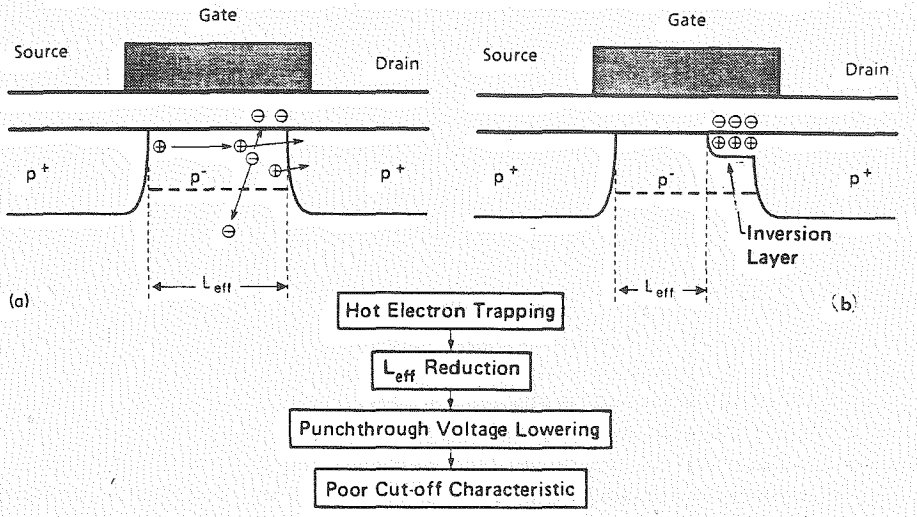


Fig. 9-65 Schematic diagram of the hot electron injection into the gate oxide of PMOSFETs.[136] (© 1987 IEEE).

eventually inverts the n-type Si surface near the drain, resulting in an extension of the p^+ drain region, as illustrated in Fig. 9-65. As a result, the effective channel length appears to shrink. This reduction in L_{eff} is even more severe in the buried-channel PMOS devices (used in CMOS technologies that have an n^+ polysilicon gate), because their buried-channel nature makes them more vulnerable to source-drain punchthrough, and consequently increases subthreshold leakage. This effect is called hot-electron-induced punchthrough (HEIP).[135,136]

Figure 9-66 shows the effects of hot-carrier stress (i.e., decrease in punchthrough

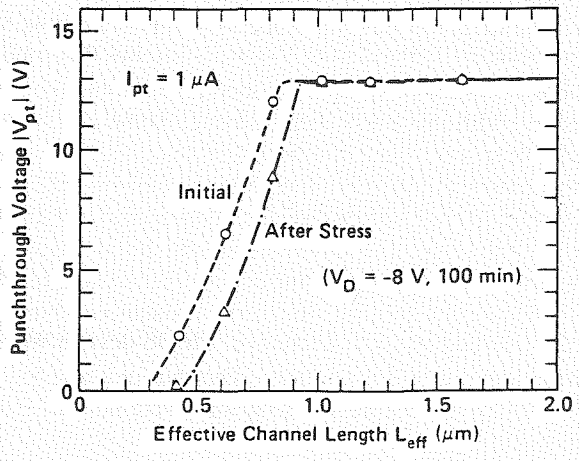


Fig. 9-66 Punchthrough voltage as a function of initial effective length for PMOS devices before and after stressing.[136] (© 1987 IEEE).

voltage as a function of stress time) in buried-channel PMOSFETs whose values of L_{eff} prior to stressing are 0.8 and 0.6 μm.[135] After only 2 hours of stress, the punchthrough voltage of the 0.6 μm device is seen to have decreased by a factor of two. A short lifetime (<10 days) was exhibited by the 0.6 μm PMOS devices under the worst-case condition for evaluating the submicron PMOSFET hot-carrier reliability. That is, the stress condition was chosen to give maximum gate current (i.e., V_{DS} = -8V and V_{GS} = -1.2V). The change in V_T (with V_T in this case defined by the V_{GS} value when I_D = -10 nA) after stressing was measured in the low gate voltage condition as a measure of HEIP-induced L_{eff} reduction, with the source and drain terminals interchanged (as this represents the worst-case stress condition for evaluating hot-carrier reliability in PMOSFETs). A model of PMOSFET hot-carrier degradation based on the shortening of L_{eff} due to hot-carrier damage is given in ref. 139.

A second effect arising from the negative charge buildup in the oxide of PMOS devices is also observed. That is, this buildup tends to reduce the magnitude of V_T. The combined effects of a smaller value of L_{eff} and a smaller V_T value result in an increase in I_D, which is usually desirable in MOSFETs, but in some circuit applications may play a detrimental role.

It should also be noted that, contrary to what occurs in NMOS, both of the above effects tend to saturate in PMOS. This can be understood by noting that the electrons trapped near the drain reduce the gate electric field present there in PMOS devices. Eventually, hot electrons injected into the oxide are drifted less strongly through the oxide to the gate, or they may even be repelled back into the Si without causing further damage. As a result, this effect does not appear to limit PMOS devices fabricated with p^+ polysilicon gates as long as L_{eff} is $\geq$0.6 μm.[134] Still, LDD structures appear to be needed for p-channel devices fabricated with n^+ polysilicon gates when L_{eff} gets to be 0.8 μm or less.[135] Figure 9-67 shows the hot-carrier lifetime improvement gained by using a p-channel LDD. It compares conventional 0.95 μm PMOS devices with 0.8 μm LDD PMOSFETs. The LDD devices have superior reliability despite their shorter gate lengths, and an extrapolation of the data to V_{DS} = 5V implies that lifetimes in excess of 10 years are possible. As noted in section, the incorporation of a PMOS LDD structure generally requires an additional mask step, although some processes that avoid this penalty (e.g., disposable spacers) have also been reported.

It should be noted, however, that a recent report indicates that graded-drain (i.e., LDD) PMOS structures exhibit *worse* hot-carrier degradation than abrupt drain (non-LDD) PMOSFETs.[137] Thus, the use of an LDD in PMOS might not automatically guarantee longer lifetimes, and separate solutions to the hot-carrier problem in PMOSFETs might have to be considered. For example, the hot-carrier resistance of buried-channel PMOS devices is reported to increase as t_{ox} is decreased. PMOS devices with gate oxides of 7 nm thickness exhibit very little degradation. The thin oxide appears to prevent the damaged region caused by trapped hot electrons that were injected into the oxide from spreading along channel toward the source.[138]

The LDD structure also helps to reduce subthreshold leakage in PMOS devices, where excessive junction depths caused by lateral diffusion have been a problem. The use of LDD basically increases L_{eff}, hence compensating for the L_{eff} reduction caused

by hot-electron injection. A new LDD structure for PMOS devices, called a *halo* LDD, is described in chap. 5.[63] In this structure, a deeper phosphorus implant is placed below the lightly doped drain-extension p-type implant. The punchthrough resistance of the PMOS device is reported to be significantly improved by this LDD structure.

9.13.3 Hot-Electron Induced Degradation in p^+-Poly PMOSFETs

Because of short-channel effects, when the channel length of PMOS devices gets to 0.5 μm or smaller, it is likely that n^+ poly will no longer be suitable as the gate material. Instead, PMOS devices will have to be fabricated with p^+ poly gates (see chap. 5). With a p^+ poly gate, PMOSFETs become surface channel devices. LDDs may still be need to be used in such devices, but their main roles will be to permit shallower source/drain junctions to be created immediately adjacent to the active channel region, and to provide electrical conductivity between the active channel region and the p^+ source/drain.

As noted in chap. 5, in surface-channel PMOSFETs, boron penetration from the p^+ poly into the substrate through the thin gate oxide can occur during high temperature annealing. This boron penetration causes a positive V_T shift and an increase in S_t. It also can cause such other undesirable effects as an increase in electron trapping, a decrease in low-field hole mobility,[186] and a degradation of current drive due to polysilicon depletion.*[187] Furthermore, it has been reported that boron penetration is enhanced when a BF_2^+ implant is used to dope the poly as compared to a B^+ implantation. This implies that the fluorine contained in the gate poly enhances boron penetration. Since this results in a large amount of boron incorporation in the gate oxide, the oxide reliability (and the hot-carrier device reliability) of such PMOSFETs comes into question. A report by Mogami indicates that the hot-carrier lifetime in both B^+-doped and BF_2^+ doped p-type poly devices is independent of the amount of boron penetration.[140] On the other hand, the lifetime of BF_2^+ doped devices is about 100 times longer than that of B^+ doped devices. This may be due to the larger critical energy for electron trap generation during hot-carrier stress in BF_2^+ gates, as this should result in less electron trapping in the gate oxide for the BF_2^+ gate. It has been reported that a fluorine-incorporated oxide has high immunity to hot-carrier-related degradation because the Si-F bonding energy is higher than Si-H bonding energy (see section 9.14).

One suggested way to way to suppress boron penetration of the gate dielectric is to use silicon oxynitride instead of SiO_2. Such oxynitride layers can be fabricated with

* The *polysilicon depletion effect* occurs when boron, as it penetrates the gate oxide, crosses from the p^+ polysilicon gate into the oxide. The concentration of boron in the region of poly near the poly-oxide interface is thereby reduced. This causes a depletion region to arise in the poly adjacent to the poly-oxide interface. Its presence is manifested as an increase in the "equivalent" gate oxide thickness.[188] As the dopant concentration at the polysilicon/oxide interface is reduced, this effect increases. Thus, a gate oxide material that suppresses boron penetration of the gate dielectric would also mitigate this problem.

a number of different techniques, and the effectiveness in suppressing boron penetration of each of them is compared in reference 141. The details of oxynitride formation are discussed in section 9.14. Another approach to limiting boron penetration is to reduce the thermal budget after the boron is implanted into the gate poly. Figure 9-68 shows the degree of boron penetration as a function of gate oxide thickness, post-implant anneal temperature, and B^+ or BF_2^+ implant.[143] Yet another technique involves *in-situ* boron doping of the gate poly, followed by selective doping of the NMOSFET gate poly.[142] This method yields less boron penetration than boron doping of the poly by implantation followed by an activation anneal. Finally, a co-implantation of boron and fluorine into an as-deposited amorphous-silicon gate layer is reported to result in minimum boron diffusion into the silicon substrate while providing a reduction in the interface state density at the oxide/silicon interface.[144]

9.14 STRENGTHENING OXIDES AGAINST HOT-CARRIER INJECTION

Altering the structure of the MOSFET is one approach to reducing the degradation effects of hot-carrier injection. Another is to make the gate oxide and oxide-silicon interface more resistant to hot carriers. This strategy entails development of gate dielectrics with the following two properties: (1) the as-formed films should exhibit very low densities of oxide trapping centers, interface state traps, and fixed oxide charge; and

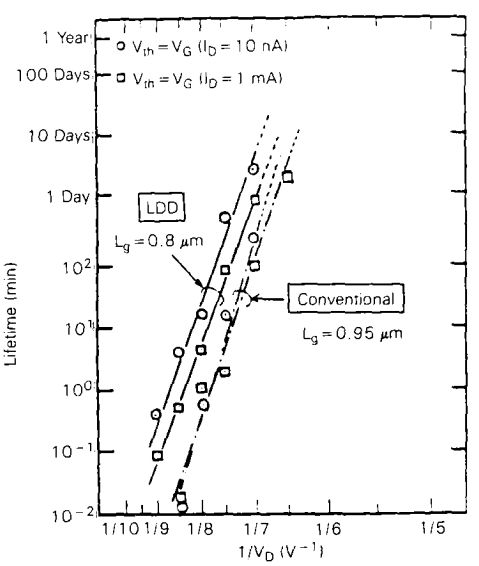


Fig. 9-67 Device lifetime versus reciprocal stressing drain voltage for submicron LDD PMOS devices. Data for conventional PMOS device is also plotted for comparison.[136] (© 1987 IEEE).

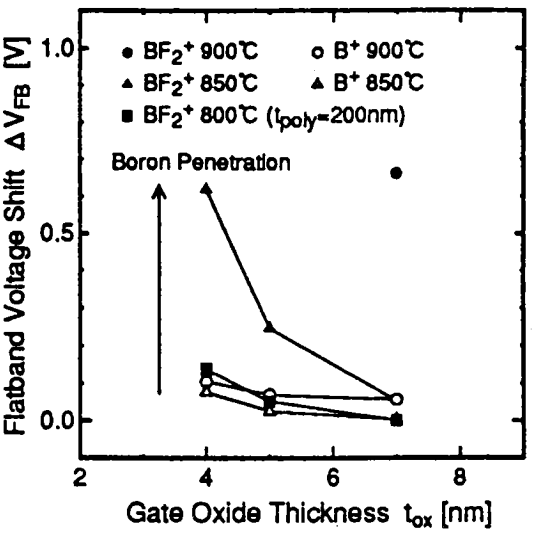


Fig. 9-68 Dependence of ΔV_{FB} due to boron penetration, determined from C-V characteristics, on t_{OX}. For t_{OX} ≤5-nm thick, samples with 900°C annealing after BF_2^+ implantation, V_{FB} values are unmeasurable because of their anomalous C-V curves, due to excess large boron penetration.[143] (© 1993 IEEE).

(2) these densities should not increase when the MOSFETs in which these films are used as gate dielectrics are subjected to hot-carrier stressing. In films with such properties the density of states available for occupancy by hot carriers injected into the oxide is small - both in the as-formed films and after being subjected to hot-carrier stressing. The ability of injected hot-carriers to cause device-degradation is thereby limited, since the number of trapping sites they can occupy is small. In this section we describe the processing techniques developed for the purpose of attaining the above goals.

9.14.1 Minimizing Hydrogen and H_2O at the Si-SiO_2 Interface

We have previously mentioned that the presence of excess hydrogen and/or water in MOSFET gate oxides has been identified as a factor which makes the devices more prone to hot-carrier degradation. Here we consider this topic in more detail.

9.14.1.1. The role of hydrogen in hot-carrier degradation. Excess hydrogen at the Si-SiO_2 interface is reported to be one of the culprits responsible for the increase in interface state density in MOSFETs during hot-carrier stressing. Earlier we indicated that atomic hydrogen appears to play a benign role in MOSFETs behavior. That is, hydrogen tends to passivate the Si-SiO_2 interface traps, most probably by the formation of Si-H bonds which eliminate dangling bonds at the Si-SiO_2 interface.

However, it is also known that such Si-H bonds can easily be broken by high-energy electrons. Hence, such bonds are likely to be broken by injected hot-carriers or ionizing radiation, thereby generating interface states.

Some amount of hydrogen is always incorporated into the MOS gate oxide because of its almost universal presence during processing. Especially large amounts of hydrogen are present during such elevated-temperature processing steps as: epitaxial growth; LPCVD of Si_3N_4; deposition of TEOS; reflow in steam; and the final sintering step in hydrogen (100 percent H_2 or 4-percent H_2 [forming gas]). Hydrogen can also incorporated into the gate oxide during high-temperature anneal steps when hydrogen-rich layers (e.g., PECVD nitride) are present on the wafer. It has been observed that charge-to-breakdown, trap generation rates, and hot-carrier degradation are worse in gate oxides which have been exposed to hydrogen at high temperature.

In early studies of hot electrons, water diffusion experiments into very dry oxides suggested that the amount of trapped charge was related to the water content in the oxide. As the moisture in the oxide was increased, so did the number of trapped electrons. In these experiments, hydrogen evolution was also observed to occur. The amount of hydrogen lost was found to have a 1:1 correspondence with the amount of negative charge produced.[153] In avalanche experiments, interface state generation for wet oxides showed a substantial increase over dry oxides, suggesting that hydrogen or water-containing trapping centers were also involved. Further evidence that implicated hydrogen was found from experiments in which the final sinter of MOSFETs was performed in various ambients (i.e., in either hydrogen or nitrogen at 450°C for 30 minutes). As shown in Fig. 9-69, devices annealed in hydrogen display a higher degradation rate, with device lifetimes improving by two orders of magnitude when

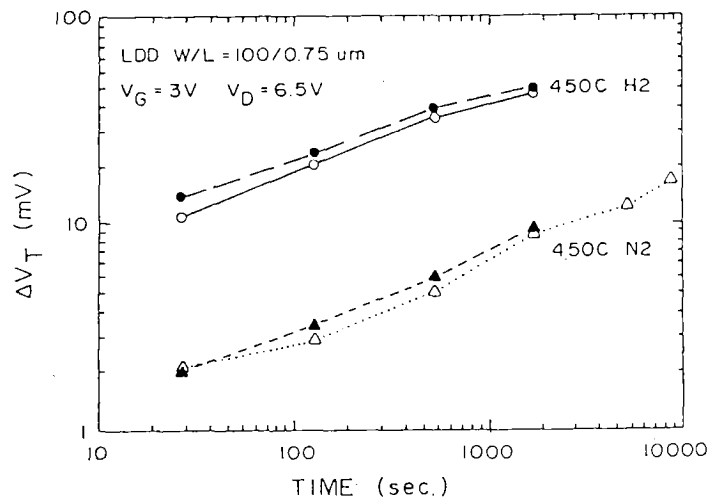


Fig. 9-69 Comparison of degradation rates in LDD devices annealed in hydrogen and nitrogen ambients (450°C 30 min). The degradation rate is greatly reduced in nitrogen-annealed devices. The two lines in each group correspond to different wafers from the same split.[154] (© IEEE 1985.)

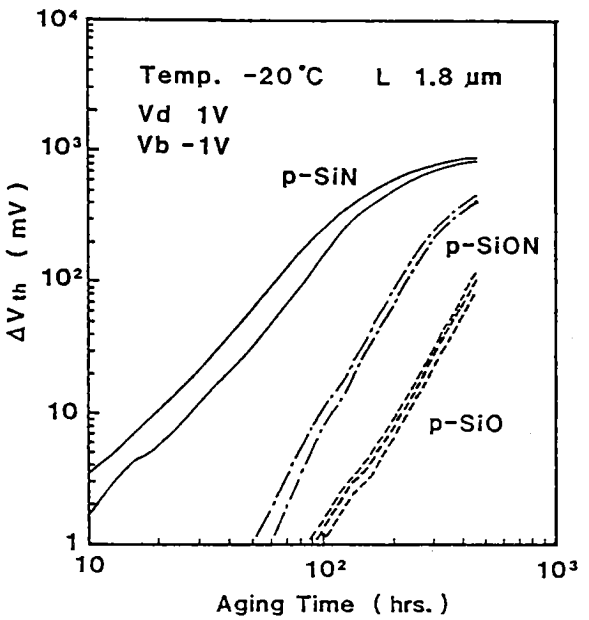


Fig. 9-70 Threshold voltage shift in MOSFETs under hot-carrier stressing with three different passivation layers.[155] (© IEEE 1986.)

nitrogen is used as the sintering ambient.[154] Another study showed that the susceptibility of the MOSFET to hot-carrier degradation after a hydrogen sinter step could be reduced by performing an additional anneal in nitrogen. This nitrogen anneal apparently permits the hydrogen introduced in the first sinter step to be released. Again, the conclusion is that excess hydrogen in the gate oxide is involved in increasing hot-carrier degradation.[189]

Devices encapsulated with PECVD nitride films have also been reported to exhibit larger hot-carrier degradation effects than those encapsulated with PECVD SiO_2 (see Fig. 9-70). The PECVD SiN films are known to contain large amounts of hydrogen. It has been proposed that hydrogen atoms from such encapsulation layers diffuse into the gate oxide layer and generate hot electron traps. The SiN layer must be present on the device during electrical stressing in order to induce the majority of the instability.[155]

9.14.1.2 Hydrogen introduced at high temperatures. Experiments have demonstrated that hydrogen introduced into MOSFET gate oxides at high temperatures (e.g. ≥800°C) exacerbates the hot-carrier degradation observed when hydrogen is introduced at lower temperatures.[156] That is, anneals in H_2 at 900°C for one hour reduce the charge-to-breakdown Q_{BD} in MOS-Cs by an order of magnitude, and also cause the electron trap density to increase by close to an order of magnitude. Finally, the interface state density after hot-carrier stressing also increases by an order of

magnitude in the hydrogen-annealed structures (see Fig. 9-71). In this figure we also see that exposure of the wafers to an 825°C hydrogen-containing ambient (in an LPCVD SiN deposition process) causes a smaller, but still significant, increase in the interface trap density. One conclusion from these observations is that operations which expose the gate oxide to hydrogen at such high temperatures should, if possible, be avoided. One example of avoiding high temperature hydrogen involves the replacement of PSG with BPSG as an interlevel dielectric film. This switch is beneficial because the reflow of BPSG (a process used to planarize and smooth contact hole steps in the interlevel dielectric prior to metal deposition) can be performed in a nitrogen ambient rather than in the steam ambient required for the reflow of phosphosilicate glass.

A possible model for why high-temperature hydrogen makes the $Si\text{-}SiO_2$ interface more susceptible to hot-carrier damage is that the Si-H or Si-OH bonds in the oxide bulk and in the $Si\text{-}SiO_2$ interface act as precursors for the formation of new electron traps. These new traps might be generated when H bonds are broken, either directly by the energetic carriers, or indirectly by a reaction of the bonded H with atomic hydrogen (whereupon molecular hydrogen is released). In either case, a dangling bond is formed in the oxide. It should be mentioned that Si-H bonds are also present in samples which are subjected only to the low temperature 400°C sintering anneal but not to the higher-temperature hydrogen step, which in this case is the 825°C LPCVD SiN deposition process. At 400°C, however, the hydrogen is likely to form Si-H bonds only with silicon atoms which have a dangling bond. At higher temperatures, however, hydrogen may react with strained Si-Si or Si-O bonds to form additional H bonds. This model is

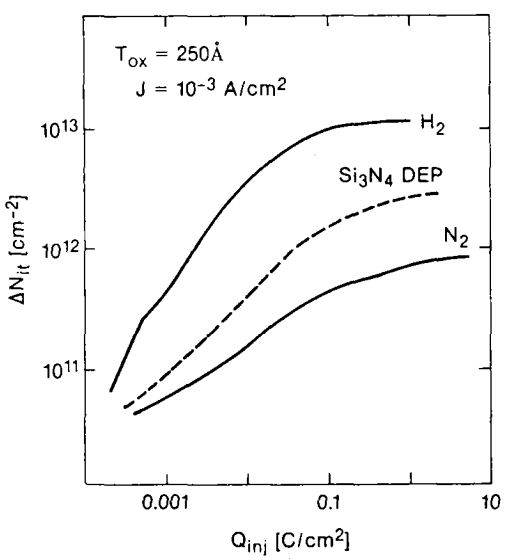


Fig. 9-71 The generated interface trap density ΔD_{it} as a function of the injected charge at tunneling injection, for hydrogenated (H_2) and control (N_2) wafers, and for a wafer which was exposed to 825°C LPCVD of Si_3N_4 (dashed line). The Si_3N_4 layer of 1000°C was deposited after PSG deposition and immediately removed by a dry etch.[156] (© IEEE 1988.)

supported by SIMS results which showed that oxides annealed in deuterium (an isotope of hydrogen) at 900°C had a deuterium content 40 times greater than the content in oxides which were annealed in deuterium at 400°C.

9.14.1.3 Moisture in Gate Oxides as a Contributor to Hot-Carrier Degradation. The presence of H_2O in gate oxides has also been found to increase their susceptibility to hot-carrier degradation, for reasons closely related to those involving atomic hydrogen. The degradation is thought to be due to the diffusion of either H, OH, H_2, or H_2O whereupon these species form Si-H and Si-OH bonds at the Si-SiO_2 interface (which are precursors of interface states [•Si•] generated by hot-carrier stress).[157]

Early experimental studies showed that if minute quantities of water-related impurities (such as H and OH) are introduced into gate oxides, hot-carrier degradation of these oxides is accelerated.[158] Later, it was observed that ultra-dry oxides (gate oxides grown in ambients of less than 1 ppm water content) exhibit significantly less hot-carrier degradation than do oxides grown in conventional dry-oxide growth processes.[157] This result again indicates that removing water and hydrogen from the gate oxide and from high-temperature fabrication processes following the gate oxide formation improves MOS device endurance against hot carriers. Ohji also noted that after gate oxide growth the ultra-dry oxides were nevertheless exposed to hydrogen (i.e., during the 620°C poly CVD process, as a result of the pyrolysis of SiH_4). However, the ultra-dry oxides still exhibited less hot-carrier degradation than the conventional dry oxides. This suggests that hydrogen may have a less damaging effect than water. That is, hydrogen may diffuse rapidly through the oxide to terminate interface traps, but may not remain in the oxide itself, whereas water may be retained as part of the SiO_2 network.[159]

It is also appears that water and/or SiOH can diffuse into gate oxides from PECVD TEOS films[160] or from spin-on glass (SOG) films.[161] Such moisture is incorporated in the TEOS films as a result of the deposition process, and into the SOG films by absorption from the atmosphere after deposition.

9.14.2 Gate-Oxide Damage by Metal Etching

It has also been reported that gate-oxide hot-carrier reliability is worsened by radiation damage generated by RIE etching of the second-layer metal of a two-level metal interconnect. A sinter temperature of 375°C following this etch process was not high enough to anneal the damage.[162] That is, after hot-carrier stressing a large increase in the interface state density was observed after hot-carrier stressing and the metal etch step, and the 375°C sinter. While raising the sinter temperature to 450°C improved the device aging characteristics, the damage was not entirely removed unless the sinter temperature was increased to more than 500°C. This not only indicates that the device degradation is due to a buildup of radiation damage caused by second-layer metal patterning, but also that developing a diffusion-barrier-layer technology that can withstand sinter temperatures in excess of 500°C would be quite useful.

9.14.3 Nitrided Oxides

Besides reducing the hydrogen and moisture content, another general approach to strengthen gate oxides is to modify their chemical composition. The main such modification involves thermally incorporating nitrogen into the gate oxide films, as is discussed in this section.

9.14.3.1 Thermal nitridation of SiO_2. Thermal nitridation of SiO_2 films (as first introduced by Ito et al.,[166] and Naiman et al.[167]), was carried out by exposing the oxide films to pure ammonia (NH_3) at atmospheric pressure and high temperatures (900-1200°C). This results in the formation of a "nitrided" oxide film with an index of refraction higher than that of the oxide. Analysis by Auger emission spectroscopy (AES) indicates that for nitridation times of ~1 hour most of the film still consists of silicon dioxide, with nitridation occurring primarily at the silicon surface and the silicon/dielectric interface (see Fig. 9-72).[165] The nitridation mechanisms appear to be the addition of nitrogen atoms to the SiO_2 lattice and silicon nitride growth at the silicon-oxide interface. That is, a fast reaction between SiO_2 and NH_3 apparently occurs at the top surface of the oxide film, but NH_3 also rapidly diffuses to the Si-SiO_2 interface, where it reacts with the silicon. In the heavy nitridation process described above (i.e., nitridation carried out in an NH_3 atmospheric-pressure ambient at 900-1200°C), nearly stoichiometric silicon nitride is formed in the surface and interfacial regions. After a few minutes, however, the nitridation reaction mainly continues in the bulk, with the surface and interface nitrogen concentration remaining essentially constant. In the bulk of the oxide layer, only partially conversion to nitride usually occurs, probably because the surface nitrogen layer impedes the diffusivity of the NH_3 into the bulk.[163,164,165]

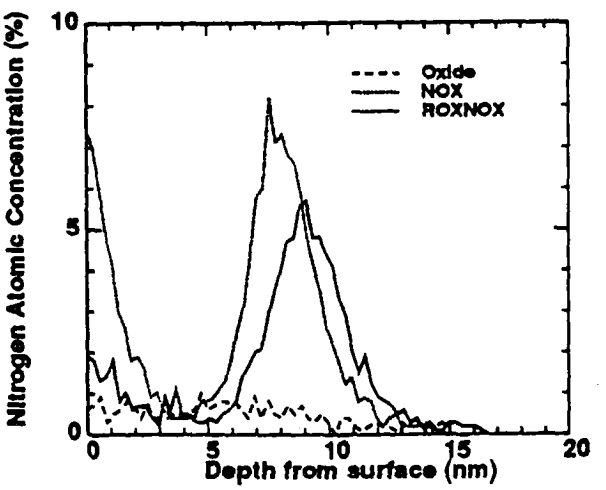


Fig. 9-72 Auger profiles of nitrogen in 11 nm low-pressure furnace nitrided and reoxidized nitrided oxide.[221] (© IEEE 1992.)

9.14.3.2 NH_3 Oxides as MOSFET Gate Dielectrics. When gate oxide films are subjected to such nitridation, they exhibit stronger resistance to interface-state generation (ΔD_{it}) under hot-carrier stress and provide a barrier to various dopants and contaminants (including boron and sodium). The increased stability of the Si-SiO_2 interface is due to the pileup of nitrogen at the interfacial region.[182] The nitrogen atoms are thought to terminate dangling Si bonds at the Si-SiO_2 interface, and/or dangling bonds in the bulk associated with Si vacancies. The Si-N bonds which are formed are harder to break than Si-H bonds. Unfortunately, if the nitridation is carried out as described above (i.e., a 1 hr furnace operation at 1000°C at atmospheric-pressure in an anhydrous-NH_3 ambient), both a high fixed-charge density Q_f[168] and a large number of electron traps Q_{ot} are generally produced in the dielectric film.[169] These charges and traps result in shifted threshold voltages, degraded inversion-layer mobilities, and reduced stability under hot-carrier stressing. Thus, oxide layers nitrided in this manner are not suitable for use as gate dielectrics in MOSFETs.

Later, however, it was discovered that a reoxidation step following nitridation reduces the density of electron traps.[170, 171, 172] (The re-oxidation process consists of exposing the nitrided oxides to a high-temperature oxidizing ambient.) The names given to such gate dielectrics are thus *re-oxidized nitrided oxides*, *ROXNOX*, *RNO*, or *ONO*. (Note that the thickness of the nitrided oxide films remains basically unchanged during the re-oxidation step. That is, as long as the nitrogen-rich layer at the Si-SiO_2 interface remains intact, oxygen is prevented from reaching the silicon and growing more SiO_2 during the reoxidizing cycle. However, if excessive re-oxidation takes place, the dense interface layer may dissipate, permitting thermal oxidation to resume. In such cases, the film thickness will increase during re-oxidation.)

