**79.** Courtesy of KLA-Tencor Inc.
**80.** *ibid., Ref. 55.*
**81.** *ibid., Ref. 39.*
**82.** W. O'Mara, "The Inevitable Shakeout in the CMP Equipment Market," *Solid State Technology*, May 1998, p. 84.
**83.** F. Kaufman *et al.,* "Chemical Mechanical Polishing for Fabricating Patterned W Metal Features as Chip Interconnects," *J. Electrochemical Society,* Vol. **138**, November 1993, p. 3460.
**84.** K. Wijekoon *et al.,* "Tungsten CMP Process Developed," *Solid State Technology,* April 1998, p. 53.
**85.** I. Kim *et al.,* "Optimized Process Developed for Tungsten CMP," *Semiconductor International,* November 1996, p. 119.
**86.** S. Roy, "Advanced Metal & Intercon. Syst. for ULSI Applications," 1995, p. 733.
**87.** *ibid. Ref. 29.*
**88.** J.-F. Wang "CMP of Cu Multilevel Interconnects," *Procs. VMIC* 1995, p. 505.
**89.** R. DeJule, "CMP Challenges Below Quarter Micron," *Semiconductor International,* Nov. 1997, p. 54.
**90.** D.R. Evans, "Electrochemical Interaction Between Copper and Barrier Materials During CMP," *Mtg. Abs. Electrochem. Soc. Mtg.,* Fall 1996, Abs. No. 504,
**91.** A. Beverina *et al.,* "Copper Photocrrosion Phenomenon during Post-CMP Cleaning," *Electrochem. Soc. Letts.,* March 2000, p. 156.
**92.** A.R. Seturaman *et al.,* "Copper vs Aluminum: A Planarization Perspective," *Semiconductor International,* June 1996, p. 177.
**93.** J. Nuetzel *et al.,* "Conventional and Slurry Free CMP of Polymer Layers for Gbit DRAM," *Mtg. Abs. Electrochem. Soc.,* Fall 1999, Abs. 1232.
**94.** W.-T. Tseng, "CMP of Fluorinated Silicon Dioxide: Is it Necessary and Feasible?" *Solid State Technology*, February, 1997, p. 61.
**95.** *ibid. Ref. 36.*
**100.** S.J. Fang *et al.,* "Optimization of the CMP Process for Pre-Metal Dielectrics," *J. Electrochem. Soc.,* February 2000, p. 682.
**101.** T. Marbeiter, *et al.,* "An Update: Transition to 300-mm CMP," *Semiconductor International,* November 1998, p. 76.
**102.** *ibid. Ref. 101.*
**103.** M. Nandakumar *et al.,* "Shallow Trench Isolation for Advanced ULSI CMOS Technologies," *Tech. Dig. IEDM,* 1998, p. 134.
**104.** L. Peters, "Choices and Challenges for Shallow Trench Isolation." *Semiconductor International,* April 1999, p. 71.
**105.** G. Badenes *et al.,* "A New Dummy-Free Shallow Trench Isolation Concept for Mixed-Signal Applications," *J. Electrochem. Soc.,* **147**, Oct. 2000, p. 3827.

**TSMC-1034**
**U.S. Patent 8,587,076**

# Chapter 9

# SHALLOW TRENCH ISOLATION (STI)

Despite advances made to decrease bird's beak and channel-stop encroachment in LOCOS isolation (by poly-buffered LOCOS and other approaches), such techniques eventually become inadequate for deep-submicron IC technologies. In particular, the scalability of LOCOS for sub-half-micron CMOS technologies comes into question. The limitations of LOCOS include: the penetration of the LOCOS bird's beak into the device active regions; non-planarity; LOCOS field-oxide-thinning at narrow isolation spacings; and stress-induced silicon defects. (LOCOS-based isolation technologies are described in more detail in Vols. 1 [2nd Ed], 2, and 3 of *Silicon Processing for the VLSI Era.*) The successor to LOCOS in CMOS is *shallow trench isolation* (STI). In STI a relatively shallow trench is first etched into the Si substrate (0.3-0.5-$\mu$m deep), which is then refilled with an insulator material (following a short thermal-oxidation growth step that forms a thin film of $SiO_2$ on the trench walls to control the Si-$SiO_2$ interface quality). The surface is planarized by CMP after refill to complete the isolation structure. Figure 9-1a depicts a shallow-trench isolation (STI) structure in CMOS.

Since the bird's beak is entirely eliminated in STI and the redistribution of channel-stop-dopant is reduced, smaller isolation spacing is also possible than with LOCOS. In addition, the field oxide in STI is fully recessed, offering the potential of a completely planar surface after the isolation structure formation. (Note that while CMP can also be used to reduce topography resulting from the LOCOS process, the oxide encroachment with LOCOS is not eliminated by such CMP.) 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. 9, *Silicon Processing for the VLSI Era*, Vol. 3). Also, MOSFETs made with fully-recessed oxide isolation exhibit larger $g_m$ and smaller $S_t$ values than those using semi-recessed LOCOS. Finally, the fully-recessed STI structures have trenches whose lower corners, though rounded, have

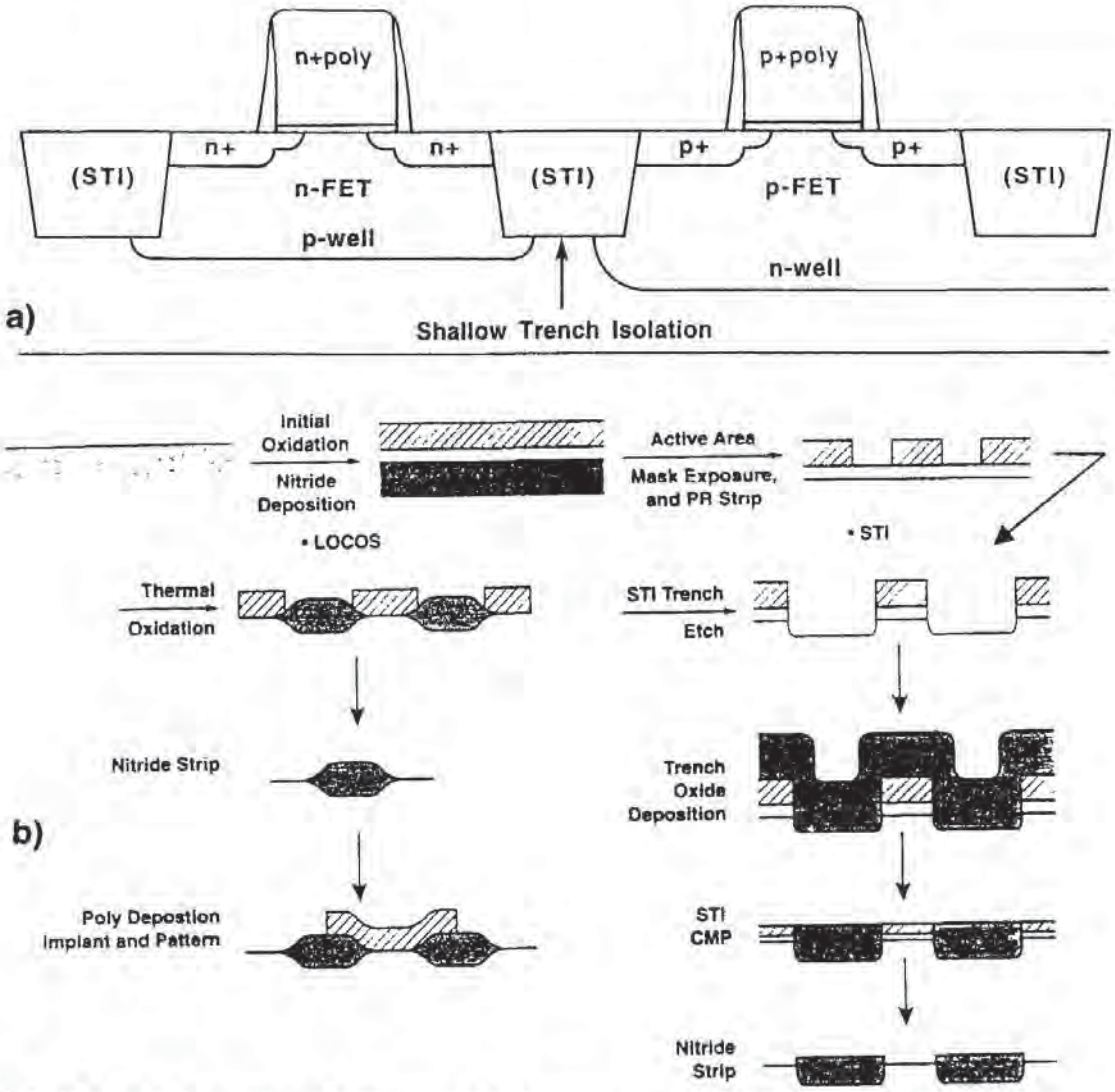



**Fig. 9-1** a) Schematic of STI in a CMOS structure. b) LOCOS vs. STI process flow, indicating that STI requires more steps.

a smaller curvature than exists with semi-recessed LOCOS field oxides. These trench corners provide improved isolation efficiency due to the so-called "corner effect."[1] 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 voltage 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 the fabrication and process integration of STI structures arise. For one thing, the STI process involves a larger number of steps than does LOCOS isolation process (Fig. 9-1b). Yet, the overall benefits of STI over LOCOS provided the impetus for the industry to find ways to

overcome these challenges. STI began to find use in mainstream IC technology when it was integrated into 166 MHz Pentium microprocessors (which used a 0.35-$\mu$m process) and the PowerPC 603e chip (which was fabricated with a 0.45-$\mu$m process). In this chapter we will describe the details of the STI process fabrication sequence, the process integration challenges that have been overcome, and those that it faces going forward.

## 9.1 EARLY SHALLOW-TRENCH-ISOLATION STRUCTURES

The earliest example of shallow-trench-and-refill is the so-called *buried-oxide* (or BOX) isolation technology. The original BOX technique (BOX-I) was introduced in 1981,[2] and an improved version was presented in 1983 (BOX-II).[3] In the latter, shallow trenches (0.5-0.8 $\mu$m deep) were etched with an orientation-dependent silicon-etch to provide an inclination of the trench sidewalls of ~60°. Next a channel-stop-implant step was performed. (The sloped sidewalls of the trenches ensure that they receive sufficient doping during the channel-stop implant step to prevent weak inversion of the silicon-sidewall surface. Thus, subthreshold leakage from source-to-drain along this parasitic conduction path is suppressed.) A thin thermal oxide (20-50-nm-thick) was then grown to passivate the trench sidewalls. Next, CVD oxide was deposited to fill the trenches. This layer was etched back, leaving oxide only in the recesses. Ideally, the top of the oxide 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-$\mu$m-deep trenches that are narrower than about 2-$\mu$m.

**2.** Inversion of the silicon at the sidewalls of $p$-type active regions could still occur, especially if vertical trench sidewalls were used to reduce the isolation spacing. This problem was made worse by boron segregation into the thin, sidewall-passivating thermal oxide. This effect lowers the boron concentration at the sidewall surface, as will be discussed in following section.

**3.** Despite complex planarization schemes, the early BOX processes were unable to yield adequately planar surfaces. (In particular, global planarity was poor.)

4. The BOX isolation structures could negatively impact the device operating behavior.[4,5,6,7] 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. 9-2a. (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 behind Fig. 9-2, which lies in the plane of the paper.) For this particular configuration of the polysilicon gate, the portion of the gate poly 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. 9-2 and along the axis of the device, so that it includes the source and drain regions.

In the ideal case of a shallow-trench-isolated MOSFET, the gate poly runs flat across the isolation, as shown in Fig. 9-2b. Consequently, the part of the gate poly over the field region creates more of a two-dimensional fringing electric-field on the corner and sidewall of the silicon mesa than in the LOCOS case of Fig. 9-2a. This establishes a *corner parasitic-MOSFET* in parallel with the main device.[8] The parasitic device turns-on at gate voltages lower than those required by the main channel (because the same voltage value occurring at the corner produces a larger electric field at that point, inducing a higher concentration of

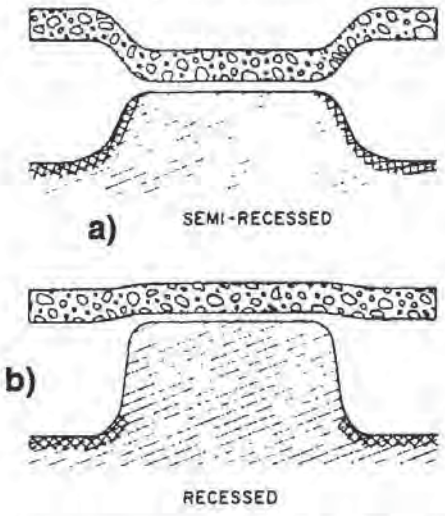



**Fig. 9-2** a) In LOCOS isolation structures, the field oxide rises above the substrate. b) In trench isolation structures, the field oxide is flat with respect to the substrate (unless the trench oxide is overetched; in that case, a downward step exists at the edge of the active regions – see, for example Fig. 9-3).

inversion-layer charge than the same voltage induces in the nominal channel region). This causes a "hump" or "kink" in the drain-current versus gate-voltage subthreshold curves (see Fig. 9-3). 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 top surface of the silicon mesa (i.e., the surface of the active region of a MOSFET IC). In that case, the polysilicon gate wraps around the corner of the mesa as shown in Fig. 9-3. 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 a 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. This exacerbates the parasitic sidewall conduction described above and also enhances the inverse narrow-width effect (see Chap. 6, *Silicon Processing for the VLSI Era* - Vol. 3).

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 continued in an effort to overcome the above problems. Several enabling technologies were developed to overcome the drawbacks of the early shallow-trench isolation processes. These included: *dry-etch processes* that allowed trenches with controlled sidewall slopes to be produced; *deposition techniques* that permitted the trenches to be filled without the formation of voids; and *chemical-mechanical polishing* (CMP) methods that could provide global planarization of the wafer surface at the conclusion of the STI process. We will describe these technologies in more detail in the next section, emphasizing how their use enabled the implementation of STI. Figure 9-4 shows an SEM of an STI structure.

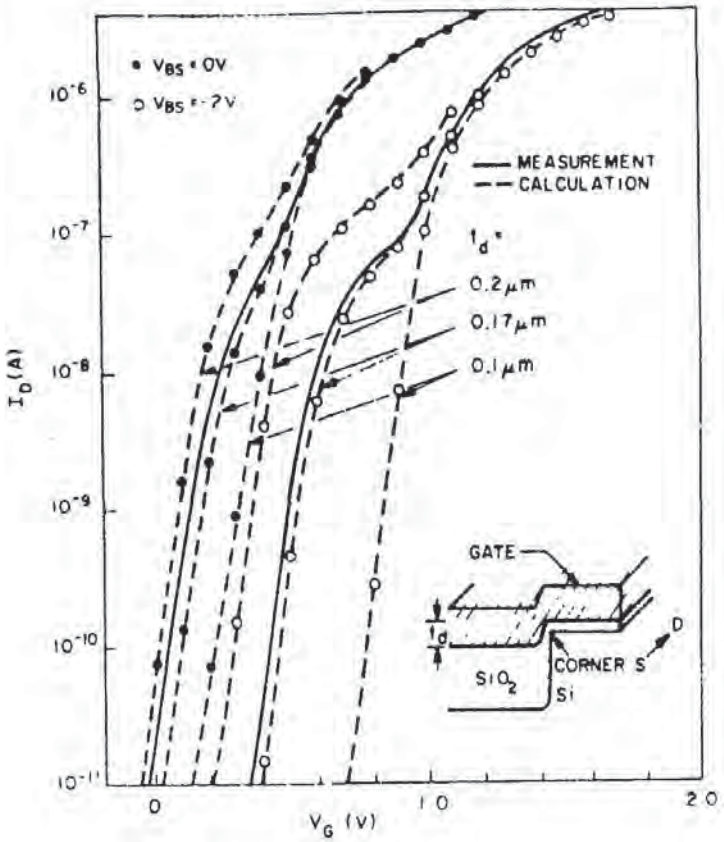



**Fig. 9-3** Measured and simulated drain current versus gate potential for devices having a downward step in the field oxide ($t_d$).[9] (© 1981 IEEE)

## 9.2 STI-ENABLING PROCESSES

Progress in overcoming the four major problems of early-STI isolation structures is summarized as follows:

**1.** The problem of void formation when refilling narrow trenches with CVD-oxide was ameliorated in two ways: **1**) dry-etch processes which could etch the trenches with a controlled slope (and rounded-bottom trench corners) were developed to replace the wet-etch techniques used to etch the trenches in early-STI structures; and **2**) various so-called "gap-filling" CVD techniques were developed that could fill the trenches with CVD-oxides without creation of voids. Such gap-filling-CVD processes include: *LPCVD TEOS*; *atmospheric-pressure CVD of ozone-TEOS*; and *high-density-plasma CVD* (*HDP-CVD*). Note that the 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.** The introduction of *chemical-mechanical polishing* (*CMP*) allowed the

oxide that is deposited into the trench to be polished back, leaving a planar surface over the wafer. The CMP process was developed in the late 1980s at IBM, and was introduced to the rest of the IC industry in the early 1990's. Details of the CMP process are provided in Chap. 8.

**3.** As described above, the subthreshold characteristics of narrow MOSFETs (e.g., $\leq$ 0.5-$\mu$m-wide) isolated by refilled shallow-trench structures are sensitive to the shape of the upper corners of the trenches.[10] That is, 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 Sect. 5.2.2 in *Silicon Processing for the VLSI Era*, Vol. 3). 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. 9-3. 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 surface 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.

## 9.3 SHALLOW TRENCH ISOLATION FOR CMOS

Trench isolation has also been explored for twin-tub CMOS as a means of implementing isolation structures, primarily as a means for achieving smaller interwell ($n^+$-to-$p^+$) isolation spacing. (Other advantages of trench isolation over LOCOS were described earlier.) A trench at the $n$-well-to-$p$-well interface prevents dopant interdiffusion where the two wells join. Thus, dopant compensa-

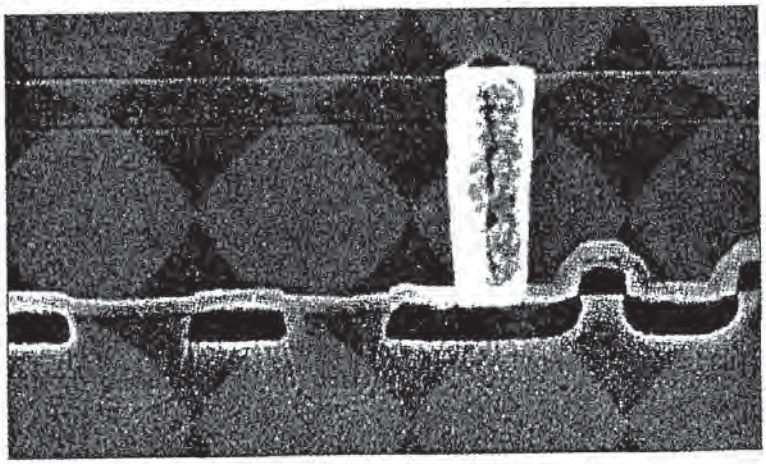


**Fig. 9-4** SEM of a shallow-trench isolation structure.

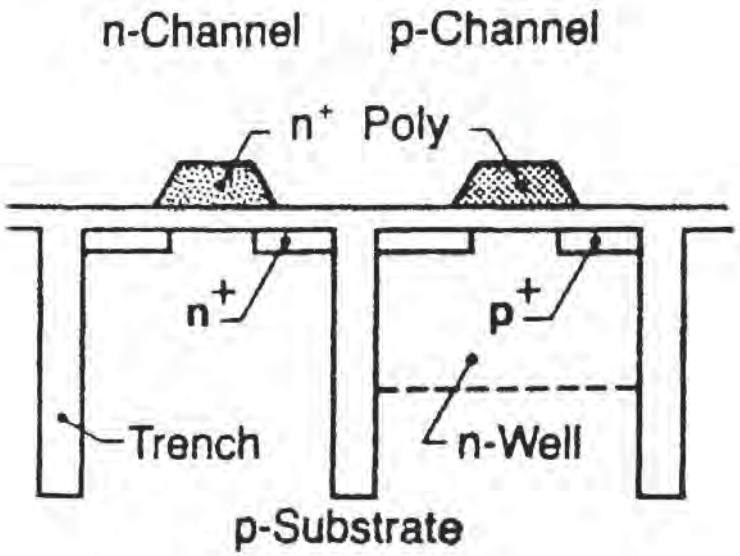



**Fig. 9-5** Schematic of an *n*-well CMOS with deep trench isolation.[11] (© 1986 IEEE).

tion (which lowers the surface dopant concentration at the well borders) is eliminated. This permits larger parasitic-MOSFET-threshold-voltage values ($V_{TFiso}$) to be maintained at smaller isolation spacings. Shallow and deep trenches of submicron widths have been used to reduce $p^+$-to-$n^+$ spacings.

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. 9-5). 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 was not widely implemented for such applications. Shallow-trench-isolation structures, however, became the standard isolation structure for CMOS at about the 0.25-$\mu$m node.

It should be noted, however, that while their use permits smaller $n^+$-to-$p^+$ spacing, the presence of trench structures also gives rise to additional leakage paths not present in LOCOS-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. 9-6a).

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 $+V_{DD}$, and the trench is narrow (e.g., $\leq 1$-$\mu$m), inversion along the sidewall of the trench between the source and drain of the

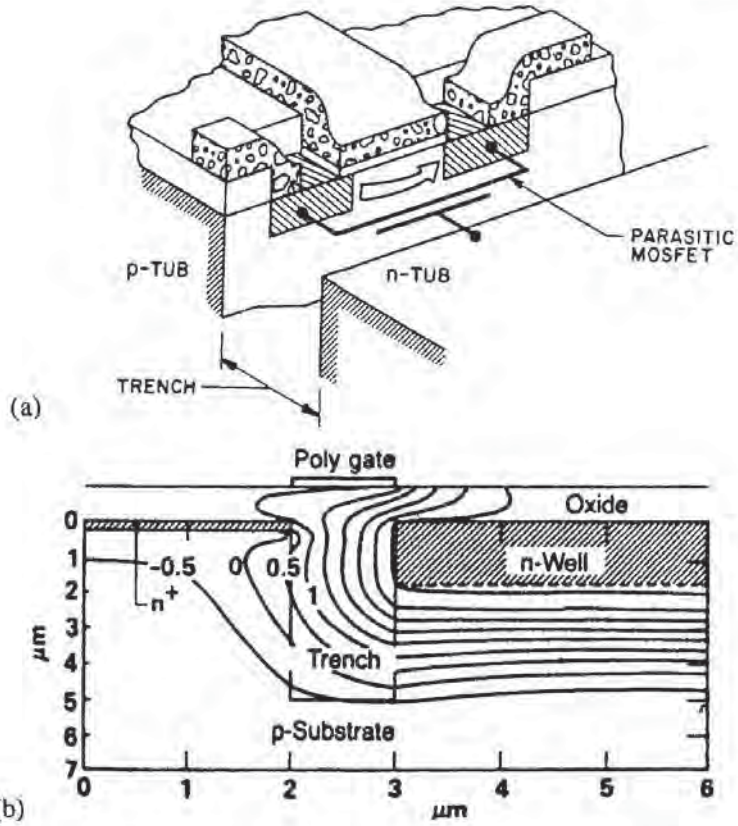



**Fig. 9-6** a) A parasitic-sidewall-MOSFET with the *n*-well as the gate and the trench serving as the gate oxide.[12] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission. b) Cross-sectional view of a simulation of the 2-D equipotential contours in an *n*-well CMOS with trench isolation. The simulation is done at $Q_{fc} = 5x10^{10}$ cm$^{-2}$, $N_A = 9x10^{14}$ cm$^{-3}$, $V_{well} = 3V$, $V_{GS} = 0V$, $V_{sub} = -1V$.[13] (© 1983 IEEE)

MOSFET can easily occur. As seen in Fig. 9-6a, this is the trench sidewall that is outside of, but facing the *n*-well. In Fig. 9-6b, the simulated 2-D potential contours in a trench-isolated CMOS structure are shown. Note that the potential on the trench sidewall 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 well edge (as shown in Fig. 9-6a). 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 well edge.

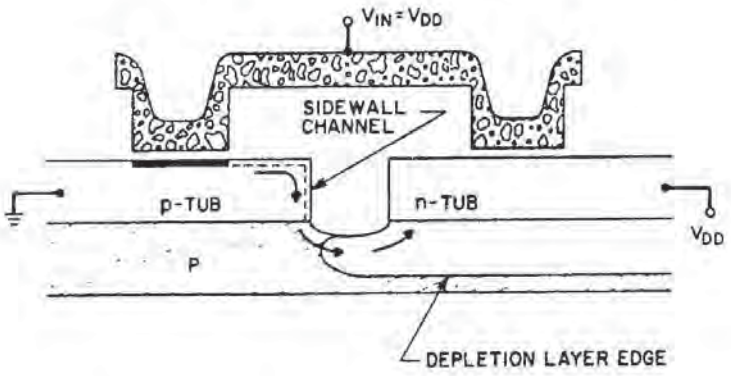



**Fig. 9-7** A trench-isolated *p*-well and *n*-well with a parasitic vertical-sidewall-MOSFET (with the *n*-well as the gate) and a parasitic-horizontal-MOSFET (with the $V_{in}$ poly-line as the gate).[14] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission.

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. 9-7.

Assuming that the leakage path of Fig. 9-7 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. 9-8 for the case of two series-connected CMOS inverters which suffer from inadequate CMOS isolation.[15] 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., $V_{in1} = 5V$), 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. 9-8 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*-well, then along the inverted trench sidewall, then under the poly of gate 2 (note that for clarity, in Fig. 9-8 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 $Si/SiO_2$ interface at the trench-sidewall. The interface fixed or trapped charge that exists at the $Si/SiO_2$ interface is positive, which tends to lower the inversion-threshold of $p$-channel regions (in this case the $p$-well). To minimize this interface charge, a thin, good-quality thermal-oxide-layer must be grown inside the trench. This oxide growth process, however, gives rise to the second problem. That is, during thermal oxidation, some of the boron dopant in the $p$-well segregates into the oxide. Thus, the B surface-concentration at the interface is reduced, further decreasing the inversion-threshold. As a result, inversion of $p$-well trench-sidewalls occurs at lower threshold-voltages than do those of an $n$-well for comparable well-doping concentrations. In 0.25-$\mu$m (and smaller) CMOS technologies, this last problem is circumvented by forming the well-regions with an implant after the STI-structure is completed. A high-energy-implant process is used to implant well-dopants through the STI-oxide (and in

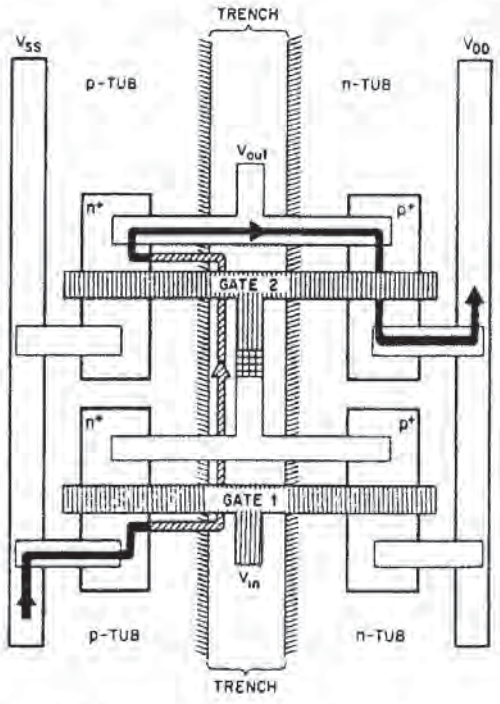



**Fig. 9-8** A dc leakage path through two inverter stages in series. The solid black path is a good conducting route and the shaded path is a parasitic route. The parasitic consists of a field oxide corner under Gate-1 (cf. Fig. 9-7), a sidewall path along the trench (cf. Fig. 9-6a), and a corner path under the $V_{out}$ strap.[15] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 John Wiley & Sons. Reprinted by permission.

the active regions). Thus, the boron-segregation problem that would occur during the trench-liner-oxidation process is avoided.

To counteract intra-well and inter-well leakage in trench-isolated-CMOS several steps are necessary. First, the surface-leakage-path must be suppressed by use of sufficient channel-stop-doping between the active-device and well-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).

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. 9-9a).[16] 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.*[17] 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, since phosphorus piles up during the brief thermal oxidation of the trench prior to CVD-oxide-refill.

In this selective-sidewall-doping process, a nitride/pad-oxide mask-stack is first formed and patterned. Next, after shallow trenches are etched by RIE to a depth of 0.3-0.5-$\mu$m, 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 or BSG away from the *n*-well-sidewall and trench-bottom (see Fig. 9-9b). 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 was used for trench widths as small as 0.8-$\mu$m (i.e., used, for example in a 0.35-$\mu$m process). For the smaller-width trenches used in 0.25-$\mu$m and smaller technology nodes, the problem is handled

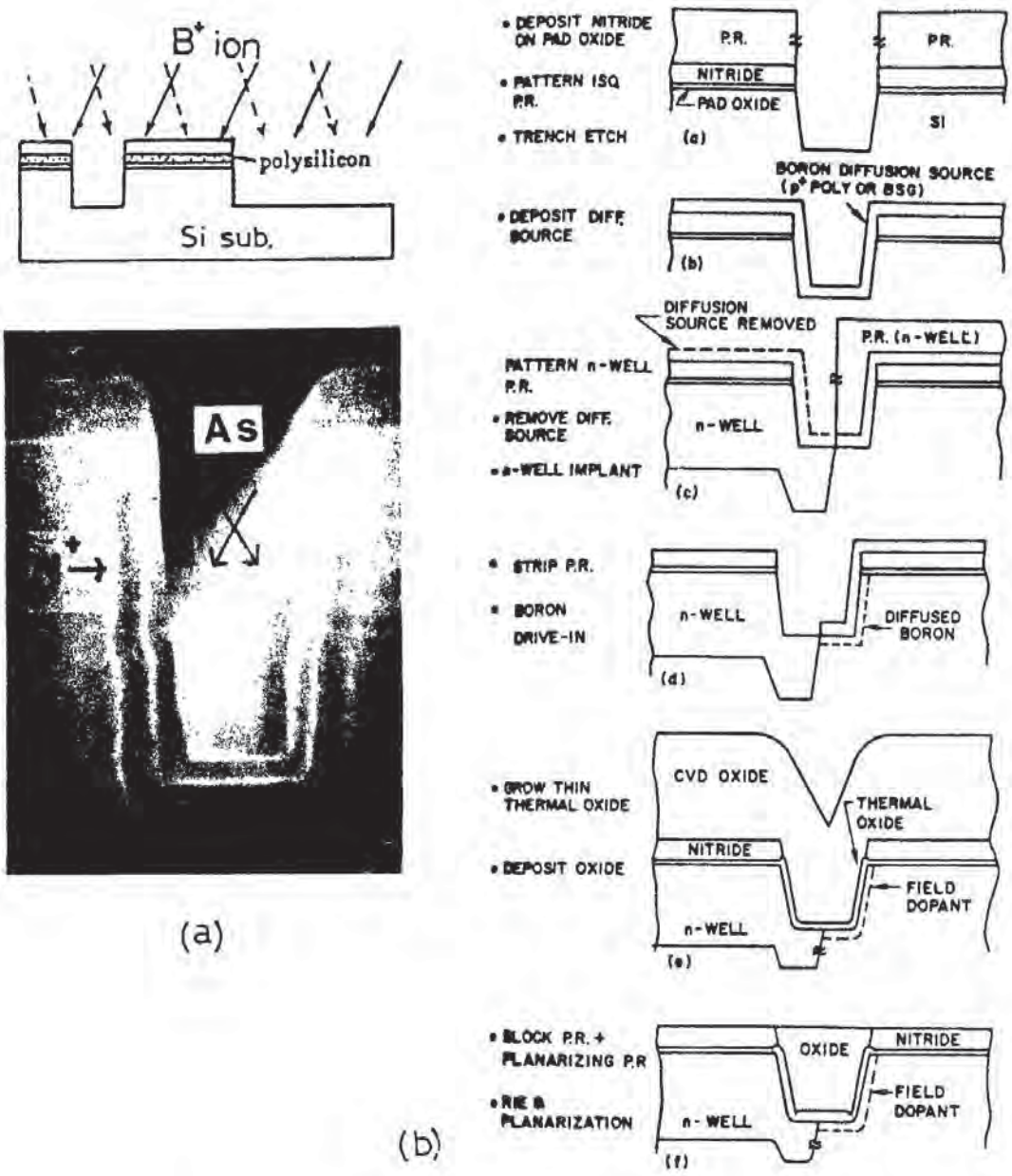



**Fig. 9-9** a) Schematic of oblique ion implantation and SEM cross section result of this procedure.[16] (© 1987 IEEE) b) Shallow trench isolation process flow using a doped oxide as diffusion source of trench sidewall doping.[17] (© 1988 IEEE)

differently (e.g., the typical minimum trench width in a 0.18-$\mu$m process is about 0.4-$\mu$m with an active-area width of about 0.6-$\mu$m). That is, the $p$-well-doping implant-step (carried out after the STI structure is completed, using a high-energy-implant process) also places boron channel-stop-dopants at the bottom of the trench in sufficient concentration to prevent an inversion path from occurring

there. In addition, the boron-doping at the silicon surface in the channel-region of the NMOSFETs required to prevent DIBL in deep-submicron MOSFETs is high enough to keep the sidewalls of the trench from inverting.

## 9.4 DETAILS OF THE PROCESS FLOW TO FORM A BASELINE SHALLOW-TRENCH-ISOLATION (STI) STRUCTURE

In this section we present a process for fabricating STI structures. Note that the process sequence we describe is that of a baseline process, and one that is likely carried out for 0.25-$\mu$m CMOS generations. Some modifications can be made to this baseline process (even at the 0.25-$\mu$m technology node), and we will describe these variations in a later section. As ICs are scaled further, other modifications to the STI process will have to be made, and we will also discuss the need for such changes in the sections that follow this one that cover those process integration challenges in more detail.

The baseline sequence of steps for forming STI structures is given as follows:

1. The first several steps performed in fabricating an STI structure are the same as those used in the LOCOS isolation process sequence, with the exception being that the trenches for shallow trench isolation are the first structures formed on the bare silicon starting wafer. (On the other hand, in the conventional twin-well-CMOS process that used LOCOS, the wells were the first structures created on the bare-Si starting wafer.) In the STI process sequence, the bare, starting Si wafer is first cleaned (e.g., with an RCA clean), and then a pad oxide is thermally grown on the Si surface (typically to a thickness of 10-20-nm). Note that the role of the pad oxide layer is to prevent stresses that would exist if the silicon nitride layer (which is deposited next) were deposited directly on the bare-Si wafer. Such stresses could cause defects in the Si-substrate to occur. Note, however, that in the STI process the stresses associated with the formation of the STI structure are smaller than those involved in forming a LOCOS isolation structure. That is, in the LOCOS process, a long, high-temperature operation is performed to grow the field-oxide. During this field-oxide-growth step, the stresses between the nitride and the Si-substrate can be severe, and the pad oxide must therefore be sufficiently thick to keep these stresses from getting large enough to cause defects in the Si-substrate. In the STI process, there is no high-temperature field-oxide-growth process, and thus the stresses between the nitride film and the Si surface will not be as large as those in the LOCOS process. Hence,

the pad oxide in the STI process can be significantly thinner than the ones used in the LOCOS process (10-nm-thick vs. 25-40-nm-thick).

**2.** A silicon-nitride layer is then deposited on the pad-oxide layer (Fig. 9-10a). The silicon nitride plays two roles in the fabrication of the STI structure: **i**) it serves as a hard mask for etching the trench; and **ii**) it acts a CMP polish-stop-layer when the trench-fill dielectric material is planarized with a CMP process. The nitride-layer-deposition step is typically performed with an LPCVD process that deposits a silicon-nitride layer that is 100-200-nm thick. Note that the LPCVD of silicon nitride is a batch process that is simultaneously carried out carried out in a furnace tube on many wafers at once. As a result, the backsides of the wafers are also exposed, and silicon nitride is deposited on the back of the wafers (as well as on the front). Hence, this backside-nitride must somehow later be removed.

**3.** Then photoresist is applied and *Mask #1* is used to pattern the regions on the wafer surface where STI trenches will be opened.

**4.** The nitride and pad-oxide are then anisotropically etched, using a plasma-etching process (Fig. 9-10b). Usually this etch is carried out in a dry etcher designed for etching dielectric materials (*Etcher #1*).

**5.** The trenches are then etched (Fig. 9-10c). This silicon-etch step is also carried using a plasma-etching process. The goals of this etch process include: **a**) etching the trenches to the correct depth; **b**) controllably producing trenches with the sidewalls having the desired slope (e.g., from 70-85°, depending on the technology node, the trench depth, and the type of circuit being fabricated); **c**) producing trenches with rounded bottom corners; and **d**) having the etch process produce smooth sidewalls and minimum Si damage.

In general, it is easier to fill trenches without forming voids in the trench-fill material if the trench openings are sloped. Obviously, this requires more Si area for the trenches, and thus as ICs are scaled, the sidewall profiles will have to become more vertical. Gasses used for etching the trenches are mixtures of HBr and other gases (e.g., $HBr/Cl_2/O_2$, or $HBr/NF_3/He/O_2$).[18]

The trench-etch process is usually carried out in a second etcher (*Etcher #2*), different from the one used to etch the nitride and oxide layers (*Etcher #1*). That is, *Etcher #2* is usually a modified version of a gate-etch tool (i.e., one in which polysilicon is etched).

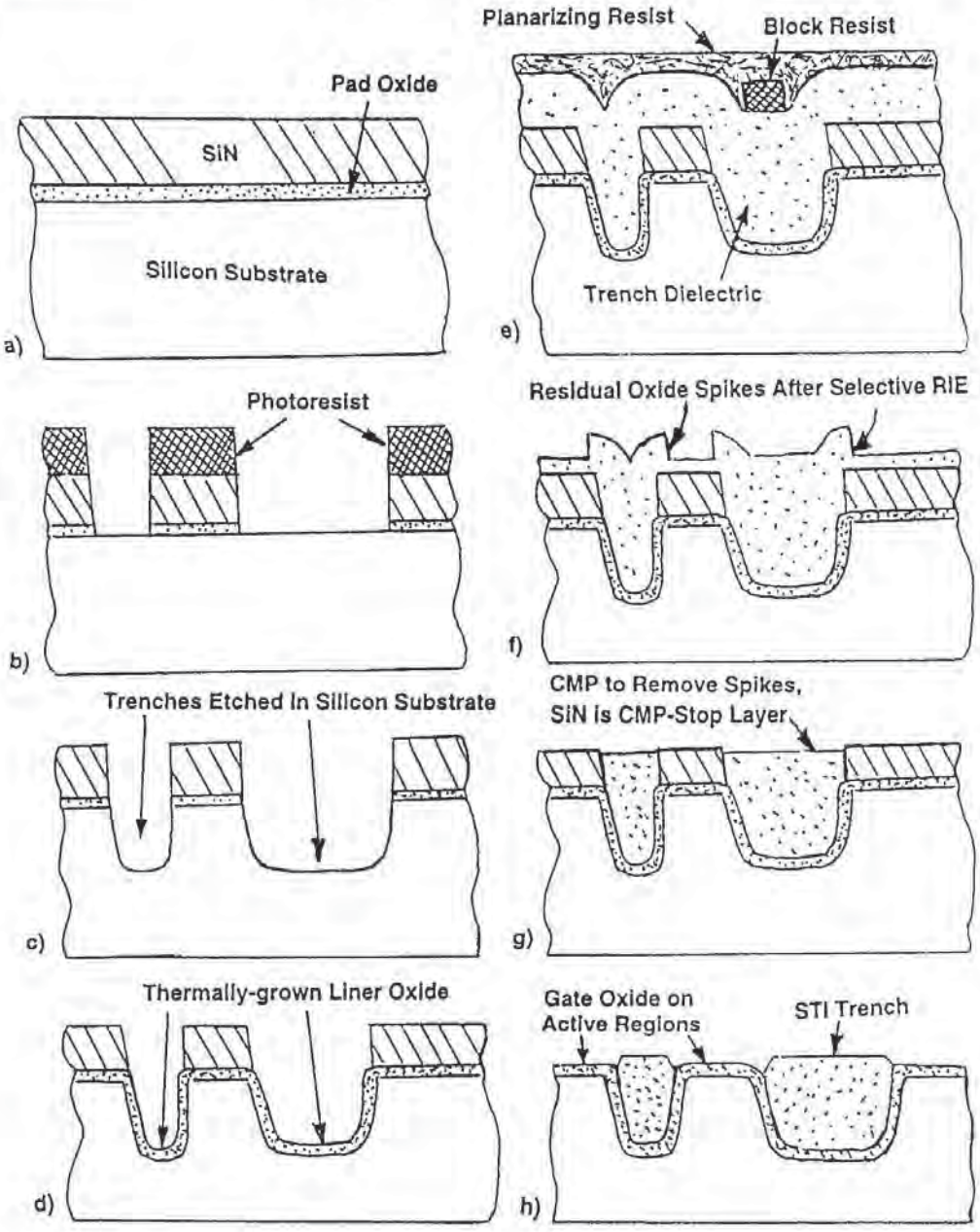



**Fig. 9-10** Baseline process flow used to form an STI structure: a) Form pad oxide by thermal oxidation, then deposit a silicon-nitride layer by LPCVD; b) Pattern the nitride and pad-oxide layers; c) Etch trenches, strip resist; d) Grow liner-oxide on the trench walls, using thermal oxidation; e) Deposit the trench dielectric & densify it, then deposit a block resist layer & pattern it. Finally deposit the planarizing resist layer; f) Use a selective-RIE-etch process to simultaneously remove the resist and the trench-oxide-layer, leaving residual oxide spikes; g) CMP the CVD-oxide-layer to abrade spikes and planarize its surface. The silicon nitride acts as a CMP-stop-layer; h) Strip the nitride and then grow the gate-oxide on active regions. The STI structure is now completed.

Note, however, that these two etch steps could also be carried out in a single etcher. Significant enhancements to process productivity (in the form of higher throughput and lower cost) could be obtained if the opening of the silicon-nitride hard-mask-layer and the etching of the trenches were carried out *in-situ* in a single etcher (and such a process has been termed *in-situ hard-mask-open*). Etch tools capable of *in-situ* hard-mask-open are available from Applied Materials, Lam Research, and TEL America, and some such processes have reportedly been put into production by 2000.[19] The main challenge of successfully implementing an *in-situ* hard-mask-open process is selectivity to the photoresist.[20]

The trench etching process is difficult for a number of the following reasons:

**a.** The trench depths across the wafer must have uniform depth to ensure correct circuit operation (the dependence of circuit performance on trench depth will be discussed later, in the section that deals with CMP-dishing problems). However, some dry-etch processes exhibit *aspect-ratio-dependent-etching* (ARDE) characteristics, in which material from wider-spaced trenches is etched more rapidly than from narrower trenches. Thus, etch-process development must be employed to find etch-process conditions in which ARDE effects are limited.

**b.** The trench etch process is a 'blind' process, in that there is no etch-stop layer at the bottom of the trench, nor is there any material present at the trench bottom that will create an end-point-signal. Hence, an etch process must be available that is so well-controlled

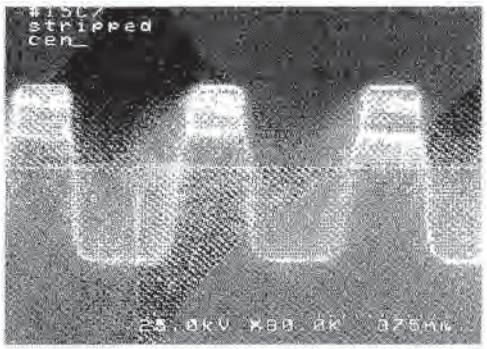



**Fig. 9-11** SEM of etched trenches of an STI structure with trench walls having an 87° slope and bottom corner rounding. Courtesy of Lam Research.

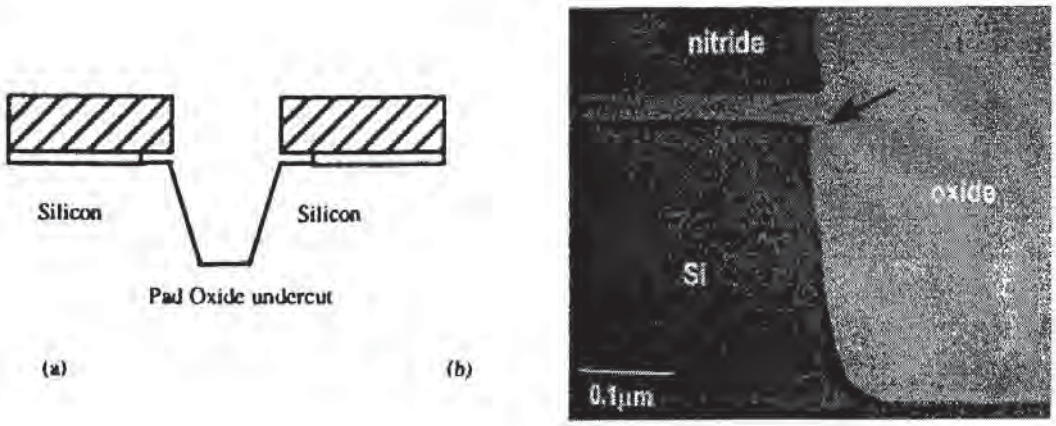



**Fig. 9-12** a) Pad-oxide trench-undercut in the STI-baseline process. b) TEM of an as-etched trench profile showing the sharp corner of the active Si region at the top of the trench.

that a timed etch can be relied upon to produce trenches with the correct depth, from wafer-to-wafer, and across a wafer.

**c.** The etch process should yield smooth trench sidewalls with angles of 70–85°, rounded bottom-corners (to minimize stress), and a residue-free silicon-surface after etch (see Fig. 9-11). Trenches with rounded-bottom corners minimizes stress to the silicon lattice encountered during the trench-liner-oxidation process of Step **8** and the trench-filling-process of Step **9**. It was reported that using gas etch mixtures consisting of $HBr/NF_3/He/O_2$ gave improved corner rounding results than mixtures containing $HBr/Cl_2$.[21]

**6.** After the trenches are etched, the resist is stripped.

**7.** The pad oxide is next undercut with a short HF dip (Fig. 9-12a). The reason this step is carried out has to do with the fact that upon its completion, a silicon dioxide layer is thermally grown on the exposed trench surfaces. This so-called *liner-oxide* layer serves several important roles. First, the process of growing this oxide consumes any residual damage caused by the plasma-trench-etch process. Second, the thermal oxide passivates the Si surface of the trench, providing as stable interface between the Si and the trench fill material. (The threshold voltage of the parasitic field MOSFETs must be kept high enough so that leakage paths at the Si-surfaces of the trenches are not electrically activated during normal device operation. To ensure that this requirement is met, the trench-Si/filler-material interface must be electrically stable. A thermal oxide in

contact with the Si surface provides the most electrically stable of interfaces.) Third, at the top corner of the active-region (i.e., where the active-region and the isolation-region meet, see Fig. 9-12b) the Si normally has a sharp corner after the trench etching has been completed. Such a sharp corner will enhance the electric field lines at this corner and degrade the MOSFET turn-off characteristics (as will be discussed in more detail later). To suppress this effect, the top-sharp-corner of the silicon active-region must therefore be rounded. Device simulations have demonstrated that the parasitic double-hump in the $I_D$-$V_{GS}$ curves can be eliminated by having a trench corner with a radius of 50-nm if a trench with no more than an 80° sidewall slope is specified.[22] This can be done in a number of ways, but here, we will assume that such a "corner-rounding" process is also performed by the thermal-oxidation step that forms the liner-oxide.

It has been demonstrated that the top-sharp-corner will be more effectively rounded by the thermal-oxidation step that forms the liner if the pad-oxide is undercut prior to performing the thermal-oxidation liner-step. Hence, as shown in Fig. 9-12a, this pad-oxide-undercutting step is performed next, i.e., by dipping the wafer in HF to etch-back the pad-oxide to a controlled extent.

**8.** The thermal-oxidation step that forms the liner oxide and rounds the sharp-top-corner of the trench is performed next. The oxide grown in this procedure is typically 40-50-nm thick, and it is grown in dry-$O_2$. It has also been reported that maximum corner-rounding is achieved if the liner-oxide is grown with a high-temperature (1050-1100°C) rapid-thermal-oxidation (RTO) process (Fig. 9-13).

Note that a thin layer of the nitride is converted to oxide during this liner



(a) Lower temp+Hcl (b) High temp RTO

**Fig. 9-13** Cross-sectional TEM of the top trench corner after pad oxide undercut & liner oxidation: HCl in the ambient (a); and higher-temp RTP-oxidation (b) improve corner rounding.

oxidation step, an effect that must be taken into account when the nitride is later stripped. Also note that in DRAMs fabricated with STI, a thin silicon-nitride liner may also be deposited by CVD prior to filling the trench with a CVD oxide. This nitride liner helps decrease leakage current, a critical parameter in DRAM operation.

**9.** Next, a CVD-dielectric film is used to fill the trenches. Since the deposition is a blanket process, some of the material is deposited onto the active regions that are still covered with nitride (i.e., 8000-9000-Å of gap-fill dielectric are deposited when filling a 5000-Å-deep [0.5-$\mu$m] trench). As a result, the material must be removed from the active regions after the deposition process is completed (Fig. 9-10d). As mentioned earlier, a number of CVD processes have been pursued for filling the trenches without void formation, including: LPCVD TEOS; (sub)atmospheric-pressure CVD of ozone/TEOS; and high-density plasma CVD (HDP-CVD).

**10.** The dielectric material is densified at about 900°C. This step is carried out so that the trench-fill dielectric will exhibit etch rates comparable to that of thermal $SiO_2$. This characteristic is necessary so that excessive loss of the trench-fill dielectric is not incurred during later HF deglazing steps (i.e., steps **15**, **17**, and **20**). This densification process reduces the wet-etch rate by about 20%.

**11.** A photoresist layer is next spun-onto the wafer. This resist is then patterned to form a reverse-mask of the active regions and a planarizing-resist layer is deposited over the patterned reverse-masked wafer (Fig. 9-10e). The oxide and the resist are etched simultaneously in an RIE-etch-back step that etches resist and oxide at the same rate. This reduces oxide thickness in the field areas. The result of partially planarizing the surface with this REI step is that the time needed to planarize the wafer surface during the subsequent CMP step is decreased. A shorter-duration CMP step decreases the problem of dishing (as discussed further in Sec. 9.5.3).

**12.** After the resist is patterned and the planarizing-resist layer is deposited, the wafer is etched back with an RIE process described in Step **11**, and as shown in Fig. 9-10f.

**13.** Any remaining photoresist is then stripped. Small spikes of oxide, however, remain near the edges of the active/field regions (see also Fig. 9-25). These must be removed to produce a globally planarized STI surface.

**14.** A CMP-step is used to polish back the dielectric layer to remove these

spikes, which is continued until the nitride is reached (the nitride acts as a CMP-stop layer, Fig. 9-10g). A major issue involved with this CMP-step is dishing of the trench-fill-dielectric in wide trench regions. The reverse-mask procedure of Steps **11,12,** and **13** is one method used to mitigate the dishing problem. Other dishing-suppression techniques are discussed in Section 9.5.4.

**15.** The nitride is then stripped, using a hot-phosphoric acid, wet-cleaning process. But, it should be noted that the stripping process is actually done using two wet-etching steps. That is, the thin layer of nitride that was converted to an oxynitride layer during the trench-liner-formation thermal-oxidation-step must be removed prior to carrying out the nitride-stripping-step itself. The oxidized layer on the nitride will not etch in hot phosphoric acid, and unless it is removed, the phosphoric acid will thus not strip the nitride. Thus, a short HF-etch step is first performed to remove the oxide layer on the nitride. Note, also that the CVD-oxide that has filled the trench is also attacked by this HF-etch-step. Since the CVD-trench-oxide etches more rapidly than does the oxide layer on the nitride, this HF-etch-step must be carefully optimized. Another factor that should be considered when determining this HF-etch is the possibility of there being insufficient CVD-oxide removal from the nitride by the CMP process. If the oxide on the nitride remaining after the CMP step is too thick, the HF-etch of the process in this step may not remove all of it, making the nitride-layer-etching process ineffective, leading to potentially significant yield degradation.

**16.** After removing the oxide layer on the nitride, the nitride can be stripped by the hot phosphoric acid. There are several reasons why a wet process is used to strip the nitride, rather than a dry process. First, the wet process is inexpensive and selective (i.e., the hot phosphoric acid will not attack the underlying pad oxide, or Si substrate). Second, a wet process will remove the nitride deposited on the back of the wafer earlier by LPCVD, as well as the nitride on the front wafer surface.

**17.** Normally the pad oxide is stripped in HF at this point. Again, the oxide in the trench is also attacked by this HF etch step. This etching plays an important role in forming the topography at the edge of the active region, where the polysilicon gate is deposited on both the active area and the trench region (see Fig. 9-14). This topic will be considered again later.

**18.** Grow sacrificial oxide, about 20 nm thick, to protect the exposed Si surface of the active areas from contamination during the implant process.

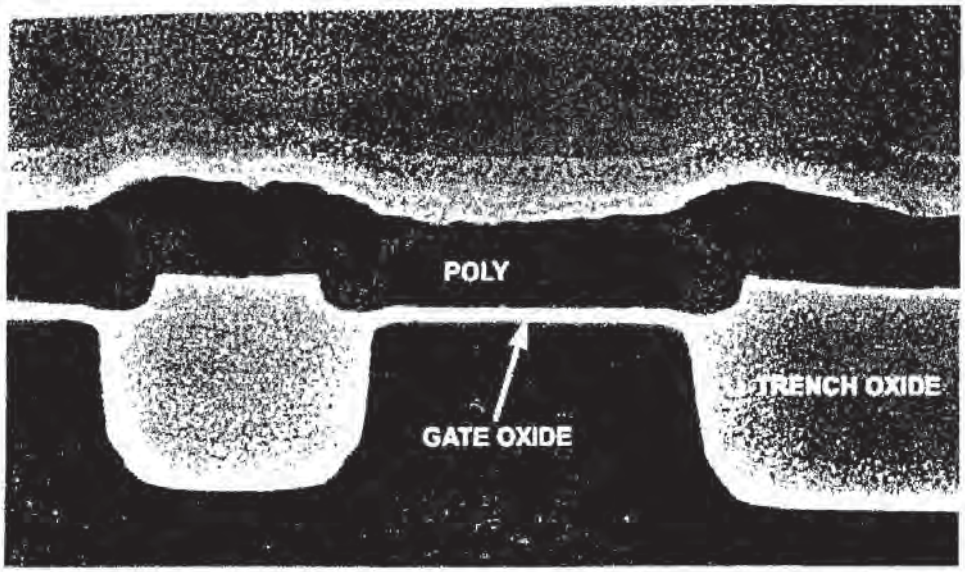



**Fig. 9-14** SEM showing polysilicon crossing the region at the edge of the active device region and the trench in an STI structure.

Note that another reason for stripping the pad oxide and growing a sacrificial oxide is that an effect similar to the Kooi-effect (that occurs during LOCOS isolation - see Vols. 2 and 3 for more details on the Kooi effect) has also been reported to possibly occur during STI processes. That is, if the trench-oxide is densified in an ambient containing $O_2$, this $O_2$ can react with residual $H_2$ in the trench oxide film or in the SiN layers present on the chip to form $H_2O$. This $H_2O$ in turn can react with the SiN to form $NH_3$ which would lead to nitridation of the Si as in the Kooi effect. The net result of such a phenomenon would be that the gate-oxide grown where such a nitrided region existed could be thinner there, as shown in Fig. 9-15.[23] To avoid the possibility of such gate oxide thinning (and a resultant degradation of the gate oxide integrity [GOI]), a sacrificial oxide is grown which will consume any

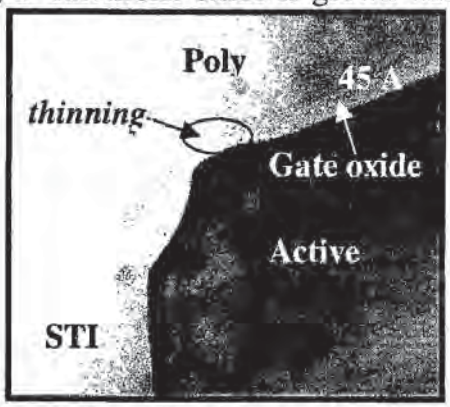



**Fig. 9-15** TEM showing thinning of the gate oxide near the edge of an active region in an STI structure due to the Kooi effect.[23]

nitrided regions that formed on the active Si surface regions.

**19.** The wells are now created by performing a high-energy-implant to place dopants within the Si substrate in the active regions, and beneath the STI structures to produce a channel-stop layer of dopants. In addition, the $V_T$-adjust- and punchthrough-implants are carried out at this time.

**20.** The sacrificial oxide is then etched, again using a wet (HF-etch) process. Once again, the oxide in the trench is subjected to an HF etchant, further reducing the thickness of the oxide in the trench (and also further modifying the trench edge corner region - Fig. 9-16). By the way, it should not necessarily be construed that the HF etches will remove enough oxide in the trench regions to cause the trench oxide to be recessed. While this might be a possibility, not much oxide thickness loss occurs as a result of the HF etches. The more significant effect is on the region of the trench oxide near the trench top corner.

**21.** Finally, the gate-oxide growth process is carried out (Fig. 9-10h).

## 9.5 ISSUES INVOLVED WITH STI PROCESS INTEGRATION

### 9.5.1 Issues Involved With Etching Trenches for STI

Some of the issues involving trench etching for a baseline-STI structure were discussed in Section 9.4. Here, we elaborate on some other issues that arise as more advanced STI structures are needed. In the baseline-STI structures, the trench sidewalls were typically sloped, with slopes ranging from 70-85°. As the trench widths are scaled to ever decreasing dimensions, the trench sidewall profile requirements are becoming steeper, with slopes as high as 88° being called for in some applications.[24]

One consequence of this change is that the *in-situ* hard-mask-open etch process may become more feasible. That is, in the baseline process of Sect. 9.4,

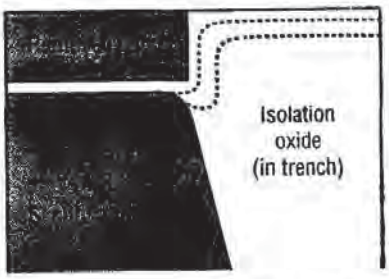



**Fig. 9-16** Topography of the top corner of the Si active region and the oxide surface in the trench. The dotted lines represent the removal of oxide that occurs during the wet HF dips before the sacrificial oxide and gate oxide growth steps.

STI structures had sidewall profiles with smaller slopes (e.g., 75-80°). If the trenches were etched while the resist remained on the wafer, the polymer deposition on the sidewalls during the trench etch would lead to increased ARDE effects, and this would have contributed to trench-depth and profile-variations between trenches of different widths. Thus, the two-etcher trench-etch-processes (with the resist being stripped after the 1st trench-etch step) were more desirable. To produce steeper trench profiles, lower-pressure etch-processes with higher ion-bombardment-fluxes are used. In these high-ion-bombardment-flux processes, there is much less polymer deposition needed on the trench sidewalls to protect them from lateral etching. As a result, the presence of the resist during the trench-etching stage can be more tolerated, opening the door for the *in-situ* hard-mask-open single-etcher solution. Note that the steep trench-sidewall profiles are also controlled by using low-pressure, high-density-plasma etchers. The gases employed seek to exploit a non-depositing etch process, with gas mixtures for example consisting of $Cl_2$ and Ar.

Another trend as devices are scaled, is toward using shallower trench depths. The baseline-STI structures used trenches with depths of ~0.5 $\mu$m. As devices approach the 0.1-$\mu$m node, the trench depths are expected to be decreased to < 0.2-$\mu$m, at least for some circuit applications at that technology node.[25] The trench depth is set by two main considerations:

**1.** The oxide thickness in the trench (field-oxide), together with the channel-stop-doping under the field-oxide determine the threshold voltage of the parasitic-isolation-region-MOSFETs. To keep this threshold-voltage high enough to prevent significant isolation leakage, a minimum field-oxide - thickness (i.e., trench depth) is needed. As the operating voltages are scaled for future device generations, smaller threshold voltages of the parasitic-isolation-region-MOSFET will be acceptable. Thus, a shallower STI-trench will be feasible.

**2.** The trench-oxide-depth impacts the capacitance value of the polysilicon and 1st-level-interconnect-metal to the substrate. The thinner the trench depth, the higher the capacitance. However, use of smaller operating voltages will also allow thinner STI-trench-depths to be used, with the exact depth value to be determined by the circuit application. Note that the severity of dishing problem, to be discussed in Sect. 9.5.4, will also play a role in determining how much the trench depth will be permitted to be reduced.

The contamination caused by the trench-etch-process is also an issue. Since an isolation structure is being fabricated, concerns regarding metallic contamination

increase with ever-decreasing device dimensions. To address this problem, Lam Research reportedly replaced all ceramic chamber components with quartz (as the ceramic material may contain trace levels of metal that could be released into the etch-gas ambient during the etch process, and then these could redeposit on the exposed Si of the trench walls). Applied Materials attacked this question by implementing a Clean Chemistry™ using fluorine throughout the etch, which eliminates the need for frequent dry-cleans of the chamber.

As already mentioned in the baseline-STI process sequence, maintaining *trench-depth-uniformity* across a die (and from wafer-to-wafer), is a challenging problem. However, merely controlling the *trench-depth* is also particularly important, because the uniformity of fill and also uniformity of CMP, in turn, depend on trench-depth-uniformity. The trench etch is essentially a blind etch, with no etch-stop-layer, nor any material differences in the Si substrate regions available to provide an end-point-detection signal. Thus, in the 2000-timeframe, the etch-depth-control must rely only on a timed-etch process.

Depth- and profile-control-nonuniformity are functions of reactor design and etch-gas chemistry. The stability of the chamber conditions also affect the trench-etch-rate-uniformity. The etch-rate will vary for different patterns, structures, open areas, and different microloading conditions. Thus, the specific etch-rate may need to be re-calibrated, depending on a particular set of conditions being faced. This indicates that there is currently a great opportunity for a production-worthy endpoint-detection-system for the STI-trench-etching process.[26]

### 9.5.2 Corner-Rounding Techniques in STI Structures

In Steps **7** and **8** of the baseline-STI process we described how the top of corner of the STI trench must be rounded so that adequate leakage performance charac-

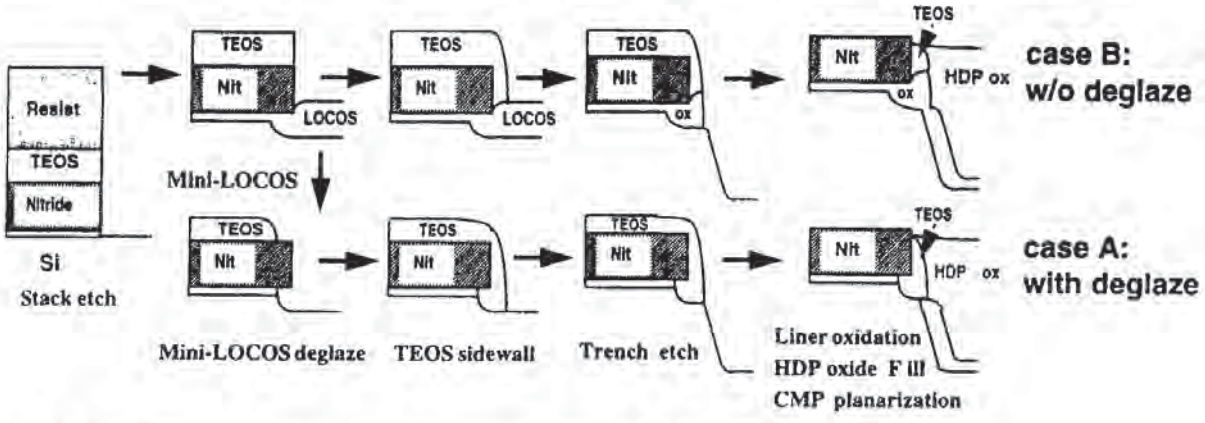



**Fig. 9-17** Process sequence of STI using mini-LOCOS edge.[29] (© IEEE 1997)

teristics will be exhibited by CMOS ICs that employ STI. The corner-rounding process described in the baseline-STI flow involved undercutting the pad-oxide just before performing the rounding (aka *liner*) oxide procedure. The best corner-rounding results are obtained if an RTO step in dry-$O_2$ at 1050-1100°C is used.[27] In the paper that described this step, several other corner-rounding procedures were also studied. These included: **a**) adding HCl to the dry-$O_2$-RTO process;[28] **b**) carrying out a post-CMP-oxidation in dry-$O_2$ at 1050-1100°C;[29] and **c**) using a "mini-LOCOS" process (see Fig. 9-17).[30] Good rounding is observed in the post-CMP-oxidation-process, but it comes at the expense of reducing the active-device area to some degree. The mini-LOCOS process ads considerable process complexity. A hydrogen-annealing procedure was also reported as an alternative method of achieving corner rounding. In this approach, the wafer is annealed at 1100°C after the trenches are etched (but before being filled). It is proposed that at this temperature, Si-surface atoms migrate and establish positions that represent the structure with the lowest free energy (namely a sharp corner becomes a rounded corner, see Fig. 9-18).[31,32] A novel approach to rounding the trench corner during the trench process itself is described in Sect. 9.5.6.

### 9.5.3 Trench-Fill Dielectrics for STI

We know that the trenches of the STI structure are filled with a dielectric, and that deposition of oxide-like films by CVD has been the chief technique for

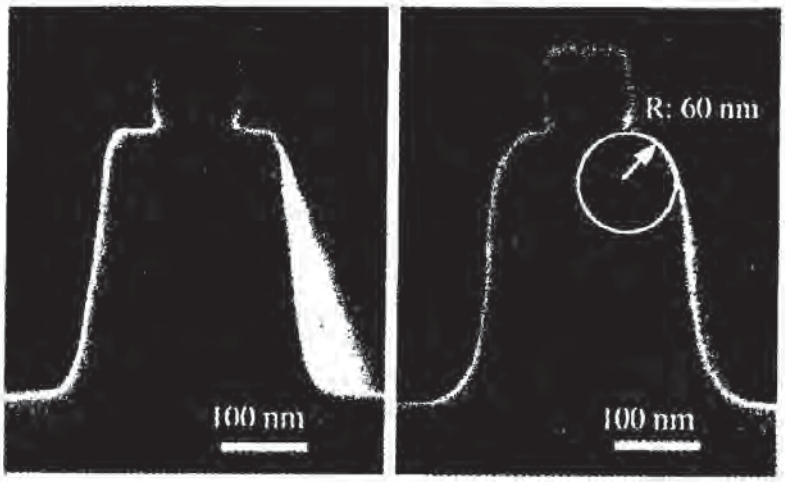



**Fig. 9-18** Cross-sectional SEM of before (a) and after (b) $H_2$-annealing. Annealing condition was 950°C, 380 torr, 60 sec. Silicon migration phenomenon can round out STI's top-corner-rounding-radius up to 60-nm.[32] (© IEEE 1998)

filling the trenches. A report published by Nag *et al.*,[33] compared several such dielectrics deposited by such methods for their suitability as the trench insulating material. The criteria they used to evaluate these materials were as follows:

**1.** The deposited material must be void-free, to avoid opening seams during HF-deglazing steps (i.e., such voids could later lead to poly filaments).