A model for why such re-oxidation reduces the electron trap density was proposed by Hori, Iwasaki, and Tsuji.[173,174] They observed that hydrogen-containing species (i.e., NH_x, -H, and -OH) are formed by the dissociation of NH_3 during the nitridation step. When such species are incorporated into the SiO_2 film they are known to cause electron traps (see Fig. 9-73).[179] Hori et al. also found that hydrogen concentration in SiO_2 films increases both as a function of temperature and nitridation time in NH_3 (see Fig. 9-74), as do flatband voltage shifts. This would explain how the NH_3 nitridation step is responsible for causing electron traps in the oxide, and also the correlation between the increase in hydrogen concentration and flatband voltage shift. Hori et al. also found that the hydrogen concentrations in the nitrided films are reduced during the re-oxidation step (see Fig. 9-74b - where the arrow shows the effect of a re-oxidation at 1150°C for 60 sec of an oxide nitrided at 950°C for 60 sec. - and Fig. 9-75), and that hydrogen concentration levels comparable to those of control oxides can be reached by sufficient re-oxidation. At the same time, they determined that the nitrogen concentration at the Si-SiO_2 interface (which is responsible for strengthening the interface against interface state generation) is only decreased slightly by the re-oxidation step (see Fig. 9-75, which shows the AES profiles of an NH_3-furnace nitrided oxide [a] immediately after nitridation, and [b] after an additional 100 min. 1000°C reoxidation step).[182] Thus, they concluded that "the re-oxidation procedure seems to be one of most promising processes for obtaining nitrided oxides of excellent integrity because it can eliminate the H-

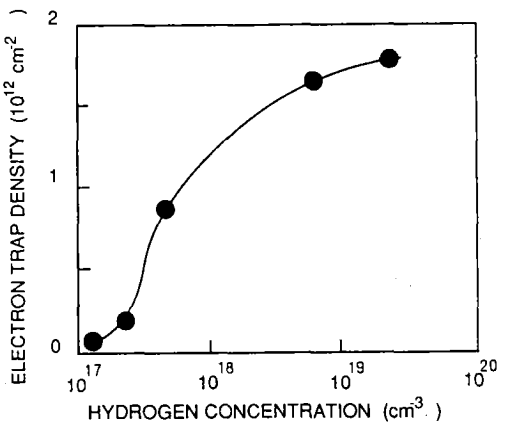


Fig. 9-73 The relationship between hydrogen contained in the oxide and the electron trap density, measured by nuclear reaction analysis and avalanche injection, respectively.[179] (© IEEE 1993.)

containing species while keeping the nitrogen concentrations in the interfacial region of the film unchanged."

9.14.3.3 Low-Pressure and RTP NH_3 Oxides (Lightly Nitrided Reoxidized NH_3 Nitrided Oxides). While reoxidation reduces the density of electron traps in films produced by high-temperature (>900°C) atmospheric-pressure furnace nitridation, the resulting ROXNOX films still exhibit some significant drawbacks. First, the density of electron traps may not be sufficiently reduced in case of either excessive nitridation or insufficient reoxidation. Furthermore, the reoxidation of nitrided oxides formed by heavy nitridation does not reduce the fixed charge densities to levels required by high-performance scaled MOSFETs with high inversion-layer mobilities.

As a result, modifications to the nitridation process have been sought to overcome these limitations, including low-pressure furnace nitridation processes, RTP nitridation and re-oxidation processes, and replacing NH_3 with N_2O in the nitridation step. We will the discuss the first two of these approaches here, and N_2O oxides in the next section. The intent in all of these approaches is to reduce both the concentration of hydrogen and nitrogen in the film. That is, Q_f, D_{it}, and Q_{ot} all increase, peak, and turnaround with increasing nitridation. The buildup and turnaround in these electrical parameters occur more gradually in oxides subjected to low-pressure and RTP nitridation than in atmospheric nitridation.[180] (In fact, early nitrided oxides used heavy nitridation to exploit the turnaround in Q_f to minimize the fixed positive charge. Despite this approach the value of Q_f in such films generally nevertheless remains higher than the levels required by high-performance MOSFETs with high inversion-layer mobilities.[181]) Furthermore, when coupled with a reoxidation step, low-pressure and RTP nitrided oxides exhibit virtually no interface-state generation under hot-carrier stress.[152] Thus,

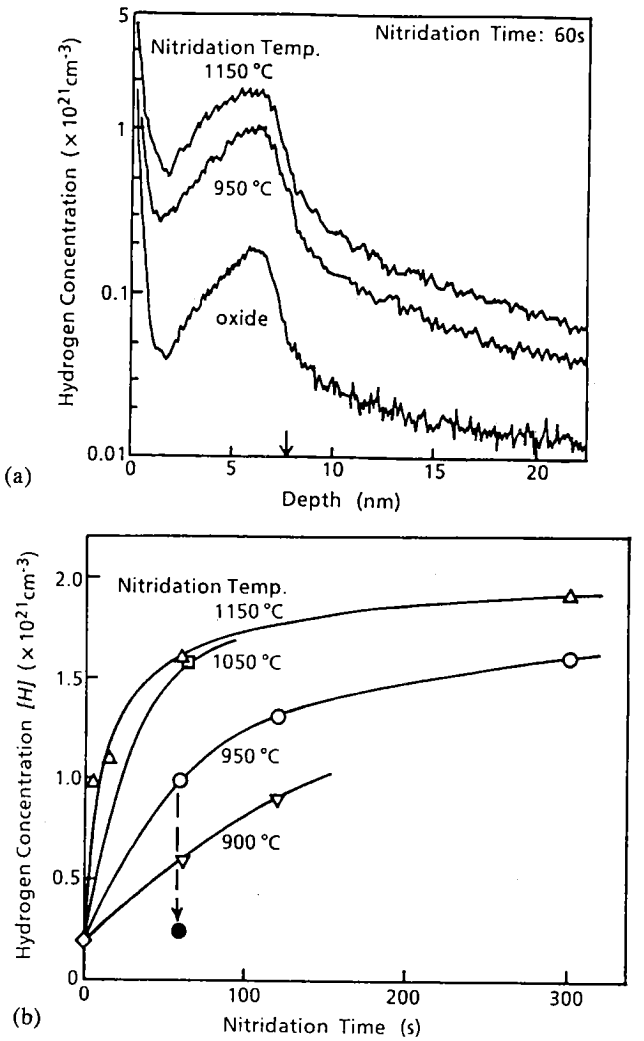


Fig. 9-74 (a) Hydrogen SIMS depth profiles of an oxide and oxides nitrided at 950 and 1150°C for 60 sec. The thickness of the starting oxide is 8 nm. The hydrogen concentration increases monotonically as temperature increases. (b) Hydrogen concentration [H] versus nitridation time for oxides nitrided at 900, 950, 1050, and 1150°C. [H] increases monotonically as nitridation proceeds. The dashed arrow and ▪ in the figure indicate the effect of reoxidation at 1150°C for 60 sec following a nitridation at 950°C for 60 sec.[173] (© IEEE 1988.)

relatively light nitridations in the pre-turnaround regime can be used to optimize the electrical properties of the dielectric. A comprehensive study of the effect of interfacial nitrogen concentration [N_{it}] on MOSFET characteristics has shown that optimum MOSFET device behavior is obtained with *very lightly nitrided* oxides, or when [N_{it}] is in the range of 0.2-1 atomic percent, as obtained by SIMS.[204]

In the low-pressure furnace processes the maximum temperature was also limited to 950°C[175] and 850°C,[176] respectively. To form a 10-nm-thick gate oxide with the 850°C low-pressure furnace nitridation process, the initial oxide film growth time is 55 minutes. This is followed by a 1 hour nitridation step in NH_3 at 0.1 atmospheric pressure, and a 3 hour re-oxidation step. The reoxidation step reduces the nitrogen concentration at the surface and bulk of the film, but without significantly reducing the nitrogen concentration near the interface, or without increasing the film thickness. Such oxides exhibit less than 8 atomic percent of nitrogen in the interfacial nitride region, and a negligible concentration of nitrogen in the oxide bulk.

Several RTP ROXNOX processes have also been reported.[174,177,178,207,208] An example of the Si, O, and N profiles in a 10 nm-thick ROXNOX film formed by one of these processes is shown in Fig. 9-76. We see that the nitrogen concentration at the Si-

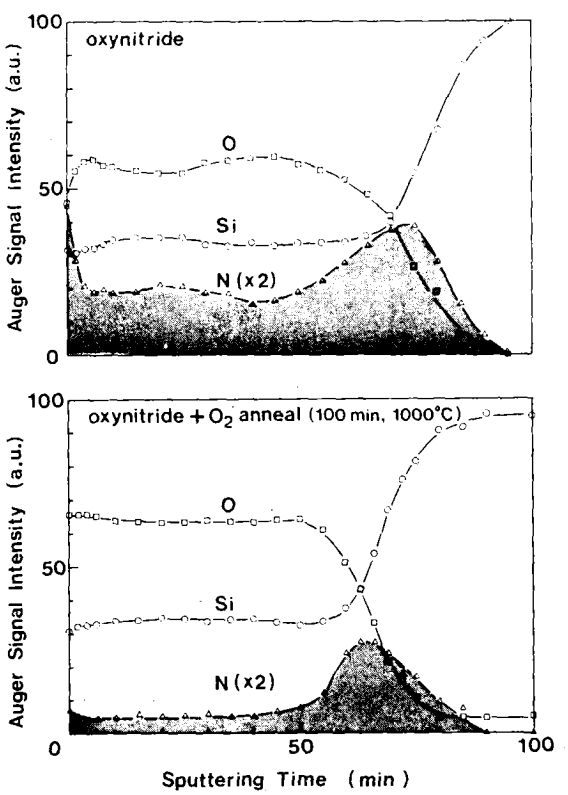


Fig. 9-75 Auger electron spectroscopy profiles of oxygen, silicon, and nitrogen in a furnace nitrided film immediately after NH_3 nitridation, and after reoxidation.[182] (© IEEE 1988.)

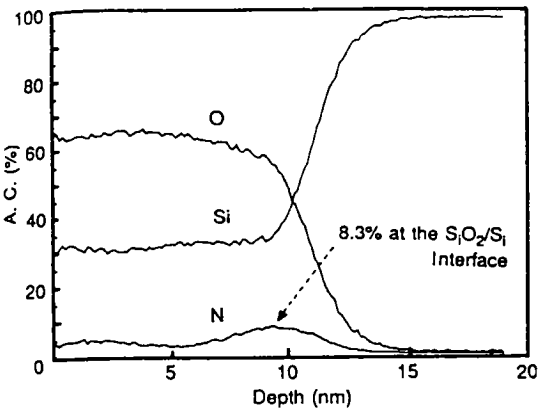


Fig. 9-76 Si, O, and N profiles in a10-nm-thick RTP ROXNOX film.[177] (© IEEE 1986.)

SiO_2 interface is lower than that found in an atmospheric-pressure furnace-nitrided oxide. The RTP reoxidation is also effective in reducing the hydrogen concentration resulting from the RTP NH_3 nitridation step, as is shown in Fig. 9-77. Note that in some RTP processes the oxide is grown in a furnace operation, while in others RTP is also used to grow the initial oxide films. The range of temperatures and times used to nitridate thin gate oxide films using RTP is 950-1150°C and 15-300 sec, respectively. The range of temperatures used to reoxidize these nitrided films with RTP and dry oxygen is also 950-1150°C. Typically, RTP reoxidation times are 60 sec, and temperatures at the high end of this range are used. The specific processing conditions for a particular gate dielectric

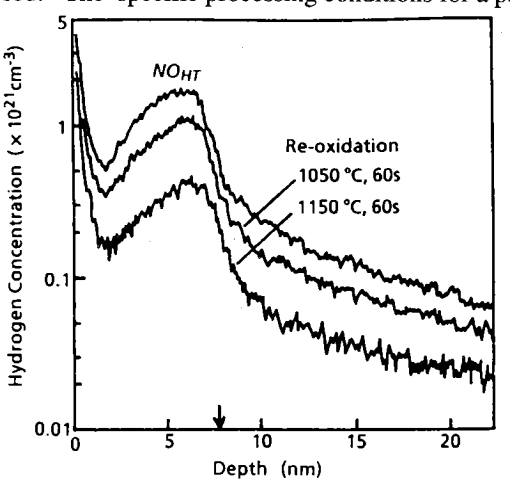


Fig. 9-77 Hydrogen SIMS depth profiles of nitrided oxide (NO_{HT}) and nitrided oxide reoxidized at 1050°C and 1150°C for 60 sec. An arrow indicates the position of the Si-SiO_2 interface.[174] (© IEEE 1989.)

film must be established by trading off electrical performance parameters because each combination of nitridation and reoxidation conditions will yield a different set of Q_f, D_{it}, and Q_{ot} values.

Sufficient reoxidation can return the fixed charge density back to the original values of a pure oxide. However, such reoxidation may be excessive, insofar as it may decrease the interface immunity to hot-carrier stress and/or it may lead to a relatively large increase in the film thickness, for the reasons cited earlier. Thus, ROXNOX films are generally reoxidized only to a degree in which the fixed-oxide-charge density remains somewhat higher than that of pure oxides. These fixed charge levels contribute to mobility degradation and reduction of the transconductance. Hence, post-reoxidation annealing in an inert ambient has been explored as a technique to further reduce Q_f without degrading the other electrical parameters of the ROXNOX film. One report examined the effects of RTP reoxidation and inert post-reoxidation annealing.[183] In another, a post-reoxidation anneal was added to the 850°C low-pressure furnace nitridation process described earlier, and has been is named a *fixed-charge anneal* (FCA).[184,185] The FCA is a nitrogen anneal which is performed after the polysilicon gate film is deposited, but before the poly doping implant step. Since it appears that it takes only a short time to significantly reduce Q_f in ROXNOX films once the temperature reaches 1000°C, the FCA can be carried out in either a furnace or by RTP. In ref. 184, 20 and 30 sec RTP anneals in N_2 at 1000°C were used with 45Å- and 65Å-thick films, respectively. The effective mobility of electrons is also increased by more than 12% in such FCA-treated ROXNOX NMOSFETS, while hole mobilities are increased by more than 30%. This is due to the reduction of Q_f in the films.

9.14.3.4 Oxynitridation of Silicon in N_2O Ambients. Although ROXNOX films exhibit better hot-carrier resistance than do pure oxides, some practical issues make them troublesome to manufacture. First, the ROXNOX formation sequence is complex and difficult to optimize, especially for CMOS applications. For example, the combination of nitridation and reoxidation steps used in the ROXNOX process may leave excess hydrogen in the film. In that case, the density of electron traps will be higher than in a pure oxide film. Since electron trapping is the dominant degradation mechanism in PMOS devices, PMOSFETs with such films will exhibit worse hot-carrier degradation.[216] On the other hand, if the reoxidation step is excessive, the film can lose its strength against interface-state generation by hot-carriers. Thus, NMOSFETs containing such films would be prone to more hot-carrier degradation. The gate dielectric thickness might also grow too thick. Another drawback is that the best NH_3-nitrided films always exhibit higher densities of positive fixed charge (Q_f) than do the best SiO_2 films. The fixed charge arises as a result of the nitridation process. Finally, the ROXNOX process generally involves multiple high-temperature steps, which are undesirable for ULSI CMOS applications.

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

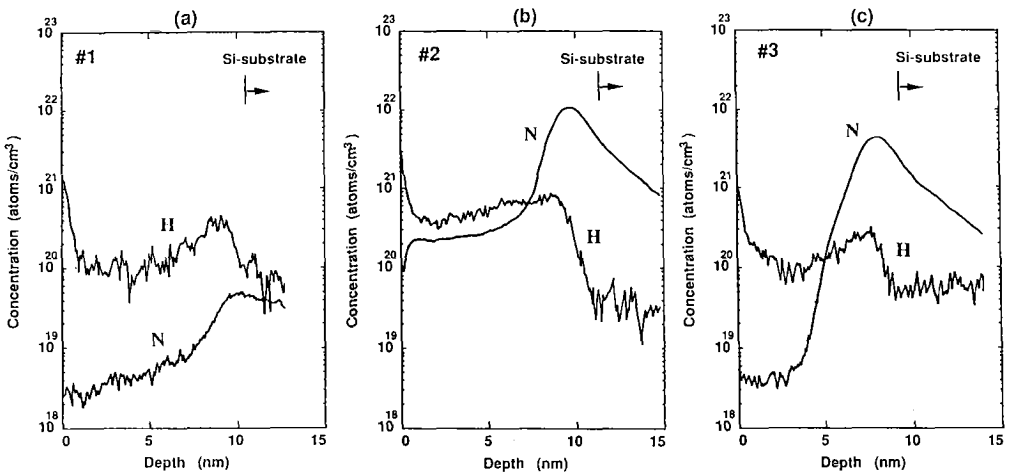


Fig. 9-78 SIMS depth profiles of N and H in three kinds of oxide films: (a) pure SiO_2 (#1); (b) NH_3-nitrided SiO_2 (#2); and (c) N_2O-nitrided SiO_2 (#3). (© IEEE 1990).

This approach was first employed to grow a gate dielectric by directly reacting N_2O at the silicon surface in an RTP system.[190,191,196,199] Since there is no hydrogen in N_2O, this process results in films with significantly less incorporated hydrogen than NH_3-nitride films. Figure 9-78 shows SIMS profiles of dielectric films prepared by: (a) thermally oxidizing silicon; (b) nitriding an oxide film in NH_3; and (c) growing a film on silicon by RTP in an N_2O ambient.[191] We see that the hydrogen concentration in the N_2O film is significantly smaller than in NH_3 film. It is also evident that the nitrogen concentration in the N_2O film is highest at the Si-SiO_2 interface. Early studies of such films concluded that they are self-limiting in thickness, as is shown in Fig. 9-79a.[190,200] The build up of the nitrogen concentration was thought to block the oxidation reaction that occurs simultaneously at the interface. Later work showed that the thickness was not self-limiting (see Fig. 9-79b),[201,202] and that the thickness in fact, increases approximately as $(\text{time})^{1/2}$. This seems to indicate that the growth is instead limited by the diffusion of N_2O through the oxide bulk (rather than through the nitrogen-rich layer).[202] The discrepancy in the apparent growth mechanisms may have been explained by Okada et al., who observed significant retardation of the growth only at relatively high (~0.9 at %) interfacial nitrogen concentrations, which is higher than observed in some N_2O furnace processes.[203] However, the interfacial reactions may still dominate growth rates for short times and thin oxides.

Nevertheless, the nitrogen profile in the N_2O film is similar to that in the ROXNOX film, yet this is achieved with only a one-step process (i.e., no reoxidation is necessary). The peak nitrogen concentration (at the Si-SiO_2 interface) is ~0.5%, which is less than that found in furnace-nitrided in NH_3 under atmospheric pressure (2-10 at.%). Further-

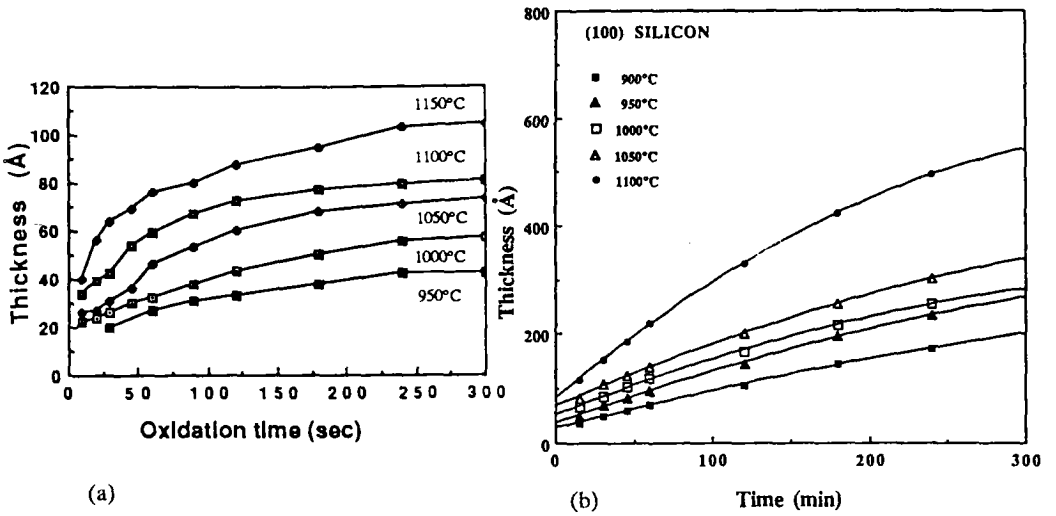


Fig. 9-79 (a) N_2O oxynitride thickness versus oxidation time using RTP for various process temperatures. Ramping rate was 200°C/sec. Oxide thickness was measured at a fixed refractive index (1.46).[190] (© IEEE 1990.) (b) N_2O oxynitride thickness versus time in a furnace anneal process. (© Electrochemical Society 1993.)[201]

more, residual hydrogen incorporation into the film is no longer a concern. This is evidenced by the low electron trapping density exhibited by such films. NMOSFETs fabricated with such films also exhibited less tendency toward hot-carrier degradation than those made with control oxides (i.e., they showed smaller values of ΔD_{it}), and were also less susceptible to the creation of electron traps throughout the oxide under hot-carrier stressing.[192]

However, several limitations were still observed in these N_2O-grown oxynitrides. First, larger initial values of D_{it} were observed than in pure oxide films (Fig. 9-80), even though the rate of their generation under hot-carrier stressing was much smaller (see also Fig. 9-80).[192] Second, the initial Q_f values were also somewhat higher. Third, the slow growth-rate of such films make them only practical for ULSI applications which use dielectric films thinner than 100Å (see Fig. 9-79). Fourth, the maximum mobility of electrons in MOSFETs made with such films is ~5% less than in MOSFETs with oxide gate dielectrics (but the mobility is actually higher at large lateral electric fields - see Fig. 9-81).[193] The most severe limitation, however, is that such RTP N_2O-grown films exhibit unacceptably large nonuniformities in composition and thickness across a wafer.[194] Such non-uniformities would limit the use of such films in ULSI MOS devices. Temperature nonuniformities in the RTP chamber, gas depletion, and heat-transfer effects were some of the phenomena suspected of leading to these nonuniformities.

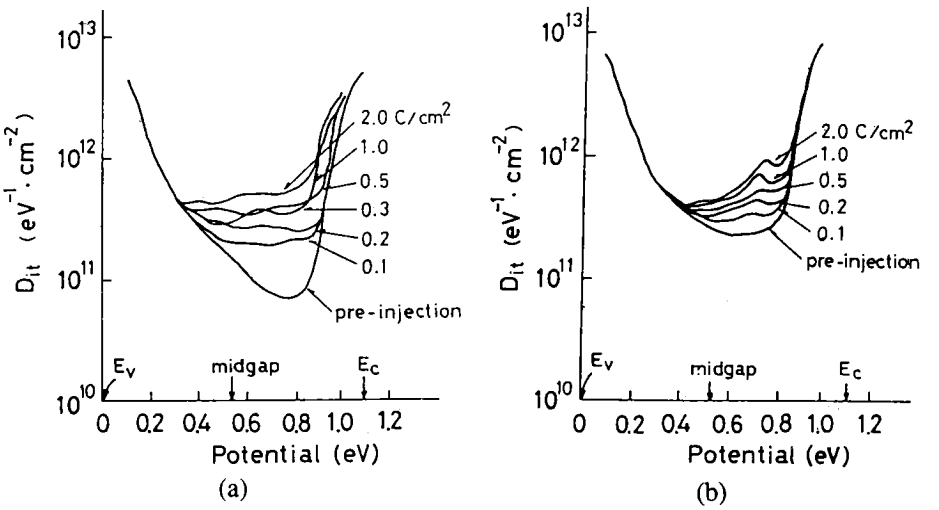


Fig. 9-80 Changes in the interface trap densities of MOS capacitors with (a) pure SiO_2 and (b) SiO_XN_Y by Fowler-Nordheim electron injection. Electrons were injected from the gate electrodes into the oxide films at the current density of ~10 mA/cm^2.[192] (© IEEE 1991.)

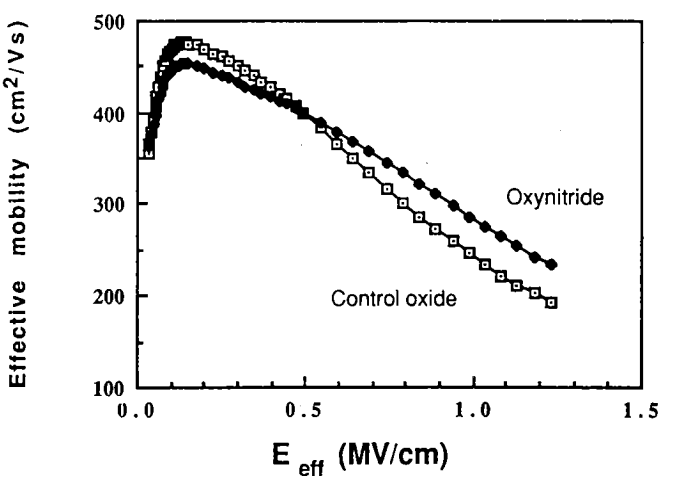


Fig. 9-81 Effective electron mobility versus effective normal field of NMOSFETs with 65Å-thick gate dielectric. The mobility values were calculated using the linear region I_D versus V_{GS} curves (V_{DS} = 100 mV), with a Z/L' ratio of 75-μm/ 60-μm. The maximum effective mobility of the oxynitride is 5% lower than that of the control oxide. However, the high-field mobility of the oxynitride shows a 10% improvement.[193] (© IEEE 1991.)

Nevertheless, it was found that growth of such N_2O dielectrics in a furnace (rather than by RTP) resulted in films with thickness uniformities comparable to those of pure oxide films, and with adequate compositional uniformity as well.[195,197] The process described in these reports was a 950°C furnace step in N_2O, which grew 60Å-thick dielectric films having a nitrogen concentration at the Si-SiO_2 interface of ~3%. Hot-carrier degradation of PMOSFETs with such films was excellent since electron-trap densities in these films was also low.

9.14.3.5. Oxynitridation of Oxides in N_2O. Unfortunately, the use of an N_2O furnace process to directly grow a dielectric film on bare silicon still results in layers that were too thin for reasonable thermal budgets. That is, at 950°C only 50Å films are grown in 40 minutes. To increase the oxide thickness, higher oxidation temperatures and/or extended times would be required. However, such high-thermal-budget processing steps are undesirable in ULSI applications. In addition, the charge-to-breakdown of such direct-N_2O furnace oxides is significantly degraded.[202,205,206]

To overcome these limitations, it was decided to instead investigate a two-step operation, i.e., in which oxides are first grown in O_2, and are then subjected to an oxynitridation step in N_2O. When N_2O is used as a post-oxidation annealing ambient, a wide range of final oxide thicknesses is achievable with much smaller thermal budgets. Processes using this two-step N_2O sequence have been described using atmospheric-pressure furnace,[198,202,209,217] low-pressure-furnace,[210] and RTP[211] conditions. Figure 9-82 shows an example of such a two-step process performed in a conventional furnace at atmospheric pressure.[209] Typically, the oxide is grown in dry O_2 in a furnace at 850-900°C. The N_2O step is usually carried out at 800-850°C for 5-40 minutes. In the process described in ref. 217, 50Å SiO_2 films are grown in dry O_2 at 850°C for 30 minutes. During the second step in N_2O at 950°C, an additional 35Å of oxide is grown to form films with a final thickness of 85Å.

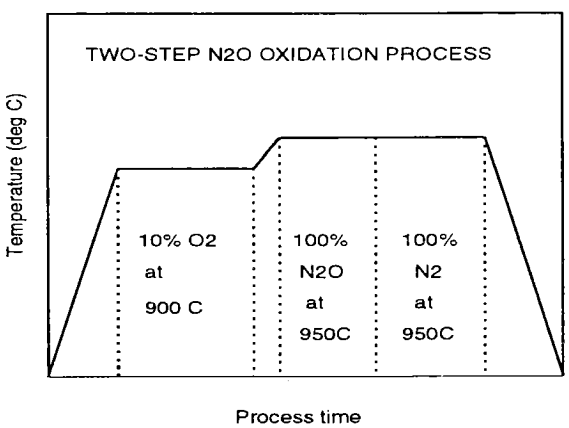


Fig. 9-82 Two-step oxidation and N_2O nitridation process in a conventional furnace.[209] (© IEEE 1993.)

PMOSFETs made by oxynitridation of SiO_2 in N_2O also exhibit enhanced hot-carrier immunity compared to devices made with pure oxide. This is thought to be due to the fact that electron trapping is suppressed in devices with such N_2O dielectrics.[218] Consequently, it has been suggested that the optimum N_2O oxynitride process is oxidation in O_2 followed by nitridation in N_2O, rather than oxidation in N_2O alone.

In the low-pressure approach[210] it was reported that a minimum pressure of between 1 and 50 torr of N_2O was needed before a measurable amount of nitrogen was incorporated into the oxide. However, at low pressures of N_2O just beyond the minimum needed for the onset of nitrogen incorporation, the concentration of nitrogen at the interface could be well controlled, but at the same time, minimal additional oxide was grown during the oxynitridation step. Thus, tight control of the final film thickness could be maintained.

As noted in a previous paragraph, the oxide film continues to grow during the second (i.e., N_2O) step. However, if re-oxidation of N_2O oxides is carried out after they are formed, unusual behavior occurs. That is, new SiO_2 grows at the $Si\text{-}SiO_2$ interface, and the existing N-rich layer remains intact and is displaced away from the interface by growth of the new oxide (see Fig. 9-83).[223] The resulting film does not have as good hot-carrier resistance as films that have not been reoxidized, and hence the step seems to be pointless. However, understanding the phenomenon that takes place during reoxidation of the N_2O oxide is important because it is something that does occur when the poly reox step is carried out after poly patterning. That is, a sidewall oxide is grown

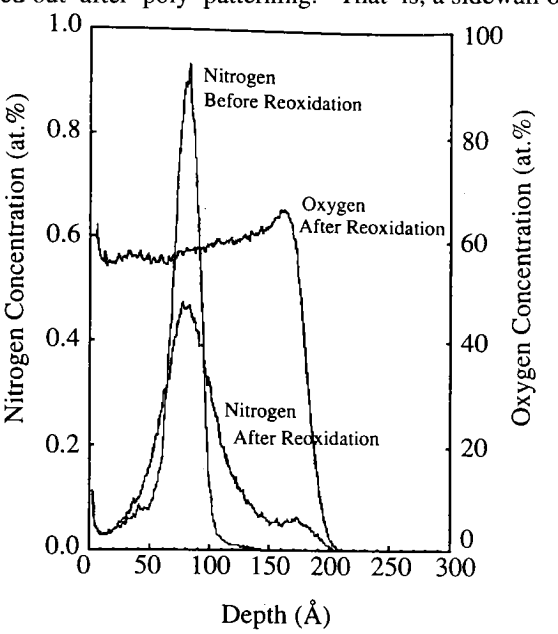


Fig. 9-83 SIMS profile of the oxynitride film grown from 100% N_2O at 1100°C before reoxidation, and after 85 min reoxidation.[223] Reprinted with permission of the Electrochemical Society.

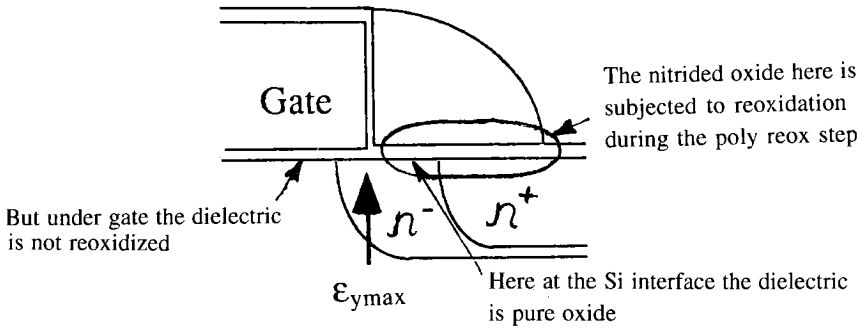


Fig. 9-84 LDD MOSFET with N_2O nitrided oxide as gate deielectric. The position of ε_{ymax} is such that hot carriers are injected into the gate dielectric under the spacer. If the gate dielectric at that location has been reoxidized such that pure oxide exists instead of nitrided oxide at the silicon-dielectric interface, the device will exhibit the same hot-carrier resistance as an LDD MOSFET made with a pure oxide.

during the poly reox step. Thus, in fact, it also represents a reoxidation of the N_2O oxide that will exist under the spacers of an LDD MOSFET. The reox step therefore grows an oxide under the N-rich layer of the oxide film that will exist under the spacers (see Fig. 9-84). Now, hot carriers are injected into the dielectric and spacers near the polysilicon edge, as shown in Fig. 9-84 (where the position of the simulated peak electric field of an LDD in which the n^+ region does not overlap the gate edge, is depicted). The region of maximum hot-carrier damage in the device is located at that point. Since the oxide at the Si-SiO_2 interface is no longer an N_2O oxide, but has been returned to a pure oxide, LDD MOSFETs fabricated with such reoxidized N_2O gate dielectrics exhibit hot-carrier degradation characteristics no better than LDD MOSFETs made with a pure oxide gate dielectric.[224]

A novel device structure has been proposed that would mitigate this effect. That is, we noted earlier that one form of degradation caused by the fixed oxide charge in nitrided oxides is a reduction in the peak channel mobility of both P- and NMOSFETs. This is apparently due to the presence of nitrogen at the interface.[221] While special process techniques (such as the fixed-charge anneal[185]) have been used to reduce this problem, a novel LDD structure has also been reported to overcome it. That is, a pure gate oxide is used under the gate of a MOSFET to yield the highest mobility of carriers in the channel, but a nitrided oxide process is used to reoxidize the poly sidewall after patterning (see Fig. 9-85). Since most gate-oxide hot-carrier-damage occurs outside of the channel region (i.e., under the spacer), by having a strengthened interface in those locations the device simultaneously exhibits significantly improved hot-carrier resistance and a higher drive-current capability compared to an LDD MOSFET made with an oxide spacer and a conventional poly reox process.[222]

We should also mention that adding nitrogen to gate oxides has also been reported by several other novel techniques, including implantation of nitrogen through the polysilicon gate,[212] N_2O nitridation of CVD oxides,[213] oxidation of deposited oxide on nitrided surface (ODON),[214] and oxidation of the silicon surface beneath a CVD oxynitride.[215]

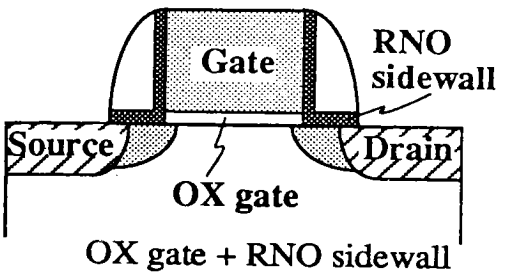


Fig. 9-85 Novel LDD MOSFET that has a pure oxide gate dielectric and a reoxidized nitrided poly sidewall. The latter feature is created during the reox of the poly (i.e., after poly patterning, but prior to spacer dielectric film deposition).[222] (© IEEE 1991.)

9.14.3.6 Fluorinated Oxides. The presence of fluorine in the gate oxide has also been reported to increase the hot electron resistance of devices fabricated with such oxides.[225,226] Such fluorine may inadvertently be introduced from a BF_2^+ source/drain implant, or from an LPCVD-W or LPCVD WSi_X process. In the latter two processes fluorine is produced during the reaction of WF_6 and SiH_4.[227] Other reports confirm that a deliberate incorporation of fluorine into the gate oxide (from implanting fluorine into a polysilicon film, and then diffusing it into the gate oxide) produces a more hot-electron resistant interface.[228,229] The increased hot-carrier resistance of the F-containing oxides is thought to be due to fluorine atoms which form Si-F bonds. These are stronger than Si-H bonds. The presence of the fluorine implies that some of the Si-H bonds are replaced by Si-F bonds, thus yielding a stronger oxide.

Unfortunately, the presence of fluorine in the gate oxide increases the penetration of boron (from p^+ poly gates) though the gate oxides. Although various approaches have been examined to suppress this effect, many of them rely on methods that prevent fluorine from entering the gate oxide. As a result, use of nitrided oxides to suppress hot-carrier degradation has been pursued much more vigorously than use of fluorinated oxides. Nevertheless, a recent report has indicated that MOSFETs using a WSi_X polycide process exhibit significantly improved gate oxide hot-carrier hardness and higher carrier mobility as a result of the fluorine incorporated by the WSi_X polycide process.[230]

REFERENCES

1. T.Y. Chan, P.K. Ko, and C. Hu, "Dependence of channel electric field on device scaling," *IEEE Electron Dev. Lett.*, **EDL-6**, October 1985, p. 551.
2. P.K. Ko, R.S. Muller, and C. Hu, "A Unified Model for Hot-Electron Currents in MOSFETs," *Tech. Dig. IEDM,* 1981, p. 600.
3. C. Sodini, P.K. Ko, and J.L. Moll, "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Trans. on Electron Dev.,* **ED-31**, October 1984, p. 1386.

4. D.K. Schroeder, *Advanced MOS Devices*, Vol VII in the Modular Series on Solid-State Devices, Reading MA, Addison-Wesley, 1983.
5. P.K. Chatterjee, *Tech. Dig. IEDM*, 1979, p. 14.
6. M.L. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings IEEE,* December 1986, p. 1715.
7. T.Y. Chan, P.K. Ko, and C. Hu, "A Simple Method to Characterize Substrate Current in MOSFETs," *IEEE Electron Dev. Letts.*, **EDL-5**, December 1984, p. 505.
8. P. K. Ko, "Hot Electron Effects in MOSFETs," Doctoral Thesis, Dept. of EECS, University of California, Berkeley, June 1972.
9. J. Chung et al., "New Insight into Hot-Electron Currents in Deep-Submicrometer MOSFETs," *Tech. Dig. IEDM,* 1988, p. 200.
10. C. Hu et al., "Hot-electron-induced MOSFET degradation - model, monitor, and improvement," *IEEE Trans. Electron Dev.*, **ED-32**, February 1985, p. 375.
11. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation due to Hot Carrier Injection," *IEEE Trans. Electron Dev. Letts.*, **EDL-4**, p. 111, 1983.
12. E.H. Nicollian and J.R. Brews, *MOS Physics and Technology*, Wiley, 1983.
13. K.R. Hoffmann, W. Weber, C. Werner, and G. Dorda, "Hot-Electron and Hole-Emission Effects in Short N-MOSFETs," *IEEE Trans. Electron Dev.*, **ED-32**, March 1985, p. 691.
14. M. Yoshida et al., "Increase in Resistance to Hot Carriers in Thin Oxide MOSFETs," *Tech. Dig. IEDM*, 1985, p. 254.
15. J. Chung et al., "Hot-Electron Currents in Deep-Submicrometer MOSFETs," *Tech. Dig. IEDM*, 1988, p. 200.
16. P. Woerlee et al., "The Impact of Scaling on Hot-Carrier Degradation and Supply Voltage of Deep Submicron NMOS Transistors," *Tech. Dig. IEDM,* 1991, p. 537.
17. C. Hu, "Hot-Carrier Effects," Chap. 3, in *Advanced MOS Device Physics*, N.G. Einspruch and G. Gildenblat, Eds., Vol. 18, VLSI Electronics Microstructure Science, Academic Press, San Diego, CA, 1989, p. 119-160.
18. C. Hu et al., "Hot-electron-induced MOSFET degradation - model, monitor, and improvement," *IEEE Trans. Electron Dev.*, **ED-32**, February 1985, p. 375.
19. P. Yang and S. Aur, "Modeling of Device Lifetime due to Hot Carrier Effects," *Proc. of 1985 Internat. Symp. on VLSI Tech.*, p. 227.
20. R.A. Chapman et al., "An 0.8 μm CMOS Technology for High Performance Logic Applications," *Tech. Dig. IEDM* 1987, p. 362.
21. T.H. Ning et al., "1 μm MOSFET VLSI Technology: Part IV - Hot Electron Design Constraints," *IEEE Trans. Electron Dev.*, **ED-26**, 346 (1979).
22. T.Y. Chan, C. I. Chiang, and H. Gaw, "New Insight into Hot-Electron-Induced Degradation of *n*-MOSFETs," *Tech. Dig. IEDM,* 1988, p. 196.
23. J.Y. Choi, P.K. Ko, and C. Hu, "Hot-Carrier-Induced MOSFET Degradation: AC versus DC stressing," *Dig. Tech. Papers Symp. VLSI Tech. 1987*, p. 45.
24. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation due to Hot Carrier Injection," *IEEE Trans. Electron Dev. Letts.*, **EDL-4**, p. 111, 1983.
25. R. Woltjer and G. Paulzen, "Universal Description of Hot-Carrier-Induced Interface States in NMOSFETs," *Tech. Dig. IEDM,* 1992, p. 535.

26. F.-C. Hsu and K.Y. Chiu, "Hot-Electron Substrate-Current Generation During Switching Transient," *IEEE Trans. Electron Dev.*, **ED-32**, (1985), p. 375.

27. K.R. Hoffmann, W. Weber, C. Werner, and G. Dorda, "Hot-Carrier Degradation Mechanism in N-MOSFETs," *Tech. Dig. IEDM*, 1984, p. 104.

28. W. Weber, C. Werner, and A. Schwerin, "Lifetimes and Substrate Currents in Static and Dynamic Hot-Carrier Degradation," *Tech. Dig. IEDM,* 1986, p. 390.

29. K.L. Chen, S. Saller, and R. Shah, "The Case of AC Stress in the Hot-Carrier Effect," *IEEE Trans. Electron Dev.*, **ED-33**, March 1986, p. 424.

30. B.S. Doyle et al., "Dynamic Channel Hot-Carrier Degradation of MOS Transistors by Enhanced Electron-Hole Injection into the Oxide," **EDL-8**, May 1987, p. 237.

31. J.Y. Choi, P.K. Ko, and C. Hu, "Hot-carrier induced MOSFET degradation under AC stress," *IEEE Trans. Electron Dev. Letts.*, **EDL-8**, p. 333, 1987.

32. R. Bellens et al., "Analysis of Mechanisms for the Enhanced Degradation During AC Hot Carrier Stress of MOSFETs,"*Tech. Dig. IEDM,* 1988, p. 212.

33. W. Weber et al., "Dynamic Degradation in MOSFETs - Part II: Application in the Circuit Environment," **ED-38**, (1991), p. 1859.

34. K.R. Mistry et al., "Circuit Design Guidelines for *n*-Channel MOSFET Hot Carrier Robustness," *IEEE Trans. Electron Dev.*, **ED-40**, July 1993, p. 1284.

35. J. Winnerl et al., "Influence of Transistor Degradation on CMOS Performance and Impact on Lifetime Criterion," *Tech. Dig. IEDM,* 1988, p. 204.

36. S.C. Hu and M.P. Brassington, "Test Structure to Measure HC Degradation in CMOS Circuits," *IEEE Trans. Electron Dev.*, **ED-38,** August 1991 p. 1958.

37. M.M. Kuo et al., "Simulation of MOSFET lifetime under AC Hot-Electron Stress," *IEEE Trans. Electron Dev.*, **ED-35**, July 1988, p. 1004.

39. S. Aur, D. Hocevar, and P. Yang, "HOTRON - A Hot-Electron Effect Simulator," *Proc. ICCAD,* 1987, p. 256.

40. E.R. Minami et al., "Prediction of Hot-Carrier Degradation in Digital CMOS VLSI by Timing Simulation," *Tech. Dig. IEDM,* 1992, p. 539.

41. K.N. Quader, P.K. Ko, and C. Hu, "Projecting CMOS Circuit Hot-Carrier Reliability from DC Device Lifetime," *Tech. Dig. IEDM*, 1993, p. 511.

42. H. Hazama, M. Iwase, and S. Takagi, "Hot-Carrier Reliability in Deep Submicrometer MOSFETs," *Tech Dig. IEDM*, 1990, p. 569.

43. M.-L. Chen, "CMOS Hot-Carrier Protection with LDD," *Semiconductor Internat.*, April 1988, p. 88.

44. C.Y. Lu et al., "High Performance Salicide Shallow-Junction CMOS Devices for Submicrometer VLSI Application in Twin-Tub VI, **ED-36**, *IEEE Trans. Electron Dev.*, November 1989, p. 2530.

45. E. Takeda, H. Kume, and T. Toyabe, " Submicrometer MOSFET structure for minimizing hot carrier generation," *IEEE Trans. Electron Dev.* , **ED-29**, (1982), p. 611.

46. H. Mikoshiba, "Comparison of Drain Structures in *n*-Channel MOSFETs," *IEEE Trans. Electron Dev.*, Jan. 1986, p. 140.

47. M. Koyonagi, H. Kaneko, and S. Shinizu, "Optimum design of n^+-n- double-diffused drain MOSFET to reduce hot-carrier emission," *IEEE Trans. Electron Dev.*, **ED-32,** (1985), p. 562.

48. S. Ogura et al., "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate FET,"*IEEE Trans. Electron Dev.*, **ED-27**, 1980, p. 1359.

49. P.J. Tsang et al., "Fabrication of High-Performance LDD FETs with Oxide Sidewall Spacer Technology," *IEEE Trans. Electron Dev.*, **ED-29**, April 1982, p. 590.

50. K. Marayam, J. C. Lee, and C. Hu, "A Model for the Electric Field in Lightly-Doped Drain Structures," *IEEE Trans. Electron Dev.*, **ED-34**, July 1987, p. 1509.

51. S. H. Dhong and E. J. Petrillo, "Sidewall Spacer Technology for MOS and Bipolar Devices," *J. Electrochem. Soc.*, Feb. 1986, p. 389.

52. J. Sweeney et al., "Substrate Bias Dependent Leakage in LDD MOSFETs," *Tech. Dig. IEDM,* 1988, p. 230.

53. M. Weling and C. Gabriel, "Contact and Spacer Etch Dependent Junction Leakage in Sub-Micron CMOS Technology," *Ext. Abs. Mtg. Electrochem. Soc.,* Spring 1994, Abs. # 206, p. 324.

54. S. Onishi et al., "Mechanism of the Sidewall Process Induced Junction Leakage Current of LDD Structure," *J. Electro.Chem Soc.,* **Vol. 138**, May 1991, p. 1439.

55. F. Hsu and K. -Y. Chiu, "Evaluation of LDD MOSFETs Based on Hot-Electron-Induced Degradation," *IEEE Electron Dev. Letts.*, **EDL-5** (1984) p. 162.

56. H. Mikoshiba, "Comparison of Drain Structures in *n*-Channel MOSFETs," *IEEE Trans. Electron Dev.*, Jan. 1986, p. 140.

57. H. Ishiuchi et al., "Measurement of Intrinsic Capacitance of Lightly Doped Drain (LDD) MOSFETs, *IEEE Trans. Electron Dev.*, **ED-32**, November 1985, p. 2238.

58. T. Smedes and F.M. Klaassen, "Effects of the Lightly Doped Drain Configuration on Capacitance Characteristics of Submicron MOSFETs," *Tech. Dig. IEDM*, 1990, p. 197.

59. F.C. Hsu and H.R. Grinolds, "Structure-enhanced MOSFET Degradation Due to Hot-Electron Injection," *IEEE Electron Dev. Letts.*, **EDL-5** (1984) p. 71.

60. R. Izawa and E. Takeda, "The Impact of n^- Drain Length and Gate/Drain-Source Overlap on Submircmeter LDD Devices for VLSI," *IEEE Electron Dev. Letts.*, **EDL-8**, October 1987, p. 480.

61. A. Hamada et al., "n^- Source/Drain Compensation Effects in Submicrometer LDD MOS Devices," *IEEE Electron Dev. Letts.*, **EDL-8**, September 1987, p. 398.

62. P. Ko et al., "The Effects of Weak Gate-to-Drain (Source) Overlap on MOSFET Characteristics," *Tech. Dig. IEDM*, 1986, p. 292.

63. P.K. Ko et al., "Enhancement of hot-electron currents in graded-gate oxide (GGO)-MOSFETs," *Tech. Dig. IEDM*, 1984, p. 88.

64. T. Y. Chan, A.T. Wu, P.K. Ho, and C. Hu, "Effects of the Gate-to-Drain/Source Overlap on MOSFET Characteristics," *IEEE Electron Dev. Letts.*, **EDL-8**, July 1987, p. 326.

65. L.C. Parrillo et al., "Disposable Polysilicon LDD Spacer Technology," *IEEE Trans. Electron Dev.*, **ED-38**, January 1991, p. 39.

66. J. Pfiester et al., "Poly-gate sidewall oxidation-induced submicrometer MOSFET degradation," *IEEE Electron Dev. Letts.*, **EDL-10**, August 1989, p. 367.

67. L. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.

68. T. Mizuno et al., "A New Degradation Mechanism of Current Drivability and Reliability of Asymmetrical LDD MOSFETs," *Tech. Dig. IEDM,* 1985, p. 250.

69. Y. Oowaki et al., "Analysis of LDD Transistor Asymmetry Susceptibility in VLSI Circuits," *Tech. Dig. IEDM*, 1985, p. 492.

70. T. Chan, et al. "Asymmetrical Characteristics in LDD and Minimum-overlap MOSFETs," *IEEE Electron Dev. Letts.*, **EDL-7**, January 1986, p. 16.

71. R. Gregor, "Some Consequences of Ion Beam Shadowing in CMOS Source/Drain Formation," *IEEE Electron Dev. Letts.*, **EDL-7**, December 1986, p. 677.

72. T. Hamamoto et al, "Asymmetry of the Substrate Current Characteristics Enhanced by Gate Bird's Beak," *IEEE Symp. VLSI Tech., Dig. of Tech. Papers*, p. 67, 1990.

73. G. Krieger wt al., "Shadowing Effects Due to Tilted Arsenic Source/Drain Implant," *IEEE Trans. Electron Dev.*, **ED-36**, November 1989, p. 122458.

74. J. Pfiester et al., "Asymmetrical High Field Effects in Submicron MOSFETs," *Tech. Dig. IEDM*, 1987, p. 51.

75. F.K. Baker and J. Pfiester, "The Influence of Tilted Source-Drain Implants on High Field Effects in Submicron MOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2119.

76. B.S. Doyle, C. Bergonzoni, and A. Boudou, "The Influence of Gate Edge Shape on the Degradation in Hot-Carrier Stressing of *n*-Channel Transistors," *IEEE Electron Dev. Letts.*, **EDL-12**, July 1991, p. 363.

77. J. Hu and J. Moll, "Submicrometer device design for hot-electron reliability and performance," *IEEE Trans. Electron Dev. Letts.*, **EDL-6**, July, 1985, p. 350.

78. Y. Toyoshima et al., "Mechanism of Hot-Electron Induced Degradation in LDD-MOS," *Tech. Dig. IEDM*, 1984, p. 786.

79. J. Sweeney et al., "Substrate Bias Dependent Leakage in LDD MOSFETs," *Tech. Dig. IEDM*, 1988, p. 230.

80. L.C. Parrillo et al., "A versatile, high-performance, double-level-poly, double-level-metal, 1.2 μm CMOS technology," *Tech. Dig. IEDM,* 1986, p. 244.

81. J. R. Pfiester et al., "LDD MOSFETs using Disposable Sidewall Spacer Technology," *IEEE Electron Dev. Letts.*, **EDL-9,** 1988, p. 189.

82. J. R. Pfiester et al., "A poly-framed LDD sub-half-micrometer CMOS technology," *IEEE Electron Dev. Letts.*, **EDL-11,** 1990, p. 529.

83. L.C. Parrillo et al., "An advanced 0.5 μm CMOS disposable spacer technology," *VLSI Symp. Dig.*, p. 31, 1989.

84. L.C. Parrillo et al., "Disposable Polysilicon LDD Spacer Technology," *IEEE Trans. Electron Dev.*, **ED-38**, January 1991, p. 39.

85. J. R. Pfiester et al., "An integrated 0.5 μm CMOS disposable TiN LDD/Salicide spacer technology," *Tech. Dig. IEDM,* 1989, p. 781.

86. H. Ronkainen et al., "The Use of Disposable Double Spacers and Self-Aligned $CoSi_2$ for LDD MOSFET Fabrication," *IEEE Electron Dev. Letts.*, **EDL-12,** 1991, p. 125.

87. W. Hansch, C. Mazure, A. Lill, and M.K. Orlowski, "Hot-Carrier Hardness Analysis of Submicrometer LDD Devices," *IEEE Trans. Electron Dev.*, **ED-512**, March 1991, p. 512.

88. M. Kinugawa et al., "Submicron MLDD MOSFET for 5V Operation." *1985 Symp. VLSI Tech., Dig. of Tech Papers,* p.116.

89. G. Krieger et al., "Moderately Doped NMOS (M-LDD) - Hot Electron and Current Drive Optimization," *IEEE Trans. Electron Dev.*, **ED-38**, January 1991, p. 121.

90. J.J. Sanchez, K.K. Hsueh, and T.A. DeMassa, "Drain-Engineered Hot-Electron-Resistant Device Structures: A Review," *IEEE Trans. Electron Dev.*, **ED-36**, June 1989, p. 1125.

91. Y. Tsunashima et al., "Metal-coated Lightly-Doped Drain (MLD) MOSFETs for Submicron VLSI," *1985 Symp. VLSI Tech., Dig. of Tech Papers,* p.114.

92. M.-L. Chen et al., "Suppression of Hot-Carrier Effects in Submicron CMO Technology," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2210.

93. H. Grinolds, M. Kinugawa, and M. Kakumu, "Reliability and Performance of Submicron LDD NMOSFETs with Buried As *n*- Impurity Profiles," *Tech. Dig. IEDM,* 1985, p. 246.

94. S. Jain et al., "Sloped-junction LDD (SJLDD), MOSFET structures for improved hot-carrier reliability," *IEEE Electron Dev. Lett.*, Oct 1988, p. 539.

95. C. Wei et al., "Buried and Graded/Buried LDD Structures for Improved Hot-Electron Reliability," *IEEE Electron Dev. Letts.*, **EDL-7**, June 1986, p. 380.

96. J.M. Ford and D. Stemple, "The Effects of Arsenic Drain Profile on Submicrometer Salicide MOSFETs," *IEEE Trans. Electron Dev.* , **ED-35**, March 1988, p. 302.

97. T. Toyoshima et al., "Profiled Lightly Doped Drain (PLDD) Structure for High Reliability NMOSFETs," *Proc. of VLSI Symposium,* 1985, p. 362.

98. R. Simonton, "Ion Implantation Applications in CMOS Process Technology," in *Ion Implantation Technology,* J.F. Zeigler, Ed., Elsevier Science Pub., New York, 1992.

99. C. Codella and S. Ogura, "Halo Doping Effects in Submicron DI-LDD Design," *Tech. Dig. IEDM,* 1985, p. 230.

100. T. Hori and K. Kurimoto, "A New Half-Micron*p*-Channel MOSFET with LATIPS (Large-Tilt-Angle Implanted Punchtrhough Stopper), *Tech. Dig. IEDM,* 1988, p. 394.

101. T. Hori et al., "A Self-Aligned Pocket Implantation (SPI) Technology for 0.2 micron Dual-Gate CMOS," *Tech. Dig. IEDM,* 1991 p. 641.

102. T. Hori et al., "High-Performance Dual Gate CMOS Utilizing a Novel Self-Aligned Pocket Implantation (SPI) Technology," *IEEE Trans. Electron Dev.,* **ED-40**, September 1993, p. 1673.

103. C. Duvvury et al., "Reliability Design of p^+ Pocket Implant LDD Transistors," *Tech. Dig. IEDM*, 1990, p. 215.

104. T. Huang et al., "A novel submicron LDD transistor with inverse-T gate structure," *Tech. Dig. IEDM*, 1986., p. 742.

105. T. Huang et al., "A new LDD transistor with inverse-T gate structure," *IEEE Electron Dev. Letts.,* April 1987., p. 151.

106. M. Stinson and C. Osburn, "Effects of Ion Implantation on Deep-submicrometer Drain-engineered MOSFET Technologies," *IEEE Trans. Electron Dev.,* **ED-38**, March 1991, p. 487.

107. R. Izawa, T. Kure, and E. Takeda, "Impact of the Gate-Overlapped Device (GOLD) for Deep Submicrometer VLSI," *Tech. Dig. IEDM,* 1987, p. 38.

108. R. Izawa, T. Kure, and E. Takeda, "Impact of the Gate-Overlapped Device (GOLD) for Deep Submicrometer VLSI," *IEEE Trans. Electron Dev.,* **ED-35**, December 1988, p. 2088.

109. D.S. Wen et al., "A self-inverse T-gate fully overlapped LDD device for sub-half micron CMOS," *Tech. Dig. IEDM,* 1989, p. 765.

110. J.E. Moon et al., "A New LDD structure: total overlap with polysilicon spacer (TOPS)," *IEEE Electron Dev. Letts.,* **EDL-11**, 1990, p. 221.

111. I.-C. Chen, C.C. Wei, and C.W. Teng, "Simple Gate-to-Drain Overlapped MOSFETs Using Poly Spacers for High Immunity to Channel Hot-Electron Degradation," *IEEE Electron Device Letters*, **EDL-11**, February 1990, p. 78.

112. I. Chen, R. Chapman , and C. Teng, "A Sub-half micron Partially Gate-to-Drain Overlapped MOSFET Optimized for High Performance," *Tech. Dig. IEDM,* 1991 p. 545.

113. M. Chen et al., "Self-Aligned Silicided Inverse -T Gate LDD Devices for Sub-Half Micron CMOS," *Tech. Dig. IEDM*, 1990, p. 829.

114. J. R. Pfiester et al., "An self-aligned LDD/channel implanted ITLDD process with selectively-deposited poly gates for CMOS VLSI," *Tech. Dig. IEDM,* 1989, p. 769.

115. J. R. Pfiester et al., "An ITLDD CMOS process with self-aligned reverse sequence LDD/channel implantation," *IEEE Trans. Electron Dev.,* **ED-38,** Nov. 1991, p. 2460.

116. T. Hori et al., "A new submicron MOSFET with LATID (LArge-Tilt-angle, Implanted Drain) structure," *Symp.VLSI Tech., Dig. of Tech. Papers,* 1988, p. 15.

117. T. Hori et al., "1/4 μm LATID (LArge-Tilt-angle, Implanted Drain) technology for 3.3V operation," *Tech. Dig. IEDM,* 1989, p. 777.

118. F. Matsuoka et al., "Drain Structure Optimization for Highly Reliable Deep Submicron *n*MOSFETs with 3.3V High Performance Operation on the Scaling Trend," *Tech. Dig. IEDM,* 1990, p. 832.

119. T. Hori et al., "Gate-capacitance Characteristics of Deep-submicron LATID MOSFETs," *Tech. Dig. IEDM,* 1991, p. 375.

120. T. Hori et al., "Deep Submicrometer LATID Technology," *IEEE Trans. Electron Dev.,* **ED-39,** October 1992, p. 2312.

121. M. Inuishi et al., "Gate Capacitance Characteristics of Gate/*n*- Overlap LDD Transistor with High Performance and High Reliability," *Tech. Dig. IEDM*, 1991, p. 371.

122. A. Hori et al., "High Carrier Velocity and Reliability of Quarter-Micron SPI (Self-Aligned Pocket Implantation) MOSFETs," *Tech. Dig. IEDM*, 1992, p. 699.

123. T. Hori et al., "High-Performance Dual Gate CMOS Utilizing a Novel Self-Aligned Pocket Implantation (SPI) Technology," *IEEE Trans. Electron Dev.,* **ED-40**, September 1993, p. 1673.

124. K. Kurimoto and S. Odanaka, "A T-Gate Overlapped LDD Device with High Circuit Performance and High Reliability,"*Tech. Dig. IEDM*, 1991, p. 541.

125. M. Inuishi et al., "A High Performance and Highly Reliable Dual Gate CMOS with Gate/n^- Overlap LDD Applicable to Cryogenic Operation," *Tech. Dig. IEDM*, 1989, p. 773.

126. T. Mizuno et al., "Si_3N_4/SiO_2 Spacer Induced High Reliability in LDD MOSFET and its Simple Degradation Model," *Tech. Dig. IEDM,* 1988, p. 234.

127. T. Mizuno et al., "High Dielectric LDD Spacer Technology for High Performance MOSFET Using Gate-Fringing Field Effects," *Tech. Dig. IEDM,* 1989, p. 613.

128. T. Mizuno et al., "Gate-Fringing Field Effects on High Performance in High Dielectric LDD Spacer MOSFET s," *IEEE Trans. Electron Dev.,* April 1992, p. 982.

129. J.C. Guo et al., "Performance and Reliability Evaluation of High-Dielectric LDD Spacer on Deep Sub-Micrometer LDD MOSFET," *IEEE Trans. Electron Dev.,* **ED-41**, July 1994, p. 1239.

130. T. Buti et al., "Asymmetrical Halo Source- GOLD Drain (HS-GOLD) Deep Sub-half Micron *n*-MOSFET Design for Reliability and Performance," *Tech. Dig. IEDM,* 1989, p. 617.

131. T. Buti et al., "A New Asymmetrical Halo Source- GOLD Drain (HS-GOLD) Deep Sub-half Micron *n*-MOSFET Design for Reliability and Performance,"*IEEE Trans. Electron Dev.*, **ED-38**, August 1991., p. 1757.

132. K.K. Ng and G.W. Taylor, "Effects of Hot-Carrier Trapping in *n*- and *p*-Channel MOSFETs," *IEEE Trans. Electron Dev.*, **ED-30**, (1983), p. 871.