**2.** Post-deposition HF-deglazing also reduces the trench-fill oxide thickness. This can expose silicon at trench corners and cause leakage. Therefore a comparison of the etch rates of the trench-fill-oxides relative to thermal oxide must be established.

**3.** Film stress, shrinkage, and composition after the high-temperature anneal (densification) affect the integration.

**4.** The CMP-polish-rate relative to nitride and polish uniformity of the oxide impacts the trench-fill-oxide thickness remaining after CMP that influences the "corner effect" problem.

**5.** Isolation characteristics of PMOS and NMOS diodes have to be tested to verify the electrical performance of trench-fill dielectrics.

In this study four films types were compared: LPCVD TEOS; HDP-CVD oxide; SACVD $O_3$-TEOS; and an HSQ-SOG (see Vol. 1, 2nd Ed for more information on HSQ-SOGs, as well as Chap. 14 of this text). In general, they concluded that the HDP-CVD oxides were the most suitable of these four films for STI-trench filling. Table 9-1 summarizes the results of their study. Trenches 0.5-$\mu$m-deep and widths as small as 0.22-$\mu$m could be filled without forming voids.

**Table 9-1 TRENCH-FILL DIELECTRIC EVALUATION CRITERIA**

| Criteria | SACVD | LPTEOS | HDP | SOG |
|---|---|---|---|---|
| | | | | |
| Gap-fill | OK | No | OK | OK |
| Hi-T Anneal required | Yes | Yes | No | Yes |
| As-dep'd HF-etch rate vs. thermal ox. | 5 | 2.1 | 1.1 | |
| Densified (850C) HF etch | 1.3 | 1.1 | 1.1 | 1.1 |
| CMP rate vs. SiN | 5.1:1 | 5.1:1 | 5.1:1 | 5.2:1 |
| Shrinkage at 850C | 5-6% | 3-4% | 0-0.8% | 22-28% |
| Moisture Absorption | Yes (cure) | No | No | Yes (cure) |
| Process | Seed/Cap | 1 step | 1 step | Seed/Cap |

Some more recent reports, however, have indicated that some of the other films might still be good candidates as trench-fill materials. For example, a study by Xia *et al.*, demonstrated that a two-step deposition of a BPSG film using a sub-atmospheric (SA)CVD-process of $O_3$-TEOS is also capable of void-free gap filling of trenches whose widths are as small as 0.06-$\mu$m and whose depths are 0.35-$\mu$m deep.[34,35] Figure 9-19 shows a SACVD chamber used to deposit such films.

Nevertheless, since the HDP-CVD method appears to have emerged as the technology of choice for depositing STI trench oxides, some comments should be made about this process (and the potential impact that the HDP-CVD method can have on the STI structure). Note that Novellus (SPEED® HDP-CVD model) and Applied Materials are vendors of HDP-CVD tools for dielectric deposition.

The basis for the improved gap-filling capability by films deposited with HDP-CVD is that HDP sources provide a high-density of low-energy ions to the wafer surface. These ions simultaneously sputter the film being deposited by CVD. The sputtering removes film material from some regions on the surface more rapidly than from other regions. This effect is exploited to yield films that fill narrow gaps without voids. Another benefit of HDP-CVD is that films can be deposited at lower temperatures. This mechanism is described in detail here.

The film being deposited by CVD initially covers all the surfaces of the wafer conformally (including the sidewalls and bottom of any trenches or contact holes). As the film grows thicker, an overhang develops at the top of the trench. This will close-off the opening of the trench before the dielectric material completely fills it, leaving a void. If the overhang can be prevented from

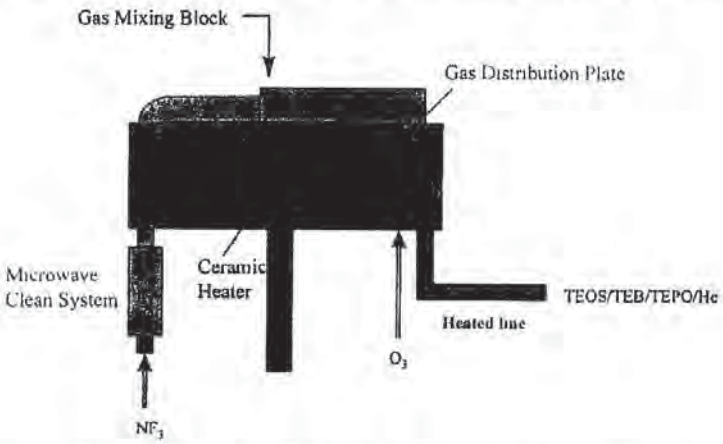



**Fig. 9-19** Schematic of Applied Materials SACVD Giga-Fill BPSG chamber, equipped with a high-temperature heater and remote microwave system for predeposition cleaning.[34]

forming, it may be possible to fill the trench without void formation. This is accomplished in an HDP source by the bombardment of the wafer surface with a high density of ions. For depositing oxides by HDP-CVD, a mixture of $SiH_4$, $O_2$, and Ar (or He) gases is most commonly used. The HDP source excites the oxygen and argon gas molecules. A bias applied to the wafer pulls the energetic ions out of the plasma and directs them toward the wafer surface. The oxygen reacts with silane to form $SiO_2$ while the Ar ions simultaneously sputter deposited material away. A high-density-plasma source is needed to maintain a high-ion-density at the relatively low pressures (a few mTorr) needed to obtain the directionality required for good gap-fill. The good gap-fill is achieved as described below.

The sputter yield is higher for surfaces that face the incoming ion beams at an angle, than at those surfaces perpendicular to the incoming ion directions. The maximum sputter-etching-rate occurs on surfaces at about 45° to the normal. Thus, a deposition-to-etch-ratio exists that varies with location on the wafer surface. The net deposition-rate is higher on flat surfaces (the field-regions of the wafer and the trench bottoms), than it is on the trench corners. (These angular surface regions are the 45° facets formed by depositing dielectric film at the top of the trenches.) By adjusting the deposition-to-etch-ratio, the trench openings can be kept from closing up while the dielectric film fills the trenches from the bottom up (Fig. 9-20). Sputtering from the via bottoms also causes re-deposition of some material onto the sidewalls. This helps re-distribute the film so the gap is

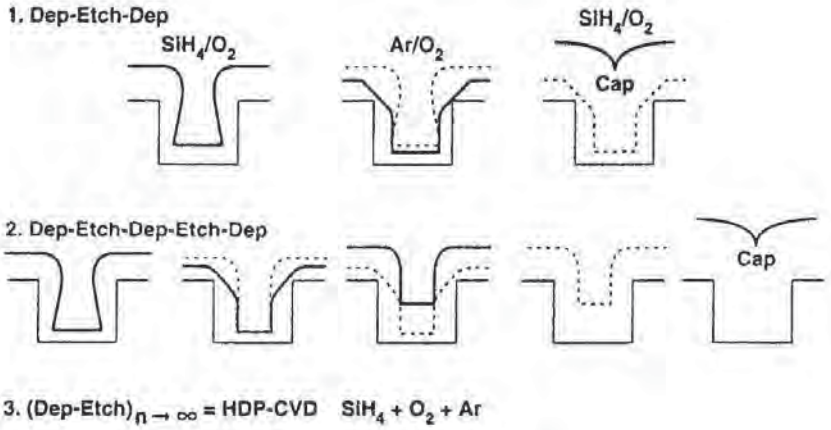



**Fig. 9-20** High-density plasma CVD achieves good gap filling through simultaneous deposition and etching. The deposition-to-etch-ratio is usually kept between 0.14 and 0.33.[43] Reprinted with permission of Semiconductor International.

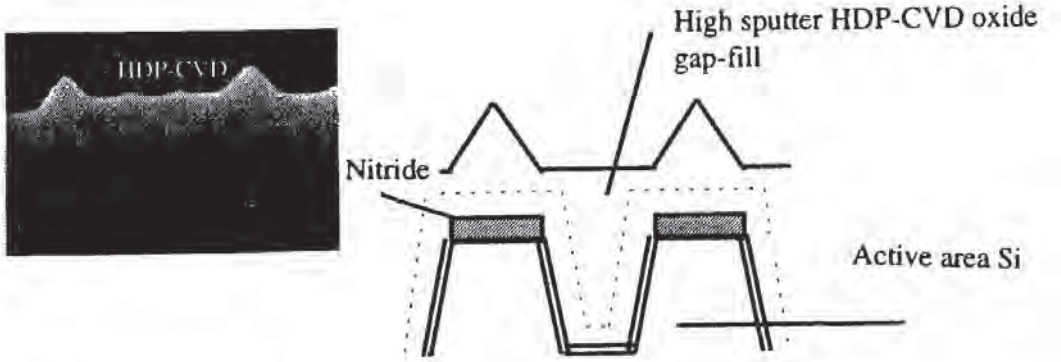



**Fig. 9-21** The surface topography of a dielectric film deposited by HDP-CVD showing the sharp spikes that remain as a result of the dep-etch nature of the HDP-CVD process.

filled without forming a void. To achieve good filling, the ratio of deposition to removal rate should be no larger than 3.0. The wafers are also heated by the ion bombardment. Thus, the deposition-rate is limited by the highest sputter-etch-rate that can be tolerated. The process is run at low pressures (< 10 mtorr), which allows ions striking the wafer surface to cross the sheath with no collisions. They arrive with the full energy they receive from the voltage drop across the sheath.

At the conclusion of the deposition process, the surface of the dielectric will generally still contain sharp ridges over the underlying metal lines (Fig. 9-21). CMP is needed to restore the wafer surface to a planar condition (although the amount of CMP required is reduced from that needed for a dielectric film deposited with a non-HDP-CVD process).

One issue that arises with HDP-CVD is that the combination of deposition and etching can sputter the top of the nitride corners, in effect "clipping them." One technique that suppresses this effect is to use He instead of Ar as the sputtering species in the HDP-CVD process. Another approach is to decrease the deposition-to-(sputter)-etching ratio of the process. Since this reduces the gap-filling capability, a trade-off must be made to optimize the overall deposition process.

### 9.5.4 Dishing Problem Associated with the CMP of STI Trench Dielectrics

Another CMP problem that occurs when fabricating STI structures is *dishing*. The effect involves the thinning of trench oxide regions that are exposed during the overpolish time that occurs after the CMP-stop layer is reached (Fig. 9-22). A silicon nitride layer is used to act as a CMP-stop layer. The thinning of the exposed oxide areas between the nitride-CMP-stop features is known to be highly dependent on feature size, with the dishing effect becoming larger as the

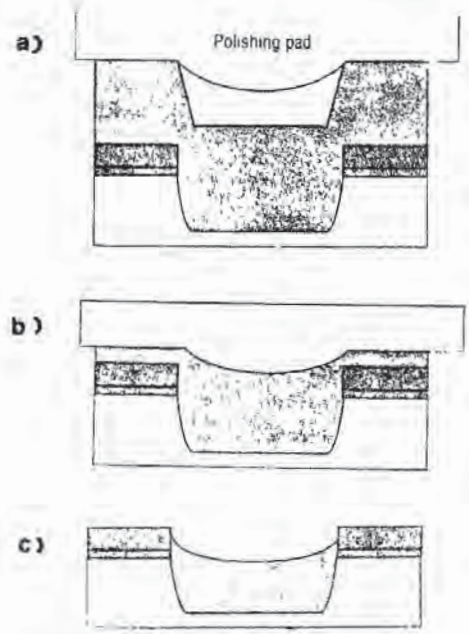



**Fig. 9-23** Origin of the dishing effect in CMP: (a) polishing pad flexes into the trench opening; (b) formation of a concave surface; (c) trench area remains "dished" after planarization. Reprinted with permission of Solid State Technology. Published by PennWell.

width of the field oxide is increased (Fig. 9-24). For example, in the CMP process of Fig. 9-24, dishing causes the field oxide thickness to decrease by more than 50-nm when the field oxide width exceeds 100-$\mu$m.[36] Dishing occurs because the polishing pad is not perfectly rigid, and because oxide removal is faster than nitride removal. If a pad exhibited perfect rigidity, CMP of the exposed oxide regions would produce a perfectly planar surface (because the exposed recessed regions would not be subjected to any polishing pressure). However, since actual pads exhibit some flexibility, above-zero polishing pressures are produced in recessed areas (especially in the middle of wide exposed regions) when polishing continues after the nitride has been exposed. Thus, some oxide is removed from these regions, and a "dished" region is observed after the polish step is completed. To minimize dishing, the silicon dioxide polish should be stopped as soon as the oxide over the nitride clears. Accordingly, the quality of an STI-CMP-process is measured by the control of trench-oxide and nitride thickness. This requires tight control of the step height between the trench-oxide surface and silicon surface after nitride strip.

Dishing must be minimized to maintain planarity for subsequent wafer processing. It still represents a significant limitation of the STI process. A number of

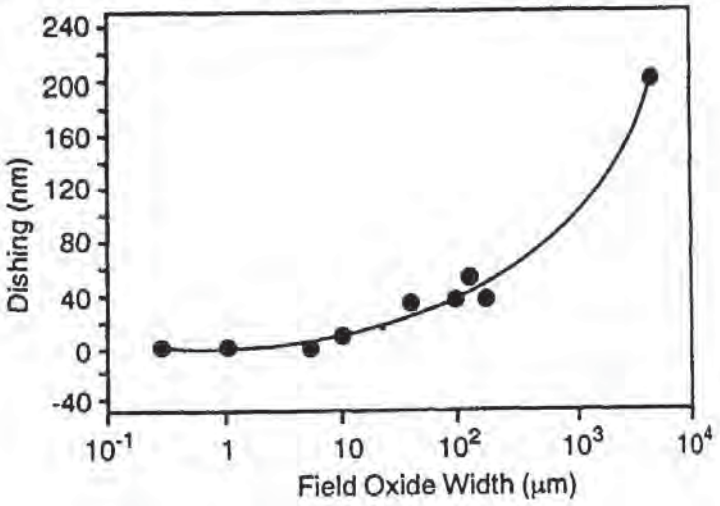



**Fig. 9-24** Dishing versus field-oxide-region width for a specific oxide CMP process[36] (©1993 IEEE).

approaches have been proposed to decrease dishing, including: **1**) a combination RIE/CMP process (as described in Sect 9.4); **2**) dummy active-regions; **3**) adding an extra nitride layer onto the trench-dielectric after it has been deposited; [37,38] **4**) adding an extra nitride layer onto the dielectric after it has been deposited, and using a reverse block-mask;[39] **5**) implementing a direct-polish CMP-process using either a linear-polishing tool, or a rotating-polishing tool with a harder polishing pad and an efficient endpoint-detection-system.

As described earlier, in the *RIE/CMP approach* the goal is to reduce the oxide thickness over active areas prior to CMP, so that a shorter CMP-oxide-removal step is needed. Since CMP is performed for a shorter time, the degree of dishing in wide trench regions is reduced. In the first step of the RIE/CMP process, a mask that contains a pattern that is the reverse of the trench-opening mask is used (but, with the exception that active-regions below a specified nar-row width are also not exposed - Fig. 9-24).[40] This mask is aligned so that its opaque patterns cover the trench regions on the chip surface. That is, this mask exposes the active regions of the chip, leaving the trench regions covered with resist (Fig. 9-10e). After the subsequent deposition of another planarizing resist layer (see Fig. 9-24), an etchback step is carried out using RIE. This process re-moves resist and dielectric material at the same rate, and so thins the oxide covering the active regions. Note that oxide in the trenches is not etched while oxide over the active regions is being removed, because the trench oxide is still covered with resist.

After this RIE step the resist is stripped, leaving the oxide over the active areas partially removed, Now, when the CMP step is performed, it only has to remove a thinner oxide layer, reducing the CMP time This not only reduces

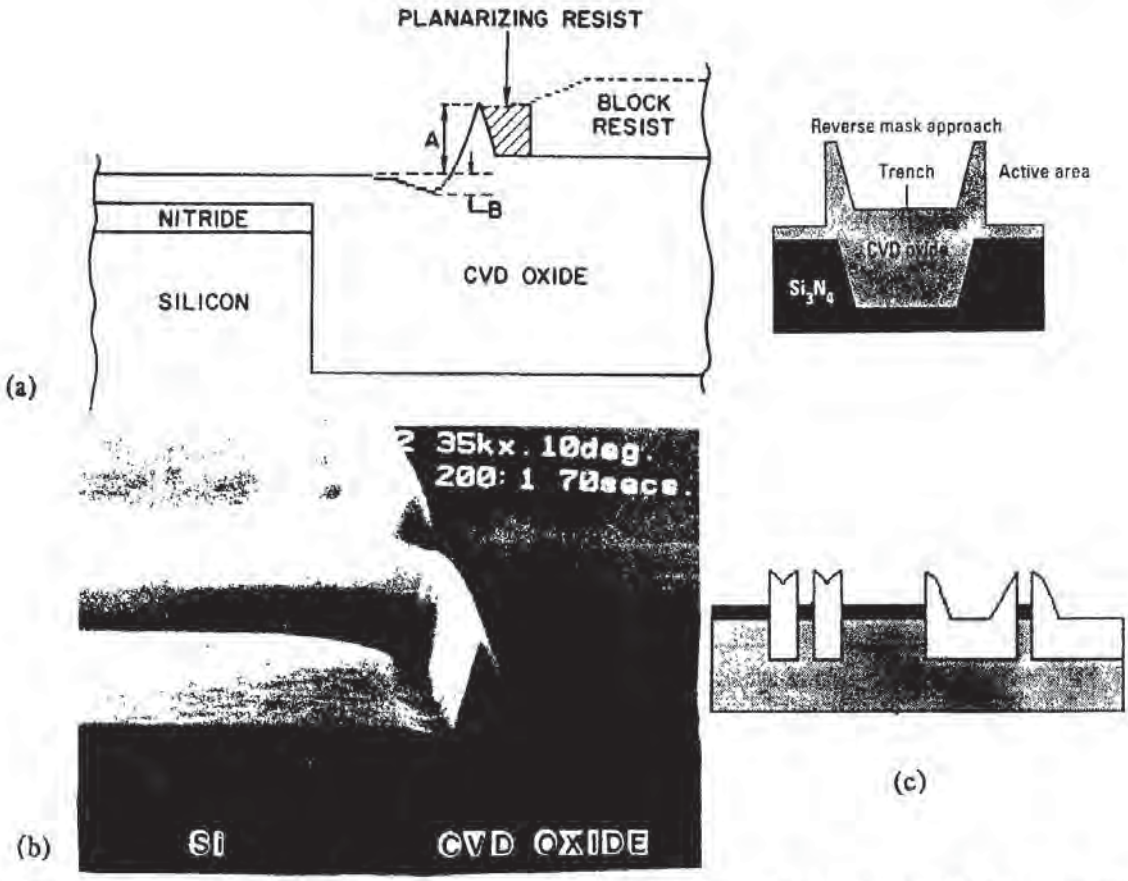



**Fig. 9-25** Spikes in the dielectric layer after the RIE step of the RIE/CMP process: a) Simulated cross-section after RIE; b) SEM cross-section after RIE and resist strip, showing dielectric spikes; c) Schematics of the dielectric spikes after RIE and resist strip.

dishing, but reduces CMP cost (since a shortened CMP step allows the throughput of the CMP procedure to be increased). Note also that dimensions on the reverse-trench mask are actually slightly smaller than the dimensions of the active-regions themselves. Thus, even when the mask is aligned to the active-regions, a small border of the active-region oxide remains covered by the resist. As a result, at the conclusion of the RIE step, small "spikes" of oxide remain on the wafer surface (Fig. 9-25).[41,42] But, these spikes can be effectively removed by the CMP step, allowing a planarized oxide surface over the chip to be obtained. The main drawback of the RIE/CMP method is that it increases the complexity (and thus also the cost) of the process used to planarize the trench oxide.

In the *dummy-active-region approach*, some areas of the trench-region are not etched. Instead, as shown in Fig. 9-26,[42] regions of Si remain in the center of trench. This has the same effect as if the trench width was reduced, and the degree to which dishing occurs decreases with trench width. Consequently, the

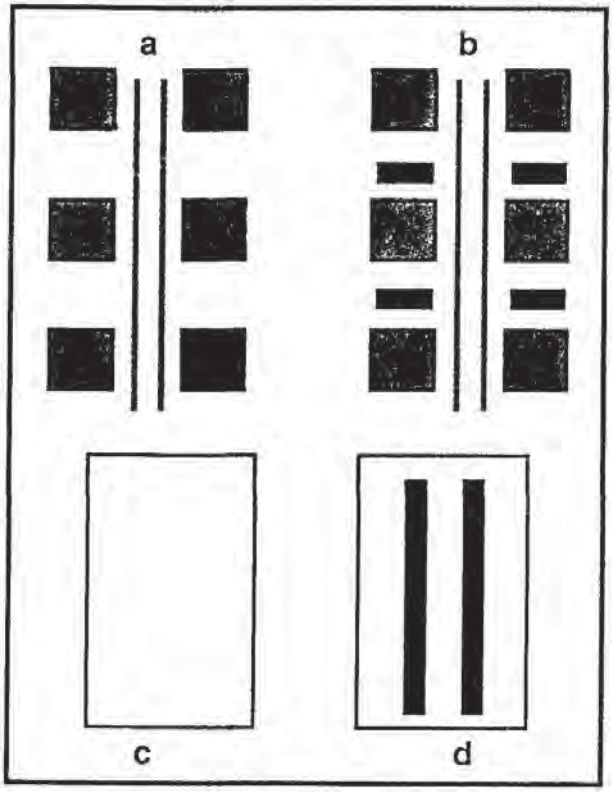



**Fig. 9-26** Dummy (b) and dishing support features of small (a) and large wafer (c) features can prevent dishing.[50] Reprinted by permission of Semiconductor International.

dummy structure method is also effective in reducing dishing effects to an acceptable degree. The CMP-removal rate over regions that are densely packed with active areas is also slower than over regions containing small isolated areas. Thus, to expose the nitride-CMP-stop-layer over such densely packed regions, the nitride over small isolated active-areas gets stripped, causing damage to Si in areas beneath the stripped nitride. As a result, dummy-active-regions must also be added to areas around isolated active regions. The dummy-structure technique, however, poses several difficulties. First, it imposes electrical, design, and space constraints, with the main problem being that several times as many structures must be added within a die over previous designs. This complicates the design and mask making process, and significantly raises defect levels. In addition, such dummy structures in trenches cannot be used for analog- and mixed-signal chips because the coupling capacitance over trenches becomes variable, and this disrupts the correct operation of analog circuitry.

In Procedure **3** used to reduce dishing during trench-CMP, a thin layer of silicon nitride is blanket-deposited onto the trench-oxide before it is polished-back with CMP (Fig. 9-27). A slurry with low selectivity between nitride and

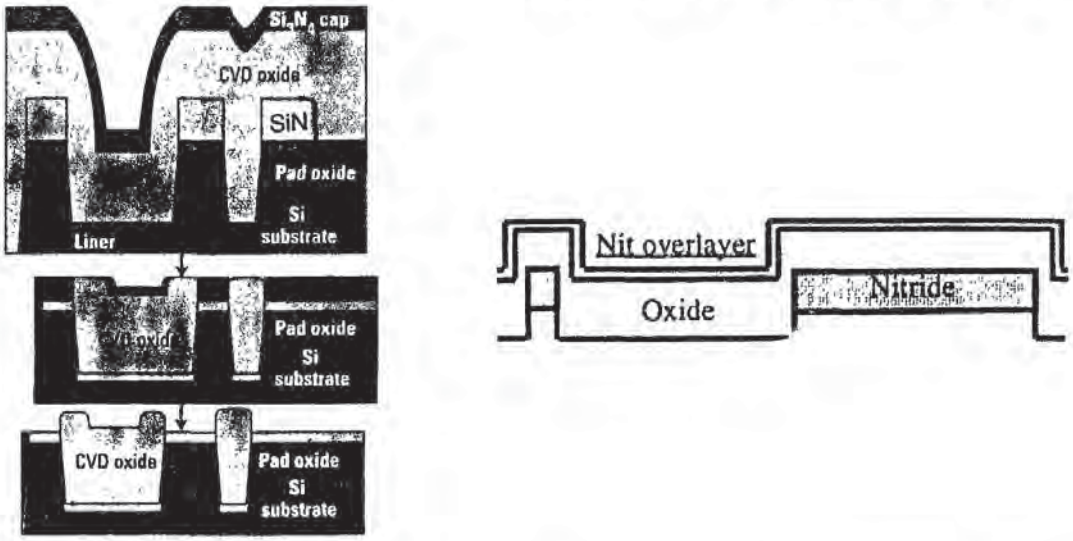



**Fig. 9-27** *Nitride cap* approach to minimize dishing during CMP of the trench-fill oxide.

oxide is used for the CMP process, and thus the thin-nitride-layer on the top of the oxide covering the active areas (which are "up" regions) is polished away. This exposes the oxide under the nitride in the active-areas to the CMP mechanism. However, the surface of oxide that fills the trenches is recessed below the surface of the oxide on the active-regions (as these are "down" regions). Hence, the SiN on these recessed-regions is not polished away when the areas of nitride on the active-regions are removed. This SiN thus remains on the wafer to act as a polish-stop-layer when the trench-oxide-thickness is reduced by the CMP process to the degree that it is essentially planar with the surface of the active-regions. In this way dishing over wide trench regions is decreased.

In Procedure **4**, the thin nitride layer is again deposited on top of the trench oxide surface before CMP, as is done in Procedure **3**.[44] However, in Procedure **4**, this thin nitride is also patterned with a mask such that after this lithography step is completed, the nitride on the active area regions is exposed (Fig. 9-28). This nitride is then etched, so that it is no longer present on the active-areas of the chip. However, the thin nitride does remain covering the trench regions, where it acts as a polish-stop-layer during the oxide-planarization CMP-step. Thus, dishing of the oxide in the trenches is again suppressed.

Use of the *direct-CMP method* avoids the added complexity of the previous methods used to combat the trench-dishing problem. That is, in the direct-CMP-approach the trench dielectric is etched directly after deposition and densification without applying any of the above procedures. To permit the trench-oxide to be directly polished, the CMP process itself must exhibit reduced tendency to dishing (putting the burden the CMP tool, the slurry, and the overall CMP process).

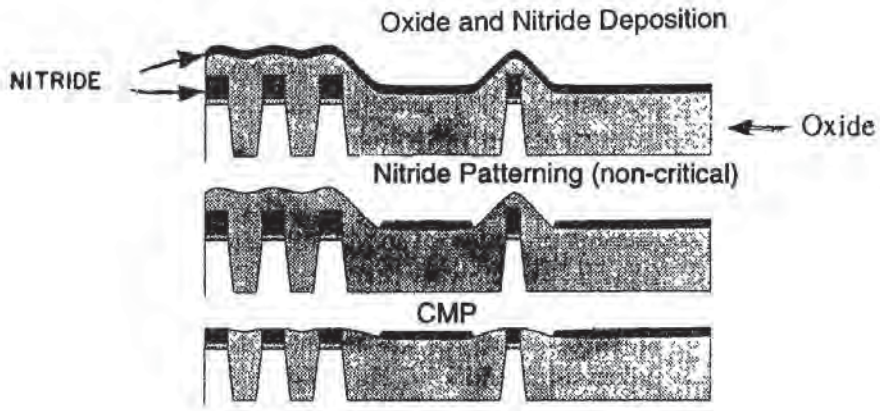



**Fig. 9-28** *Patterned second-nitride layer* approach to minimize dishing during CMP of trench-fill oxide.

The three methods being pursued to allow direct-CMP are: **1**) the use of *linear polishing tools*; **2**) the use of *rotary polishing tools with a slurry that exhibits relatively low selectivity between oxide and nitride*; and **3**) the use of *rotary polishing tools with a fixed abrasive pad.* Linear polishers employ higher pad velocities (300–500-ft/min versus 34 ft/min for rotary pad speeds) and lower polishing pressures (3–5 psi versus 7 psi for rotary polishers). In the linear polishing process, a *highly selective slurry* (with cerium oxide particles dispersed in an aqueous suspension of material that inhibits silicon-nitride removal) is also employed.[45] This set of process conditions significantly reduces dishing effects (maximum oxide removal across a die was 20-nm) and decreases the nitride removal over isolated active regions.[46,56] On the other hand, dishing is reportedly decreased when *rotary polishing tools* are used by selecting a different set of process conditions and CMP consumables. That is, decreased dishing effects are obtained by using a harder polishing pad, a relatively non-selective slurry, efficient endpoint detection method, and an optimized CMP process.[47,48] The use of a fixed-abrasive pad for the direct-CMP method of STI trenches has also reportedly shown good results.[49]

The degree to which a CMP process used to polish STI trenches causes dishing can be evaluated with the help of a special STI density- and pitch-pattern test-wafer developed by MIT.[55] These test-wafers have 4800-Å-deep trenches, 8000-Å-TEOS oxide deposited over 1960-Å of silicon nitride and 150-Å of thermal liner-oxide. Process engineers can use this test wafer to measure the effects of trench dishing and nitride erosion.

### 9.5.5 Corner Engineering and Its Effects on MOSFET TURN-OFF Characteristics of MOSFETs Fabricated with STI

As we described in Section 9.1, the relationship of the elevation of the isolation dielectric surface to that of the active region of the MOSFET can impact the TURN-OFF behavior of the MOSFET. Specifically, if there is a recess of the isolation-region dielectric with respect to the active-region-edge at the top corner of the trench-edge, the polysilicon gate will wrap around the corner as shown in Figs. 9-3 and 9-29.[51] This can cause the leakage current of the MOSFET in the subthreshold region of operation to increase above the value it would exhibit if the polysilicon and the isolation were at the same elevation (and there then would be no poly wrap-around).

In the sequence that described the baseline-STI structure we indicated that there are three HF-deglaze steps that typically occur in such a process flow. The effect of the latter two (see Steps **17** and **20**) is to cause the dielectric in the trench to be etched at the same time the pad oxide and sacrificial oxides are etched (in Steps **16** and **19**, respectively). The effect on the topography of the dielectric at the top corner of the trench is shown in Fig. 9-30, in which it is seen that a *divot* is formed in the trench oxide at the active-area edge.[52] This divot can have four related and detrimental effects on the active MOSFET. First, it causes poly wrap-around, that may lead to increased leakage of the MOSFET operating in the subthreshold region. As described in the baseline-STI process, however, if sufficient rounding of the top corner is implemented, the enhanced subthreshold leakage (as evidenced by the "double-hump" or "kink" in the MOSFET TURN-OFF characteristics) can be suppressed. Second, the wrap-around causes the MOSFET to behave like a MOSFET which has an isolation region recessed

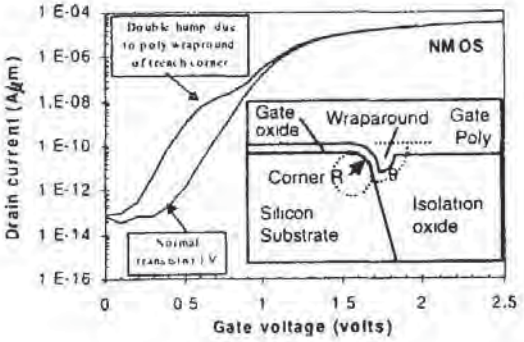



**Fig. 9-29** "Kink" in MOSFET $I_D$-$V_{GS}$ characteristic due to parasitic corner conduction caused by poly wraparound.

below the silicon surface. This enhances the so-called reverse-narrow-width effect, in which the threshold voltage decreases with MOSFET channel width, (see Fig. 9-31).[53] Note that this figure also shows that corner rounding reduces this effect, even if some poly wraparound occurs. Third, the threshold voltage of the MOSFET decreases as the depth of the divot is increased (Fig. 9-30). These effects produce unacceptably large device threshold voltage variations (± 200 mV). Fourth, the rounded edge encroaches into the active region of the device (i.e., under the nitride edge). This encroachment reduces the electrical width of the MOSFET. As the 0.1-$\mu$m technology node is approached, this width-reduction can represent a high percentage of the width of a minimum-sized MOSFET.

One reported technique that is used to combat this problem involves implementing a new corner rounding process that is accomplished after the nitride hard mask is opened (but before the silicon trench etch is performed). That is, a new lithographic technique from Clariant-AZ termed RELACS,[54] causes the trench width to be slightly reduced at the time the silicon nitride layer is opened. This is accomplished by applying a material to the resist pattern after it has been exposed and developed, and this material remains attached as a thin layer to the sidewalls of the resist pattern. This makes the openings of the resist pattern that define the silicon-nitride patterns slightly smaller (Fig. 9-32a). After the nitride layer has been opened, a short etch of the silicon trench surface undercuts the nitride edge, leaving a rounded top corner on the Si trench corner surface. The remainder of the trench is etched with an anisotropic etch. The overall trench profile after these etches is shown in Figs. 9-32b and 9-32c (here with the nitride left in place to show the location of the rounding with respect to the active area defined by the silicon nitride edge). The top corner rounding is clearly visible.

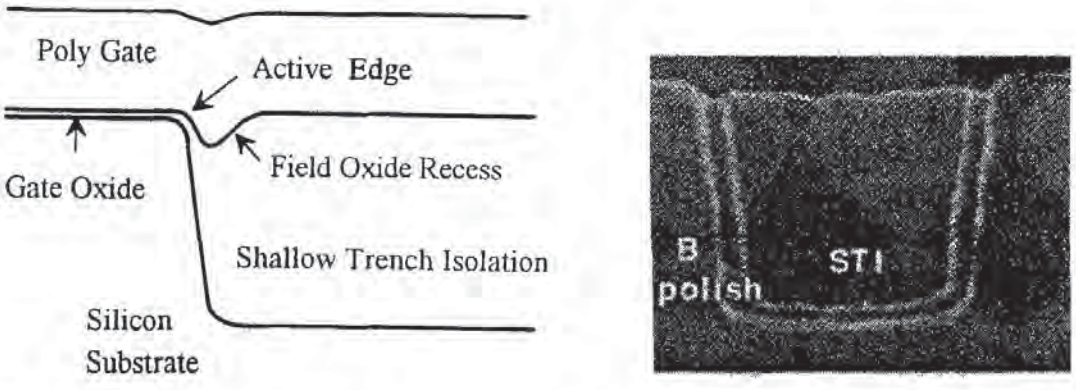



**Fig. 9-30** Field-oxide recess (or *divot*) formed at the active edge of an STI structure.

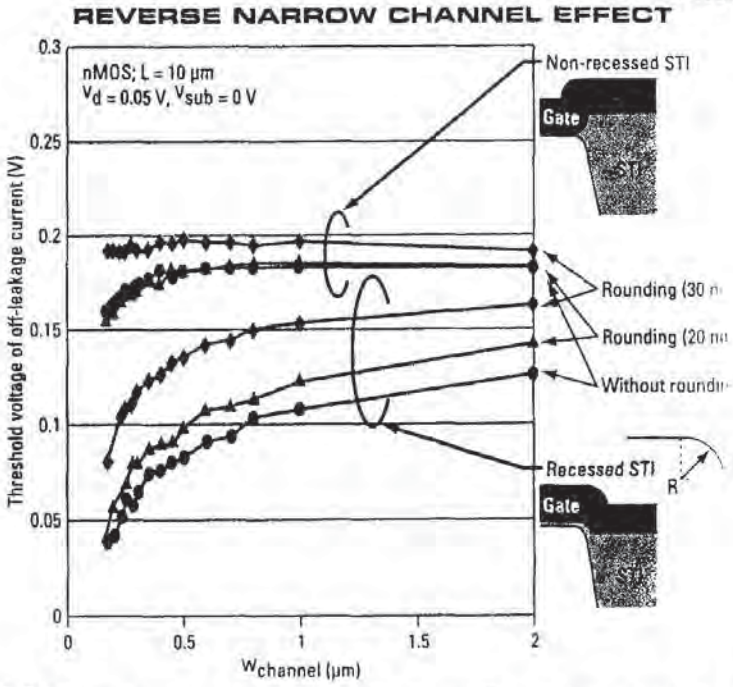



**Fig. 9-31** Corner rounding helps eliminate the reverse narrow channel effect that causes subthreshold leakage. Recess of the isolation region below the silicon surface exacerbates the problem.[26] Reprinted with permission of Semiconductor International.

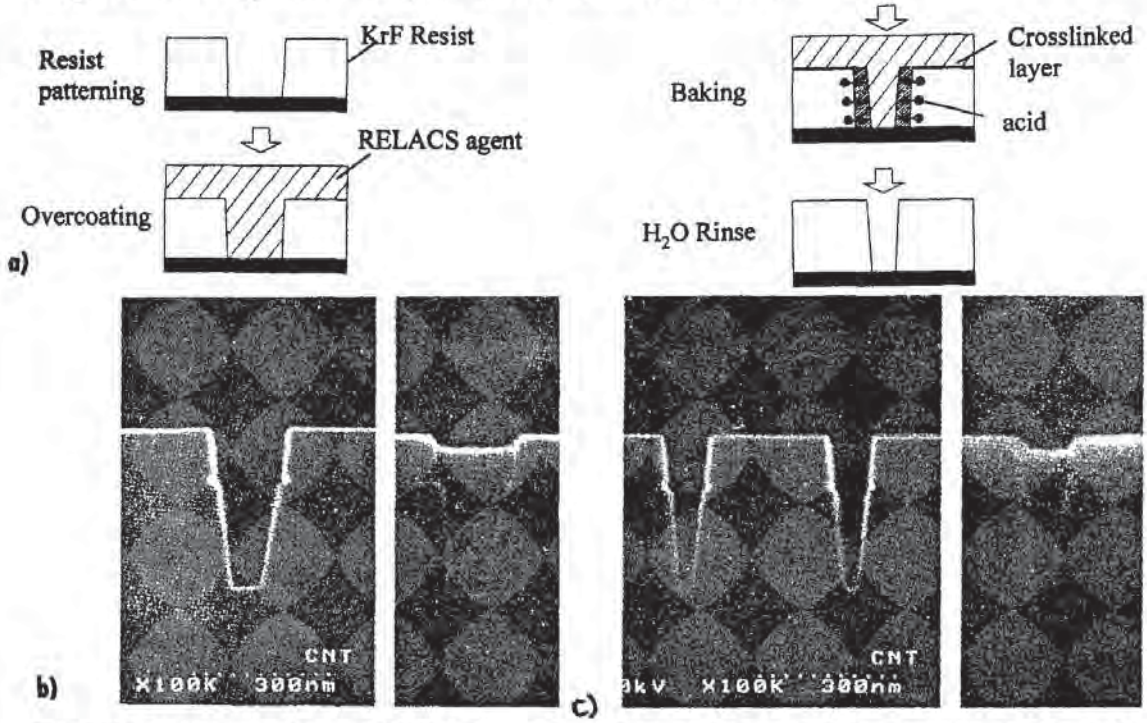



**Fig. 9-32** a) Process flow of RELACS. (b & c) Top corner rounding provided during etch in a TCP system. An *in-situ* hard mask/rounding/trench etch scheme was used. The top corner is rounded to ~15 nm radius. b) 180-nm-wide trenches as etched and after fill and CMP; c) 100-nmwide trenches as etched and after fill and CMP.[54]

By providing the top-corner-rounding during the etch, the encroachment under the nitride edge by the rounded corner produced in the baseline-STI sequence is eliminated, and the severity of the poly wrap-around phenomenon is also reduced. Furthermore, the corner is rounded without encroachment. Another benefit of this technique is that less variability in the threshold-voltage value is observed from-wafer-to-wafer. The reverse narrow-channel effect becomes more severe if the trench-oxide encroaches beneath the silicon-nitride edge.

## REFERENCES

**1.** E. DuBois *et al.*, "A Study of the Electrical Performances of Isolation Structures," *IEEE Trans. Electron Dev,*. **ED-37**(6), p. 1477, June 1990.
**2.** 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.
**3.** T. Shibata *et al*, *IEDM Tech. Dig.* 1983, p. 27.
**4.** D. C. Cole *et al.*, "The Use of Simulation in Semiconductor Technology Development," *Solid State Electronics*, **33**, No. 6, p. 591, 1990.
**5.** D. Foty, J. Mandelman, and T. Furukawa, "Behavior of an NMOS Trench-Isolated Corner Parasitic Device at Low Temperature," *Electrochemical Soc. Mtg. Abs.*, Fall, 1989, p. 295.
**6.** T. Furukawa, J. Mandelman, "Process and Device Simulation of Trench Isolation Corner Parasitic Device," *Electrochem. Soc. Mtg. Abs.*, Fall 1988, p. 329.
**7.** N. Shigyo *et al.*, "Three-Dimensional Analysis of Subthreshold Swing and Transconductance for Fully Recessed Oxide (Trench) Isolated 0.25-$\mu$m-Width MOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, July 1988, p. 944.
**8.** A. Bryant *et al.*, "The Current-Carrying Corner Inherent to Trench Isolation," *IEEE Electron Dev. Letts.*, **EDL-14**, August 1993, p. 412.
**9.** T. Iizuka *et al.*, "Double Threshold MOSFETs in Bird's-Beak Free Structures," *Tech. Dig. IEDM* 1981, p. 384.
**10.** K. Ohe *et al.*, "Narrow-Width Effects of Shallow-Trench-Isolated CMOS with $n^+$ Polysi Gate," *IEEE Trans. Electron Dev,*. **ED-36**, June 1989, p. 1110.
**11.** J.Y. Chen, "CMOS - The Emerging Technology," *IEEE Circuits and Devices Magazine*, March 1986, p. 16.
**12.** J.R. Brews, "The Submicron MOSFET," in *High-Speed Semiconductor Devices*, Ed. S.M. Sze, John Wiley & Sons, New York, 1990, p. 183.
**13.** K.M. Cham *et al.*, "Characterization and Modeling of the Trench Surface Inversion Problem for the Trench Isolated CMOS technology," *Tech. Dig. IEDM* 1983, p. 23.
**14.** *ibid. Ref. 12.*
**15.** *ibid. Ref. 12, p. 185.*

**16.** 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.

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

**18.** S. Ambadi *et al.*, "Impacting Device Performance and Yield Through Sacrificial Oxidation Improvements," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 713.

**19.** S. Lassig *et al.*, "An Integrated etch Approach as STI Evolves for the 100-nm Regime," *Solid State Technology*, July 2000, p. 157.

**20.** L. Peters, "Gaining Control Over STI Processes," *Semiconductor International*, March 2000, p. 74.

**21.** A. Ambadi *et al.*, "Development of Shallow Trenches with Round Bottoms and Factors Affecting Etch Profile and Uniformity," *Mtg. Abs. Electrochem.. Soc.*, Fall 1999, Abs. 1312.

**22.** N. Nandakumar *et al.*, "Shallow Trench Isolation for Advanced ULSI CMOS Technologies," *Tech.. Dig. IEDM*, 1998, p. 133.

**23.** N. Balasubramanian, "Gate Oxide Thinning in MOS Structures with Shallow Trench Isolation," *Mtg. Abs. Electrochem. Soc.*, Spring 2000, Abs. 454.

**24.** *ibid., Ref. 20.*

**25.** *ibid., Ref. 19.*

**26.** L. Peters, "Choices and Challenges for Shallow Trench Isolation." *Semiconductor International*, April 1999, p. 69.

**27.** *ibid., Ref. 22.*

**28.** A. Chatterjee, *J. Vac. Sci. Technol.*, 1997, p. 1936.

**29.** C.P. Chang *et al.*, "A Highly Manufacturable Corner Rounding Solution for 0.18-$\mu$m Shallow Trench Isolation," *Tech. Dig. IEDM*, 1997, p. 661.

**30.** A. Chatterjee *et al.*, "Shallow Trench Isolation Using LOCOS Edge for Preventing Corner Effects for 0.25/0.18-$\mu$m CMOS Technology & Beyond," *Tech. Dig. IEDM*, 1996 p. 829.

**31.** T. Sato *et al.*, "Trench Transformation Technology using Hydrogen annealing for Realizing Highly reliable Device Structure with Thin Dielectric Films," *Dig. Symp. VLSI Tech.*, 1998, p. 206.

**32.** T. Sato, "Novel Corner-Rounding Process for Shallow Trench Isolation Utilizing Micro-Structure Transformation of Silicon," *Tech. Dig. IEDM*, 1998, p. 137.

**33.** S. Nag *et al.*, "Comparative Evaluation of Gap-Fill Dielectrics in Shallow Trench Isolation for Sub-0.25 $\mu$m Technologies," *Tech. Dig. IEDM*, 1996, p. 841.

**34.** L.-Q. Xia *et al.*, "High Aspect Ratio Trench filling Using Two-Step Subatmospheric Chemical Vapor Deposited Borophosphosilicate Glass for <0.18-μm *Device Application," J. Electrochem. Soc.*, May 1999 p. 1884,

**35.** L.-Q. Xia *et al.*, "High Temperature Subatmospheric CVD Deposited Undoped Silicate Glass," *J. Electrochem. Soc.*, March 1999 p. 1181.

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

**37.** *ibid., Ref. 22.*

**38.** *ibid. Ref. 26.*

**39.** G. Badenes *et al*, "A New Dummy-free Shallow Trench Isolation Concept for Mixed-Signal Applications," *J. Electrochem. Soc.*, October 2000, p. 3827.

**40.** *ibid., Ref. 22.*

**41.** *ibid., Ref. 26.*

**42.** P. Tsui *et al.*, "A Manufacturable and Modular 0.25-μm CMOS Platform Technology," *Dig. Symp. VLSI Technology,* 1998, p. 152.

**43.** P. Singer, "The Future of Dielectric CVD: High-Density Plasmas?" *Semiconductor International*, July 1997, p. 127.

**44.** *ibid. Ref. 39.*

**45.** J.Y. Kim. "Novel STI CMP Process Using High Selective Ceria Slurry with Thin Nitride stopper," *Mtg. Abs. Electrochem. Soc.*, Spring 2000, Abs. 379.

**46**. B. Withers *et al.*, "Wide Margin CMP for STI," *Solid State Technology*, July 1998, p. 173.

**47.** *ibid. Ref. 26.*

**48.** *ibid. Ref. 26.*

**49.** J. Schlueter, "Trench Warfare: CMP and Shallow Trench Isolation," *Semiconductor International,* October 1999, p. 123.

**50.** R. DeJule, "CMP Challenges Below a Quarter Micron," *Semiconductor International,* November 1997, p. 54.

**51.** *ibid. Ref. 49.*

**52.** *ibid., Ref. 19.*

**53**. *ibid., Ref. 48.*

**54.** T. Toyoshima *et al.*, "0.1-μm-Level Contact-Hole Pattern Formation with KrF Lithography by Resolution Enhancement Lithography Assisted by Chemical Shrink (RELACS)," *Tech. Dig. IEDM,* 1998, p. 333.

**55.** Massachusetts Institute of Technology.

**56.** E. Zhao and C. Shan Xu, "Direct CMP for STI," *Semiconductor International,* June 2001, p. 145.

# Chapter 10

# SILICON-GERMANIUM (Si:Ge) TECHNOLOGY FOR HIGH-PERFORMANCE TELECOMMUNICATIONS ICs

Silicon-Germanium (Si:Ge) BiCMOS technology has become an important technology with many new and exciting product applications. Si:Ge-based heterojunction-bipolar-transistors (HBTs) are now found in a wide variety of technology offerings, and are incorporated in the product roadmaps of many telecommunications companies. In this chapter we describe the process technology issues of fabricating Si:Ge-HBTs and Si:Ge-BiCMOS ICs.

## 10.1 INTRODUCTION TO TELECOMMUNICATIONS IC APPLICATIONS

There is a rapidly growing market for wide-bandwidth and wireless telecommunications. That is, consumer demand for data from the Internet, for digital audio and video, and for video-on-demand is on the rise. In general, data is delivered to the consumer in two ways: *guided-wave propagation* and *free-space transmission*. Guided-wave propagation occurs along optical-fiber links and coaxial cables. Free-space transmission is the territory of radio, television broadcasting, satellite links, local multi-point distribution systems, and wireless local-area networks. (For example, such wireless applications as cellular phones currently operate in the 2.5-10-GHz range.) For free-space propagation, expensive installations are not necessarily required, but governments do regulate the use of bandwidth and frequency with a firm hand.

Since most of the *wireless* part of the frequency spectrum has been allocated, new or large bandwidth applications will be forced to resort to unallocated higher carrier frequencies. For example, a new wireless local-area network may have to operate at 60 GHz. Besides possible economy, a second attraction of higher frequencies is that they use smaller antennas. Because antenna efficiency is proportional to carrier frequency, receivers of higher carrier frequencies need only compact antennas. For instance, an antenna just 10-cm in length suffices for

collision-avoidance automotive radar at 77-GHz.

As for *guided-wave systems*, cable has been deployed, and once it is in place, bandwidth is relatively unregulated. In this kind of system (such as an optical-fiber data-link), the development of high-frequency components is mainly a response to the call for higher data rates. As optical-fiber trunk lines are more fully exploited, the next generation of communications channels will come to operate at 40, 80, or 100 Gb/s, well above the current 2.5-10 Gb/s. In fact, 100 Gb/s links could be in production as soon as 2010.

These data rates raise another obstacle. The signals must be received and processed at high speed and then eventually slowed down to match silicon CMOS processing speeds of, at best, 1.5-Gb/s. Such a receiving circuit requires complexity of at least the LSI level.

There is also a trend to higher operation frequencies in mobile communications and high-speed data-links. Future optical communications systems will also operate at over 10-Gb/sec and microwave/millimeter-wave systems will be used for mobile communication and intelligent traffic-control. The diverse range of applications in the broadband space includes *local multi-point distribution systems* (LMDS) that distribute information within major metropolitan centers via microwave transmission; *global-positioning systems* (GPS); *wide-* and *local-area networks*; *cellular base stations* and *handsets*,

**Table 10-1 Performance Characteristic Comparison of GaAs and Si:Ge-HBTs**

| Characteristic | GaAs HBT | Si:Ge HBT |
|---|---|---|
| **GAIN** | | |
| $f_t$, GHz | 50 | 30 |
| $f_{max}$, GHz | 70 | 62 |
| $G_{max}$ @ 2 GHz, dB | 20 | 23 |
| **NOISE** | | |
| $F_{min}$ @ 2 GHz | 1.5 | 0.5 |
| $G_{assoc}$ | > 20 | > 0.5 |
| 1/f corner | 1-10 | 0.1-1 |
| **POWER** | | |
| $V_{be}$, V | 1.4 | 0.85 |
| $V_{knee}$, V | 0.5 | 0.25 |
| **FEATURE SIZE** ($\mu$m) | 2.0 | 0.5 |

including *cellular, 3rd-generation* (3G) *personal-communications systems* (PCS) for wireless telephony; *asynchronous transfer mode* (ATM) *switching and multiplexing for information transfer*; and the *synchronous optical network* (SONET) protocols and systems for optical communication.

In 1990, conventional wisdom predicted that devices built using silicon-based technology could operate only at high frequencies below 2-GHz. As noted above, there are many applications in telecommunications, radar, and data transmission that need higher frequencies. It was felt that all of these applications would require devices built using the more expensive III-V-based materials. In fact, since that time such wireless and high-performance (wide bandwidth) applications have turned increasingly to gallium arsenide (GaAs) and other III-V compound-semiconductor technologies to fabricate the high-frequency devices that such applications require. However, in the wireless personal communications systems that use GaAs devices, designers have had to use many technologies for one system (such as GaAs technology for the power amplifier [PA], low-noise amplifier [LNA] and other rf components), and CMOS-silicon-VLSI-technology for the baseband operations. To overcome this issue, in 1992 IBM decided to develop a competitive, commercially-acceptable analog technology for such applications based entirely on silicon (that is to say, strictly speaking, on silicon devices fabricated with some added *silicon-germanium* [Si:Ge] layers).[36]

The premise for this technical and commercial breakthrough was to transform standard Si into a heterojunction semiconductor material by doping the Si lattice with germanium (Ge). Such a film can be created by epitaxially depositing such a Si:Ge layer onto a Si substrate. This epitaxial Si:Ge layer can be used as the base region of a *heterojunction bipolar transistor* (HBT). Table 10-1 compares the performance of such Si:Ge-HBTs to those made with GaAs. It can be seen that the maximum operating frequency of operation of Si:Ge-HBTs is close to those made from GaAs.

When IBM decided to pursue this path, they realized that there were many process integration challenges that would have to be met. However, if the venture was successful, it would become possible for silicon-based devices to compete with III-V semiconductor performance while simultaneously maintaining the low-cost advantages traditionally associated with Si. Furthermore, the intent was to make the ultra-high-speed Si:Ge-HBTs fully compatible with existing CMOS technologies. This would enable ICs to be designed with a BiCMOS process that integrated these ultra-high-frequency (bipolar) HBTs with CMOS on the same chip. In this fashion, a technology would be created that combines the radio-frequency (rf) front-end circuits with CMOS signal-processing circuitry. Since

such a BiCMOS technology would offer both high-speed HBTs and CMOS on the same chip, it would be ideally suited for providing the high-performance and low-power characteristics required by wireless applications. In addition, it would be possible to manufacture the products on 200-mm wafers, with low defect densities, and in mature silicon-IC production facilities.

The overall approach to establishing such a technology was to develop a process in which the fabrication sequence of the Si:Ge-HBT became a module that could be integrated into a state-of-the-art CMOS process with no change to the CMOS flow or parameters. If this could be possible, it would represent an attractive route to cost-effective, single-chip solutions for the wireless market. That is, the cost of manufacturing Si:Ge-ICs is only marginally higher than for ICs in Si alone, while the fabrication of ICs made of GaAs or InP is considerably higher. In addition, such an approach would also pave the way for achieving high levels of integration not possible with GaAs technologies. One of the pioneers of Si:Ge technology, Bernard Meyerson of IBM, was asked in 1999 about the future of Si:Ge technology. He replied, "In the next 3-5 years, the entire world will move from global systems for mobile communications to code division multiple access (CDMA) with added focus on high-speed data-links."[1]

### 10.1.1 Basic Research Underpins IBM's Si:Ge Bid

The decision by IBM to launch an IC product-line for wide-bandwidth and telecommunications applications based on Si:Ge-technology was bolstered by more than a decade of research that was performed at the IBM Watson Research Center on Si:Ge. The original intention when this research was launched in the mid-1980's was to extend bipolar technology for mainframe computing applications. While IBM eventually abandoned bipolar technology in favor of CMOS for this application, it nevertheless became apparent that due to their high-frequency performance and low-noise characteristics, Si:Ge *heterojunction bipolar transistors* (Si:Ge-HBTs) were also ideally suited for telecommunications applications. Thus, IBM decided to continue to fund the Si:Ge technology program, with the new application in mind.

IBM published the first papers on Si:Ge devices in 1989. The devices described in these papers demonstrated much better performance than conventional silicon devices. Maximum clock speeds were 75-GHz compared to 25 GHz for the fastest silicon devices. This high speed could be traded off for improved linearity and lower power. That is, a Si:Ge device running at 20-GHz requires only 10-20% of the power needed to run a conventional silicon device at the same speed. This reduction in power is critical to the operation of battery-

operated systems like cell phones. Operating at lower power improves the linearity, lowers the amount of noise generated, and increases reliability.

By 2000 Si:Ge-based HBTs were available with a below-10-ps ECL-gate-delay, and a cutoff frequency of over 100 GHz. In fact, with these operating characteristics, it is expected that Si:Ge-HBTs will gradually replace GaAs devices in such key antenna interface circuits as low-noise amplifiers, switches, and power amplifiers. Indeed, the development of Si:Ge-technology has opened the entire frequency range to silicon-based devices. That is, it is now possible to build complex high-speed systems on a chip which combine complex logic arrays, high-speed bipolar devices, and passive components. These ICs offer unmatched performance yet can be built as routine products in conventional commercial production fabs. Furthermore, essentially any new process step developed for silicon can be incorporated into Si:Ge-technology.

The first Si:Ge commercial process was qualified for production at IBM in 1996. This BiCMOS process revolved around an HBT with a maximum available power-gain frequency of 65-GHz, together with various active and passive support elements that allowed designers to produce a wide array of high-performance analog ICs.[2] The BiCMOS process used a pre-existing ASIC CMOS process to merge large CMOS digital logic-functions with a high-performance HBT. A variety of products have since been introduced using ICs built with this process. For example, in the second half of 1999, Alcatel announced a 10-Gb/sec SONET transceiver built with IBM's Si:Ge-technology. By the year 2000, IBM had developed its third generation of Si:Ge-technology, which includes 0.4-$\mu$m and 0.25-$\mu$m BiCMOS.[3,4] A cross-sectional SEM of a 0.2-$\mu$m Si:Ge HBT is shown in Fig. 10-1.[5]

By 2000 a large number of companies had joined the rush to carry out research, development, or production on Si:Ge-HBTs. While the adoption of Si:Ge ICs by 2001 is still at an early stage, the overall market for such products was rapidly growing due to the lower-cost, high-performance opportunity created at the boundary of Si and GaAs.[6] At the beginning of 1999, total sales of Si:Ge ICs was believed to be only $3.7 million. But, it is predicted that by 2005 the market for Si:Ge chips alone will exceed $1 billion, and the systems into which they will be inserted will have sales of between $5-$10 billion. Increasingly, participants in the analog segment of the IC industry are developing proprietary process recipes to compete for market share in this lucrative business.

However, there are some barriers that have limited the growth of the market for this technology (although, as we shall note, this situation is changing):

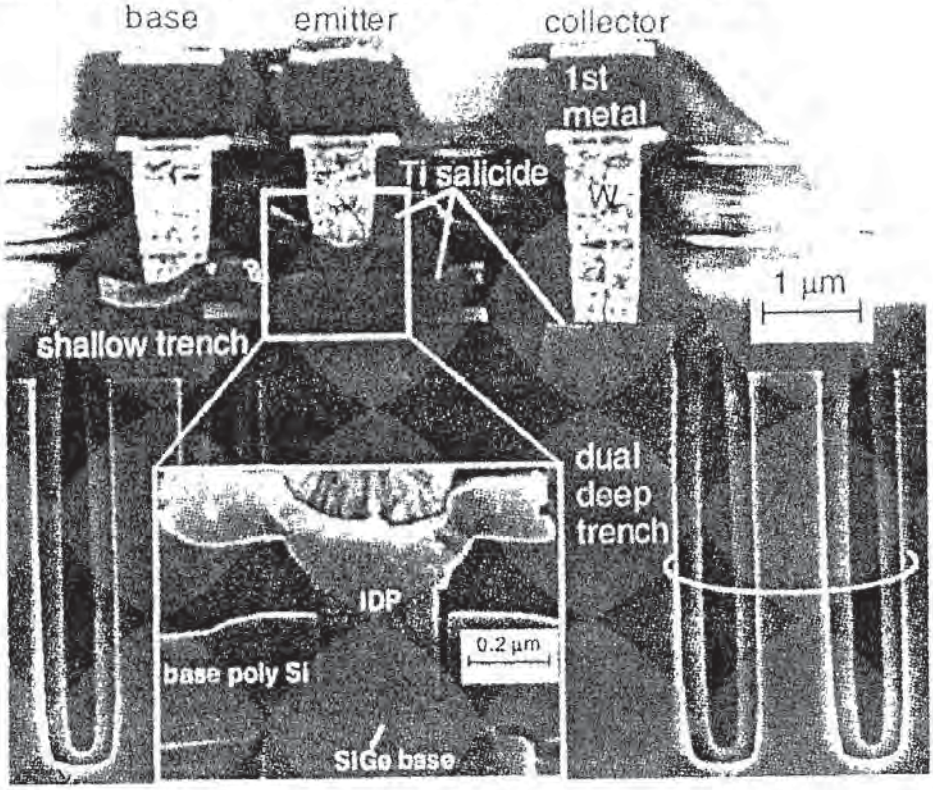



**Fig. 10-1** SEM cross-sectional view of a 0.2-$\mu$m Si:Ge-HBT.[5]

**1.** The intellectual property is narrowly held by only a few organizations.

**2.** Until recently, relatively few manufacturers or foundries had developed the process-technology expertise for manufacturing such devices. However, by late-2000, most state-of-the-art BiCMOS foundries had announced Si/Si:Ge-HBTs as their baseline bipolar technologies.[37] For example, the Taiwanese foundry, TSMC, announced in early 2001 that they had licensed Si:Ge technology from Conexant.

**3.** The devices being made with Si:Ge are still primarily bipolar HBTs. The lack of commercially-available Si:Ge CMOS structures is therefore also an impediment. However, development of MOS-based Si:Ge devices has been described.[7]

## 10.2 INTRODUCTION TO HETEROJUNCTION BIPOLAR TRANSISTORS

As the vertical dimensions of the bipolar transistor were being scaled, it was predicted that some serious device-operational limits would be encountered, even if such innovations as the use polysilicon emitters were implemented (see Vol. 2, *Silicon Processing for the VLSI Era; Process Integration - Chapter 7 - Bipolar and BiCMOS Process Integration* for a detailed discussion on bipolar transistor device physics and the evolution of bipolar transistor technology). The following

scenario illustrates this point: When the base-width of the bipolar transistor is decreased, it's doping level must be increased in order to control punchthrough, and to maintain a sufficiently low-value of base resistance. Since the doping level in the emitter must remain high in order for a sufficient emitter-injection efficiency to be maintained, the emitter-base capacitance $C_{EB}$ will also increase. The current density will thus need to be increased so that the propagation delay associated with this extra capacitance can be overcome. To handle the increased current density without suffering the Kirk effect, it will also be necessary to raise the doping in the collector. Consequently, even though such a device will have a higher $f_T$, it will have a narrower operating-current range. Furthermore, the increased doping levels will reduce the values of the emitter-base and collector base breakdown-voltages ($BV_{EBO}$ and $BV_{CEO}$, respectively), to such a degree that it will not be possible to use the device in many practical applications.

One technique that can be used to overcome the above limits is to build bipolar heterojunction transistors (HBTs). An HBT is a bipolar transistor in which the emitter, base, or collector use materials with different energy bandgaps. Bipolar transistor speed is determined mostly by the time a charge-carrier (electron or hole) takes to pass through the base of the device and the time the junction capacitances take to charge up. Thinner base-region layers, higher charge-carrier velocities, and higher current densities all increase the frequency of operation, as do lower parasitic capacitances and lower device resistances.

Recognized metrics for the speed of transistors are the following:

1. *Current-gain cutoff frequency* ($f_T$): This is the frequency at which the current gain of the active-device falls to 1. The highest clock rate at which the technology based on these active devices can operate is related to $f_T$. The value of $f_T$ is strongly influenced by minimum device dimensions in the direction of current flow and by device capacitances.

2. *Maximum frequency of oscillation* ($f_{max}$): This is the maximum frequency at which it is possible to make an oscillator out of a given device. It is also referred to as the *unity power gain cutoff frequency.* Unlike $f_T$ (which is influenced by minimum device dimensions in the direction of current flow and device capacitances), the value of $f_{max}$ is instead quite sensitive to parasitic elements external to the core of the device, such as the transistor contacts.

High values of both $f_T$ and $f_{max}$ are desirable for both digital and analog designs. While $f_T$ and $f_{max}$ values can be traded off against each other, balanced values are ideal for a general-purpose IC. Normally, small *gate lengths* in

MOSFETs and small *base widths* in bipolar transistors lead to high cutoff frequencies. If NMOSFETs are compared to Si:Ge bipolar transistors, it is found that NMOS devices with gate lengths of 0.18-$\mu$m have an $f_T$ of about 70-GHz and an $f_{max}$ of less than 40-GHz, while the most advanced Si:Ge-HBTs have cutoff frequencies in the 80-100-GHz range. It is predicted that Si:Ge-technology will be scaled to provide Si:Ge-HBTs with 150-GHz $f_T$-values by about 2005.[8] (Note that indium-phosphide [InP]-HBTs though not discussed here, have been fabricated with an $f_T$ of 250-GHz, and an $f_{max}$ over 1-THz [1000-GHz].)

## 10.3 HETEROJUNCTION BIPOLAR TRANSISTORS WITH LARGE-BANDGAP EMITTERS

One approach to fabricating HBTs is to build transistors with emitter materials whose band gaps are larger than that of the material used in the base. This technique appears promising because the emitter-injection-efficiency, $\gamma$, depends on the band-gap difference according to the equation:[9]

$$\gamma = (n_e \nu_e / p_b \nu_h) \exp(\Delta E_g / kT) \qquad \textbf{(10.1)}$$

where $n_e$ is the emitter electron-concentration, $\nu_e$ is the effective velocity of electrons injected into the base, $p_b$ is the base hole-concentration, $\nu_h$ is the corresponding hole velocity injected into the emitter, and $\Delta E_g$ is the difference in the band gaps of the emitter and base. From Fig. 10-2, it can be seen that the larger band gap in the emitter gives rise to a larger potential-energy barrier for holes going from the base to the emitter than for electrons going from the emitter to the base. Thus, if the band gap of the emitter in silicon devices is larger than that of the base, when $\Delta E_g$ is 0.2 eV or more, $\exp(\Delta E_g/kT)$ will exceed 2000. As a result, a sufficiently-high emitter-injection efficiency will be assured, no matter what doping levels exist in the base and emitter.

For example, doping levels in the base can be increased to above $10^{19}$ $cm^{-3}$, and in the emitter they can be dropped to $5 \times 10^{17}$ $c\,m^{-3}$ at the emitter-base interface (Fig. 10-3a1). Under such conditions, many other device characteristics are also improved (among the improvements are a lower base resistance, a higher punchthrough voltage, a higher Early voltage, and a decreased E-B junction capacitance). However, for such devices to be successful in silicon technology, the technology must be compatible with advanced silicon bipolar processing.