133. T. Tsuuchiya and J. Frey, "Relationship Between Hot-Electrons/Holes and Degradation for *p*- and *n*- Channel MOSFETs," *IEEE Trans. Electron Dev.*, **EDL-6**, p. 8, 1986.

134. Y. Hiruta et al., "Impact of Hot-Electron Trapping of Half-Micron PMOSFETs with p^+ Poly -Si Gate," *Tech. Dig. IEDM,* 1986, p. 718.

135. M. Koyanagi et al., "Investigation and Reduction of Hot-Electron Induced Punchthrough (HEIP) Effect in Submicron PMOSFETs," *Tech. Dig. IEDM,* 1986, p. 722.

136. M. Koyonagi et al., "Hot-Electron Induced Punchthrough (HEIP) Effect in Submicron PMOSFETs," *IEEE Trans. Electron Dev.*, **ED-34**, 1987, p. 839.

137. B.S. Doyle and K.R. Mistry, "Anomalous Hot-Carrier Behavior for LDD *p*-Channel Transistors," *IEEE Electorn Dev. LEtts.,* **EDL-14**, November 1993, p. 536.

138. S. Odanaka and A. Hiroki, "Gate-Oxide Thickness Dependence of Hot-Carrier-Induced Degradation in Buried *p*-MOSFETs," *Tech. Dig. IEDM,* 1990, p. 565.

139. R. Woltjer and G.M. Paulzen ,"Oxide Charge Generation During Hot-Carrier Degradation in PMOSTs," *Tech. Dig. IEDM*, 1993, p. 713.

140. T. Mogami et al., "Hot-Carrier Effects in Surface-Channel PMOSFETs with BF_2^+ or Boron-implanted Gates," *Tech. Dig. IEDM*, 1991, p. 533.

141. A.B. Joshi et al., "Oxynitride Gate Dielectrics for p^+-Polysilicon Gate MOS Devices," *IEEE Electron Dev. Letts.*, **EDL-14**, December 1993, p. 560.

142. T. Eguchi et al., "New Dual Gate Doping Process using In-Situ Boron Doped-Si for Deep Submicron CMOS Device," *Tech. Dig. IEDM*, 1993, p. 831.

143. K. Uwasawa et al., "Scaling Limitations of Gate Oxide in p^+ Polysilicon Gate MOS Structures for Sub-Quarter Micron CMOS Devices," *Tech. Dig. IEDM* 1993, p. 895.

144. H.-H. Tseng, et al., "The Effect of Silicon Gate Microstructure and Gate Oxide Process on Threshold Voltage Instabilities p^+ -Gate *p*-Channel MOSFETs with Fluorine Incorporation," *IEEE Trans. Electron Dev.*, **ED-39**, July 1992, p. 1687.

145. T. Noguchi et al., "Parasitic Resistance Characteristics for Optimum Design of Half Micron MOSFETs," *Tech. Dig. IEDM,* 1986, p. 730.

146. J.M. Pimbley, M. Ghezzo, H.G. Parks, and D.M. Brown, *Advanced CMOS Process Technology,* Vol. 19, VLSI Electronics Microstructure Science, Ch. 5, "Relibility," Academic Press, San Diego, CA, 1989.

147. *ibid.* Ref. 17, this chapter, p. 130.

148. F.H. Hsu and K.Y. Chiu, "A Comparative Study of Tunneling, Substrate Hot-Electron, and Channel Hot-Electron Injection Induced Degradation in Thin-Gate MOSFETs," *Tech. Dig. IEDM*, 1984, p. 96.

149. E. Takeda, A. Shimizu, and T. Hagiwara, "Role of hot-hole injection in hot-carrier effects and the small degraded channel region in MOSFETs." *IEEE Electron Dev. Letts.*, **EDL-4,** p. 329, (1984).

150. T.K. Horiuchi, H. Mikoshiba, K. Nakamura, and F. Hamano, "A Simple Method to Evaluate Device Lifetime Due to Hot-Carrier Effect Under Dynamic Stress," *IEEE Trans. Electron Dev. Letts.*, **EDL-7**, p. 377, (1986).

151. P.M. Lee et al., "Circuit Aging Simulator (CAS)," *Tech. Dig. IEDM*, 1988, p. 134.

152. G.J. Dunn and S.A. Scott, "Channel Hot-Carrier Stressing of Reoxidized Nitrided Silicon Dioxide," *IEEE Trans. Electron Dev.*, **ED-37**, July 1990, p. 1719.

153. R. Haruta et al., "Improvement of Hardness of MOS Capacitors to Electron-Beam Irradiation and Hot-Electron Injection by Ultra-Dry Oxidation of Silicon," *IEEE Electron Dev. Letts.*, **EDL-10,** January 1989, p. 27.

154. F. Hsu and K. Chiu, "Effects of Device Processing on Hot-Electron Induced Device Degradation," in *VLSI Symp. Tech. Dig.*, 1985, p. 108.

155. J. Mitsuhashi, S. Nakao, and T. Matsukawa, "Mechanical Stress and Hydrogen Effects on Hot Carrier Injection," *Tech. Dig. IEDM,* 1986, p. 387.

156. Y. Nissan-Cohen and T. Gorczyca, "The Effect of Hydrogen on Trap Generation, Positive Charge Trapping, and Time-Dependent Dielectric Breakdown of Gate Oxides," *IEEE Electron Dev. Letts..*, **EDL-9,** June 1988, p. 288.

157. Y. Ohji et al., "Effects of Minute Impurities (H, OH, F) on SiO_2/Si Interface as Investigated by Nuclear Resonant Reaction and Electron Spin Resonance," *IEEE Trans. Electron Dev.*, **ED-37** August 1990, p. 1635.

158. R.B. Fair and R.C. Sun, "Threshold voltage instability in MOSFETs due to channel hot-hole emission," *IEEE Trans. Electron Dev.*, **ED-28** p. 83, January 1981.

159. R. Haruta et al., "Improvement of Hardness of MOS Capacitors to Electron-Beam Irradiation and Hot-Electron Injection by Ultra-Dry Oxidation of Silicon," *IEEE Electron Dev. Letts.*, **EDL-10,** January 1989, p. 27.

160. M.T. Takagi, I. Yoshii, and K. Hashimoto, "Characterization of Hot-Carrier-Induced Degradation of MOSFETs Enhanced by H_2O Diffusion for Multilevel Interconnection Processing," *Tech. Dig. IEDM*, 1992, p. 703.

161. N. Lifshitz and G. Smolinsky, "Hot-Carrier Aging of the MOS Transistor in the Presence of Spin-on Glass as at the Interlevel Dielectric," *IEEE Electron. Dev. Letts..*, **EDL-12**, March 1991, p. 140.

162. M.L. Chen et al., "Hot-carrier aging in two-level metal processing," *Tech. Dig. IEDM*, p. 55, 1987.

163. Y. Hayafuji and K. Kajiwara, "Nitridation of Silicon and Oxidized Silicon," *J. Electrochem. Soc.* , **129**, (1982).

164. Y. Hayafuji, K. Kajiwara, and S. Usui, "Shrinkage and Growth of Oxidation Stacking Faults During Thermal Nitridation of Si and Oxidized Silicon," *J. Appl. Phys.*, **63,** 8639 (1982).

165. J.A. Nemetz and R.F. Tressler, "Thermal Nitridation of Silicon and Silicon Dioxide for Thin Gate Insulators," *Solid State Technology*, p. 209, September 1983.

166. T. Ito, T. Nozaki, and H. Ishikawa, "Direct thermal nitridation of silicon oxide films in anhydrous ammonia gas," *J. Electrochem. Soc.* , **127**, September 1980, p. 2053.

167. M.L. Naiman et al., "Properties of thin oxynitride gate dielectrics produced by thermal nitridation of silicon dioxide," *Tech. Dig. IEDM,* 1980, p. 562.

168. C.-T. Chen et al., "Study of electrical characteristics on thermally nitrided SiO_2 films," *J. Electrochem. Soc.* , p. 875, 1984.

169. S.-T. Chang, N.M. Johnson, and S.A. Lyon, "Capture and tunnel emission of electrons by deep levels in ultrathin nitrided oxides on silicon," *J. Appl. Phys. Lett.*, **44**, March 1984, p. 316.

170. S.K. Lai, J. Lee, and V.K. Dham, "Electrical properties of nitrided oxide systems for use in gate dielectrics and EEPROM," *Tech. Dig. IEDM*, p. 180, 1983.

171. S.S. Wong et al., "Composition and electrical properties of nitrided-oxide and re-oxidized nitrided oxide," in *Proc. Symp. Silicon Nitride Thin Insulating Films,* **ECS vol. 83-8**, p. 346, 1983.

172. F.C. Hsu and K.-Y. Chiu, "A comparative study of tunneling, substrate hot-electron and channel hot-electron injection induced degradation in thin-gate MOSFETs," *Tech. Dig. IEDM,* p. 96, 1984.

173. T. Hori, H. Iwasaki, and K. Tsuji, "Charge-Trapping Properties of Ultrathin Nitrided Oxides Prepared by Rapid Thermal Annealing," *IEEE Trans. Electron Dev.* , **ED-36**, July 1988, p. 904.

174. T. Hori, H. Iwasaki, and K. Tsuji, "Electrical and Physical Properties of Ultrathin Reoxidized Nitrided Oxides Prepared by Rapid Thermal Processing," *IEEE Trans. Electron Dev.* , **ED-36**, February 1989, p. 340.

175. W. Yang, R. Jayaraman, & C.G. Sodini, "Optimization of low-pressure nitridation/re-oxidation of SiO_2 for scaled MOS devices," *IEEE Trans. Elect. Dev.*, **ED-35**, p. 939, 1988.

176. B. J. Gross, K.S. Krisch, and C.G. Sodini, "An Optimized 850°C Low-Pressure-Furnace Reoxidized Nitrided Oxide (ROXNOX) Process," *IEEE Trans. Elect. Dev.*, **ED-38**, September 1991, p. 2036.

177. H.S. Momose et al., "Hot-Carrier Related Phenomena for N- and PMOSFETs with Nitrided Gate Oxide by RTP." *Tech. Dig. IEDM,* 1989, p. 267.

178. A. Wu et al., "Gate bias polarity dependence of charge trapping and time dependent dielectric breakdown in nitrided and reoxidized nitrided oxides," *IEEE Electron Dev. Letts.*, **EDL-10**, p. 443, October 1989.

179. E. Takeda et al., "VLSI Reliability Challenges: From Device Physics to Wafer Scale Systems," *Proceedings of the IEEE*, May 1993, p. 653.

180. W. Yang, R. Jayaraman, and C.G. Sodini, "Optimization of Low-Pressure Nitridation/Reoxidation of SiO_2 for Scaled MOS Devices," *IEEE Trans. Electron Dev.,* **ED-35**, July 1988, p. 935.

181. M.A. Schmidt et al., "Inversion layer mobility of MOSFETs with nitrided oxide gate dielectrics," *IEEE Trans. Electron Dev.,* **ED-35**, 1988, p. 1627.

182. T. Kaga and T. Hagiwara, "Short- and Long-Term Reliability of Nitrided Oxide MISFETs," *IEEE Trans. Electron Dev.,* **ED-35**, 1988, p. 929.

183. P. J. Wright, A. Kermani, and K.C. Saraswat, "Nitridation and Post-Nitridation Anneals of SiO_2 for Ultrathin Dielectrics," *IEEE Trans. Electron Dev.*, **ED-37**, August 1990, p. 1836.

184. H. Fang, et al., "Low-Temperature Furnace-Grown Reoxidized Nitrided Oxide Gate Dielectrics as a Barrier to Boron Penetration," *IEEE Electron Dev. Letts.*, **EDL-13,** April 1992, p. 217.

185. H. Fang et al., "Ultrathin Furnace Reoxidized Nitrided Oxide Gate Dielectrics for Extreme Submicrometer CMOS Technology," *Tech. Dig. IEDM*, 1992, p. 623.

186. J.R. Pfiester et al., "The effects of boron penetration on p^+ polysilicon gated PMOS devices," *IEEE Trans. Electron Dev.*, **ED-37**, p. 1842, 1990.

187. J.S. Cable, R.A. Mann and J.C.S. Woo, "Impurity barrier properties of reoxidized nitrided oxide films for use with p^+ polysilicon gates," *IEEE Electron Dev. Letts.*, **EDL-12**, p. 128, 1991.

188. C. Wong et al., "Doping of n^+ and p^+ Polysilicon gates in dual-gate CMOS process," *IEDM Tech. Dig.* , 1988, p. 238.

189. F.C. Hsu, J. Hui, and K.Y. Chiu, "Effect of Final Annealing on Hot-Electron-Induced MOSFET Degradation," *IEEE Electron Dev. Letts.*, **EDL-6**, July 1985, p. 369.

190. H. Hwang et al., "Electrical and Reliability Characteristics of Ultrathin Oxynitride Gate Dielectric Prepared by Rapid Thermal Processing in N_20, *Tech. Dig. IEDM*, 1990, p. 421.

191. A. Uchiyama et al., "High Performance Dual-gate Sub-halfmicron CMOSFETs with 6 nm-thick Nitrided SiO_2 Films in an N_2O Ambient," *Tech. Dig. IEDM*, 1990, p. 425.

192. H. Fukuda et al., "Novel N_2O - oxynitridation technology for forming highly reliable EEPROM tunnel oxide films," *IEEE Electron Device Letters*, **EDL-12** (1991) p. 587.

193. H. Hwang et al., "Improved reliability characteristics of submicrometer MOSFETs with oxynitride gate dielectric prepared by rapid thermal oxidation," *IEEE Electron Device Letters*, **EDL-12**, (1991), p. 495.

194. T.Y. Chu et al., "Thickness and Compositional Nonuniformities of Ultrathin Oxides Grown by Rapid Thermal Oxidation of Silicon in N_2O, *J. Electrochemical Soc.*, **138**, June 1991, p. 113.

195. W. Ting et al., "MOS Characteristics of Ultrathin SiO_2 Prepared by Oxidizing Si in N_2O," *IEEE Electron Device Letters*, **EDL-12** (1991), p. 416.

196. H. Fukuda, T. Arakawa, and S. Ohno, "Highly reliable thin nitrided SiO_2 films formed by rapid thermal processing in an N_2O Ambient," *Jap. Jnl. of Appl. Phys.*, **vol. 29**, p. 2333, 1990.

197. G.Q. Lo et al., "Improved Performance and Reliability of MOSFETs with Ultrathin Gate Oxides Prepared by Conventional Furnace Oxidation of Si in Pure N_2O Ambient," *Symp. VLSI Tech.* p. 43, 1991.

198. J. Ahn W. Ting, and D.-K. Kwong, "Furnace nitridation of thermal SiO_2 in pure N_2O ambient for ULSI Applications," *IEEE Electron Device Letters*, **EDL-13**, 1992, p. 117.

199. H. Hwang et al., "Electrical Characteristics of Ultrathin Oxynitride Gate Dielectric Prepared by Rapid Thermal Oxidation of Si in N_2O, *Appl. Phys. Lett.*, **57**, 1990, p. 1010.

200. Z. Liu et al., "The Effects of Furnace Annealing on MOSFETs," *Tech. Dig. IEDM*, 1992, p. 625.

201. S.C. Sun, T.S. Chen, and H.Y. Chang, "Thermal Oxidation of Silicon in Pure N_2O," *Ext. Abs. Mtg. Electrochem. Soc.*, Spring 1993, Abs. #749, p. 1080.

202. N.S. Saks et al., "Characteristics of Oxynitrides Grown in N_2O," *Ext. Abs. Mtg. Electrochem. Soc.*, Spring 1994, Abs. #162, p. 251.

203. Y. Okada et al., "Evaluation of Interfacial Nitrogen Concentration of RTP Oxynitrides by Reoxidation," *J. Electrochem. Soc.*, **140**, p. L87, June 1993.

204. H.S. Momose et al., "Very Lightly Nitrided Oxide MOSFETs for Deep-Sub-Micron CMOS Devices," *Tech. Dig. IEDM*, 1991, p. 359.

205. A. Joshi et al., "High-Field Breakdown in Thin Oxides Grown in N_2O Ambient," *IEEE Trans. Electron Dev.* , **ED-40**, p. 1437, August 1993.

206. Y. Okada et al., "The relationship between growth conditions, nitrogen profile, and charge-to-breakdown of gate oxynitrides grown from pure N_2O," *Appl. Phys. Letts.*, **63**, p. 194, 1993.

207. T. Hori, S. Akamatsu, and Y. Odake, "Deep-Submicrometer CMOS Technology with reoxidized or annealed nitrided-gate dielectrics prepared by rapid thermal processing," *IEEE Trans. on Electron Devices*, **ED-39**, January 1992, p. 118.

208. T. Hori, T. Yasui, and S. Akamatsu, "Improved hot-carrier immunity in submicrometer MOSFETs with reoxidized nitrided-gate dielectrics prepared by rapid thermal processing," *IEEE Trans. on Electron Devices*, **ED-39**, January 1992, p. 134.

209. H.G. Pomp et al., "Lightly N_2O nitrided dielectrics grown in a conventional furnace for E^2PROM and 0.25 μ m CMOS, *Tech. Dig. IEDM*, 1993, p. 463.

210. M.S. Soo et al., "Very Light Nitridation of Thin Gate Oxide in Low Pressure N_2O Ambients," *Ext. Abs. Mtg. Electrochem. Soc.*, Spring 1994, Abs. #168, p. 262.

211. Y. Okada et al., "The Performance and Reliability of 0.4 Micron MOSFETs with Gate Oxynitrides Grown by Rapid Thermal Processing Using Mixtures of N_2O and O_2," *IEEE Trans. Electron Dev.* , **ED-41,** p. 191, February 1994.

212. S. Haddad and M.S. Liang, "Improvement of Thin Gate Oxide Integrity Using Through-Silicon-Gate Ion Implantation," *IEEE Electron Dev. Letts.*, **EDL-8**, p. 58, 1987.

213. J. Ahn et al., "High Quality Thin Gate Oxide Prepared by Annealing Low-Pressure CVD SiO_2 in N_2O," *Appl. Phys. Letts.*, **59**, p. 283, 1991.

214. I. Rahat, J. Shappir, and D. Ben-Atar, "Novel Interface Nitridation Process for Thin Gate Oxides," *IEEE Trans. Electron Dev.*, **ED-40**, November 1993, p. 2047.

215. L. Manchanda et al., "A Boron-Retarding and High Interface Quality Thin Gate Dielectric for Deep-Submicron Devices," *Tech. Dig. IEDM*, 1993, p. 459.

216. B. Doyle and A. Philipossian, "*p*-Channel Hot-Carrier Optimization of RNO Gate Dielectrics Through the Reoxidation Step," *IEEE Electron Dev. Letts.*, **EDL-14**, April 1993 p. 161.

217. G.Q. Lo, J. Ahn, and D.L. Kwong, "Improved Hot-Carrier Immunity in CMOS Analog Device with N_2O-Nitrided Gate Oxides," *IEEE Trans. Electron Dev.*, **EDL-13**, September 1992, p. 457.

218. G.Q. Lo et al., "P-Channel MOSFETs with ultrathin N_2O gate Oxides," *IEEE Electron Device Letters*, **EDL-13** (1992), p. 111.

219. Z. Liu et al., "Improvement of Charge Trapping Characteristics of N_2O-Annealed and Reoxidized N_2O-Annealed Thin Oxides," *IEEE Electron Dev. Letts.*, **EDL-13**, October 1992, p. 519.

220. Z. Liu et al., "Effects of N_2O Anneal and Reoxidation on Thermal Oxide Characteristics," *IEEE Electron Dev. Letts.*, **EDL-13**, August 1992, p. 402.

221. C.G. Sodini and K.S. Krisch, "Silicon Oxynitride Gate Dielectrics for Scaled CMOS," *Tech. Dig. IEDM,*, 1992, p. 617.

222. S. Kusonoki et al., "Hot-Carrier-Resistant Structure by Reoxidized Nitrided Oxide Sidewall for Highly Reliable and High-Performance LDD MOSFETs," *Tech. Dig. IEDM*, 1991, p. 649.

223. Y. Okada et al., "Evaluation of Interfacial Nitrogen Concentration of RTP Oxynitrides by Reoxidation," *J. Electrochem. Soc.* , **140**, June 1993, p. L87.

224. Y. Okada and P.J. Tobin, "Hot-Carrier Degradation of LDD MOSFETs with Gate Oxynitride Grown in N2O," *IEEE Electron Dev. Letts.*, **EDL-15**, July 1994, p. 233.

225. E. F. Da Silva, *Tech. Dig. IEDM*, 1987, p. 848.

226. P. Wright et al., "Hot-Electron Immunity of SiO_2 Dielectrics with Fluorine Incorporation," *IEEE Electron Device Letters*, **EDL-10**, August 1989, p. 347.

227. V. Jain, D. Pramanik, K.Y. Chang, and C. Hu., "Improved Sub-micron CMOS Device Performance in CVD Tungsten Silicide," *1991 Symp. VLSI Tech.*, p. 91.

228. Y. Nishioka et al., "Hot-Electron Hardened Si-Gate MOSFET Utilizing F Implantation," *IEEE Electron Device Letters*, **EDL-10**, April 1989, p. 141.

229. N. Kasai, P. Wright, and K.C. Saraswat, "Hot-Carrier Degradation Characteristics for Fluorine Incorporated NMOSFETs," *IEEE Trans. Electron Dev.*, **ED-37**, June 1990 p. 1426.

230. I.C. Chen et al., "Performance and Reliability Enhancement for CVD Tungsten Polycided CMOS Transistors Due to Fluorine Incorporated in the Gate Oxide," *IEEE Electron Dev. Letts.*, **EDL-15**, September 1994.

PROBLEMS

1. Explain the difference between CORS and CAS.

2. Explain the difference between electron traps in SiO_2 and interface state traps.

3. Qualitatively describe the advantages and disadvantages of LDD MOSFETs.

4. Calculate $\mathcal{E}_{ymax}$ in a non-LDD MOSFET having a 1 μm channel length, t_{ox} = 200 nm, r_j = 0.2 μm, V_T = 0.7V, V_{GS} = 5V, and V_{DS} = 5V. (b) Use Eq. 9-4 to calculate I_{sub} in this device if I_D = 100 μA.

5. Calculate $\mathcal{E}_{ymax}$ in an LDD MOSFET having a 1 μm channel length, t_{ox} = 200 nm, r_j = 0.2 μm, V_T = 0.7V, V_{GS} = 5V, and V_{DS} = 5V using Eq. 9-13. Assume the L_{n^-} region in this device is 0.2 μm in length.

6. (a) An *n*-channel MOSFET has a tox of 20 nm, an L_{eff} = 1 μm, r_j = 0.2 μm, and V_T = 0.6V. If the device is biased at V_{GS} = 3V and V_{DS} = 5V, calculate the saturation voltage V_{DSsat} and the maximum electric field. If this device is built using an LDD and the n^- length of the LDD is 0.2 μm, find the decrease in the maximum electric field in the LDD device. If the FET channel width is 10 μm and the S/D sheet resistance in the LDD region is 100 Ω/square, calculate the additional S/D resistance due to the LDD insertion and compare it with the intrinsic FET resistance.

7. List some advantages of using disposable spacers for fabricating LDD MOSFETs. What are some of their drawbacks?

8. Discuss the advantages and disadvantages of the following device structures compared to conventional LDD MOSFETs:

a) Fully overlapped drain structures
b) Buried LDD structures
c) Large-angle-tilt implanted drain structures
d) LDD MOSFETs with high-dielectric spacers
e) Asymmetrical MOSFET structures (Halo-Source GOLD)

9. Why have LDD structures not been necessary in PMOS devices until deep submicron gate lengths are encountered?

10. Explain briefly why N_2O nitrided oxides have proved to be superior to NH_3-nitrided gate dielectrics.

APPENDIX A

MATRIX METHODS

A.1 INTRODUCTION TO MATRICES

Matrices and their algebra are vital to modern engineering mathematics. They represent an enormously effective means by which data and other mathematical quantities of interest can be manipulated in an elegant and efficient manner. The practical use of finite difference and finite elements are greatly facilitated by the use of matrices. Hence, we present here a few of the important basic concepts of matrices and matrix algebra.

A *matrix* is a rectangular array of elements arranged in rows and columns. The elements may be numbers, variables, functions, derivatives, or even other matrices. The following set of m linear algebraic equations serves to illustrate a common notational convention used relative to matrix representation of a linear system of equations.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ \ldots\ + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ \ldots\ + a_{2n}x_n &= b_2 \\ \vdots \qquad \vdots \qquad\qquad \vdots \qquad &\vdots \\ \vdots \qquad \vdots \qquad\qquad \vdots \qquad &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ \ldots\ + a_{mn}x_n &= b_m \end{aligned} \tag{A - 1}$$

The set of variables $x_1, x_2, \ldots, x_n$, and a set of m dependent variables $b_1, b_2, \ldots b_m$ are related through the mxn coefficients $a_{11}, a_{12}, \ldots, a_{mn}$. Using matrix notation we write Eq. A-1 more compactly as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{Bmatrix} \tag{A - 2}$$

The mxn array of a's is referred to as the coefficient matrix $[\mathbf{A}]$ or $[\mathbf{a}_{i,j}]$ and the independent variables as the $\{\mathbf{x}\}$ or $\{\mathbf{x}_j\}$ vector. In addition, the dependent quantities on the right-hand side are grouped in a separate array called the $\{\mathbf{B}\}$ or $\{\mathbf{b}_{i,j}\}$ vector.

A.1.1 Vectors and Scalars. Matrices with only one row and n columns ($1xn$) or one column and m rows ($mx1$) are called vectors. The symbol { } is used to indicate a row or column vector. Matrices with only one element (1x1) are called scalars. For example, the 3x1 matrix $\{\mathbf{A}\}$

$$\{\mathbf{A}\} = \begin{Bmatrix} 2 \\ 10 \\ -3 \end{Bmatrix}_{3\times 1} = \{a_{i,j}\}, \quad i = 1,2,3 \;\; j = 1 \qquad \text{(A - 3)}$$

is a *column vector.*

A.1.2 The Square Matrix. A matrix having an equal number of columns and rows is called a *square matrix.* Square matrices play an important role in engineering and have properties that are unique. For example, the determinant and inverse can only be defined if the matrix is a square matrix.

A.1.3 The Identity Matrix. The identity matrix is a square matrix whose off-diagonal elements are equal to zero and diagonal elements are equal to one. Since identity matrices are unique, they are designated by the letter I and sometimes written as

$$[\mathbf{I}]_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \text{(A - 4)}$$

A.1.4 The Diagonal Matrix. Diagonal matrices are square matrices whose off-diagonal elements are equal to zero and whose diagonal elements can take on any value. These matrices are very important and occur frequently when one solves systems of ordinary differential equations. Furthermore, "diagonalizing" is used to mean that a nondiagonal square matrix is reduced to one which is diagonal.

A.1.5 The Augmented Matrix. This is a matrix that can be formed from two other matrices. Consider the matrices [**A**] and [**B**]:

$$[\mathbf{A}] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad [\mathbf{B}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Then a third matrix [**C**] is the augmented matrix if

$$[\mathbf{C}] = \left[\begin{array}{cc:cc} a_{11} & a_{12} & 1 & 0 \\ a_{21} & a_{22} & 0 & 1 \end{array}\right] \qquad \text{(A - 5)}$$

Note that [**C**] can only be defined if both [**A**] and [**B**] have exactly the same number of rows. The dashed line in [**C**] is used to separate the elements of [**A**] and [**B**]. Augmented matrices are useful in solving sets of linear algebraic equations.

A.1.6 The Upper-Triangular Matrix. This is a square matrix often denoted as [**U**] in which all elements below the diagonal are zero and the rest of the elements can take on any value. That is:

$$[\mathbf{U}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{mn} \end{bmatrix} \qquad \text{(A - 6)}$$

The solution of linear algebraic equations using the Gaussian elimination method results in an upper-triangular matrix.

A.1.7 The Lower-Triangular Matrix. This is a square matrix often denoted as [**L**] in which all elements above the diagonal are zero and the rest of the elements can take on any value. That is:

$$[\mathbf{L}] = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad \text{(A - 7)}$$

A.2 MATRIX MULTIPLICATION

If we wish to obtain the product of two matrices [**A**] and [**B**], which is equal to a third matrix [**C**]:

$$[\mathbf{C}] = [\mathbf{A}]\,[\mathbf{B}] \qquad \text{(A - 8)}$$

matrix [**C**] can only be defined if the number of rows of matrix [**B**] is equal to the number of columns of matrix [**A**]. That is,

$$[\mathbf{C}]_{mxL} = [\mathbf{A}]_{mxn}\,[\mathbf{B}]_{nxL} \qquad \text{(A - 9)}$$

Here m is the number of rows of matrices [**A**] and [**C**]; n is the number of columns of [**A**] and the number of rows of [**B**]; L is the number of columns of [**B**] and [**C**].

To illustrate the procedure for finding the products of the matrices consider the following example:

$$[\mathbf{A}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \qquad [\mathbf{B}] = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

Let us define

$$[\mathbf{C}] = [\mathbf{A}]_{2x3}[\mathbf{B}]_{3x2} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \qquad \text{(A - 10a)}$$

where

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ c_{12} &= a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ c_{22} &= a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{aligned} \qquad \text{(A - 10b)}$$

We see that the process involved in multiplying two matrices causes us to use an *entire row* of [**A**] and an *entire column* of [**B**] to obtain a *single element* of [**C**]. That is, "c_{11}," is determined by combining the products of the elements in the first column of [**B**] with the elements of the first row in [**A**], "c_{12}" is determined by combining the products of the elements in the second column of [**B**] with the elements of the first row in [**A**]; ... etc.

Example A-1: Find the product [**A**] [**B**] for the following matrices.

$$[\mathbf{A}] = \begin{bmatrix} 2 & 1 \\ 3 & -4 \end{bmatrix} \qquad [\mathbf{B}] = \begin{bmatrix} a & c & 0 \\ b & d & 0 \end{bmatrix}$$

Solution:

$$[\mathbf{C}]_{2x3} = \begin{bmatrix} 2 & 1 \\ 3 & -4 \end{bmatrix} \begin{bmatrix} a & c & 0 \\ b & d & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2a + b & 2c + d & 0 \\ 3a - 4b & 3c - 4d & 0 \end{bmatrix}$$

A.3 MATRIX INVERSION

There are instances where the solution of matrices involves the calculation of the inverse matrix. The inverse of a matrix [**A**] is a property of square nonsingular matrices (determinant is not zero) and is denoted [$\mathbf{A}^{-1}$]. In addition,

$$[\mathbf{A}][\mathbf{A}^{-1}] = [\mathbf{I}] \qquad \text{(A - 11)}$$

Note that only square matrices can have an inverse.