Most approaches to heterojunction fabrication are based on growing an epitaxial layer on the Si substrate, such that the epi material has a different band gap than the Si, but also has a lattice constant that is at least approximately

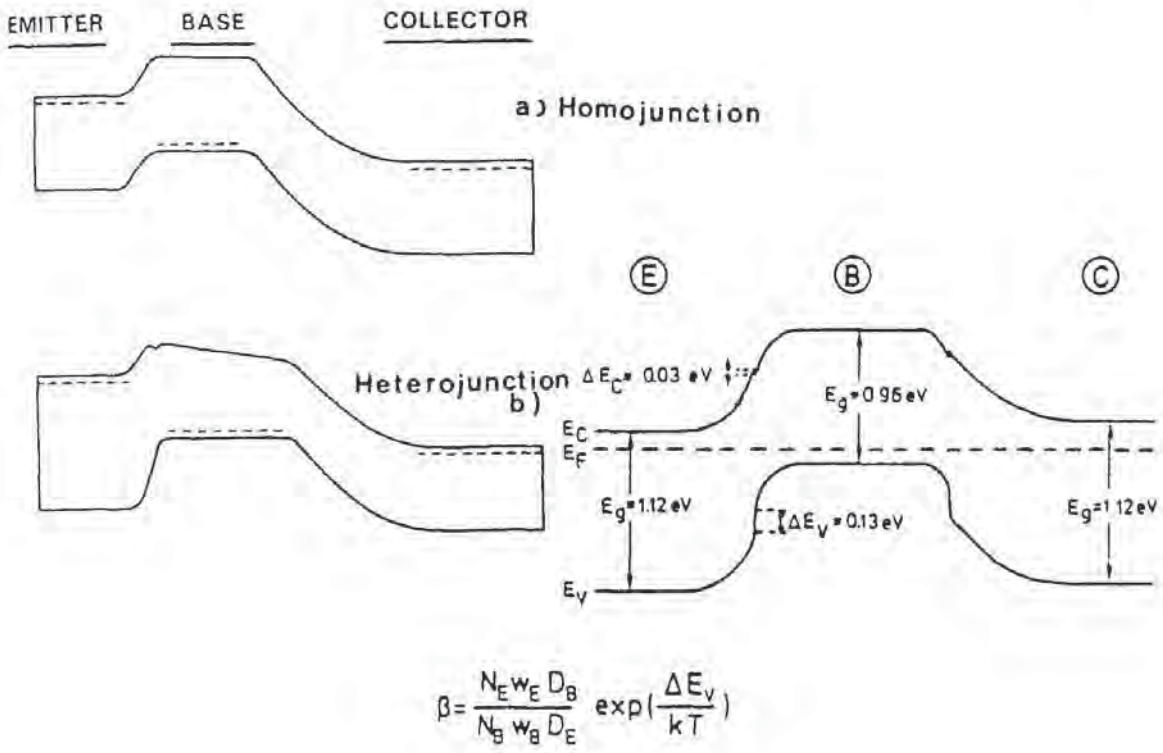



**Fig. 10-2** Band diagrams of representative *npn:* a) homojunction, and; b) heterojunction bipolar transistors.[9] (© 1989 IEEE)

matched to the lattice constants of the Si substrate. If such deposited layers are sufficiently thin, the requirement for exact lattice-match is relaxed. That is, thin layers with an almost-matched lattice constant will deform so that they can accommodate to the lattice constant of the substrate. In this deformed condition they are either "coherently strained" or "pseudomorphic" layers. The latter are thermodynamically stable for thicknesses below a strain-dependent "critical" value. However, it is important that subsequent processing temperatures be restricted to prevent the stress of the layers from being relaxed through the formation of an array of misfit dislocations.

From the point of view of process integration, the most attractive approach to building devices with hetero-E-B junctions is to increase the band gap of the emitter (since this represents the only change from a conventional bipolar process). Several methods for implementing this approach have been attempted, including the use of the following materials as the wider band-gap emitters: GaP;[10] semi-insulating polycrystalline-silicon (SIPOS, $E_g$ ~1.7 eV);[11] oxygen-

doped Si-epitaxial films (OXSEF, $E_g$ ~1.7 eV);[12] β-SiC ($E_g$ ~2.2 eV);[13] and phosphorus-doped hydrogenated micro-crystalline silicon (μc-Si, $E_g$ ~1.7 eV).[14]

Despite an intense effort to develop such processes, no completely successful fabrication procedure using an emitter with a larger bandgap was demonstrated. Although numerous high-gain transistors have been reported, these devices exhibited other problems, including high-temperature requirements, high bulk and contact resistivities, high emitter resistances, or poor stability during the multilevel-interconnect fabrication steps. The most promising candidate of this group was the μc-Si emitter. Fujioka *et al.* reported achieving reduced emitter resistances with such structures and built ECL gate arrays with them using a non-self-aligned emitter structure.[14] Basic gate delays of 295-ps at 6.5-mW/gate were obtained.

## 10.4 HETEROJUNCTION BIPOLAR TRANSISTORS WITH SMALLER BANDGAP BASE REGIONS (Si:Ge BASE)

An alternative approach to using a *larger bandgap* material for the emitter is to use a material in the base of the transistor that has a *smaller bandgap* than that of

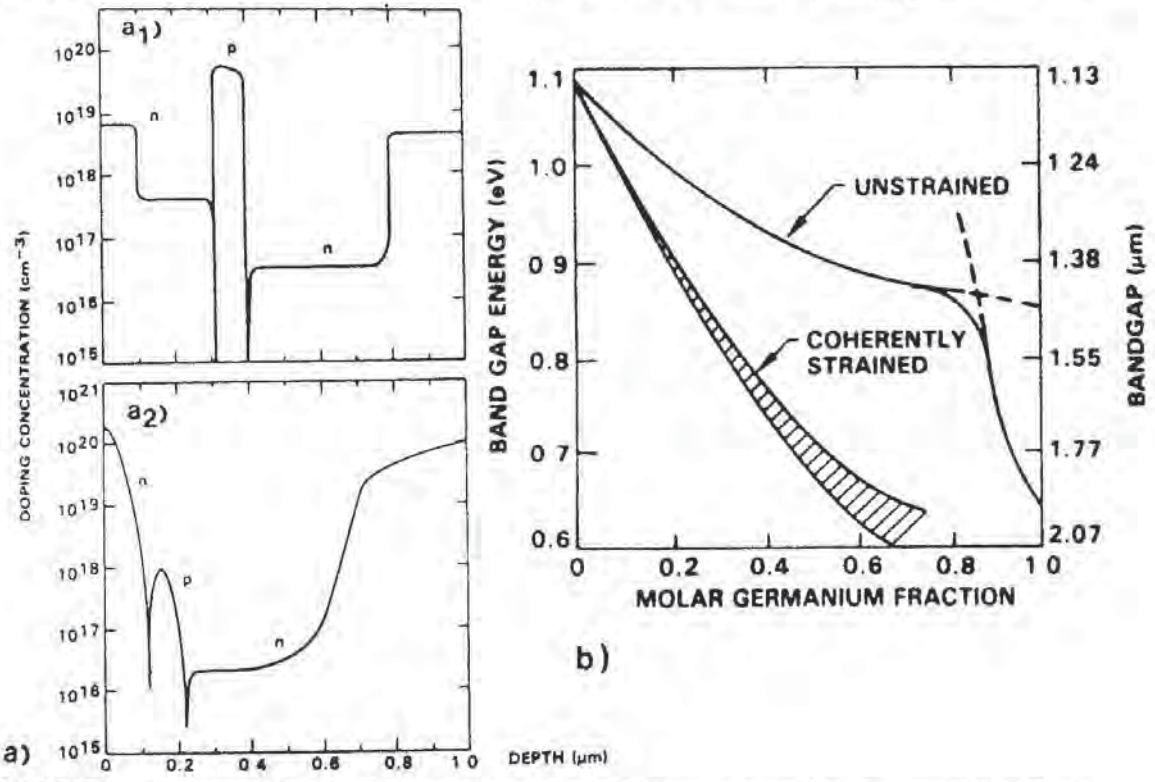



**Fig. 10-3** a) Representative doping profiles for npn heterojunction bipolar transistors.[9] (© 1989 IEEE); b) Band-gap energy versus alloy composition for Si:Ge alloys.[19] (after People and Bean Copyright 1986 AIOP)

silicon. Layers of Si to which Ge has been added exhibit such reduced bandgaps (Fig. 10-3b). For example, pseudomorphic epitaxial $Si_xGe_{1-x}$-alloy layers of up to 100-nm in thickness can be formed with a Ge content of 20%, and the $\Delta E_g$ between the emitter and base in Si transistors fabricated with such base layers (i.e., *$Si_xGe_{1-x}$-base-HBTs*) can be as large as 0.1 eV.[15]

Pseudomorphic epitaxial $Si_xGe_{1-x}$ layers can be deposited by means of molecular beam epitaxy (MBE) or limited-reaction-processing (LRP) CVD.[17] The most successful commercial CVD methods have involved ultra-high-vacuum-CVD (UHV-CVD),[16] or low-pressure CVD (LP-CVD). Note that appropriate levels of boron doping must also be incorporated into these films during the epi-growth process of the base-region in order to form the *p*-base of an *npn*-HBT. Only *npn-HBTs* are being pursued at present because *pnp*-HBT devices do not benefit substantially from the use of a Si:Ge base (due to a less favorable band-offset). We will describe the details of the UHV-CVD process used to form the Si:Ge-epitaxial-layers in the next section.

It should also be noted that to add a Si:Ge layer to a conventional Si device requires a special piece of semiconductor equipment (i.e., an MBE machine, or a special CVD reactor), which in 2000 cost approximately $2.5 million. However, the only other way to make a major performance improvement in a family of semiconductor devices is to shrink the geometry of the devices. Implementing a device shrink would typically require installing a whole new generation of process equipment and would cost more than $750 million!

HBTs with narrow-band-gap $Si_xGe_{1-x}$ bases are able to achieve very high performances (e.g., $f_T$'s as high as 100-GHz have been built with a 20%-Ge-concentration in a 100-nm base, a 0.2-$\mu$m-wide Si emitter, and a collector-current-density of $1x10^5$ A/cm$^2$).[18] Conversely, $Si_xGe_{1-x}$-base-HBTs offer the potential benefits of the same performance characteristics as those of conventional homojunction transistors, but with a relaxed profile (i.e., the base can be doped more heavily for a lower base-sheet resistance, or it can be made thicker to achieve lower doping levels with higher breakdown voltages).

The chief drawback of the reduced-band-gap-base approach is that it is more difficult to integrate into existing advanced bipolar processes. The metastable nature of the $Si_xGe_{1-x}$ films places constraints on the Ge content, layer thickness, subsequent annealing cycles, and thermal oxidation. On the other hand, this method does not suffer from the other problems exhibited when wide-band-gap-emitter materials are used. In addition, to provide the ultrathin base profiles for future bipolar devices, the conventional ion-implantation approaches are becoming increasingly difficult to implement successfully, while the low-

temperature epitaxial processes show great promise. We will discuss the process technology and process integration issues of Si:Ge ICs in the following sections.

## 10.5 PROCESS TECHNOLOGY FOR FABRICATING Si:Ge-HBTS

As noted in the previous section, Si:Ge heterojunction bipolar transistors have the same general structure as homojunction bipolar transistors, with the exception that they contain a Si:Ge layer in place of a Si layer in the intrinsic-base-region of the transistor. This Si:Ge layer is formed by epitaxial deposition onto a Si substrate, and a polysilicon layer is deposited onto the Si:Ge base-layer to serve as the polysilicon emitter. Since only *npn*-HBTs are fabricated (i.e., as noted earlier, *pnp* HBTs do not exhibit substantial performance improvement over their homojunction *pnp* counterparts, so there is no advantage to manufacturing them), the *p*-type intrinsic base is fabricated by also incorporating boron into the Si:Ge layer as it is being grown.

The use of an epitaxial layer as a specific structure of a bipolar transistor (i.e., the base region), is a significant departure from past uses of epitaxial layers in the fabrication of Si ICs. That is, in conventional (i.e., non-Si:Ge) IC fabrication, epitaxial layers have been used to form a lightly-doped Si layer on a heavily-doped Si substrate. That is, the epitaxial layer becomes a part of the *substrate* in which the device is formed, but *does not act* as a *specific device structure* itself. The fact that the epitaxial Si:Ge layer in an HBT is an active-device-structure-layer allows one of the fundamental limitations of bipolar device fabrication to be overcome, namely the fact that the base region of modern homojunction bipolar transistors must be formed using an ion-implantation process.

The ion implantation process produces a base-doping-profile with a skewed Gaussian distribution (due to the Gaussian-distribution of the as-implanted dopants, plus channeling effects associated with the ion-implant process). By using an epitaxial process to form the base region, it is theoretically possible to grow a *uniformly-doped* layer with atomically abrupt boundaries, and all of this doping can be confined within the Si:Ge layer. In practice, such atomically abrupt boundaries are not possible, but one can create a boron-doping profile within the Si:Ge layer that is much more uniformly doped than are the base regions of bipolar transistors created with an ion-implantation process. Note that the Si:Ge epitaxial layers must be grown at temperatures low enough that the Si:Ge layers retain their pseudomorphic structure (500-550°C). Furthermore, by performing the epitaxial-deposition process at such low temperatures the outdiffusion of the dopant from the layer during the epitaxial growth process is essentially suppressed, keeping the boron well-confined within the Si:Ge layer.

A low-temperature-epitaxy (LTE) technique using ultra-high-vacuum chemical vapor deposition (UHV/CVD) in a multi-wafer reactor is the most mature process used to grow epitaxial layers of Si:Ge-alloys for use in HBTs.[21] It is reported that device-quality epitaxy can be achieved at a growth temperature of 550°C. This LTE technique originated in the early 1980's with Meyerson's work at IBM on a novel UHV/CVD reactor. The Si:Ge epitaxy process has two stages: **1**) preparation of the Si substrate surface so that an epitaxial layer can be grown on it at a low temperatures; and **2**) growth of the layer itself.

To enable growth of an epitaxial layer on a Si surface, the surface must be free of native oxide and other contaminants. To eliminate the need to use high temperatures it is necessary to ensure that the surface of the Si is free of such species prior to initiating the epitaxial-growth-stage of the process. A simple wet chemical procedure is carried out to remove any such native oxide from the Si surface and at the same time to passivate the wafer surface with hydrogen. This process entails etching the wafer for 10-15-sec in dilute (10:1) $H_2O$/HF. The layer of hydrogen adsorbed on the Si surface during this wet etch acts to retard the growth of native oxide on the Si surface. That is, after the 10-sec HF dip, there is a passivation layer consisting of hydrogen-terminated bonds across the silicon surface. The presence of the hydrogen reduces the reactivity of the growth interface by approximately 13 orders of magnitude from that of a bare-silicon surface with respect to its oxidation rate in ambient air. Since the wafer surfaces are oxide free and passivated as-pulled from the HF dip, the wafers may be loaded directly into the growth chamber (Fig. 10-4) after this HF-last step. This

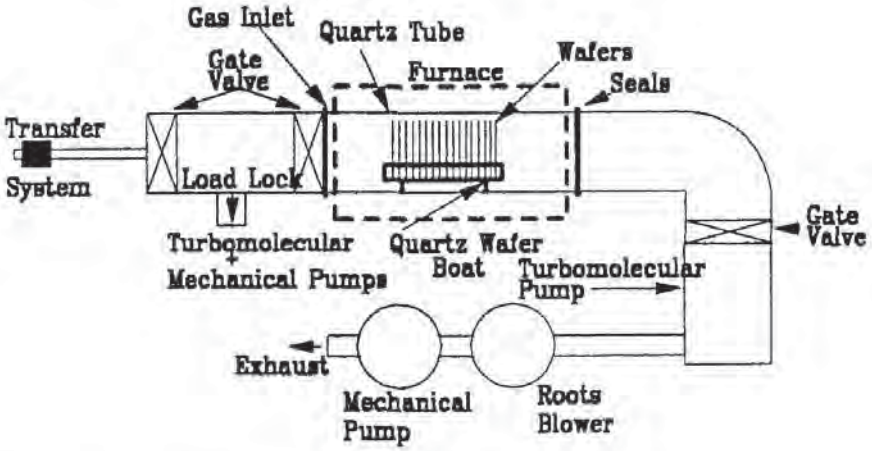



**Fig. 10-4** Cross-section schematic of a UHV/CVD reactor.[15]

surface-preparation step has been shown to reliably keep the Si surface free of native oxide until the growth process can be begun (provided the temperature/time history of the load-sequence prevents hydrogen desorption until growth is initiated). Note that bare-silicon wafers removed from an HF bath are hydrophobic (i.e., they do not wet). Thus, the surface readily de-wets after being pulled from the HF. This makes it easy to eliminate residual HF from such a surface, allowing this removal process to be automated.

The batch of wafers (typically 25) is introduced into the loadlock of the UHV-CVD tool. After a pumpdown to $10^{-6}$ torr, the wafers are transferred into the growth chamber under an $H_2$ flow and growth is commenced immediately. The growth chamber is pumped down to a base pressure of $10^{-9}$ torr (UHV), and the precursor gases (silane, germane, and diborane) plus hydrogen are introduced. Films are deposited at temperatures in the range of 500-600°C. During deposition the pressure is maintained at $10^{-3}$ torr. The growth rate during this process is only about 0.4-4.0-nm/min. This very-low growth rate provides precise control of the layer thickness (on the order of 1-2 atomic layers in this instance), doping concentration, and formation of a pseudomorphic strained SiGe layer. Versions of this system are available for processing 200-mm wafers (based on IBM designs), from Unaxis (formerly Balzers Process Systems) and CVD Equipment Corp (Fig. 10-5). To ensure these tools are robust and productive in a manufacturing environment, good vacuum practice and contamination-prevention methods must be incorporated when designing them. Note that these reactors are batch systems that process up to 25 wafers at a time. The industry, however, is moving toward single-wafer processing, and single-wafer Si:Ge deposition tools are thus also available.

High-quality Si- and SiGe-epitaxial-films can be formed at such low temperatures for the following reason.[36] The hydrogen-terminated silicon wafer surface desorbs its hydrogen at temperatures between 500-600°C. The silane ($SiH_4$) that is fed into the reactor chamber and that adsorbs on the wafer surface also decomposes at these temperatures. Of the products of this decomposition, silicon incorporates into the crystal, and the hydrogen remains adsorbed on the silicon surface, replenishing the hydrogen that was initially there. Thus, at these temperatures the silicon-epitaxial-layer grows while maintaining a hydrogen-passivated surface. Thus, the surface is protected from ambient oxygen contamination, allowing high quality epitaxial layers to form. (Note that silicon can undergo solid-phase-recrystallization at reasonable rates in the range of 500-700°C). However, at higher temperatures (> 650°C), the hydrogen passivating film is removed instantaneously, allowing even low residual oxygen content in

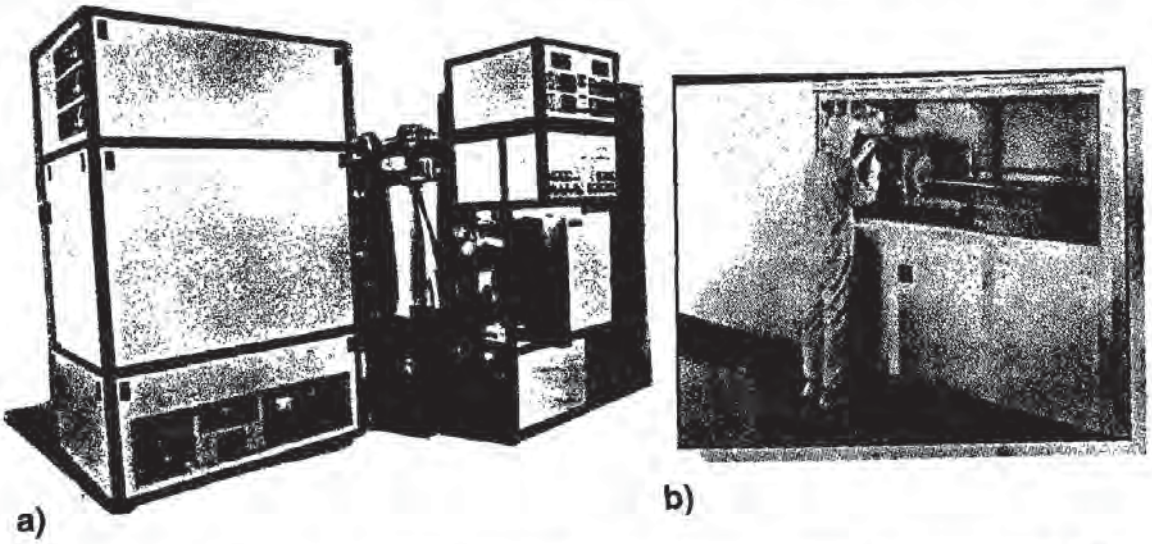



**Fig. 10-5** a) Commercial UHV/CVD Si"Ge deposition system. Courtesy of CVD Systems; b) Commercial Low-energy plasma-enhanced CVD system for Si:Ge deposition. Courtesy of Unaxis Ltd.

the growth ambient to cause the formation of native oxide on the silicon surface. This causes high defect densities in the epitaxial film. As a result, to obtain high-quality films when silicon wafers are heated to temperatures greater than 650°C (initially hydrogen-passivated or not), a high-temperature pre-treatment step just prior to epitaxial growth is needed to remove the now-present native oxide.

Single-wafer, ultra-high-vacuum, rapid-thermal-CVD (UHV/RTCVD) has been reported as a means for achieving LTE.[22] An advantage of this single-wafer technology is that the process can be initiated and terminated by rapid changes in the substrate temperature, rather than by changing gases. Figure 10-6 shows a schematic of such a system.[37]

Conventional single-wafer epi reactors can also be operated with an LP-CVD process at low temperatures (about 600°C) to produce strained-layer Si:Ge-alloys at a deposition-rate of about 10-nm/min. Note that in single-wafer, load-locked, RTP tools, the surface-cleaning of the Si-substrate prior to the Si:Ge-epitaxial-growth stage is carried out by a high-temperature RTP step which sublimates SiO present on the wafer surface (instead of relying on an HF-etch as is done in the batch-UHV/CVD process). It must be rigorously ensured that the ambient during both the cleaning stage and the growth stage of the epitaxial-growth process in the single-wafer reactor is kept completely free of oxygen and moisture. This is accomplished by using a loadlock, point-of-use gas-purification, and ultra-clean gas handling. Two vendors that supply single-wafer reactors in which Si:Ge can be deposited by LP-CVD are ASM International[23,24] and Applied Materials.[39,40]

A substantial number of these reactors are reportedly being sold to the industry.[25] ASMI claims they have 75% of the installed Si:Ge-reactors in 2000 (ASMI Epsilon E2000®).

Etching the Si:Ge layers also presents some challenges. While Si:Ge can be etched with plasma processes that use chemistries similar to those used to etch Si, the fine features and profiles required to maintain high-frequency performance, place increased demands on the processes. Novel chemistries can be an aid, and endpoint-detection plays a vital role. Figure 10-7 shows a super-self-aligned selectively-grown-epitaxial Si:Ge-base transistor.

The use of *selective-epitaxial-growth* (SEG) to grow the Si:Ge-base of the HBT has been pursued to achieve higher $f_T$ and $f_{max}$ values. For example, a double-polysilicon self-aligned-HBT with a low-resistivity Si:Ge-base was fabricated having a 88-GHz $f_T$-value and a 78-GHz $f_{max}$-value. Si:Ge-HBTs with such performance characteristics would make then suitable for wireless communications up to 10-GHz and high-speed data-links at 40-Gbit/sec and above.[26] Another report described how such a Si:Ge-SEG-process was used to build a 0.2-$\mu$m, self-aligned Si:Ge-HBT that was then integrated into an SOI BiCMOS technology. The base Si:Ge layer in this HBT was deposited by UHV/CVD. The HBT in this report had an $f_{max}$ of 180-GHz.[27]

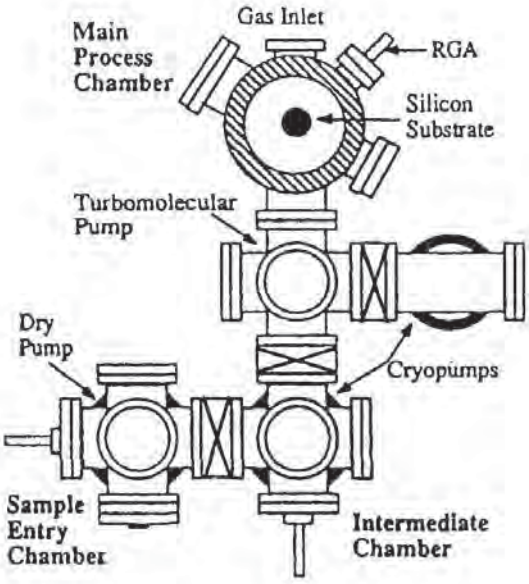



**Fig. 10-6** Schematic of a UHV/RTCVD epitaxial system.[22] Reprinted with permission of the American Physical Society.

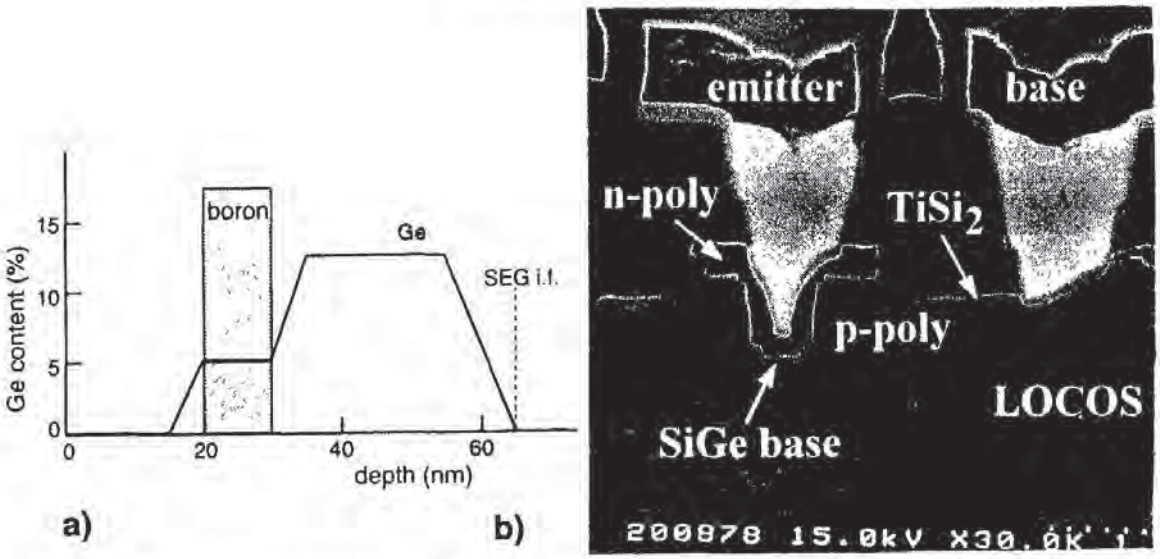



**Fig. 10-7** a) Schematic of the Ge profile in the Si/Si:Ge layer of a Si:Ge-HBT;[27] b) SEM cross-section of the emitter-base complex of a 0.2-$\mu$m Si:Ge-HBT.[26]

## 10.6 PROCESS INTEGRATION ISSUES FOR BiCMOS TECHNOLOGY WITH Si:Ge-HBTs

The previous section dealt with the process technology of forming epitaxial layers of Si:Ge on Si-substrates. Here we consider the issues involved with integrating such an epitaxial-growth-step into a CMOS process so that HBT structures can be fabricated that are integrated on the same chip as CMOS circuitry. That is, a BiCMOS process is implemented instead of a CMOS process. The goal of creating such a BiCMOS technology is to end up with a process that results in no change (or minimal change) to the standard-CMOS flow, and that the electrical characteristics of the CMOS in a BiCMOS structure matches the baseline-CMOS as closely as possible (and conversely, that the CMOS process steps do not degrade the performance-characteristics of the HBT).

As noted earlier, basic physics explains that when a Si:Ge-lattice is formed, the atomic size difference between Si and Ge atoms necessarily results in the formation of a layer that will create a lattice-mismatch if the layer is formed on a Si substrate acting as a growth template. This lattice-mismatch must be dealt with when integrating this layer with the overall fabrication sequence. While a certain level of strain is required for the bandgap-reduction of the Si:Ge-layer (and this bandgap-reduction is a driving force behind HBT performance improvement), the strain must be kept low enough to avoid a generation of lattice defects and unstable Si:Ge films. This limitation has been extensively studied and the Matthews and Blakeslee curves relate film-thickness to strain in a Si/Si:Ge

system.[34] To allow the Si:Ge-layer to be integrated with the remainder of the processing steps used to fabricate standard CMOS, this implies that the Si:Ge-strained-layer must be able to tolerate total emitter-thermal-cycles such as an 850°C furnace-anneal followed by a short 900-1000°C rapid-thermal-anneal.[28]

## 10.7 HIGH-QUALITY Si:Ge EPITAXIAL FILM GROWTH ON PATTERNED WAFERS

One of the most significant issues facing the integration of an HBT device into a CMOS process is that of being able to deposit high-quality epitaxial films onto wafers containing pre-existing oxide-isolation-patterns. The use of wafers patterned prior to epitaxial deposition is the process factor that separates this approach from the approach that is used to fabricate discrete HBTs or mesa bipolar transistors that contain a Si:Ge-epitaxial- base-region. The problem is that the parts of the wafer covered by oxide are hydrophyillic (these surfaces *do* wet). The wetted surface makes it difficult to remove residual-HF without a DI-water rinse-and-dry step. But, such a water rinse would degrade the passivation quality of the hydrogen-termination layer present on the regions of the wafer where silicon is exposed (through hydrolysis of less stable F-terminated surface and/or kink sites which cannot be ideally passivated).

A pragmatic two-step solution when depositing Si:Ge-films in the UHV/CVD reactor is to first assure that the entire exposed wafer surfaces (front and back) are composed of silicon. This is done by depositing a by a polysilicon "protect" layer on the wafer. Then windows are cut into this layer where epitaxial growth is required. The polysilicon layer also serves to protect the field isolation and/or critical oxides from HF attack during the HF-passivation dip, and to reduce the overall moisture load in the UHV/CVD chamber. As a result, the basic approach is to provide a substrate which resembles as closely as possible the ideal of an all-silicon, hydrogen-passivated surface of a bare Si-wafer.

## 10.8 THERMAL STABILITY PROCESS INTEGRATION ISSUES

It was noted previously that it must be possible to maintain a nearly defect-free Si:Ge-strained-layer on Si throughout the entire BiCMOS process. Although unstrained Si:Ge layers also exhibit significant bandgap-reduction, the presence of an extensive misfit dislocation network in the material would preclude the fabrication of high quality bipolar devices. However, by adopting effective epitaxial growth techniques and optimizing the Si:Ge-strained-layer characteristics, it is possible to minimize strain-induced-defects. It turns out that Si:Ge films with doping levels of Ge up to 13% are able to maintain their

mechanical stability under annealing conditions of up to 900°C for 30 min, making it feasible to use such Si:Ge layers in BiCMOS processes (provided the layers are not subjected to temperatures and times above this limit).

## 10.9 COMPATIBILITY WITH ION-IMPLANTATION PROCESSES

There have been some questions concerning the compatibility of a Si:Ge layer with subsequent ion-implantation steps. That is, one might wonder if introducing significant amounts of lattice damage to a strained layer through implantation would degrade its integrity. Such implantation damage could provide a source of defect nucleation sites, which in turn could lead to deleterious formation of an extensive misfit dislocation network upon annealing. It turns out, however, that Si:Ge layers which are mechanically stable are apparently able to undergo implantation and anneal cycles with no additional defect-generation or movement beyond that expected in single-crystal silicon.

The two processes likely to cause problems are the moderate-energy, low-dose implants used to selectively implant dopants for the sub-collector of the HBT (phosphorus, at 100-keV, $1x10^{13}/cm^2$), and the high-dose, low-energy implants used to form the extrinsic base of the HBT and the source/drain regions of the PMOSFETs (boron, 5-keV, $1x10^{15}/cm^2$). The effects of these implants on the Si:Ge-layers was examined, and it was found that no significant yield impact was observed. A examination of the Si:Ge-strained-layer after ion implantation was studied using transmission electron microscopy (TEM). There was no evidence of misfit-dislocation formation within the device regions.

## 10.10 COMPATIBILITY WITH THERMAL OXIDATION PROCESSES

It must also be possible to oxidize exposed regions of Si to create an insulator in the integrated BiCMOS process. Studies indicate that the presence of Ge in the Si dramatically enhances the oxidation rate, and that Ge segregates to the growth-interface during oxidation. Since local-oxidation is generally used to provide electrical junction isolation, the Ge "pile-up" effect is likely to lead to high levels of leakage and potential reliability concerns under electrical stress. A generally accepted way to control both of these effects is to use an all-Si capping-layer for oxide growth and isolation purposes.

A low-thermal-budget CMOS process is required for thermal-cycle compatibility with a narrow-base *npn* process. After the base-epitaxial-layer is grown, the thermal-cycle should be kept to a minimum, ideally consisting of only the bipolar *npn*-emitter-cycle, which entails a furnace anneal and an RTA anneal. The two largest post-gate-etch thermal-cycles in a CMOS process are the

polysilicon reoxidation step, and the activation anneal performed after the source/drain implantation step. Both of these steps in conventional CMOS are typically furnace anneals at temperatures > 900°C, which causes significant diffusion of the base. However, the poly reoxidation step has been recently carried out at somewhat lower temperatures (850°C), and short RTP anneals have generally been replacing furnace S/D anneal steps. As the Si:Ge processes have been evolving through several generations, the thermal budget issues have been handled differently. We will discuss this subject again in Sects. 10.11 and 10.12.

## 10.11 VERTICAL SCALING OF Si:Ge HBTS

Using Si:Ge and bandgap-engineering, bipolar transistors will continue to provide circuit designers with the performance necessary to meet the stringent circuit and systems requirements of the next decade. In demonstration of this trend, companies are leveraging the learning from 50-GHz-$f_T$ Si:Ge-devices to develop 75 to 90-GHz-$f_T$ Si:Ge-devices. Continued progress in process and device design will certainly push performance of the Si:Ge-HBT to higher levels in the future.

Higher $f_T$ performance is principally achieved through the vertical scaling of dopants and dimensions in the HBT device. Such scaling primarily involves: **1**) a reduction in transit time through the base and collector by reduction of the vertical-layer dimensions; and **2**) a commensurate increase in Ge-grading across the (now thinner) neutral base.

Key to vertical dimension reduction is the minimization of thermal cycles in post-base-deposition processing. The modular integration trend exemplified by the base-after-gate process sequence (described in Sect. 10.12) provides one important aspect to obtaining vertically scaled HBTs. In addition, movement towards RTP processing is favorable to achieving low overall thermal budgets.

Another technique for reducing the base-width-dimension is to incorporate carbon into the base epitaxy. Carbon incorporation at concentrations lower than $10^{20}/cm^3$ has been shown to significantly reduce B diffusion (by lowering the Si self-interstitial concentration within the C-rich regions).[37]

The choice of emitter dopant is closely linked to the device thermal-cycles. Various emitter dopants have been described in the literature, including implanted arsenic or phosphorus emitters, and *in situ*-doped phosphorus emitters. Because thermal-cycles required to diffuse and activate arsenic emitters will adversely impact base-widths as thinner bases are used, *in-situ*-doped phosphorus emitters will become prevalent in future Si:Ge-HBT device designs targeting greater than 100-GHz-$f_T$.[38]

## 10.12 BASE-DURING-GATE AND BASE-AFTER-GATE BiCMOS PROCESS FLOWS

As noted earlier, the goal of producing a BiCMOS process that integrates HBTs into a CMOS process flow is to ensure that the electrical characteristics of the CMOS devices remains as close to those obtained in the CMOS process alone. Therefore, the process must maintain the thermal excursions needed to produce the CMOS structures. In the first two generations of BiCMOS technologies developed at IBM this was accomplished by sharing layers and thermalcycles. That is, the HBT was fabricated *during* the fabrication of the CMOS layers, and this sequence was termed *base-during-gate* (BDGate). A flow diagram of the base-during-gate process is shown in Fig. 10-8.[29] To avoid significant widening of the base during the post-base-deposition thermal-processes of poly-reoxidation and NMOSFET-S/D-annealing, a low-temperature poly-reox-step was used, together with a low-thermal-budget CMOS-process technology. The BDGate flow worked well down to the 0.5-$\mu$m-BiCMOS technology node, but it was difficult to extend it into the next, scaled generations. This was because the thermal-cycles needed in the next generations became too high for the HBT-device.

As a result, a new flow, termed *base-after-gate* (BAGate) was developed. In the base-after-gate sequence the thermal-cycles for the CMOS-devices and the HBTs were decoupled. This is achieved by building the HBT after the CMOS is fabricated. That is, the CMOS is largely fabricated in one complete block before the Si:Ge-base is deposited. Therefore, the CMOS-technology steps can be completely copied without modification. The bipolar-transistor is fabricated in a complete block after the CMOS without sharing any thermal-cycles (except for a final emitter drive-in RTA). Since the HBT-processing is performed at low temperatures, thermal effects on the FETs are minimal. A flow diagram of the base-after-gate process for the 0.25-$\mu$m-BiCMOS process is given in Fig. 10-9. By decoupling the fabrication of the CMOS- and HBT-structures, the migration from generation-to-generation becomes easier.[30,35,38] The modularity of the BAGate process supports rapidly developing derivatives and low-cost alternative BiCMOS technologies. The baseline-CMOS processes are also incorporating more analog elements (capacitors, inductors, and resistors) because of the expanding use of CMOS for analog products. This trend is favorable for BiCMOS technology because fewer processing steps need to be added to the CMOS to have an analog-Si:Ge-BiCMOS process.

Both processes employ a planarized deep-trench-plus-shallow-trench isolation-structure fabricated with the aid of CMP technology. The Si:Ge is blanket deposited so that it is single-crystal over the exposed silicon-regions and polysilicon over the field-oxide isolation-regions. Standard CMOS processing is

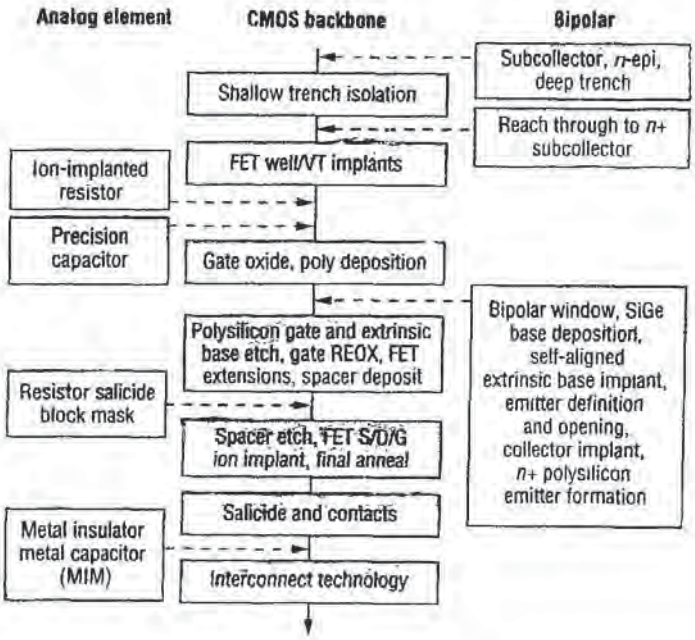



**Fig. 10-8** CMOS process flow for fabricating *Base-During-Gate* Si:Ge BiCMOS structure.[29]

used, including the use of titanium salicide. The metallization is planarized Al-Cu metal, patterned by RIE, with tungsten interlevel plugs. Fig. 10-10 shows a schematic cross-section of this BiCMOS structure.

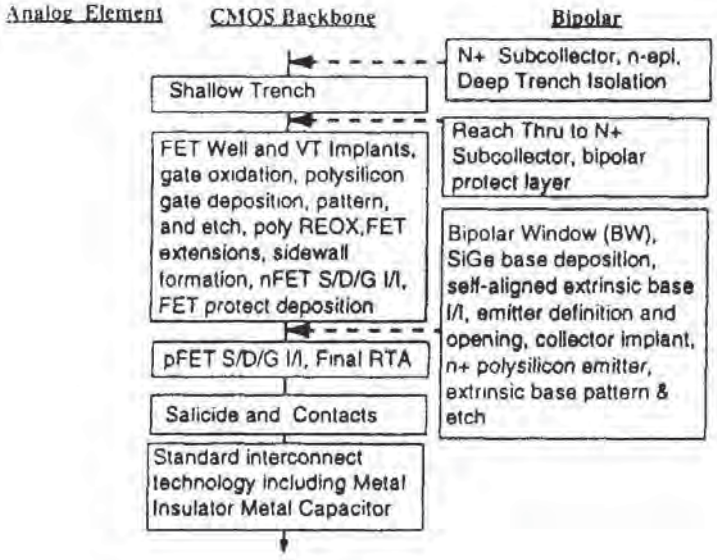



**Fig. 10-9** CMOS process flow for fabricating *base-after-gate* Si:Ge BiCMOS structures.[29]

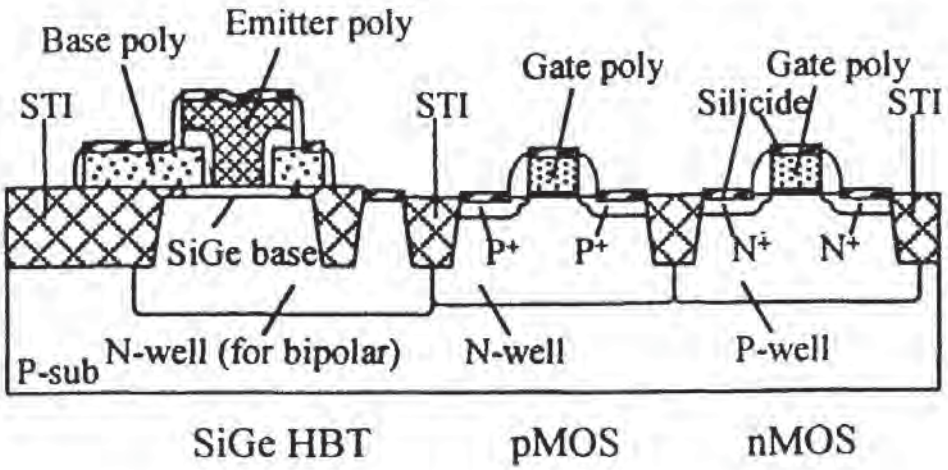
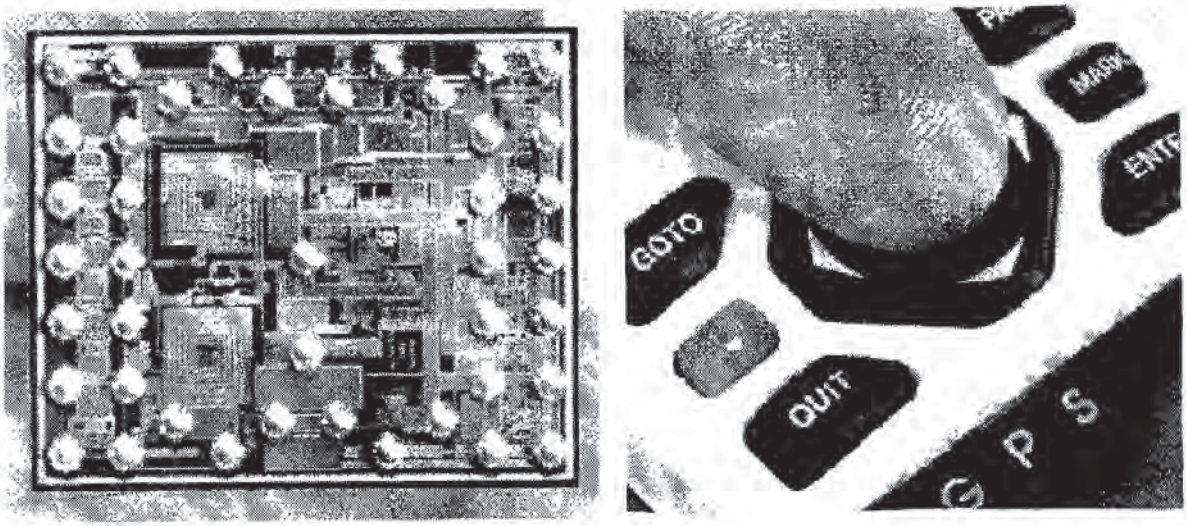



**Fig. 10-10** Schematic cross-sectional view of the Si:Ge BiCMOS. The Si:Ge HBT has a self-aligned structure.[16]

## 10.13 EXAMPLES OF THE MODULAR INTEGRATION Si:Ge-HBTS INTO A BiCMOS PROCESS FLOW

The first Si:Ge-IC that IBM offered commercially was built using 0.25-$\mu$m geometries and combined 1200 Si:Ge-bipolar-transistors with over million CMOS-transistors. This IC was the highest-performance, lowest-noise *data-channel-circuit* in the world. It is used to read data from high-speed hard-drives. The IBM fab at Burlington VT was processing 75 wafers per day of this device (starting in 1999), with yields above 75%.[20]




a) b)

**Fig. 10-11** a) Single-chip GPS Receiver; b) Relative size of GPS Receiver.[8]

In 1998 IBM described the characteristics of a 1.8-million-transistor ASIC chip, fabricated using a Si:Ge-BiCMOS technology.[31] Another description of the modular integration of an HBT that used a Si:Ge-base-layer (in which carbon was also incorporated) into a standard 0.25-$\mu$m-CMOS process to yield a BiCMOS technology is given in Ref. 32. Neither the HBT nor the CMOS parameters were adversely impacted by the additional thermal budget of the complete BiCMOS process. The technology is low-cost, has high-performance, and allows re-use of existing CMOS designs. The Si/Si:Ge-HBTs in this technology are able to perform at very high speeds; $f_{max}$ = 65-GHz. Because such high device-switching-speeds are not necessary for most wireless circuits operating at 900-MHz and 2.4-GHz, the Si:Ge leverage comes from being able to trade the excess speed for improvement in other device attributes; most notably, low-power operation (Fig. 10-11). Another example of integrating Si:Ge-HBTs into a CMOS process is given in Ref. 33. In this process four additional masking steps were needed to create this BiCMOS process. The result is a BiCMOS technology that includes a super self-aligned (SSA) Si:Ge-HBT with a peak $f_T$ of 52-GHz and a peak $f_{max}$ of 70-GHz. A 0.18-$\mu$m rf-BiCMOS technology with a 73-GHz $f_T$ S i:Ge-HBT was described in late 2000.[18] A two-step annealing technique is used to solve the thermal budget tradeoff between Si:Ge HBTs and CMOS. IBM has also devoted significant effort to develop high-quality capacitors and inductors, which can be integrated into the Si:Ge-process. These passive components allow IBM to build the world's only single chip Global Positioning Satellite (GPS) receiver (Fig. 10-11).[8]

A recent report described the effects of scaling on the performance characteristic of deep-submicron Si:Ge HBTs (Fig. 10-12).[41]

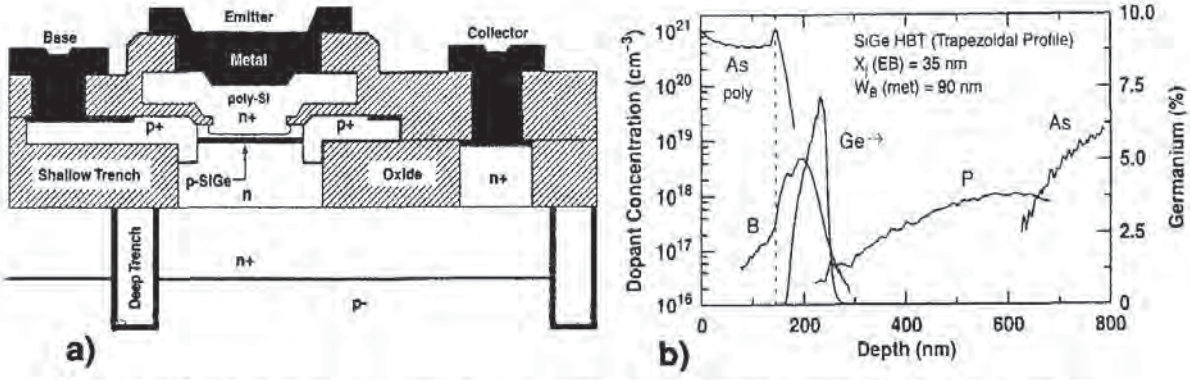



**Fig. 10-12** a) Schematic cross-section of deep-submicron Si:Ge HBT; b) SIMS doping profiles of the dee-submicron Si:Ge HBT.[41] (© 2002 IEEE)

## REFERENCES


1. M. Helms, Chip: Semiconductor Industry Source Book, 1999, pp. 26-27, www.bps-IT.com.
2. D.C. Ahlgren *et al.*, "Manufacturability of an Integrated SiGe HBT Technology for the Analog and Wireless Marketplace," *Tech Dig. IEDM* 1996, p. 859.
3. S. St. Onge *et al.*, "A 0.24-$\mu$m SiGe BiCMOS Mixed Signal RF Technology Featuring a 47 GHz $f_T$ HBT and 0.18 $\mu$m-$L_{eff}$ CMOS," BCTM Proc. 1999.
4. G. Niu, *et al.*, "SiGe Profile Design Tradeoffs for RF Circuit applications," *Tech. Dig. IEDM* 1999 p. 573.
5. K. Washio *et al.*, "A 0.2 $\mu$m Self-Aligned SiGe HBT Featuring 107-GHz $f_{max}$ and 6.7-ps ECL," *Tech. Dig. IEDM* 1999 p. 558.
6. A. Catalano, *Solid State Technology,* September 2000, p. 59.
7. Y.-C. Chao *et al.*, "Enhanced Performance of Sub-100 nm CMOSFETs Using Strained Epitaxial Si-Ge," *Tech. Dig. IEDM*, 2000 p. 753.
8. L.E. Larson, "Device and Technology Requirements for Next Generation Communications Systems," *Tech. Dig. IEDM*, 2000 p. 737.
9. J. Woo and J. Plummer, *IEEE Trans. Electron Dev.,* August 1988, p. 1311.
10. S.L Wright *et al.*, "Molecular Beam Epitaxy Growth of GaP on Si," *J. Appl. Physics*, Vol. 55, April 1984, p. 2916.
11. Y.H. Kwark and R.M. Swanson, "*n*-Type SIPOS and Poly-Silicon Emitters," *Solid-State Electronics,* Vol. 30, November 1987, p. 1121.
12. M. Takahashi *et al., IEEE Electron Dev. Letts.*, October 1987, p. 475.
13. T. Sugii *et al., IEEE Electron Dev. Letts.*, February 1988, p. 87.
14. H. Fujioka *et al., tech. Dig. IEDM*, 1988, p. 574.
15. D. L. Harame *et al.*, "Si/Si-Ge Epitaxial-Base Transistors - Part I: Materials, Physics and Circuits," *IEEE Trans. Electron Dev,.* Vol. 42 March 1995, p. 455.
16. B.S. Meyerson, "Low-Temperature Silicon Epitaxy by Ultrahigh Vacuum Chemical Vapor Deposition," *Appl. Phys. Letts.,* Vol. 47, p. 797 (1985).
17. J.F. Gibbons *et al.*, "Limited Reaction Processing: Silicon Epitaxy," *Appl. Phys. Letts.*, Vol. 10, p. 721 (1985).
18. T. Hashimoto *et al.*, "A 73 GHz 0.18 $\mu$m $f_T$ RF-Si:Ge BiCMOS Technology Considering Thermal Budget Trade-Off and With Reduced Boron-Spike Effect on HBT Characteristics," *Tech Dig. IEDM*, 2000, p. 149.
19. R. People and J.C. Bean, *Appl. Phys. Letts.,* **48**, p. 538 (1986).
20. B.S. Meyerson, "Silicon Germanium Technology: A 'Material of the Future' Gone Mainstream," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference,* Fall 2000, Boston MA.
21. B.S. Meyerson, "Low-Temperature Si & SiGe Epitaxy by Ultrahigh Vacuum/CVD: Process Fundamentals," *IBM J. Res. Dev.,* Vol. 34, p. 806 (1990).

**22**. K.E. Violette *et al*, "Low-Temperature Selective Silicon Epitaxy by Ultrahigh Vacuum Rapid Thermal Chemical Vapor Deposition Using $Si_2H_6$, $H_2$ and HCl," *Appl. Phys. Letts.*, Vol. 68, p. 66 (1996).
**23**. "Motorola's Next Si:Ge Wireless Technology Will Use the ASM Epsilon E2000 Reactor." *Press Release: ASMl*, March 1, 2001, Bilthoven, Netherlands.
**24**. D.J. Meyer and J. Italiano, "Productive Single-Wafer Si:Ge Processing," *Meeting Abstracts of the Fall Mtg. 1999, Electrochem. Society*, Abs. No. 1314.
**25**. ibid. Ref. 6.
**26**. J. Bock *et al.*, "Si-Ge Bipolar Technology for Mixed Digital and Analogue RF Applications," *Tech. Dig. IEDM* 2000 p. 745.
**27**. K. Washio *et al.*, "A 0.2 μm 180-GHz - 6.7 ps-ECL SOI/HRS Self-Aligned SEG Si-Ge HBT/CMOS Technology for Microwave and High-Speed Digital Applications," *Tech. Dig. IEDM* 2000 p. 741.
**28**. D. L. Harame *et al.*, "Si/Si-Ge Epitaxial-Base Transistors - Part II: Process Integration and Analog Applications," *IEEE Trans. Electron Dev.* **42** Mar 1995, p. 469.
**29**. D.C. Ahlgren and B. Jaganathan, "SiGe for Mainstream Semiconductor Manufacturing," *Solid State Technology,* January 2000, p. 53.
**30**. D. Harame *et al.*, "Trends in SiGe BiCMOS Process Integration," *Meeting Abstracts of the Fall 1999 Mtg. Electrochemical Society,* Abs. No. 1333.
**31**. R.A. Johnson *et al.*, "1.8 Million Transistor CMOS ASIC Fabricated in a Si-Ge BiCMOS Technology," *Tech. Dig. IEDM* 1998, p. 217.
**32**. K.E. Ewald *et al.*, "Modular Integration of High-Performance SiGe:C HBTs in a Deep Submicron, Epi-Free CMOS Process," *Tech Dig. IEDM* 1999 p. 561.
**33**. C.A. King, "Very-Low Cost Graded SiGe-Base Bipolar Transistors for a High Performance Modular BiCMOS Process," *Tech Dig. IEDM* 1999 p. 565.
**34**. S.R. Stiffler *et al.*, "The Thermal Stability of Si-Ge Films Deposited by UHV/CVD," *Appl. Phys. Lett.*, **70**, p. 1416, 1991.
**35**. D.C. Ahlgren *et al.*, "Si-Ge Fuels High-Speed Communications," *Semiconductor International,* November 2000, p. 62.
**36.** D.L. Harame and. B.S, Meyerson, "The Early History of IBM's SiGe Mixed Signal Technology," *IEEE Trans. Electron Dev.,* **48**, November 2001, p. 2555.
**37.** H. Rucker *et al.*, "Dopant Diffusion in C-doped Si and SiGe: Physical Model and Experimental Verification," *Tech. Dig. IEDM*, 1999, p. 345.
**38.** D.L. Harame *et al.*, "Current Status and Future Trends in SiGe BiCMOS Technology," *IEEE Trans. Electron Dev.*, **48**, November 2001, p. 2575.
**39**. A.V. Samilov *et al.*, "Dedicated Pre-Clean Reduces Epi's Thermal Budget," *Semiconductor International,* November 2000, p. 73.
**40.** V. Loup *et al.*, "Fine Control of Low-Temperature Epitaxial Growth," *Solid State Technology,* July 2001, p. 91.
**41.** S.Zhang *et al.*, "Effects of Geometrical Scaling on Frequency Response and Noise Performance of SiGe HBTs," *IEEE Trans. Electron Dev.* Mar. 2001, p. 429.

# Chapter 11

# SILICON-ON-INSULATOR (SOI) TECHNOLOGY

In this chapter the characteristics of silicon-on-insulator (SOI) technologies fabricated with deep-submicron MOSFETs are described. The focus is on the SOI-devices themselves as well as on the process technologies used to fabricate both the starting SOI-wafers and the SOI ICs. An argument is also presented as to why SOI technology may provide a path for continuing to advance CMOS ICs for several more generations. That is, implementation of SOI may make it possible to exceed the limits facing bulk-silicon-CMOS in the near term.

## 11.1 WHAT IS SILICON-ON-INSULATOR (SOI)?

*Silicon-on-insulator* (*SOI*) circuits consist of ICs made from single-device islands dielectrically isolated from the substrate and each other. Each such island contains a single transistor, with the isolated transistor being either a MOSFET or a bipolar transistor. However, since about 90% of the ICs manufactured in 2002 are based on CMOS-technology, the discussion in this chapter will be restricted to SOI technology involving MOSFETs. Figure 11-1 shows a cross-section of a MOSFET fabricated on an SOI wafer that also uses STI as the isolation structure to provide lateral isolation. Note that in the case of a MOSFET fabricated on an

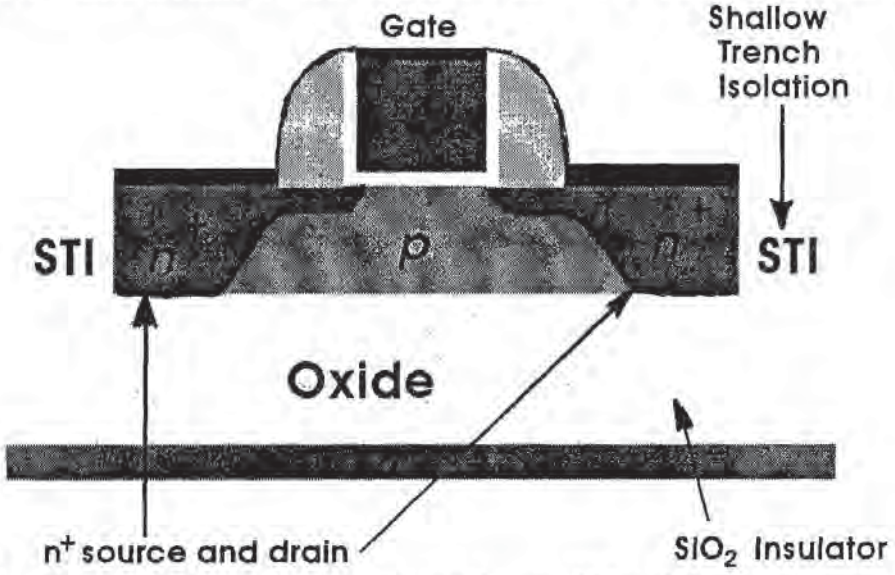



**Fig. 11-1** Cross-section of a MOSFET fabricated on an SOI wafer.

SOI wafer, a *buried oxide* (*BOX*) underneath the device provides vertical isolation, and a trench oxide (in this example), provides the lateral isolation. Since the trench oxide extends all the way through the silicon device layer atop the BOX (and a thermal oxide exists on the wafer top surface), dielectric material continuously surrounds each MOSFET on all sides. Thus, complete electrical isolation is provided for the MOSFET with respect to all other devices, and the substrate, on the wafer.

### 11.1.1 Isolation Issues in CMOS Fabricated in Bulk-Silicon Wafers

The electrical isolation requirements of CMOS ICs built in SOI are quite different from those encountered when CMOS ICs are fabricated in bulk-silicon. To see why, the circuit strategies and structures needed for isolation purposes in CMOS circuits on bulk-silicon must be examined. For reference purposes a single-MOSFET built in a bulk-silicon-substrate is shown in Fig. 11-2a (with various device structures and regions defined). Figure 11-2b shows two NMOSFETs built on a *p*-doped substrate, together with the parasitic-transistor that exists in the field-region separating the active devices. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs). Figure 11-2c shows an *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. This figure also shows the parasitic MOSFETs that are associated with that structure (as well as the *pn*-junction that separates the *n*-wells from the *p*-substrate). Again, a semirecessed-LOCOS field-oxide plays a role in maintaining electrical isolation among the MOSFETs of the CMOS structure. (It should be noted that shallow-trench isolation [STI] is used in place of LOCOS to form the field-oxide in CMOS technologies below 0.35-$\mu$m. Nevertheless, the isolation issues described here for CMOS on bulk-silicon remain the same, regardless how the field-oxide is formed). Finally, in Fig. 11-2d a cross-section of a CMOS IC structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents.

Using the IC device structures described above, the isolation issues involved in CMOS ICs can be undertaken. The first item to be noted is that the *pn*-junction separating a well from the substrate must be kept under reverse-bias to prevent current from flowing from the power-supply-line to ground through forward-bias of this junction. Second, inter-device isolation for devices within a well (*intra-well isolation)* or within the substrate regions, entails the use of a field-oxide to suppress spurious surface-leakage effects between active devices within a well. Third, a field-oxide and channel-stop implants are also used to prevent leakage

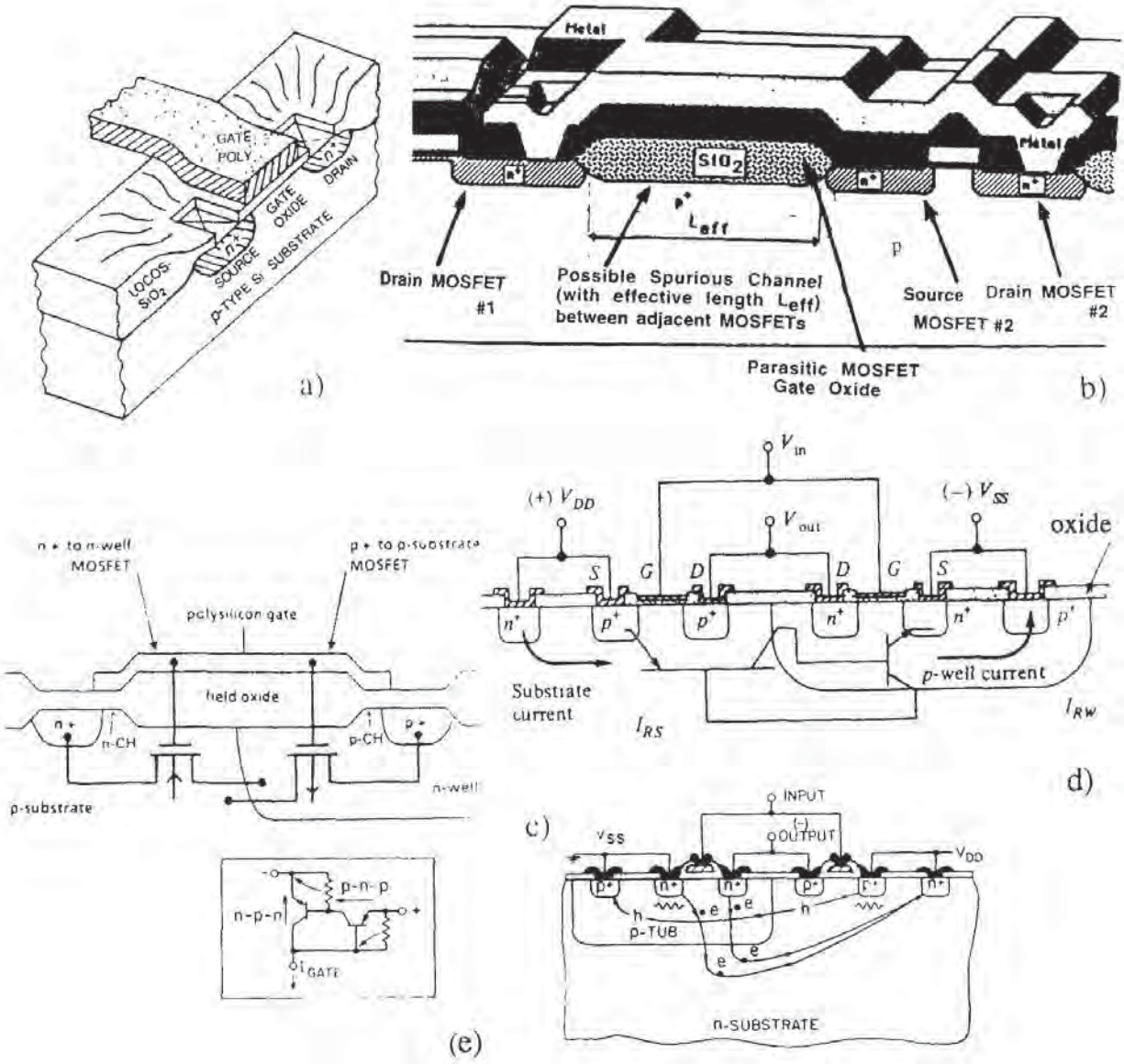



**Fig. 11-2** a) A single MOSFET built in a bulk-silicon substrate, with various device structures and regions defined; b) Two MOSFETs (e.g., both NMOSFETs) built in a substrate having one type of doping (i.e., on a *p*-doped substrate), together with the parasitic transistor that exists in the field-region separating these MOSFETs. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs); c) An *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. Also shown in the figure are the parasitic MOSFETs that are associated with that structure, as well as the *pn* junction the separates the *n*-wells from the *p*-substrate; d) A cross-section of a CMOS-IC-structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents; e) The parasitic bipolar transistors that exist in CMOS structures can be *turned on*, which can cause the CMOS IC to go into a latchup state.

between devices among different wells (*inter-well isolation*). Yet, despite employing these isolation structures, intra-well- and inter-well-leakage still occurs to some degree in CMOS ICs made in bulk-silicon (and the problem gets worse as the devices are brought closer together). Note that Volume 3 can be consulted for a more detailed discussion of isolation issues in CMOS built in bulk-silicon wafers. In addition, the parasitic-bipolar-transistors that exist in CMOS structures can be *turned-on*, and this can cause the CMOS IC to go into a *latchup state* (Fig. 11-2e). That is, latchup can occur because two parasitic-bipolar-transistors are always simultaneously formed in the course of fabricating a CMOS structure. In bulk-CMOS, these two transistors are connected in such a way that if the current gain of the loop formed by these two devices is larger than unity, latchup can occur. Latchup can be triggered by different mechanisms, such as node- voltage overshoots, displacement-current, junction-avalanching, and pho-tocurrents. It is true that latchup is a well-understood phenomenon which can be dealt with in bulk-CMOS using proper processing and design methodologies. However, with the constant reduction in feature spacing, latchup immunity in bulk CMOS ICs remains a continuing challenge.

### 11.1.2 Isolation in SOI CMOS

The isolation issues of CMOS fabricated in SOI wafers are much less complex than those involving CMOS ICs in bulk-silicon. This is because each MOSFET in SOI is built in an island of silicon on an insulating substrate. Hence, all of the previously-mentioned possible isolation causes are eliminated. Let us explain this assertion in more detail.

First, the requirement of keeping the *pn*-junction that separates the CMOS well regions from the substrate under reverse-bias is eliminated. That is, this requirement no longer exists because a continuous oxide barrier exists in SOI between the well-regions and the substrate. Consequently, current cannot flow between the power and ground lines, even if the junction in question becomes forward-biased. Second, the possibility of leakage current between devices within a well (intra-well-leakage) and from well-to-well (inter-well-leakage) is eliminated. This is because the parasitic-field-region-MOSFET does not exist (as a result of the field oxide extending through the silicon film and joining the buried oxide). In other words, the channel between the source and drain of any parasitic-field-region-MOSFET is electrically interrupted by the oxide moat surrounding each MOSFET within an SOI IC. Consequently, leakage current due to: 1) field-inversion of the channel of a parasitic-field-region-MOSFET; or

**Table 11-1 ITRS Roadmap of the Required MOSFET Junction Characteristics**

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2008 | 2010 | 2014 |
|---|---|---|---|---|---|---|---|---|---|
| Gate size (nm) | 120 | 100 | 85 | 80 | 70 | 65 | 45 | 32 | 22 |
| **S/D junction depths (nm)** | | | | | | | | | |
| Contact | 65–125 | 55–105 | 45–90 | 43–5 | 38–75 | 35–70 | 30–55 | 20–40 | 15–35 |
| Extension/channel | 36–60 | 30–50 | 25–43 | 24–40 | 20–35 | 20–33 | 16–26 | 11–19 | 8–13 |
| SSRD channel | 18–30 | 15–25 | 12–22 | 12–20 | 10–18 | 10–17 | 8–13 | 5–10 | 4–8 |
| **Extension doping** | | | | | | | | | |
| Sheet resistance (Ω/sq ) | 310–760 | 280–730 | 250–700 | 240–675 | 220–650 | 200–625 | 150–525 | 120–450 | 100–400 |
| Resistivity ($10^{-3}$Ω-cm) | 1.1–4.6 | 0.84–3.6 | 0 62–3.0 | 0.57–2 7 | 0.44–2.3 | 0 40–2.1 | 0.24–1 4 | 0 13–0 86 | 0 08–0 52 |
| **Lateral extension junction abruptness (nm/decade)** | | | | | | | | | |
| For $V_t$-roll-off control | 12 | 10 | 8 5 | 8 | 7 | 6.5 | 4.5 | 3.2 | 2 2 |
| For spreading resistance control | 4 8 | 4 1 | 3 4 | 2 9 | 2 7 | 2 4 | 1 25 | 0 8 | 0 5 |

2) parasitic-MOSFET subsurface-punchthrough, do not exist in SOI-technologies. Finally, the existence of vertical- (as well as lateral-) isolation within SOI ICs eliminates the possibility of latchup. That is, the current paths from the collectors of the parasitic-bipolar-transistors to the bases of the complimentary bipolar-transistors (which set up feedback paths in CMOS in bulk-silicon that cause the circuits to enter latchup), are interrupted by the vertical oxide of the SOI. Hence, latchup can no longer occur in SOI ICs.