Several methods for inverting matrices are available, including: the *Cramer's rule* method; the *Gauss-Jordan method;* the *reducing matrix method*; and the *partitioning method.* Here we show the Gauss-Jordan method - an extension of the Gauss-Jordan method for solving matrix equations.

Suppose we have two 3x3 square matrices **A** and **B** such that $[\mathbf{A}]^{-1} = [\mathbf{B}]$, then by definition [**A**] [**B**] = [**B**] [**A**] = [**I**], or more explicitly,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11}b_{12}b_{13} \\ b_{21}b_{22}b_{23} \\ b_{31}b_{32}b_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then if we form an augmented matrix [[**A**] : [**I**]]

$$[[\mathbf{A}] \vdots [\mathbf{I}]] = \left[\begin{array}{ccc:ccc} a_{11} & a_{12} & a_{13} & 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 1 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 1 \end{array}\right]$$

we can use the Gauss-Jordan method to reduce this augmented matrix to the form:

$$[[\mathbf{I}] \vdots [\mathbf{B}]] = \left[\begin{array}{ccc:c} 1 & 0 & 0 & b_{11}b_{12}b_{13} \\ 0 & 1 & 0 & b_{21}b_{22}b_{23} \\ 0 & 0 & 1 & b_{31}b_{32}b_{33} \end{array}\right]$$

where [**B**] is the inverse of [**A**]. This method of determining the inverse is summarized in one step for an *nxn* matrix:

$$[[\mathbf{A}] : [\mathbf{I}]] \text{ ---> } [[\mathbf{I}] : [\mathbf{B}]] \qquad \text{(A - 12)}$$

Example A-2: Use the Gauss-Jordan elimination method to determine the inverse of [**A**]:

$$[\mathbf{A}] = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$$

Solution: First, multiply the first row by -2 and add it to the second row

$$\left[\begin{array}{cc:cc} 1 & 2 & 1 & 0 \\ 2 & -1 & 0 & 1 \end{array}\right] \quad \text{---->} \quad \left[\begin{array}{cc:cc} 1 & 2 & 1 & 0 \\ 0 & -5 & -2 & 1 \end{array}\right] \; -2R_1 + R_2$$

Next, divide row two by -5

$$\left[\begin{array}{cc:cc} 1 & 2 & 1 & 0 \\ 0 & 1 & 2/5 & -1/5 \end{array}\right] - R_2/5 \quad \text{---->} \quad \left[\begin{array}{cc:cc} 1 & 0 & 1/5 & 2/5 \\ 0 & 1 & 2/5 & -1/5 \end{array}\right] -2R_2 + R_1$$

Finally, multiply row two by -2 and add it to the first row to get the final augmented matrix. The inverse of [**A**] is simply equal to the right-hand portion of this augmented matrix, namely,

$$[A]^{-1} = \frac{1}{5}\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$$

A.4 MATRIX METHODS FOR SOLVING SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS

As noted in the text, the solution of simultaneous linear algebraic equations is probably one of the most important procedures in modern engineering computations. For example, the finite element method and the finite difference method both involve the solution of systems of linear algebraic equations. Two classes of methods are available for solving a set of linear algebraic equations: *direct methods*, and *iterative methods*. The direct methods are based on elimination, and include Cramer's Rule, Gauss elimination method, the Gauss-Jordan elimination method, and Crout's Method. The iterative methods include Jacobi's method (and its variations), the Gauss-Seidel method (and its variations), the successive-overrelaxation (SOR) method, the alternating-direction implicit (ADI), and Stone's strongly implicit method. An good (but compressed) survey by Duff on the subject of the various matrix methods for solving a set of linear algebraic equations is listed in reference 1. A text dealing with this topic is A. George and J.W.H. Liu, *Computer Solution of Large Sparse Positive Definite Systems,* Prentice-Hall, Inc. Englewood Cliffs, N.J. (1981).

A.4.1 Direct Methods

A.4.1.1 Cramer's Rule. This method is widely taught as a procedure for solving a set of simultaneous linear algebraic equations. However, since it is likely to be familiar to readers and because it less efficient than other techniques, we do not emphasize it here.

A.4.1.2 Gauss-Jordan Elimination Method. This method is basically an extension of the Gauss method presented in the next section. The difference is that there is no need for back-substitution using this method. This is accomplished quite simply by eliminating the coefficients above the diagonal in the coefficient matrix. Consider the typical case of three equations with three unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \text{(A - 13)}$$

The Gauss-Jordan process operates on the augmented matrix

$$\left[\begin{array}{ccc:c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array}\right] \qquad \text{(A - 14)}$$

By successive operations of dividing rows by numbers, interchanging rows, and subtracting a multiple of one row from another, we try to reduce this matrix to the form

$$\left[\begin{array}{ccc:c} 1 & 0 & 0 & b_1^* \\ 0 & 1 & 0 & b_2^* \\ 0 & 0 & 1 & b_3^* \end{array}\right] \qquad \text{(A - 15)}$$

in which case the solution is readily obtained as

$$\begin{aligned} x_1 &= b_1^* \\ x_2 &= b_2^* \\ x_3 &= b_3^* \end{aligned} \qquad \text{(A - 16)}$$

If the matrix of the coefficients [**A**] is nonsingular, the process described can be carried out in full. Although this method requires more computational effort than the Gauss method, we present it because it provides a direct method for solving the inverse problem discussed earlier.

Example A-3: Solve the following set using the Gauss-Jordan method:

$$\begin{aligned} 2x_1 - 4x_2 + 6x_3 &= 5 \\ x_1 + 3x_2 - 7x_3 &= 2 \\ 7x_1 + 5x_2 + 9x_3 &= 4 \end{aligned}$$

Solution: Forming the augmented matrix, we proceed by reducing the first column to that of an identity matrix. Thus

$$\left[\begin{array}{ccc:c} 2 & -4 & 6 & 5 \\ 1 & 3 & -7 & 2 \\ 7 & 5 & 9 & 4 \end{array}\right]$$

Next, divide row one by 2

$$\text{---->} \left[\begin{array}{ccc:c} 1 & -2 & 3 & 2.5 \\ 1 & 3 & -7 & 2 \\ 7 & 5 & 9 & 4 \end{array}\right] \; R_1/2$$

Then add the negative of row one to row two, and multiply row one by -7 and add it to row three:

$$\left[\begin{array}{cccc} 1 & -2 & 3 & 2.5 \\ 0 & 5 & -10 & -0.5 \\ 0 & 19 & -12 & -13.5 \end{array}\right] \begin{array}{l} \\ -R_1 + R_2 \\ -7R_1 + R_3 \end{array}$$

Now divide row two by 5:

$$\text{---->} \left[\begin{array}{cccc} 1 & -2 & 3 & 2.5 \\ 0 & 1 & -2 & -0.1 \\ 0 & -1 & 9 & -13.5 \end{array}\right] R_2/5$$

Then multiply row two by 2 and add it to row one, and also multiply row two by -19 and add it to row three:

$$\left[\begin{array}{cccc} 1 & 0 & -1 & 2.3 \\ 0 & 1 & -2 & -0.1 \\ 0 & 0 & 26 & -11.6 \end{array}\right] \begin{array}{l} 2R_2 + R_1 \\ \\ -19R_2 + R_3 \end{array}$$

Next, divide row three by 26:

$$\text{--->} \left[\begin{array}{cccc} 1 & 0 & -1 & 2.3 \\ 0 & 1 & -2 & -0.1 \\ 0 & 0 & 1 & -0.44 \end{array}\right] R_3/26$$

Finally, add row three to row one, and also multiply row three by 2 and add it to row two.

$$\left[\begin{array}{cccc} 1 & 0 & 0 & 1.85 \\ 0 & 1 & 0 & -0.99 \\ 0 & 0 & 1 & -0.45 \end{array}\right] \begin{array}{l} R_3 + R_1 \\ 2R_3 + R_2 \\ \\ \end{array}$$

Note that this form is exactly the same as is given in Eq. A-15. The solution to the set of equations is then given as

$$\begin{aligned} x_1 &= 1.85 \\ x_2 &= -0.99 \\ x_3 &= -0.45 \end{aligned}$$

Note also that roundoff errors may be introduced depending on the number of significant figures used in manipulating the matrices. This may become significant if a large set of equations is to be solved.

A.4.1.3 Gauss Elimination Method. An alternative to the Gauss-Jordan method is the simple *Gauss method.* This is one of the most popular and efficient methods for solving an *nxn* system of equations. Since less computations are needed for the Gauss elimination method than for the Gauss-Jordan method, the simple Gauss method is preferred. Here we use the same steps as in the Gauss-Jordan method, but the goal is to produce a matrix with zeros *below the diagonal.* Each unknown is then determined from this matrix by back-substitution. That is, we first seek to achieve a matrix of the form

$$\begin{bmatrix} 1 & b_{12} & b_{13} & \vdots & c_1^* \\ 0 & 1 & b_{23} & \vdots & c_2^* \\ 0 & 0 & 1 & \vdots & c_3^* \end{bmatrix} \qquad \text{(A - 17)}$$

The corresponding set of equations from this matrix is then

$$\begin{aligned} x_1 + b_{12}x_2 + b_{13}x_3 &= c_1 \\ x_2 + b_{23}x_3 &= c_2 \\ x_3 &= c_3. \end{aligned} \qquad \text{(A - 18)}$$

Now x_3 can be obtained from the last equation and this value can be substituted in the next-to-last equation. This allows us to solve for x_2, which is then substituted together with the known value of x_3 in the first equation, to obtain x_1.

Example A-4: Solve the following set of equations using the Gauss Elimination method:

$$\begin{aligned} 2x_2 + 4x_3 &= 6 \\ 4x_1 + x_2 - 3x_3 &= 1 \\ 3x_1 - 8x_2 + 2x_3 &= 2 \end{aligned}$$

Solution: Using the augmented matrix approach and switching the second and first equations (so that $a_{11} \neq 0$) gives

$$\begin{bmatrix} 4 & 1 & -3 & \vdots & 1 \\ 0 & 2 & 4 & \vdots & 6 \\ 3 & -8 & 2 & \vdots & 2 \end{bmatrix}$$

We begin by dividing the first row by 4: thus

$$\begin{bmatrix} 1 & 0.25 & -0.75 & \vdots & 0.25 \\ 0 & 2 & 4 & \vdots & 6 \\ 3 & -8 & 2 & \vdots & 2 \end{bmatrix} \; R_1/4$$

Multiplying the first row by -3 and adding it to the third row eliminates x_1 from all equations except the first. Hence

$$\left[\begin{array}{cccc} 1 & 0.25 & -0.75 & 0.25 \\ 0 & 2 & 4 & 6 \\ 0 & -8.75 & 4.25 & 1.25 \end{array}\right] \quad -3R_1 + R_3$$

The second row is now divided by 2 to give

$$\left[\begin{array}{cccc} 1 & 0.25 & -0.75 & 0.25 \\ 0 & 1 & 2 & 3 \\ 0 & -8.75 & 4.25 & 1.25 \end{array}\right] \quad R_2/2$$

Finally, multiplying the second row by 8.75 and adding it to the third row eliminates x_2 from the third equation. That is,

$$\left[\begin{array}{cccc} 1 & 0.25 & -0.75 & 0.25 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 21.75 & 27.5 \end{array}\right] \quad 8.75R_2 + R_3$$

Obviously, the original set of equations has now been transformed to an upper-triangular form. Hence, dividing row 3 by 21.75 and expressing the set in algebraic form yields

$$\begin{aligned} x_1 + 0.25x_2 - 0.75x_3 &= 0.25 \\ x_2 + 2x_3 &= 3 \\ x_3 &= 1.264 \end{aligned}$$

These equations are now solved by back-substitution to give

$$\begin{aligned} x_3 &= 1.264 \\ x_2 &= 3 - 2\,(1.264) = 0.472 \\ x_1 &= 0.25 - 0.25\,(0.472) + 0.75\,(1.264) = 1.080 \end{aligned}$$

Note that in this and the other direct solution method example there is no rounding off error and the answer is exact.

A.4.1.4 Cholesky's Method. The classical Gaussian elimination method described above is not suitable for the systems of equations that must be solved in microelectronics because the size of the coefficient matrix is generally both large and sparse (most of its coefficients are zero). In addition, in many cases a set of tridiagonal linear matrix equations result. Therefore, some modifications of the classical Gaussian elimination algorithm are introduced to account for the zero entries and to utilize the relative simplicity of the tridiagonal set. One common approach is to decompose the

augmented matrix into the product of a unit upper-triangular matrix and a lower-triangular matrix (Cholesky's method, also called Crout's method, or the LU-decomposition method). In Cholesky's method (also known as Crout's method) *triangular factorization* is applied to the coefficient matrix e.g., [**A**] of the equation

$$[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\} \qquad \text{(A - 19)}$$

which yields

$$[\mathbf{A}] = [\mathbf{L}]\,[\mathbf{L}^T] \qquad \text{(A - 20)}$$

where [**L**] is a *lower triangular matrix* and $[\mathbf{L}^T]$ is the transpose of the [**L**] matrix, and $[\mathbf{L}^T]$ is also an *upper triangular matrix* (see section A.1 of this appendix for definitions of these terms). Using (A-20) in (A-19) we get

$$[\mathbf{L}]\,[\mathbf{L}^T]\,\{\mathbf{x}\} = \{\mathbf{b}\} \qquad \text{(A - 21)}$$

and by substituting $\{\mathbf{y}\} = [\mathbf{L}^T]\,\{\mathbf{x}\}$, we can find {**x**} by solving the triangular systems

$$[\mathbf{L}]\{\mathbf{y}\} = \{\mathbf{b}\} \qquad \text{(A - 22)}$$

and

$$[\mathbf{L}^T]\,\{\mathbf{x}\} = \{\mathbf{y}\}. \qquad \text{(A - 23)}$$

Example A-5: Solve the following matrix equation for {**x**} using the Cholesky method.

$$\begin{bmatrix} 4 & 1 & 2 & 1/2 & 2 \\ 1 & 1/2 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 \\ 1/2 & 0 & 0 & 5/8 & 0 \\ 2 & 0 & 0 & 0 & 16 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{Bmatrix} = \begin{Bmatrix} 7 \\ 3 \\ 7 \\ -4 \\ -4 \end{Bmatrix}$$

Solution: The coefficient matrix of the above equation can be factored to yield (A-20) where [**L**] in (A-20) in this case is

$$[\mathbf{L}] = \begin{bmatrix} 2 & & & & \mathbf{0} \\ 0.50 & 0.50 & & & \\ 1 & -1 & 1 & & \\ 0.25 & -0.25 & -0.50 & 0.50 & \\ 1 & -1 & -2 & -3 & 1 \end{bmatrix}$$

Solving $[\mathbf{L}]\{\mathbf{y}\} = \{\mathbf{b}\}$, we obtain

$$\{y\} = \begin{Bmatrix} 3.5 \\ 2.5 \\ 6 \\ -2.5 \\ --0.5 \end{Bmatrix}$$

and then solving $[L^T]\{x\} = \{y\}$ yields

$$\{x\} = \begin{Bmatrix} 2 \\ 2 \\ 1 \\ -0.8 \\ --0.5 \end{Bmatrix}$$

A.4.2 Iterative Methods

Iteration is defined as a computational procedure in which the desired result is approached through a repeated cycle of operations, each of which more closely approximates the desired result. Iterative methods can be used to solve a set of simultaneous linear algebraic equations. Such methods yield an approximation to the solution of a system of such equations, but unlike the direct methods they may not always yield a solution, even if the determinant of the coefficient matrix is not zero. Consequently, for these techniques to work, certain other conditions must be considered:

1. The set of equations must posses a strong diagonal. This is a necessary but not a sufficient condition for a solution to be found.

2. A sufficient condition for a solution to be found is that the absolute value of the diagonal coefficient in any equation must be greater than the sum of the absolute values of all other coefficients appearing in that equation.

These two conditions can often be satisfied for a variety of practical problems encountered in engineering.

Iterative techniques are often used when dealing with sparse matrices. They offer the advantage that less computer memory is required for solving a set of algebraic equations than would be required using direct techniques (such as Gaussian elimination). Another important advantage of iterative methods is that round-off errors in the solution are limited to those incurred in the final iteration, and are normally negligible. This is because the values used in the previous iteration are treated as initial estimates and updated independently of the rounding error. On the other hand, direct solution methods are subject to cumulative round-off errors which can be come significant for large coefficient matrices or ill-conditioned systems of equations. The main disadvantage of iterative methods is that they depend on the numerical conditioning of the matrix and may fail to converge in some situations.

Here we present two of the most basic iterative methods in some detail. This presentation is designed to give readers some flavor for how iterative methods arrive at a solution. The more advanced methods, which are the methods more likely to be found in simulation software packages, are mentioned briefly in the final section, with references that provide more details on their procedures. Their complexity puts them outside of the scope of this text.

A.4.2.1 Jacobi's Method. One of the simplest iterative methods for solving linear algebraic equations is Jacobi's method. The following example illustrates this technique:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \qquad \text{(A - 24a)}$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \qquad \text{(A - 24b)}$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \qquad \text{(A - 24c)}$$

The solution process begins by solving for x_1 from Eq. A-124a, then x_2 is determined from Eq. A-24b, and finally x_3 is determined from Eq. A-24c. In addition, a dummy index k is introduced to number the iterations performed in approximating the solution. In particular the equations used to get the next iteration from the last one are

$$\begin{aligned} x_1^{k+1} &= (b_1/a_{11}) - (a_{12}/a_{11})x_2^k - (a_{13}/a_{11})x_3^k \\ x_2^{k+1} &= (b_2/a_{22}) - (a_{21}/a_{22})x_1^k - (a_{23}/a_{22})x_3^k \\ x_3^{k+1} &= (b_3/a_{33}) - (a_{31}/a_{33})x_2^k - (a_{32}/a_{33})x_3^k \end{aligned} \qquad \text{(A - 25)}$$

Thus, if we assume an initial vector $\{x\}^k = \{x_1^k, x_2^k, x_3^k\}$ and substitute it into Eq. A-25, a new vector $\{x\}^{k+1}$ is calculated. This process is continued until sufficient accuracy is achieved. That is, until

$$\{x\}^{k+1} \cong \{x\}^k, \quad k = 0, \ldots, N \qquad \text{(A - 26)}$$

where N is the number of iterations performed. In general, for an *nxn* set, Eq. A-25 can be expressed in the following matrix form:

$$\begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}_{k+1} = \begin{Bmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{Bmatrix} - \begin{bmatrix} 0 & A_{12} & \cdots & A_{1n} \\ A_{21} & 0 & \cdots & A_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ A_{n1} & A_{n2} & \cdots & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}_k \qquad \text{(A - 27)}$$

where

$$B_i = b_i/a_{ii} \qquad i = 1, \ldots. \ n$$
$$A_{i,j} = a_{ij}/a_{ii}, \qquad i = 1, \ldots. \ n, \qquad j = 1, \ \ldots \ , n$$
$$k = 0, \ldots , N$$

The reader is to be reminded that Eq. A-27 is guaranteed to converge if condition 2 is met. Otherwise a solution may or may not be possible. Also, computer software packages for solving sparse matrices using the Jacobi conjugate gradient method (a variation of the Jacobi method) are available - such as ELLPACK, developed at Purdue University.[2]

Example A-6: Solve the following set of linear algebraic equations using the Jacobi method:

$$\begin{aligned} 5x_1 - 2x_2 + x_3 &= 4 \\ x_1 + 4x_2 - 2x_3 &= 3 \\ x_1 + 2x_2 + 4x_3 &= 17 \end{aligned}$$

Solution: First check the diagonal coefficients and determine whether the iterative procedure is suitable. Thus

$$|5| > |-2| + |1| \ \text{---}> \ 5 > 3 \ \text{O.K.}$$

$$|4| > |1| + |-2| \ \text{---}> \ 4 > 3 \ \text{O.K.}$$

$$|4| > |1| + |2| \ \text{---}> \ 5 > 3 \ \text{O.K.}$$

Obviously, the iterative approach will converge. Therefore, expressing the set in the form of Eq. A-20 yields

$$\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix}_{k+1} = \begin{Bmatrix} 0.8 \\ 0.75 \\ 4.25 \end{Bmatrix} - \begin{bmatrix} 0 & -0.4 & 0.2 \\ 0.25 & 0 & -0.5 \\ 0.25 & 0.5 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix}_k \qquad \text{(A - 28)}$$

Assuming an initial iteration $\{x\}_0 = \{0\}$ and substituting into equation yields our first approximation to the solution. That is,

$$\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix}_1 = \begin{Bmatrix} 0.8 \\ 0.75 \\ 4.25 \end{Bmatrix} - \begin{bmatrix} 0 & -0.4 & 0.2 \\ 0.25 & 0 & -0.5 \\ 0.25 & 0.5 & 0 \end{bmatrix} \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}_0 = \begin{Bmatrix} 0.8 \\ 0.75 \\ 4.25 \end{Bmatrix}_1$$

Substituting $\{x\}_1$ into Eq. A-28 gives the second approximation to the $\{x\}$ vector:

$$\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix}_2 = \begin{Bmatrix} 0.8 \\ 0.75 \\ 4.25 \end{Bmatrix} - \begin{bmatrix} 0 & -0.4 & 0.2 \\ 0.25 & 0 & -0.5 \\ 0.25 & 0.5 & 0 \end{bmatrix} \begin{Bmatrix} 0.8 \\ 0.75 \\ 4.25 \end{Bmatrix}_1 = \begin{Bmatrix} 0.25 \\ 2.675 \\ 3.675 \end{Bmatrix}_2$$

The process is continued until successive values of each vector are sufficiently close in magnitude. For the example at hand, the first ten iterations are shown in Table A-1.

Table A-1. Successive Approximations of Solution (Jacobi's Method)

Variable	Iteration 1	2	3	4	5	6	7	8	9	10
x_1	0.8	0.25	1.14	1.24	1.02	0.92	0.98	1.02	1.01	0.99
x_2	0.75	2.68	2.53	1.89	1.79	1.99	2.07	2.02	1.98	1.99
x_3	4.25	3.68	2.85	2.70	2.99	3.10	3.02	2.97	2.98	3.01

A.4.2.2 Gauss-Seidel Method. This method converges more rapidly than the Jacobi method. However, it is constrained by the same conditions outlines in the previous section. The Gauss-Seidel solution procedure presupposes that the new solution values are a better approximation to the exact solution than the initial assumed values. This is true for most problems. Consequently for the 3x3 set given by Eq. A-19 the solution for the new values becomes

$$\begin{aligned} x_1^{k+1} &= (b_1/a_{11}) - (a_{12}/a_{11})x_2^{k} - (a_{13}/a_{11})x_3^{k} \\ x_2^{k+1} &= (b_2/a_{22}) - (a_{21}/a_{22})x_1^{k+1} - (a_{23}/a_{22})x_3^{k} \\ x_3^{k+1} &= (b_3/a_{33}) - (a_{32}/a_{33})x_1^{k+1} - (a_{32}/a_{33})x_2^{k+1} \end{aligned} \qquad \text{(A - 29)}$$

Comparison of Eq. A-25 with A-29 clearly shows that one need not assume an initial x_1 value. Instead, x_1 is computed by assuming x_2 and x_3. Then the new x_1 value and the assumed x_1 value are used to calculate a new x_2 value. This procedure is continued by substituting the new x_1 and x_2 values to determine a new x_3 value.

Note that the Gauss-Seidel method will converge under the same conditions required for convergence of the Jacobi method. For many engineering problems, this condition can be satisfied. For example, the Poisson equation is readily solved using the Gauss-Seidel method. The Gauss-Seidel method (or variations thereof) are more widely used

than the Jacobi method whenever an iterative solution is sought, because it converges faster.

Example A-7: Rework Example A-6 using the Gauss-Seidel method.

Solution: Since the Jacobi method converged for this particular example, the Gauss-Seidel technique should also converge. Consequently, expressing the set in the form of Eq. A-29 yields:

$$\begin{aligned} x_1^{k+1} &= 0.8 + 0.4x_2^{k} - 0.2x_3^{k} \\ x_2^{k+1} &= 0.75 - 0.25x_1^{k+1} + 0.5x_3^{k} \\ x_3^{k+1} &= 4.25 - 0.25x_1^{k+1} - 0.5x_2^{k+1} \end{aligned} \qquad \text{(A - 30)}$$

Now, assuming initial values of $x_2 = 0$ and $x_3 = 0$ and substituting into Eq. A-30 gives for $k = 0$

$$\begin{aligned} x_1^{1} &= 0.8 + 0 - 0 = 0.8 \\ x_2^{1} &= 0.75 - 0.250\,(0.8) + 0 = 0.55 \\ x_3^{1} &= 4.25 - 0.250\,(0.8) - 0.5\,(0.55) = 3.775 \end{aligned}$$

This process is continued by substituting the new x_2^1 and x_3^1 into Eq. A-30 permitting the determination of yet another set of x values. Computations for the first seven iterations are summarized in Table A-2.

Table A-2. Successive Approximations of Solution (Gauss-Seidel Method)

Variable	Iteration 1	2	3	4	5	6	7
x_1	0.8	0.265	1.249	0.956	1.002	1.001	0.999
x_2	0.55	2.571	1.887	2.008	2.003	1.999	2.000
x_3	3.775	2.898	2.994	3.007	2.998	3.000	3.000

Note that the results obtained in five iterations are more accurate than those obtained in ten iterations using the Jacobi method.

A.4.2.3 Advanced Iterative Solution Techniques. Although, in principle, the simple iterative methods (Jacobi or Gauss-Seidel) can be used to solve a set of linear algebraic equations, in practice, several other methods are better suited for this task. That is, other techniques with increased rate of convergence have been developed, and it is generally one of these that are typically used in the software packages that solve such equations.

The first of these advanced iterative methods the *successive over-relaxation* (SOR) method (and its variations, such as the *block SOR method*).[3] The SOR method is

essentially an enhanced form of the Gauss-Seidel method. It converges more rapidly as a result of multiplying the *residual* (the difference between the latest estimate of the unknown and the previous estimate) by a relaxation factor ω so that

$$x_{ij}^{k+1} = x_{ij}^{k} + \omega (x_{ij}^{k+1} - x_{ij}^{k}) \qquad \text{(A - 31)}$$

The optimum relaxation factor is chosen to provide either a minimum error in a specific number of iterations or to ensure convergence of the solution to a required tolerance in the minimum number of iterations. The procedure for selecting ω starts with $\omega=1$ (equivalent to the Gauss-Seidel method), but then increases or decreases ω by 0.1 (or less) in the succeeding iterations until the residuals monitored change sign, which is an indication of an oscillating solution.

The other two advanced iterative methods are implicit methods: (1) the *alternating-direction implicit* methods;[4] and (2) *Stone's strongly implicit method.*[5] Both were developed to solve discretized elliptic partial differential equations, such as the Poisson equation.

The Jacobi, Gauss-Seidel and SOR methods can also be significantly accelerated by applying several other methods including the conjugate gradient method, the conjugate residual (GCR) method, the Chebyshev method, and the incomplete Cholesky-conjugate gradient (ICCG) method.[6]

A.5 NON-LINEAR ALGEBRAIC EQUATIONS

We noted in section that the set of discretized equations at each node of a device grid is a set of nonlinear algebraic equations. (Nonlinear algebraic equations are defined as those which contain powers of variable(s) and/or transcendental functions.) To solve such equations, the object is to find the so-called zeros of the equation. That is, given

$$y = f(x) \qquad \text{(A - 32)}$$

we seek values of x such that y is zero. The x values that result in y = 0 are referred to as the roots of the function f(x). No general algebraic method is available for solving all nonlinear algebraic equations, and a variety of iterative techniques with varying degrees of accuracy and rates of convergence have been developed (including the one we present here, the *Newton-Raphson method*). The degree of accuracy attained for a given root is directly related to the amount of computational effort and the method being used.

We show this method because it provides an example of an iterative numerical procedure used to solve a nonlinear algebraic equation and it illustrates the principle of convergence.

A.5.1 Graphical method

This method does not linearize the nonlinear equations, but in fact solves for the roots directly using a graphical solution method. It's chief role is to establish a range of

values for the real roots of a given function (i.e., it provides no information relative to complex roots). That is, it is mainly used to find first approximations of the roots; these are then used as the initial estimate in more accurate numerical techniques. Basically, the function f(x) is evaluated at a number of points, and the results are plotted. The locations where the function crosses the x-axis are the roots. Figure A-1a a shows how the graphical method is applied for approximating the real roots of the function $y = x^3 - 6x^2 + 11x - 6$. If a more accurate estimate is needed, the scales of the ordinate and abscissa can be altered, as shown in Fig. A-1b.

A.5.2 Newton-Raphson Method

Iterative methods are generally used to find more accurate values of the roots of nonlinear algebraic equations. One of the more popular iterative methods is the *Newton-Raphson method.* It offers the benefit of converging to an accurate solution relatively quickly. However, it suffers from the limitation that an initial value close to the root must first be estimated and the derivative of the function must be evaluated. That is, if the initial estimate of the root is not sufficiently close to the actual root, the method may not always converge to a root.

To find the k+1th iteration of the approximation of the root of a nonlinear algebraic equation using the Newton-Raphson method, the following equation is used:

$$x^{k+1} = x^k - f(x^k) / f'(x^k) \qquad \text{(A - 33)}$$

where x^{k+1} = the approximate root after k+1 iterations
x^k = the approximate root after k iterations
$f(x^k)$ = functional value at x^k
$f'(x^k)$ = first-derivative of the function at x^k
k = 1, 2, ...

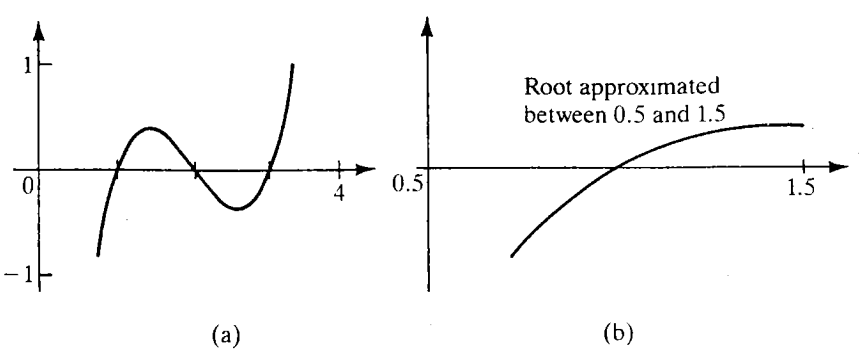


Fig. A-1 Procedure used in approximating the real roots of the function $y = x^3 - 6x^2 + 11x - 6$.