Despite the fact that intra-well and inter-well leakage paths are interrupted in SOI CMOS structures, new *intra-device* parasitic transistor paths inherent in SOI CMOS structures can still produce leakage current, as is discussed in Sect. 11.7.

## 11.2 THE "END OF THE ROAD FOR BULK CMOS" SCENARIO PREDICTED BY THE 2001 INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

The *2001 International Technology Roadmap for Semiconductors* (ITRS-2001) indicates that planar CMOS transistors on bulk-silicon are expected to reach their limits at gate lengths of about 50-nm in the 2005-2006 time frame. That is, the ITRS-2001 analysis documents a clear "end-of-roadmap" event-horizon at the onset of IC device scaling to 50-nm gate-lengths.[1] The premise of the ITRS study was to predict, as far as possible, the characteristics of an extension of the industry's "standard technology," namely, planar-CMOS-transistors fabricated

on bulk-silicon wafers. The study concludes that many process and materials constraints combine to force this change in the technology path. That is, the conclusions drawn in this study are that sometime between 2005 and 2008 nearly all of the 30 or so transistor characteristics required for the extension of Moore's Law are listed as "no known solutions." This is a clear signal that IC transistor technology requires fundamental changes in the next half-decade. In this section the principal issues limiting *junction-formation* in sub-100-nm CMOS ICs on bulk-silicon will be detailed. As part of the discussion to be presented in the next section (Section 11.3), it will be described how some of these constraints can be removed or relaxed through the use of SOI. That is, by implementing CMOS-on-SOI a path may be provided to allow CMOS ICs to be fabricated beyond the 50-nm gate-length-limit faced by CMOS-on-bulk-silicon. An extension of as much as a decade beyond the "end-of-the roa-map" scenario posited by the ITRS-2001 may be possible by fabricating CMOS ICs on SOI-wafers.

The three principal challenges for forming shallow-junctions in CMOS on bulk-silicon wafers are as follows:

1. Limiting dopant-diffusion in the source/drain regions to shallow depths;
2. Activating a sufficient percentage of the dopants implanted into the source/drain regions to provide source/drain regions that exhibit sufficiently low contact-resistance and sheet-resistivities:
3. Obtaining source/drain junctions that have highly-abrupt vertical and lateral doping-profiles.

The ITRS-2001 values for these key transistor parameters, as a function of the device generation (which correlates to the gate-length of that generation and the expected time of its introduction) are given in Table 11-1.

### 11.2.1 Source/Drain Junction Scaling

As MOSFET gate lengths ($L_L$) are scaled, the other device structures (i.e., the gate-oxide-thickness and the junction-depth of the source/drain extension) are also scaled proportionally. To maintain such "proportional scaling" for CMOS when the gate-length decreases below 100-nm, the design rule used in the ITRS-2001 study was that the junction depth of the source/drain extension $X_{j(extension)}$ would scale according to: $X_{j(extension)} \sim 0.55 \times L_L$ within a range of ±25% (see Fig. 11-2a for a graphic that depicts the region referred to as the source/drain extension). Controlling junction depths to less than 55-nm for sub-100-nm MOSFETs will push the limits of sub-keV boron implantation and "spike" anneals.[2]

### 11.2.2 Dopant Activation in Shallow Source/Drain Junctions

In the ITRS-2001 model, it is expected that ways will be found to produce highly doped source/drain extension-regions having a sufficiently low resistivity (to reduce the sheet-resistance of these regions) at the same time the depth of the junctions of these regions is reduced. If this can be done (i.e., if the resistivity of the source/drain extension regions could somehow be reduced from the value of $10^{-3}$ Ω-cm [that is produced in the source/drain extension-regions of 0.15-μm NMOS transistors by using low-energy boron implantation {0.2-2.0-keV} and "spike" anneals] to say $10^{-4}$ Ω-cm) a sheet-resistance versus junction-depth curve (as plotted in Fig. 11-3 for the value of $\rho = 10^{-4}$ Ω-cm) could be obtained.[3] Overlaid on these sheet-resistance versus junction-depth curves (with $\rho$ as the parameter), are the industry expectations of the value of $\rho$ for: **1**) high-performance ICs; and **2**) cost-focus (i.e., lower-cost) ICs, as taken from the ITRS-2001. That is, the values of $\rho$ that the ITRS-2001 roadmap expects in order to be able to fabricate CMOS ICs on bulk-silicon for smaller gate-length-devices, decreases from about $10^{-3}$ Ω-cm for NMOSFETs with gate lengths of 0.14-μm, down to

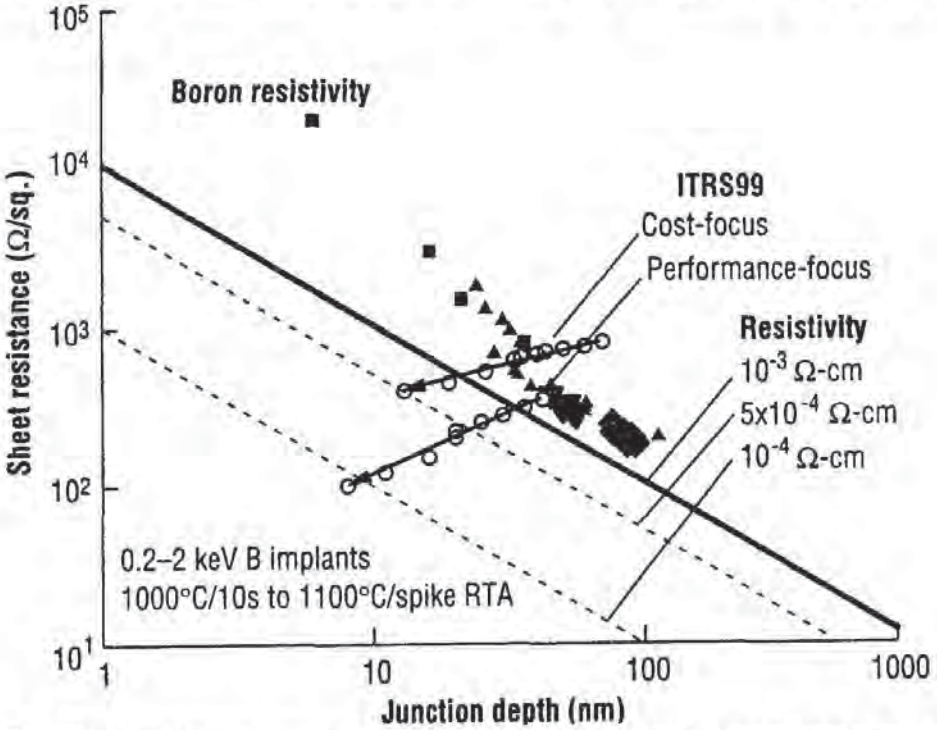



**Fig. 11-3** Sheet-resistance and junction-depths for low-energy boron implants and spike-implant anneals, and industry expectations for S/D extension junctions (starting with 140-nm gates to 22-nm gates at the head of the arrows) for cost-focus and performance focus devices. The dopant resistivity for sub-keV boron implant and spike-RTP process (in CZ-Si) is limited by a "wall" at about $10^{-3}$ Ω-cm.

about $10^{-4}$ Ω-cm for NMOSFETs with 22-nm gate lengths (corresponding to a junction-depth of 8-nm and a sheet resistance of 100 Ω/sq). Some alternative doping technology may be required to achieve such low resistivity. Possible alternative doping techniques include: **a**) low-temperature (or selective) epitaxy; **b**) diffusion from a doped-glass source; **c**) diffusion from a low-temperature Si-Ge alloy; **d**) direct gas-source-diffusion in an RTP-system; **e**) gas-immersion-laser-diffusion (GILD); and **f**) planar-doping-layers using atomic-layer-epitaxy (ALE). Of these many possible processes, low-temperature-epitaxy may hold some key advantages, as it can also be done in a selective manner (which may possibly minimize masking steps). However, with selective epitaxy, separate hard-masks are required for *n*- and *p*-S/D-regions. On the other hand, GILD with projection masking (PGILD) does not require resists or hard-masking and may become an important doping technology.

At this time, however, it is not clear whether any of the above methods of reducing the value of $\rho$ in the source/drain extension regions of NMOSFETs will be able to reduce $\rho$ down to $10^{-4}$ Ω-cm. In fact, the resistivities of junctions formed by the present day "best methods" of sub-keV boron implants and spike anneals appears to be limited by the "activation-wall" to a value no smaller than $10^{-3}$ Ω-cm.[4]

### 11.2.3 Doping-Profile Abruptness in Shallow Source/Drain Junctions

To permit sub-100-nm MOSFETs to exhibit the performance characteristics expected in the ITRS-2001 roadmap, the total source-to-drain resistance of such MOSFETs must be kept sufficiently small.[5] Controlling this total source-to-drain resistance-value in deep-submicron MOSFETs requires that special attention be paid to the contact-resistance and spreading-resistance at the junction of the source extension and the channel. As shown in Fig. 11-4, parasitic-resistances exist in the path between the metal-to-Si substrate interface and the region in the device where the actual transistor action begins (see also Sect. 5.4, Chap. 5).

In the planar MOSFET, the device-region in which transistor-action occurs is the channel-region under the gate. The current flowing from the metal-interconnect lines to the channel-region must be transported from the metal-Si contact to the edge of the channel through the source or drain regions (Fig. 11-4). In the MOSFET, current enters the contact perpendicular to the wafer surface, and then travels parallel to the surface in order to reach the channel. The total parasitic-series-resistance, $R_S$, of the current path from the contact to the edge of the channel can be modeled by the sum of four components,

$$R_S = R_{co} + R_{sh} + R_{sp} + R_{ac} \qquad (11.1)$$

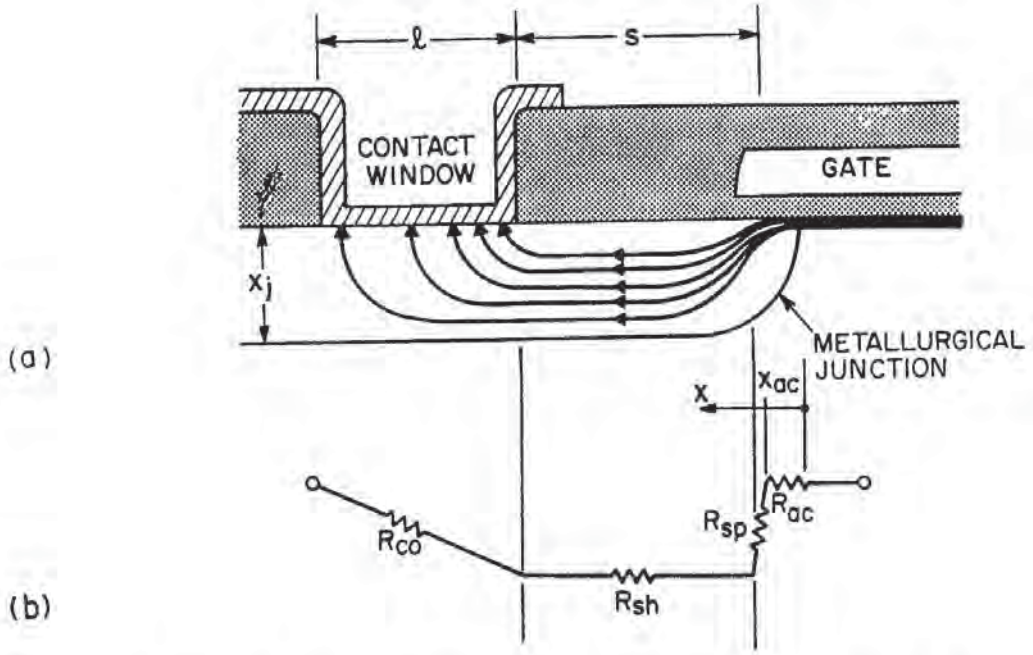



**Fig. 11-4** Schematic drawing showing: a) current paths within the source and drain regions, of a planar MOSFET and; b) their corresponding representative series resistance components.[5] (© 1979 IEEE)

where $R_{co}$ is the *contact-resistance* between the metal and the source/drain region, $R_{sh}$ is the *sheet-resistance* of the bulk-region of the source and drain, $R_{sp}$ describes the *resistance of the current lines crowding near the channel end of the source*, and $R_{ac}$ is the *accumulation-layer resistance*. Qualitatively, for a non-abrupt junction, $R_{sp}$ and $R_{ac}$ arise because, after leaving the channel, the current does not immediately spread into the bulk of the source-drain regions (the parasitic-resistance-effect in these bulk regions is $R_{sh}$). Thus, $R_{sp}$ and $R_{ac}$ exist because the accumulation-layer conductance is much higher than the semiconductor conductance in the vicinity of the junction (see again Sect 5.3).

The key issue for sub-100-nm MOSFETs is the abruptness of the doping-profile at the lateral edge of the source extension and its effect on $R_{sp}$ and $R_{ac}$. The ITRS-2001 study calls for lateral-doping-profiles that decrease into the channel region at a slope of ~3 nm/decade in concentration, for 100-nm gates. Obtaining a lateral-doping-profile abruptness of less than 3-nm/decade presents severe challenges for bulk silicon doping by implantation and rapid thermal annealing. As-implanted profiles for 0.2-keV boron into pre-amorphized silicon have a vertical profile steepness of 2.5-nm/decade.[6] The key S/D-extension profile in the lateral direction, still remains to be characterized, and will be further broadened by thermal annealing in actual transistors.

## 11.3 WHY SOI?

Silicon-on-insulator (SOI) technology has long been used in many specialized applications, such as radiation-hardened or high-voltage integrated circuits. It is only in recent years (since about the mid-1990s), however, that SOI has emerged as serious contender for low-power, high-performance applications.[7] The power consumption, P, in CMOS ICs is given approximately by:

$$P \sim C V^2 f + V I_{leakage} \quad (11.2)$$

where C is the total capacitance, V is the supply voltage, $f$ is the frequency, and $I_{leakage}$ is the standby-leakage-current. Thus, in low-power applications the supply voltage must be reduced to decrease both terms of this equation. This power-supply reduction has had to follow the evolution in battery-voltage and capacity. The SEMATECH roadmap target has been to reach 0.9V by the end of the century. However, at such a low voltage, the IC performance is also reduced due to a smaller value of $I_D$ in MOSFETs. As will be seen, however, the use of SOI allows the drain current to be significantly increased, even at such low voltages. SOI therefore also becomes appropriate for emerging system-on-chip ICs used for smart power, as well as for micro-electromechanical systems (MEMS), mixed-signal, and integrated-optics applications.

SOI also provides a path to permit the scaling of ICs to go forward. That is, to continue to increase performance and keep up with Moore's Law, several cumulative approaches have been adopted. The evolution of IC technology in the near future which uses this strategy is summarized in Fig. 11-5. This diagram indicates that integrating copper wiring and SOI will represent the next step toward the continued improvement in IC performance. Here we list the advant-

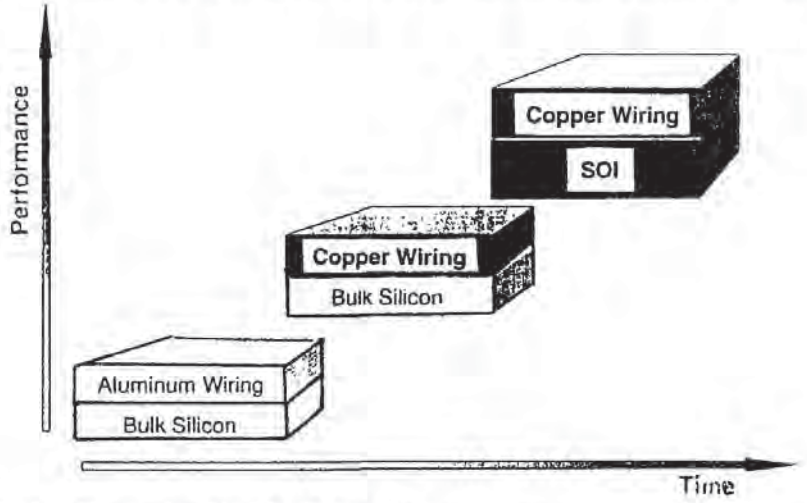



**Fig. 11-5** Technology migration for future IC progress.

ages offered by CMOS-onSOI which contribute to such performance improvement.

1. **Improved Speed-Performance.** The speed performance of SOI ICs is enhanced by 20-35% compared to the same circuits fabricated in CMOS on bulk-silicon if they are operated at the same voltages. (The reasons for this increase in speed performance is due to a number of factors that are listed later as other advantages of SOI.) Thus, SOI circuits built using the design rules of "generation-$n$" of some CMOS technology in bulk silicon will operate as well as the ICs of "generation-$n$ + 1" of CMOS-on-bulk-silicon. This also implies that if a roadblock occurs in some upcoming generation of CMOS on bulk silicon, by having the capability to produce these ICs in SOI, it will be possible to fabricate ICs that allow that roadblock to be exceeded for at least one generation (if not more).

2. **Reduced Power-Consumption.** If CMOS-ICs in bulk-silicon and CMOS ICs in SOI are operated to give the same speed-performance, there is a 35-70% reduction in power-consumption in the CMOS-on-SOI ICs. This advantage provides CMOS-on-SOI a huge competitive edge for low-voltage/low-power applications operated with a one-battery supply. Figure 11-6 depicts this performance-versus-power tradeoff issue.

3. **Reduced Junction-Capacitance.** One reason for the increased performance of CMOS-on-SOI is that the junction-capacitance of the MOSFETs is significantly reduced. For example, a depletion-capacitance exists at the drain-substrate junction, and its value is a function of the

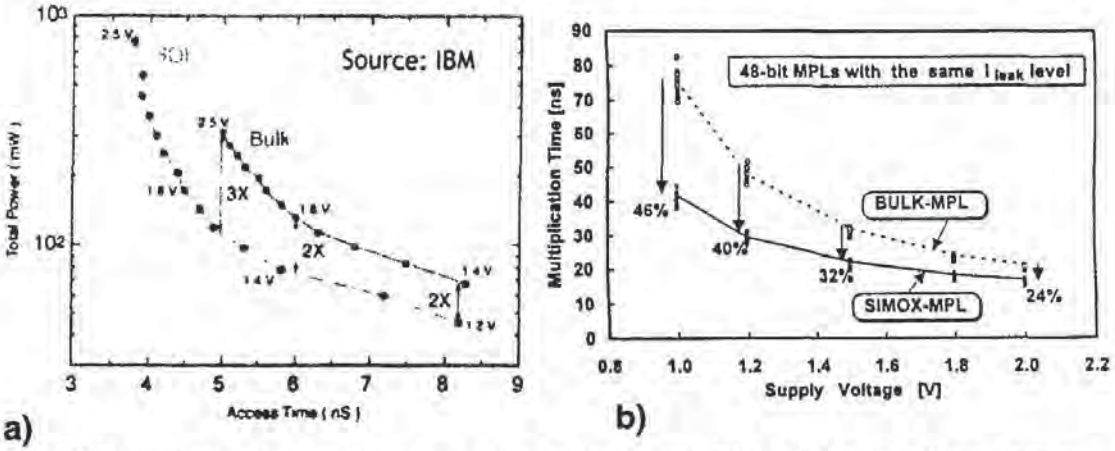



**Fig. 11-6** a) Power drain and access time for bulk and SOI CMOS for drive voltages ranging from 1.2 V; b) Comparison of 48-bit-multiplier performance between bulk and SOI using a 0.25-$\mu$m CMOS technology.

substrate doping-concentration. The higher the substrate doping-concentration, the higher the capacitance. Modern submicron circuits increase the substrate doping-concentrations with each new technology generation, which contributes to an increased junction-capacitance. However, in SOI circuits, the maximum capacitance between the junctions and the substrate is the capacitance of the buried-insulator. This capacitance is proportional to the dielectric constant of the capacitance material. Silicon dioxide (which is used as the buried insulator), has a dielectric constant $k_{SiO_2} = 3.9$ which is three times smaller than that of silicon $k_{Si} = 11.7$ Therefore, a source/drain junction located on a buried-oxide gives a parasitic-capacitance approximately three times smaller than that of a bulk-junction. Since the SOI buried-insulator-thickness does not scale with device scaling, parasitic-capacitances do not necessarily increase as SOI technology progresses.

As a result of the above phenomena, and depending on the circuit layout, the total capacitance of a CMOS SOI-IC can be reduced by 15-50% compared with the same IC fabricated in bulk-silicon. In fully-depleted SOI (FD-SOI) MOSFETs (described in Sect. 11.5.1), the source/drain-regions butt-up against buried oxide (BOX), and thus no *pn*-junction exists at that device-region boundary (and hence there is no depletion-region that needs to charged and discharged there). Even in partially-depleted SOI devices (PD-SOI), the depletion-region of the junction of the source/drain-regions extends to the BOX, and does not increase in volume beyond that point once this depth is reached. Thus, even in PD-SOIs the maximum junction-capacitance of the source-to-body, $C_{jSB}$, and drain-to-body, $C_{jDB}$, is limited to a smaller value than that existing a comparably-sized MOSFET in bulk-silicon. The wiring-capacitance (i.e., associated with interconnect lines running over the field regions of an IC) is also slightly reduced by the presence of the BOX in SOI wafers.

**4. Latchup Elimination.** As noted in Sect. 11.1.2, the phenomenon of latchup is eliminated in CMOS-on-SOI ICs and the techniques needed to suppress latchup in bulk CMOS are unnecessary. This also means that the use of epitaxial-wafers with a heavily-doped-substrate to control latchup in some applications is no longer needed.

**5. Reduction of Interwell-Isolation-Spacing.** Due to the elimination of the possibility of latchup, and the ability to retain adequately low intra-well and inter-well leakage-currents due to the dielectric isolation, it is possible to reduce the isolation-spacing design rules, compared to those needed in

CMOS ICs in bulk-silicon (especially the $n^+$-$p^+$ spacing). In addition, the contact-overlap-spacing can be smaller. Thus, the packing density of the CMOS SOI-IC is increased, allowing circuits with greater functional-density to be fabricated on the same area of silicon. There is also no need for high-energy implantation of well-dopants.

6. **Reduced Soft-Error-Sensitivity.** SOI reduces the substrate-collection-volume for alpha-particle- and cosmic-ray-generated electron-hole tracks, hence lowering soft-error-sensitivity. In particular, SOI- technology for ECL-bipolar/CMOS ICs is preferred because of the high sensitivity of the ECL peripheral-circuitry to radiation.[8]

7. **Improved MOSFET Turn-Off Characteristic (Steeper Subthreshold-Slope).** The subthreshold slope, $S_t$, of MOSFETs built on SOI-wafers is much steeper than that of MOSFETs in bulk-Si-wafers. Subthreshold-slopes as small as 62-mV/decade have been observed in SOI MOSFETs.[9] The small value of $S_t$ implies that an SOI-MOSFET can be turned-off with a smaller gate-voltage swing. This allows faster switching in MOSFETs, along with the use of lower threshold-voltages. A lower threshold-voltage allows lower power-supply-voltages to be used, and an increase in the drive-currents of SOI-MOSFETs compared to MOSFETs built in bulk-silicon.

8. **Reduced Leakage-Current Arising from Carrier-Generation in Junction Depletion Regions.** The ability to make the source/drain junction-areas of SOI-MOSFETs smaller by approximately two decades, also decreases the volume of the depletion-regions of these junctions. Thus, less leakage - current arising from the generation of carriers in these regions is produced, which reduces the standby leakage-current of SOI-ICs to a comparable degree. The negligible leakage current makes SOI ideal for low-power ICs.

9. **Compatibility with Bulk-Si CMOS Process Tools & Processes.** The SOI process-technology is compatible with the conventional process-technology used to fabricate CMOS ICs in bulk-silicon. Thus, the tools and facilities used to fabricate CMOS-on-bulk-silicon can also be used to fabricate CMOS-on-SOI. The SOI process-sequence is even somewhat simpler than that used to fabricate CMOS on bulk-silicon. That is, there is no need for the well-formation process in SOI-ICs, nor any need for channel-stop-implants to suppress lateral-surface-leakage between devices due to field-inversion.

10. **Advantages for System-On-Chip (SOC).** A high-resistivity substrate under the BOX can be used, which can significantly reduce the device

capacitance and the loss of passive components, such as inductors and transmission-lines, commonly used in analog- and rf-devices.[10] Therefore, SOI offers advantages over bulk-Si to realize a system-on-a-chip by integrating high-performance digital and analog/rf circuits on the same chip.

**11. Increased Drain-Current.** The drain current of SOI-MOSFETs, $I_D$, is also increased due to the significantly lower doping that can be used in the channel. The ability to tolerate lower-channel-doping occurs because of the natural punchthrough suppression arising from the backgating effects of the insulator interface.

**12. Improved Circuit Reliability Due to Elimination of Junction-Spiking.** When contact is made to a MOSFET in bulk-CMOS, unwanted reactions between the metal or the silicide can take place, leading to the problem of "junction-spiking" (and the attendant rise in uncontrolled leakage currents). While barrier materials normally control this problem, the potential for its occurrence is still there in bulk-CMOS. If MOSFETs are fabricated in thin SOI material, the source and drain junctions extend to the buried insulator. In that case, no leakage will be produced if some uncontrolled metal-silicon reaction occurs.

### 11.3.1 The Path of Using SOI May Allow Extension of the ITRS Roadmap

Many of the challenges for CMOS transistors in bulk-silicon discussed in Sect. 11.2 are relaxed or removed when SOI-wafers are used instead of bulk-silicon wafers, and SOI technologies offer new options for transistor design. With the 20-50% increase in transistor speed that can be achieved by switching to SOI, the performance of ICs can be increased without the cost of a shift to a higher lithography resolution. SOI "thin-body" CMOS transistors provide excellent control for the off-current (leakage) characteristics that limit the endurance of battery-operated computer and communication devices. Furthermore, soft-error-immunity and radiation-hardness of SOI devices is seen as a key requirement for expansion of high-data-rate network-servers that form the core of Internet and other global data-transfer-links.

In addition, SOI-wafers are opening a path for the extension of planar CMOS transistor design by perhaps a decade beyond the "end-of-roadmap" limits foreseen in the ITRS-2001 study for 50-nm gates on bulk-silicon. Operational SOI transistors have been studied with gate lengths as small as 25-nm and Si channel thicknesses of 1-nm.

**TABLE 11-2 COMPARISON OF KEY ISSUES FOR SUB-100-nm CMOS FOR BULK-SILICON AND FULLY-DEPLETED SOI MOSFETS**

| Critical Issue | Bulk silicon | SOI |
|---|---|---|
| Junction depth | RTP/(LTP) | Set by Si-SOI film |
| Dopant activation | Thermal budget on vertical and lateral diffusion | Lateral diffusion limit |
| Junction abruptness | Critical vertical and lateral abruptness limits | Carrier injection still requires abrupt lateral junction in thin-body |
| Contacts | Extended S/D | Extended S/D |
| Off-current | Requires halo | Dual gate |
| Threshold control | SSRD (super-steep retrograde doped channels) or SiGe gate | SiGe gate |
| Well doping | High-energy implant | No well |

To begin, it is easier to control the junction-depths of the source/drain extensions in SOI devices. In bulk-silicon, obtaining shallow-junctions requires careful control of thermal budgets, with the use of rapid-thermal-annealing-"spikes" or laser-annealing to limit both the vertical- and lateral-dopant-diffusion, while obtaining high degrees of dopant-activation and adequate defect annealing. With fully-depleted, thin SOI (FD-SOI) MOSFETs (discussed in Sect. 11.5.2), the source/drain-extension junction-depth is determined by the Si-SOI-layer thickness. This means that the thermal-budget-constraints on vertical-junction-depth are essentially removed for fully-depleted-SOI CMOS devices. Second, the ability to activate dopants in the source/drain-regions of CMOS in bulk-silicon is limited by the thermal-budget on vertical- and lateral-diffusion effects, while in FD-SOI MOSFETs only the lateral-diffusion sets a limit on this thermal-budget (a potentially less difficult problem to solve). Third, in CMOS-on bulk-silicon both the vertical- and lateral-abruptness of the source/drain-extension doping-profile needs to be controlled, whereas in FD-SOI, only the abruptness of the lateral-doping-profile of the source/drain extension-region needs to be controlled. Fourth, the well-doping in CMOS-on-bulk-silicon needs to be carried out with high-energy ion-implanters, but in CMOS-on-SOI, no wells even need to be formed. Finally, SOI-technologies can be developed to include new device structures, such as dual-gate and high-channel-mobility (with

Si:Ge-CMOS), that provide new tools for transistor designers to increase signal speed while decreasing off-current leakage-levels. The transistor characteristics and process tradeoffs for bulk-silicon- and SOI-wafers are summarized in Table 11-2. The process challenges shift to controlling lateral-diffusion and obtaining abrupt-junction-interfaces.

One of the critical processes that will be needed to allow thin-SOI-wafers to serve as the substrates for sub-50-nm CMOS is the ability to fabricate thin-SOI wafers with high-levels of control over the SOI-Si device-layer-thickness, electrical-properties, and interface-characteristics. The Smart-Cut® and Nanocleave® processes described in Sects. 11.6.3.2 and 11.6.3.3, respectively, are two approaches that may allow these goals to be met.

The timeline for the use of the various types of "thin" SOI MOSFETs can be linked to the projections made by the ITRS-2001 (Fig. 11-7). The thickness of the Si-to-SOI surface layer ($T_{Si}$) will shrink according to the approximate design rules of $T_{Si} \sim 2 \times X_{j(S/D\text{contact region})}$ for partially depleted SOI MOSFETs (see Fig. 11-2a, which shows what is meant by the terms *"S/D contact-region"* and *"S/D extension region"*), and $T_{Si} \sim 1 \times X_{j(S/D\text{extension})}$. Up to the year 2000 timeframe, production SOI-ICs were being manufactured using PD-SOI.

For the time between 2001 to 2005 (i.e., for MOSFETs with gate lengths of 0.15-$\mu$m down to 0.07-$\mu$m), the Si-SOI layer thicknesses being used is expected to cause the type of SOI FETs to be a mix of partially-depleted (PD) CMOS for mainstream logic and ASIC devices, and fully-depleted (FD) SOI for high-performance processor and server/router devices.

For CMOS devices with gate-lengths in the range of 50-nm (beginning in about 2006), the multitude of process and performance issues described in the ITRS study will preclude the use of planar-CMOS on bulk-Si as well as in partially-depleted SOI. By migrating mainstream transistor structures to fully-depleted SOI, it may also be possible to extend CMOS circuit technologies a full decade beyond the "end-of-roadmap" for planar-CMOS on bulk-silicon. "Thin-body" SOI, with gates sizes at 20-nm on Si-SOI layers of ~10-nm are now at the threshold of research devices. If scaling continues at the historical rate of a 0.7 shrink every three years, these devices would be in production by about 2015.

## 11.4 HISTORY OF SILICON-ON-INSULATOR (SOI) TECHNOLOGY

Silicon-on-insulator technology has been available since the late 1960's, and has been used since that time for military integrated-circuit applications. However, it has taken more than 30 years for the first mainstream commercial SOI devices

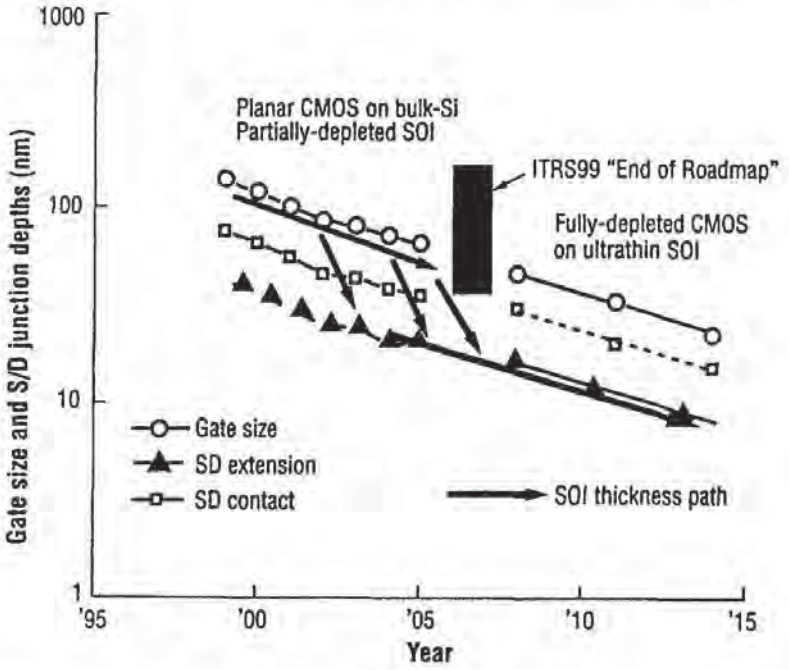



**Fig. 11-7** Timeline of Si-SOI layer thickness linked to junction depth and gate size predictions from ITRS-2001.

to be offered. As described in Sect. 11.3, the advantages of SOI technology are well known. First, circuits fabricated in SOI wafers have reduced parasitic-capacitance when compared to bulk or epi wafers. Less capacitance translates to lower power or higher speed. Second, SOI devices have improved radiation-induced single-event-upset (SEU) immunity, and thus they are useful for space applications. Even large memory arrays that are used in terrestrial applications are subject to SEU and may require SOI circuits. Third, SOI devices are completely free of latchup. Finally, the fabrication process in SOI wafers can be simplified by reducing the number of masks by as much as 30%. However, SOI wafers are more expensive to fabricate and may contain more defects than bulk or epi-silicon. As the requirements for improved speed, reduced power, and less expensive processing have become more important, a keen interest in SOI technology has been re-kindled among mainstream device suppliers.[41]

Currently there are a number of useful approaches for making SOI wafers. The first group includes the "legacy" SOI technologies of *dielectric isolation* (see Vol. 2, Chap. 2, for more information on the dielectric-isolation-technology) and *silicon-on-sapphire* (*SOS*) that were developed in the 1960's, and that even today, still find limited application. (As we will note, an improved SOS technology is currently being used to fabricate high-performance rf-devices.) The more modern SOI-technologies are based on either a buried-insulator-layer

created by ion-implantation technology (e.g., *Separation by IMplanted OXygen* [*SIMOX*]), or by a *wafer bonding* (*WB*) technique. (The insulating substrate for SOS wafers is sapphire [i.e., very pure single-crystal, aluminum oxide, $Al_2O_3$]. On the other hand, the insulator for the SIMOX and wafer-bonding technologies is silicon dioxide.) Various (and to-date, unsuccessful) attempts have been made to grow single-crystal silicon (or large-grain polycrystalline-silicon) directly on amorphous insulators.[42] When devices were first built on SOI-wafers, small islands of silicon were formed (typically by etch techniques) on top of the insulator. In modern SOI ICs, either LOCOS or STI is used to complete the dielectric isolation. Individual devices are then fashioned in the islands and these devices are then interconnected in a conventional manner. SOI-technology has been the subject of intense research and development in the late 1990s. A number of symposia that address SOI technology are held on a regular basis.[44,45] For more information on the evolution of SOI circuits, see Sect. 11.8.

## 11.5 SILICON-ON-INSULATOR DEVICES

The MOSFETs built using SOI-substrates are categorized in a number of ways. The first such grouping differentiates between so-called *thick-SOI-MOSFETs* (built on SOI-substrates whose Si layer is thicker than 1 micron - i.e., > 1.0-$\mu$m), and *thin-SOI-MOSFETs* (in which this layer is less than 1 micron thick – i.e., < 1.0-$\mu$m). The applications for "thick" SOI include power-IC switches, rf bipolar-signal-processors, and a wide variety of MEMS and optoelectronic devices. Thick-SOI devices were also historically used for a variety of military applications (e.g., radiation-hardened ICs) and for such commercial-applications as high-voltage and high-power devices. The principal SOI-material being presently used for VLSI and ULSI CMOS signal, logic, and memory applications is "thin" SOI (50-1000-nm thick [0.005-1-$\mu$m]. The "thick" SOI market segment, at only $50 million/year is a very small piece of the $110 billion/year commercial, mainstream CMOS IC market (and is not likely to grow significantly in the future). Hence, we will not further deal with such devices, but instead will focus on "thin" SOI devices in the rest of this section.

Within the category of "thin" SOI devices, there are a number of additional subcategories, namely *partially-depleted SOI-MOSFETs* and *fully-depleted SOI-MOSFETs*. The thickness of the silicon-device-layer of the SOI-structure can be chosen so that the SOI-MOSFETs will operate in one or the other of these modes. Each of these SOI device types will be discussed separately in more detail. Note once again that one of the ways that MOSFETs fabricated in SOI differ from the MOSFETs fabricated in a bulk (i.e., non-SOI) substrate is that the

channel region (or body) of the SOI-MOSFET is not automatically connected to any external terminal (i.e., intrinsically the body of the SOI MOSFET is electrically floating). This aspect of SOI-MOSFETs (a floating body), leads to so-called *floating-body-effects,* which can be detrimental to device performance. Floating-body effects in SOI-MOSFETs will be described in more detail later.

### 11.5.1 Partially-Depleted Thin-SOI-MOSFETs (PD-SOI)

*Partially-depleted, thin-SOI-MOSFETs* (PD-SOIs) are given this name because when the maximum gate-bias is applied to the MOSFET, the depth of the resultant depletion-region in the channel of such devices is shallower than the thickness of the silicon layer (Fig. 11-8). Note that "thick" SOI MOSFETs are also operated as partially-depleted MOSFETs.

As mentioned earlier, the body of SOI-MOSFETs is intrinsically floating. That is, unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating (Fig. 11-9a). However, if the SOI-MOSFET body is grounded (i.e., by independent body contacts or by body-source ties), partially depleted SOI MOSFETs behave very much like bulk-Si MOSFETs, and most of the standard $I_D(V_{GS},V_{DS})$ equations and design concepts apply. If body-contacts are not supplied, so-called *floating-body effects* arise, leading to potentially detrimental consequences (including the introduction of a *kink* in the dc I-V characteristics, a reduction in breakdown-voltage, and the causing of hysteresis and instability during dynamic operations). The issue of floating-body

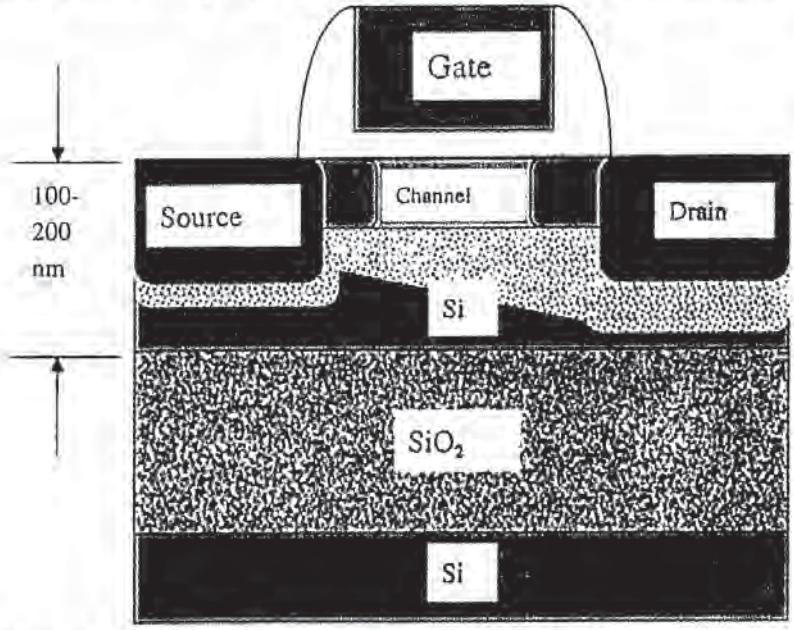



**Fig. 11-8** Partially-depleted transistor in a Si layer over an isolation layer.

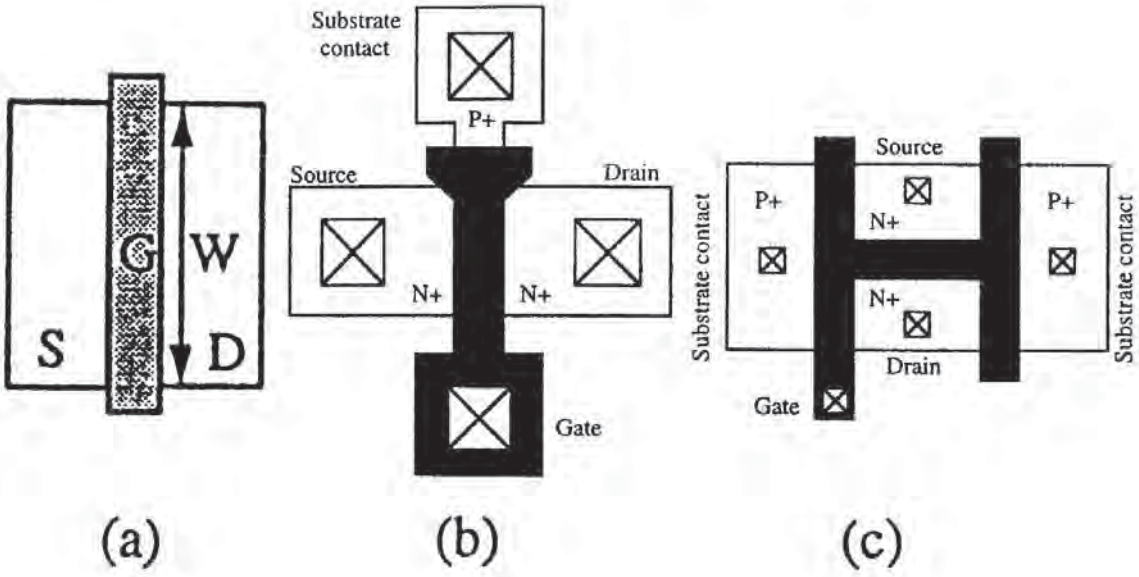



**Fig. 11-9** a) Top-down view of an SOI-MOSFET in which no provision is made for contacting the body. Unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating; b) MOSFET with body-contact made on one side of the MOSFET channel; c) An "H-gate" SOI-MOSFET (i.e., a body-contact is made on each side of the MOSFET channel).

effects in partially depleted SOI MOSFETs is described in further in Sect. 11.9.

Note that body- and substrate-contacts were routinely used when "thick" SOI devices were fabricated, as shown in Fig. 11-9b. To avoid the possible device-performance-degradation that floating-body-effects could cause, early partially-depleted, thin-SOI-MOSFETs also used either separate body-contacts or source-body ties, again as shown in Fig. 11-9b. However, since the body-region of the MOSFET beneath the channel can exhibit a high resistance, a single body-contact tied to ground may not be adequate. In some applications a body contact is made on both sides of the MOSFET channel, forming a so-called "H-gate" (Fig. 11-9c).[11] Figure 11-10 shows how a body-contact is made to an SOI-NMOSFET both from a top view, and from a cross-sectional view. Figure 11-11 shows the layout of a DRAM sense-amplifier-circuit fabricated in SOI, in which a body-to-source-tie is employed.

While body-contacts (and body-source-ties) can be used to avoid the problems associated with the floating-body-effect, these structures consume valuable silicon-circuit-area. That is, the global use of body-contacts on every SOI MOSFET in a circuit may severely degrade the density (and hence also the

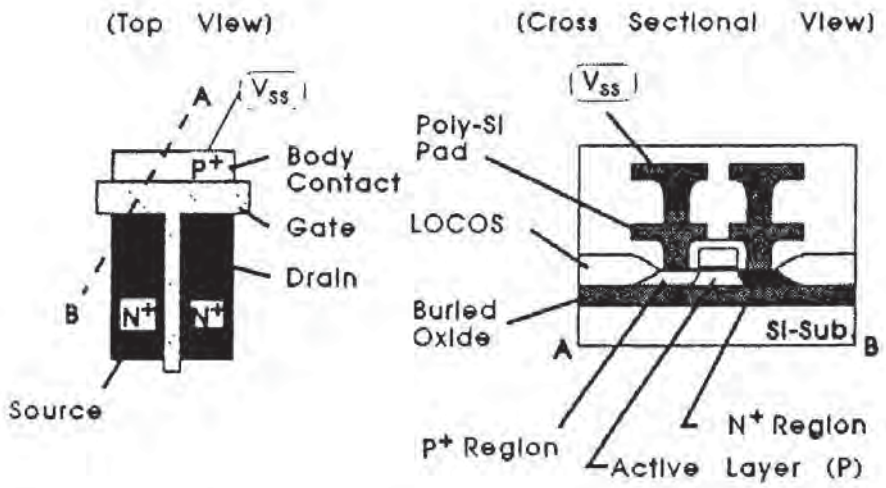



**Fig. 11-10** A top-down and diagonal cross-sectional view of an SOI-NMOSFET with a body-contact made on one side of the MOSFET channel.

performance), of a large-scale IC design. As a result, in more recent SOI ICs built using partially-depleted thin SOI MOSFETs, attempts were made to find ways to eliminate the need to use such body-contacts. As shall be seen in Sect. 11.9, it has been demonstrated that in some circuit applications, such body-contacts may not need to used. That is, the device-performance-degradation caused by floating-body-effects may be small enough to allow the elimination of body-contacts from such SOI circuits.

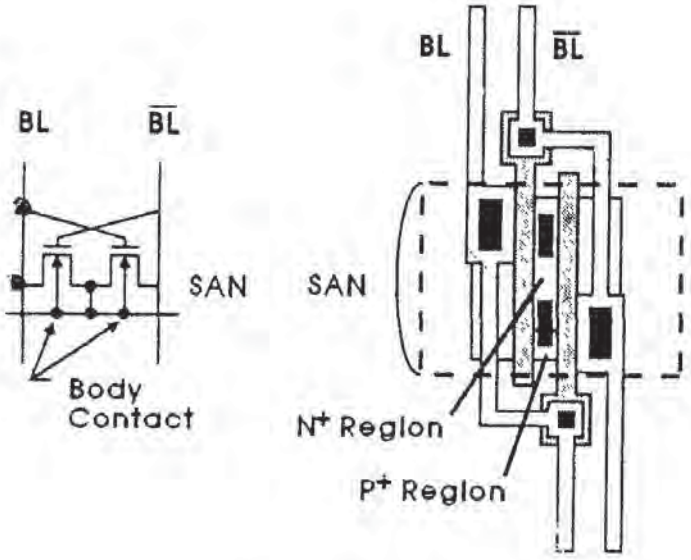



**Fig. 11-11** Layout of a SOI-NMOSFET DRAM sense amplifier incorporating a body contact.

When CMOS circuits are built with partially-depleted thin-SOI-MOSFETs (and are powered with the same voltages), switching-speeds are observed to increase by 20-30% compared to equivalent circuits in bulk-silicon devices. Alternatively, partially-depleted SOI devices can be operated at lower voltages and lower battery-power-drains with a significantly-smaller speed-penalty than bulk-silicon devices. As noted earlier, since partially-depleted SOI-MOSFETs operate electrically in essentially the same manner as bulk MOSFETs, an immediately useful advantage is that circuit designs from bulk-silicon devices can be imported into the SOI environment with only a modest investment in circuit redesign. As of 2000, the SOI circuits being offered commercially were largely based on partially-depleted SOI-MOSFETs.

### 11.5.2 Fully-Depleted Thin SOI MOSFETs (FD-SOIs)

In *fully-depleted, thin-SOI-MOSFETs* (FD-SOIs) the depth of the depletion-region in the channel of the transistor extends entirely through the thickness of the surface-silicon-layer of the SOI-structure (Fig. 11-13b), even when there is no-bias applied to the gate. (In addition, the junction-depth formed in the source/drain-regions in fully-depleted SOI-structures is also as deep as the Si-surface-layer-thickness. Hence, there is no *pn*-junction at the [source/drain]/BOX interface, and thus there is no depletion-region at that interface to contribute to the source/body- and drain/body-capacitance.) This also implies that there is less overall junction-surface area than in a CMOS-structure formed in bulk-silicon (which results in reduced leakage-currents and junction-capacitances). In the FD-SOI case, the depletion-charge is thus constant and cannot extend further into the body as the gate-bias is increased. As a result, the application of a gate-bias results in more inversion-layer-charge for the same gate-bias as in a bulk-Si MOSFET. This has several beneficial effects, including: **1**) drain current is increased; and **2**) subthreshold-slope becomes steeper (Fig. 11-12). However, the $I_D(V_{GS},V_{DS})$ equations used to model bulk MOSFET and PD-SOI MOSFETS are no longer valid, and new I-V relationships need to be used for FD-SOIs. Although circuit-designs need to take full account of the special conditions of operation in an SOI layer, the highest gains in circuit-speed, reduced-power requirements, and the highest level of soft-error-immunity are realized with fully-depleted SOI-devices. That is, FD-SOI devices exhibit the following advantages over bulk and partially-depleted SOI-MOSFETs: **1**) sharper subthreshold slope;[12] **2**) higher drive-current; **3**) higher-transconductance; **4**) reduced parasitic-capacitance; and **5**) reduced short-channel-effects.[13] The latter benefit is gained because some of the charge that would otherwise be needed to support the

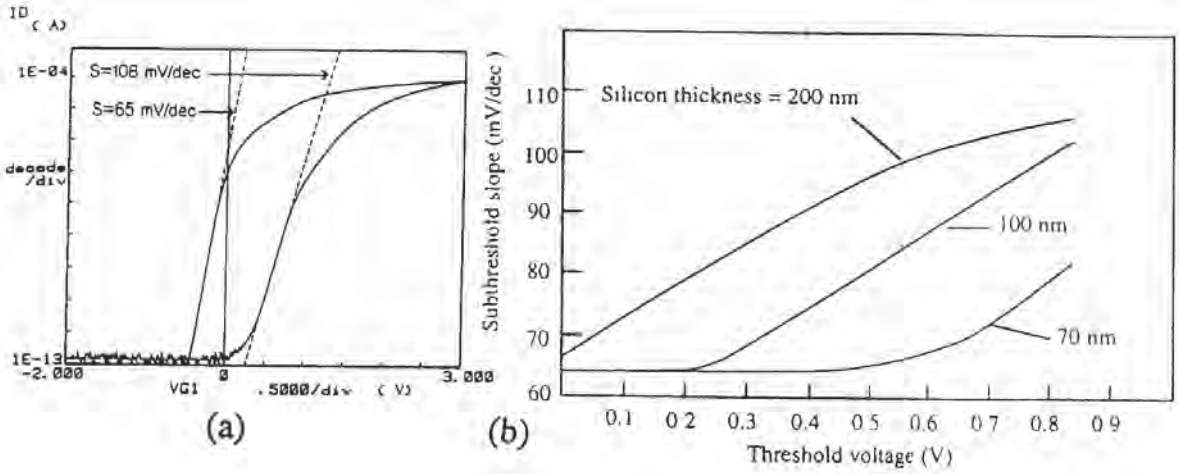



**Fig. 11-12** a) Log-plot of drain-current $I_D$ versus gate-voltage in SOI-MOSFETs made with two different silicon film thicknesses: 90-nm (left curve) and 450-nm (right curve). The inverse-subthreshold-slope, $S_t$ (in millivolts per decade) is indicated. $V_{DS}$ = 1 V; b) $S_t$ as a function of threshold voltage for various SOI film thicknesses (*n*-channel device).

channel depletion-region-charge is instead produced by charge in the substrate under the BOX. However, as of 2001, fully-depleted SOI MOSFETs were not yet being widely implemented in commercially available ICs.

The reasons fully-depleted SOI MOSFETs were not being used in commercial production by 2001 is that these devices must meet the requirement that the silicon-film-thickness must always remain well below the depletion-region-width in the channel, during all modes of the MOSFET operation (i.e., from when the device is OFF through inversion). This dictates the use of a low threshold-voltage-value, and the ability to keep the Si-SOI surface-layer-thickness well below the depletion-region-width under conditions of thickness variation. That is, it must be possible to control the Si-surface-layer-thickness to within 5-10-nm.[14] This imposes severe limitations on the manufacturability of the device. (Use of partially-depleted SOI MOSFETs in an SOI-circuit, on the other hand, alleviates the constraint on $V_T$ and its sensitivity, allowing the channel-doping-profile to be tailored to meet any desired $V_T$ value, thus easing the manufacturing problem.) Nevertheless, by the end of 2001 this situation was starting to change. Intel announced a 50-nm FD-SOI structure, and said that they would use FD-SOI MOSFETs to build ICs from 2005 and beyond.[66] In addition to FD-SOI, the Intel MOSFETs incorporate elevated source/drains (see Sect. 5.3.2).

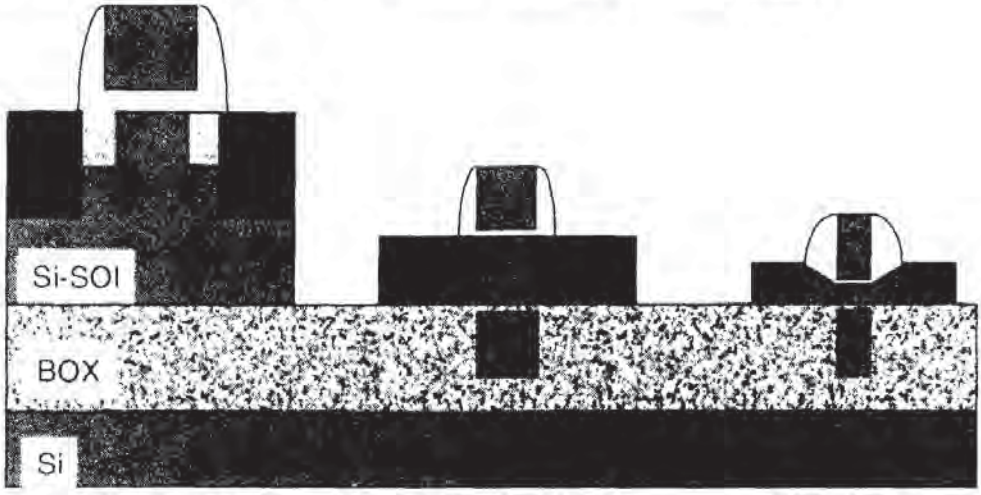



**Fig. 11-13** Schematics of partially-, fully-depleted, and "thin-body" (double gate) SOI MOSFETs.

In FD-SOIs some of the gate-charge is balanced by charge in the depletion-region of the channel, but additional charge will arise in the substrate beneath the BOX layer. This induced charge in the substrate will have the effect of trying to establish an inversion layer *along the interface between the Si surface layer and the BOX* (i.e., the substrate beneath the BOX acts as a *back-gate*). As a result, a coupling exists between the potentials on the top-gate and the back-gate, and the electrical characteristics of the channel along the back-gate vary with bias applied to the top-gate. For example, leakage along the back-gate-channel will be caused by the application of a bias on the top-gate. In fact, when the Si-surface-layer of an FD-SOI becomes too thin, the top-gate eventually can no adequately longer turn-off the FD-SOI MOSFET. In such cases it may be necessary to fabricate these FD-SOI MOSFETs with a dual-gate structure and a thinned body-region (as discussed in Sect. 11.5.3.2 and as shown schematically in Fig. 11-13c). Such dual-gate FD-SOI-devices offer the prospect of excellent control of the off-current of small channel-length SOI-MOSFETs.

A dual-gate FD-SOI has both a top-gate and bottom-gate (Fig. 11-13c). The channel is sufficiently thin so as to be fully-depleted and an inversion-layer is established at both the top and bottom gates. To minimize short-channel-effects the channel-thickness must be considerably smaller than the gate-length.[48] For gate lengths smaller than 50-nm, this will require thin silicon layers of 20-nm or

less. The potential transconductance of a dual-gate device is approximately twice that of a bulk-device because of the presence of two gates. Also, the subthreshold slope, $S_t$, is potentially much improved over a conventional MOS device. Several theoretical studies have shown that the silicon thickness for a single-gate FD-SOI device must be about 1/2 that of a dual-gate device for equivalent short-channel effects.[48]

However, by 2000, dual-gate FD-SOIs were still in their infancy, and were far from being ready for consideration as devices in commercial ICs. That is, there are major problems in the practical manufacturing of dual-gate FD-SOIs. In a horizontal configuration, a gate must be fabricated below a thin silicon layer and accurate alignment of the top-gate over the bottom-gate must be achieved. All of these steps are very difficult.

### 11.5.3 Process Technologies Used to Fabricate Fully-Depleted SOI Devices

As noted earlier, one reason it is difficult to manufacture FD-SOI is that it must be possible to fabricate very thin Si-SOI surface-layers and to control their thickness to within 5-10-nm. Here some of the process sequences that have been developed to try to meet these requirements are listed.

**11.5.3.1 FD-SOI Fabrication by Thinning the Si-SOI Surface Layer by Thermal Oxidation:** An early process for fabricating FD-SOI devices was reported in 1992.[15] In this work a 220-nm-thick Si-SOI layer formed with a SIMOX process was the starting substrate. This layer was thinned to a 40-nm-thickness by carrying out a blanket thermal-oxidation of the Si-SOI layer and then etching off the oxide in an HF etch step. A LOCOS isolation-structure was used to provide lateral isolation. After forming the gate, a titanium salicide process was used to form $TiSi_2$ on the source-drain regions and on the polysilicon gate. The silicon in the source-drain regions was completely converted to $TiSi_2$ so that the silicide layer lies on top of the buried-oxide. A silicidation that fully consumes the SOI layer in the source/drain regions was chosen for ease of control. Figure 11-14 shows a cross-section of the finished device.

**11.5.3.2 FD-SOI Fabrication by a LOCOS-Recess of the MOSFET Channel-Region:** As the thickness of the Si-SOI layer ($T_{Si}$) has to become even thinner than the 40-nm-thick Si layer used in the FD-SOI device of the previous section, the parasitic resistance of the source/drain regions of such devices becomes too large to obtain the required device performance characteristics (even if these regions

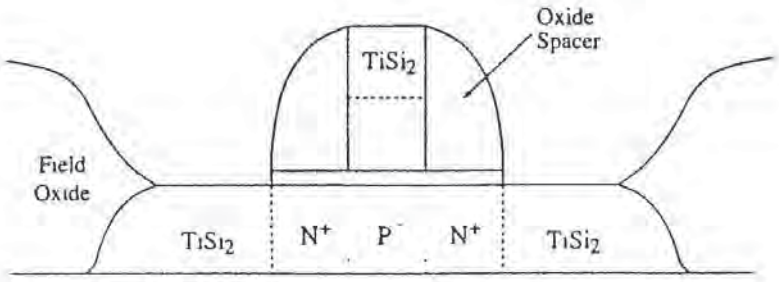



**Fig. 11-14** Cross-section of the ultrathin SOI transistor with silicided source/drain and gate.

are silicided). An approach which allows the S/D regions to remain thicker while reducing the thickness of the body has thus been developed. In this method, the body region of the FD-SOI device is thinned by using LOCOS process, as shown in Fig. 11-15.[16] In one example, a SOI wafer with a $T_{Si}$ of 195-nm was the starting material. A LOCOS process was used to thin the body region of the FD-SOI device to a 110-nm thickness. MESA active islands were formed by selectively etching back the Si-SOI to the BOX layer. These MESA islands were then oxidized to provide the lateral isolation.[17] Figure 11-16 shows an SEM of a completely fabricated FD-SOI device.

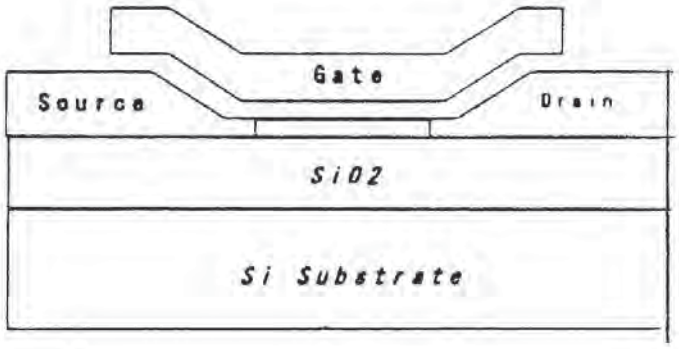



**Fig. 11-15** Cross-section of a LOCOS-thinned "thin-body" SOI MOSFET.

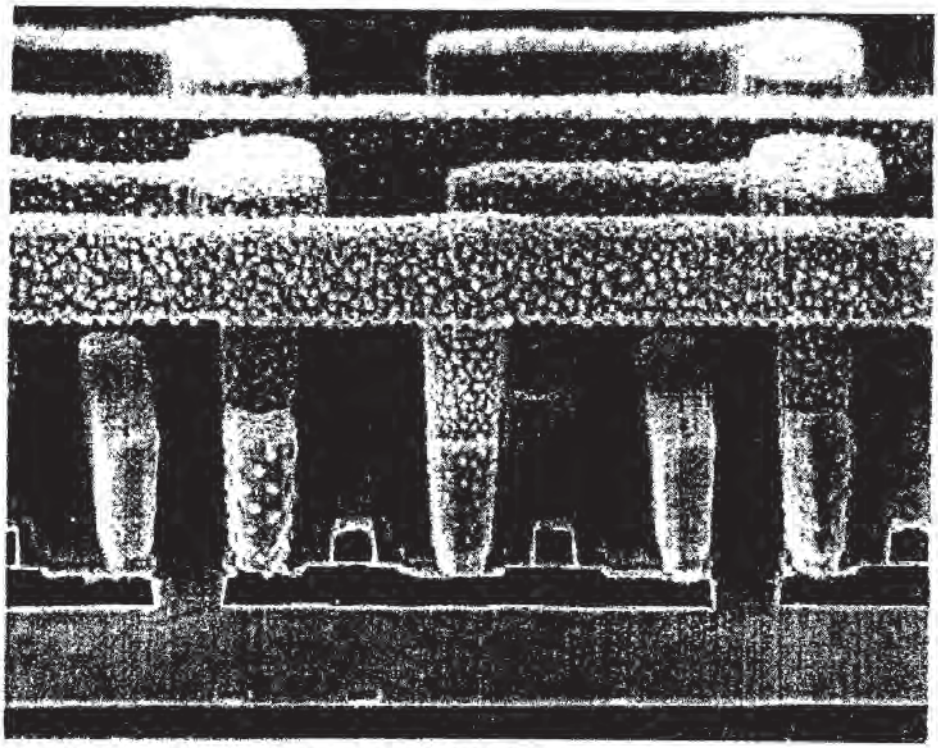


**Fig. 11-16** SEM of completely fabricated SOI-MOSFET.

## 11.6 FABRICATION TECHNOLOGIES USED TO PRODUCE STARTING SOI-WAFERS

There a number of methods for fabricating starting SOI-wafers. Here those being currently used to make SOI wafers will be discussed. The annual production of SOI wafers in 1998 was only 10-million square-inches of silicon wafer. However, it is predicted that the total SOI-wafer-demand will grow by a compound annual growth rate (CAGR) of 55% between 1999 and 2005. In 2005, this should make the production of SOI wafers reach 187-million square-inches. The SOI merchant-wafer market registered $100 in revenue in 1999 (this did not include the captive SOI wafer production by such companies as IBM).

### 11.6.1 Silicon-on-Sapphire (SOS)

The work on silicon-on-sapphire (SOS) dates back to the 1960s when there was interest in this technology for its well-known immunity to radiation damage. Sapphire is insulating, and so circuits built in SOS substrates have high speed and excellent device-isolation. That is, unlike in other SOI approaches (where a buried oxide-layer is used to provide vertical isolation), the entire sapphire substrate is an insulating material. This reduces microwave losses and makes SOS an ideal candidate for the integration of rf-circuits for wireless

communications. Furthermore, the high thermal-conductivity of sapphire facilitates thermal dissipation through the substrate, which alleviates the self-heating effects of the other SOI technologies as described in Sect. 11.9. Device isolation can be easily achieved through simply removing the conducting silicon between the transistors by etching down to the insulating sapphire-substrate. SOS circuits were historically used for radiation-hardened applications. There are several excellent review articles on this subject and the reader is referred to them for additional details.[18,19]

Since SOS technology is limited to fabrication of MOS devices, a (100) orientation of the crystalline-silicon film is desirable. This orientation has been hetero-epitaxially grown on several different sapphire (i.e., crystalline alumina) orientations, specifically (0112), (1012), and (1102). The lattice-parameter mismatch between the Si-epi layer and the sapphire-substrate is sufficiently large to cause the generation of misfit dislocations, stacking faults, and edge dislocations in the Si-epi layer. The defect-concentration is heaviest near the substrate/epi interface and decreases as the film grows.[20] Figure 11-17 shows the stacking fault density as a function of epitaxial-film-thickness in a thick SOS film. However, current requirements for fabrication employ so-called *fully-depleted MOSFETs* (which are defined in Sect. 11.5) and these must be built in a Si-epi films only 0.1–0.3-$\mu$m thick. These films can be highly faulted, and require additional processing to improve their crystalline quality.

The accepted way to improve the film quality is the technique of amorphization (by implanting a high dose of Si into the deposited silicon

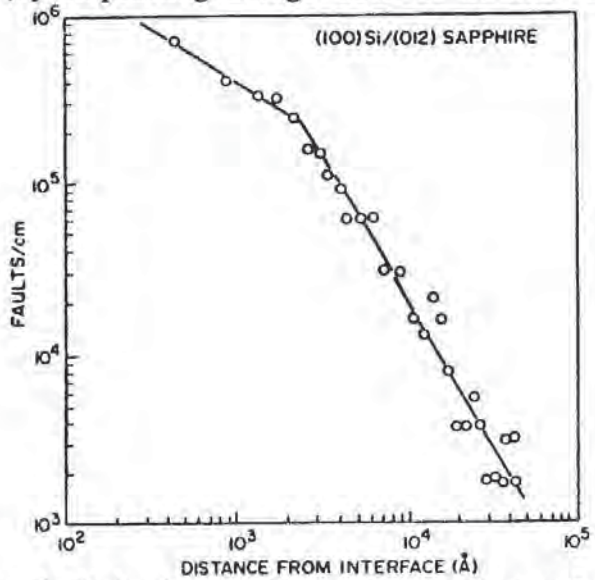



**Fig. 11-17** Stacking-fault-density as a function of the distance from the silicon/sapphire interface in an MOS structure.[21] Reprinted by permission of the American Physical Soc.

epitaxial layer) followed by *solid-phase-epitaxial regrowth (SPE)*.[22] In *thick SOS films* only a single amorphization is required. That is, an ion implantation step causes the region near the interface to become amorphized. Then the undamaged single-crystal-Si near the surface is used as seed material for the epitaxial regrowth of the amorphous layer. Since the regrowth is primarily along the <100> direction, stacking faults and twins do not propagate, leaving a controllable number of dislocations in the film.