For a system of nonlinear algebraic equations, the Newton-Raphson technique can also be used to find the roots. For a set of equations, the equation above in matrix form is

$$\{\mathbf{x}\}^{k+1} = \{\mathbf{x}\}^{k} - [\mathbf{J}]_k^{-1} \{\mathbf{f}\}^{k} \qquad \text{(A - 34)}$$

where [J] is the square matrix of the partial derivatives called the *Jacobian* (and $[\mathbf{J}]^{-1}$ is the inverse of the Jacobian).

Example A-8: Finding the Roots of a Set of Nonlinear Algebraic Equations Using the Newton-Raphson Method. Estimate one set of roots for the following system of nonlinear algebraic equations:

$$x_1^2 + x_2^2 = 18 \qquad f_1(x_1, x_2) = x_1^2 + x_2^2 - 18$$
$$x_1 - x_2 = 0 \qquad f_2(x_1, x_2) = x_1 - x_2$$

Solution: This system has an exact solution of $x_1 = \pm 3$, $x_2 = \pm 3$. However, let us use the Newton-Raphson method to estimate these roots. If a graphical procedure is used to obtain an initial estimate of the roots, we might get $x_1 = 2$ and $x_2 = 2$.

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 2 \end{Bmatrix}_1$$

Then the Jacobian is determined and its inverse is evaluated, or

$$[\mathbf{J}] = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 \\ 1 & -1 \end{bmatrix}$$

and

$$[\mathbf{J}]^{-1} = \frac{1}{2x_1 + 2x_2} \begin{bmatrix} 2x_1 & 2x_2 \\ 1 & -1 \end{bmatrix}$$

Therefore, evaluating [J] at the initial assumed values of $x_1 = 2$ and $x_2 = 2$ gives

$$[\mathbf{J}]_1^{-1} = \frac{1}{8} \begin{bmatrix} 1 & 4 \\ 1 & -4 \end{bmatrix}$$

The functions are now evaluated to give

$$\begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix}_1 = \begin{Bmatrix} -10 \\ 0 \end{Bmatrix}_1$$

Substituting into

$$[\mathbf{x}]^{k+1} = [\mathbf{x}]^{k} - [\mathbf{J}]_k^{-1}\} \ \mathbf{f}^{k}$$

gives

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}_2 = \begin{Bmatrix} 2 \\ 2 \end{Bmatrix}_1 - \frac{1}{8}\begin{bmatrix} 1 & 4 \\ 1 & -4 \end{bmatrix}_1 \begin{Bmatrix} -10 \\ 0 \end{Bmatrix}_1 = \begin{Bmatrix} 3.25 \\ 3.25 \end{Bmatrix}_2$$

The inverse of the Jacobian and the functions are now evaluated at $x_1 = 3.25$ and $x_2 = 3.25$; then, using Eq. A-34, we get

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}_3 = \begin{Bmatrix} 3.25 \\ 3.25 \end{Bmatrix}_2 - \frac{1}{13}\begin{bmatrix} 1 & 2.5 \\ 1 & -2.5 \end{bmatrix}_2 \begin{Bmatrix} 3.125 \\ 0 \end{Bmatrix}_2 = \begin{Bmatrix} 3.0096 \\ 3.0096 \end{Bmatrix}_3$$

Clearly in just two iterations, we have achieved an acceptable level of accuracy.

References

1. I.S. Duff, "A Survey of Sparse Matrix Research, "*Proceedings IEEE,* **65-4**, p. 500 (1977).
2. J.R. Rice, *Mathematical Software III.* New York, Academic Press, 1977, p. 319.
3. R.S. Varga, "Matrix Iterative Analysis," Prentice-Hall, Englewood Cliffs, 1962.
4. E.L. Wachspress, "Iterative Solution of Elliptic Systems," Prentice-Hall, Englewood Cliffs, 1966.
5. H.L. Stone, "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations," *SIAM J. Numer. Anal.*, Vol. 5 p. 530, 1968.
6. Grimes et al., "ITPACK 2A - A Fortran Implementation of Adaptive Accelerated Iterative Methods for Solving Large Sparse Linear Systems," University of Texas, Austin, Report No. CNA-164, 1980.

APPENDIX B

MAXIMIZING THE CIRCUIT PERFORMANCE OF LONG-CHANNEL MOSFETS THROUGH DEVICE DESIGN AND PROCESSING TECHNOLOGY

The desired circuit characteristics of active MOSFETs in digital ICs include the following: high output-current drive, precisely-manufacturable and stable threshold voltage (V_T), very small subthreshold current (I_{Dst}) at $V_{GS} = 0$, high gate-oxide breakdown voltage, high drain/body breakdown voltage, low source/drain-to-substrate junction capacitances, and high reliability. For the parasitic (field) MOSFET structures threshold voltages (V_{TF}) need to be large enough to prevent leakage between devices. For long-channel MOS devices (L>2.0 μm), the relationship between these electrical parameters and the various MOS device structure is provided by the relatively simple long-channel models of MOS circuit behavior given in chap. 4.

B.1 CHARACTERISTICS OF DIGITAL IC MOSFETS

It is useful to list the desired circuit characteristics of submicron MOSFETs in digital VLSI and ULSI CMOS chips. Depending on the circuit application, the relative importance of each characteristic can vary. In subsequent sections of this appendix we provide examples of how these device characteristics impact circuit performance. Specifically, we would like a MOSFET designed for use in digital applications to exhibit the following characteristics:

1. High output-drive current (e.g., I_{Dsat} >400 μA/μm) when the device is ON.
2. Symmetrical I_D behavior (i.e., I_D should be unchanged if source and drain are reversed).
3. Low leakage current when the device is OFF (e.g., I_D of individual devices should be 10^{-9}-10^{-12} A when the applied gate voltage is zero).
4. Small parasitic capacitances (i.e., junction and overlap capacitance).
5. Small parasitic resistance (the sum of the contact resistance and the source/drain tip series resistance [of an LDD] should be much smaller than the channel resistance).
6. Uniform threshold voltage V_T (e.g., ±0.1V) throughout each circuit, as well as from circuit to circuit. V_T value should also remain stable (i.e., within ±10% of its original value) over the expected operating lifetime of the chip.
7. Small-valued body-effect coefficient (i.e., this makes V_T less sensitive to source-body bias).

8. High device reliability (e.g., devices should not exhibit hot-carrier-induced degradation or inadequate time-dependent dielectric breakdown, TDDB).

B.1.1 Output Current (I_{Dsat}) and Transconductance (g_m). The long-channel models indicate that I_{Dsat} in MOSFETs can be impacted by various device parameters. First, increasing the gate width linearly increases I_{Dsat}, making this the most obvious way to raise drive current. However, when minimum-sized devices must be used (e.g., for maximum packing density), this avenue may not be available, and the impact of the other parameters on drive current must be examined.

From the dependence of I_{Dsat} on μC_{ox}, both the mobility of the carriers in the channel and the gate-oxide capacitance should be as high as possible. Since electron mobility is greater than hole mobility, circuits using NMOS devices will exhibit higher performance than those built with PMOS devices. In fact, long-channel NMOSFETs of the same width provide roughly 2.5x the drive current of PMOSFETs. In addition, the mobility of carriers decreases as the doping concentration of the channel increases beyond ~$1x10^{17}$ cm^{-3}. Hence, lighter doped channel regions are also favored.

Because the value of C_{ox} is inversely proportional to t_{ox}, as thin a gate oxide as possible (commensurate with oxide breakdown and reliability considerations) is normally used. I_{Dsat} is also inversely proportional to L, and minimum channel lengths are therefore desirable. On the other hand, as the channel lengths become shorter, device operation will be altered so that new models must be developed, as described in section 5.5.

Although the equations indicate that (V_{GS} - V_T) should be as large as possible for maximum I_{Dsat}, $V_{GS} = V_{DD}$ is usually fixed by system specifications and material limitations, and is not alterable by the device designer. Similarly, V_T is selected primarily by other circuit considerations.

B.1.2 Effect of Parasitic Series Resistance on MOSFET Performance. Parasitic resistances consisting of the metal-to-Si contact resistance and the series resistance of the path in the source and drain between the contact and channel also exist in MOSFETs. Since they are in series with the channel, their value R_S is added to the channel resistance R_{ch} to give the total MOS dc output resistance. This implies that R_S may adversely impact I_D and the transconductance (g_m). An approach for minimizing this impact is to keep R_S small compared to R_{ch}. For example, the maximum value of R_S could be limited to 10% of R_{ch}. (While this 10% percent limit may seem arbitrary, it is approximately equal to the R_S/R_{ch} ratio in 1.0 μm MOS devices. Thus, maintaining this ratio as devices shrink seems reasonable.)

Since R_S is normally much smaller than R_{ch} in long-channel MOSFETs this goal is automatically satisfied. However, as devices get smaller, R_{ch} (which is directly proportional to L) also shrinks, while R_S grows larger (because of decreased contact sizes and shallower source/drain junction depths). Eventually, it becomes necessary to determine the channel lengths at which R_S may be a significant fraction of the total MOS resistance. Here we summarize how estimates of R_S have been derived for devices with L<1 μm.

To begin, the relation used to calculate the MOS channel resistance, R_{ch} (with the device operating in the linear region of the I_D vs. V_{DS} curves), is normalized to unit width, and is therefore expressed in units of Ω-μm. Then, the value of R_{ch} in the linear region is selected to represent the channel impedance, as it provides the worst case comparison with R_S. An expression from which R_{ch}(linear) can be calculated is

$$R_{ch} \text{ (linear region)} = [(L_{eff} + \{V_{DS}/E_c\})/[\mu_o C_{ox}(V_{GS} - V_T - V_{DS}/2)] \qquad \text{(B - 1)}$$

where $\mathcal{E}_c$ represents the critical field at the onset of velocity saturation (see chap. 5). Note that this equation is a modified form of the MOSFET channel resistance equation, which includes in the numerator the term V_{DS}/E_c, whose origin stems from the mobility dependence on the longitudinal electric field. That is, in Eq. B-1, instead of including $\mathcal{E}_c$ in a factor also containing L_{eff} and V_{DS}, the term V_{DS}/E_c is added to L_{eff} to obtain a simpler formula. This leads to a mobility decrease as being equivalent to an increase in L_{eff} for channel resistance or transconductance calculations.

A graph of the values of R_{ch} versus channel length calculated using this equation for NMOS devices and PMOS devices is shown in Fig. B-1, assuming V_{GS} = 5 V. In this figure, the 10% value of R_{ch} is also plotted. We note that as the devices get smaller, the value of R_{ch} also decreases. These graphs indicate that the maximum allowed values for R_S for submicron devices will thus be in the 300-500 Ω-μm range.

A more sophisticated constraint than the simple limit on R_S was proposed by Ng and Lynch.[1] They performed an analysis in which an estimate of the decrease in switching speed was calculated as R_S was increased and L_{eff} decreased. Overall, they found that the intrinsic series-resistance effects will negatively impact the speed performance of MOS devices as the devices are scaled below 0.5 μm. Nevertheless, they predict that it will be possible to build devices that demonstrate improvement of speed performance as they are scaled to 0.15 μm. The use of self-aligned, silicided source/drain regions to reduce the contact resistance will be important, as will the development of processes that produce junctions with more steeply graded doping profiles. In addition, contact technologies that provide smaller contact resistance and lower sheet resistances of the diffusion layer under the contact window (or the silicided region) will need to be developed.

B.2 Desired Circuit Characteristics of IC Interconnects

Circuit performance characteristics of the IC interconnect scheme are also important. These include:

1. Low sheet resistance values of the conductor lines.
2. Low conductor-substrate parasitic-capacitance values.
3. High reliability of the interconnect system.

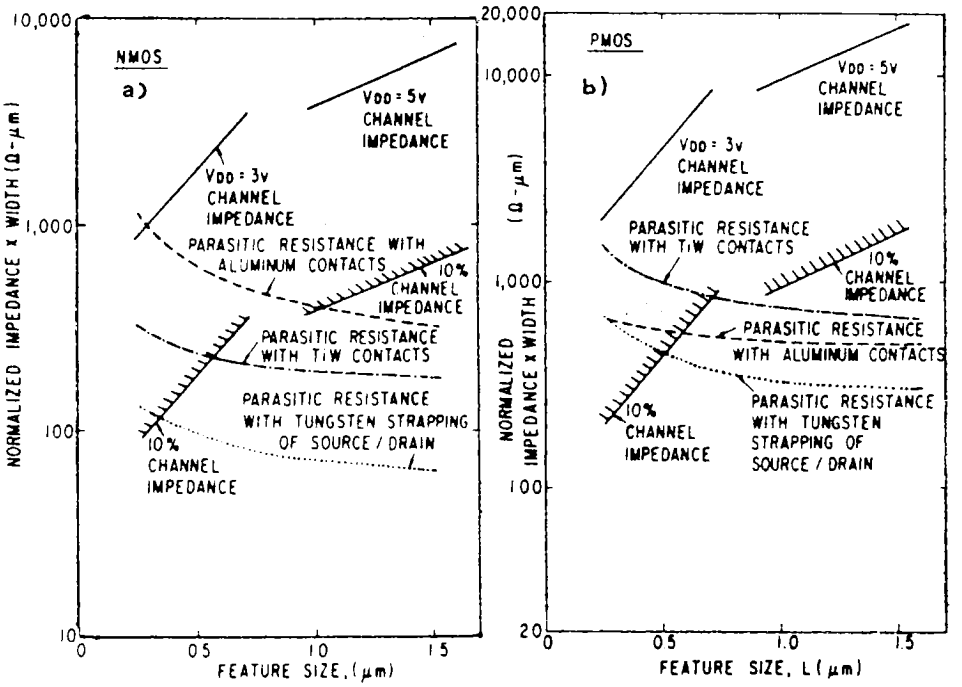


Fig. B-1 Plot of minimum channel resistance and parasitic series resistance versus design rule feature for (a) NMOS, and (b) PMOS. The parasitic residences are calculated for three different contact technologies: (1) Al --- ; (2) Ti:W alloy ··· ; and (3) self-aligned W strapping of source and drain.[4] (© 1986 IEEE).

B.2.1 Circuit Speed as Impacted by MOSFET Device and Interconnect Conductor Characteristics. Speed performance is especially important in densely packed static logic and dynamic logic (digital MOS) ICs. *Static* logic circuits* are often found in the data path and control sections of digital ICs, and examples are the digital multiplier and the static arithmetic logic unit.* *Dynamic* logic circuits** are functionally the same as static logic circuits, but employ dynamic circuit

* In CMOS static-logic-circuits the logic function of each gate is implemented in both *p*- and *n*-channel devices. The use of complimentary MOSFETs in each gate results in larger area circuits per function and larger input capacitive loads since both *p*- and *n*-channel devices must be driven. However, the dual-device types give a major advantage to fully-static logic circuits, namely they exhibit a large noise immunity (which is necessary for fast turnaround designs, where detailed circuit design is not practical).

** Dynamic logic circuits offer the advantage of speed and functional density at the expense of extra design time. The output nodes of the logic gates are precharged to a particular level when the clock has a value 0. During this time (i.e., when the clock value is 0) the inputs to the circuits also change. Then, when the clock goes to value 1, the output may be pulled down to the complementary value, depending on the input conditions. Both functional density and speed are improved because it is not necessary to use a pull-up *p*-type transistor. This leads to almost a 50% savings in silicon area compared to static logic, and a reduction in output capacitance.

techniques such as domino logic. An example of this type of circuit is the ALU in the Belmac 32-bit processor.

A fundamental measure of digital circuit speed is the delay time per gate τ_d. We will perform a rudimentary analysis of the factors that impact τ_d in a MOS logic gate so that the main device characteristics that establish this parameter can be identified.

In MOS logic gates the MOSFET output current (assumed to be I_{Dsat}) is used to discharge the load capacitance C_L (as shown in Fig. B-2). If the incremental change in output voltage is defined as ΔV and the incremental discharge time as Δt (and if we neglect the current through R_1)

$$I_{Dsat} = C_L \Delta V/\Delta t \qquad \text{(B - 2)}$$

where $\Delta Q = C_L \Delta V$ is the charge stored in the capacitor. Therefore, the delay time

$$\tau_d \sim \Delta t = C_L \Delta V/ I_{Dsat} \qquad \text{(B - 3)}$$

if I_{Dsat} is constant. This indicates that an increase in I_{Dsat} will decrease the logic gate delay time, and exemplifies the desirability of large I_{Dsat} in digital MOSFETs.

By using the simplified form for I_{Dsat} (see Eq. 4-68), and approximating $V_{GS} - V_T = V_{DD}$, we have

$$I_{Dsat} = \mu Z C_{OX} V_{DD}^2/2L \qquad \text{(B - 4)}$$

where μ, C_{OX}, L, and Z are the carrier mobility, oxide capacitance (in units of capacitance per unit area), channel length, and width of the MOSFET, respectively.

Next, assume the logic swing ΔV is equal to the power-supply voltage V_{DD}, and C_L is roughly the sum of the gate capacitance of the next stage ($C_{OX}ZL$) and any other parasitic capacitances C(parasitic) of the next stage and the interconnect path. That is

$$C_L \sim C_{OX} Z L + C(\text{parasitic}) \qquad \text{(B - 5)}$$

If $C(\text{parasitic}) << C_{OX} Z L$, the load capacitance is essentially just the gate capacitance

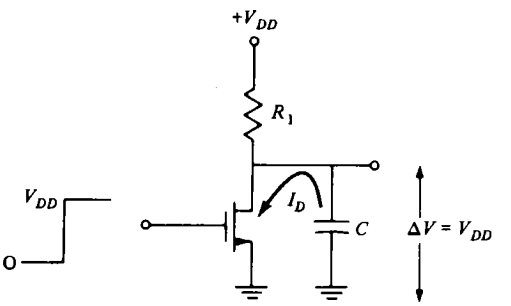


Fig. B-2 Load capacitance discharged to output current in an FET circuit.

of the next MOSFET. In this case, substituting Eqs. B-4 and B-5 into Eq. B-3 and simplifying, we find

$$\tau_d = 2L^2/\mu V_{DD} \qquad \text{(B - 6)}$$

Equation B-6 shows that if a circuit is dominated by intrinsic capacitance the delay time is proportional to L^2. Thus, significant improvement in switching speed can be realized by downsizing L, but decreasing t_{ox} will have no impact.

On the other hand, if C(parasitic)>>$C_{ox}ZL$, τ_d will be decreased if t_{ox} is made thinner, because this will make I_{Dsat} grow larger, and the improvement will not be offset by the increase in C_{ox} [since C_{ox} is insignificant compared to C(parasitic)].

In either case, we can see that the speed of MOSFET logic gates can generally be increased by increasing I_{Dsat} and/or decreasing C_L. If C_{ox} dominates, shrinking the MOSFET dimensions will decrease the gate delay time. If C(parasitic) dominates, gate delay will be decreased by decreasing C(parasitic). Thus, it is also useful to examine the nature of C(parasitic).

B.2.2 Components of the Parasitic Capacitance. There are several effects in a MOS IC that contribute to C(parasitic), including the gate-drain overlap capacitance C_{gdo}, drain-substrate junction capacitances C_j,* and the capacitance associated with the interconnect lines (or wiring) C_w. In long-channel MOSFETs G_{gdo} was relatively small, but in submicron MOSFETs C_{gdo} can become significant. This is because C_{gdo} can be increased by such drain structures as LDDs. While these are designed to reduce hot-electron degradation in MOSFETs, the best structures in terms of hot-electron resistance may give rise to the largest C_{dgo} (see chap. 5). C_j is reduced by reducing the drain-to-substrate junction area and also by keeping the substrate region under the drain as lightly doped as possible.

In large VLSI chips, the delay caused by the C_wR_w product can dominate, where R_w is the resistance of the interconnect line. The delay resulting from various interconnect materials per millimeter of length running over a 1 μm thick field oxide is graphed for some of these materials (assuming a 1-cm-long interconnect line) in Fig. B-3. This shows, for example, that if a polysilicon line having a 20 Ω/sq sheet resistance is used to connect one corner of a 0.7-x-0.7 cm chip to the diagonal corner, the propagation delay will be 1.4 μs. Even if Al is used as the interconnect layer, the delay caused by 1-cm-long interconnection lines will limit the circuit speeds once device dimensions fall below ~2 μm.

The latter data indicate that in order for circuit performance to be increased as device dimensions shrink, two goals must be met. First, the materials used for transmitting

* When the MOSFET in Fig. B-2 is turned on, electrons flow through it to the positively charged load capacitance C_L tied to the drain. The electrons discharge C_L, causing the drain voltage to drop. As the drain voltage falls, the drain-to-substrate junction must also discharge. Thus, the drain-to-substrate junction capacitance adds to the load capacitance and, therefore, increases τ_d.

signals over long distances on a chip must have the lowest possible resistance values. This has been the key motivation in the drive to replace polysilicon interconnects (20 Ω/sq) with polycide interconnects (1.7 Ω/sq for $TiSi_2$). It may eventually cause refractory metals (0.5 Ω/sq for W) to be used as the first-level interconnect layer in multilevel-interconnect systems. In addition, it also points outs why Al will continue to be used as an interconnect material, even as devices shrink further.

Second, the length of interconnect lines on a chip must be made as short as possible. The RC delay can be shown to be proportional to the square of the length of the interconnect line. That is,

$$R_w = (\rho\, l)/(w\, t_m) \tag{B - 7}$$

and

$$C_w = \varepsilon\, w\, l / t_{ox}, \tag{B - 8}$$

and therefore

$$R_w C_w = (\rho\, \varepsilon\, l^2)/(t_{ox}\, t_m) \tag{B - 9}$$

where ρ is the resistivity, l the interconnect line length, w the line width, ε the permittivity of the insulator film, t_m the thickness of the metal, and t_{ox} the thickness of the oxide. Multilevel-interconnect structures can effectively be used to reduce the longest interconnect lines.

B.2.3 Drain Current in OFF-State MOSFETs. The drain current of the MOSFET in the OFF state (i.e., I_D when V_{GS} = 0V) is an undesired, but nevertheless important leakage current device parameter in many digital CMOS IC applications (including DRAMs, SRAMs, dynamic logic circuits, and portable systems). In

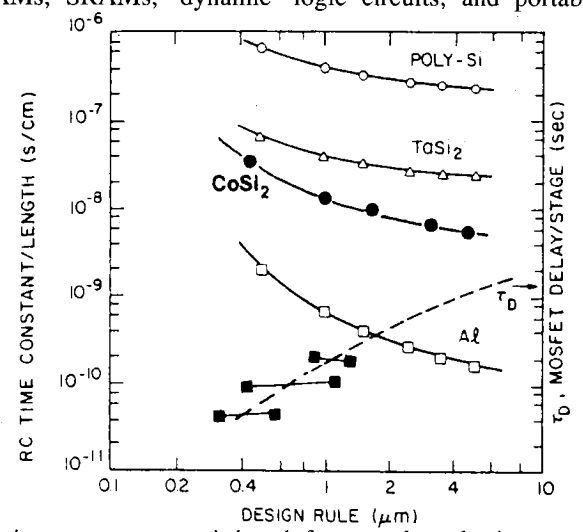


Fig. B-3 $R_W C_W$ time constant per unit length for several conductive materials as a function of feature size. Also shown is delay per stage of NMOS ring oscillators as a function of feature size. The $R_W C_W$ time constant is calculated assuming a field oxide thickness of 1 μm. Copyright AIOP.[5] Reprinted with permission.

addition, in any large device-count MOS IC, leakage currents in OFF-state transistors must be kept small to keep heating under control. (The standby power consumed by devices in the OFF state adds to the total power consumed by the IC, increasing heat dissipation problems in the chip and package).

In DRAMs, SRAMs, and dynamic-logic circuits, charge is stored on nodal capacitances isolated by access transistors or transmission gates (Fig. B-4 shows a transmission gate in a dynamic CMOS circuit). Once the capacitance is charged, the transmission gate is turned off by setting $V_{GS} = 0V$. The charged nodes must not be discharged during the time of a clock cycle by a leaky, nominally-OFF transmission-gate MOSFET. The allowable leakage is determined by how long a time the transmission gate must hold a voltage on a capacitive node. In portable systems, low standby-power-dissipation is needed for long operating times under battery-power.

The value of I_D when $V_{GS} = 0V$ consists of several components, including the subthreshold leakage current of the MOSFET I_{Dst}, the leakage current of the reverse-biased drain-substrate junction, gate-induced drain leakage (GIDL), and leakage currents due to loss of device isolation. In well-designed devices and IC processes it is possible to keep the latter three components smaller than I_{Dst} (procedures to accomplish this are discussed in chaps. 4 and 5). The value of I_{Dst} in long-channel devices when $V_{GS} = 0V$ (per unit width of the channel) depends on the device threshold voltage V_T, the subthreshold swing S_t, and the temperature of operation. As devices shrink, several short-channel effects arise that can increase the long-channel I_{Dst} value.

B.2.4 The Need for Precise Control and Stability of V_T in MOSFETs. Since the MOSFET models of section 4.5 indicate that I_{Dsat} is proportional to $(V_{GS} - V_T)$, it might appear that V_T should be as close to zero volts as possible. However, as V_T decreases, I_{Dst} will increase at $V_{GS} = 0V$ (see section 4.7), and the noise margin of the logic gate will decrease. Thus, the minimum V_T is

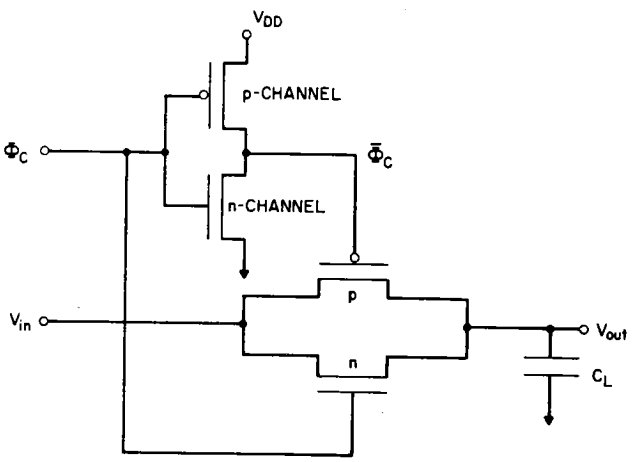


Fig. B-4 CMOS transmission gate driven by a clock signal φ_C, and its complement φ_C, generated locally by an inverter.

generally selected to ensure either a sufficiently small I_{Dst} when $V_{GS} = 0$, or a sufficiently large noise margin.

Once V_T is selected, the need for V_T stability during the normal lifetime of the device (e.g., 10 years) can be inferred from the previous discussion. If V_T decreases, I_{Dst} will increase (giving rise to problems associated with excess leakage), and noise margin will decrease. On the other hand, if V_T increases, I_{Dsat} will decrease since $(V_{GS} - V_T)$ will become smaller. Hot-carrier trapping in the gate oxide is the chief mechanism causing shifts in V_T during the life of the device, a topic discussed in chap. 9.

Another example of the importance of being able to control V_T involves semiconductor memory devices. In these circuits, charge flows from the memory cells to the sense amplifier. The sense amplifier is a delicately balanced flip-flop whose voltage-sensing capability is directly related to the threshold-voltage variation between the transistors. Hence, such circuits would not function reliably unless V_T in the circuit devices was highly uniform and stable.

B.2.5 Junction Breakdown Voltage (Drain-to-Substrate). The source and drain regions are heavily doped to minimize their resistivities. Thus, the breakdown voltage of the drain-to-substrate junction will be determined by the lighter doping concentration of the substrate. The breakdown voltage decreases as the doping increases. Hence, as substrate doping is decreased, the junction breakdown voltage will increase.

While the current in breakdown also represents a form of subthreshold I_D, it turns out that the $\mathcal{E}$-field needed to cause breakdown of the drain junction in submicron MOSFETs is smaller than that allowed by hot-carrier degradation limits. (As is discussed in chap. 9, hot-carrier degradation becomes excessive at $\mathcal{E}_{ymax} \sim 0.2$ MV/cm, whereas drain breakdown is estimated to occur for fields of the order of $\mathcal{E}_{ymax} \sim 0.6$ MV/cm.)[2] Thus, breakdown voltage is not considered to be a major limitation of submicron MOSFET design.[3]

B.2.6 Gate-Oxide Leakage and Gate Breakdown Voltage. Submicron MOSFETs with $L = 1.0$-0.5 μm use gate-oxide thicknesses of 12-25 nm. High-quality SiO_2 films typically break down at electric fields of 8-12 MV/cm (the exact value is a function of oxidation and anneal conditions, oxide charges, surface crystallographic orientation, surface preparation and a number of other factors). This corresponds to 8-12 V across a 10-nm-thick oxide.

As discussed in chap. 6, oxide breakdown may also occur at smaller electric-field values, due to process-induced flaws in the gate oxide. Such defects include: metal precipitates on the silicon surface prior to oxide growth; high defect density in the silicon lattice at the substrate surface (e.g., stacking faults and dislocations); pinholes and weak spots created in the gate oxide by particulates; thinning of the oxide during growth caused by the Kooi effect; and oxide wearout due to failure mechanisms related to hot-electron injection.

Charge transport through such oxides without breakdown can occur under high electric fields by Fowler-Nordheim tunneling (i.e., electrons tunnel to the oxide conduction band, and are then transported by drift current to the gate). Below around 4

nm (at less than 3V), some electrons will tunnel directly through the oxide. If oxides thinner than ~3 nm are used, tunneling increases to the degree that an inversion layer cannot form in the MOS capacitor (i.e., the tunneling current removes the inversion layer carriers faster than they are supplied by thermal generation). Direct tunneling therefore sets a fundamental lower limit of about 3 nm for the thickness of the gate oxide. A search for alternative gate dielectric materials to mitigate this limitation is being conducted.