For *thinner SOS films* a *double SPE (DSPE)* process may be required. In this approach, the Si-region near the sapphire-interface is first amorphized by a high-energy ion-implantation step, and the surface-regions are used as a seed for regrowth.[23] This is followed by a lower-energy ion-implantation step that then amorphizes the surface-region of the Si. The previously re-crystallized bottom layer is now utilized as the seed to re-grow the surface-amorphous-layer. Devices fabricated in such material show higher surface mobilities and lower junction leakage than either untreated SOS, or material obtained from a single SPE-process.[24] Peregrine Semiconductor, which markets high-performance SOS-rf devices for the commercial market, builds these devices using standard ULSI processing techniques on 200-mm SOS wafers. SOS films may require additional characterization over that used for conventional epi-films. These methods are comprehensively reviewed by Cullen *et al*,[25] and Munteanu *et al.*[26]

### 11.6.2 SOI by Separation by Implanted Oxygen (SIMOX)

A widely used approach to manufacture SOI wafers is that of *Separation by IMplanted Oxygen*, or SIMOX. This process was first proposed by Izumi of NTT in 1978 (and who also coined the acronym SIMOX).[27] The SIMOX technology

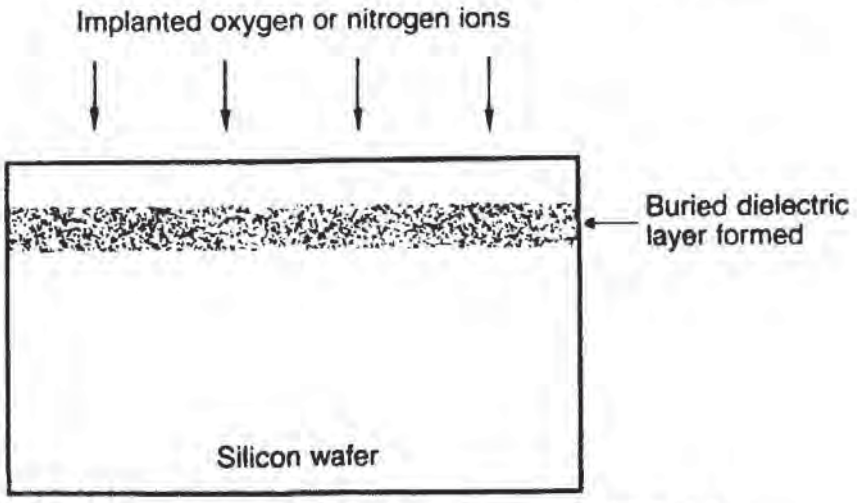



**Fig. 11-18** Implanted oxygen ion process for the formation of a SIMOX wafer.

is based on the ion-implanting a high-dose of oxygen ions ($O^+$) into a silicon wafer (~$1.4x10^{18}$ $O^+$ ions/$cm^2$), such that the peak concentration lies about a half micron beneath the silicon surface. After implantation these wafers are subjected to a high-temperature-anneal which forms a continuous stoichiometric subsurface-layer of silicon-dioxide (Fig. 11-18). This *buried-oxide* (or BOX) layer electrically separates the top and bottom portions of the wafer, and is typically 350-nm thick. The single-crystal-region of silicon above the BOX is typically 250-nm thick. The result is a wafer with a thin surface-silicon-film of high-crystalline-quality, which also has an abrupt interface to a buried layer of stoichiometric-$SiO_2$.

The SIMOX process has undergone significant evolution since it was first introduced. In early-SIMOX the material was of very-poor quality. That is, the defect-density (dislocations, oxide-precipitates, and polycrystalline-silicon-inclusions) was so high that epitaxial-silicon had to be grown on top of the silicon overlayer in order to obtain a silicon film good enough for device fabrication. In addition, at that time, the implantation of sufficient oxygen to form a buried-oxide-layer in a 2x2 $cm^2$ area on a silicon-wafer required 24 hours (using a conventional implanter). Thus, most SOI researchers viewed SIMOX as just a laboratory curiosity. Nevertheless, when NTT produced a 1-k SRAM using SIMOX in 1982, the SOI community began to consider SIMOX from a less skeptical perspective.

In 1985, Eaton (now Axcelis) produced the first high-current oxygen-implanter, the NV-200, according to specifications provided by NTT. This machine could implant $O^+$ ions at a beam-current of 100-mA and at energies up to 200-keV. The availability of such an implanter was a key step toward allowing SIMOX development to go forward.

Before 1985, SIMOX wafers were annealed at 1150°C (which is the highest temperature available in conventional furnaces equipped with quartz tubes). In 1985, it was demonstrated that annealing at 1300°C in a furnace equipped with either a polysilicon or a silicon-carbide tube (and even at higher temperatures - up to 1405°C - in a lamp-annealing system) could dramatically improve the quality of the material, and produce atomically sharp silicon/silicon-dioxide interfaces. Other improvements (to be discussed in more detail later), included the use of multiple-implants, and a lower oxygen-dose. Eventually, by the late 1980s, SIMOX-wafers began to be produced commercially by publicly-traded companies (e.g., Ibis Technology).

As a result of all of this progress, three forms of SIMOX have been developed. The first two forms (*high-dose SIMOX* and *low-dose SIMOX*) use

conventional ion-implantation technology to implant a high-concentration of oxygen. A third version of the SIMOX process, which uses a plasma immersion implanter for the oxygen implant, is called *Separation by Plasma IMplantation of OXygen* (SPIMOX). The latter technique is a potentially low-cost process for fabricating SOI substrates, because the oxygen atoms are implanted by *plasma-immersion ion-implantation* (PIII, see Fig. 11-26). The whole wafer is implanted at once and implantation is thus independent of wafer size. This results in high wafer throughput. A possible drawback of this technique is the lack of ion-beam-selection and the fact that this process may cause more contamination in the SOI wafer than would a conventional ion-implanter.[28] High-dose SIMOX was the first technique to be developed, and it was followed more recently by the low-dose SIMOX-process. The high-dose SIMOX and low-dose SIMOX forms are described in more detail here.

Although SIMOX wafers can be obtained from several commercial suppliers, the high-dose SIMOX wafers are still significantly more costly than bulk-Si or even epi-Si wafers.[29] However, the development of the low-dose SIMOX process appears to significantly cut the premium that must be paid for SIMOX wafers. In fact, it is estimated that SIMOX wafers in the future may not cost any more than epitaxial wafers (whereas in 2000 they cost 2-3x the cost of epi wafers). Future cost-reductions in SIMOX wafers will come from such advances as increased beam-current (e.g., 2x that of present implanter-technology). Note that the SIMOX process has only two major steps: **1**) oxygen implantation; and **2**) annealing. Since SIMOX wafers are commercially available from various suppliers, introducing SIMOX wafers into production does not require any major infrastructure changes in a fab. Furthermore, after years of development work, SIMOX-SOI-wafer and device-fabrication processes have achieved sufficiently high-quality and low-cost to be applied commercially. In August 1998, IBM announced it would apply its leading-edge CMOS technology to commercialize SOI-CMOS products based on low-dose-SIMOX SOI.

### 11.6.2.1 High-Dose SIMOX

Figure 11-18 shows a schematic of the high-dose SIMOX process. The sequence of steps involved in the high-dose SIMOX process is as follows:

**1.** The starting material is typically a (100) device-quality Si-wafer. The wafer is first subjected to a high dose ($1\text{-}2\text{x}10^{18}/\text{cm}^2$) oxygen ($O^+$) ion-implantation step at high enough energy (150–300 keV) so that the peak (projected range) of the implant is deep within the silicon (0.3–0.5-$\mu$m).

This process is carried out using a dedicated high-current oxygen-ion-implanter, typically with a beam current of 100-mA. For reasons described in the next paragraph, the wafer is maintained at a temperature of ~600°C during the oxygen-implant step.

Note that stoichiometric-silicon-dioxide contains $4.4x10^{22}$ oxygen atoms/$cm^3$. Therefore, from such considerations alone, an implantation dose of $4.4x10^{17}/cm^2$ should be sufficient to produce a 100-nm-thick buried-oxide-layer. Unfortunately, due to the statistical nature of ion implantation, the oxygen-profile in silicon does not have a box shape, but rather a Gaussian profile. Thus, these implanted-oxygen atoms will spread out over a wider range than 100-nm, and silicon-dioxide-stoichiometry is not reached. If the wafer was annealed after such a $1\text{-}2x10^{18}/cm^2$ oxygen-dose, oxygen precipitates would be formed at a depth equal to the depth of the maximum oxygen-concentration, but no continuous layer of $SiO_2$ would be formed. Experiments have shown that a dose of $1.4x10^{18}$ oxygen atoms/$cm^2$ must be implanted in order to create a continuous-buried-oxide-layer. The standard dose (which is the most common) is $1.8x10^{18}/cm^2$, and produces a 400-nm-thick buried-oxide-layer upon annealing. As the dose is increased above $1.8x10^{18}/cm^2$, the overall oxygen-implant-profile no longer shows a higher peak value, but instead becomes broadened. This is due to the fact that at the elevated temperature of the implant, oxygen can readily diffuse to the $Si/SiO_2$ interface, where oxidation occurs. The dose of oxygen at which the continuous-buried-oxide-layer begins to form ($1.4x10^{18}/cm^2$) is called the "*critical dose*."

**2.** The wafers are then given a post-implant anneal for sufficient time (3–5 -hours) at an extremely high temperature (1300-1350°C [in furnaces equipped with poly-crystalline-silicon or silicon-carbide tubes], are now considered standard annealing conditions) in $N_2$ (with about 2% oxygen added, to protect the silicon from the pitting phenomenon which occurs when silicon is annealed in pure-nitrogen at high temperatures), or Ar. It is generally agreed that the higher the annealing temperature, the better the quality of the SIMOX material. An anneal in pure-nitrogen can also be performed if a cap-layer of oxide (usually deposited by CVD) is formed on the wafer after the implant, and prior to the anneal.

This extremely-high-temperature annealing-step has three functions: **1)** It drives the chemical reaction which forms a continuous-buried-amorphous-oxide (BOX) layer of silicon-dioxide near the peak of the

implantation. (The choice of a sufficiently-high $O^+$-implant-dose and energy provide the required oxygen concentration to enable stoichiometric $SiO_2$ to be formed); **2**) It removes many of the defects (dislocations) formed during the ion-implantation step; and **3**) It dissolves any $SiO_x$-precipitates in the silicon device-surface-layer and denudes this layer of oxygen as well (see Chap. 2 for a discussion of denuded-layer formation in CZ-silicon-wafers). To allow integrated-circuit devices to be fabricated on the wafer after the formation of the BOX, a thin layer of crystalline-Si must remain above the BOX-layer. The BOX-layer depth is defined by the depth of the ion-implantation and it's thickness is limited to about 0.5-$\mu$m.

The evolution of the SIMOX structure during the high-temperature-anneal (as mentioned above) is quite complicated, but it is useful to summarize the phenomena that occur from a thermodynamic perspective. During annealing, both dissolution of the oxide-precipitates and precipitation of the dissolved-oxide takes place in the oxygen-rich silicon-layers. Oxygen diffuses out of the top silicon-surface, initially denuding the zone near the silicon-surface. Within the bulk of the silicon, small precipitates dissolve into silicon and large precipitates grow from the dissolved oxygen. At any given temperature (and for a given concentration of oxygen in the silicon), there exists a *critical precipitate radius*, $r_c$, below which a precipitate will disappear, and above which it is stable. The critical precipitate radius increases with temperature, and becomes essentially infinite at very-high temperatures. At temperatures above 1300°C the only stable precipitate is the buried-oxide-layer itself, which has an essentially infinite radius-of-curvature. The complete dissolution of all precipitates at annealing temperatures greater than 1300°C, and the diffusion of dissolved oxygen towards the buried-oxide-layer explain the thickness increase of the BOX when high-temperature annealing is performed.

The fact that a thin-layer of high-quality crystalline-silicon must exist above the BOX-layer implies that the lattice structure of the silicon in this region must maintain a high-degree of crystallinity in the face of the high-dose oxygen-implant step. The high oxygen-implantation-dose does amorphize the layer of silicon that lies above the projected range. However, it would be unacceptable if complete amorphization occurred, and a polycrystalline layer was formed during further annealing. This would happen if the temperature of the wafer was kept too low during the oxygen implant step. On the other hand, by keeping wafers at a temperature greater than 500°C during the implant (by a combination of

beam power and external heating), the amorphization damage anneals out during the implantation process ("self-annealing"), and a high-quality crystal-structure is maintained in the silicon layer at the wafer surface. By maintaining such an elevated temperature during the implant, the implanted oxygen also tends to diffuse away from the vicinity of the projected range to form a broad, flat concentration plateau. Note that the most commonly used implantation temperatures range between 600°C and 650°C. This is because $SiO_2$ precipitates form in the silicon overlayer at higher temperatures (700-800°C), mostly near the bottom interface. However, large precipitates are also observed near the surface region of the silicon film (and this therefore sets an upper limit of 700°C on the implantation temperature).

**3.** If the device requires thicker silicon layer than can be achieved with the depth of ion-implanted oxygen, an epitaxial-Si layer can be deposited to provide this additional thickness.

The are several shortcomings associated with the high-dose SIMOX technology, including the following:

**1.** The formation of SIMOX wafers requires extremely high doses of oxygen, and this necessitates very-long implantation times (which, as we shall see, contributes to the high-cost of SIMOX-wafers). The need to anneal the wafers at the extreme temperatures exceeding 1300°C may also be a limitation for 300-mm wafers.

**2.** Transition metals may be sputtered from the beam-line or end-station components during the implantation. These metals can deposit onto the SIMOX wafer surface and can then be driven into the silicon by the wafer-heating and subsequent thermal processing. Their presence in the silicon will enhance junction-leakage and reduce the minority-carrier-lifetime of the devices. Metallic contamination is present in all implant processes, but since high-dose SIMOX implants are nearly a thousand times greater than those of conventional high-dose source/drain implants, the metal contamination problem is potentially increased by approximately the same degree.

This issue was especially troublesome during the early days of SIMOX, when implanters were not yet built to limit transition-metal-sputtering. Since that time, however, great efforts were made to reduce the metal contamination sources, and silicon or silicon-dioxide shields have been placed in implanters such that the beam never "sees" any metallic part. The

design of high-temperature furnaces used to anneal the SIMOX material has also been optimized in order to minimize the diffusion of metallic contaminants from the furnace heating-elements into the furnace tubes.

**3.** Pinholes in the buried-oxide-layer will occur if a particle is present on the wafer during the oxygen-implantation-process. Such particles screen the oxygen implant at that point. The result is a breach in the buried-oxide-layer that represents a flaw in the insulating barrier at that point. Careful wafer cleaning and attention to particulate control in the implant process can control this problem.

**4.** Despite the high-temperature anneal, the surface silicon may retain a dislocation density on the order of $10^3/cm^2$. The presence of such defects can cause reduced oxide-integrity and increased junction-leakage. However, devices fabricated in the best SIMOX material have been shown to be able to build excellent quality CMOS and bipolar ULSI circuits.[30]

**5.** SIMOX-wafers are more costly than bulk-Si-wafers, although the cost is expected to sharply decrease as the popularity of this technique expands. The cost reduction will be partly achieved by improvement in the implant equipment used to fabricate SIMOX-wafers. For example, SIMOX-dedicated ion-implanters are now available from IBIS Technology. Such implanters can furnish adequate throughput and acceptably-low levels of metallic-contamination. These production tools have an output of 100-mA of $O^+$ at an acceleration voltage 210-keV. The total transition-metal contamination on the wafers (from Ti to Zn) has been measured to be less than $5x10^{10}$ atoms/$cm^2$.

### 11.6.2.2 Low-Dose SIMOX

A recent trend with SIMOX fabrication is to use a lower oxygen-implant-dose than is used in the high-dose SIMOX approach. With the low-dose technique, improved, lower-cost SOI substrates are obtained. For example, a typical modern SIMOX implanter with a nominal 50-mA of $O^+$ beam current has a maximum output of only about 25 wafers of 200-mm diameter per day, round the clock operation. The low-dose SIMOX process also drastically increases the crystal quality of the silicon layer above the stoichiometric BOX layer. The oxygen implant dose is reduced to about $4x10^{17}$-$1.0x10^{18}$ ions/$cm^2$, and at an energy around 120-keV. Note that only a very-narrow process window close to the dose of $4.0x10^{17}$ ions/$cm^2$ will produce a continuous, precipitate-free BOX layer.

Table 11-3 compares the process conditions used to produce SIMOX with high-, medium-, and low-doses of oxygen.

A silicon-layer between 150-200-nm deep is produced, but the lower oxygen-implant-dose also reduces the BOX layer to thicknesses of only 90-150-nm (depending on the selected oxygen dose).[31] A high-temperature anneal in Ar (i.e., at 1350°C) is then carried out, just as is done in the high-dose SIMOX process. However, this high-temperature-anneal in an inert ambient is followed by a subsequent high-temperature anneal in a gas mixture of inert gas and 30-60% oxygen (at ~1350°C).

The oxygen in the furnace-ambient at this high-temperature both oxidizes some of the superficial silicon layer, but also diffuses through the surface silicon layer and reacts at the buried-oxide/Si-interface to form new oxide within the Si (Fig. 11-19). This has the effect of increasing the thickness of the BOX-layer.[31] This process of growing the BOX-layer to a greater thickness by annealing the low-dose SIMOX wafers in oxygen is called an *internal thermal oxidation process* (*ITOX*). Note that as long as the oxide grown on the surface of the silicon during ITOX is thinner than 500-nm, there exists a linear-relationship between the thickness of this oxide layer ($t_{ox}$) and the thickness increase in the buried

**TABLE 11-3 Physical and Electrical Properties of Commercial SIMOX**

| Parameter | Units | Standard Dose SIMOX | Medium Dose SIMOX | Low Dose SIMOX |
|---|---|---|---|---|
| Silicon Thickness | A° | 1500 - 2000 | 1500 - 2200 | 1500 - 1800 |
| BOX Thickness | A° | 370 - 400 | 170 - 190 | 90 - 150 |
| HF-Defects Density | $cm^{-2}$ | < 1 | < 1 | < 1 |
| Dislocation Density | $cm^{-2}$ | $< 10^7$ | $< 10^6$ | $< 10^5$ |
| Stacking Fault Tetrahedara | $cm^{-2}$ | $< 10^5$ | $10^5 - 10^6$ | $10^5 - 10^6$ |
| Surface Roughness (10um x 10um) | A° | 7 - 10 | 5 - 7 | 3 - 5 |
| Metallics (TXRF) | $cm^{-2}$ | $< 10^{11}$ | $< 10^{11}$ | $< 10^{11}$ |
| Short Density | $cm^{-2}$ | < 0.1 | < 2 | < 2 |
| Electron Mobility | $cm^2/V\text{-}s$ | 850 - 1000 | 850 - 1000 | 850 - 1000 |
| Hole Mobility | $cm^2/V\text{-}s$ | 200 - 325 | 200 - 325 | 200 - 325 |
| Oxide Charge | $C/cm^{-2}$ | $2 - 3 \times 10^{11}$ | $1.5 - 3 \times 10^{11}$ | $1 - 2.5 \times 10^{11}$ |
| Interface State Density | $cm^{-2}eV^{-1}$ | $2\text{-}3 \times 10^{11}$ | $1 - 3 \times 10^{11}$ | $1 - 2.5 \times 10^{11}$ |
| Transconductance (Gmsat) | mS | 400 - 440 | 900 - 1200 | 1000 - 1300 |

oxide $\Delta t_{BOX}$): $t_{ox} = 0.06\ \Delta t_{BOX}$ (Fig. 11-19c). The internal-oxide also grows at the expense of the bottom of the silicon overlayer.

In the low-dose SIMOX process, the BOX formation consists of oxygen precipitation, growth of the precipitates, and their coalescence during the high-temperature-anneal process. This sequence of events occurs because the oxygen implantation dose is less than the $SiO_2$ stoichiometry, even at the depth of the peak-oxygen-concentration, and a stoichiometric-$SiO_2$-buried-layer is required to be formed to produce effective dielectric isolation. By carrying out the ITOX process, the BOX in the low-dose-SIMOX-process is made thicker and exhibits better isolation characteristics. The upper part of the BOX after the ITOX process is essentially a thermal-oxide-layer. Since the upper segment of an ITOX-SIMOX layer is essentially a thermal oxide, such important BOX properties as the breakdown-field, show a marked improvement (i.e., the breakdown-field is increased from 3-4-MV/cm for the case of conventional low-dose SIMOX to more than 5-MV/cm for ITOX-SIMOX, see Fig. 11-19d).[49] This improvement can be attributed to the BOX-thickness incremental-increase due to ITOX. In addition, the ITOX reduces the size and density of silicon pipes in the BOX, which is the main cause of BOX breakdown.[50] In 2001 IBIS Technology licensed a low-dose-SIMOX-technology from IBM called Advantox® MLD.

In 2000, IBM said recent advances in the production of low-dose-SIMOX-wafers will cut incremental processing costs of SOI to only 3% over the costs of producing ICs on bulk-Si-wafers. Low-dose-SIMOX-wafers are commercially available from such vendors as Ibis, Komatsu, and Nippon Steel. For example, the Ibis-1000 model is currently available. The next generation oxygen implanter, the Ibis-2000 (which can accommodate both 200-mm and 300-mm wafers) was under development in 2001, and is forecast to be ready for delivery in the first half of 2002. Figure 11-20 is a photograph of a Power4™ processor made by IBM using a SIMOX-SOI-based process.

### 11.6.3 Wafer Bonding

Another popular technology for producing SOI wafers is *wafer bonding*.[32] Wafer bonding is a simple concept. Basically, wafer bonding refers to the phenomenon that mirror-polished, flat and clean wafers of almost any material, when brought into contact at room temperature, are locally attracted to each other by van der Waals forces and adhere, or *bond* to each other. The bonding process is started by locally applying a slight pressure to the wafer pair, and the bonded area spreads laterally over the whole wafer area in a few seconds. Because the bonding achieved at room temperature is typically relatively weak, wafer pairs bonded

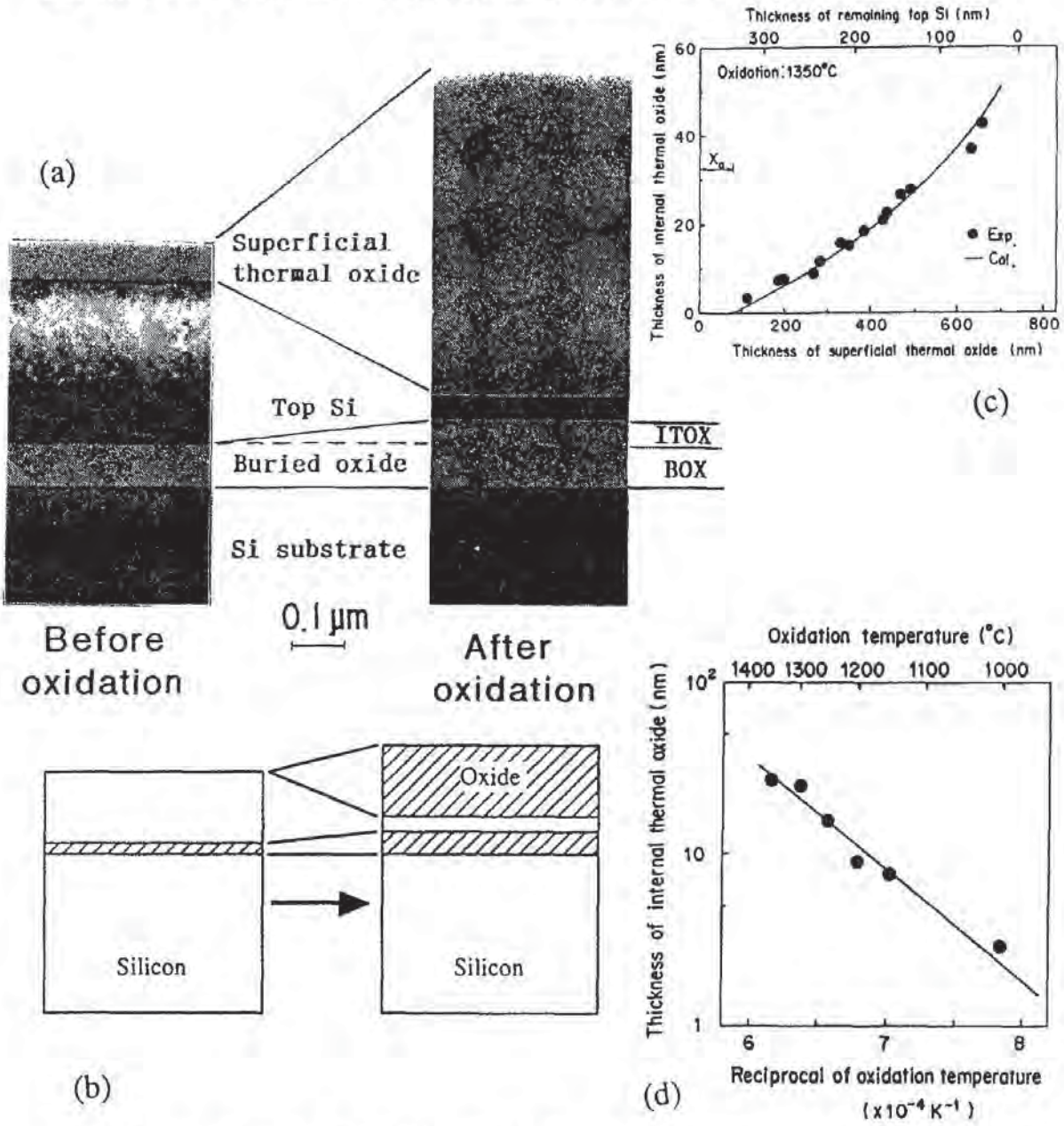



**Fig. 11-19** a) An XTEM of SIMOX wafers before and after high-temperature-oxidation at 1350°C; b) A schematic drawing of SIMOX wafers before and after high-temperature oxidation at 1350°C; c) Relationship between the thickness of the ITOX layer and the superficial thermal-oxide-layer formed by high-temperature-oxidation at 1350°C; d) Relationship between the thickness of the ITOX-layer and the reciprocal of the oxidation-temperature.

at room temperature are usually subjected to a heat treatment to strengthen the bonding interface.

In the application of the bonding-process to form SOI-substrates, two thermally-oxidized-Si wafers (or one oxidized-wafer and one bare-Si-wafer) are first

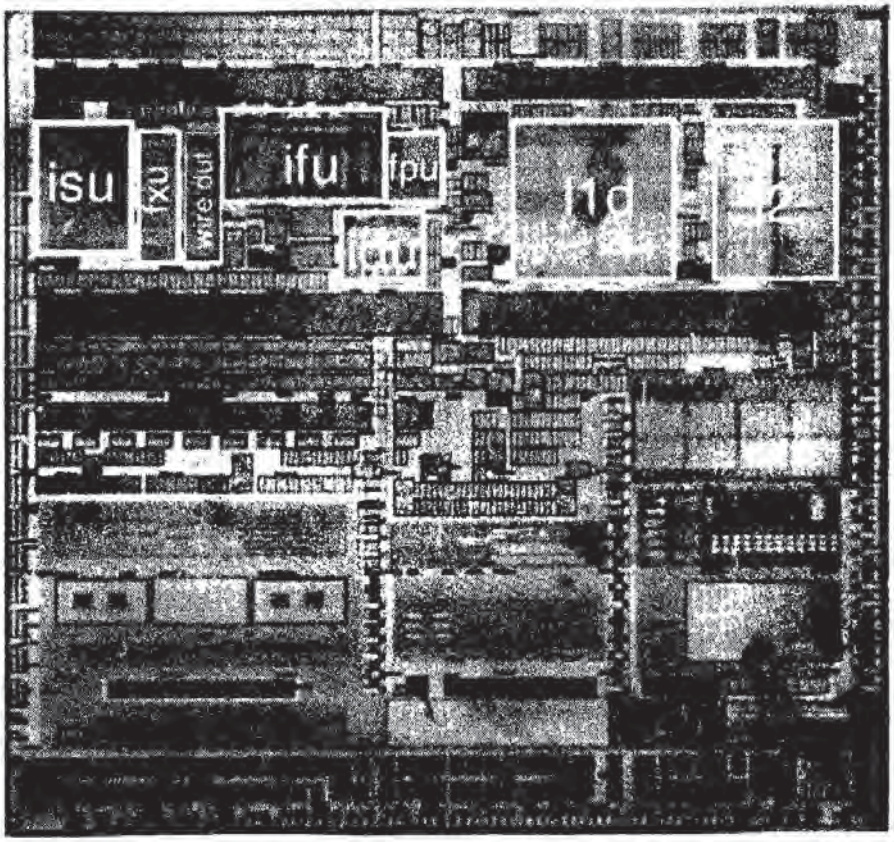



**Fig. 11-20** Photograph of an IBM Power4™ processor using a SIMOX-SOI-based process.

first bonded by applying pressure to the wafers at room temperature (Fig. 11-21). A subsequent furnace anneal is used to increase the bonding strength. After such bonding, most of the silicon from one of the wafers is removed, leaving a thinned

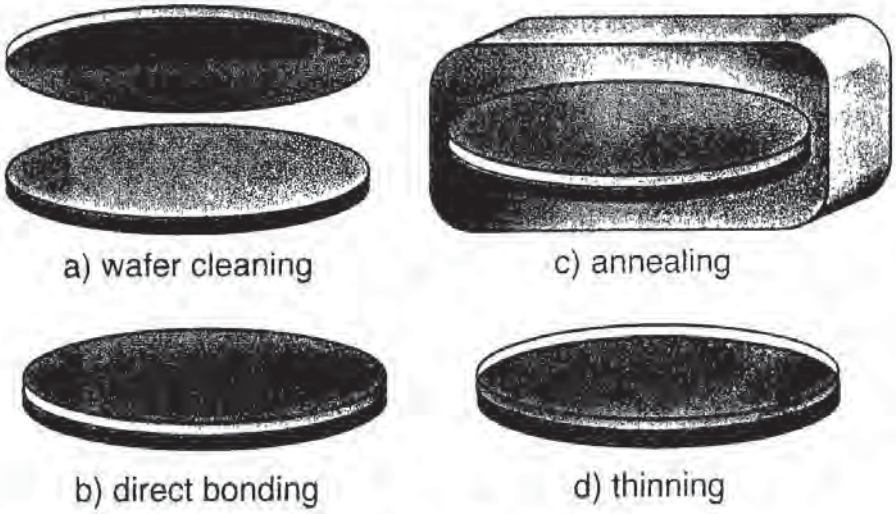



**Fig. 11-21** Generic direct-wafer-bonding process flow.

layer of silicon from one of the wafers. The IC devices are built in this remaining thinned-silicon-layer, which now exists over the oxide present at the interface of the bonded wafers (i.e., a buried-oxide). The wide-availability of chemical-mechanical polishing (CMP) in IC processing, as well as a variety of other thinning approaches has led to wide and diverse use of wafer bonding. Unlike SIMOX, the wafer-bonding technique can form BOX-layers up to 50-$\mu$m thick. A book has recently become available dealing with the topic of wafer bonding.[33]

There are several approaches for achieving bonding. In one method, two oxidized-silicon-wafers are held together (face-to-face) by applying external pressure to the wafers while they are exposed to an oxidizing-ambient at about 700°C.[34] Bonding is believed to take place by the conversion of oxygen trapped between the wafers, into $SiO_2$. The conversion consumes the oxygen and creates a partial-vacuum that forces the wafers into intimate contact, allowing the chemical-bonding-reaction to proceed. The chemical-reaction is believed to be the polymerization of silanol bonds to form siloxane networks.

A second approach applies a moderate voltage across two oxidized-wafers placed face-to-face at temperatures in the range of 1100–1200°C.[35] Since no mechanical force is required when the wafers are first placed in the furnace, this method is less likely to introduce crystalline-defects. A third approach uses one oxidized-wafer and one bare-wafer. Both are treated with an acid solution that forms O-H groups on the surface of the wafers. After drying the wafers are placed face-to-face at room-temperature in clean air. After this treatment a self-adhesive contact is formed which holds the wafers together during the bonding process. The bonding is completed by annealing the two wafers at 1100°C in a nitrogen-ambient for 4 hours.[36]

Recently it has been demonstrated that bonding can occur at temperatures as low as 150°C, if the process is performed in a moderate vacuum (e.g., several torr).[37] It is believed that the vacuum significantly reduces the amount of trapped nitrogen at the interface. Trapped nitrogen can prevent an intimate contact of the bonding-surfaces during annealing, and hinders formation of the required covalent-bonds. It is also possible to achieve full-bonding-energy directly at room-temperature under ultra-high vacuum conditions if the wafer surfaces are free of adsorbates.[38]

It should be mentioned that particles or protrusions on either surface being bonded, do prevent bonding, (at least locally) around these disturbances. These lead to locally-unbonded-areas that are frequently termed *interface bubbles* or *voids*. Such voids are, of course, undesirable, since they can lead to delamination of devices located in imperfectly-bonded-areas during chip-assembly and

packaging. Voids with a diameter of several millimeters are readily created by particles 1-$\mu$m or less in size (and these are easily revealed by infrared or ultrasonic imaging). Therefore particles on wafers have to be avoided by proper cleaning and by carrying out all bonding procedures in an ultra-clean environment (Fig. 11-22).

**11.6.3.1 Bond-and-Etch-back SOI (BESOI):** Once the wafers are bonded it is necessary to remove most of the silicon from one of the wafers to produce the SOI structure. There are several proposals being evaluated for this removal process. The *bond and etch-back SOI* (BESOI) technique is used to produce relatively thick-films of both oxide and silicon. The BESOI wafers typically have thicker, yet better-quality silicon and buried-oxide-layers than do SIMOX wafers.

The silicon-layer on BESOI substrates can be thinned in one of several ways. Mechanical grinding and polishing can be used to thin the Si to a thickness down to ~1-$\mu$m, but only to within 10-30% uniformity. Better uniformity can be achieved by using an etchstop. Several etch-stoplayers have been developed, including: a boron-doped-layer, a silicon-germanium-layer, a carbon-implanted-layer, and a porous-silicon-layer.

In the first of these etchstop methods, a lightly-doped B-region can act as an etchstop-layer when it is formed on a heavily boron-doped-substrate (Fig. 11-23). That is, the wafer that will ultimately contain the device consists of a lightly-B-doped epitaxial-layer on a heavily-B-doped-wafer. After oxidation and bonding of the lightly-doped surface of this wafer to another wafer, the heavily doped

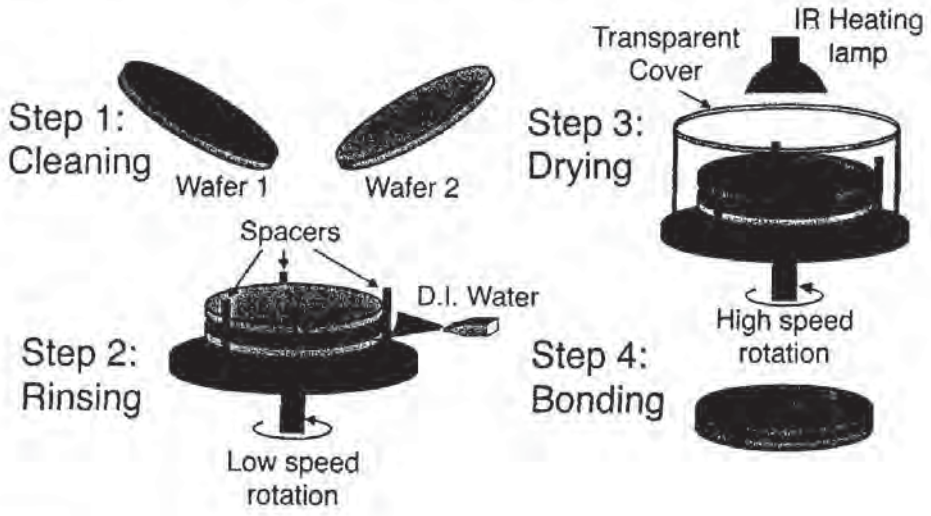



**Fig. 11-22** Schematic of the microcleanroom setup and procedure.

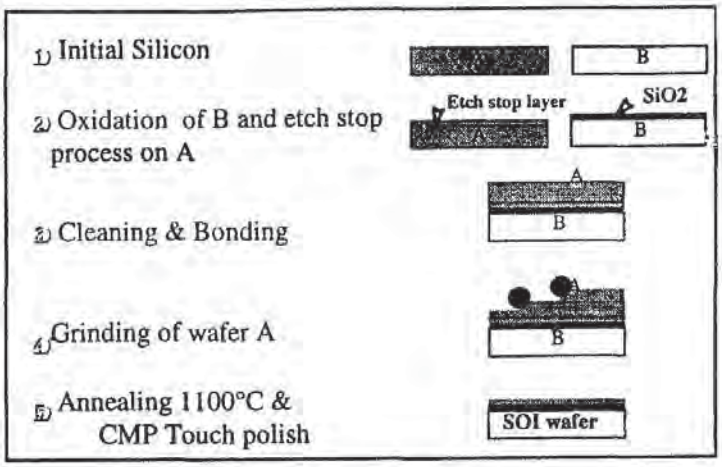



**Fig. 11-23** Schematic of the Bonded-and-Etched-Back SOI technology (*BESOI*).

substrate is removed by an etch process that preferentially etches such heavily doped silicon. The epi-layer that remains bonded to the second wafer determines the thickness of the active-device-layer and the sum of the thickness of the oxides on the two wafers determines the thickness of the buried-oxide. One of the issues with this approach is that two silicon-wafers are required to make one device-SOI- wafer. Another is that the uniformity of the thickness of the silicon-layer depends on the etch-stop-selectivity (which can vary, depending on the etch-stop method used, from 10 to $10^5$).

Such BESOI methods are widely used for "thick" SOI applications (i.e., > 1-$\mu$m), such as power devices (high-voltage and high-temperature devices) and micromechanical machines (MEMS). The power-devices are used for military, industrial, and aerospace applications. The share of the thick-film segment of the SOI market dropped from 80% in 1995 to 53% of the market in 1999 and was taken over by the thin-film segment. Future thick-film SOI applications will be driven by industrial and automotive applications.[54]

**11.6.3.2 Hydrogen-Implantation-Induced Layer Splitting to Create SOI Substrates (SMART-CUT®):** A recent (and elegant) wafer-bonding-based method has been developed for forming SOI substrates using a technique to obtain very-uniform and thin regions of single-crystal-silicon. This technique employs a process called *hydrogen-implantation-induced layer splitting*.[39] The process is also referred to by its inventor M. Bruel (and by the company that sells such SOI

substrates [SOITEC SA, France]), as *Smart-Cut.*® The Smart-Cut process also allows the sacrificial-wafer to be reused and only one wafer is therefore required to make a bonded SOI substrate. (The SOI material made with the Smart-Cut process is termed UNIBOND.®) There are two reasons why this process may be significantly less-expensive than the process used to fabricate SIMOX-wafers. First, the sacrificial wafer can be re-used. Second, because the implant-dose is an order of magnitude lower than even one in the low-dose SIMOX-process (and the implant is done at room temperature in a conventional implanter), the cost of the implant step is significantly lower, compared to those used in SIMOX.

Figure 11-24 shows how the process works. Two device-quality-wafers (labeled A and B) start the process. Wafer-A is oxidized (to a thickness of about 0.5-$\mu$m) to form what will become the buried-oxide of the SOI-structure. Then this wafer is ion-implanted with a high-dose of hydrogen ions ($\approx 5 \times 10^{16}$ $H^+$ ions/cm$^2$) through the oxide, using a conventional high-current-implanter. The peak-concentration of this implanted hydrogen-dose is at a projected range $R_p$ of about 0.2-$\mu$m below the oxide-surface. At this point, only micro-cavities or micro-bubbles are formed at the $R_p$ of the implantation. For reasons explained next, the depth of the $H^+$ ion implantation determines the thickness of silicon above the BOX.

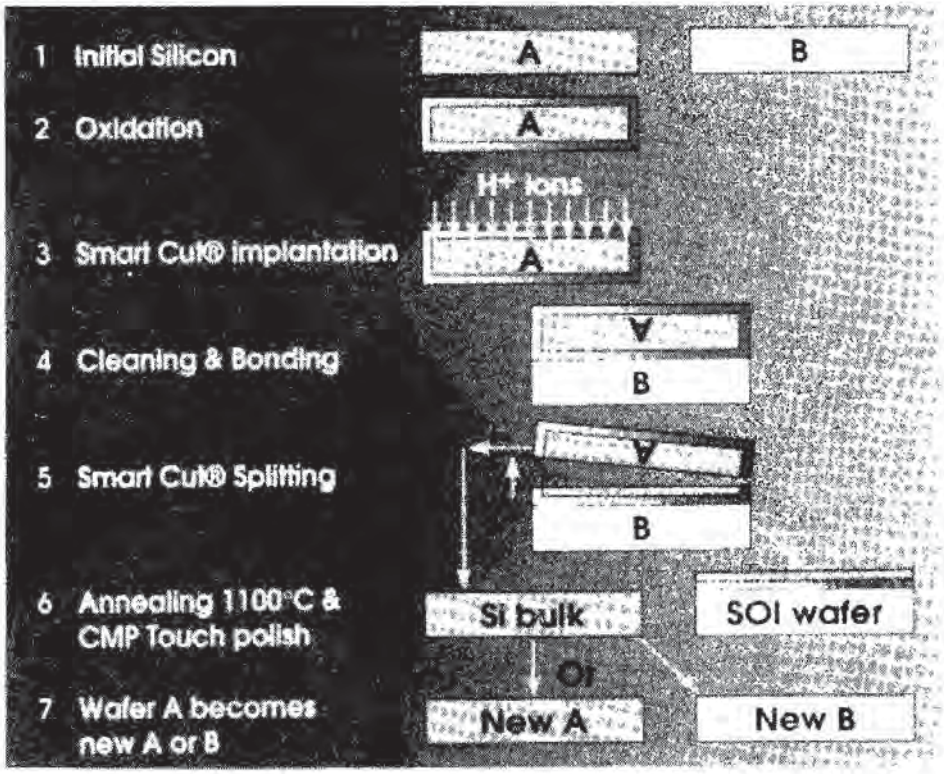



**Fig. 11-24** Process flow for fabricating Smart-Cut® wafers.

After implanting the hydrogen into Wafer-A, both Wafers A and B are cleaned and low-temperature bonded face-to-face. The bonded wafers are next annealed (at 400-600°C). During this anneal the implanted-hydrogen precipitates, and this precipitation-process merges the buried network of bubbles and voids at the vicinity of the depth of $R_p$. As these hydrogen-filled microcavities continue to grow and merge, they eventually form a continuous microcrack along the line corresponding to the peak of the implanted-hydrogen. At that point Wafer-A delaminates from along that line, and Wafer-B (which now has a BOX-layer covered by the silicon layer from Wafer-A) becomes the SOI wafer. Meanwhile, Wafer-A remains available to be used as the sacrificial wafer of another UNIBOND substrate. Both wafers are then annealed at 1100°C to remove defects. This is followed by a "light" CMP (or "touch-polish") to erase the surface-roughness resulting from the layer-splitting process. Wafer-A, which is now only a fraction of a micron thinner than it was before the start of the process is used as Wafer-B when the process starts over. The use of ion implantation for layer separation improves the layer-thickness-uniformity over that of bond-and-etch-back SOI-wafers. The method is used to prepare commercially available 200-mm SOI wafers with a silicon thickness above the BOX in the 100-nm range and a uniformity of better than 5%. The dislocation density is $< 10^2/cm^3$, the surface roughness is $< 2$ Å (rms), and metal contamination is $< 5x10^{10}/cm^3$.[40]

**11.6.3.3 Nanocleave® Wafer-Bonding SOI:** The fabrication of starting wafers for SOI can also be carried out in other ways. One of these processes, named *Nanocleave®* was introduced by Silicon Genesis Inc. in 1998.[43] The process combines growing a SiGe/Si epitaxial layer, low-temperature wafer-bonding, and a room-temperature cleaving-process to produce sub-100-nm-thick layers of single-crystal-silicon over a buried-oxide substrate. One of the key advantages of the Nanocleave process is that it can produce such thin SOI-layers with a uniformity of several percent over an entire wafer. This type of uniformity is necessary for the successful fabrication of fully-depleted CMOS-SOI-FETs. Other thinning techniques (such as etch-back or CMP have increasing difficulty maintaining such SOI-layer thickness-control when silicon layers of sub-100-nm thickness are being produced. The Nanocleave process uses an atomic-scale cleaving-process that produces cleaved surfaces with an as-cleaved surface roughness less than 5-Å (as measured by AFM). In addition, the Nanocleave process potentially offers a method that is lower in cost than any of the other fabrication processes used to fabricate SOI-substrates. The Nanocleave technique also offers another path toward the capability of fabricating both "ultrathin" Si-

SOI-layers (10-to-50-nm-thick Si) and "nanothin" Si-SOI (5-to-10-nm-thick Si) required for the MOSFETs with gate-lengths smaller than 50-nm that are scheduled to begin production by 2010.

The Nanocleave process begins with the formation of a *cleave-plane*, the growth of a thermal-oxide-layer and Si-device layer on a high-quality "donor" wafer (Fig. 11-25a).[44] This "donor" wafer is then bonded to an inexpensive oxidized "handle" wafer (on which the device layer will ultimately reside).

The cleave-plane is created by first epitaxially growing a layer of Si-Ge on the "handle wafer, and then continuing the growth of this layer with an epitaxial-Si device layer (Fig. 11-26a).[63] The buried Si-Ge-layer is a compressive-strain-layer of a non-dislocated Si-Ge alloy, which contains no porous-regions, inclusions, or voids, as shown in Fig. 11-26a. When device wafer is separated, the resulting separated surface has a surface roughness, as measured by AFM profiling of 0.2-nm, or less.

After the cleave plane has been formed and a thermal-oxide is grown on the surface of the wafer with the cleave-plane (which will become part of the buried-oxide-layer in the finished SOI-device), the "donor" wafer is bonded to the

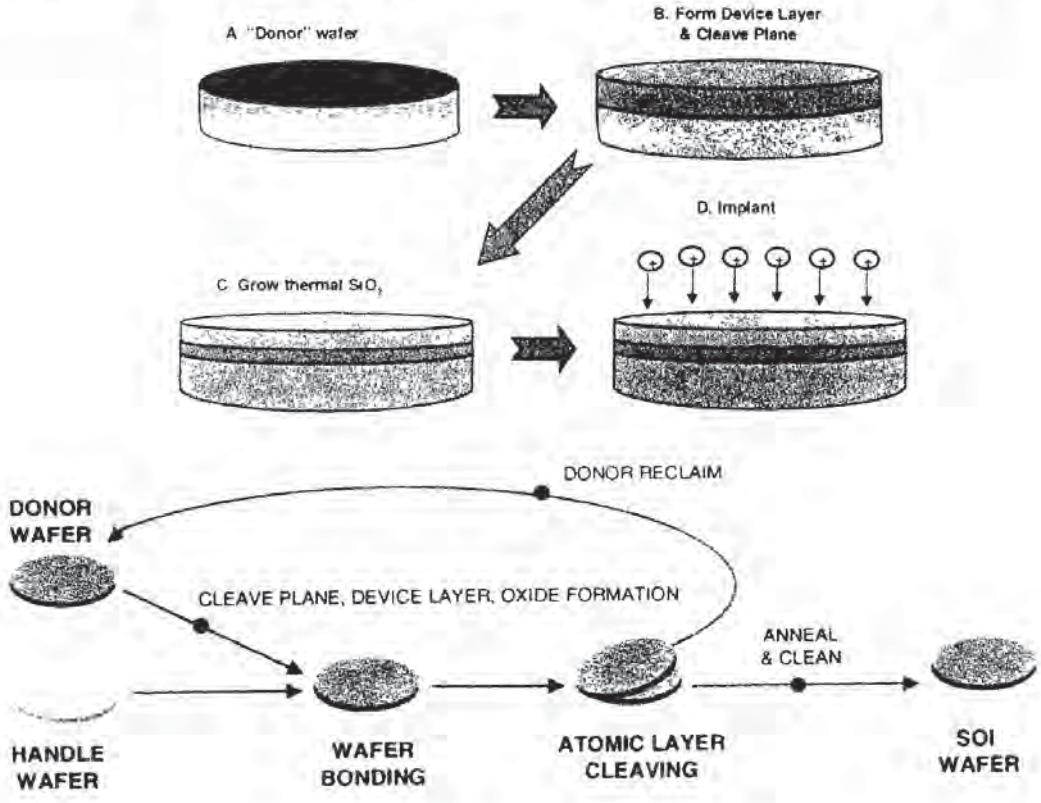



**Fig. 11-25** Process flow of the Nanocleave® SOI technology.

"handle" wafer (which also contains a surface-oxide layer). Plasma-bonding techniques are employed to increase the initial bond-strength to sufficient levels for the room-temperature process used to transfer the silicon-layer from the "donor" to the "handle" wafer. This transfer process (by cleaving the device layer from the "donor" wafer from the "handle" wafer) is carried out next (Fig. 11-25b). Since the bond-strength is significantly greater than the interface-strength along the cleave plane, a novel cleaving process is used to separate the Si device layer from the "donor" wafer (which results in the creation of an SOI structure on the "handle" wafer). In the Nanocleave process, the donor wafer cleaving is done at room temperature (unlike in the Smart-Cut-process that uses an elevated temperature anneal for the cleaving step). In the Nanocleave process, the drive force is a static pressure of $N_2$ gas. A mechanical fixture secures the bonded wafer pair as a high-pressure nitrogen-gas is introduced along the strained Si-Ge layer. That is, cleaving is controlled by fracture along a cleave-front within the Ge-rich Si-layer. A thin-film of material from the "donor" wafer remains bonded to the substrate "handle" wafer, producing an SOI-structure. The result is an SOI-wafer with a silicon-layer whose thickness uniformity is within 1% across a 200-mm wafer. Furthermore, the atomic-layer cleaving-process is claimed to produce an as-cleaved surface roughness of less than 1-nm (typically 2-5 Å).[63] This is an order-of-magnitude smoother than the 80 Å of hydrogen-induced thermal cleaving obtained with the Smart-Cut technique. The smoother surface in the Nanocleave process is acceptable for many IC applications with no additional

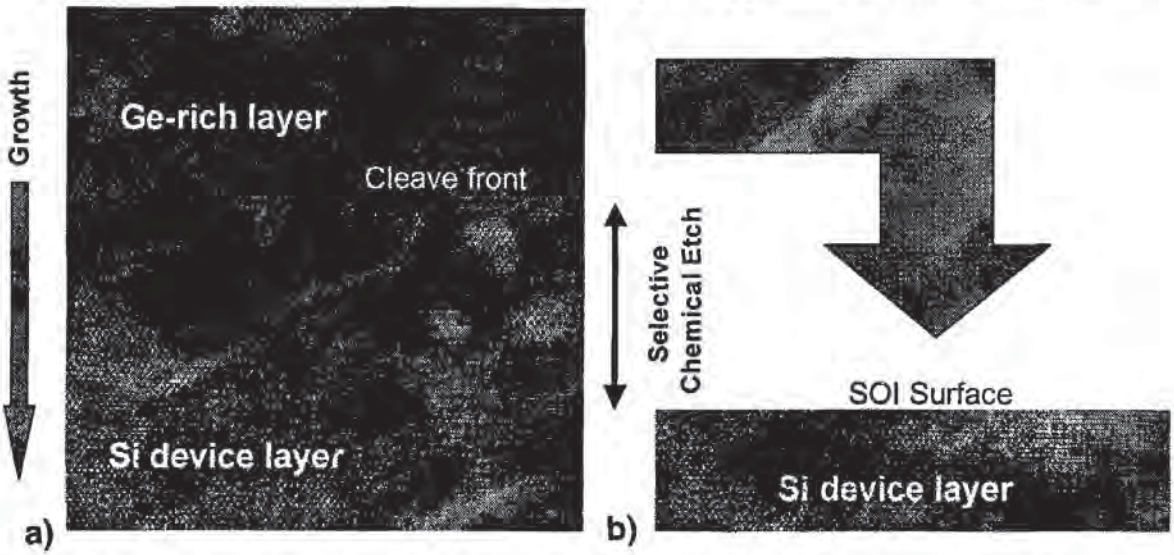



**Fig. 11-26** a) HR-TEM image of the boundary area of a Ge-rich, epitaxial strain-layer and a crystalline-Si-device-layer in a state prior to atomic-layer-cleaving; b) The residual Ge-rich layer on the as-cleaved-wafers are subsequently removed through a selective etch step.

surface polishing. The edge of the SOI-layer has a smooth and regular character without the need for edge polishing. The residual Ge-rich layer on the as-cleaved wafers is subsequently removed through a selective etch step (11-26b).

A fully automated tool that separates bonded donor-handle wafer-pairs is available for this purpose. The high-quality wafer can be used to produce many thin-film transfers, reducing the process cost. Another factor that contributes to further cost-reduction arises from the fact that the "handle"-wafer is separated from the device-layer by the buried-oxide-layer. Consequently, a "handle" wafer with considerably relaxed electrical and chemical specifications (and therefore lower cost) can be used in this process.

Post-cleave-processes for layer-thickness-adjustment include: **1**) Si layer thinning using etching processes; and **2**) layer-thickening by growth of additional Si or Si-Ge multilayers. The layer-thickness-control for thickened wafers is better than 0.5% with a within-wafer thickness-uniformity of 2% or better (Fig. 11-27).

### 11.6.3.4 ELTRAN® Wafer Bonded SOI

A relatively new SOI structure (originated and developed by CANON Inc.), which is also based on wafer bonding, is called *ELTRAN.*® The name ELTRAN comes from *Epitaxial LAyer TRANsfer by BESOI* (Bond and Etch-back SOI method) with selective-etching of porous-silicon (Fig. 11-28).[45,46] ELTRAN wafers became commercially available in 1997 for 200-mm substrates.

The porous-silicon surface on an otherwise bare-silicon-wafer is formed by an anodizing treatment of the silicon surface. The ELTRAN method is characterized as BESOI implemented by growing an epitaxial-layer on a layer of porous-silicon. (It has been demonstrated that high-quality epitaxial-layers can be grown on porous-silicon.) The surface of the epitaxial-layer is then thermally oxidized, and this layer will eventually become the buried-oxide-layer of the SOI-structure. This wafer (with a thermally-oxidized epitaxial-film grown on porous-silicon) is then bonded to a "handle" wafer as the "donor" wafer. In early ELTRAN wafer fabrication process, the "donor" wafer material was removed after bonding by a grinding process (down to the porous-silicon-layer). The exposed porous-silicon that remains after the grinding-process is then removed by a wet-chemical etch. A mixture of HF and $H_2O_2$ is used for this etching step, since it provides an etch selectivity as high as 105:1 between the porous-silicon and the epitaxial-layer. At the end of the process, only the epi-layer remains on the top of the oxide layer and handle-wafer. The ELTRAN process has been shown to produce 470-nm-thick SOI-layers with a 14-nm thickness uniformity.

In the more recent ELTRAN process (Fig. 11-29), a double porous-silicon

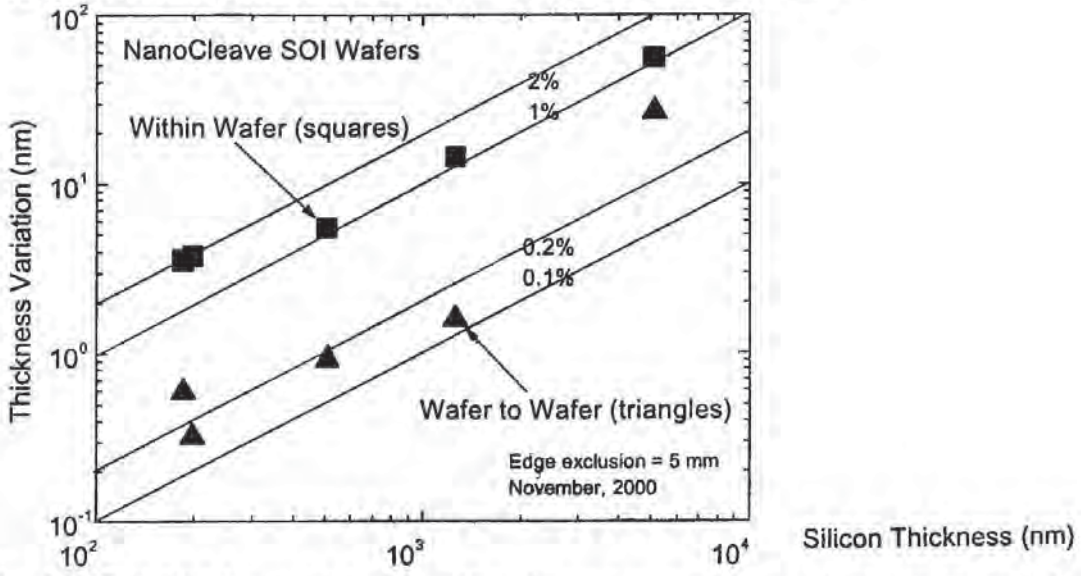



**Fig. 11-27** Si-SOI layer-thickness-uniformity and average layer-thickness-control for post-cleave thickening of as-cleaved SOI-wafers with cleaved 200-nm Si-SOI-layers.[63]

layer is grown on the "donor" wafer, with each of these porous-silicon-layers having a different porosity. The porosity of the porous-silicon-films on the wafer surface can be adjusted by changing the value of the anodization current. After bonding it becomes possible to physically separate the "donor"-wafer from the

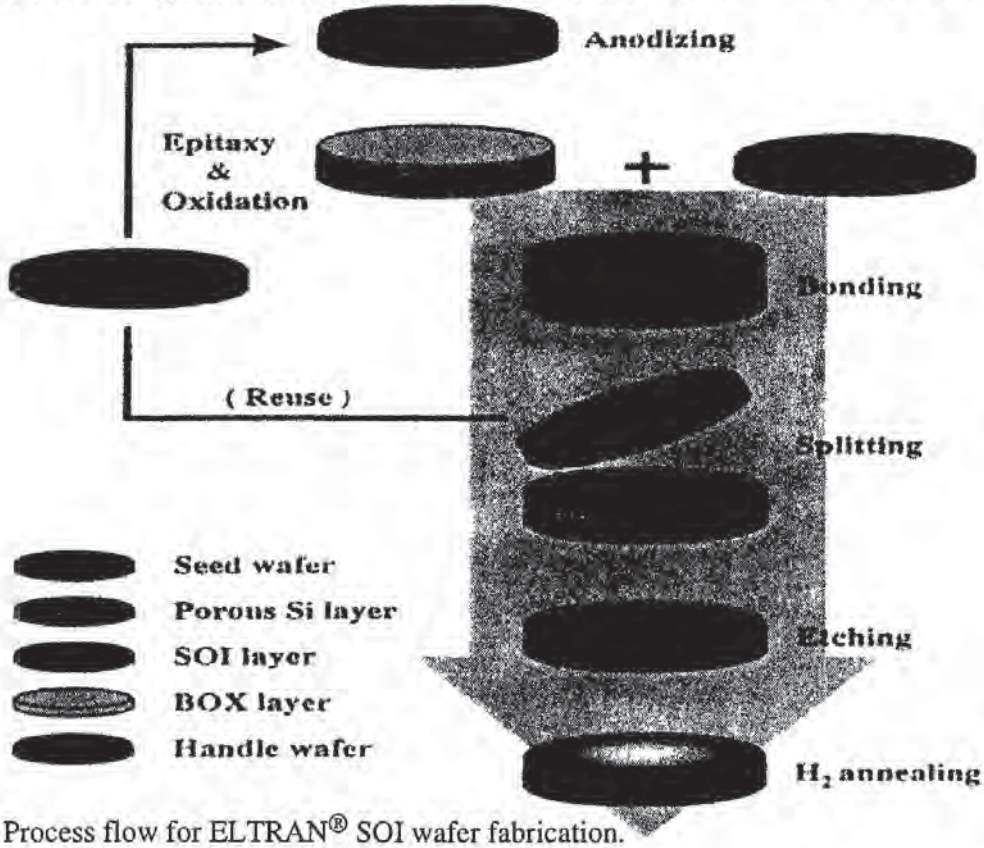



**Fig. 11-28** Process flow for ELTRAN® SOI wafer fabrication.

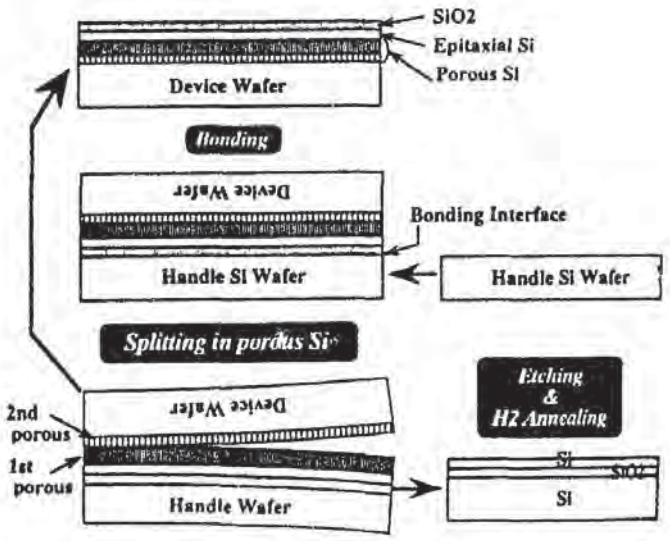



**Fig. 11-29** The ELTRAN process flow for device wafer re-usage. Bonded pairs are split within the porous-Si layer. The device wafer sides are used as a starting device wafer. The handle sides result in ELTRAN wafers after etching the porous-Si and $H_2$-annealing.

"handle" wafer using a physical splitting process (i.e., a water-jet slices the SOI-layer from the seed layer).[62] The splitting occurs along the interface between the two porous-silicon layers. That is, since the porous-layers are a double-stacked structure with different porosities, the strain-energy accumulated at the interfacial area limits splitting to an extremely narrow range (because the interface between the first and second porous layers is mechanically the weakest point in the structure). Therefore, the stress-configuration induced by the double porous-layers is responsible for the splitting mechanism.

For splitting a "water-jet" technology is used. The water jet – a small collimated water stream – is injected at the edge of the bonded-pair and splits the two wafers at the thin porous-silicon layer. The double-layered porous silicon structure contains the stress concentration, with the result that both of the split wafers (SOI and seed wafer) are covered by the porous silicon protective layers. The water-jet drives high-pressure water into narrow gaps by fluid wedging and applies uniform pressure, reducing the chance of damage. The splitting yield is reported to be over 90%, and automatic water-jet splitting-machines are in production use.

The device layer over a buried-oxide is then formed by the etchback of the porous-silicon. After the porous-silicon-film is removed, the surface of the device silicon-layer is flattened by hydrogen-annealing (see Chap. 2, for a discussion of hydrogen-annealing of Si wafers).

There is another economic benefit of the more recent, double-porous-silicon-layer process used to form ELTRAN wafers. In this process (and like in the Smart-Cut and Nanocleave processes discussed in the previous sections), after the splitting process, the "donor"-wafer becomes available to be used as another a "handle"-wafer (instead of being ground away, as in the early ELTRAN process). Canon announced that it has expanded the production of ELTRAN wafers from 2000-per-month to 10,000-per-month in November 2000. This production expansion is being supported by a greater marketing of ELTRAN wafers. To this effect Canon opened a marketing office in the U.S. to achieve this goal.[53]

## 11.7 LATERAL ISOLATION IN SOI

SOI-CMOS-structures eliminate all inter-device leakage-currents by building *n*- and *p*-type islands on an insulating substrate for *p*- and *n*-channel transistors, respectively. Field-inversion and device-to-device-punchthrough do not exist in SOI-technology, but new parasitic transistors in an SOI-structure can still give rise to leakage-currents. SOI-CMOS-technology depends on high-quality isolation. The methods used to laterally isolate SOI structures will be described here, as well as a description of the leakage paths that arise as a result of each of them. The three techniques for providing lateral isolation in SOI are: **1**) MESA isolation; **2**) LOCOS isolation; and **3**) shallow-trench isolation (STI).

### 11.7.1 MESA Isolation

MESA isolation was the early method used to create islands of silicon on an insulating substrate (Fig. 11-30). The technique is attractive because of its simplicity. That is, in MESA isolation the active-device-regions (the silicon islands) are formed into a "mesa" by using a mask-step and a silicon-etch step to remove the silicon in the field-areas down to the insulating-substrate. The SOI oxide helps as the etch-stop-layer, while anisotropic-etching allows for minimization of the distance between adjacent mesas. Passivation of the islands (mesas) is performed at the gate-oxidation step, where the gate-dielectric is grown not only on top of the silicon islands, but on their edges as well.

Although the MESA isolation method is simple, it also has some inherent problems. The first involves the etching of the mesas and the thermal-growth of the gate-oxide that is subsequently grown over the entire mesa. That is, the etching of the mesas produces silicon islands with sharp upper-corners. When the gate oxide is grown on these islands, the oxide at the top corner is thinner than on the flat areas of the mesa (typically 30-50% thinner than on the top surface - see

Vols. 1 and 2 for a discussion of oxide-thinning at corners during thermal-oxidation processes.) When the gate-material is run over the mesa and voltage is applied to turn-on the MOSFETs built in the mesa, the thin-oxide at the corner can cause premature gate-oxide-breakdown. The gate-oxide-breakdown-voltage may also be reduced on the sidewalls, since the crystal orientation may be different from that of the top surface of the silicon wafer. The growth-rate of $SiO_2$ is dependent on the silicon-crystal-orientation of the surface on which it is being grown. Furthermore, crystal-orientations other than (100)-Si generally have larger fixed-oxide-charge-densities and surface-state-charge-densities than (100)-Si, and this can decrease the threshold-voltage of MOSFETs built on such mesas.

The second problem relates to the steep sidewalls produced by an anisotropic-etch of the silicon-mesas. When a polysilicon-layer (used to form the MOSFET gate) is deposited on the mesa (typically to a thickness of 400-nm), an anisotropic-etching of the poly will leave stringers along the mesas, and this may short the gates of devices within the same mesa. Overetching is therefore needed to remove such stringers, but this can lead to poor dimensional-control of the poly gate-lengths.

The third problem involves early inversion along the top of the mesa sidewalls, which behave like lateral (edge) transistors in parallel with the top (main) device. This is due to charge-sharing of the depletion-regions caused by the poly running over the top and sides of the mesa-sidewalls. Thus, the corners show a lower threshold-voltage than the main channel of the MOSFET. This manifests itself as a kink in the subthreshold-characteristic of the drain-current. Such a deterioration of the subthreshold behavior is most pronounced in NMOSFETs in SOI-mesas, and can be alleviated by performing a boron-implant into the field-regions of the silicon wafer before etching the mesas. This boron-implant is then driven-in laterally into the active-area. The mesas are then etched, leaving the sidewalls more highly doped than the active-channel. Figure 11-31 shows a perspective drawing of a laterally-isolated MESA SOI-NMOSFET with a T-gate and a body-substrate contact.

### 11.7.2 LOCOS-Isolation for SOI

LOCOS-isolation is the second technique that has been used to produce lateral-isolation among active-silicon-islands in SOI. LOCOS is more complicated than MESA isolation, but it has the advantage of already being in wide use in bulk-CMOS technology. LOCOS-isolation in SOI is much the same as LOCOS in bulk-CMOS (see Vols. 1, 2, or 3 of *Silicon Processing for the VLSI Era* for more details on the LOCOS process.) However, there are some issues that do arise

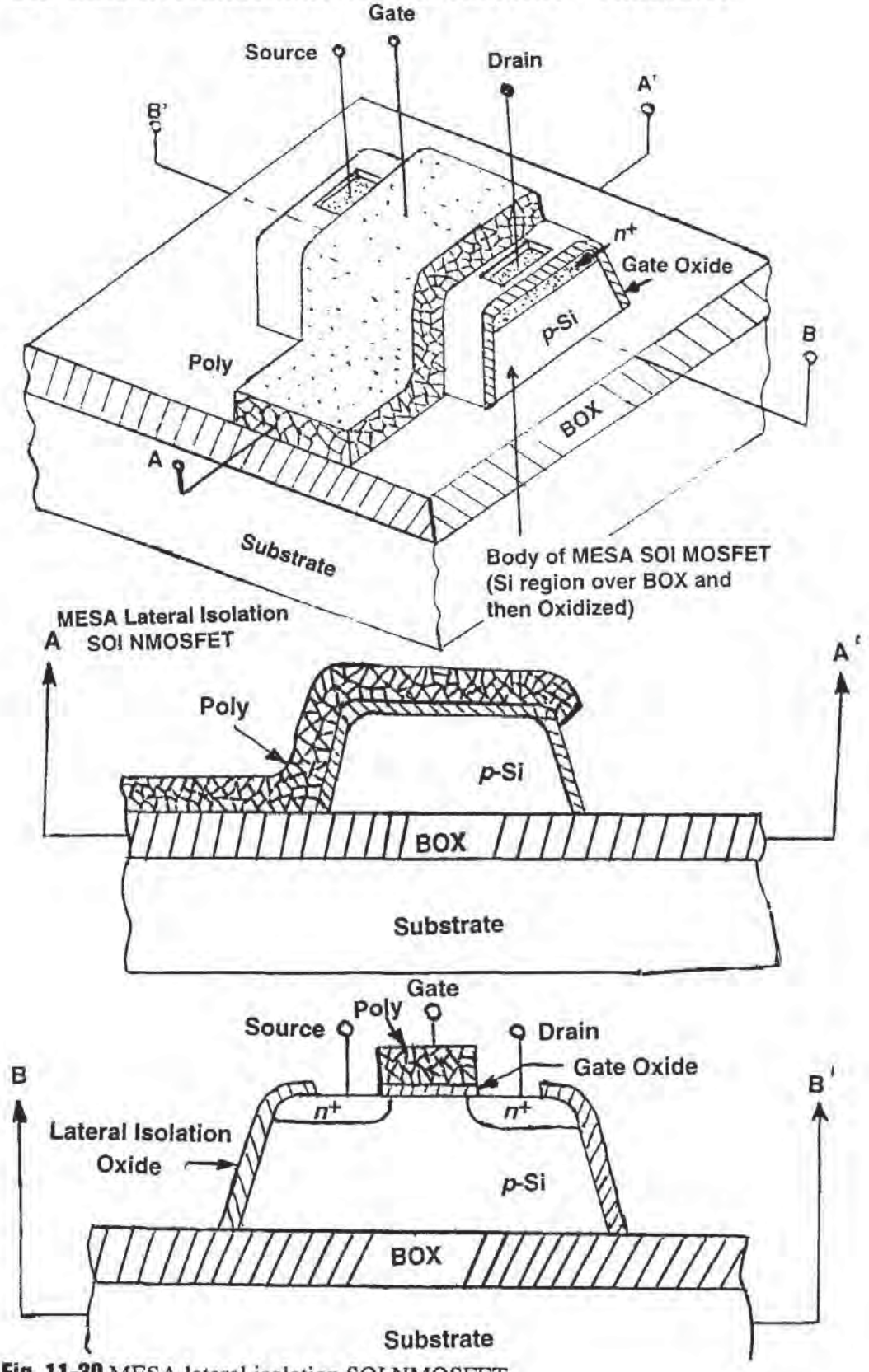
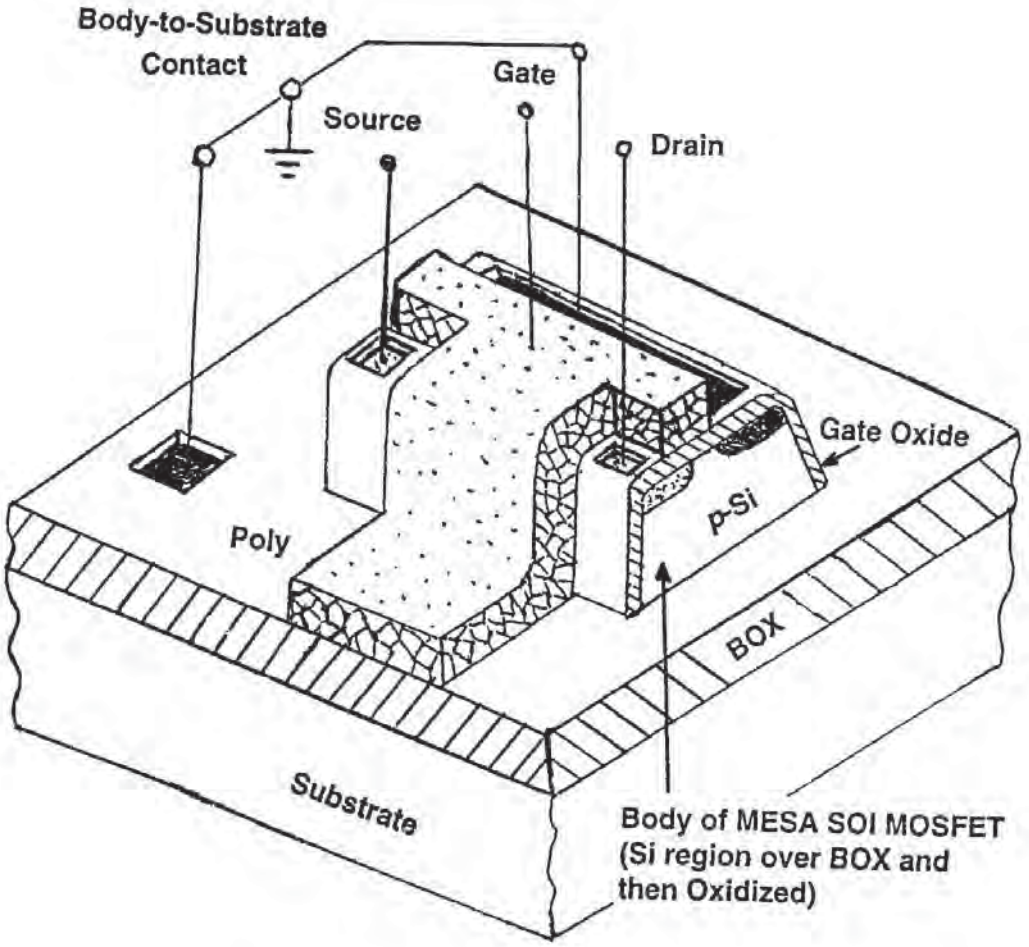



**Fig. 11-30** MESA lateral isolation SOI NMOSFET.




**11-31** Mesa lateral-isolation SOI NMOSFET with a T-gate and body-to-substrate contact.

when LOCOS is applied as the technique for producing lateral-isolation in SOI. These issues are discussed in this section.

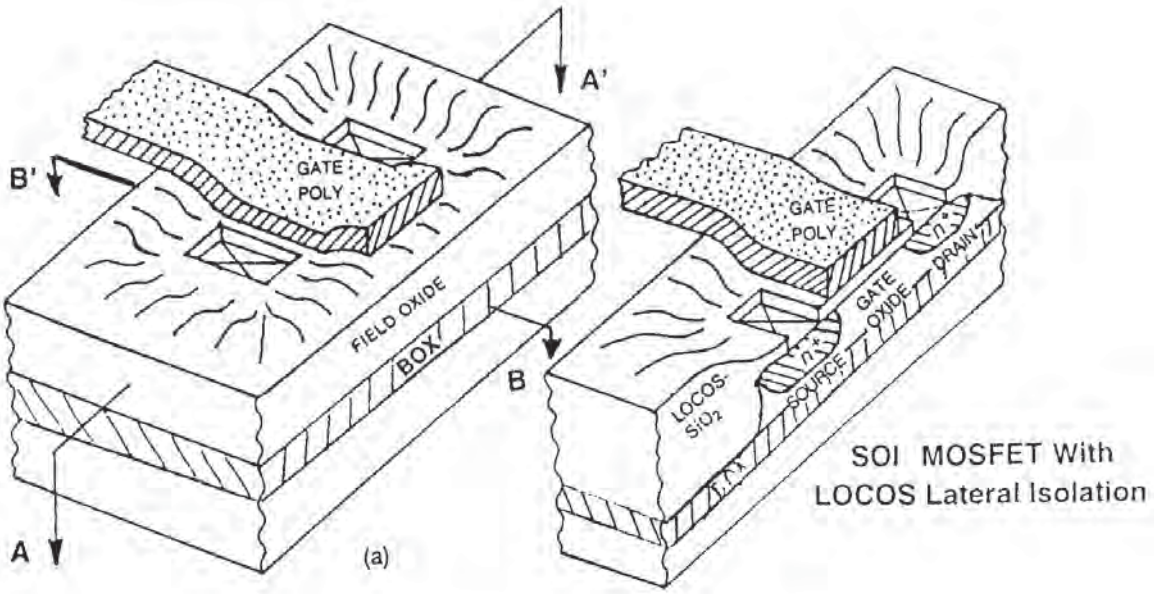
The LOCOS lateral-isolation-technique in SOI consists of the following steps (for producing NMOSFETs): a thin pad-oxide is grown, and a silicon-nitride-layer is then deposited by LPCVD onto the pad-oxide. Using a masking step, the nitride is patterned to define the active-silicon-areas. Boron is then implanted around the islands of *p*-type silicon that will contain the NMOSFETs (i.e., a channel-stop-implant). Next, the field-oxide is thermally grown. No oxide growth occurs over the active-areas which are protected by the silicon-nitride. A nitride and pad-oxide stripping step completes the process. In SOI the field-oxide-growth is carried out so that the thermally-grown field-oxide completely

consumes the silicon in the regions not covered by the nitride, with the thermally-grown field-oxide reaching through to the buried oxide (see Fig. 11-32).

The lateral-encroachment of the field-oxide (called the "bird's beak") is proportional to the thickness of the grown field-oxide. This encroachment has typical values of a fraction of a micron, and this hampers the use of the LOCOS technique for submicron bulk-CMOS processes (where field-oxide-thicknesses of 0.3-05 $\mu$m are grown). The problem of "bird's-beak"-encroachment in SOI-CMOS is even worse than in bulk-CMOS processes. This is because the oxidation-kinetics in SOI are somewhat different, particularly when the growing field-oxide reaches the buried-oxide. That is, consumption of silicon during oxide-growth on a bulk-silicon-wafer is 0.45-$t_{ox}$, where $t_{ox}$ is the oxide-thickness. That means that a thermal-field-oxidation process that produces a silicon-dioxide-layer about twice as thick as the consumed silicon-layer must be used in a bulk-CMOS. For example, a silicon-layer of about 125-nm is consumed when a 250-nm $SiO_2$ field-oxide-layer is grown.

By comparison, a 125-nm silicon-layer above the buried-oxide of an SOI-wafer is not fully oxidized if only such a 250-nm silicon-oxide growth-process is used. That is, unoxidized-filaments of silicon remain in $SiO_2$-layers grown by the 250-nm $SiO_2$-process on an SOI-wafer with a 125-nm silicon-layer above the buried-oxide. To ensure that such silicon-filaments no longer remain in the silicon-dioxide-layer, an oxidation-process equivalent to one that would grow a 350-nm-thick $SiO_2$-field-oxide on a bulk-wafer must be used. In other words, a process that grows an oxide 3-times-as-thick as the silicon-layer consumed has to be used in SOI to completely laterally-isolate the active-regions with a LOCOS process. The longer field-oxidation-process produces even more "bird's-beak" encroachment (and more lateral-boron-diffusion from the channel-stop-implant into the active-region) than a LOCOS-process in a bulk-CMOS process.

Another problem of the LOCOS method of lateral-isolation is that source-to-drain leakage-current in the subthreshold-region of operation can also be increased by the parasitic-edge-transistor that exists along the sidewalls of the silicon-active-regions in parallel with the main-channel of the SOI-MOSFETs. The problem is most severe in LOCOS-isolated SOI-NMOSFETs for the following reason: A $p^+$ channel-stop-implant is carried out before the field-oxidation to prevent such edge-transistors from being turned-on prematurely and causing excess subthreshold-leakage-current. However, during the field-oxidation step, such implanted-boron segregates into the growing-oxide on the top of the silicon-surface - and into the buried-oxide as well. As a result, the boron doping at the tip of the bottom corners of the active-device islands ends up being



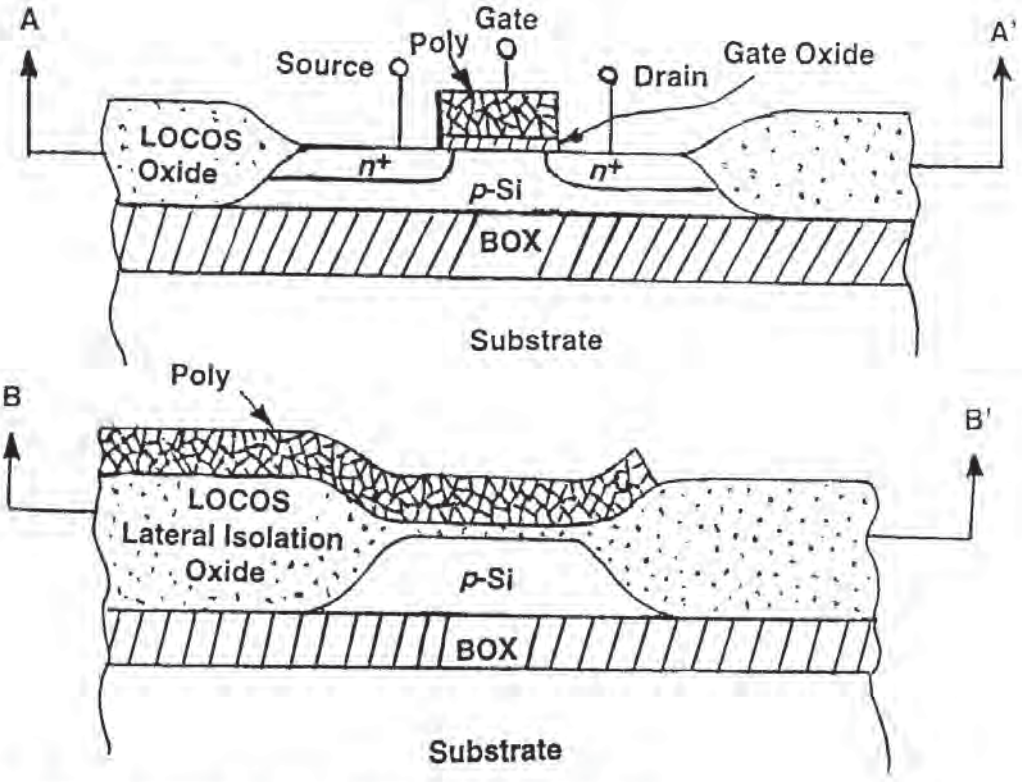




**Fig. 11-32** SOI-MOSFET that uses LOCOS lateral isolation.

quite low, allowing edge-transistor-leakage to arise. One method suggested to alleviate this problem is to perform a second boron channel-stop-implant after the LOCOS field-oxide-step has been completed. Since the implant step is carried out after the field-oxide step, such B will be implanted into both active-regions and field-regions of the wafer. The energy of the implant can be adjusted to place the peak of the B-implant-dose close to the meeting point of the grown- and buried-oxides. Thus, this implant step will not impact the doping-concentration of the main MOSFET channels near the surface of the active-regions.

### 11.7.3 Shallow-Trench Isolation (STI) for SOI

Because of the problems associated with the MESA and LOCOS isolation techniques for providing lateral isolation in SOI technology, the more advanced SOI technologies have adopted shallow-trench-isolation as the lateral-isolation method of preference (Figs. 11-1 and 11-33). That is, the active-regions of the silicon are formed as in the MESA isolation method, but in STI the regions between the mesas are filled with a CVD-dielectric-material and then the surface is planarized by CMP to yield a flat wafer surface. Oxide remains in the trenches between the mesas, and its top surface lies in the same plane as that of the silicon-islands. STI eliminates the problems created by the sidewalls of the mesas, and the issues associated with the LOCOS-isolation method for SOI. (That is, the problem of "bird's-beak" encroachment does not occur when STI is used. Also, anomalies in the oxidation kinetics associated with field-oxide growth in SOI do not have to be addressed since the field-regions between the active devices are filled using CVD rather than a thermal-oxidation process). However, the same general issues involved with shallow-trench-isolation methods in bulk-CMOS still have to be addressed when applying an STI approach to SOI. Chapter 9 of this text provides a detailed discussion of STI technology.

## 11.8 SOI PRODUCT EVOLUTION

Table 11-4 presents a list of the SOI circuits that have been introduced during the International Electron Devices Meeting (IEDM) from 1989 to 1998.[64] It provides an indicator of SOI R&D efforts. The initial publication of the gain in the speed of a relatively small IC fabricated in SOI has been followed by a series of studies of performance-enhancements of larger ICs, mostly using SIMOX wafers. In general, these ICs were fabricated using partially-depleted SOI-MOSFET technologies, and silicon- and BOX-thicknesses were all comparable (and fairly thick). In 1998, IBM announced it would apply its leading-edge SOI-CMOS technology (including copper interconnects), to commercialize SOI-CMOS prod-

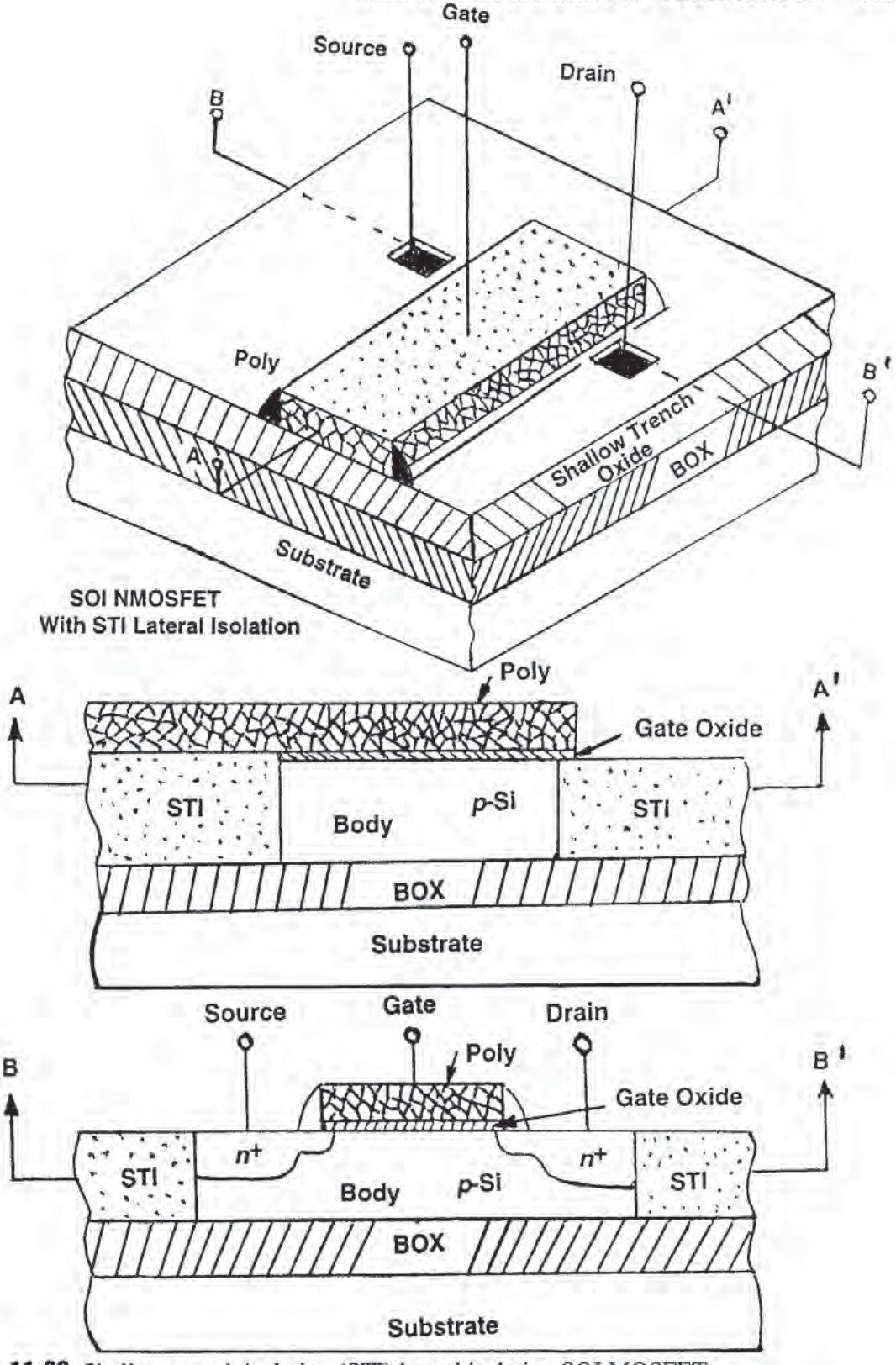



**Fig. 11-33** Shallow-trench isolation (STI) lateral isolation SOI MOSFET.

## Table 11-4 Summary of SOI Circuits Presented Annually at IEDM[64]

| | Bell Labs | IBM | Motorola | Mitsubishi | Hyundai | DEC | IBM | Intel |
|---|---|---|---|---|---|---|---|---|
| | 1989 | 1993 | 1995 | 1996 | 1997 | 1997 | 1997 | 1997 |
| Circuit | Prescaler | 0.5Mb SRAM | CPU | 16M DRAM | 1G DRAM | 1M SRAM | 4 Mb SRAM | 4 Mb SRAM |
| Technology (μm) | 0.50 | 0.50 | 0.50 | 0.50 | 0.17 | 0.35 | 0.25 | Sub-0.25 |
| $V_{DD}$ | | 1.0 | | 1.0 | | 2.0 | 1.8 | 1.8 |
| Silicon thickness | 134 | | 100 | 100 | 200 | 200 | 200? | 150 |
| BOX (nm) | 235 | | 400 | 400 | 250 | 200 | 400 | 400 |
| Lateral isolation | LOCOS | | | MESA | LOCOS | STI | STI | STI |
| Silicon layer characteristics | PD/FD | Non-depleted | PD | | Body-cont., thin Si film (75nm) beneath FOX | PD | Non-depleted SOI | PD |
| Metal levels | 1 | | 2 | | | 4 | 6 | 5 |
| Performance improvement | 50% | | | | | | 25–35% | |
| Power reduction | | | | | | 50% | 170–300% | 18% |
| Yield (if known) | Low | | Same as bulk | | | Approaching bulk | | Low |
| Conclusions and claims | | | Bulk CMOS design, performance 2 × bulk | Both body-tied and floating body MOSFETs | Not clear if circuit is functional! | 20% performance improvement | Highest-performance 0.25μm CMOS technology reported FB effects negligible | No speed advantage over bulk for sub-0.25μm logic applications |

ucts based on SIMOX. Figure 11-16 shows a SEM of a 0.18-$\mu$m IBM IC structure built using SOI. The company believes SOI processors are 30% faster than bulk-CMOS versions, and that they provide comparable circuit yields and add less than 10% to wafer-processing costs. Other companies have also announced the transition to SOI for high-speed microprocessors or low-power logic circuits. This is a positive development that supports IBM's claim that CMOS could move more towards SOI in the next few years (see Fig. 11-5).

In 1999, the presentation of new SOI devices occurred primarily at the *International Solid-State Circuits Conference* (instead of at the IEDM). Announcements were made concerning the successful fabrication of a number of SOI circuits in this ISSCC 1999 Conference. These circuits included 32- and 64-bit microprocessors, and 16-Mbit DRAMS on SOI. Depending on the technology and circuit layout performance, 10-40% enhancement was reported. A 0.18-$\mu$m SOI-CMOS technology for high-performance logic applications, however, was described by IBM in the 1999 IEDM. This SOI technology used a low-dose SIMOX substrate, shallow-trench-isolation, dual gate-oxide-thickness, dual threshold-voltage ($V_T$), dual-doped-polysilicon, cobalt-silicide, low-$k$ dielectric

in their interconnect scheme, and seven layers of copper metallization. The SOI circuit devices were partially-depleted MOSFETs, and the test chip built using this technology contained a microprocessor-circuit that demonstrated performance above 1-GHz.[58,59] These reports indicate a new-degree of maturity in SOI technology.

The 0.18-$\mu$m SOI process of IBM was qualified by that firm for production in the summer of 2000, and work was continuing to qualify a 0.13-$\mu$m SOI process in 2001. The 0.13-$\mu$m SOI technology would allow production of high-performance devices with 44-million transistors. IBM also said that it had begun shipping its AS400 servers with 0.22-$\mu$m SOI devices early in 2000, and that their newer "Turbo" Unix machines would employ 0.18-$\mu$m SOI-processors running at well over 800-MHz. The 0.18-$\mu$m designs utilize shallow-trench-isolation, dual-work-function-poly, local-interconnect, and borderless tungsten-interlayer interconnect.

At the end of 2001, IBM reported on their 2$^{nd}$ generation 0.13-$\mu$m SOI process. The power supply voltage was 1.2 V, and a thinner SOI-layer and a gate-dielectric material with a higher concentration of nitrogen were used. The interconnect structure can have as many as 8 levels of metal. A completely low-$k$ dielectric stack (i.e., with SiLK for the main dielectric material and SiC as the hardmask material and the post-CMP cap-dielectric). The Cu wires are formed using a dual-damascene process.

## 11.9 PROBLEMS ASSOCIATED WITH SOI

### 11.9.1 Floating Body Effects in Partially-Depleted SOI MOSFETs

In partially-depleted SOI MOSFETs, the depletion-region under the gate when the MOSFET is operated in inversion does not extend all the way to the buried-oxide interface. Instead, a neutral-region still exists. When the body is grounded (say, with an independent body-contact), a partially-depleted SOI-MOSFET behaves very much like a bulk-MOSFET, and most standard $I_D(V_{GS},V_{DS})$ equations and design concepts still apply. However, if a body-contact (connected by a conducting-line to ground) is not made to the SOI-MOSFET, so called *floating-body-effects* arise, some of which have a detrimental impact on the SOI MOSFET behavior. These are discussed in more detail in the following sections.

**11.9.1.1 Kink Effect in Partially Depleted SOI NMOSFETs:** The *kink effect* manifests itself as a "kink" in the output characteristics of partially-depleted SOI MOSFETs (Fig. 11-34). The kink appears in the graph of the drain current of

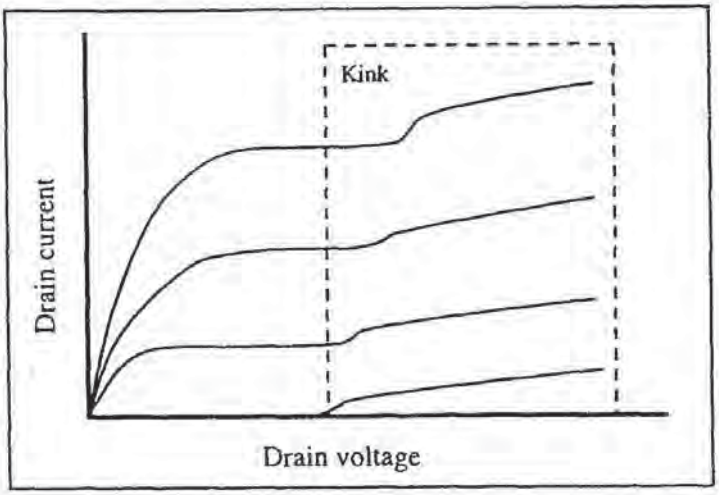



**Fig. 11-34** Illustration of the kink-effect in the output characteristics of a partially-depleted SOI-MOSFET.

such devices above some value of the drain voltage. It is often quite pronounced in SOI-NMOSFETs, but is usually absent in SOI-PMOSFETs. The kink-effect is also not observed in any bulk-MOSFETs operated at room temperature if a substrate or well contact exists. It may be observed in bulk-NMOSFETs if they are operated a low temperatures or in NMOSFETs fabricated in a floating *p*-well.

The kink-effect occurs for the following reason. When the drain voltage is large enough in an NMOSFET, the electrons near the drain are accelerated across the channel-drain depletion-region by an electric-field that is strong enough to cause electron-hole pair generation by impact-ionization (in PMOSFETs the coefficient of electron-hole pair generation by energetic-holes is much lower than that of energetic-electrons - which are the carriers in NMOSFETs). Drain voltages typically get this high when the MOSFET is operating in the saturation region (Fig. 11-34).

In a bulk-NMOSFET, electrons generated by this effect rapidly move into the drain, adding to the drain-current. The generated holes exit the device through the body-contact where they become substrate-current. In partially-depleted SOI-NMOSFETs, the generated electrons still move out of the drain and add to the drain-current. The generated holes, however, collect in the NMOSFET body, where they raise the potential of the body, with respect to the source (which is grounded). The amount by which the potential of the body is raised can be calculated, but this requires iterative solving of a set of complex equations, and

will not be described here. However, when the body-potential rises about 700-800 mV, it starts to significantly impact the output-characteristics of the NMOSFET.

That is, an increase in the body-potential has the same effect as biasing the substrate (body-effect). In other words, when the body of an NMOSFET is positively biased, it will give rise to a decrease in the threshold-voltage, $V_T$, of the NMOSFET (this is termed the *body-effect*). The decrease in $V_T$ causes an increase in the drain-current, and this increase shows up as a "kink" in the drain-current curve of the SOI-NMOSFET. The kink represents an annoyance when analog circuits are designed with SOI NMOSFETs, but can actually lead to the fabrication of faster digital-circuits.

The situation regarding floating-body-effects in fully-depleted SOI-MOSFETs is somewhat different than in partially-depleted devices for several reasons. First, in the fully-depleted SOI-MOSFET, the body of the MOSFET at threshold has the same potential as the surface of the silicon at the top $Si/SiO_2$ interface. Hence, any holes generated by impact ionization at the drain will migrate out of the source and not collect in the body (which, in a partially-depleted SOI MOSFET, will have a lower potential than the source, and will thus attract the generated holes). Hence, there is no collection of holes in the body region, and the body-effect is also insignificant. In other words, the generation of holes by impact-ionization in fully-depleted SOI-MOSFETs does not decrease $V_T$, and hence no kink-effect is observed in fully-depleted SOI-MOSFETs, even in the absence of a body contact (i.e., body contacts are thus no longer needed). Second, since the potential of the body-region is higher, the electric-field at the drain will be smaller, reducing the number of electron-hole-pairs generated by impact-ionization.

There is, however, some effect on the subthreshold-slope of both partially-depleted and fully-depleted SOI-MOSFETs that is produced by the electron-hole pairs generated by impact-ionization at the drain. That is, as the MOSFET is being turned-on (but is still in the subthreshold-regime), the generated-holes will collect in the body (which has a lower potential than the source in both partially- and fully-depleted SOI-NMOSFETs if it is being operated in the subthreshold-regime). These injected-holes charge the body and produce a forward-bias on the body-source junction that reduces the threshold-voltage. As a result, the subthreshold-current increases, and the subthreshold-slope of the $I_D(V_{GS})$ curve is not as steep (i.e., the turn-off behavior of the device is degraded).

**11.9.1.2 Transient $I_D$ Effects:** The floating-body may also produce transient-effects in the drain-current of a partially-depleted SOI-MOSFET. For example, an *overshoot* in the drain-current occurs when the MOSFET is turned-on (e.g., by increasing the gate-voltage from 0 to $V_{DD}$). What is happening is that majority-carriers are being expelled from the volume of the silicon of the floating-body under the gate that is becoming a depletion-region, and these carriers collect in the remaining neutral-region of the floating-body. This causes the potential of the floating-body to rise, and (as is described in the section on the kink-effect), this causes an increase in the drain- current. However, these excess-majority-carriers recombine and eventually equilibrium is re-established. At that point, the potential of the floating-body is returned to the value it would have without the presence of the excess injected- majority-carriers. Thus, the drain-current gradually subsides with time to a lower value. A reciprocal transient *undershoot* in the drain-current is observed when the gate voltage is switched off (e.g., if the gate-voltage decreases from $V_{DD}$ to 0).

Such short-duration transient spikes in the drain-current can impact such SOI-circuits as SRAMs and DRAMs, which use pass-gate configurations. The amount of charge that is passed by the transient floating-body-effects must be accounted for in the design of such circuits.

**11.9.1.3 Effect of Building Partially Depleted SOI-MOSFETs Without Body-Contacts:** As discussed in the previous sections, the floating-body of partially-depleted SOI- MOSFETs has an impact on various MOSFET performance characteristics, most of which appear to be undesirable. However, Shahidi made a radical proposal in 1994. He suggested that in many important commercial circuit applications, partially-depleted SOI-MOSFETs could be used without contacts to the floating-body, and that this could be accomplished by making some modifications to the circuit design. Furthermore, by leaving the body floating, not only would the density be increased, but circuit-speed would be improved as well.

Early applications of SOI were limited to space and military programs where radiation-hardness was more important than performance and cost. These products were built with partially-depleted SOI, and fixed potentials to the body were provided by well- and substrate-contacts, just as is done in bulk-CMOS technology. In commercial applications, the considerations of performance and cost become paramount, and radiation-hardness is almost never an issue. Thus, Shahidi suggested that the floating-body-effects would not significantly degrade the performance of many SOI-ICs tailored for specific applications. If that was the case, the floating-body-contacts could be left out. In fact, several bulk-SRAM

circuits and microprocessors were implemented in a partially-depleted SOI-technology that did not employ floating-body-contacts, and these SOI-ICs did indeed exhibit high-performance. In addition, they met all other commercial product specifications (such as stress, reliability, burn-in, and yield). Although some circuit-modifications to the bulk-CMOS circuit design had to be made when implementing them in SOI, these modifications resulted in a higher performance gain than would have been obtained by merely re-mapping the bulk-CMOS design into SOI.[51] The floating-body-effects on the change in drain-current were compensated for in the circuit designs by a number of solutions, and these allowed the SOI-ICs to be successfully designed and fabricated using partially-depleted SOI-MOSFETs without contacts to the floating-body.[52]

Another floating-body issue arises when designing 0.13-$\mu$m SOI-CMOS circuits (and beyond), as a result of the very thin gate-oxide-thicknesses needed for these generations of devices. That is, such thin-gate-oxides give rise to a significant amount of gate-to-body tunneling-current. The charge introduced into the floating-body modulates the body-voltage and therefore causes a so-called *history-effect*, which can become excessive on 0.1-$\mu$m SOI-CMOS circuits.[60]

This history-effect can be explained as follows: The gate-current in these circuits can change the device "ON" $V_{bs}$, but does not usually change the "OFF" $V_{bs}$. Therefore, the gate current will not impact the delay when the input signal switches from "0" to "1" for the first time, but it will impact the delay when the input signal then falls from "1" to "0" ($1^{st}$ switch $V_{bs}$) - and the impact of this delay will persist when the input signal switches the second time from "0" to "1" ($2^{nd}$ switch $V_{bs}$ – Fig. 11-35). This history-effect must be minimized in 0.1-$\mu$m SOI-CMOS by appropriate device-design.

Minimizing the history-effect can accomplished by making two modifications to the SOI-MOSFET device-design: **1**) the silicon-dioxide gate-dielectric is replaced with a dielectric having a higher *k*-value (which reduces the gate-current for an equivalent-oxide-thickness); and **2**) the value of gate-side junction capacitance $C_j$ is adjusted.[60] (That is, $C_j$ controls the drain-capacitance coupling to the body, and hence determines the $2^{nd}$ switch $V_{bs}$.) By adjusting $C_j$ at a given level of gate leakage-current, the $1^{st}$ and $2^{nd}$ switch delays can be balanced (thereby eliminating the history effect in the switching delay). From a performance point of view, a low-value of $C_j$ is beneficial. $C_j$ can be reduced effectively by scaling the SOI silicon-layer thickness.

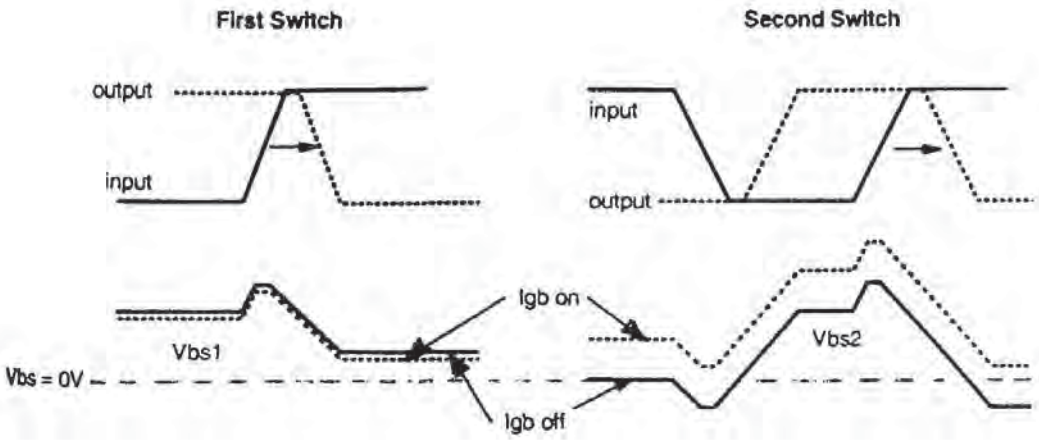



**Fig. 11-35** Definition of 1st and 2nd switch rise-to-fall delay for the NMOSFET in an SOI-CMOS inverter. (© IEEE 2000).[60]

## 11.9.2 Short-Channel-Effects in SOI-MOSFETs

SOI-MOSFETs offer good short-channel behavior, especially for devices with channel lengths smaller than 0.1-$\mu$m. Fully-depleted SOI-devices are being studied for low-power/low-voltage applications because they appear to exhibit the best short-channel behavior. However, the silicon-thickness and the buried-oxide-thickness have to be made ultrathin in single-gate fully-depleted SOI-MOSFETs to achieve the best short-channel behavior (i.e., the silicon-thickness must be considerably smaller than the junction-depth of a comparable bulk technology). Even in such ultra-thin SOI-MOSFETs, back-channel-potential control-issues can exacerbate such short-channel-effects as threshold-voltage-rolloff and this issue must be addressed.[54,55]

The problem of threshold-voltage-rolloff arises in deep-submicron single-gated fully-depleted SOI-devices because the depletion-region-charge of the source/drain-regions cannot terminate in the buried-oxide region, and thus must terminate in the substrate below the buried-oxide or in the channel-region (Fig. 11-36).[55] Because the buried-oxide is thick, a portion of the depletion-charge will terminate in the channel. Thus, the source and drain behave like fringing back-gates. If the channel is short, the fringing-fields from these parasitic back-gates can induce back-channel-conduction, which enhances the problem of threshold-voltage-rolloff with decreasing channel-length. This phenomenon has been

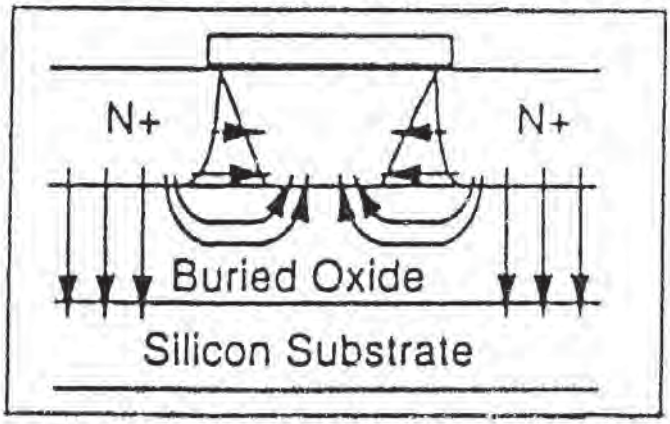



**Fig. 11-36** Electric-field lines in the source/drain and body *pn*-junctions in a short-channel SOI-MOSFET. Some of these lines terminate in the substrate beneath the buried-oxide (BOX), while others terminate in the body, producing fringing-electric-fields.[55]

studied by simulation and empirically modeled.[55] Reducing the buried-oxide thickness can also help alleviate the 2-D charge-sharing through the buried-oxide, because the bottom substrate is brought closer to the source and drain regions, provided there is no significant depletion in the substrate. Simulation studies have been presented that predict that single-gate SOI-MOSFETs, even with ultra-thin silicon channels, do not provide adequate immunity against short-channel-effects in 25-nm channel-length devices.[57]

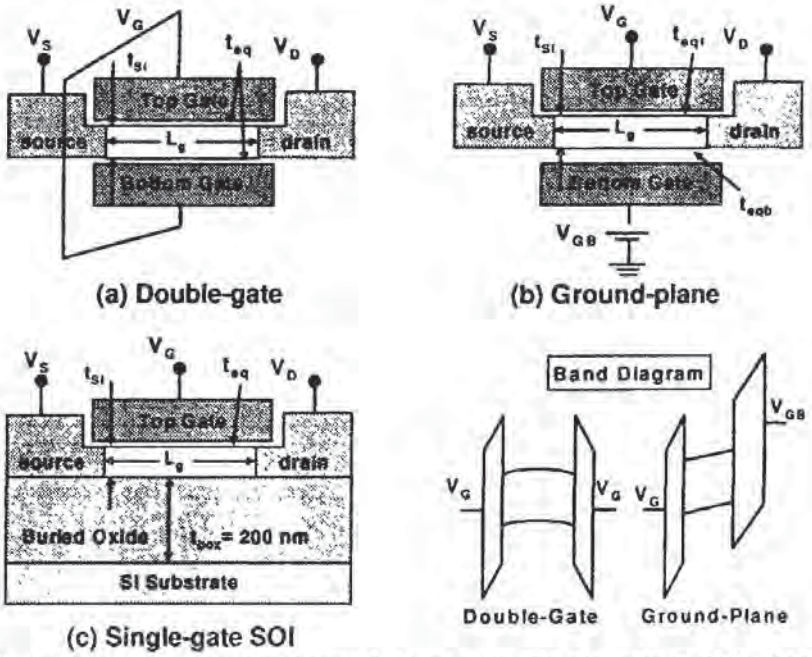



**Fig. 11-37** a) Double-gate; b) Ground-plane; c) Single-gated SOI-MOSFET. Gate material work-function is set at the mid-gap of the silicon band gap. On-chip biasing of the ground plane is assumed. The source/drain junction is abrupt. The channel length ($L_g$) is the metallurgical channel length. (© 1998 IEEE).[57]

As a result, two alternative approaches have been proposed to reduce the threshold-voltage-rolloff problem described above. That is, it has been suggested that either a *double-gate SOI-device*, or a *ground-plane SOI-device*, be used instead of a single-gate SOI-device (Fig. 11-37). Simulation studies show that double-gate SOI-MOSFETs give the best control over the threshold voltage rolloff problem, but the difficulties in fabricating double-gate SOI MOSFETs are daunting. Instead, a compromise structure (namely an SOI device with a low-resistance ground plane) has been proposed.[56,57] In the ground-plane SOI MOSFET, the low resistance ground plane can be biased with a power supply separate from that used to control the gate. This offers the flexibility of tuning the threshold voltage by means of the bottom gate bias.[57] Both double-gate and ground-plane SOI-MOSFETs are shown to able to meet the electrostatic design requirements.

### 11.9.3 Self-Heating Effects in SOI MOSFETs

SOI MOSFETs are thermally insulated from the substrate by the poorly heat-conducting buried-oxide layer. Consequently, removal of excess heat generated by Joule heating within the transistors is less efficient than in bulk-CMOS technology. As a result, the device-temperature in SOI technologies can rise substantially above what it would be in a comparable bulk technology. Although the heat transfer rate is lower in an SOI device, heat is nevertheless still transported out through a number of paths, namely: **1**) through the buried oxide layer; and **2**) laterally through the silicon-island into the contacts and metallization. As we shall see below, however, such heating effects have not proved to be a significant problem in SOI CMOS technology.

The increased operating temperature within an SOI-MOSFET can have an impact on the device-performance. The most significant effect is that the mobility of the carriers in the channel is reduced. That is, if a MOSFET is being operated in the saturation-region, and the drain voltage is increased, there will be a larger voltage drop at the drain-junction. This will increase the power dissipated in the device and its temperature will rise. Once the device temperature is increased by this effect by 50°C (to 150°C), a significant reduction in carrier-mobility will occur. This effect will manifest itself in the output characteristics of the device in the form of a drop in the drain-current with increasing drain voltage (Fig. 11-38).

However, in CMOS circuits, all of the MOSFETs normally operate in a standby-mode. Large drain-currents in any device occur only during switching, and these cease after a very brief period of time (< 1-ns). Since the thermal time constants of SOI MOSFETs are on the order of several tens of nanoseconds, it

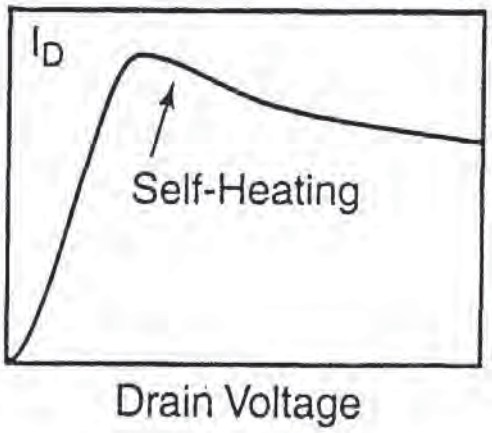



**Fig. 11-38** The effect of self-heating in SOI-MOSFETs on output characteristics of the device.

seems that this self- heating effect generally will cause no problems in CMOS digital circuits, with a few exceptions. That is, in logic-circuits that exhibit short transition times, temperature rises of only 2-3 degrees are observed compared to bulk-CMOS. However, in such circuits as off-chip drivers (that can be ON for a significant portion of the switching cycle), the temperature rise is observed to be higher (e.g., ~10 degrees).

### 11.9.4 Reliability Issues for SOI

There are also some reliability issues unique to SOI-technology that should be addressed when considering SOI for ULSI applications.[61] Such an analysis indicates that SOI-circuits do not exhibit significantly worse reliability behavior than comparable bulk-CMOS-circuits. Some of the reliability issues associated with silicon-on-insulator ICs are reviewed here.

First, bulk- and SOI-CMOS circuits appear to react comparably with respect to *off-state current degradation* when subjected to bias-temperature stress-testing. Hence, this apparently does not represent a special reliability-problem for SOI. Second, channel hot-electron-degradation (CHE) in SOI-devices is more complex than in bulk-MOSFETs because charge-trapping and interface-state-generation can occur at both the gate-oxide-interface and at the buried-oxide-interface. Third, the defects responsible for gate-oxide yield and reliability have been minimized by 2000 in state–of-the-art SOI-materials to acceptable levels. This claim is supported by TDDB tests on gate-oxides on both bulk- and SOI-materials (i.e., the oxides grown on both the bulk- and SOI- materials show

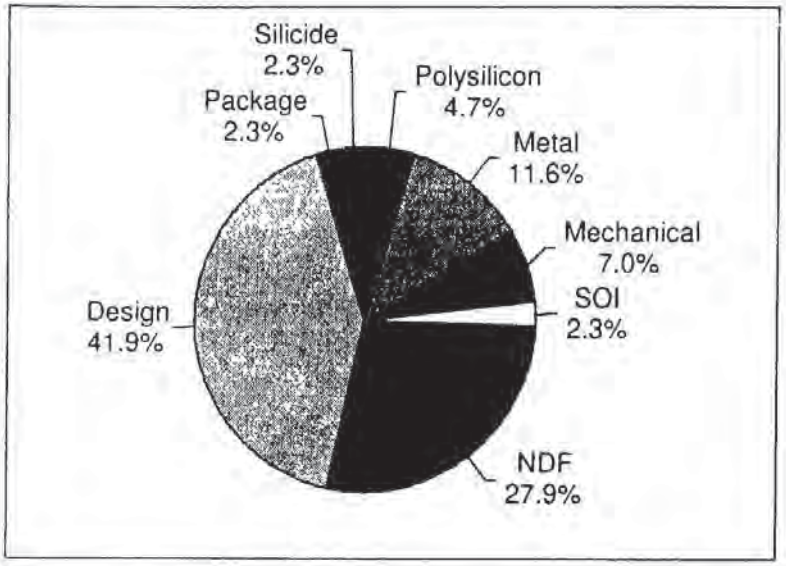



**Fig. 11-39** SOI unique defect fails (induced by burn-in stress), are negligible part of the overall failure population.[61] (© 2000 IEEE)

similar failure distributions). Fourth, SOI-devices appear to be less susceptible to charging damage than bulk-CMOS devices. Finally, when failure mechanisms in SOI-microprocessor-circuits are catalogued, the percentage of fails due to SOI-unique defects represent a negligible portion of the failure distribution (Fig. 11-39). In summary, these reports indicate that by using state-of-the-art SOI material, highly reliable products can be produced.

## REFERENCES

1. International Technology Roadmap for Semiconductors, 1999 Edition, Semiconductor Industry Association, San Jose, CA.

2. "Proc. 5th Inter. Workshop on Measurement, Characterization, and Modeling of Ultrashallow Doping Profiles in Semiconductors," *J. Vac. Sci. Technol.*, **B-18**, 2000, p. 337-604.

3. G. Shahidi, *et al.*, "Mainstreaming of the SOI Technology," *Proc. IEEE Inter. SOI Conf.*, p. 1-4, 1999.

4. A. Agarwal, "Ultra-Shallow Junction Formation by Spike Annealing in a Lamp-Based or Hot-Walled Rapid Thermal Annealing System: Effect of Ramp-Up Rate," *Materials Sci. in Semiconductor Processing*, 1, p. 237, 1998.

5. K.K. Ng and J. Lynch, "The Impact of Intrinsic Series Resistance on MOSFET Scaling," *IEEE Trans. Electron Dev.*, **ED-34**, p. 503, 1987.

**6.** M.I. Current *et. al.,* "Process Integration Issues for Doping of Ultrashallow Junctions," *J. Vac. Sci. Technol.* **B-18**, p. 468, 2000.
**7.** C.-T. Chaung *et al.*, "SOI for Digital CMOS VLSI: Design Considerations and Advances," *Proc. IEEE,* **86**, April 1998, p. 689.
**8.** S.T. Liu *et al.*, "Radiation Response of SOI Materials," *Mtg. Abs. Electrochem. Soc,*. Spring 1999, Abs. 414.
**9.** J.-P. Colinge, "Subthreshold Slope of Thin-Film SOI MOSFETs," *IEEE Electron Dev. Letts.*, **EDL-7**, pp. 244, 1986.
**10.** C.L Chen *et al.*, "High-Frequency Characterization of Sub-0.25-$\mu$m Fully-Depleted Silicon-on-Insulator MOSFETs," *IEEE Electron Dev. Letts.*, **EDL-21,** October 2000, p. 497.
**11.** *ibid.* Ref. 7.
**12.** *ibid. Ref. 9.*
**13.** J.-P. Colinge, "Fully-Depleted SOI CMOS for Analog Applications," *IEEE Trans. Electron Dev.*, **45**, May 1998, p. 1010.
**14.** G.G. Shahidi *et al.*, "A Room-Temperature 0.1-$\mu$m CMOS on SOI," *IEEE Trans. Electron Dev.*, **41,** December 1994, p. 2405.
**15.** N. Kistler *et al.*, "Sub-Quarter Micron CMOS on Ultrathin SOI, *IEEE Electron Dev. Lett.*, **EDL-12**, May 1992, p. 235.
**16.** M. Chan *et al.*, "Recessed-Channel Structure for Fabricating Ultrathin SOI MOSFET with Low Series Resistance," *IEEE Electron Dev. Lett,* **EDL-14**, January 1994, p. 22.
**17.** A. Toriumi *et al.*, "Performance and Reliability Concerns of Ultra-Thin SOI and Ultra-Thin Gate Oxide MOSFETs," *Tech. Dig. IEDM,* 1995, p. 847.
**18.** G.W. Cullen and C.C. Wang, Eds., *Heteroepitaxial Semiconductors for Electronic Devices,* Springer-Verlag, 1978.
**19.** G.W. Cullen *et al.*, *J. Crystal Growth*, **56**, 287 (1982).
**20.** M.S. Abrahams and C.J. Buiocchi, "Cross-Sectional Electron Microscopy of Silicon on Sapphire", *Appl. Phys. Lett.*, **27**, 325 (1975).
**21.** K.E. Violette, P.A. O'Neill, M.C. Ozturk, K. Christensen, and D. Maher, "Low- Temperature Selective Silicon Epitaxy by Ultra High Vacuum Rapid Thermal Chemical Vapor Deposition Using $Si_2H_6$, $H_2$ and HCl," *Appl. Phys. Lett.*, **68**, 66 (1996).
**22.** A. Gupta and P.K. Vasudev, "Recent Advances in Hetero-Epitaxial Silicon-on-Insulator Technology, Part II," *Solid State Technology,* **26**, 129 (1983).
**23.** R. Reedy, T.W. Sigmon, and L.A. Christer, "Suppressing Al Outdiffusion in Ion Implantation Amorphized and Recrystallized Silicon-on-Sapphire Films," *Appl. Phys. Lett.*, **22**, 537 (1983).
**24.** G.A. Garcia, and R.E. Reedy, "Electron Mobility Within 100 nm of the Si/Sapphire Interface in Double-Solid Phase Epitaxially Regrown SOS," *Electron. Lett.*, **22**, 537 (1986).

**25**. G.W. Cullen, M.T. Duffy, and R.K. Smeltzer, "Recent Advances in Heteroepitaxial Si on Sapphire Technology," *VLSI Science and Technology/1984*, Electrochemical Society, N.J., 1984, p. 230.
**26**. D. Munteanu *et al.*, "Silicon-on-Sapphire Material and Device Properties at High and Low Temperature," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 435.
**27**. K. Izumi *et al.*, *Electron Lett.*, **14**, (1978) 593.
**28**. J. Liu *et al.*, "Formation of Buried Oxide in Silicon Using Separation by Plasma Implantation of Oxygen," *Appl. Phys. Letts.*, **67**, p. 2361, 1995.
**29**. One supplier of SIMOX wafers is Ibis Technology Corp, Danvers, MA.
**30**. C.E. Chen *et al.*, *1984 IEDM Tech. Digest*, 701 (1984).
**31**. S. Nakashima *et al.*, "Investigations on High-Temperature Thermal Oxidation Process at Top and Bottom Interfaces of Top Si of SIMOX Wafers," *J. Electrochem. Soc.*, **143**, p. 244, January 1996.
**32**. *Wafer Bonding and SOI Technologies*, Abstracts of the Fall 1997 Meeting of the Electrochemical Society, Paris, France.
**33**. Q.-Y. Tong and U. Gosele, *Semiconductor Wafer Bonding*, VCH-Wiley, New York (1999).
**34**. J.B. Lasky *et al.*, "SOI by Bonding and Etchback," *Tech. Dig. IEDM*, 1985 p. 684.
**35**. R.C. Frye *et al*, "A Field-Assisted Bonding Process for Silicon Dielectric Isolation," *J. Electrochem. Soc*, **133**, 1673 (1986).
**36**. H. Ohashi, "Improved Dielectrically Isolated Device Integration by Silicon-Wafer Direct Bonding Technique," *Tech. Dig. IEDM*, 1986 p. 210.
**37**. Q.-Y. Tong, W.J. Kim, T.-H. Lee, and U. Gosele, "Low-Vacuum Wafer Bonding," *Electrochemical and Solid State Letter*, **1**, 52 (1998).
**38**. A. Plossl and G. Krauter, *Mater. Sci. and Engr Reports*, **R25**, 1 (1999).
**39**. A.J. Auberton-Herve, "SOI: Materials to systems," *Tech. Dig. IEDM* 1996, p. 3.
**40**. Unibond Wafer Specification Sheet. SOITEC USA Inc. Peabody, MA, (1998).
**41**. IBM Microelectronics
**42**. H.W. Lam, A.F. Tasch, Jr., and R.F. Pinzzotto, "Silicon-on-Insulator for VLSI and VHSIC," in *VLSI Electronics*, Ed. N.G. Einspruch, Vol. 4, p. 1, Academic Press, NY (1982).
**43**. S. Farrens *et al.*, "High-Quality SOI Structures by the Genesis Process," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs 396.
**44**. M.I. Current *et al.*, "Ultrashallow Junctions or Ultrathin SOI?" *Solid State Technology*, September 2000, p. 66.
**45**. K. Sakaguchi *et al.*, "ELTRAN by Splitting Porous Si Wafers," *Mtg. Abs. Electrochem. Soc* , Spring 1999, Abs. 395.

**46**. T. Yonehara *et al.*, "ELTRAN: Epitaxial Layer Transfer," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 394.
**47**. K. Suzuki *et al.*, "Scaling Theory for Double-Gate SOI MOSFETs," *IEEE Trans. Electron Dev.*, **40**, 2326 (1993).
**48**. C.H. Wann *et al.*, "A Comparative Study of Advanced MOSFET Concepts," *IEEE Trans. Electron Dev.*," **43**, 1742 (1996).
**49**. M. Tachimori *et al.*, *Proc. Electroche,m. Soc. on SOI Tech & Dev.* Vol. 96-3, p. 53 (1996).
**50**. A. Matsamura *et al.*, "Recent Progress in Low-dose SIMOX Wafers Fabricated with Internal-Thermal-Oxidation (ITOX) Process," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 391.
**51**. D.H. Allen *et al.*, ISSCC, 1999.
**52**. F. Assaderaghi and G. Shahidi, "Advanced SOI Circuits for the Microelectronics Market," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 383.
**53.** "Canon Brings Big Company Muscle to the SOI Market With Yet Another Technology," *Solid State Technology*, June 2001, p. 50 (and 167, an Ad which announces the expanded production of ELTRAN wafers).
**54.** T. Ogawa, "At Long Last, SOI Market is On the Move," *Solid State Technology*, February 2001, p. 62.
**55.** L. Su *et al.*, "Deep-Submicrometer Channel Design in Silicon-On-Insulator (SOI) MOSFETs," *IEEE Electron Dev. Letts.*, Vol. 15, September 1994, p. 366.
**56.** T. Ernst and S. Cristoloveanu, "Ground Plane Concept for Reduction of Short-Channel Effects in Fully-Depleted SOI Devices," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 431.
**57.** H.-S. P. Wong, D.J. Frank, and P.M. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFETs at the 25 nm Channel Length Generation," *Tech. Dig. IEDM* 1998, p. 407.
**58.** E. Leobandung, *et al.*, "Scalability of SOI Technology into 0.13 $\mu$m 1.2 V CMOS Generation," *Tech. Dig. IEDM* 1998, p. 403.
**59.** E. Leobandung *et al.*, "High-Performance 0.18 $\mu$m SOI CMOS Technology," *Tech. Dig. IEDM* 1999, p. 679.
**60.** S.K.H. Fung *et al.*, "Controlling Floating-Body Effects for 0.13 $\mu$m and 0.1 $\mu$m SOI CMOS," *Tech. Dig. IEDM* 2000, p. 231.
**61.** R. Bolam *et al.*, "Reliability Issues for Silicon-on-Insulator," *Tech. Dig. IEDM* 2000, p. 131.
**62.** K. Sakaguchi and T. Yonehara, "SOI Wafers Based on Epitaxial Technology," *Solid State Technology*, June 2000, p. 88.
**63.** M. Current *et al.*, "Atomic-Layer Cleaving and Non-Contact Thinning and Thickening for Fabrication of Laminated Electronic and Photonic Materials," *Proceedings 2001 Materials Research Society Spring Meeting*, Paper 18.3.

**64.** A. Kamgar and J.T. Clemens, "Isolation's Path to SOI Technology," *Solid State Technology*, October 1999, p. 109.
**65.** J.W. Sleight *et al.*, "A High Performance 0.13 $\mu$m SOI CMOS Technology with a 70 nm Silicon film and with a Second Generation Low-$k$ Cu BEOL," *Tech. Dig. IEDM*, 2001, p. 621.
**66.** R. Chau *et al.*, "A 50-nm Depleted Substrate CMOS Transistor (DST)," *Tech. Dig. IEDM,* 2001, p. 621.

# Chapter 12

# MULTILEVEL INTERCONNECTS FOR ULSI

The semiconductor industry continues to pursue the development of ever more powerful ICs. This effort basically involves a two-pronged approach. The first approach seeks to increase the speed of individual transistors through a continual reduction in the minimum size of device features. Such scaling also provides an increase in the device density on the chip. The second approach entails developing increasingly complex interconnect systems, which now use multilayer structures of metal wiring separated by interlayer dielectrics (ILDs). This enables the higher-speed devices to be interconnected (resulting in ICs with enhanced performance and system functionality). Efforts in both of these development arenas are supported by continually more sophisticated device and circuit designs.

The tradeoffs between individual device and interconnect performance depend on the details of the circuit architecture. When the IC industry was in its early stages (i.e., from the mid-1970s to about 1990), circuit design rules were in the 1-5-$\mu$m range. During that time, interconnect delays could generally be ignored because they were typically much smaller than the device switching times (which were ~50-Mz). However, by the mid-1990s, microprocessors and other large area ULSI devices with millions of circuit elements and speeds greater than 1-GHz were envisioned. Designers then began to realize that interconnect delays would become comparable to (and could eventually exceed) the intrinsic delays of the transistors. It was generally agreed that beyond the 0.25-$\mu$m technology node, the overall circuit performance would be dominated by the efficiency with which devices are connected, rather than by the speed of individual devices (Fig. 12-1).[1]

In addition, by the mid-1990s, two different technological domains had also started to collide – namely circuits designed for purely for radio-frequency (rf) applications and others designed merely for digital applications. That is, rf circuits were typically low-density and high-frequency (GHz) designs used in avionics and wireless communication, whereas digital circuits strove for high-density and covered a wide performance spectrum (1-100 MHz). The merger of performance requirements (or *mixed-signal products*), reflects a growing trend,

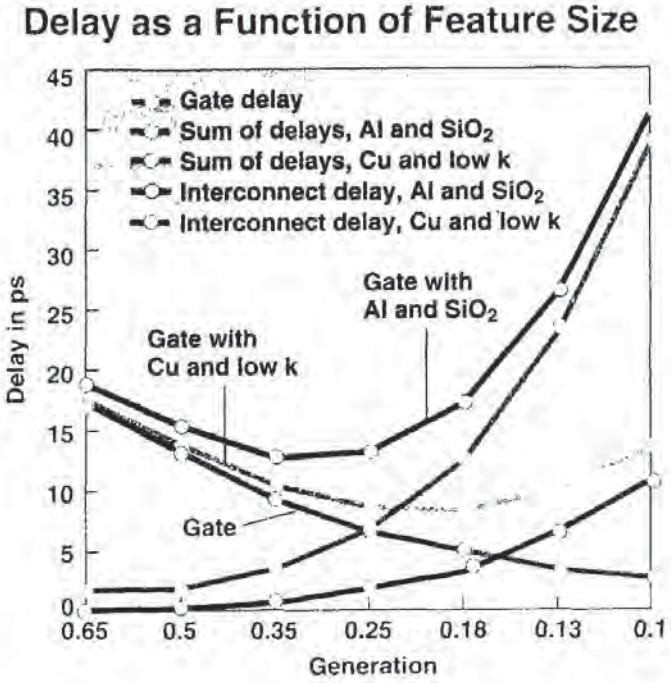



**Fig. 12-1** RC-delay as a function of feature size.[1] (© IEEE 1997)

and places additional demands on interconnect systems as performance requirements continue to escalate.

In this chapter, some of the challenges related to the performance and density issues of ULSI interconnects will be discussed. The discussion will be divided into two parts:

**1.** The performance requirements of ULSI interconnect systems;

**2.** The evolution of the materials and architecture of interconnects for ICs.

The exact feature size at which the interconnect $RC$ delay equals and then surpasses the intrinsic transistor delay depends on the details of the circuit design. Furthermore, the total overall interconnect delay $t_{int}$ is composed of four components and can be approximated by:[2]

$$t_{int} = R_0 C + 0.4\,[(RC)^{1.6} + (tof)^{1.6}]^{1/1.6} + 0.7\,RC_{in} \qquad \textbf{(12.1)}$$

where $R_0$ is the output resistance of the driver device, $tof$ is the time of flight of the electromagnetic wave in the interconnect, and $C_{in}$ is the input capacitance of the device in the next gate. These four terms contribute differently to the overall delay, depending on the gate-length and gate-width, the interconnect line-length, and other geometrical parameters. However, most signs point to the feature size of the order of 0.5-$\mu$m at which the interconnect $RC$-delay starts to become a significant part of the overall signal delay.

## 12.1 EVOLUTION OF THE STRUCTURE AND MATERIALS USED IN IC INTERCONNECTS

Interconnects for ICs have evolved from the single-level-metal system on the original IC patented by Noyce, to the multi-level-metal interconnect systems used in today's ULSI. In this section we will describe the evolution of these interconnect systems. This will point out the limitations that existed for previous generations of IC interconnect technologies, and how they were overcome by learning how to build more advanced interconnect systems. By describing the state-of-the-art of current interconnects, we will also be able to see the limitations they face, and what kinds of techniques are being explored to overcome them.