REFERENCES

1. K.K. Ng and W.T. Lynch, *IEEE Trans. Electron Dev.*, March 1987, p. 503.

2. F. M. Klaassen, "Design and Performance of Micron-Size Devices," *Solid-State Electronics,* **21** (3), 565, (1978).

3. W.S. Feng et al., *IEEE Electron Dev. Letts.*, July 1986, p. 449.

4. D.M. Brown, M. Ghezzo, and J.M. Pembley, *Proceedings of the IEEE*, Dec. 1986, p. 1678.

5. A.K. Sinha, *J. Vac, Sci. Technol.* **19**, 778, (1981).

APPENDIX C

DERIVATION OF THE BSIM I_D MODEL

The *ionic-charge model* of Ihantola and Moll described in chap. 4 is the basis of the Level 2 MOSFET model in SPICE. The I_D equation in linear region of this model is given by Eq. 4-57 and is repeated here as C-1 for convenience:

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GB} - V_{FB} - 2\varphi_B)V_{DS} - \frac{1}{2}V_{DS}^2$$

$$- \frac{2}{3}\gamma\left[(V_{DS} + V_{BS} + 2\varphi_B)^{3/2} - (V_{BS} + 2\varphi_B)^{3/2}\right] \qquad (4 - 57) \quad (C - 1)$$

This model predicts I_D in strong inversion but it is too complicated for hand calculations and even for most applications involving computer simulation of large circuits. The complexity of Eq. C-1 is due to the 3/2 power terms in the right hand side of the equation, leading to long execution times if the model is used in the SPICE program to simulate the circuit behavior of a large circuit. Thus, a simpler model for this application is desirable, even at the expense of some accuracy. The Level 3 and Level 4 (BSIM) models of the MOSFET in SPICE utilize such a simpler model. Here we show how this model is derived starting from Eq. 4-57, as is done by Sheu.[1]

The 3/2 power terms of Eq. 4-57 have their origin in the square root term of Eq. 4-10, repeated here as Eq. C-2 with V_{BC} replaced by V_{BS}

$$Q_I = -C_{ox}\left[V_{GC} - V_{FB} - 2\varphi_B - V_{BC} - \gamma(2\varphi_B + V_{BS})^{1/2}\right] \qquad (4 - 10) \quad (C - 2)$$

which, in turn is due to the expression for Q_B given by Eq. 4-9 (Eq. C-3 here)

$$Q_{Bmax} = [2\kappa_{si}\varepsilon_o N_A(2\varphi_B + V_{BS})]^{1/2} = \gamma(2\varphi_B + V_{BS})^{1/2}] \qquad (C - 3)$$

The solid line of Fig. C-1 is a plot of the $-Q_{Bmax}/C_{ox}$ term of Eq. 4-10 versus effective reverse bias in the channel.[2] Since the slope of this curve is seen not vary very much, it is reasonable to attempt to approximate the $-Q_{Bmax}/C_{ox}$ term by the first two terms of its Taylor series about the point V_{SB}.

Likewise, if we were to plot the two terms containing V_{BS} in Eq. 4-57 [which we write here as $F(V_{DS}, 2\varphi_B + V_{DS})$],

$$F(V_{DS}, 2\varphi_B + V_{BS}) = \frac{2}{3}\left[(V_{DS} + V_{BS} + 2\varphi_B)^{3/2} - (V_{BS} + 2\varphi_B)^{3/2}\right] \qquad (C - 4)$$

we would also see that the slope of their plot would not vary much. Thus, it is appropriate to approximate these terms by using the first two terms of the Taylor series of $F(V_{DS}, 2\varphi_B + V_{DS})$, which are

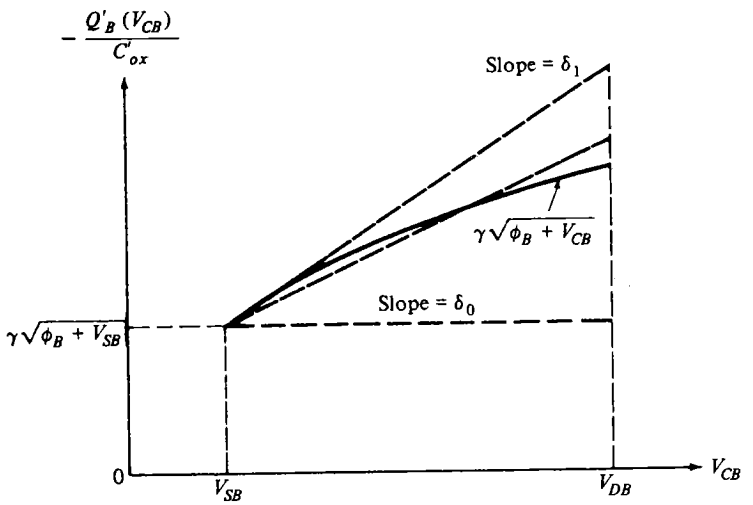


Fig. C-1 The quantity $-Q_B/C_{OX}$ in strong inversion versus effective reverse bias in the channel. Solid line: accurate calculation (i.e., using Eq. 4-56); upper broken line: first-order approximation from a Taylor expansion at $V_{CB} = V_{SB}$; middle broken line: improved first-order approximation; lower broken line: zero-order approximation (i.e., Square-law model - Eq. 4-63).[2] Reprinted with permission of McGraw-Hill Book Company.

$$F(V_{DS}, 2\varphi_B + V_{BS}) = \sqrt{2\varphi_B + V_{BS}}\, V_{DS} + \frac{0.25\, V_{DS}^2}{\sqrt{2\varphi_B + V_{BS}}} \qquad (C-5)$$

This expansion, however, is invalid if $(2\varphi_B + V_{BS}) >> V_{DS}$. To alleviate this problem, the expansion is changed to

$$F(V_{DS}, 2\varphi_B + V_{BS}) = \sqrt{2\varphi_B + V_{BS}}\, V_{DS} + \frac{0.25 a V_{DS}^2}{\sqrt{2\varphi_B + V_{BS}}} \qquad (C-6)$$

where $a\ (2\varphi_B + V_{BS})$ is determined by requiring the expansion in C-6 to give the best fit to $F\ [V_{DS}, 2\varphi_B + V_{BS}]$ in the desired voltage range.

The value of $(2\varphi_B + V_{BS})$ is considered in the range 0.7V-20.7V at 2-V increments. For each fixed value of $(2\varphi_B + V_{BS})$, the value of the parameter a is determined such that the expansion

$$\sqrt{2\varphi_B + V_{BS}}\, V_{DS} + \frac{0.25\, a\, V_{DS}^2}{\sqrt{2\varphi_B + V_{BS}}}$$

will give the best fit to $F\ [V_{DS}, 2\varphi_B + V_{BS}]$ in a least-square sense, over a range of V_{DS} from 0V to 10V at 0.5V increments. It is found that a can be accurately expressed as a function of $F\ [V_{DS}, 2\varphi_B + V_{BS}]$ in the following form:

$$1 / (1 - a) = P_1 + \{P_2\, (2\varphi_B + V_{BS})\} \qquad (C-7)$$

where P_1 and P_2 are determined by a least-square fitting over the range of $(2\varphi_B + V_{BS})$ from 0.7 to 20.7V. The results are

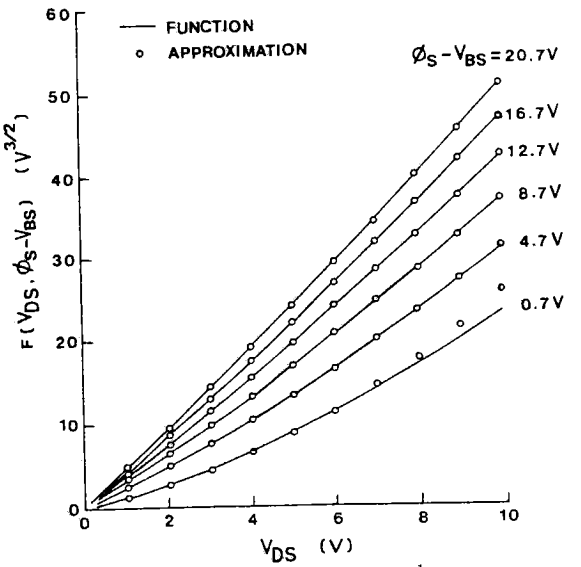


Fig. C-2 Approximating the function $F[V_{DS}, 2\varphi_B + V_{BS}]$.[1] (© IEEE 1987).

$$P_1 = 1.744 \qquad \text{(C - 8)}$$

and

$$P_2 = 0.8364. \qquad \text{(C - 9)}$$

The approximation (C-6) is compared to the function $F[V_{DS}, 2\varphi_B + V_{BS}]$ for various values of $(2\varphi_B + V_{BS})$ in Fig. C-2. The agreement between them is quite good, with the root-mean-square error of the approximation being 2 percent when the values of P_1 and P_2 given above are used.

The values of P_1 and P_2 are substituted into Eq. C-7 and this expression is then solved for *a*. The approximation (C-6) is used in Eq. C-1 for the terms given by Eq. C-4. Thus C-1 becomes (with additional terms also added for mobility degradation and velocity saturation effects as described in section 5.5):

$$I_D = \frac{\mu_0}{[\,1 + U_0(V_{GS} - V_T)} \frac{C_{ox}\frac{Z}{L}}{(1 + \frac{U_1}{L}V_{DS})}\left(V_{GS} - V_T)V_{DS} - \frac{aV_{DS}^2}{2}\right) \qquad \text{(C - 10)}$$

where

$$a = 1 + \frac{\gamma}{2\sqrt{2\varphi_B - V_{BS}}}\left[1 - \frac{1}{1.744 + 0.8364\,(2\varphi_B - V_{BS})}\right]$$

REFERENCES

1. B.J. Sheu and D.L. Scharfetter, "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. Solid-State Circuits,* **SC-22,** August 1987, p. 558.

2. Y.A. Tsividis, *Operation and Modeling of the MOS Transistor,* McGraw-Hill, New York, 1987, p. 124.

APPENDIX D

DERIVATION OF THE DISCRETIZED 2-D POISSON EQUATIONS USING THE BOX-INTEGRATION METHOD

In chapter 2 we showed how the difference-operator method (one of the two finite-difference discretization techniques) is used to discretize the 2-D Poisson equation. However, we noted that in such device simulators as GEMINI and PISCES-II the other finite-difference discretization technique (the box-integration method) is used instead, primarily because it permits non-planar device structures to be handled more easily. Here we show how the 2-D Poisson equation is discretized using the box-integration method, as well as how it facilitates the treatment of non-planar surfaces. As is done in the GEMINI device simulator, we assume a rectangular grid is used, together with the five-point approximation method. We mention, however, that in PISCES-II the box integration method is applied to an irregular (i.e., triangular-based) grid. Note also that the derivation given here is a close adaptation of the method presented by Greenfield and Dutton.[1]

In the box integration method the domain is divided into a set of rectangular subdomains, as well as a rectangular grid structure with nonuniform spacing in both spatial dimensions, as shown in Fig. D-1. The four nearest neighbor nodes are used in the discrete approximation to the continuous Poisson equation, as described in chap. 2. Figure D-2 depicts a node surrounded by its four nearest neighbors and defines the locations and potential values associated with these nodes. The perpendicular bisectors C_m of the lines joining the center node to the outside nodes defines a rectangle with four triangular subregions S_m. The finite difference approximation at the center node is derived by integrating Poisson's equation over the rectangle S. That is, the Poisson equation is first integrated twice on each side as follows:

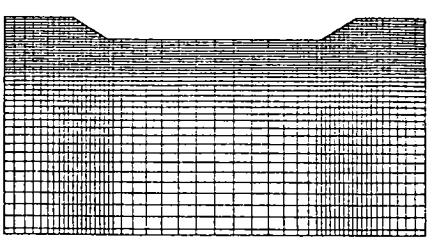


Fig. D-1 Non-uniform, non-planar rectangular grid.

$$\int_{S} \int_{y_o-\frac{h_4}{2}}^{y_o+\frac{h_2}{2}} \int_{y_o-\frac{h_1}{2}}^{x_o+\frac{h_3}{2}} \nabla \cdot [\varepsilon \nabla\varphi (x,y)]\, dx\, dy = -\int_{S} \int_{y_o-\frac{h_4}{2}}^{y_o+\frac{h_2}{2}} \int_{y_o-\frac{h_1}{2}}^{y_o+\frac{h_3}{2}} [\rho (x,y)]\, dx\, dy \qquad \text{(D - 1)}$$

By applying one of Green's theorems the double (surface) integral on the left-hand side of Eq. D-1 can be transformed into a line integral C of S. That is,

$$\int_{S} \int \nabla \cdot [\varepsilon \nabla\varphi (x,y)]\, dx\, dy = \int_{C} (\varepsilon \nabla\varphi (x,y) \cdot \hat{n}\, dl = \int_{C} (\varepsilon \frac{\partial\varphi}{\partial x} dy - \varepsilon \frac{\partial\varphi}{\partial y} dx) \qquad \text{(D - 2)}$$

The line integral on the right hand side must be split into four parts because of the discontinuous edges of the domain S. In addition, the surface integral over S (the right hand side of Eq. D-1), can also be broken up into four parts, each one representing the integral of the charge density in one of the four triangular regions of S. Thus, Eq. D-1 can be written as

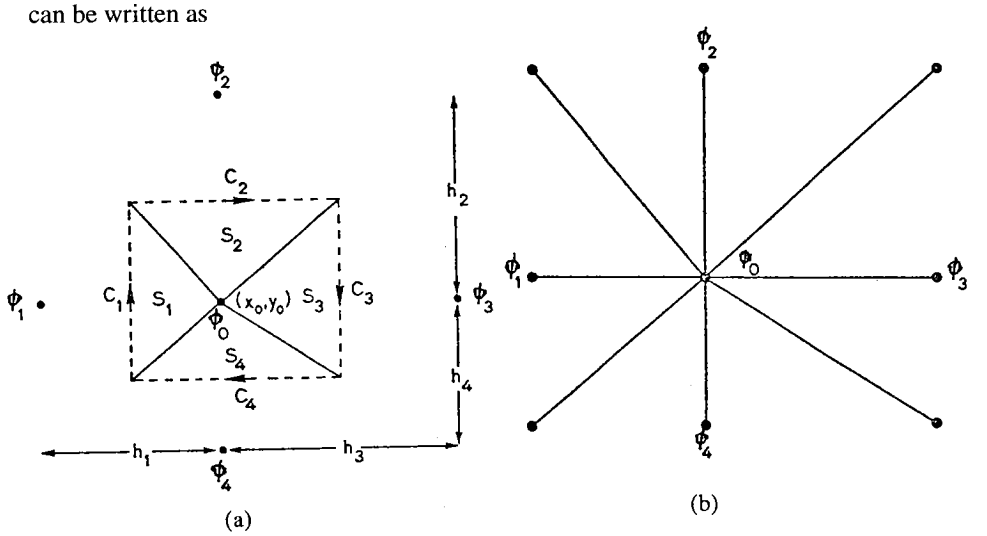


Fig. D-2 (a) A node and its four horizontally and vertically adjacent neighbors. The node locations and potentials are indicated. The C_m are perpendicular bisectors of the lines joining the center node to the outside nodes and define a rectangle S with triangular subregions S_m. (b) The eight line segments joining a node to its nearest and next nearest neighbors. Surfaces and material interfaces are composed of sequences of these segments.[1] (© IEEE 1980).

$$\int_{y-\frac{h_4}{2}}^{y+\frac{h_2}{2}} \left[\varepsilon \frac{\partial\varphi(x_o+h_3/2,y)}{\partial x}\right] dy + \int_{x-\frac{h_1}{2}}^{x+\frac{h_3}{2}} \left[\varepsilon \frac{\partial\varphi(x, y_o-h_4/2)}{\partial y}\right] dx +$$

$$\int_{y-\frac{h_4}{2}}^{y+\frac{h_2}{2}} \left[\varepsilon \frac{\partial\varphi(x_o-h_1/2,y)}{\partial x}\right] dy + \int_{x-\frac{h_1}{2}}^{x+\frac{h_3}{2}} \left[\varepsilon \frac{\partial\varphi(x,y_o+h_2/2)}{\partial y}\right] dx$$

$$= -\sum_{m=1}^{4} \left\{ \int \cdot \int \left[\rho\,(x,y,\varphi)\right] dx\, dy \right\} \qquad \text{(D - 3)}$$

Approximations are introduced to Eq. D-3 in terms of the potential values at the five nodes. The integrands on the left hand side are approximated as

$$\nabla\varphi\,(x, y) \cong \frac{\varphi_m - \varphi_o}{h_m} \qquad \text{(D - 4)}$$

and it is assumed that its value is constant over C_m (i.e., the component of the electric field $\nabla\varphi$ normal to each side of this rectangle is taken to be constant along that side).

The simplest approximation to the term on the right-hand side of Eq. D-3 is obtained by assuming that the impurity concentration and free carrier concentrations [and hence $\rho(x,y,\varphi)$] is constant in each triangular subdomain S_m, with its value being determined by evaluation at the coordinate of that node (i.e., using φ_o and N (x_o, y_o). These assumptions transform Eq. D-3 to:

$$\sum_{m=1}^{4} L_m \frac{\varphi_m - \varphi_o}{h_m} = -\rho\,(x, y, \varphi) \sum_{m=1}^{4} A_m \qquad \text{(D - 5)}$$

$$\sum_{m=1}^{4} B_m\,(\varphi_m - \varphi_o) = -\rho\,(x, y, \varphi) \qquad \text{(D - 6)}$$

where

$$B_m = L_m / (h_m\, A) \qquad \text{(D - 7)}$$

and

$$L_m = \int_{C_m} \varepsilon\, dl \qquad \text{(D - 8)}$$

and

$$A = \sum_{m=1}^{4} A_m \qquad \text{(D - 9)}$$

When the center node does not lie on the device surface or on a dielectric interface, A_m is the area of S_m. The handling of nonplanar surfaces and interfaces may be described in

terms of their effects on L_m and A_m in Eqs. D-5 to D-9. Only the calculation of L_1 and A_1 is described as the other pairs are computed in a similar fashion. Figure D-2b illustrates the eight line segments joining a node to its nearest and next nearest neighbors. We assume surfaces and interfaces are constructed from sequences of these segments.

For a node lying along the surface, a difference equation is only necessary when the boundary condition used at that node specifies the normal gradient of potential. Figure D-3 shows three orientations of the left portion of the surface which affect L_1 and A_1 in Eqs. D-5 to D-9 for a surface passing through the center node. The presence of the surface may eliminate part of C_1 and S_1 in Fig. D-2a, modifying L_1 and A_1. The dielectric permittivity below the surface does not affect S_1. In Fig. D-3a, the presence of the surface does not affect S_1. In Fig. D-3b, the surface eliminates the upper portion of S_1. In Fig. D-3c, the surface entirely eliminates S_1. Table D-1 summarizes the values of L_1 and A_1 for these cases.

Figure D-4 depicts three orientations of the left portion of the interface which affect L_1 and A_1 in Eqs. D-5 to D-9 for a dielectric interface passing through the center node. The dielectric permittivities above and below the interface are ε_i and ε_s respectively. In Fig. D-4a, S_1 is entirely within the semiconductor. In Fig. D-4b, the interface divides S_1 between the semiconductor and the insulator. In Fig. D-4c, S_1 is entirely within the insulator. The values of L_1 and A_1 for these cases are also listed in Table D-1. As an example of the derivation of the complete finite difference equation, consider Fig. D-4c where the center node lies at a bend in the dielectric interface with the interface

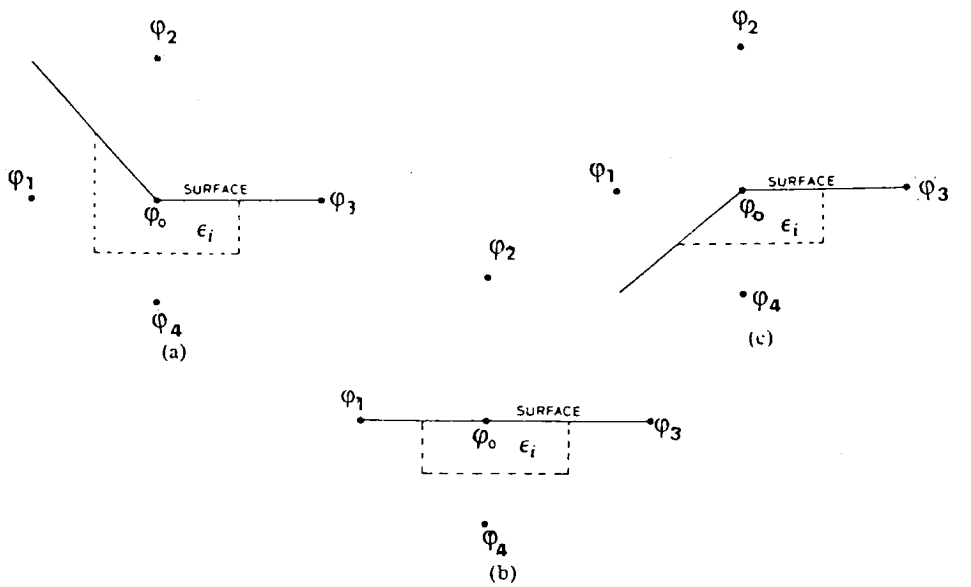


Fig. D-3 Three surface orientations which affect L_1 and A_1 in Eq. D-5 for a surface passing through the center node. (a) Surface is above S_1 and has no effect. (b) Surface passes through S_1 and eliminates its upper part. (c) Surface is below S_1 and eliminates it entirely.[1] (© IEEE 1980).

charge density Q_{tot} taken to be nonzero in the semiconductor beneath the surface. The coefficients L_m and A_m are as follows:

$$L_1 = \varepsilon_i (h_2 + h_4)/2, \qquad A_1 = 0$$
$$L_2 = \varepsilon_i (h_1 + h_3)/2, \qquad A_2 = 0$$
$$L_3 = (\varepsilon_i h_2 + \varepsilon_{si} h_4)/2, \qquad A_3 = h_3 h_4/8$$
$$L_4 = \varepsilon_i (h_1 + h_3)/2, \qquad A_4 = h_4 (h_1 + h_3)/8.$$

Table D-1 Discretization Coefficients for Figs. D-3 and D-4

	Surface Node (Figure D-3)		**Interface Node** (Figure D-4)	
	L_1	A_1	L_1	A_1
(a)	$\varepsilon_i (h_2 + h_4)/2$	$h_1(h_2 + h_4)/8$	$\varepsilon_s (h_2 + h_4)/2$	$h_1(h_2 + h_4)/8$
(b)	$\varepsilon_i h_4/2$	$h_1 h_4/8$	$(\varepsilon_i h_2 + \varepsilon_s h_4)/2$	$h_1 h_4/8$
(c)	0	0	$\varepsilon_i (h_1 + h_3)/2$	0

REFERENCES

1. J.A. Greenfield and R.W. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE Trans. Electron Dev.*, **ED-27**, p. 1520, Aug. 1980.

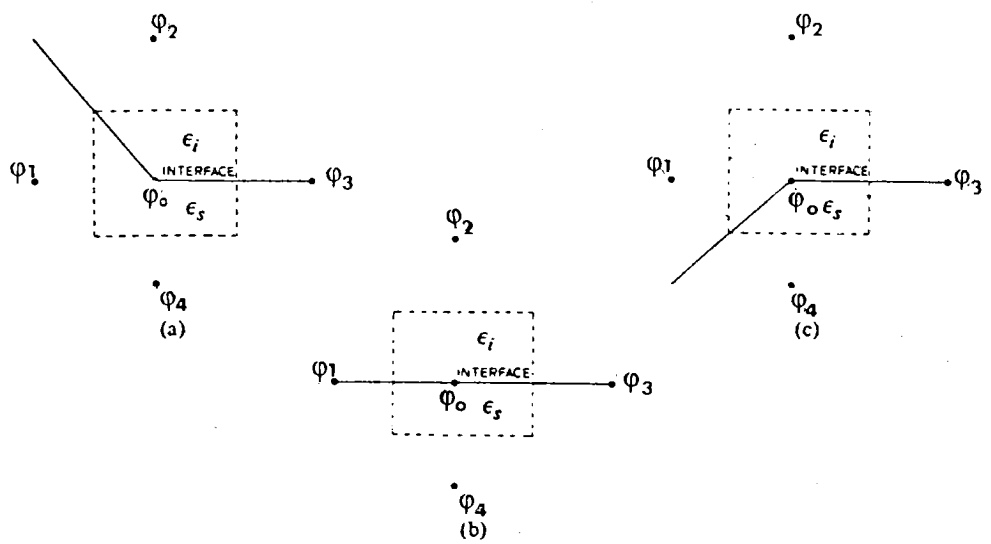


Fig. D-4 Three surface orientations which affect L_1 and A_1 in Eq. D-5 for a dielectric interface passing through the center node. (a) The interface is above S_1 and places S_1 in the semiconductor. (b) The surface passes through S_1 placing the upper portion in the insulator and the lower portion in the semiconductor. (c) The surface is below S_1 and places S_1 in the insulator.[1] (© IEEE 1980).

INDEX

NOTE: FOR MAJOR SUBJECT HEADINGS ALSO CHECK THE **TABLE OF CONTENTS**. THE LISTING ARE NOT ALWAYS DUPLICATED IN THE INDEX.

Abrupt junction approximation, 103
Accelerated testing of gate oxide reliability, 464
Accumulation in MOS devices, defined, 103
ADHOC, 585
Aging rate of MOSFET during transient stressing, 582-4
Analytic, defined, 16
Analytic solution, inadequacy of, 25-6
Antenna structures for measuring gate oxide damage due to plasma processing, 507-9
Asymmetric drive current effect in LDD MOSFETs, 604
Asymmetric MOSFETs: HS-GOLD, 636-7
Asymmetric parasitic field FETs in CMOS, 376-81, 385, 387, 391
  n^+-to-n-well FETs, 392-5
  p^+-to-p-substrate FETs, 395-7

Backward-difference approximation, 43
  in time discretization, 49
Backward-Euler method, 49
Barrier lowering analysis in MOSFET scaling, 284
Berkeley Reliability Tool (BERT), 492
Berkeley Short-Channel IGFET Model (see BSIM)
BF_2 implants, 306-8, 641
Bias-temperature stress test to monitor plasma damage, 509
"Birds-eye" plots, defined, 57
Bird's beak, defined, 337-8
  compaction schemes, 357-8
  factors which impact its length and shape, 340-4
    mask stack geometry factors, 342
  modeling its growth, 350-7
  bird's beak punchthrough (isolation lifting), 342-3
B-mode failures in gate oxides, (439,488, def.), 458, 468-82
  qualitative models of, 458-62
    how to reduce their density, 460-2
  quantitative models of, 467-82
Body effect, defined, 144
  effect of V_T-adjust implant on, 187
Body effect coefficient, 147
Boltzmann transport equation, 15
Boron penetration of gate oxides, 312-3, 641
Boron redistribution
  in channel stop implants, 335-7
  in PMOSFET channels, 306-7
  in PMOSFET source/drains, 307-8
Box integration finite-difference method, 39,40
Bridging oxygen atoms in SiO_2, 423-4
BSIM, 182, 271
  Model equations, 272-4
BSIM2, 271
BSIM3, 271
Built-in voltage, defined, 139
Bulk-charge model of MOSFET I-V behavior, 163, 165, 173-5, 178
Buried-channel MOSFETs, 292-7
Buried-LDD (B-LDD), 616-7
Burn-in, 482-91

CADDET, device simulator, 78
Capacitors for measuring gate oxide damage due to plasma processing, 508
  area intensive, 508, 509
  edge intensive, 508
Carrier injection into gate oxide, 429-38
  by channel hot electrons (CHE), 430-4
  by direct tunneling, 436-7
  by drain avalanche hot carriers (DAHC), 434-5
  by Fowler-Nordheim tunneling, 435-6

CAS, 584-5
Central-difference approximation, 43
Centroid, of V_T-adjust implant (defined), 189, 283
Channel hot electrons (CHE), 430-4
Channel-length modulation, 247, 260-1
Channel length modulation parameter, 182
Channel-stop doping in CMOS, 383
  dual channel stops, in twin-well CMOS 383
  in trench isolated CMOS, 410-13
Channel-stop implant, defined, 323, 335
  problems associated with in LOCOS, 335-7
  implant-free channel stop, 337
Channel voltage, defined, 139
Characteristic (effective) length of saturation region in MOSFET (defined), 268-9
Charge-sharing models, 211-3
Charge-sheet approximation, 112-3
Charge-sheet model, 97, 112-6, 163, 170-3, 178
Charge-to-breakdown, Q_{BD}, of gate oxides, defined, 441
  dependence on electron fluence, 453-7
Chemical-mechanical polishing, 373
Circuit aging simulator (CAS), 584-5
Circuit Oxide Reliability Simulator (CORS), 492
Classical semiconductor equations, 14-15
Co-implantation
  boron-chlorine, 336
  boron-germanium, 336-7
Complimentary error function (erfc), 20-3
Constant-current-V_T, 182, 196
Constant-field-gradient model, 264-5
Contour plots, defined, 58
Convergence, 54, 66
Corner parasitic MOSFET, 225
  in trench isolated CMOS, 369
CORS, 492
Crab-eye, in poly-buffered-LOCOS, 357
Crank-Nicholson time discretization method, 49-50, 72, 73
CREEP oxidation modeling program, 353-5
Cumulative Distribution Function, defined, 442
Cumulative Failure Percentage, defined, 442
Cumulative oxide defect density, defined, 448
Curve fitting, 6-7
Cutoff (defined), 175

Dangling bonds at Si-SiO_2 interface, 425-7
Deal Triangle, 121
Deep depletion, defined, 198 fn.
"Deglaze" nitride process, 345
Depletion approximation, 96, 103-4
Depletion in MOS-C, defined 94
  in a MOSFET, 158
Depletion-mode MOSFETs (defined), 137, 198, 199
Depletion-region, 87, 104
Device simulators (see Index under the following entries)
    CADDET
    GEMINI,
    MINIMOS
    PISCES-II
DIBL, see Drain-induced barrier lowering
Dielectric strength of gate oxides, measuring, 438
    ramped voltage test, 438-9
Diffusion , 11-12, 19-22
Direct tunneling in gate oxides, 436-7
Dirichlet boundary conditions, 63
Dispersion (log-normal plots), 445
Doping parameter, φ_B, defined, 99
Doping transformation procedure for calculating V_T, 186
Drain current in MOSFETs, defined, 144, 163-9
Drain-induced barrier lowering (DIBL), 213-222
  in parasitic field FETs, 383
  measuring DIBL, 218-9
DIBL rate parameter, *b*, and its inverse parameter, *R*, 302-4
Disposable spacers in LDDs
  oxide, 609
  poly, 608
  silicon nitride, 609-10
  TiN, 611
Dopant redistribution at well-borders in CMOS, 381
Double diffused drain (DDD) MOSFETs, 588-90
Double-implanted drain LDD (DI-LDD), 621-2
Drain avalanche hot electrons (DAHE), 435
Dual-doped poly CMOS, 312

EDTA, 459
Electrical tests for measuring oxide damage due to plasma processing, 509
Electron affinity, defined, 118
modified electron affinity, 430
Electrostatic potential, defined, 98
Energy-bands, 90-1
Energy-band diagrams, 90
of long-channel MOSFETs, 154-5
Enhancement-mode MOSFETs, 136, 137
Epitaxial CMOS, 404
Equilibrium, 14, 89, 90, 965
Erfc (complimentary error function), 20-3
Estimating gate oxide reliability using quantitative empirical model, 477-82
Euler's method, 31-2
Exposed dose, defined, 189, 282-3
Explicit time discretization, 47-8
Extrapolated threshold voltage, defined, 177, 180
Extreme (Weibull) Cumulative Distribution Function, 444

Failure rate, instantaneous, (443 def.), 472, 482
Failure Unit (FIT), (441-2 def.), 472
Fermi level, 14, 91
in heavily doped poly, 119
intrinsic, 14, 90, 98
quasi-Fermi level, 15
Fermi potential, 14
quasi-Fermi potential, 15, 166, 167
Fick's Second Law, 19-20
example of Numerical Analysis Procedure to Solve it, 70-4
simple, 20
Field-effect transistor, defined, 136
Field-induced junction, defined, 139
Field oxide, defined, 322-3
Field-oxide thinning effect, 339-40, 367
Field reduction factor in LDDs, 594-5, 612
Finite-difference method, 39-45
box integration, 39-40
difference operator, 39, 40-5
examples of solution, 59-74
Finite element method, 45-7
Fixed oxide charge, 121
Flat-band condition, defined, 91, 127
in three-terminal MOS structure, 139
Forward-difference approximation, 43
time discretization, 48
Forward Euler method, 48
Fowler-Nordheim tunneling, 435-6, 450, 453, 456, 462, 467, 477
during plasma processing, 507-9
Framed-mask poly-buffered LOCOS, (FMPBL), 362-4, 405
reverse-L-shaped FMPBL (RLS-PBL), 362-4

Galerkin matrix, 46-7
Galerkin method, 46-7
Gamma function, 447
Gate, defined, 85, 135
Gate bird's beak, 227
in LDDs, 602, 603, 634
in stacked gate oxide films, 502
Gate current, 136, 431-3
electron-temperature model of, 432-3
hot-hole gate current, 434
lucky electron model of, 432
related to oxide breakdown, 452-7
Gate-drain capacitance in overlapped LDDs, 605, 624, 632, 633-4
Gate-Drain Overlapped LDD (GOLD), 623, 625-6
Gate fringing field effect in LDDs, 634-6
Gate-induced drain leakage (GIDL), 198-200, 274
Gate-oxide characteristics, 422
Gate oxide damage by poly etching, 461-2
Gate-oxide damage during plasma processing, 504-14
Gate oxide thickness, evolution of, 447
Gate oxide thinning, 460
Gauss' Law, stated, 102-3
Gaussian function, 20-3
GEMINI, device simulator, 39, 40, 68, 78, 210, 243, 293, 301, 385
input file format, 69-70
GIDL (see gate-induced drain leakage)
Graded-buried LDDs, 617-8
Graded drains, defined, 587
why they are necessary to reduce hot-carrier degradation, 587
Graded-gate oxide (GGO), 227
Gradual Channel Approximation (GCA), 24, 89, 106, 148, 149-52, 161, 166
breakdown of, 161
in subthreshold, 150, 192
Grids, 34-8
evolution of, 35-6
local refinement of, 35
variable, 35

Halo Implants, 238, 240, 309
in LDDs, 621-2
in NMOSFETs, 238, 240
in PMOSFETs, 309-11
HCl, 500
High-dielectric spacer LDDs, 634-6
High-energy implants to form wells in CMOS, advanced well processes, 552-3
retrograde well CMOS, 539-46
Hi-wafer, 461
High-energy implantation,
to form channel-stop, 337, 344
for retrograde wells,
for PMOS punchthrough prevention, 309-11
High-pressure oxidation (HIPOX), 336, 340, 367
Hole fluence, 453
Hole generation coefficient, 462
Hole trapping in oxides, 451
Hot carriers, defined, 431-3
Hot-carrier degradation, characterizing
accelerated testing, 576-80
device parameters impacted by hot-carrier damage 573-4, 574-81
effect on circuit performance, 581
Hot-carrier degradation phenomena, 566-72
Si-H bond-breaking model, 569
Hot-hole and hot electron trapping model, 569
Hot-carrier effects in MOSFETs, defined, 559
(see also MOSFETs and LDDs)
Hot-carrier effects in PMOSFETs, 637-41
in PMOSFETs with p^+ poly, 641
models for damage in, 638
LDD structures in, 640
Hot-carrier lifetime, defined 576-7
dependence on I_{sub}, 577
under ac stressing, 580-1
Hot-carrier reliability simulation, 582-5
Hot electrons, defined, 431-3
Hot-electron induced punchthrough (HEIP), 638-9
Hot-electron injection into gate, 624-5, 633
Hot-hole gate current, 434
HOTRON, 584-5
HS-GOLD structures, 636-7
Hydrogen
in silicon nitride, 644-5
effect on hot electron degradation in MOSFETs, 643-7
high-temperature, effect of, 645-7
in NH_3 nitrided oxides, 649-54
reoxidation to remove, 649-51
passivation of dangling bonds, 125

ICGG method, 77
I_{Dsat} (defined),
Ihantola-Moll model, 174
Impact ionization which generates DAHC, 434
Implicit time discretization, 48-9
Incomplete Cholesky Conjugate method, 77
Infant mortality (failure rate curve), 443
Interface traps, 425
as a measure of oxide damage by hot-carrier injection, 477
passivation by hydrogen, 427
breaking of bonds by hot electrons, 428
physical origin of, 425
Interface states, 122-6
Interface trapped charge, 122-6
annealing out in hydrogen, 125-6
generation by breaking Si-H bonds, 569
Inverse-T LDDs (IT-LDDs) 623-5
modified IT-LDDs, 626-30
damaged re-ox layer IT-LDD, 628
FOLD, 626-7
L-shaped spacer ITLDD, 628
TOPS, 627
Inversion, defined, 94
in a MOSFET, 158
moderate inversion, defined, 116 fn
of trench sidewalls, 369, 373, 406-13
onset of inversion in parasitic field FETs in CMOS, 386, 387
strong inversion, defined, 94, 95
weak inversion, defined, 94, 100
onset of, 100
Inversion layer charge density Q_C, defined, 95, 100
Calculating, 103
in weak inversion, 114
in strong inversion, 115
Inversion layer thickness, calculating, 150-1
IRSM-CAS, 584, fn.
Isolation lifting (bird's-beak punch-through), 342
Isolation of CMOS, 373-413
establishing minimum n^+-to-p^+ spacing, 397-401

in NH_3 nitrided oxides, 649-54
reoxidation to remove, 649-51
passivation of dangling bonds, 125

ICGG method, 77
I_{Dsat} (defined),
Ihantola-Moll model, 174
Impact ionization which generates DAHC, 434
Implicit time discretization, 48-9
Incomplete Cholesky Conjugate method, 77
Infant mortality (failure rate curve), 443
Interface traps, 425
as a measure of oxide damage by hot-carrier injection, 477
passivation by hydrogen, 427
breaking of bonds by hot electrons, 428
physical origin of, 425
Interface states, 122-6
Interface trapped charge, 122-6
annealing out in hydrogen, 125-6
generation by breaking Si-H bonds, 569
Inverse-T LDDs (IT-LDDs) 623-5
modified IT-LDDs, 626-30
damaged re-ox layer IT-LDD, 628
FOLD, 626-7
L-shaped spacer ITLDD, 628
TOPS, 627
Inversion, defined, 94
in a MOSFET, 158
moderate inversion, defined, 116 fn
of trench sidewalls, 369, 373, 406-13
onset of inversion in parasitic field FETs in CMOS, 386, 387
strong inversion, defined, 94, 95
weak inversion, defined, 94, 100
onset of, 100
Inversion layer charge density Q_C, defined, 95, 100
Calculating, 103
in weak inversion, 114
in strong inversion, 115
Inversion layer thickness, calculating, 150-1
IRSM-CAS, 584, fn.
Isolation lifting (bird's-beak punch-through), 342
Isolation of CMOS, 373-413
establishing minimum n^+-to-p^+ spacing, 397-401
in n-well CMOS, 399-400
in p-well CMOS, 400-1
in twin-well CMOS, 400-1
in 0.35 and 0.25 μm CMOS, 401
experimental characterization of in CMOS, 387
improvement of with use of retrograde wells, 546-51
leakage currents in active devices and intrawell isolation loss, 376
leakage currents due to loss of isolation among opposite type FETs, 376-7
LOCOS-based, 405
trench isolation, 406-13
Iteration, solution method, 52, 53, 67
inner iteration, 67
outer iteration, 67

Jacobian matrix, 54, 66

Kinetic energy of electron, expressed in terms of an energy-band diagram, 90, 98
Kooi effect, 346-7, 460
in poly-buffered LOCOS, 357

Large-angle-tilt (LAT) implants, 310-11
of trench sidewalls in CMOS, 411-12
in PMOSFETs (LATIPS), 310-11
Large-angle-tilt implanted drain (LATID), 621, 630-4
gate overlap capacitance in, 633-4
self-aligned pocket implant, 621, 634
Lateral electric field,
models for,
in DDDs, 588-91
in LDD MOSFETs, 592-5
pseudo-2D model, 594-5
in non-LDD MOSFETs, 260-71, 561-2
Lifetime after burn-in, calculating, 485-6
Lifetime factors, 584
Lightly doped drain MOSFETs (LDDs), 591-637, 640-1
conventional LDDs
lateral electric field in, 591-5
models for $\mathcal{E}_{ymax}$ in LDDs, 592-5
pseudo-2-D model of, 594-5
fully overlapped LDD MOSFETs,
GOLD, 623, 625-6
IT-LDD, 626-30
high-dielectric spacer LDDs, 634-6

issues involved in CMOS, 606
large-angle-tilt implanted drains (LATID), 630-4
modified LDD MOSFETs,
buried LDDs, 616-7
double-implanted LDD, 621
graded-buried LDDs, 617
metal-coated LDD, 616
moderate-doped LDD, 614-16
profiled LDD, 617-21
performance limitations of, 598-608, 611-12
overcoming, 612-13
reduced drive current in 599-605
reduced hot-carrier reliability in, 606
sloped-junction LDD, 616
Linear regression, 6-7
Linearizing non-linear algebraic equations, 52-5
LOCOS (Local Oxidation of Silicon)
basic process, 330-47
fully recessed LOCOS, 331
semi-recessed LOCOS, 331
modeling, 349
empirical models of, 350
physical models of, 351-7
poly-buffered LOCOS (PBL), 357-67
Log-normal cumulative distribution function, 444, 447, 483
Long-channel MOSFETs, defined, 147-8, 152
comparison with short-channel MOSFETs, 207-8
Lucky electron model, 432

Magnetically-enhanced RIE (MERIE), 506
Mask stack, defined, 342
geometry and effects on bird's beak length, 342
stripping mask stacks, 345-6
Matrix solutions of partial-differential equations, 50-7
direct methods, 55-6
Cholesky's method,
Cramer's rule,
Crout's method
Gauss elimination method
Gauss-Jordan elimination method,
LU-decomposition method
Jacobian, 54
indirect methods, 55-6
Jacobi's method,
Gauss-Seidel method,
Successive overrelaxation method,
Alternating-direction implicit method
Maxwell-Boltzmann statistics, 90
Mean (in log-normal plots), 445
Mean-time-to-failure (MTTF), defined, 443
Metal-coated LDDs (MLD), 616
Micropores in thin oxides, 502-4
Microtrenches in gate oxide due to poly etch process, 461-2, 603
test structure to detect them, 462-3
MINIMOS, 39, 78, 161, 210, 215, 261, 385, 613
Mobile ionic charge, 120-1
Mobility, 248
surface, 248-9
Mobility degradation, 247, 248-51
Models, 1-27
empirical 5-9
physically based, 10-27
Moderate doped LDD (M-LDD), 614-6
Moderate inversion, 116 fn
Modified electron affinity, 430
Modified hole affinity, 430
MOS capacitor, defined, 85, 87
ideal, defined, 91
discussion of, 85-116
non-ideal defined, 117
discussion of, 117-8
why it is studied first, 88-9
MOSFET design strategy, 286-8
MOSFET operation,
linear region, 159
saturation region, 161, 174-5, 176, 260-74
subthreshold region, 175
MOSFET scaling, 274-88
constant electric-field scaling, 275-6
constant voltage scaling, 276
off-current scaling, 279-85
subthreshold scaling, 276-9
MOSFET SPICE models
Level-1 model, 178, 180, 271
Level-2 model, 179, 180, 271
Level-3 model, 178, 181, 271
Level-4 model, 178, 181, 271
Level-5, 271
BSIM, 182, 271
MOSFET technology evolution, 275, 421

Narrow-width effect on V_T, 222-5
due to encroachment of channel-stop dopants, 223-5
inverse narrow-width effect, 224-5
Natural-by-columns ordering, 51
Navier-Stokes equation, 352
Neumann-type boundary conditions, 63
Newton's direct method, 75, 76-7
Newton-Raphson method, 51, 52-3, 65
Nitride-clad poly-buffered LOCOS, 366-7
Nitrided oxides, 648-59
Reoxidized nitrided oxides, 649-54
Thermal nitirdation of SiO_2, 648-5
with NH_3, 648-54
with N_2O, 654-59

Off-axis implantation, effect on weak overlap in LDDs, 604-5, 606, 612
Oxidation, simulation of,
two-dimensional, empirical, 350
two-dimensional, physical
CREEP, 353-5
SOAP, 352
SUPREM IV, 349, 352, 353
three-dimensional
SMART-P 355-7
Oxides
breakdown models of, 448-482
qualitative models of intrinsic breakdown, 449-58
hole trapping model, 450-6
electron lattice damage model, 456-8
qualitative model of B-mode defect breakdown, 456-62
quantitative models of intrinsic oxide breakdown, 462-7
quantitative model of B-mode failures, 467-82
oxide thinning model, 468-82
reduced barrier height model, 467
damage from plasma processing 504-14
models for damage-causing mechanisms, 511-2
procedures for reducing, 514
lattice structure of, 423-5
material properties of, 423-5
measuring thin oxides, 502
models of thin oxide growth, 492-4
Reisman-Nicollian model, 494
native oxides, 496
stacked gate oxides, 502-4
technology of thin oxide growth,
control of growth rates, 496
HCl use in, 500
particulates in, 496
post-oxidation anneal, 501
vertical furnaces, 495
Oxide charges, 119-27
Fixed oxide charge, 121
Interface trapped charge, 122-6
annealing out in hydrogen, 125-6
generation by breaking Si-H bonds
Mobile ionic charges, 120-1
Oxide trapped charge, 126, 452
Oxide growth,
field oxide, 335-7
gate oxide,
Oxide spacers, (see Spacers)
Oxide thinning in field regions of sub-micron LOCOS structures, 339-40, 267
Oxynitride films, 341

p^+ poly-gate PMOS, 311-3,398, 641
Pad oxide, 332-3, 340, 357
Pao-Sah model, 97, 163, 166-70, 177, 178
Parasitic MOSFETs
in field regions of MOSFETs, 322-4
asymmetric in CMOS, defined, 376-8
effect of interdiffusion of dopants at well borders on V_{TFiso}, 381-4
lateral punchthrough in, 380
short-channel V_T roll-off in, 380, 393
vertical punchthrough in, 380
symmetric, 325-30
PELOX (poly-encapsulated LOCOS), 344, 365-6
nitride-clad PBL, 366-7
Perspective plots, defined, 58
Phosphoric acid, 34
Phosphorus drain, 588
PIF, 59
PISCES II, 37, 39, 40, 68, 78, 79, 80, 210, 215, 261, 385, 635
Planarizing semi-recessed LOCOS, 348-9
trench isolation structures, 369, 372
using CMP, 373
using SOG, 373
Plug implant CMOS isolation, 384, 405
PMOS devices, 288-313
buried channel devices, 292-7
LDD structures in, 308-9, 640-1
punchthrough susceptibility of, 300-4
analytical model of, 302-4

shallow junctions in, 307-8
subthreshold swing in, 297-300
suppressing short-channel effects in, 305-13
threshold adjust in, 289-92
V_T-roll-off in, 294-6
with implanted punchthrough stopper (LATIPS), 310-11
with n^+ poly gates, 289-91
with p^+ poly gates, 311-3
boron penetration through gate oxide, 398-9
counterdoping of NMOS devices, 169, 397
hot-electron degradation in, 641-2
Poisson distribution, 447, 470
Poisson equation, 13, 23-4, 27
2-D form, 27
example of solving in 2-D form using numerical analysis, 59-70
in MOS-C, 88-9
solution of in 1-D form (MOS-C) when invoking the depletion approximation, 104-5
solution of in 1-D form (MOS-C) without invoking the depletion approximation
general solution, 106-12
charge-sheet solution, 112-6
Poly-buffered LOCOS (PBL), 357-67
conventional, 357-60
voids in poly layer, 359-61
microfoot in mask stack resist, 361
modified PBL, 361-7
FMPBL, 362-4
Nitride-clad PBL, 366-7
PELOX, 344, 365-6
RLS-PBL 362-4
Polysilicon depletion, 641
Post-oxidation anneal, (POA), 501
Potential, in semiconductor, 98-9
Potential diagram defined, 99
PREDICT, process simulator, 9-10, 494
Probability density function, 442-3
Profile interchange format (PIF), 59
Profiled LDDs (P-LDD), 617-21
Profiled wells in CMOS, 552-3
PRORPS, 373
Pseudo-2D model of MOSFET
to model DIBL, 214-18
to model $\mathcal{E}_{ymax}$ in LDDs, 594--5
to model saturation, 265-71
Punchthrough, 195, 232-46
experimentally characterizing, 240-2
modeling punchthrough, 242-6
in parasitic field FETs in CMOS
lateral, 378-80, 384, 387
vertical, 378-80
in PMOSFETs, 300-4
Punchthrough prevention implant, 238-9
large-angle-tilt-implant, 310-11
Punchthrough voltage, 195, 235 (def.), 242

Quadruple-well CMOS, 386, 545, 550-1

Ramp voltage breakdown test for oxide reliability, 438-9, 475
RCA clean, 459
Reddi-Sah model, 182, 262-4
Reliability, defined, 441
Re-oxidation of poly, 227
in LDDs, 595, 602-3
in PELOX, 366
Reoxided nirided oxides, 649-54
ROXNOX, 649-54
Resistance, parasitic series
(see Parasitic resistance)
Reverse short-channel effects on V_T, 226-32

Sacrificial-gate oxide, 347-8
Saddle point of potential in MOSFET channel, 235, 301
Saturation region of operation
in MOSFETs,
long-channel, 161, 174-5, 176
short-channel, 260-71
modeling in short-channel MOSFETs, 260-74
Saturation velocity, 248-51
SCALE (Substrate Current and Lifetime Evaluator), 582
Screening tests for gate oxide reliability (burn in), 446, 482-91
Self-aligned pocket implant, 238, 240, 634
Semi-recessed LOCOS (see LOCOS)
Series resistance of lightly doped drain region of LDD, 595, 599-602
how to decrease by increasing LDD implant dose, 614-16
Shallow junctions in PMOSFETs, 307-8
Shallow trench isolation structures, 367
BOX-1 (buried oxide), 368
BOX-2 (buried-oxide-2), 368-71
problems with, 369-71

tapered trenches, 372
Short-channel effects in MOSFETs:
on threshold voltage, 208-22
barrier lowering models, 213-222
charge-sharing models, 211-3
dependence of V_T on V_{BS} in short-channel MOSFETs, 221-2
narrow-gate-width effects, 222-24
due to channel-stop dopant encroachment, 223-4
inverse narrow width effect, 224-5
reverse-short-channel effects on V_T, 226-32
on operation in saturation, 260-71
Short-channel MOSFET models, simple, analytical model, 254-60
Sidewall of trench inversion, 369, 373, 406-13
Si-SiO_2 interface, 425-9
Silicon nitride
disposable spacers, 609-11
etching phosphoric acid, 345
masking properties of, 333-4
mechanical properties of, 342
permanent spacers in LDDs, 634-6
SILO (sealed interface LOCOS), 355, 357
Sloped junction LDD (SJLDD), 616
SMART-P oxide simulator, 355-7
Snapback breakdown, 563
SOAP (Stanford Oxidation Program), 352
Sodium contamination in gate oxides, 459
Space discretization of partial-differential equations, 38-47
2-D, 43-4
finite difference method, 40-5
finite element method, 45-7
sparse matrix, 50, 64
Spacers in LDDs,
disposable, 608-11
fabrication of, 597-8
high dielectric spacers, 634-6
Spectrophotometers, UV-reflection, for thin-oxide thickness measurements, 502
Speed degradation factor, 585
SPICE, circuit simulation program, 18, 178, 180-1, 251, 271
Level-1 MOS model, 178, 180, 271
Level-2 MOS model, 179, 180, 271
Level-3 MOS model, 181, 271
Level-4 MOS model, 181
Spin-on glass (SOG) for trench filling, 373
Square-law MOSFET I-V model, 163, 165, 175-7, 178
Stacked gate oxides, 502-4
Static I-V MOSFET characteristics, defined, 153
quantitative derivation of, 162-82
Strong inversion, 94, 100
onset of, defined, 100, 117
beyond onset, 104, 105
MOSFET I-V models valid in, 173-7
Structure dependent degradation effect in conventional LDDs, 607
Substrate current, I_{sub}, 434, 563-6, 570, 576-7, 612
model of, dependent on terminal voltages, 565
Subthreshold current
"hump" (or "kink") in, due to corner parasitic MOSFET, 369
Subthreshold regime, defined, 175
Subthreshold drain currents in MOSFETs,
in long-channel MOSFETs, 191-200
in short-channel MOSFETs,
due to DIBL, 213-22
due to punchthrough, 323-46
Subthreshold swing, S_t, 196-7, 232-3
in buried-channel PMOSFETs, 297-300
in parasitic field-region MOSFETs, 324-5
modeled in BSIM, 274
SUPREM II, 72, 73, 74
SUPREM III, 71, 74
SUPREM IV, 39, 349, 352, 353, 385, 635
Surface-channel PMOSFETs, 292-3, 311-3
Surface potential, defined, 96, 99
Surface roughness, as gate oxide defect, 459
Surface states, defined, 425

Tapered shallow trench isolation structures, 372

Ta_2O_5 spacers in LDD, 635-6
Taylor Series, 40,-2, 52
TCA, 50t-DCE, 500
TEOS (tetra-ethyl-ortho-silicate) trench filling, 372, 412
spacers in LDDs, 595-6
Test structures for characterizing,
isolation leakage in MOSFETs, 326-9
isolation leakage in CMOS, full-CMOS
oxide damage due to plasma processing, 507
T-gate LDD, 634
Threshold-adjust implants, 185-88
in PMOSFETs, 289-92

Threshold voltage, V_T, in active devices
adjusting, 182-8
body effect on, 144
doping transformation procedure for calculating V_T, 186
in ideal MOS-Cs, 116-7
in non-ideal MOSFETs, 127-9
in MOSFETs with non-uniformly doped substrates, 186
in n^+ poly PMOSFETs, 289-90
in p^+ poly PMOSFETs, 289-90, 311-12
measuring V_T, 176, 182
short-channel effects on V_T, V_T-roll-off, 219-22
narrow width effects, 222-5
reverse short-channel effects, 226-32
variation with body biasing, 183
variation with substrate doping, 183-4
V_T equation in BSIM, 272
Threshold voltage in parasitic field FETs, V_{TF}, 323
nominal V_{TF}, defined, 324
in CMOS, 384-9
Time-dependent dielectric breakdown, TDDB of gate oxides, 440-1, 448, 469, 475
interface trap generation test, 477
ramped-voltage breakdown test, 438-9, 475
under constant voltage stressing, 440
under constant current stressing, 440-1
Time-to-breakdown, t_{BD}, of gate oxides, defined, 440
model to predict, 463
Time-to-failure due to plasma damage, 509
Transconductance parameter, k, defined, 176
TR-BDF2 method, 50
Trivalent silicon, 425
Tunneling,
direct, 436-7
Fowler-Nordheim, 435-6
Twin-well CMOS (see CMOS)

Variable depletion-charge MOSFET I-V Model, 163, 165, 173-5
Velocity-saturation effects, 247, 251-4
Velocity saturation region, VSR, defined, 264
Vertical Punchthrough in CMOS, 380, 401-4
Very lightly nitrided oxides, 651-9
Void formation during trench/gap filling, 369
Voltage doping transformation method for modeling punchthrough, 246

Wafer charging during plasma processing, 504-7
causes of 506-7
time to charge, 505
well engineering, defined, 552
Weak inversion, 94, 100, 112, 113
onset of weak inversion, defined, 100
approximation of Q_I in, 114-5
Weak overlap in LDDs, 602-5
due to off-axis implantation of LDD, 604-5
due to poly re-ox, 603
Wearout, defined, 443
Weibull (or Extreme) Cumulative Distribution Function, 444-5, 447
Well formation in CMOS, 523-58
after-gate implantation, 554-5
future trends in, 552-3
in epitaxial CMOS, 528-30
n-well, 527-8
p-well, 524-7
retrograde wells, 539-46
technology of well formation, 532-9
aligning features to wells, 532-5
channel stop formation, 535-7
split well drive-in process, 538-9
twin-well, 530-1
White ribbon (or Kooi) effect, 346-7
in poly-buffered LOCOS, 357
Work-function difference, 117-9

ORDER FORM

LATTICE PRESS
POST OFFICE BOX 340-W
SUNSET BEACH, CA 90742, U.S.A.

Please send the number of copies indicated below of **SILICON PROCESSING FOR THE VLSI ERA - Vols. 1, 2, and/or 3 (please specify!).** A check or money order for the full amount for the number of copies ordered is enclosed. Make checks payable to: **LATTICE PRESS.** I understand that I may return any book within 30 days for a full refund if not completely satisfied.

_____
NAME (and COMPANY - if part of your address)

_____
ADDRESS

_____
CITY

_____
STATE, ZIP CODE COUNTRY

_____
Daytime Phone Number

____ copies **Vol. 3** @ $ 94.95 each, plus $4.00 for each copy, shipping-handling fee.
(*California Residents*: Please add $5.80 per copy for Sales Tax)

____ copies **Vol. 2** @ $ 89.95 each, plus $4.00 for each copy, shipping-handling fee.
(*California Residents*: Please add $5.40 per copy for Sales Tax)

____ copies **Vol. 1** @ $59.95 each. plus $4.00 for each copy, shipping-handling fee.
(*California Residents*: Please add $3.60 per copy for Sales Tax.)

Allow 30 days for delivery

International Ordering Information:

Overseas Prices: Volume 1 - $69.95; Volume 2 - $99.95; Volume 3 - $99.95. Enclose full payment with order in Travelers Checks (US dollars), International Money Order, or by a check payable to **Lattice Press** in **US dollars, drawn on a U.S. bank.** Include a $4.00 shipping /handling fee per copy for overseas orders shipped via surface mail. (Allow 6-10 weeks delivery.) Include a $20 US shipping /handling fee per copy for overseas orders shipped by airmail. (Allow 2 weeks delivery).

ORDER FORM

LATTICE PRESS
POST OFFICE BOX 340
SUNSET BEACH, CA 90742, U.S.A.

Please send the number of copies indicated below of **SILICON PROCESSING FOR THE VLSI ERA - Vols. 1, 2, and/or 3 (please specify!).** A check or money order for the full amount for the number of copies ordered is enclosed. Make checks payable to: **LATTICE PRESS.** I understand that I may return any book within 30 days for a full refund if not completely satisfied.

_____
NAME (and COMPANY - if part of your address)

_____
ADDRESS

_____
CITY

_____
STATE, ZIP CODE COUNTRY

_____
Daytime Phone Number

____ copies **Vol. 3** @ $ 94.95 each, plus $4.00 for each copy, shipping-handling fee.

(*California Residents*: Please add $5.80 per copy for Sales Tax)

____ copies **Vol. 2** @ $ 89.95 each, plus $4.00 for each copy, shipping-handling fee.

(*California Residents*: Please add $5.40 per copy for Sales Tax)

____ copies **Vol. 1** @ $59.95 each. plus $4.00 for each copy, shipping-handling fee.

(*California Residents*: Please add $3.60 per copy for Sales Tax.)

Allow 30 days for delivery

International Ordering Information:

Overseas Prices: Volume 1 - $69.95; Volume 2 - $99.95; Volume 3 - $99.95. Enclose full payment with order in Travelers Checks (US dollars), International Money Order, or by a check payable to **Lattice Press** in **US dollars drawn, on a U.S. bank.** Include a $4.00 shipping /handling fee per copy for overseas orders shipped via surface mail. (Allow 6-10 weeks delivery.) Include a $20 US shipping /handling fee per copy for overseas orders shipped by airmail. (Allow 2 weeks delivery).

THE COMPANION VOLUMES TO THIS BOOK

Silicon Processing

for the VLSI Era

Volume 1 - Process Technology

By STANLEY WOLF and RICHARD N. TAUBER

TABLE of CONTENTS - Ch.1 Silicon: Single-Crystal Growth and Wafering; **Ch. 2** Crystalline Defects, Thermal Processing, and Gettering; **Ch. 3** Vacuum Technology for VLSI Applications; **Ch. 4** Basics of Thin Films; **Ch. 5** Silicon Epitaxial Film Growth; **Ch. 6** Chemical Vapor Deposition of Amorphous and Polycrystalline Films; **Ch. 7** Thermal Oxidation of Silicon **Ch. 8** Diffusion in Silicon; **Ch. 9** Ion Implantation for VLSI; **Ch. 10** Aluminum Thin Films and Physical Vapor Deposition in VLSI; **Ch. 11** Refractory Metals and Their Silicides in VLSI; **Ch. 12** Lithography I: Optical Photoresists - Material Properties and Process Technology; **Ch. 13** Lithography II: Optical Aligners and Photomasks; **Ch. 14** Advanced Lithography: E-Beam and X-Ray; **Ch. 15** Wet Processing: Cleaning, Etching, and Lift-Off; **Ch. 16** Dry Etching for VLSI; **Ch. 17** Material Characterization Techniques for VLSI Fabrication; **Ch. 18** Structured Approach to Design of Experiments for Process Optimization.

1986 LATTICE PRESS 660 pp. ISBN 0-9616721-3-7

Volume 2 - Process Integration

By STANLEY WOLF

TABLE of CONTENTS - Ch. 1 Process Integration for VLSI and ULSI; **Ch. 2** Isolation Technologies for Integrated Circuits; **Ch. 3** Contact Technology and Local Interconnects for VLSI; **Ch. 4**; Multilevel Interconnect Technology for VLSI and ULSI; **Ch. 5** NMOS Devices and Process Integration; **Ch. 6** CMOS Process Integration; **Ch. 7** Bipolar and BiCMOS Process Integration; **Ch. 8** Semiconductor Memory Process Integration; **Ch. 9** Process Simulation.

1990 LATTICE PRESS 752 pp. ISBN 0-9616721-4-5

Order by using form on last page of book

ISBN 0-9616721-5-3