### 12.1.1 Single-Level-Metal Interconnects in Early ICs

In 1959 Robert N. Noyce invented the concept of the monolithic-integrated-circuit (IC) while at Fairchild Semiconductor. His patent application described how planar silicon-bipolar-transistors and resistors could be interconnected with thin and narrow aluminum lines adherent to an underlying silicon-dioxide film ($SiO_2$) that covered the surface of the silicon chip. Contact holes were etched in the oxide film, and these allowed the aluminum lines to make contact to the circuit devices fabricated within the silicon substrate. After the aluminum lines were patterned, a passivation layer (e.g., a CVD P-doped $SiO_2$ film) would be deposited over the entire wafer. Figure 12-2 shows a schematic of how such a single-level-metal interconnect system was fabricated. The materials used in these early ICs (Al and $SiO_2$) were selected because they had many unique advantages. These advantages proved to be so compelling that this same combination of materials was used for many generations of IC interconnect

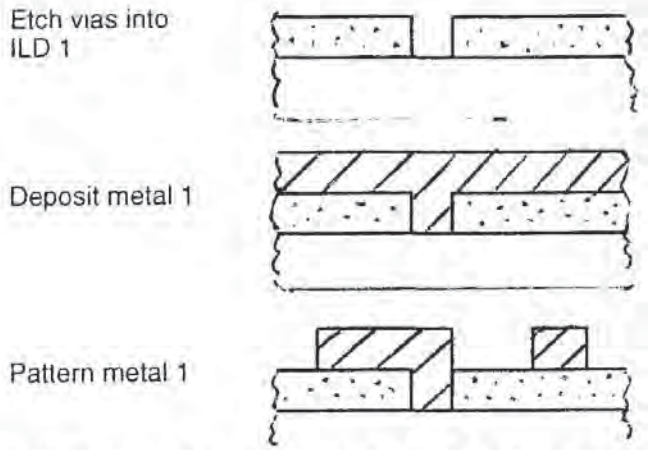



**Fig. 12-2** Schematic of process steps used to produce a single-level-metal interconnect.

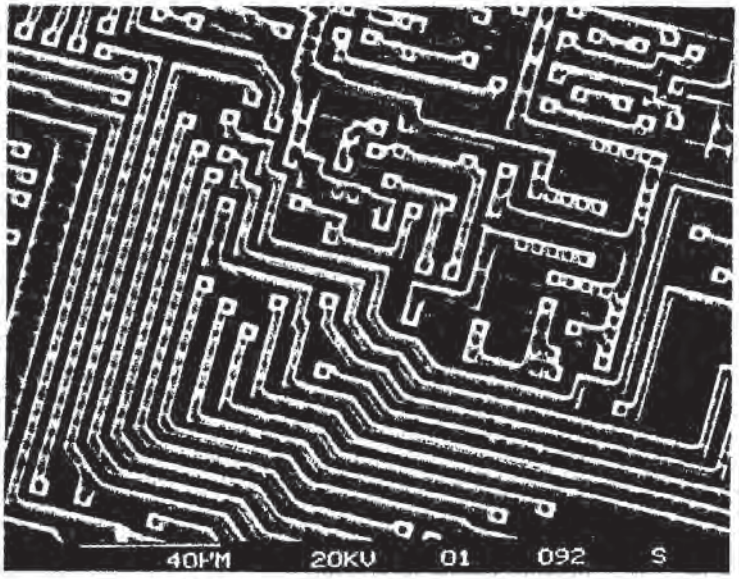



**Fig. 12-3** SEM showing the patterned metal layer on a circuit with only one layer of metal.

technology. As a point of reference for materials parameters, the resistivity of pure Al is $\rho_{Al} = 2.7\text{-}\mu\Omega\text{-cm}$, and the dielectric constant of $SiO_2$ is $k_{SiO_2} = 4.2$ (for the type of $SiO_2$ used in dielectric films under the aluminum). Later versions of this interconnect system used Al films alloyed with Si to improve contact stability. Figure 12-3 shows an SEM of a circuit using only one-level of metal.

Note that the metal wires made in this traditional way of creating the IC interconnect structure are formed by: **1**) depositing a blanket layer of the metal over the entire wafer; and **2**) selectively removing the metal from regions on the wafer where the metal is not needed (i.e., with a *subtractive* etching process).

## 12.1.2 Double-Level-Metal Interconnects in Early Bipolar ICs

As the density of ICs increased with decreasing feature size, it became necessary to have more than one level of interconnect. In NMOS ICs, the polysilicon-gate material could serve as a local-interconnect level (eventually covered with a silicide to form a polycide), and this level could be covered with a single-level of metal (Al-Si) to form a two-level interconnect system.

In bipolar technology there is no MOSFET poly-gate structure, and thus to get two-levels of interconnect, it is necessary to use two-levels of metal (Fig. 12-4). Thus, the first two-level-metal interconnect systems for ICs were used in bipolar technology. (Figure 12-5 shows the steps used to fabricate such a 2-level-metal

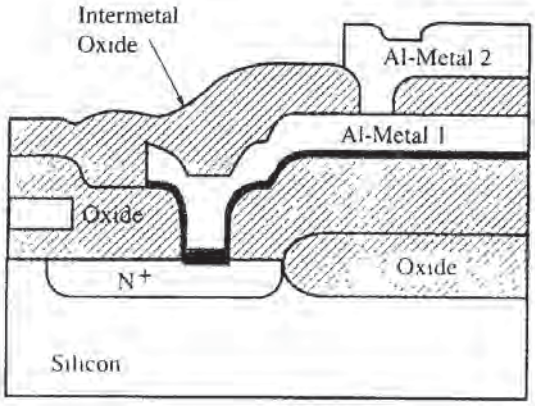



**Fig. 12-4** Schematic of a double-level-metal structure used in early bipolar-ICs.

interconnect structure.) In these 2-level-metal bipolar-ICs there was no attempt made to planarize the dielectric layer that separated the 2-levels of metal (i.e., the *interlevel dielectric*, or *ILD*). The holes opened in the ILD to allow Metal-2 to contact Metal-1 were termed *vias*. While the two-level-metal process could be successfully implemented on bipolar-ICs, the step coverage of Metal-2 as it ran over Metal-1 lines was marginally acceptable. Thus, it would not have been possible to continue to use this process to form a three-level-metal interconnect, even in bipolar-ICs.

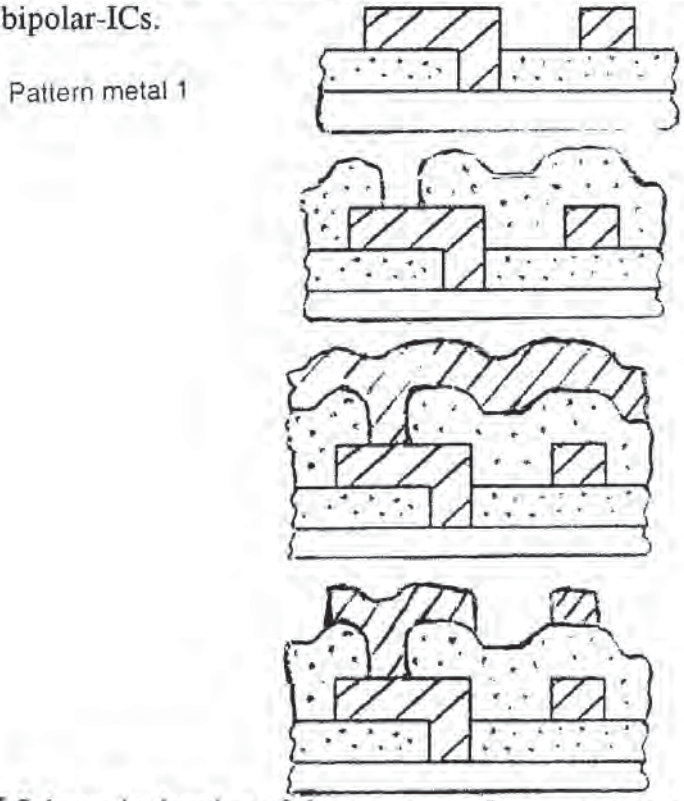



**Fig. 12-5** Schematic drawing of the sequence of process steps used to produce a double-level-metal interconnect structure as shown in Fig. 12-4.

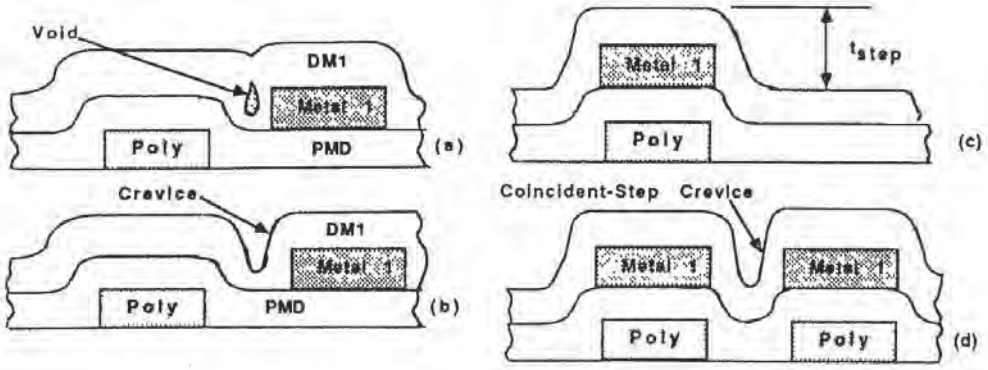



**Fig. 12-6** Topography conditions that arise in MOSFETs that make it difficult to fabricate a double-level-metal structure without some kind of planarization method.

Likewise, when CMOS ICs came to replace NMOS ICs (in the mid-1980's), it was no longer possible to use the polysilicon as a local interconnect layer in the same manner that it had been used in NMOS. This was because polysilicon in early CMOS technologies was *n*-doped, and therefore it could not be used to make contact to PMOS source/drain regions. Instead, it also became necessary to use a double-level-metal interconnect structure in these CMOS ICs. Unfortunately, the same approach to fabricating a double-level-metal structure used in the early bipolar ICs could not be used in CMOS ICs. That is, the steps created by the polysilicon layer (together with the steps caused by Metal-1) were too high to allow adequate step coverage when Metal-2 lines had to run over such (poly + Metal-1 hills), as is shown in Fig. 12-6.

### 12.1.3 Double-Level-Metal Interconnects in CMOS ICs (Partial Planarization of the ILD)

To enable double-level-metal interconnect structures to be fabricated in CMOS ICs, the fabrication sequence used for making double-level-metal interconnects in bipolar ICs had to be modified. That is, it was necessary to be able to partially planarize the ILD prior to depositing Metal-2. The partial-planarization technologies used in the CMOS ICs of the late 1980's included *resist etchback* and the use of *spin-on glass* (SOG). Details concerning these partial-planarization methods can be found in Vol. 2, Chap. 4. Furthermore, the contact holes and vias were opened with a two-step etch process. The first half (which opened the top half of the via) used an *isotropic* etch process, while the second etch step (that opened the remainder of the via) used an *anisotropic* process. This

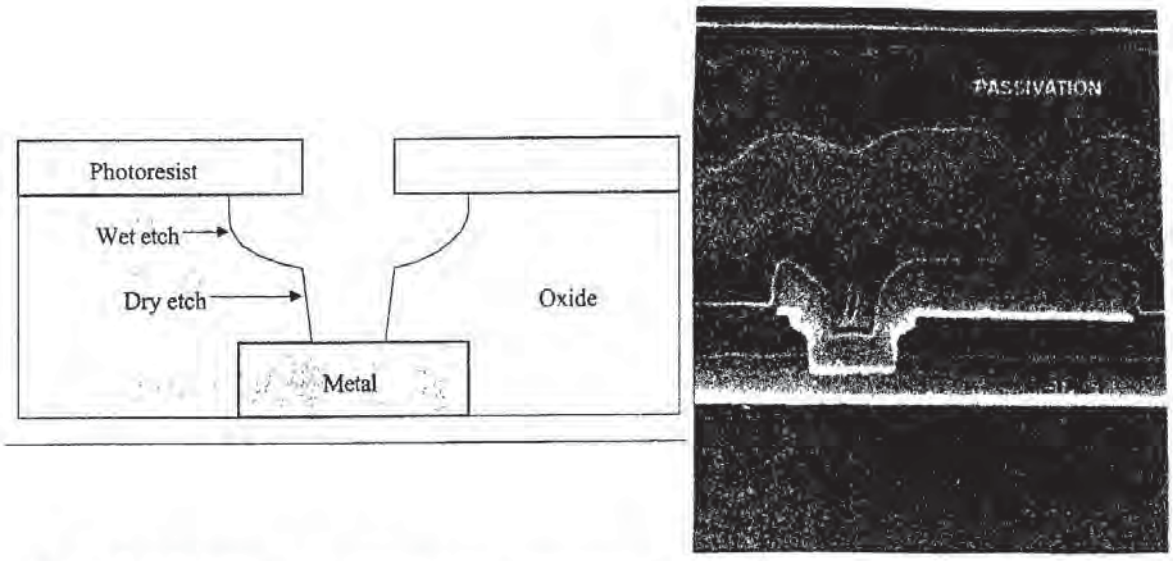



**Fig. 12-7** a) Schematic of a via with a "champagne-glass" shape; b) SEM of actual via.

two-step via-opening process produced via profiles that had a "champagne-glass" shape (Fig. 12-7). Such vias were easier to adequately fill with Metal-2. Note that Al-alloy films were typically used (e.g., Al:2%Cu) because they exhibited better electromigration characteristics than did pure-Al films. Figure 12-8 depicts a schematic of a double-level-metal structure for CMOS technology, and Fig. 12-9 shows a SEM of an actual CMOS IC with such a two-level-metal interconnect-structure. Three-level-metal structures in bipolar-ICs could also be fabricated with this approach.

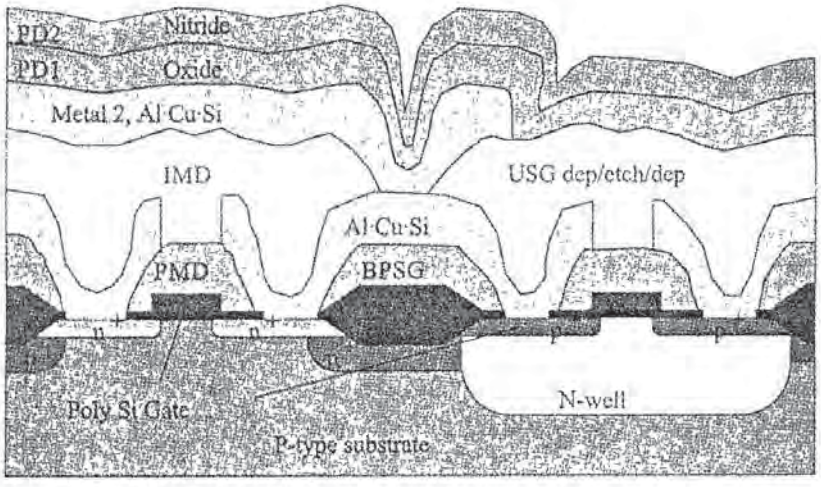



**Fig. 12-8** Schematic of a double-level-metal structure in CMOS with partial-planarization.

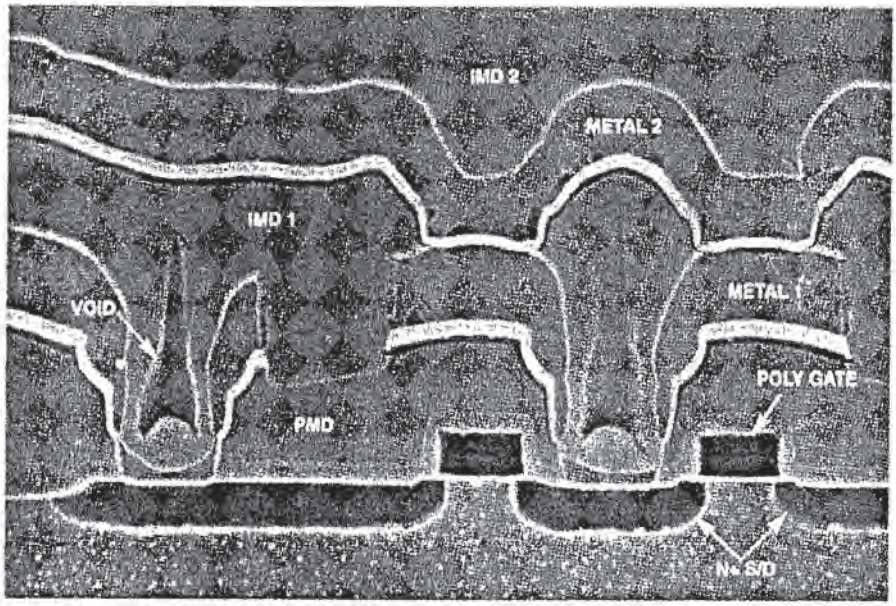



**Fig. 12-9** SEM of an actual double-level-metal structure in CMOS with partial planarization.

The partial-planarization of the ILD allowed the step-coverage of Metal-2 to be sufficiently improved, and the "champagne-glass" via profile allowed Metal-2 to adequately fill the vias. Nevertheless, the topography of the IC remained too rough to allow this technology to be extended for producing three-level-metal interconnects in CMOS-ICs.

### 12.1.4 Three-Level Metal Interconnects in CMOS ICs: (CMP of the ILD & W-Plugs)

To enable three-level-metal interconnects (or more) to be fabricated in CMOS-ICs, a technique for globally-planarizing the ILD after it was deposited was needed. The technology developed for this purpose was *chemical-mechanical-polishing (CMP)* of the CVD-$SiO_2$ ILD. With this technique, the next level of metal would not have to cover any steps, but could be deposited on a completely planar surface (Fig. 12-10).

While CMP of the ILD solved the problem of metal lines having to cross high steps, it created a new difficulty. That is, as shown in Fig. 12-10, the vias that have to be opened in the planarized ILD are of varying depths (with some being quite deep). Such deep vias could no longer be adequately filled by the traditional PVD process used to deposit metal layers (sputtering), even if a two-step via-etch process was used. To address this issue, a CVD (instead of PVD) process was

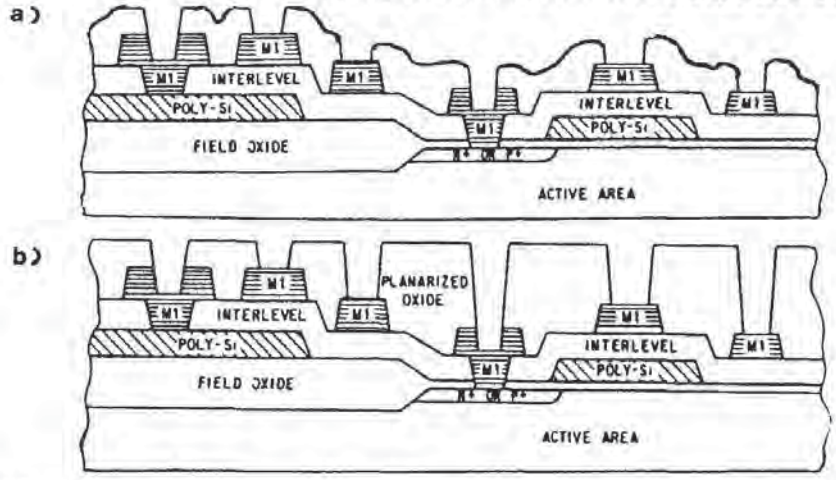



**Fig. 12-10** a) If no planarization is used, the via and contact holes between adjacent conductor levels all have approximately the same depth; b) If planarization of the intermetal dielectric layer is achieved, the vias will vary widely in depth.

used to deposit the metal to fill the vias. The metal chosen to fill the vias in this manner was tungsten (CVD-W). A blanket CVD-W process was developed that could deposit a layer of W over the entire wafer covered with a planarized ILD (Fig. 12-11), such that variable-depth-vias across the wafer could also be completely filled. This blanket-W film would then removed from the planar wafer surface (by etching or CMP), leaving W only in the vias (i.e., *W-plugs* were created). The process of filling recesses in a dielectric layer with a metal film and removing the overburden with etchback or CMP (rather than depositing, patterning, and etching the metal on top of a dielectric layer) is called a

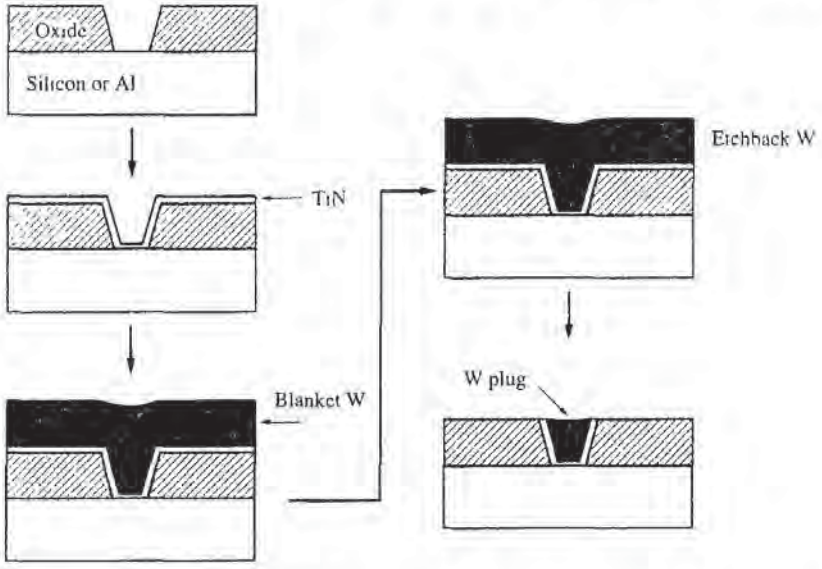



**Fig. 12-11** Schematic of the process used to form W-plugs with CVD-W and etchback.

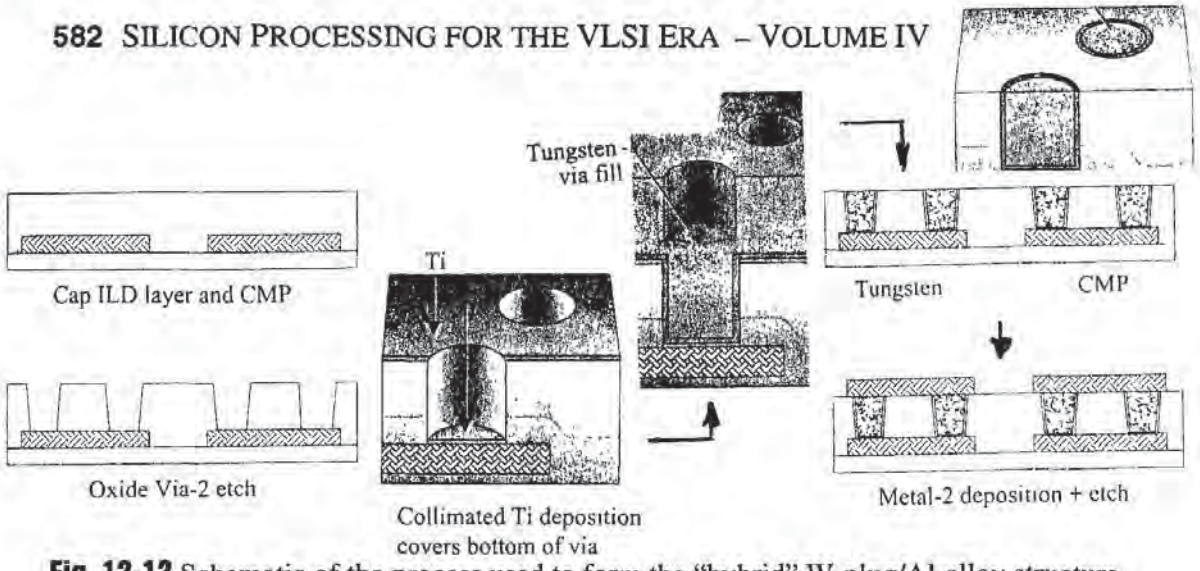



**Fig. 12-12** Schematic of the process used to form the "hybrid" W-plug/Al-alloy structure.

"*damascene*" method. Its name comes from the ancient decorative practice of inlaying metal in wood or ceramics that originated in Damascus.

The layer of aluminum alloy (e.g., Al:2%Cu) from which the interconnect lines are patterned is then deposited and etched by the usual subtractive process. Fig. 12-12 shows a schematic of the steps used to form one-level of this type of interconnect structure. Figure 12-13 shows a schematic drawing and a SEM of such an interconnect structure with stacked vias.

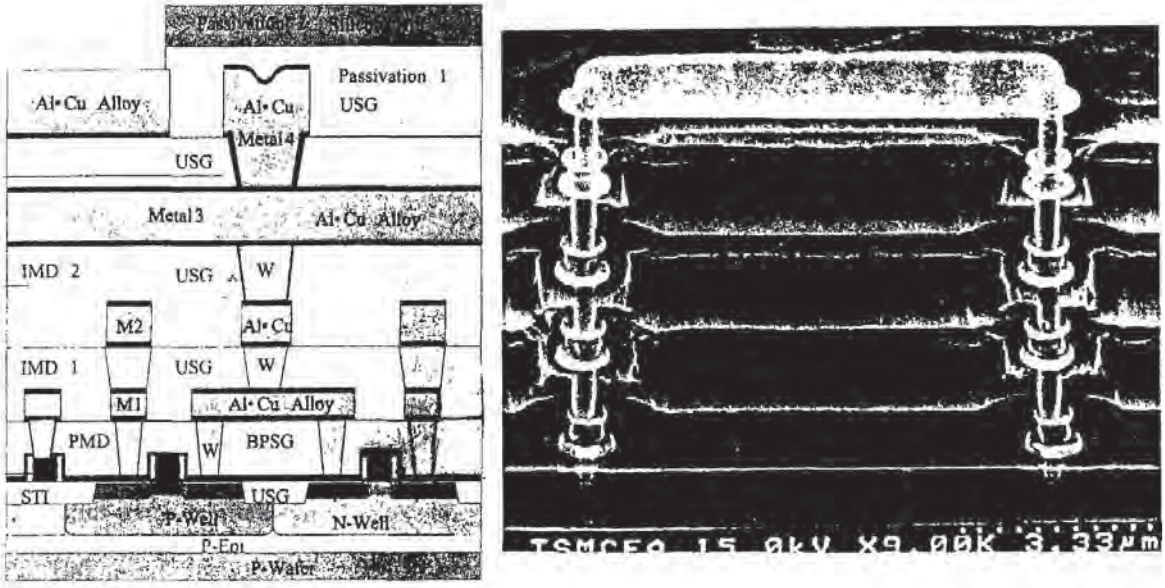



**Fig. 12-13** a) Schematic of a W-plug/Al-alloy interconnect structure with stacked vias; b) SEM of an interconnect structure with stacked vias and five-levels of metal.

Note that in several aspects this sequence involves a hybrid process-technology approach. First, two-types of metal are used, one for the plug that fills the via (W), and another to form the metal lines (Al:Cu). Second, a single-damascene process is used to remove unwanted W from atop the entire surface of the chip, while a subtractive etch-process is used to selectively remove Al:Cu from some regions of the chip surface, and thus form the interconnect lines.

The ILD-CMP/W-plug/Al:Cu-line interconnect-fabrication-approach solved some significant problems, allowing it to be used to form interconnect structures with an indefinite number of levels of metal. These solutions are: **1**) the metal lines on every level have no steps to cover; **2**) vias of varying depth can be completely filled using the CVD-W step; and **3**) the vias can be stacked on top of one another if a CMP process is used to planarize the surface of the W-plugs (which allows the density of the interconnect structure to be increased).

**12.1.4.1 Problems of W-plug/Al-line Interconnect Structures:** While this approach to fabricating multi-level interconnects represented a significant advance, it also introduced some drawbacks. First, it is a complex process, in that it uses two-types of metal, and two-different metal-deposition-processes (each with its own set of process tools). This fabrication complexity is compounded by the fact that two-CMP steps, one ILD-etch-step, and one metal-etch-step are required. Second, the overall resistance of the Al-line/W-plug-structure can grow to be too large for small sized interconnect paths (whereas this resistance value can be reduced by using a lower-resistivity conductor-material - e.g., Cu). Finally, there are some problems associated with W-contacts and vias. That is, electromigration problems in Al often occur in the vicinity W-contacts and vias. This may be partly due to mechanical stresses that can develop at W/Al interfaces (which may enhance electromigration, or promote void and hillock formation). But it is also because the different material in the contact or via serves as a point for flux-divergence. Al also has a lower electrical resistivity than W. This can be detrimental with respect to electromigration effects, especially in circuits where current flow between different materials is significant.

## 12.1.5 Dual Damascene Cu/Low-*k* Interconnects

An alternative approach to forming interconnects on ULSI ICs is the *dual-damascene approach.* The Cu/low-*k* dual-damascene interconnect scheme has been pursued because it has several advantages over the subtractive W-plug/Al-alloy interconnect approach. First, an all-Cu interconnect structure has a lower resistivity and better electromigration resistance characteristics than the W-

plug/Al-alloy system. Second, by incorporating low-*k* dielectrics, the parasitic capacitances of the interconnect structure are reduced, also improving the RC-delay. Third, only one CMP step is needed (i.e., CMP of the Cu/diffusion barrier film stack), compared to the two-CMP steps needed for the subtractive interconnect (i.e., CMP of the W-plug and the ILD dielectric layer). Fourth, the damascene approach avoids having to etch Cu lines, a difficult process to implement in manufacturing. Fifth, only a single interface exists for the via (i.e., the barrier metal in contact with the Cu film in the metal level below, versus the two interfaces at the top and bottom of the W-plug).

The Cu/low-*k* dual-damascene interconnect also has some difficult challenges, some of which are listed here. First, Cu is a fast-diffusing and device-failure-inducing contaminant, even when present in trace concentrations in the silicon. Therefore, it must be surrounded by a diffusion barrier on all sides to prevent it from entering the silicon substrate. Deposition of diffusion-barrier-materials into high-aspect-ratio trenches and vias with sufficient thickness everywhere becomes harder as the via and trench dimensions are scaled. Second, CMP of Cu (and the underlying diffusion-barrier-metal-layer) is also a complex process. Third, the deposition of Cu-layers by electroplating introduces another new, and potentially contaminant-producing process technology into the wafer fab. Fourth, developing low-*k* materials to replace CVD-$SiO_2$ as the ILD layer for the dual-damascene structure is not simple, and new materials will need to be found for several generations of low-*k* dielectric applications. These issues are discussed in detail in other chapters of this book, namely: low-*k* dielectrics (Chap. 14); issues of dual-damascene interconnects (Chap. 15); and copper deposition and CMP (Chap. 16).

The following sequence of steps is used to create a dual-damascene interconnect structure (see Fig. 15-3 in Chap. 15):

**1.** A low-*k* interlevel dielectric (ILD) film is first deposited;

**2.** Vias and trenches are next patterned and etched in the ILD (two masking and etch steps are needed to accomplish this result);

**3.** A multi-layer film (i.e., barrier-metal/Cu) is deposited into the vias and trenches of the ILD (the barrier-layer and Cu-seed-layers are deposited by PVD, and the main Cu-layer by electroplating);

**4.** The barrier-layer/Cu film is removed from regions above the surface of the ILD by CMP, leaving the barrier-layer/Cu recessed in trenches and vias;

**5.** The entire wafer surface is covered with a non-conductive diffusion-barrier-layer material (e.g., SiN) to complete the encapsulation of the Cu.

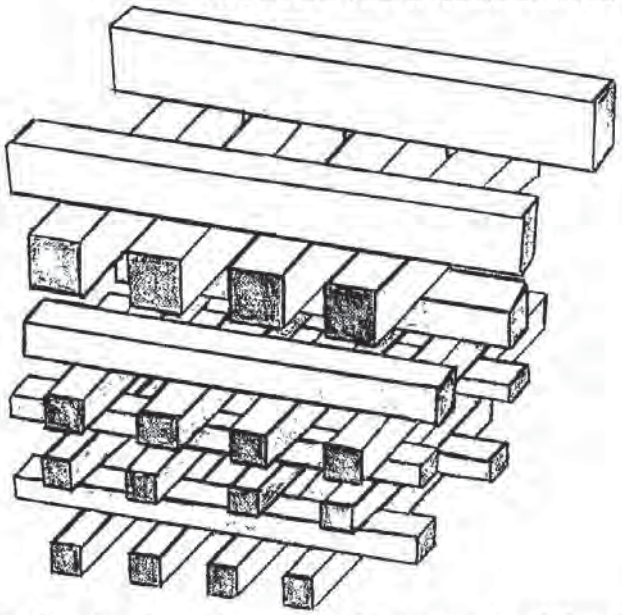


**Fig. 12-14** Schematic showing the orthogonal architecture of multilevel interconnect lines.

### 12.1.6 Other Geometrical Aspects of Multilevel ULSI Interconnects

There are two other geometrical aspects of ULSI multi-level interconnects that should be discussed: **1**) orthogonal directions of alternating metal levels; and **2**) tiers of metal lines containing the same pitch.

In a multi-level-metal interconnect structure, alternating levels are run orthogonal to the levels above and below (Fig. 12-14). This serves three purposes. First, maximum flexibility of routing signals is made possible by using such an orthogonal multilevel structure. Second, the intralevel-capacitance between levels is minimized. Third, crosstalk can be reduced. Figures 12-15 and 12-16 show examples of such an orthogonal structure in a W-plug/Al-line interconnect structure, and in a Cu dual-damascene interconnect structure, respectively. Figure 12-17 shows a cross-sectional SEM of a six-level-metal interconnect structure that reveals the use of orthogonal metal-level layout as well.

Multilevel interconnect structures are also typically arranged in *tiers*. Tiers are pairs of orthogonal metal levels with equal pitch. In some interconnect systems, only two tiers are used (i.e., a *local interconnect-tier* and a *global interconnect-tier,* Fig. 12-18a). In a *local interconnect-tier,* the wires are only used within a cell or a block. In such local tiers the finest metal-pitch allowed by the design rules is employed. In a *global interconnect-tier,* the wires are used for inter-block communication. Thus, the lines may be quite long, and are thicker and wider than those in the local interconnect tier. In interconnect structures that contain six or more levels of metal, there may be a more optimized three-tier network (i.e., one

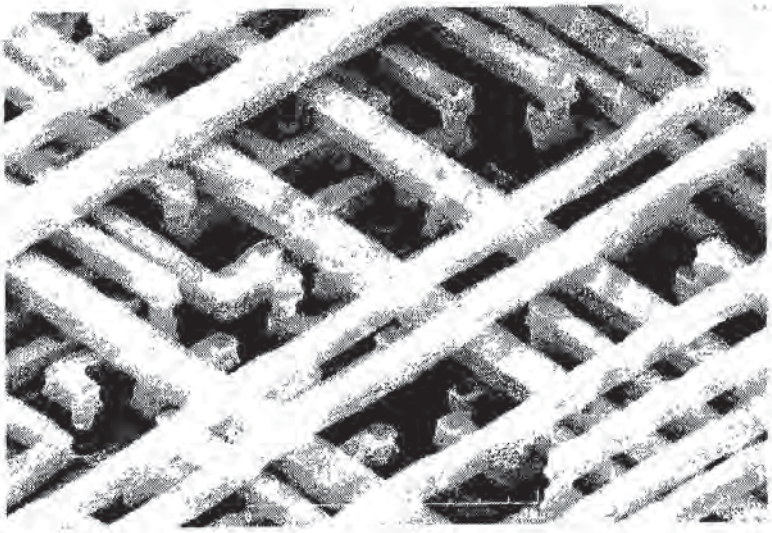


**Fig. 12-15** SEM of orthogonal architecture of metal lines in a W-plug/Al interconnect.

that contains a third, *semi-global tier*, as shown in Fig. 12-18b).

## 12.2 PERFORMANCE REQUIREMENTS OF ULSI INTERCONNECT SYSTEMS

By 2001, microprocessors for consumer PCs were operating at speeds greater than 1-GHz (and even faster in the most advanced workstations). Data in Table

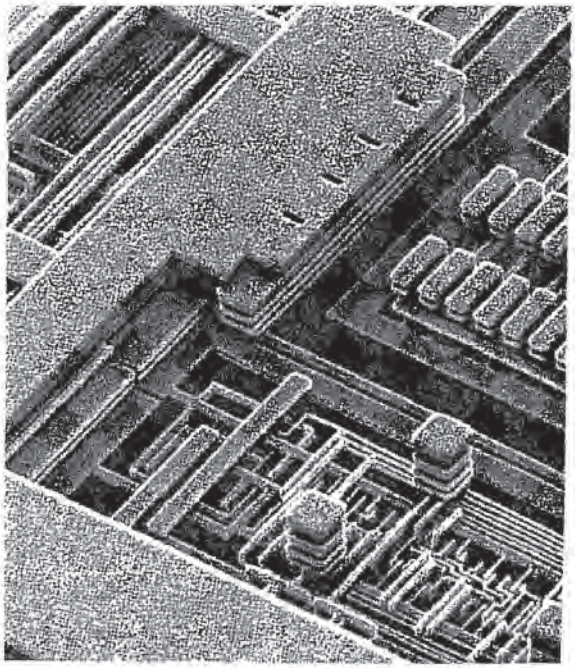


**Fig. 12-16** SEM of orthogonal architecture of metal lines in a Cu interconnect.

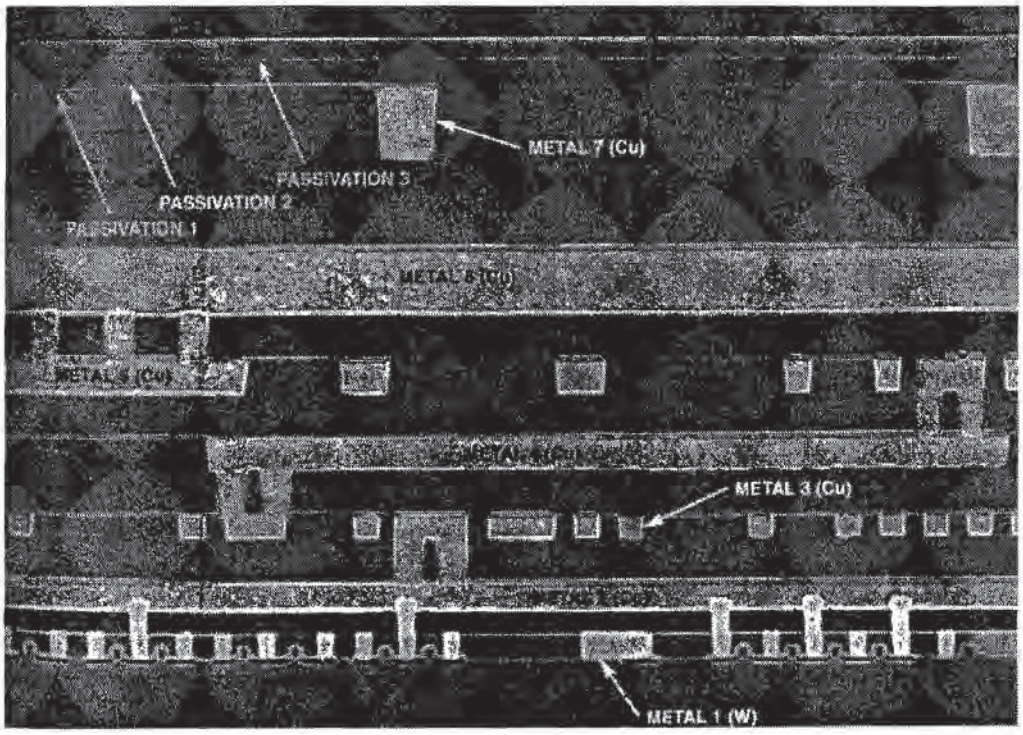



**Fig. 12-17** SEM of a cross-section of a 6-metal-level interconnect.

12-1 taken from the 1999 ITRS,[3] summarizes the projected interconnect needs of microprocessors out to 2014, with respect to performance, density, feature size, and other critical technology parameters. Here we discuss those parameters and issues involved in allowing multilevel interconnect structures to achieve the

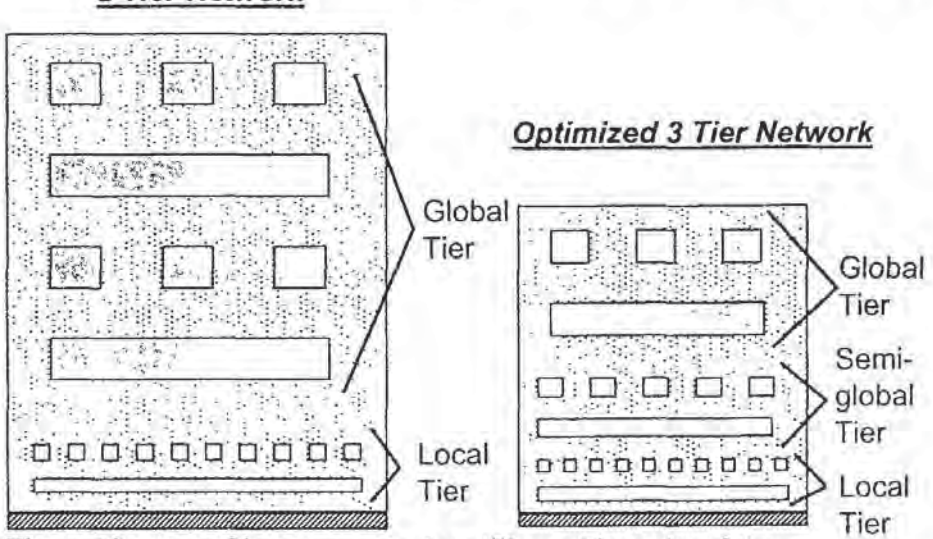



**Fig. 12-18** The architecture of interconnects also utilizes a hierarchy of *tiers*.

**TABLE 12-1 Selected Overall Technology Characteristics based on the 1999 International Technology Roadmap for Semiconductors.[3]**

| Year<br>Technology node | 1999<br>180 nm | 2002<br>130 nm | 2005<br>100 nm | 2008<br>70 nm | 2011<br>50 nm | 2014<br>35 nm |
|---|---|---|---|---|---|---|
| MPU $\frac{1}{2}$ pitch | 230 | 160 | 115 | 80 | 55 | 40 |
| MPU gate length (nm) | 140 | 85–90 | 65 | 45 | 30–32 | 20–22 |
| DRAM $\frac{1}{2}$ Pitch (nm) | 180 | 130 | 100 | 70 | 50 | 35 |
| MPU chip frequency on-chip across chip clock–MHz (Cost-performance) | 600 | 800 | 1100 | 1400 | 1800 | 2200 |
| MPU chip frequency on-chip local clock–MHz (High performance) | 1250 | 2100 | 3500 | 6000 | 10000 | 13000 |
| High volume-MPU (Cost-performance) chip size at ramp ($mm^2$) | 170 | 214 | 235 | 269 | 308 | 354 |
| High volume-MPU (Cost-performance) Mtransistors/$cm^2$ at ramp (including SRAM) | 7 | 18 | 41 | 100 | 247 | 609 |
| Memory chip size at production ($mm^2$) | 132 | 152 | 174 | 199 | 229 | 262 |
| Memory chip Gbits/$cm^2$* | 0.20 | 0.52 | 1.23 | 3.05 | 7.51 | 18.5 |
| Maximum power w/high performance heat sink (Watts) | 90 | 130 | 160 | 170 | 174 | 183 |
| Maximum power (Watts) –Battery (hand-held) | 1.4 | 2.4 | 2.0 | 2.0 | 2.2 | 2.4 |
| Minimum logic Vdd–(Maximum performance) Volts | 1.8 | 1.5 | 1.2 | 0.9 | 0.60 | 0.60 |
| Minimum logic Vdd–(Lowest power) Volts (battery power) | 1.5 | 1.2 | 0.90 | 0.60 | 0.50 | 0.30 |

performance goals listed in Table 12-1. The performance parameters that will be addressed are the following:

1. Propagation-delays *(RC*-delays) in digital circuits;
2. Power dissipation issues in ULSI interconnects;
3. Crosstalk in ULSI interconnects;
4. Contact resistance;
5. ESD protection.

## 12.2.1 Propagation-Delays in ULSI Multilevel Interconnects

As noted in the introduction of this chapter, for technology nodes beyond 0.25-$\mu$m the overall performance of ICs will generally become limited by the performance of the interconnect system on the chip. The most widely used figure of merit of an interconnect system is its $R_{int}C_{int}$-*delay*, which represents a measure of the time-delay for signal propagation. $R_{int}$ denotes the resistance of the interconnect metal-line and $C_{int}$ denotes the effective capacitance between the line and its surroundings. For the multi-level interconnect systems found in ULSI chips at the 0.25-$\mu$m technology nodes and beyond, the estimated effect of various geometrical and materials parameters on the $R_{int}C_{int}$ delay can be modeled using the test-structure shown in Fig. 12-19.[11] This test-structure consists of a set of metal lines between two grounded metal planes, all embedded in the interlayer-dielectric medium (ILD). Let us assign the following symbols to those aspects of the interconnect system depicted by this test-structure.

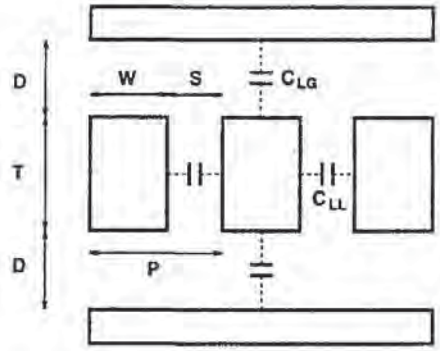



**Fig. 12-19** Simplified *RC*-delay test-structure.[11] (© IEEE 1995)

That is, the metal-line *width* is $w$, the metal-line *thickness* is $t$, the *spacing* between the lines is $s$, the metal-line *pitch* is $p$ ($p = w + s$), and the *thickness of the ILD between the top of the metal lines and the grounded planes* is $d$. Also let the *length* of the metal lines be $l$, and the *resistivity* of the metal be $\rho$ Some equations can now be derived that allow the contributions made to $R_{int}$ and $C_{int}$ from: **1**) the materials properties; and **2**) the interconnect architecture, to be expressed. In these derivations the pitch $p$ and the metal thickness $t$ will serve serve as the parameters in the equations. In that way it will be possible to set $w = ap$, and $d = bt$, where $a$ and $b$ are constants appropriate for the given geometry.

Using this approach, the resistance of a metal line is calculated from:

$$R_{int} = \rho l/wt = \rho l/apt \qquad \textbf{(12.2)}$$

Next, the capacitance between the metal lines and the grounded planes $C_{LG}$ (separated by an ILD whose relative dielectric constant is $k$), is calculated from:

$$C_{LG} = k\varepsilon_o wl/d = k\varepsilon_o apl/bt \qquad \textbf{(12.3)}$$

and the intralevel capacitance $C_{LL}$ (assuming the same ILD) is found from:

$$C_{LL} = k\varepsilon_o tl/s = k\varepsilon_o tl/(p - w) = k\varepsilon_o tl/(1 - a)p \qquad \textbf{(12.4)}$$

where $\varepsilon_o$ is the permittivity of free space. In Eqs. 12-3 and 12-4, $C_{LG}$ represents the "vertical" capacitance between a metal line and a ground plane above (or below), and $C_{LL}$ represents the "horizontal" capacitance between two neighboring lines. The total capacitance $C_{int}$ of a metal line relative to its surroundings is thus written as:

$$C_{int} = 2(C_{LG} + C_{LL}) = 2k\varepsilon_o l\{[ap/bt] + [t/(1 - a)p]\} \qquad \textbf{(12.5)}$$

For the $R_{int}C_{int}$ time delay constant, it follows that:

$$R_{int}C_{int} = 2\,R_{int}\,(C_{LG} + C_{LL}) \qquad \textbf{(12.6)}$$

By combining Eqs. 12-2, 12-3, 12-4, and 12-6, we get:

$$R_{int}C_{int} = 2\rho k\varepsilon_o l^2\{[1/bt^2] + [1/a(1 - a)p^2]\} \quad \textbf{(12.7)}$$

This equation can also be rewritten as:

$$R_{int}C_{int} = 2\rho k\varepsilon_o(l^2/p^2)\{[p^2/bt^2] + [1/a(1 - a)]\} \quad \textbf{(12.8)}$$

The right-hand side of Eq. 12-7 has a natural interpretation, in that the factor $2\rho k\varepsilon_o$ represents the materials contributions, and the factor $l^2\{[1/bt^2] + [1/a(1 - a)p^2]\}$ represents the geometrical contributions, to the interconnect time delay.

Equation 12-8 can also be examined from a slightly different perspective if one defines the aspect ratio $A$ of the metal lines as the ratio of line-thickness to line-width; i.e., $A = t/w$, or equivalently $A = t/ap$. In terms of $A$, Eq. 12-8 can be expressed as:

$$R_{int}C_{int} = 2\rho k\varepsilon_o(l^2/p^2)\{[1/a^2bA^2] + [1/a(1 - a)]\} \quad \textbf{(12.9)}$$

It can be observed that the term $1/a^2bA^2$ in Eq. 12-9 represents the "vertical" contribution to the total capacitance, and $1/a(1 - a)$ represents the "horizontal" contribution to the total capacitance.

Several conclusions about the performance of a multilevel interconnect system can be drawn by further examination of Eqs. 12-7 to 12-9. Two separate cases can be considered: **1**) where the same ILD is used in the interlevel and intralevel spaces (i.e., the value of $k$ is constant throughout); and; **2**) where the spaces between metal lines are assumed to be filled with a material of a lower $k$.

**12.2.1.1 Case When the *k*-Value of the ILD is Constant Throughout the Interconnect System:** In this case, from Eqs. 12-8 and 12-9, it can be seen that for a fixed length $l$ and a fixed aspect-ratio $A$, the $RC$-delay increases quadratically with decreasing feature size (i.e., decreasing pitch $p$). These two equations also indicate that for $A << 1$, the interlayer "vertical" capacitance is more important to the $RC$-delay than the intralayer (or "horizontal") capacitance., whereas the opposite is true for $A >> 1$. Also note that for the simplest geometry (where $a = 0.5$ and $b = 1$ (i.e., $d = t$), Eq. 12-9 reduces to:

$$R_{int}C_{int} = 8\rho k\varepsilon_o(l^2/p^2)\{[1/A^2] + [1]\} \quad \textbf{(12.10)}$$

**12.2.1.2 Case When the *k*-Values of the Interlevel-ILD and the Intralevel-ILD are Different:** In this case, one ILD is used between the metal lines and ground planes (whose "vertical" [between levels] $k$-value is given by $k_{LG}$), and another ILD is used between the metal lines themselves (whose "horizontal" [intralevel] $k$-value is given by $k_{LL}$). In that case, Eq. 12-9 becomes:

$$R_{int}C_{int} = 2\rho\varepsilon_o(l^2/p^2)\{[\,k_{LG}/a^2bA^2] + k_{LL}[1/a(1 - a)]\} \quad \textbf{(12.11)}$$

Again, for the simplest geometry where $a = 0.5$ and $b = 1$ (i.e., $d = t$), Eq. 12-11 reduces to:

$$R_{int}C_{int} = 8\rho\varepsilon_o(l^2/p^2)\{[\,k_{LG}/A^2] + [k_{LL}]\} \qquad \textbf{(12.12)}$$

Equation 12-12 demonstrates several important conclusions. First, if a high aspect ratio interconnect structure is used on a given level (i.e., $A$ is large), the large value of A will mitigate the influence of $k_{LG}$ (i.e., it doesn't matter as much if an ILD with a high or low value of $k_{LG}$ is used between interconnect levels). However, the value of $k_{LL}$ plays significant in determining $R_{int}C_{int}$ for this case. Second, Eqs. 12-11 and 12-12 indicate that the contribution of the intralayer capacitance to the total capacitance (namely, that due to the term $1/\{a[1-a]\}$), is minimized when $a = 0.5$; i.e., when $w$ is one half the pitch $p$.

While Eqs. 12-11 to 12-12 offer some general insight for minimizing the $R_{int}C_{int}$-delay, the interconnect structure can also be treated as consisting of two tiers, termed: **1**) the *local interconnect tier*; and **2**) the *global interconnect tier* (see Sect. 12.1.6). There are some differences in the ways in which the minimum $RC$-delay can be achieved in each of these tiers. The issues of are considered next.

**12.2.1.3 Minimizing the *RC*-Delay in Local-Interconnect Tiers:** The local interconnect tiers are those used to route interconnect lines *within* a cell or block, and in which the finest metal pitch allowed by the design rules are employed. In these tiers, the routing lines are so short that it can be assumed that $R_{int} << R_{trans}$ and $C_{int} >> C_{trans}$ (i.e., the wire resistance of lines shorter than about 1-mm is much less than the transistor on-resistance $R_{trans}$). In addition, the lines are so close that the largest component of $C_{int}$ is intralevel capacitance, $C_{LL}$. In this case, the $RC$-delay in local interconnects is not found from Eqs. 12-2 through 12-4 (which assume the $RC$ is controlled by $R_{int}$ and $C_{int}$) but instead is described by the following equation:

$$R_{trans}C_{int} \sim R_{trans}\,k\varepsilon_o tl/s \qquad \textbf{(12.13)}$$

Thus, unlike in the global interconnect tiers where the $RC$-delay is controlled by $R_{int}C_{int}$ (and where the delay is thus dependent on the *square* of the interconnect length $l$), in the local interconnect tiers the delay is only *linearly-dependent* on the interconnect length. In addition, in this region $C_{int}$ is dominated by the intralevel coupling between adjacent lines. Furthermore, in the local interconnect regions it is also clear that changes in the metal resistance will not play a big role in altering the $RC$ delay, since $R_{trans} >> R_{int}$. This means that in this line-length regime, the resistance and capacitance are decoupled. Consequently, factors that reduce $C_{int}$ (such as lowering the value of $k$ between metal lines, and reducing

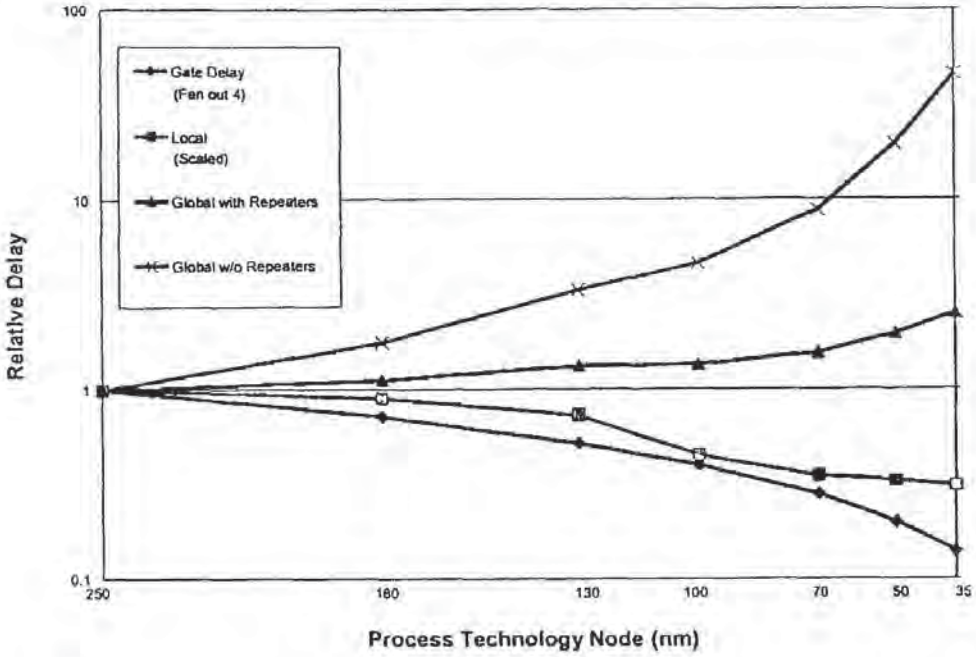



**Fig. 12-20** Interconnect versus delay, from ITRS99.[3]

the height of the metal lines), will help decrease the *RC*-delay (provided that the interconnect resistance of the metal line does not exceed that of $R_{trans}$ due to the reduction in the height of the metal lines). Figure 12-20 shows it is predicted that the relative delay of local interconnect tiers will remain roughly comparable to the gate delay as devices are scaled down toward 50-nm gate-lengths.

**12.2.1.4 Minimizing the *RC*-Delay in the Global-Interconnect Tiers:** The factors that minimize the *RC*-delay in the global tiers of the interconnect structure are somewhat different than the factors that govern minimizing *RC*-delay in the local interconnect tiers. (The *global interconnect tiers* are those used for inter-block communication employing the upper metal levels.) First, there are two coupling regimes that are encountered in global interconnect tiers – the *intralevel regime* and the *interlevel regime*. Figure 12-21a depicts the conditions that exist in the intralevel regime (which is typical of the denser interconnect lines found in the lower levels of global interconnect tiers). In this regime the intralevel capacitance dominates ($C_{LL} >> C_{LG}$). In the interlevel regime (which tends to exist in the upper levels of the global interconnect region, where wide wires are used to lower the resistance of long runs – Fig. 12-21b), the interlevel capacitance dominates ($C_{LG} >> C_{LL}$).

In the *intralevel global interconnect-regime*, $R_{int}C_{int}$ is calculated by combining Eq. 12-2 and 12-4, to get:

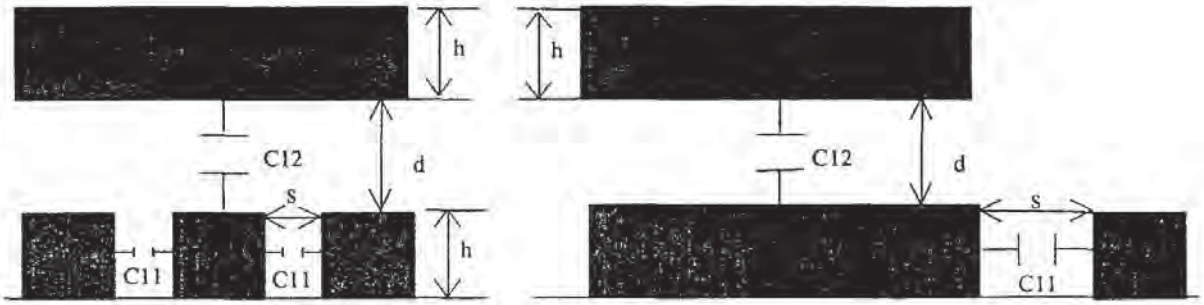



**Fig. 12-21** Capacitance coupling contributions for interconnect systems: a) Dense interconnect case where major coupling is between minimum-spaced lines; and b) Wide interconnect lines where the major coupling is between metal-levels.

$$R_{int}C_{int} = 2R_{int}C_{LL} = 2\,k\varepsilon_o\rho l^2/ws \qquad \textbf{(12.14)}$$

In the *interlevel global interconnect-regime* the interlevel-capacitance dominates, and hence $R_{int}C_{int}$ is calculated by combining Eqs. 12-2 and 12-3:

$$R_{int}C_{int} = 2\,k\varepsilon_o\rho l^2/td \qquad \textbf{(12.15)}$$

We see that in both cases the $RC$ delay is strongly impacted by the length of the interconnect line (i.e., $RC$ increases as the square of the line-length).*

In many circuits it is undesirable to operate under these conditions (i.e., in which the interconnect limits the propagation delay of the circuit). To overcome this limitation, some circuit designs include *repeaters* or *buffers* along the interconnect lines that exceed a critical length. The use of repeaters (while not desirable from a circuit-density standpoint), is nevertheless an essential technique for optimizing signal propagation. Implementing repeaters can even extend the usefulness of a given process technology. Their use will optimally halve the delay of a long line, with a small penalty of the repeater-delay and some device area.[13] Figure 12-22 shows how the addition of repeaters to an interconnect system can reduce the number of levels of metal as the functional density of ICs is increased.[16] In another example, Figure 12-23 depicts the interconnect delay contribution vs. feature size for varying repeater insertion lengths of 3000-$\mu$m

* Note that the intermediate situation (i.e., where both intralevel- and interlevel-capacitance effects are comparable in magnitude) is treated by Eqs. 12-8 and 12-9, since this situation lies between the two extremes and includes capacitive contributions both in- and between-metals-levels.

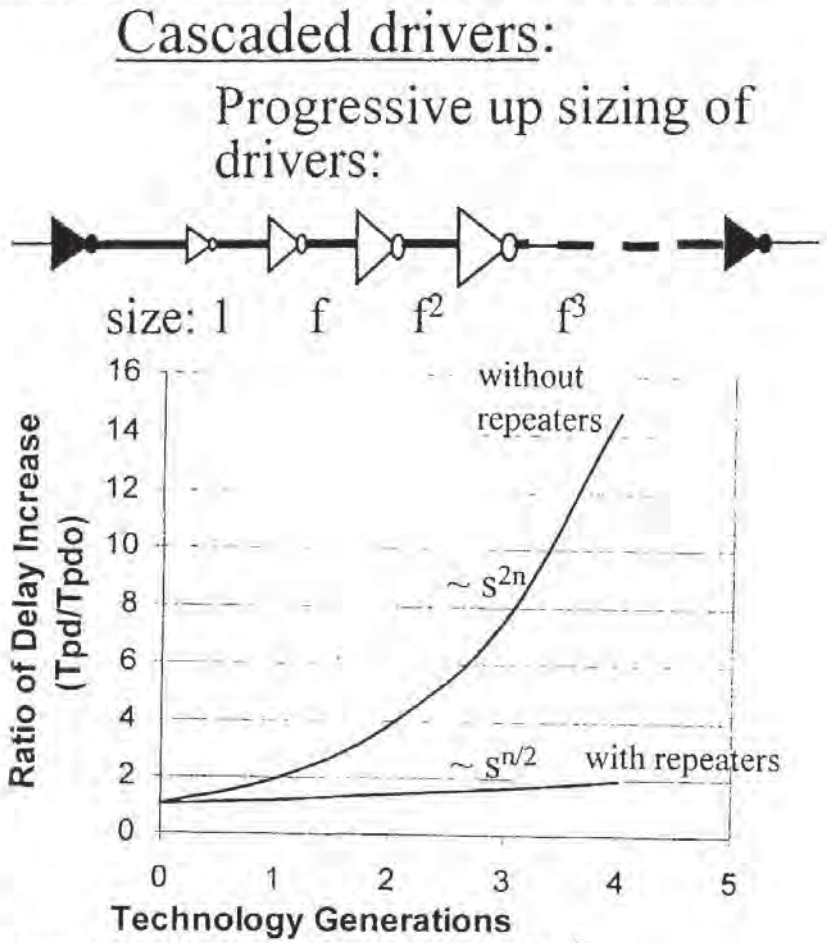



**Fig. 12-22** Impact of repeaters on global propagation delay.[5]

and 5000-$\mu$m.[14] Note that reducing the spacing at which repeaters are placed can compensate for delays due to $RC$ parasitics and circumvent the need for new materials (e.g., Cu and low-$k$ dielectrics).

On the other hand, where long interconnect lines must be used, large pitches $p$ must be employed, and the lines should be made wider and thicker. This implies that long interconnect lines should only be implemented in the upper interconnect metal layers where there is this room for increased pitch (wider lines and spaces), and where the lines can be made thicker. This is the approach that is commonly used today.

Thus, a multi-level interconnect system requires a hierarchy of scales, with the smallest lines in the first metal-level, and progressively wider (and perhaps thicker) and longer lines in the upper levels. One disadvantage of using thicker and wider lines in the upper levels is that fewer metal lines can be fabricated on any level. Consequently, to achieve the same functionality in the face of continued feature size scaling, the number of metal levels used in interconnect systems will also continue to grow.

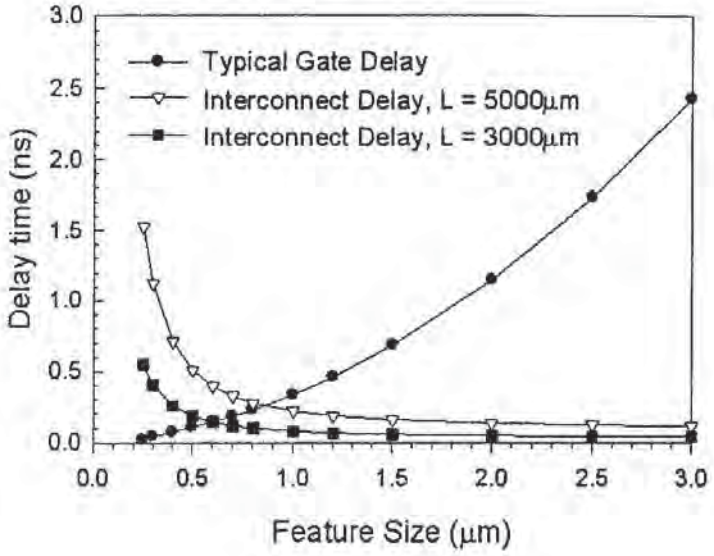



**Fig. 12-23** Gate and interconnect delay versus feature size. Interconnect delay is shown for repeater spacings (L) of 3000-$\mu$m and 5000-$\mu$m.[14]

On the materials side of this issue, it is evident that the lowest $R_{int}C_{in}$-delay in global interconnect-layers will be achieved if a metal with the lowest possible $\rho$ is used together with an ILD having the lowest possible $k$-value. With respect to the metal, this means using copper. (The details of fabricating metal interconnects with copper are described in Chap. 16). With regard to the ILD, it is especially beneficial to selectively reduce $k_{LL}$ between the metal lines. We will explain how this can be accomplished in Chap. 14.

### 12.2.2 Power-Dissipation Issues in ULSI Interconnects

As the requirements for performance in ULSI continue to increase, the power demands will also arise unless the overall system capacitance, resistance and/or operating voltages can be reduced. The power-dissipation associated with a functioning IC can be calculated using either of the following two relationships:

$$P \sim f\, C_{total} V^2 \tag{12.16}$$

or

$$P \sim I_{total}^2\, R_{total} \tag{12.17}$$

where:

P = power dissipated (watts)
f = operating frequency (Hz)
$C_{total}$ = total IC capacitance
= $C_{trans} + C_{int}$ (farads)
$R_{total}$ = total IC resistance
= $R_{trans} + R_{int}$ (ohms)

V = IC operating voltage (volts)
$I_{total}$ = IC operating current (amperes)
$C_{trans}$, $C_{int}$ = transistor & interconnect capacitance contributions
$R_{trans}$, $R_{int}$ = transistor & interconnect resistance contributions

Although Eqs. 12-16 and 12-17 indicate that the IC operating voltage plays the greatest role in determining the power dissipation, its selection is based on circuit and system design issues. On the other hand, the system resistance and capacitance are the chief parameters that can be controlled by the fabrication technology. Currently, the transistor resistive losses are responsible for most of the chip power dissipation (and these are discussed in Chaps. 1 and 5). However, as the ICs are being operated at higher and higher frequencies, an appreciable fraction of the power dissipation and voltage drops are associated with the resistive component of the interconnect. Due to the decreased size of interconnects on ULSI (~200-nm minimum feature sizes in 2000), interconnect resistance is increasing and losses can become significant if the wiring is not designed properly. The overall resistance of the interconnect includes both the line resistance and the resistance of one or more contacts and/or vias along the wiring path. Joule-heating caused by current flowing through the resistive interconnect enhances electromigration and represents a growing reliability concern. The power dissipated at the higher-resistance transition-regions of the contact or via interface is of special concern, especially for large pulsed currents. Since power-dissipation in interconnects is a growing problem, it is necessary to ensure interconnects are properly optimized to reduce such power-dissipation.

### 12.2.3 Crosstalk in ULSI Interconnects

Another issue that impacts the performance of ULSI interconnects is *crosstalk.* Interconnect lines of substantial length act as "antennae" which experience unwanted voltage pulses induced by the rapid voltage swings in neighboring wires of close proximity (which arise from the propagation of digital pulse waveforms along the driven lines). Induced voltages due to intralevel-coupling (the largest crosstalk component between dense wires in the same level) can easily achieve 20% of the supply voltage for very rapidly changing clock waveforms.[4] Figure 12-24 shows a simulation of the amount of crosstalk experienced by an undriven wire that is 1-mm long, separated by 100-nm from an adjacent interconnect line of the same length and driven with a very fast clock edge.[5] V1 is the driving waveform and V2 is the crosstalk voltage induced in the undriven line. The rate of rise of the clock edge imposes a large induced voltage

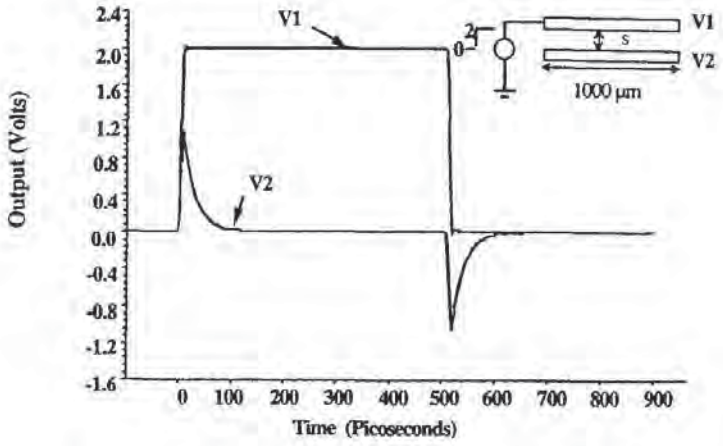



**Fig. 12-24** Electrical modeling result showing the crosstalk voltage induced from a driven interconnect-line (V1) into an undriven line (V2) seeing a 20-psec waveform-edge. The lines were spaced at s = 1000-nm with a common length of 1-mm.[10]

which is related to the parasitics of the interconnect. In this example, where the clock-edge rise time is 20-psec, the magnitude of the imposed crosstalk voltage can be as much as 60% of the voltage waveform V1 (2 x 0.6 = 1.2V). As the intralevel-spacing between interconnects decreases, the magnitude of the crosstalk noise can become substantial, as is evident from the simulation data shown in Fig. 12-25.

It is more difficult to establish the level of crosstalk in dynamic systems than in the example given above. More sophisticated circuit simulators are required to

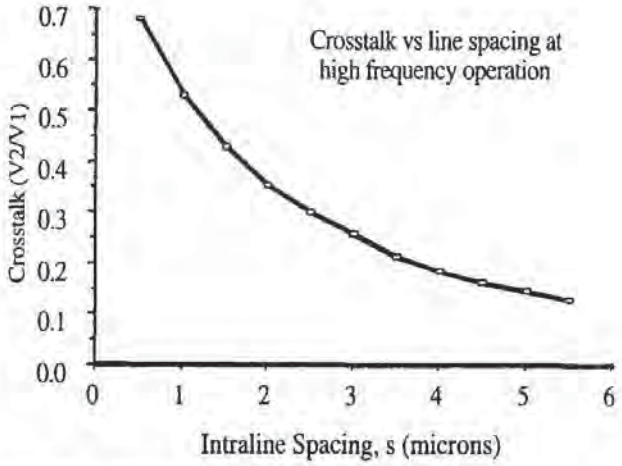



**Fig. 12-25** Simulation results showing that the magnitude of the crosstalk between 2 interconnects separated by different line spacings, s.[10] Conditions described in Fig. 12-24.

estimate the impact of such crosstalk on circuit device performance. The potential of surrounding wires, wire resistance, wire length, and the capacitive load all affect the amplitude of crosstalk signals. Nevertheless, it is possible to reduce intralevel crosstalk by changing the relative dielectric environment. That is, by inserting an embedded low-$k$ dielectric between dense intralevel wires, electric-field lines can be diverted to the short orthogonal lengths of underlying interlevel wiring (lowering capacitive-coupling between intralevel interconnects).

Crosstalk will become a significantly greater problem as ULSI mixed-signal devices begin to emerge in the marketplace. It is predicted that noisy digital wires and circuit elements will cause a number of problems for analog elements that are in close-proximity to such sources of noise. New techniques involving such novel electrical-isolation-methods as shielding and signal conditioning will be needed. This is because, in some circuits, analog devices must be immune to such "noise" down to levels that are three to six orders of magnitude smaller than the power-supply voltage-levels.

### 12.2.4 Contact Resistance In ULSI Interconnects

The magnitude and uniformity of contact and via resistance is another forthcoming challenge (see Vol. 2, Chap. 3 for more background information on contact and via resistance). Contact and via resistances can be substantial and have a significant impact on the overall wiring resistance of the interconnect system. These resistance contributions act in the same way as adding length to an interconnect line, and thus can factor heavily into voltage drops and delays in the wiring. For 0.25-$\mu$m contacts with an interfacial-contact-resistivity $\rho_c$ to heavily doped silicon of $5x10^{-8}$-$1x10^{-7}$-$\Omega$-cm$^2$, such contact-resistance values are approximately 100-200-ohms/contact.[6] Vias for systems using Al alloys with W-plugs have practical metal-to-metal contact resistivities of approximately $2x10^{-9}$ $\Omega$-cm$^2$ and a resistance of about 5-$\Omega$/via at the same geometry.[7] Vias which use copper metallization, however, are expected to show substantially lower contact-resistance because they have only one critical barrier interface per via (instead of two interfaces when W-plugs and Al alloy interconnect lines are used), and because the interfacial-contact-resistivity should be lower than $1x10^{-9}$-$\Omega$-cm$^2$.[8]

Cleaning prior to the deposition of the contact- and via-filling conductor-films and control of dimensional tolerances are critical to the reproducibility of via or contact resistance. It is predicted that production variability will increase with the migration to smaller feature sizes and have a negative impact on resistance. This

resistance must be accounted for in the delay analysis when multiple levels with a large number of contacts and vias are used.

### 12.2.5 Electrostatic Discharge Protection Issues of Interconnects

Electrostatic discharge events (ESD) can cause large transient currents to flow in an IC. This can cause "fusing" in upper wiring levels, thus jeopardizing the long-term reliability of the IC. ESD robustness has been shown to be a severe reliability concern by several research groups, for interconnect generations with metal features smaller than 0.25-$\mu$m.[9,10] Since short term discharges can exceed 2000 volts for durations of hundreds of nanoseconds, it has been shown interconnects are becoming more prone to failure due to this type of electrical stress. Copper interconnects exhibit a substantially better resistance to this type of failure, when compared to aluminum interconnects, due to the higher melting point and thermal conductivity of copper.

## 12.3 SUMMARY OF USLI INTERCONNECT ISSUES AND FUTURE TRENDS

Based on the discussion of this chapter, the following trends are likely to be encountered in ULSI interconnect systems:

1. If the materials of an interconnect system are not changed (i.e., the metal resistivity $\rho$ and ILD-$k$ values remain the same), the $RC$-delay will continue to increase as device dimensions (and therefore interconnect pitch $p$) are reduced. There are two main reasons why this increase in $RC$ occurs:

   a. Increasing the aspect ratio $A$ of metal lines (to keep $R$ constant) beyond about 2 produces diminished improvements in $RC$ - even for a constant pitch $p$ (because $C_{LL}$ rises while $R$ remains constant). However, if the pitch is reduced, this makes $C_{LL}$ grow even larger, which, in turn causes an additional rise in $RC$.

   b. The metal line lengths tend to scale less strongly than does the metal pitch $p$. Hence, even as devices keep shrinking, the ICs will continue to contain some rather long lines (which will have larger $C_{total}$/unit length, for lines with the same length).

   We noted earlier, however, that in short- and intermediate-length local interconnects, changes in the metal resistance will typically not play a major role in the $RC$-delay (since in this regime $R_{trans} >> R_{int}$). For such cases, the resistance and capacitance are decoupled. That is, factors that reduce interconnect capacitance (such as a reduction in the dielectric

constant of the ILD, and/or a reduction in the height of the metal lines), will help improve *RC* delay (as long as the interconnect resistance increase due to reducing the metal film thickness does not approach $R_{trans}$).

**2.** If $\rho$ and/or $k$ can be decreased (especially a reduction in the value of $k$ of the ILD used in intralevel metal spaces $k_{LL}$) this will cause a decrease in *RC*. Hence, there will be continual effort expended to integrate Cu and low-$k$ dielectrics into ULSI interconnect systems. The issues involved with incorporating low-$k$ ILDs are covered in Chap. 14, and the challenges encountered when attempting to implement Cu-metal-interconnects are described in Chap. 16. Figure 12-26 shows a possible migration path for materials use in IC interconnect structures.[15]

**3.** More levels of metal will need to be used as the feature sizes are decreased.

**4.** Optimizing interconnect length is a powerful method for reducing *RC*-delay. This can be influenced by design philosophy and device architecture. Use of repeaters, while not desirable from a density standpoint, is another key approach for optimizing signal-propagation, and can extend the use of a given process technology. Since incorporation of repeaters into an interconnect system is a circuit-design issue, this approach not considered further in this text.

**5.** Increasing the metal line aspect-ratios and adding more levels of interconnect are two methods for improving interconnect performance and density. However, these approaches will reach practical limits in a few

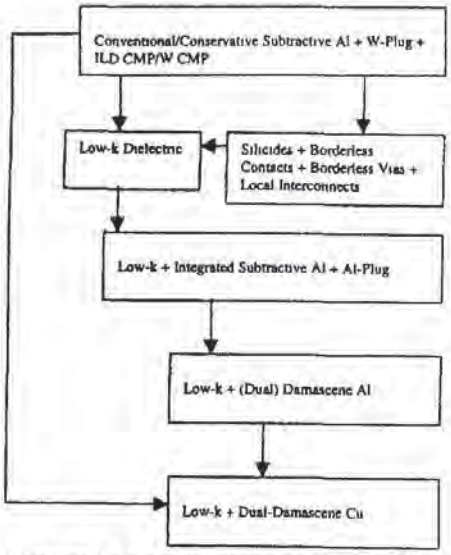



**Fig. 12-26** Schematic illustrating an interconnect migration.[15]

generations.[11] (For example, continuing to increase the aspect ratio $A$ above $A = 2$ will bring only diminished improvements in $RC$ because the reduction in the metal resistance will be offset by an increase in intralevel capacitance.) New interconnect materials, improved circuit-design techniques, and possibly novel interconnect schemes will be needed to address the growing limitation that interconnects pose to ULSI performance. Circuit design needs to more accurately model interconnect performance to avoid overdesign or unexpected performance limitations, and automatic layout tools need to make more efficient use of existing interconnects. The novel interconnect approaches include RF-transmission and optical-interconnects. More information about these novel techniques can be found in Ref. 12.

## REFERENCES

**1.** S.C. Sun, "Process Technologies for Advanced Metallization and Interconnect Systems," *Tech. Dig. IEDM*, 1997, p. 765.

**2.** R. Liu *et al., Proc. Int. Interconnect Technol. Conf.* (IEEE), 1999, p. 21.

**3.** International Technology Roadmap for Semiconductors, 1999 Edition, Semiconductor Industry Association, San Jose, CA.

**4.** C. Cregut *et al.,* "Low-*k* Dielectrics Influence on Crosstalk: EM Analysis and Characterization," *Proc. of the International Interconnect Technol. Conf.*, San Francisco CA, June 1-3, 1998, p. 59.

**5.** H.P. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison Wesley Publ., Reading MA, 1990.

**6.** I.S. Park, *et al.,* "Fabrication of 256 Mbit DRAM with a Fully Planarized Double Level Interconnection," *Proc. 12th VLSI Multilevel Interconnect Conf. (VMIC)*, Santa Clara CA, 1995, p. 45.

**7.** K. Kikuta *et al.,* "Comparitive Study of W-plug, Al-plug, and Al-Dual-Damascene for 0.18 $\mu$m ULSI Multilevel Interconnect Technologies," *Proc. of the Internat. Interconnect Technol. Conf.,* San Francisco CA, June 1-3, 1998, p. 140.

**8.** J. Zhang *et al.,* "CVD Cu Process Integration for sub-0.25 $\mu$m Technologies," *Proc. of the Internat. Interconnect Technol. Conf.,* San Francisco CA, June 1-3, 1998, p. 163.

**9.** S.H. Voldman, "ESD Robustness and Scaling Implications of Aluminum and Copper Interconnects in Advanced Semiconductor Technology," *IEEE EOS/ESD Symposium , 1997, p. 316.*

**10.** K. Banerjee *et al.*, "Characterization of VLSI Circuit Interconnect Heating and Failure under ESD Conditions," *1996 IEEE 34th International Reliability Physics Proceedings*, April 30-May 2, 1996, Dallas, TX, p. 237.
**11.** M.T. Bohr, "Interconnect Scaling – The Real Limiter to High Performance ULSI," *Tech. Dig. IEDM*, 1995, p. 241.
**12.** M.E. Thomas and R.H. Havemann, "Overview of Interconnect," Ch.12, in *Handbook of Semiconductor Manufacturing Technology*, Y. Nishi and R. Doering, Eds. Marcel Dekker, New York, 2000, pp. 303-4.
**13.** H.P. Bakoglu and J.D. Meindl, " Optimal Interconnection Circuits for VLSI," *IEEE Trans. Electron Dev.*, May 1985, p. 903.
**14.** J.C. Eble *et al*, "Minimum Repeater Count, Size, and Energy Dissipation for Gigascale Integration (GSI) Interconnects," *Proceedings of the International Interconnect Conf.*, San Francisco CA, 1998, p. 56.
**15.** J. Givens, "Engineering CMP for Manufacturability and Yield," *Abs. Mtg. Electrochem. Soc.*, Fall 1999, Abs. 1275.
**16.** S. Wills, "Interconnect Architecture," in *IEDM* 2000 *Short Course*: *Advanced Interconnects: Design, Process, Integration*, p. 27.

# Chapter 13

# POLYCIDES AND SALICIDES OF $TiSi_2$, $CoSi_2$, AND NiSi

Silicon forms many stable metallic and semiconducting compounds with metals, and such compounds are termed *silicides*. Several metal-silicides exhibit low-resistivity and high thermal-stability, making them suitable materials for several ULSI applications. The formation kinetics, film properties, and potential applications such silicides have been extensively examined by Murarka[1] and by Nicolet and Lau.[2] In this chapter the use of silicides for the purposes of fabricating polycides and salicides in ULSI applications is discussed.

## 13.1 POLYCIDES AND SALICIDES

The term *polycide* is a condensed form of the words polysilicon-silicide. Polycides are widely used for gate structures in MOSFETs that contain a non-self-aligned silicide. In this structure, a silicide is typically blanket-deposited over a doped polysilicon film and the composite film is patterned together (Fig. 13-1). Polycide gates can achieve a gate sheet-resistance smaller than 10-Ω/sq, compared to that of a polysilicon gate without a silicide of more than 40-Ω/sq. Even with its considerably lower-resistivity, the polysilicon gate with a silicide can only be used for local-routing over short-distances in deep-submicron ICs.

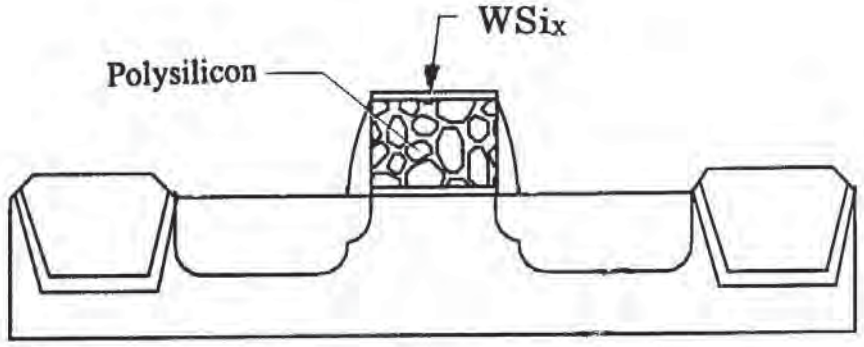



**Fig. 13-1** Schematic cross-section of a polycide gate on a MOSFET.

Still, such a local-interconnect can be beneficial when only one or two levels of metal are used in a circuit - since the poly routing can help to reduce die size by as much as 20%. In fact, the incorporation of a polycide allowed the conductivity of poly lines serving as local interconnects to be so significantly reduced that the invention of the polycide-gate-structure was responsible for delaying implementation of multilevel-metallization in CMOS by several years.

Several types of polycides have been employed in IC technology, including $WSi_2$, $TaSi_2$, and $MoSi_2$. All are refractory-metal-based, and exhibit adequate thermal stability and resistance to processing chemicals. Since such polycides are formed as part of the gate fabrication sequence, they must be able to withstand the temperatures of any processes carried out after gate formation (typically 800-1000°C). Consequently, as $WSi_2$, $TaSi_2$, and $MoSi_2$ films remain stable at these elevated temperatures, they can be used for such gate polycides.

However, the silicides of Ti and Co (i.e., $TiSi_2$ and $CoSi_2$) are unstable above about 900°C (i.e., they agglomerate above ~900°C). Thus, it is not possible to use these materials as polycides if the gate formation sequence is followed by a thermal step that exceeds 900°C. Even if a process is used in which this temperature is not exceeded, neither $TiSi_2$ or $CoSi_2$ is a good candidate for a polycide that is patterned at the same time as the polysilicon gate. This is primarily due to an issue that arises during the patterning of the poly gate. That is, silicides composed of $TiSi_2$ are attacked by the Cl-based chemistry used for polysilicon dry-etching, and this makes linewidth control difficult if $TiSi_2$ must be etched during the patterning of the critical gate level. $CoSi_2$, on the other hand, lacks a volatile reaction by-product and is thus is difficult to pattern by dry etching. In addition, both $TiSi_2$ and $CoSi_2$ are vulnerable to processing chemicals, especially to HF solutions.

However, $TiSi_2$ and $CoSi_2$ are ideal candidates for *self-aligned-silicide* (also referred to as *salicide*) applications. As will be seen, such salicide processes are performed *after* the polysilicon lines for the gate structure have been patterned, and the source/drain regions of the MOSFET have been fabricated. Since the formation of the $TiSi_2$ or $CoSi_2$ in the salicide process are self-aligned techniques that require no etching, the etching difficulties mentioned above are avoided (as is exposure to temperatures > 900°C), permitting these materials to be used.

As the minimum dimensions of an IC approached 1-$\mu$m, the conventional contact structures used up to that point (see Chap. 3, *Silicon Processing for the VLSI Era*, Vol. 2 for details on contact formation in VLSI ICs) began to limit device performance in several ways. First, it was not possible to minimize the contact resistance if the contact hole was also of the minimum feature size, and

problems with cleaning the small contact holes became another concern. In addition, the area of the source/drain-regions could not be minimized because the contact-hole had to be aligned to these regions with a separate masking step, and extra area thus had to be allocated for misalignment. (The larger area that had to be used for such source/drain-regions also resulted in increased source/drain-to-substrate junction-capacitance, which slowed-down device speed).

Finally, when non-minimum-width MOSFETs were fabricated with conventional contacts, several small, uniform-sized contact holes were usually used rather than one wider contact hole. (The reason for this is that if all of the contact holes on an entire chip are of the same size, they are more likely to clear simultaneously during the etching process.) A problem with using either one wide contact hole (as well as if several contact holes of equal size are used) is that a metal film must be used to contact each contact hole. The result is that valuable routing area is thus removed from the chip surface (because it must be used for running metal connections to the contacts), and this results in an increase in the die size. The main approach used to alleviate this problem has been to employ a self-aligned silicide on the source/drain-regions (where these silicides are formed at the same time as the polycide structure is created). The process of simultaneously fabricating such self-aligned-silicides and polycides on the gate is referred to as a *salicide* (self-aligned silicide) process.

## 13.2 POLYCIDES FABRICATED BY CVD OF TUNGSTEN SILICIDE ($WSi_X$)

As mentioned earlier, various refractory metal silicides including $WSi_X$, $TaSi_2$, and $MoSi_2$ were explored in the 1980s as possible shunt materials in polycide structures. Tungsten silicide ($WSi_X$) films ended up being the most widely adopted for this application, with CVD being the chosen method of forming the $WSi_X$ thin-films. Polycides with $WSi_X$ films now find widespread use as word-line and bit-line interconnects in IC memory chips. Note that unlike CVD-W, which is used for plug/interconnect applications (and is thus never exposed to temperatures above about 500°C), these $WSi_X$ films are deposited on polysilicon during the gate formation process. This means that such $WSi_X$ films will have to have to be able to endure temperatures up to 800-1000°C during BPSG reflow, source/drain implant-activation, and poly reoxidation. $WSi_X$ has also found some application as a stand-alone adhesion layer for blanket CVD-tungsten films.

A process sequence for producing the $WSi_X$ polycide gate-structure is depicted in Fig. 13-2. It shows that $WSi_X$ is blanket-deposited onto the doped polysilicon film, and then this $WSi_X$/poly multilayer is etched to form the polycide gate-structure. The CVD deposition of $WSi_X$ won out over other candidate processes

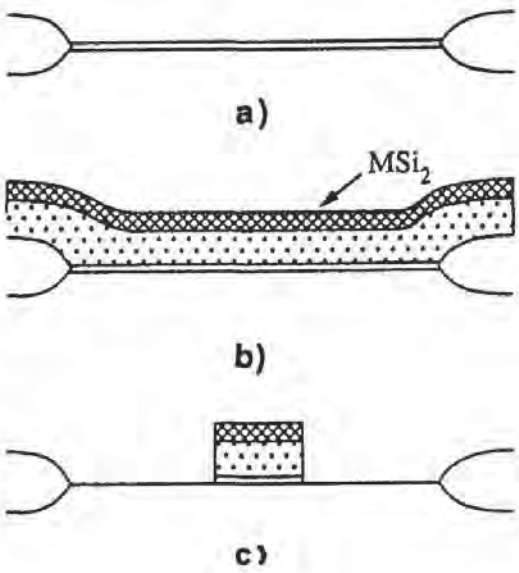



**Fig. 13-2** Process sequence for producing a $WSi_x$-polycide-structure: a) Gate oxide growth; b) Polysilicon and $WSi_x$ deposition by CVD; c) Patterning the polycide.

(i.e., evaporation and sputtering of $WSi_x$) for the following reasons: **a**) this process can produce high-purity $WSi_x$ films without the need for high-vacuum deposition equipment; **b**) its throughput is acceptable; **c**) better step-coverage is obtained than with PVD; and **d**) good wafer-to-wafer and run-to-run uniformity is attainable. The first chemical reaction used to deposit $WSi_x$ was the following, and it is capable of producing films with the benefits listed above:

$$WF_6\ (gas) + 2\ SiH_4\ (gas) \rightarrow WSi_2\ (solid) + 6\ HF\ (gas) + H_2\ (gas) \qquad \textbf{(13-1)}$$

Tungsten-silicide film deposition using this reaction is carried out at pressures between 50-300 mtorr (LPCVD) and at deposition temperatures between 300-400°C.[3] This reaction is similar to the silane reduction of $WF_6$ used to produce CVD-W films. As mentioned earlier, if it is desired that $WSi_x$ be deposited, a higher flow rate of $SiH_4$ is needed. This ensures that the deposited film is $WSi_x$ and not W. During film deposition, $WSi_2$ forms with excess Si collecting in the grain boundaries. Since an excess of silicon exists in such films, the chemical formula to denote them is given as $WSi_x$. However, unless $x \geq 2.0$, $WSi_x$ films formed with this reaction are prone to cracking and peeling away from the underlying poly during later high-temperature steps. That is, the excess Si in the as-deposited silicide film avoids the possibility of consuming some of the underlying poly material, which could cause the above problems. Thus, in practice a $SiH_4/WF_6$ flow-ratio of more than 10:1 is used to ensure a deposition

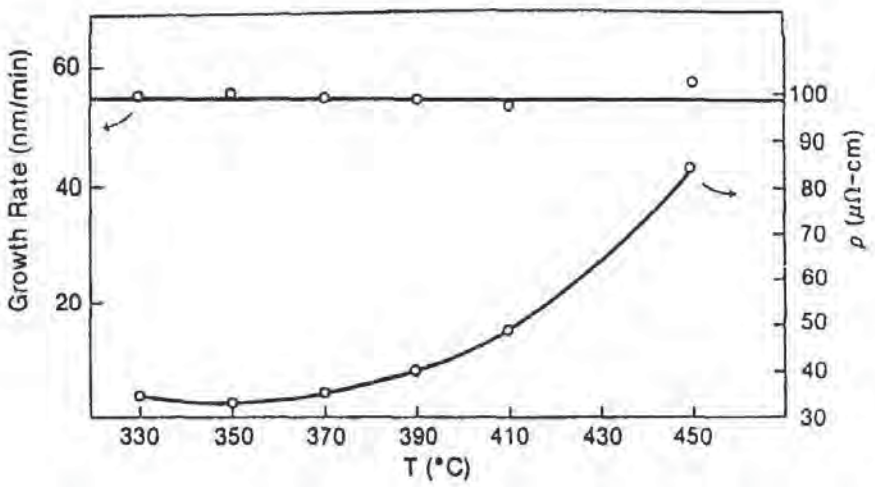



**Fig. 13-3** Resistivity of CVD-$WSi_x$ versus x, where x varies from 2.2 to 2.6.

with x ~2.2–2.6. As-deposited $WSi_x$-films exhibit high resistivity (~500-$\mu\Omega$-cm), but this drops to about 50-$\mu\Omega$-cm after an RTP anneal at 900°C. Note that the resistivity of CVD-$WSi_x$ also depends on its stoichiometry as is shown in Fig. 13-3, where x varies from 2.2 to 2.6. The resistivity of the film increases as it becomes progressively richer in Si.

It was found that CVD-$WSi_x$ films formed by the reaction of Eq. 13-1 had high fluorine-concentrations (~$10^{20}/cm^3$). This creates problems when such films are used with gate oxides thinner than about 20-nm, because low-field oxide breakdown and threshold-voltage shifts are observed.[4] Such problems arise because some of the fluorine gets incorporated into the gate oxide during anneal of the polycide stack. As a result, an alternative chemistry using dichlorosilane (DCS - $SiH_2Cl_2$) instead of $SiH_4$ was developed for CVD of $WSi_x$:[4]

$$WF_6(gas) + 3.5SiH_2Cl_2(gas) \rightarrow WSi_2(sol) + 1.5SiCl_4(gas) + 6HF(gas) + HCl(gas) \quad \textbf{(13-2)}$$

This DCS process is also carried at low pressure (LPCVD) in the temperature range of 570–600°C. The fluorine content is much lower than in films formed with the silane reduction of Eq. 13-1, and the chlorine content is also quite low. The resistivity is comparable in both types of films. However, the step-coverage is better in DCS films, and its deposition-rate is three to five times faster. Peeling and cracking is also less severe in $WSi_x$ films formed from the reaction of $WF_6$ and DCS. DCS has thus largely replaced silane for CVD-$WSi_x$ processes.

Exposure of the $WSi_x$ films to high-temperature $O_2$-ambients (900°C) will cause a dense adherent layer of $SiO_2$ to grow on the silicide while the polycide structure underneath remains intact. The oxide forms by reaction of oxygen with the excess-silicon dissolved in the silicide film, and continues after this silicon is

consumed by diffusion of additional silicon from the polysilicon to the surface of the $WSi_X$. The oxide-formation reaction occurs at the oxide/silicide interface.

CVD-$WSi_X$ films are deposited in cold-wall reactors. Batch processes were initially used but single-wafer systems have also been introduced. A process that integrates polysilicon deposition, polysilicon doping, and $WSi_X$ deposition in a single cluster tool is available.[6] Advantages of such a configuration include cleaner processing and less cleanroom floor space.

## 13.3 SALICIDE (SELF-ALIGNED-SILICIDE) STRUCTURES

In the salicide process, the entire surface area of the source and drain regions of a MOSFET are covered with a conductor film (i.e., a metal silicide). This approach becomes even more attractive if such a film can be formed using a self-aligned process, in that a self-aligned process does not entail any additional masking steps. The process sequence performed to fabricate such a self-aligned conducting silicide on the source/drain-regions (and to simultaneously form a conducting silicide on gate-polysilicon structure [polycide]) is as follows:

**1**. After the source- and drain-regions have been implanted and annealed to form the source/drain-junctions (Fig. 13-4a), CVD-$SiO_2$ sidewall-spacers on the polysilicon gate are fabricated (i.e., by CVD deposition of an oxide, followed by an etchback step which leaves spacers on the poly sidewalls - Fig. 13-4b).

**2**. The metal film (of either Ti or Co) used to form the silicide is deposited (typically by sputtering).

**3**. Next, the wafer is heated, which causes the silicide reaction to occur wherever the metal is in contact with the silicon. Everywhere else, the metal remains unreacted (Fig. 13-4c). Note that RTP is extensively used in this first silicidation step ($1^{st}$ RTP-step). At the conclusion of this step, the silicide is typically in a high-resistance phase. A second RTP-step (carried out after the unreacted metal is removed in step **4**), is performed at a higher temperature than the $1^{st}$ RTP step. This transforms the silicide into the desired low-resistivity phase.

**4**. The unreacted-metal is selectively removed through the use of a wet-etchant that does not attack the silicide, the silicon-substrate, or the $SiO_2$ (Fig. 13-4d). As a result, each exposed source and drain region, and each exposed polysilicon gate region is now completely covered by a silicide film, but no metal or silicide film remains elsewhere.

5. As noted in step 3, after the unreacted metal is stripped, a 2nd RTP step is performed at a higher temperature than the 1st RTP step. This higher temperature process transforms the high-resistance phase of the silicide remaining on the source/drain- and poly-regions to the desired low-resistance silicide-phase.

6. A dielectric layer is deposited onto the silicide, and contact holes are opened in it down to the silicide layer.

7. Metal is deposited into the contact holes to make contact with the silicide (Fig. 13-4e).

The goals of the above salicide process are as follows:

1. The silicide should meet the required low sheet-resistivity specifications called for by the particular technology node being fabricated.

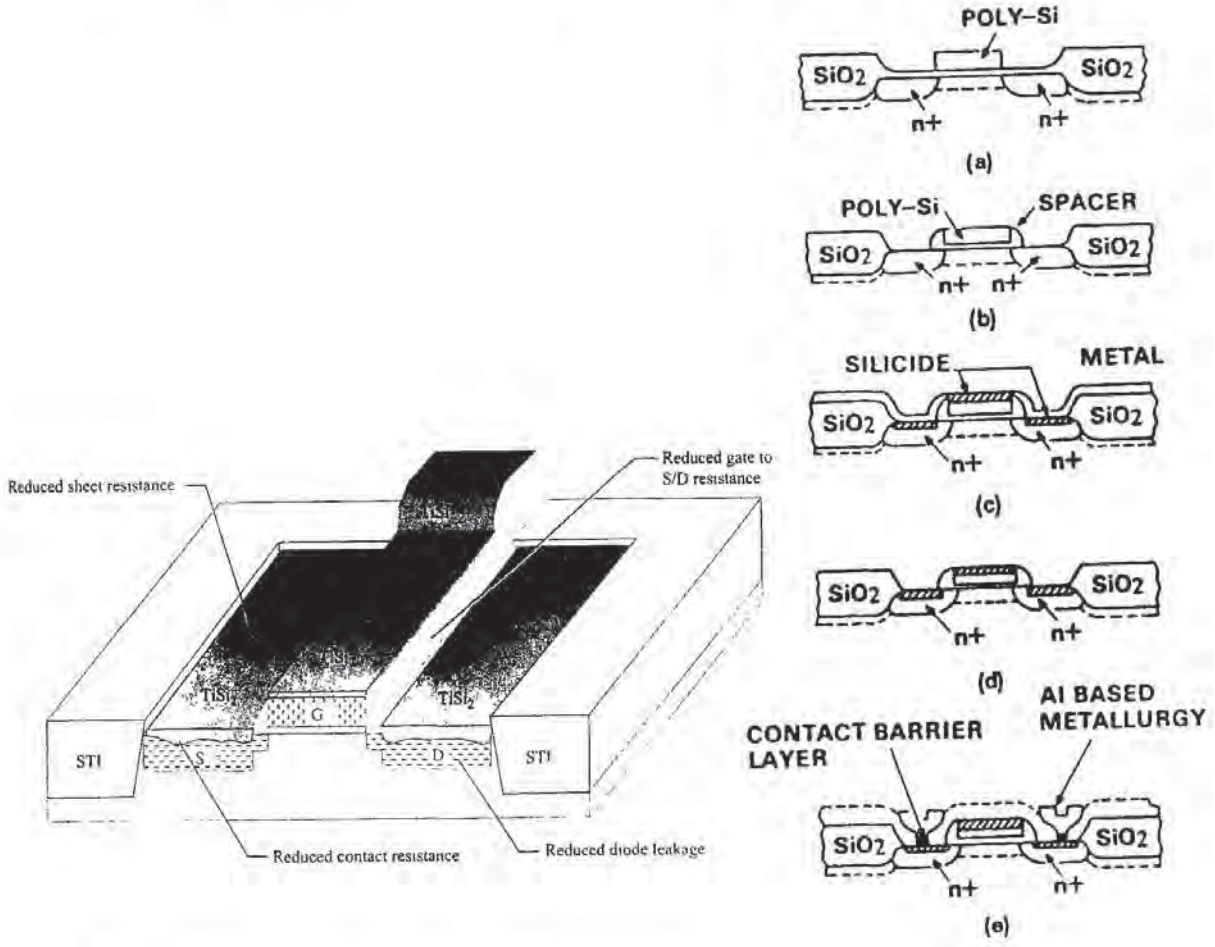



**Fig. 13-4** Salicide processing steps and final structure.

2. Low junction leakage should exist on all source/drain geometries.

3. No bridging should occur across isolation regions or across gate sidewall spacers.

4. A low specific-contact-resistivity should exist for the silicide/Si contact.

The above contact-formation-method offers several advantages over the conventional contact structure:

1. The value of the parasitic-series-resistance component due to the bulk sheet-resistance of the source- and drain-regions $R_{sh}$ becomes negligibly small because the $\rho_{sh}$ value of the silicide is typically 1-2-Ω/sq, whereas that of the diffused junction region alone is 40-120 Ω/sq.

2. The contact area of the silicide and the Si is much larger than that of the metal-Si area in a conventional contact in a device of the same total area. Thus, the value of the contact resistance, $R_{co}$ arising from the contact between the silicide and the Si substrate is smaller than in a conventional contact. That is, $R_{co}$ is equal to the specific contact resistivity $\rho_c$ divided by the contact area.

3. The contact hole is now used to make contact between the metal and the silicide, but the value of $\rho_c$ exhibited by silicide-metal interfaces is typically $\leq 1 \times 10^{-9}$-Ω-cm$^2$. This is two orders of magnitude lower than nominal metal-to-silicon specific contact resistivities, ~$1 \times 10^{-7}$-Ω-cm$^2$, obtained at or near solid-solubility silicon-doping-limits. Hence, the contribution to $R_{co}$ due to the metal-silicide contact resistance at the contact hole is also usually negligibly small.

As it became apparent that a self-aligned contact-structure would be beneficial for reducing the values of the parasitic-series-resistances of the conventional contact structure, some other processing trends also allowed a new group of materials to be considered as contact metals to Si. These materials were $TiSi_2$ and the group-VIII silicides $CoSi_2$ and NiSi. When it also became necessary to fabricate shallower junctions, processing technology had to be performed at lower temperatures to avoid the formation of deep junctions. When the maximum temperature after the silicide formation was thus restricted to less than 900°C, the door was opened to the use of these silicide films (since all of them are unstable at temperatures higher than 900°C).

There are several important reasons why $TiSi_2$ and the group-VIII metal silicides are especially attractive as candidates for self-aligned ohmic contacts and local interconnects to silicon. These are as follows:

**1.** Such silicides exhibit lower resistivities than the refractory-metal silicides of W, Ta, and Mo. That is, the respective resistivities are 13-20-$\mu\Omega$-cm for $TiSi_2$ (with C-54-phase crystal structure), 16-18-$\mu\Omega$-cm for $CoSi_2$, and 15-18-$\mu\Omega$-cm for NiSi.

**2.** All of the group-VIII metals react with Si at 600°C or less, and with the exception of Ti, there is no reaction of the metal with the masking oxide during the silicide formation. Even with Ti, the reaction is minimal below 600°C, and the reaction products are soluble in the Ti-removal etch.[8]

**3.** When the silicides are formed at such low temperatures, the metal atoms diffuse into the Si, and react with it. As the metal atoms diffuse into the $SiO_2$, no Si is available for reaction, and hence, the metal over device areas covered with $SiO_2$ remains unreacted. This also prevents lateral formation of the silicide over the narrow oxide-spacer-region during the silicide formation-step.*

**4.** For the cases of Co and Ni (but not for Ti), dopant atoms in the silicon are "snowplowed" into the silicon substrate as the $CoSi_2$ and NiSi are formed. Such snowplowing maintains a high doping concentration at the silicide-Si interface, which helps produce low-resistance contacts.

**5.** In all cases, after the silicide has been formed, the unreacted metal can be etched off from the oxide surface using chemical etchants that do not attack the oxide, the silicide, or the substrate silicon. This last characteristic allows self-aligned contacts to be formed through the use of these silicides.

The focus of most development for the salicide process has revolved around $TiSi_2$ and $CoSi_2$. Both of these silicides exhibit lower resistivities than silicides formed with the refractory metals, W, Ta, and Mo and they can withstand process temperatures in excess of 800°C (such temperatures are frequently used to reflow the doped glasses deposited over the salicided source/drain-regions).

The salicide process exhibits a limitation related to the fact that the gate and the source/drain-silicides are formed at the same time. On the gate, it is desirable for the silicide to have the lowest possible sheet resistance (so that the gate electrode will also possess a low interconnect-resistance). To achieve this, a thick

* There is an exception to this statement. During $TiSi_2$ formation at temperatures above 600°C, Si diffuses into the Ti. The diffused Si then reacts to form $TiSi_2$ over the oxide-spacer regions, resulting in shorting of the gate and source/drain regions. Procedures for avoiding this effect are discussed later.

Silicide-layer on the gate is needed. Over the source/drain-regions, however, the silicide can be only of limited thickness, in order to prevent excess consumption of the substrate-silicon by silicide formation. Thus, a thicker silicide, though favorable at the gate level, is detrimental to contact formation (and vice-versa). Going forward, this tradeoff between sheet-resistance and silicon-consumption represents a key scaling issue, in that thicker salicide films are needed to achieve low sheet-resistance-values, while shallower junction-depths require thinner silicide films.

Another problem is that it is difficult to metallize shallow junctions (≤ 200-nm-thick), because of the junction leakage that is observed when the silicidation process is used. The implant damage (caused during the source/drain junction-formation) may not be sufficiently annealed-out by the lower-temperature silicide-formation step. This damaged region may remain within the depletion-region of the shallow junction, giving rise to higher leakage. In addition, the RIE-process used to open the contact-holes in the doped oxide (deposited after silicide formation), may damage the silicide.[9] Techniques to reduce or eliminate this last set of problems are discussed in later sections. A final problem of salicide technology is that it degrades the effectiveness of some electrostatic-discharge-protection devices used with CMOS circuits.

## 13.4 THE EFFECTS OF SALICIDES ON MOSFET PERFORMANCE

A self-aligned-silicide on the source/drain-region lowers the contact-resistance because the metal-Si contact area is increased from the area of the contact window(s) to the entire source/drain area. For example, consider the contact resistance $R_{co}$ ($\Omega$) of an *n*MOSFET with a source/drain area of 1-$\mu$m x 1-$\mu$m and a contact window the size of 0.4-$\mu$m, with and without a silicide on the source/drain-region. Assume that the specific-contact-resistivity is identical for a silicide/$n^+$-Si contact and a metal/$n^+$-Si contact (e.g., in this case $1\times10^{-7}$-$\Omega$-cm$^2$), and $0.2\times10^{-8}$-$\Omega$-cm$^2$ for the metal/silicide contact. (Note that the value of the specific-contact-resistivity of metal-to-silicide is much smaller than the specific contact resistivity of silicide or metal to $n^+$-Si.)

Thus, if no silicide is formed on the source/drain region, then only the contact resistance $R_{co}$ of the metal/$n^+$-Si contact exists, and its value is calculated from:

$$R_{co}([\text{metal}/n^+\text{-Si}]\ \text{contact}) = (\text{specific contact resistivity})/\text{contact area}$$
$$= (1\times10^{-7}\text{-}\Omega\text{-cm}^2)/(0.16\times10^{-8}\ \text{cm}^2) = 65\ \Omega \qquad \textbf{(13.3)}$$

On the other hand, if a silicide has been formed that covers the entire source/drain-regions, the contact-resistance will be the sum of the two contacts in series. That is, it will be sum of the contact resistance $R_{co}$([metal/silicide] contact), and $R_{co}$([silicide/$n^+$-Si] contact). The value of $R_{co}$([metal/silicide] contact) is found from:

$$R_{co}([\text{metal/silicide}]\ \text{contact}) = (1\text{x}10^{-7}\text{-}\Omega\text{-cm}^2)/(1\text{x}10^{-8}\ \text{cm}^2) = 10\ \Omega \quad \textbf{(13.4)}$$

The value of $R_{co}$([silicide/$n^+$-Si] contact), is found from:

$$R_{co}([\text{silicide/}n^+\text{-Si}]\ \text{contact}) = (0.2\text{x}10^{-8}\text{-}\Omega\text{-cm}^2)/(0.16\text{x}10^{-8}\ \text{cm}^2) = 1.2\ \Omega \quad \textbf{(13.5)}$$

The sum of these two contact resistances is:

$$R_{co}(\text{total}) = R_{co}([\text{metal/silicide}]\ \text{contact}) + R_{co}([\text{silicide/}n^+\text{Si}]\ \text{contact})$$

$$= 11.2\ \Omega \quad \textbf{(13.6)}$$

Therefore, even for a nominal-sized MOSFET (i.e., in the case above, having a width of 1.0-$\mu$m), the self-aligned-silicide significantly decreases the contact resistance.

However, even a contact-resistance of 65-$\Omega$ is only a small fraction of the channel-resistance. Reducing this quantity even further does not significantly increase device performance of the nominally-sized MOSFETs of a circuit. On the other hand, for wider MOSFETs (used, for example, to supply more current to subsequent stages of the circuit, or to provide drive-current for off-circuit loads), the benefit can much more significant. Assume that a transistor is 10-$\mu$m wide (instead of being only 1.0-$\mu$m wide, as in the above example), and that this MOSFET is capable of delivering 4-mA of current when $V_{GS}$ is 3.3 V. Let us see how the use of a self-aligned-silicide on the source- and drain-regions impacts the performance of such a device (compared to one that does not). We will do this by considering the effect of the silicide on each side of the device separately, namely first on the drain-side, and then on the source-side.

The effect of the contact-resistance $R_{co}$ on the *drain-side* of such a wider MOSFET device is to reduce the drain voltage $V_{DS}$ by the amount of $R_{co}I_D$. If the MOSFET is operating in the saturation region (as is the case in most digital circuits, when the MOSFET is ON), then the reduction in $V_{DS}$ has little effect. If the transistor is operating in the linear region (as in analog circuits), then the output current $I_D$ is reduced from:

$$I_D = V_{DS}/R_{channel} \quad \textbf{(13.7)}$$

to:

$$I_D = (V_{DS} - R_{co}I_D)/R_{channel} \quad \textbf{(13.8)}$$

or:

$$I_D = V_{DS}/(R_{co} + R_{channel}) \quad \textbf{(13.9)}$$

Therefore, the effective resistance in the linear channel has been increased by the amount $R_{co}$. For this MOSFET, the channel resistance in the linear region, $R_{channel}$, can be calculated as $V_{Dsat}/I_D$, or approximately 250-Ω. If no silicide exists, and only a single metal/$n^+$-Si, 0.4-$\mu$m contact is made to this drain region, the contribution of $R_{co}$, is about 25%.

The effect on the performance of the device due to the $R_{co}$ on the *source-side* of the device is more significant. That is, a resistance on the source-side of the MOSFET not only increases the drain-voltage by the same amount as on the drain-side, but it also raises the potential at the source by $R_{co}I_D$. The effective gate voltage $V_{GS}$ is then reduced by the same amount:

$$V'_{GS} = V_{GS} - R_{co}I_D \quad \textbf{(13.10)}$$

A reduction in effective gate voltage is equivalent to a reduction in transconductance:

$$g_m = \Delta I_D/\Delta V_{GS} \quad \textbf{(13.11)}$$

$$g'_m = g_m/(1 + R_{co}g_m) \quad \textbf{(13.12)}$$

A MOSFET with a $g_m$ = 300-$\mu$S/$\mu$m and an $R_{co}$ = 65-Ω, has a $g'_m = 0.98\text{-}g_m$, or only a 2% loss in drive current for a 1-$\mu$m-wide MOSFET. But for a 10-$\mu$m-wide MOSFET, a 20% loss for the drive current results. Therefore, for wide MOSFETs, either multiple contacts, or a self-aligned silicide is needed to reduce the parasitic resistance.

The contact-resistance will degrade the device performance more severely on MOSFETs with smaller minimum dimensions. At the 0.25-$\mu$m node, the contact window will be about 0.3-$\mu$m wide, making $R_{co}$ ~140-Ω, compared to the 65-Ωin the example in which the contact window was 0.4-$\mu$m wide. At the 0.18-$\mu$m technology node, $R_{co}$ will be about 250-Ωwithout a self-aligned silicide. Even for a 1-$\mu$m-wide MOSFET, the $g'_m$ will be reduced to only about $0.9g_m$. If a self-aligned silicide is used, $R_{co}$ drops to about 18-Ω in the 0.25-$\mu$m node, and to about 25-Ω in the 0.18-$\mu$m node.

## 13.5 SALICIDES FORMED WITH TITANIUM SILICIDE ($TiSi_2$)

Salicides formed with titanium silicide $TiSi_2$ are currently being used extensively in the production of 0.25-$\mu$m CMOS devices. $TiSi_2$ is attractive for this salicide application because it exhibits low resistivity, and because it can reduce native-

oxide layers (making it the only known refractory metal that can reliably form a silicide on both polycrystalline and single-crystal silicon through a thermal reaction). Furthermore, devices fabricated with titanium silicide on the gate electrode are more resistant to high-field-induced hot-electron degradation than are conventional poly-Si gate devices.[10] It is conjectured that the $TiSi_2$ is an effective getter for the hydrogen atoms introduced during the hydrogen anneal. Less hydrogen is thus incorporated into the gate oxide, and this improves the hot-electron reliability.

Some aspects of the $TiSi_2$ process are not as favorable, including:

1. The reactivity of Ti with $SiO_2$ can cause unwanted reactions with the oxide spacers during the silicide-formation process (a problem referred to as *bridging*).
2. $TiSi_2$ is less thermally stable than $WSi_2$ or $MoSi_2$ (i.e., it agglomerates at T > 900°C).
3. Because Ti films have a high propensity to oxidize, the silicide must be formed in ambients that are free of oxygen.
4. For line-widths smaller than about 0.25-$\mu$m it becomes difficult to form low-sheet-resistivity $TiSi_2$ films (and this phenomenon is termed the *$TiSi_2$ narrow-line effect*).
5. Silicon is consumed by the silicide reaction (i.e., about 30-nm of Si is consumed by 10-nm of Ti to form 35-nm of $TiSi_2$). Silicon consumption during silicide-formation becomes a major issue in the process of forming very shallow source/drain-junctions.
6. During silicide formation, dopants in the junctions redistribute themselves. This represents an especially severe problem with $TiSi_2$.

Each of these issues will be discussed in the upcoming sections.

### 13.5.1 Reaction Kinetics of $TiSi_2$ Formation (and the Problem of Bridging)

In practice, a two-step formation-process is typically used to form $TiSi_2$.[11] The first step (1st RTP step) is carried out at a temperature is between 620-680°C in nitrogen for 15-60 sec. This short RTP step transforms the Ti in contact with Si (or polysilicon) into the metastable, high-resistivity C-49 phase-$TiSi_2$:

$$Ti + 2Si \dashrightarrow TiSi_2 \ (C\text{-}49) \qquad \textbf{(13.13)}$$

Following this step, selective etching is used to remove the unreacted Ti (e.g., in a room-temperature mixture of DI $H_2O$, 30% $H_2O_2$, and $NH_4OH$ [5:1:1]). Then a second RTP-step at a temperature of 800°C in nitrogen for 30-60 sec is used to

lower the $TiSi_2$ sheet-resistance by transforming the metastable C-49-phase into the low-energy, low-resistivity orthogonal C-54-$TiSi_2$ phase:[12]

$$TiSi_2 \text{ (C-49)} \dashrightarrow TiSi_2 \text{ (C-54)} \qquad \textbf{(13.14)}$$

The resistivity of C-49-phase-$TiSi_2$ is about 3-4 times higher than that for C-54-phase-$TiSi_2$ (60-70-$\mu\Omega$-cm vs 13-20-$\mu\Omega$-cm). Phase-transformation occurs at 700-750°C, but it will be seen why the process is not carried out in a single step.

The reason why a two-step process is used to form C-54-phase-$TiSi_2$ involves the problem termed *bridging*. That is, oxide-spacers at the edges of the poly-silicon are formed to separate the silicided gate and source/drain-regions. But these spacers are typically only 200-300-nm wide. As a result, any lateral-formation of silicide can easily bridge the separation and cause the gate to be come shorted to the source/drain. Figure 13-5 illustrates the formation of $TiSi_2$ along the sidewall of the gate, resulting in a short between the gate and the source/drain. Such lateral-formation of $TiSi_2$ along the gate-sidewall oxide-spacer occurs because silicon is the moving species of this reaction. Thus, it diffuses into the Ti-regions that cover the spacer-oxide, where it subsequently reacts with the Ti. However, when annealing is performed at a lower temperature (and for short times), the Si does not diffuse as far, and this inhibits the lateral $TiSi_2$ reaction along the spacer. However, at least 600°C is needed to cause Ti to react with Si, even to form the C-49 phase. At this temperature, the Si-diffusion will already produce bridging - even if RTP is used.

It has also been reported that the surface of the $SiO_2$ sidewall spacers is covered with $CF_x$ polymers after the dry-etch step used to from the spacers (carried out in a $CHF_3/C_2F_6$ plasma). The presence of these polymers enhances the lateral Ti-silicide growth over oxide-spacers and can result in bridging. Therefore a cleaning procedure (not based on a fluorine-based chemistry) should be used remove these polymers. A suggested post-spacer-etch cleaning-process in this report was an $O_2$ plasma + 2% diluted-HF in $H_2O$ + APM (ammonia-hydroxide/hydrogen-peroxide mixture). This process was able to remove the fluorocarbon polymers while consuming less than 2.8-nm of the oxide-spacers.[13]

As a result, in addition to using the low-temperature RTP process, the 1st RTP step is also carried out in an $N_2$ ambient. This causes the Ti to absorb a significant amount of $N_2$ (e.g., more than 20atom%) before extensive Si diffusion occurs. The $N_2$ is absorbed preferentially at the Ti grain-boundaries, which "stuffs" the grain-boundary diffusion paths. This further reduces the diffusivity of Si in the Ti, and essentially suppresses the lateral silicide-reaction. Annealing in pure $N_2$ or pure forming gas (95% $N_2$ + 5% $H_2$), allows C-49-phase-$TiSi_2$ form-

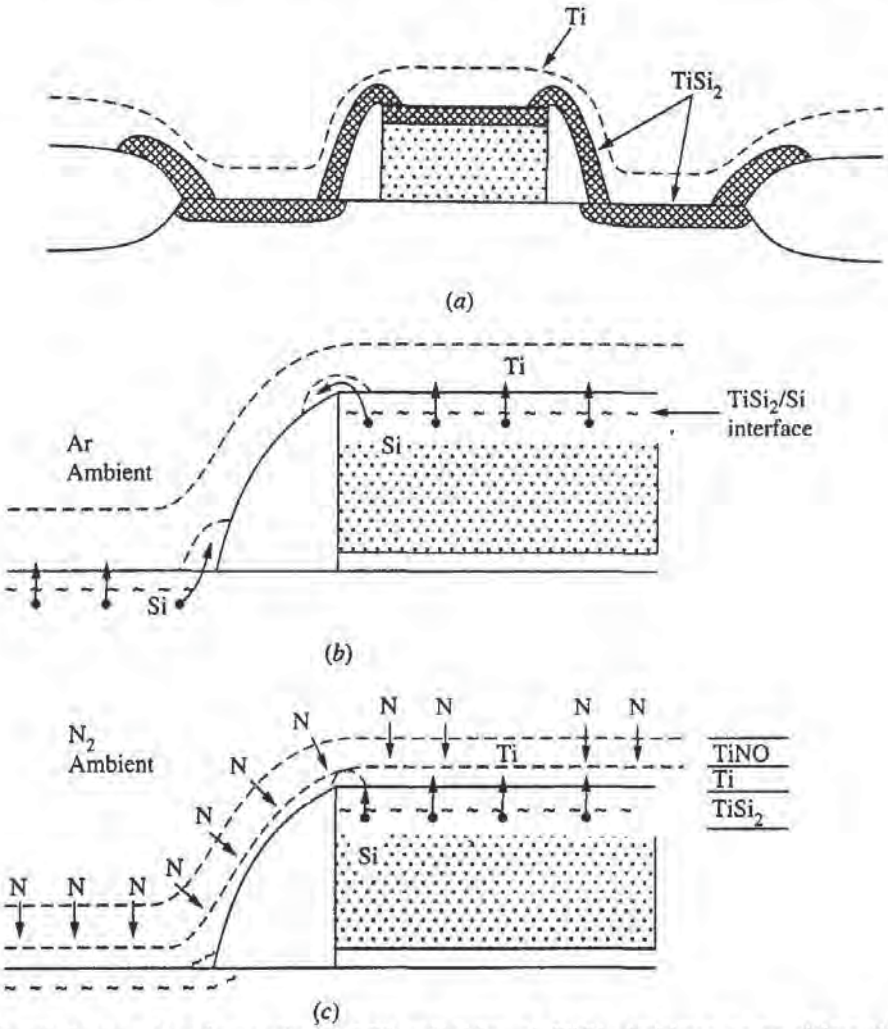



**Fig. 13-5** "Bridging" during the formation of Ti salicide: a) $TiSi_2$ bridges over sidewall spacer and a short circuit is formed between the source-and-gate and the drain-and-gate; b) During $TiSi_2$ formation, Si diffuses into the Ti to form $TiSi_2$. Under favorable conditions, Ti can diffuse a long distance and form $TiSi_2$ on the sidewall spacer; c) In a nitrogen ambient, nitrogen diffuses into the Ti and blocks the diffusion of Si, reducing or eliminating bridging. TiN and/or Ti-O-N are formed on top of the $TiSi_2$.

ation without bridging (provided the temperature does not exceed ~680°C). It is also important to control the $N_2$ ambient so that it contains less than 5 ppma of oxygen or water to avoid unwanted oxidation of the Ti film.[15] The absorbed $N_2$ also reacts with the Ti to form TiN at the Ti surface.[14,15]

Another reason for performing the 1st RTP step at the temperatures below 700°C is that at higher-temperatures, Ti and the spacer $SiO_2$ will react rapidly enough to form significant amounts of titanium oxides. If any residues of this

reaction remain on the oxide spacer, they can degrade device performance by compromising the oxide integrity or by producing bridging.

### 13.5.2 The $TiSi_2$ Narrow-Line Effect

As the polysilicon lines onto which the $TiSi_2$ is formed are made narrower, the sheet resistivity of the $TiSi_2$ is found to increase above the 13-20-$\mu\Omega$-cm exhibited by C-54 phase-$TiSi_2$ (Fig. 13-6). As a result, the resistance of the polycide-coated gate-poly lines no longer meet the resistance specifications for the corresponding technology nodes. The factor that causes this increase in $TiSi_2$ sheet resistivity on narrow polysilicon-lines is that the transformation of the C-49-phase into the C-54-phase is incomplete. It was observed that after the 1st RTP-step, small-grained (0.2-$\mu$m) C-49-$TiSi_2$ was formed on both wide (> 10-$\mu$m) and narrow (0.4-$\mu$m) polysilicon lines. After a 725°C anneal, the $TiSi_2$ on the wide polysilicon-lines transformed into large-grained (1 to 10-$\mu$m) C-54 phase. However, on narrow-lines only smaller-grained (< 1-$\mu$m) C-54-phase-$TiSi_2$ was observed, and only on a few lines. At higher temperatures (775°C anneal), C-54-phase-$TiSi_2$ was formed on both wide- and narrow-lines. At 800°C, $TiSi_2$ started to agglomerate, and the line resistance rapidly degraded. The incomplete transformation of C-49 to C-54 on narrow lines causes the line to have some regions along the line consisting of C-49 and others of C-54 (Fig. 13-7), resulting in an overall increase (and variability) in the $TiSi_2$ sheet-resistance.

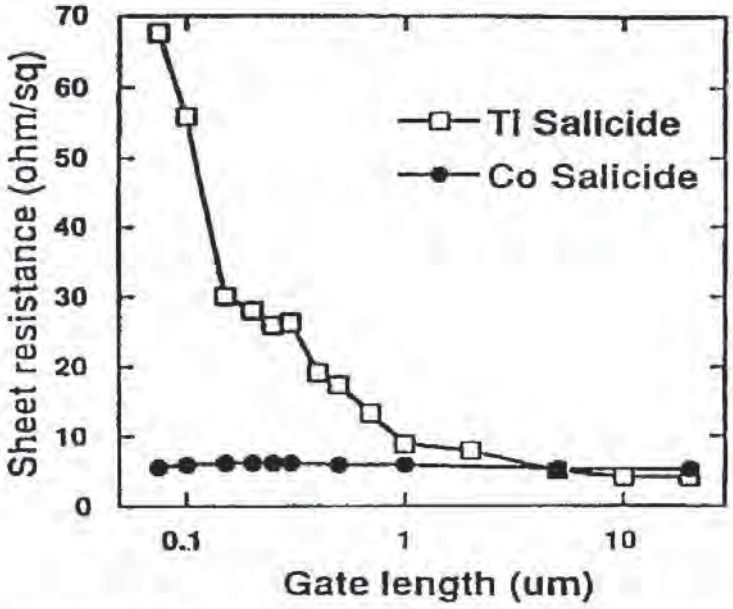



**Fig. 13-6** The $TiSi_2$ narrow-width effect: As the polysilicon lines onto which the $TiSi_2$ is formed are made smaller (i.e., < 1 $\mu$m), the sheet resistivity of the $TiSi_2$ increases.

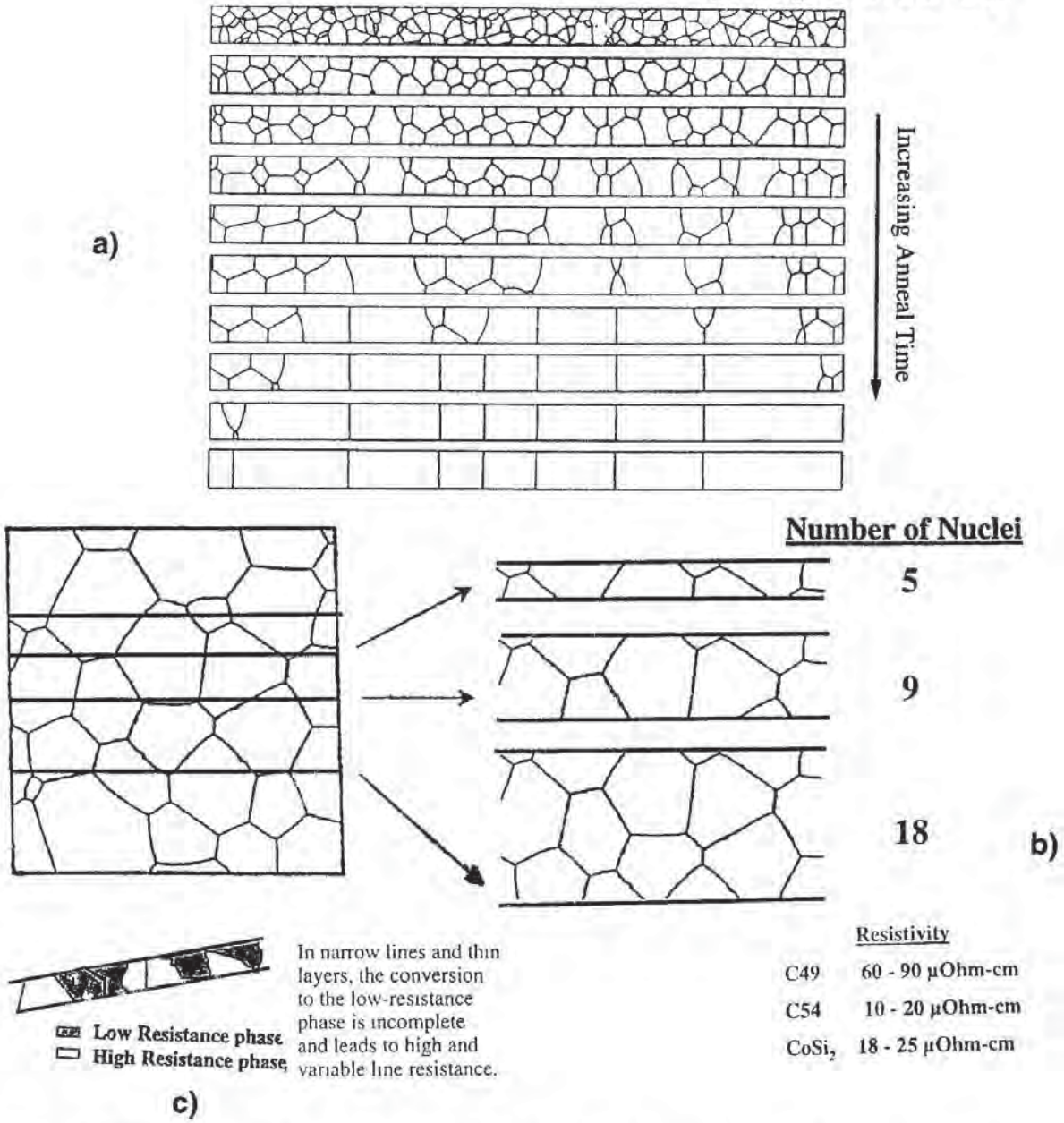



**Fig. 13-7** a) Simulation of grain-growth within a line with increasing annealing time, starting with small grains; b) When the as-deposited grains are already large (relative to the line width) few nucleation sites for the C-54 phase exist. In fact, the narrower the line, the fewer the number of nuclei; c) Thus, the conversion to C-54-phase from C-49-phase is incomplete, and the final $TiSi_2$ sheet-resistance is a combination of low and high resistance phases, and is therefore higher (and variable).

The reason why there is difficulty in completely transforming C-49-phase into C-54-phase is based on the fact that on narrow polysilicon-lines the small grains of the C-49-phase are comparable to the width of the polysilicon lines.[16] As a

result, there is a low-density of grain boundaries - in particular triple-grain boundaries - to function as nucleation sites for C-54-phase-$TiSi_2$ grain growth. (It has been shown by that the nucleation of C-54-phase-$TiSi_2$ preferentially occurs at triple-grain junctions [or grain-boundaries].)[17] In addition, on narrow-lines, the grain growth is limited by the linewidth, and only a fraction of the line is covered by the C-54-phase grown from the few nuclei available. On the other hand, on wide-lines, there are an adequate number of such nucleation sites per unit length of the line to allow C-54 grains to nucleate and grow to a size that transforms the entire $TiSi_2$ along the line to the C-54-phase (see Fig. 13-7). One way to solve the problem is to heat the $TiSi_2$ lines to > 750°C, to accelerate the nucleation of the C-54-phase. Unfortunately, thin films of $TiSi_2$ agglomerate at such temperatures, rendering this solution unusable. The linewidth and C-49 grain-size-dependence of the C-49-to-C-54 transformation-kinetics are in excellent agreement with a nucleation-density-model based on these assumptions.[18]

As shown in Fig. 13-8, as the polysilicon linewidth and $TiSi_2$ thickness decreases, it takes higher temperatures to transform the C-49-phase entirely into C-54-phase. Unfortunately, as the thickness of the film is decreased, the temperature at which it begins to agglomerate also decreases. As a result, the breadth of the process window (here expressed as the allowed variation in the time at which the 2nd RTP step is performed at the appropriate temperature) is also reduced. Without modifying the process conditions in some way, at some point it becomes no longer possible to produce $TiSi_2$ films with the sheet-resistance values required to meet the process specifications.

### 13.5.3 Proposed Process Modifications to Mitigate the $TiSi_2$ Narrow-Line Effect

Several techniques have been pursued as solutions to the narrow-line effect. One method is to replace $TiSi_2$ with $CoSi_2$, as $CoSi_2$ does not exhibit this effect. (Salicide formation with $CoSi_2$ is covered in Sect. 13.6.) However, to allow $TiSi_2$ to continue to be used on polysilicon lines below 0.25-$\mu$m, it has been proposed that the grain-size of the C-49-phase in the $TiSi_2$ be decreased. This will increase the density of C-54 nucleation sites in the film, allowing the C-49-phase to be transformed into C-54-phase even on narrower lines (Fig. 13-7a). The method that has been studied to decrease the grain sizes of the $TiSi_2$ C-49-phase is to perform a *pre-amorphization implant* (or *PAI*). That is, a shallow implant is made into the Si-substrate and poly to amorphize the top 30-nm of the silicon before the Ti is deposited (using either Ge, Si, or Xe as the implant species, Fig. 13-9).

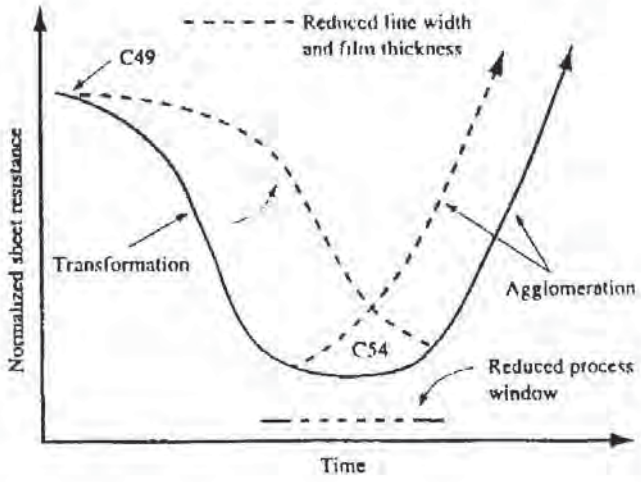



**Fig. 13-8** For a thinner polysilicon line, the time required to transform the C-49-phase to the C-54-phase increases (i.e., at some chosen, constant temperature), but the time it takes for the $TiSi_2$ film on the line to undergo agglomeration decreases as the line width is decreased. Hence, the process-window for forming C-54-phase on thinner lines eventually closes.

This approach has been shown to induce a smaller C-49 grain size (~0.07-$\mu$m) compared to the 0.2-$\mu$m grain size found in the process without the PAI. This enables transformation of silicide films to the low-sheet-resistance-phase on structures down to 0.1-$\mu$m (Fig. 13-10).[19] It should be noted that such PAI steps can also cause increased junction leakage, due to damage by the implant of the Si substrate. Thus, the integration of PAI into a fabrication sequence involves careful design of the process conditions.[20]

An alternative approach to PAI which may allow the continued use of $TiSi_2$ on narrow polysilicon-lines is the implantation of refractory metal impurities into the Si surface before depositing the Ti, close to the Ti/Si interface.[21] Reports of using Mo implants for this purpose have been published.[22] This work indicates that the reaction between Ti and Mo-doped-polysilicon bypasses the C-49-phase, growing C-54-phase directly at low temperatures. On the contrary, however, the reaction between Ti and the Mo-doped single-crystal (100) silicon proceeds by the conventional sequence of C-49 growth first. To achieve a single-step growth of C-54 phase $TiSi_2$ on single-crystal-Si, it was found necessary to combine the Mo-implant into the Si-substrate with a PAI. Such a one-step Mo + PAI achieves low-sheet resistance on poly lines as small as 0.6-$\mu$m wide. Use of a single silicide-anneal step needed for the Mo + PAI process resulted in higher

concentrations of B at the silicide *p*MOS interface, and nearly an order of magnitude improvement in the *p*MOS silicide-to-source/drain contact-resistance.

### 13.5.4 Junction Dopant Redistribution and Silicon Consumption During $TiSi_2$ Formation

Note also that $TiSi_2$ in contact with Si was originally thought to be capable of being subjected to temperatures up to ~900°C. ($TiSi_2$ agglomerates at higher temperatures.) It was observed, however, that if the temperature exceeds 800°C, the value of $\rho_c$ to $p^+$ Si becomes too large.[23] However, the $\rho_c$ of $TiSi_2$-to-$n^+$ Si can be less than $10^{-6}$-Ω-cm$^2$ for anneal temperatures of 800°C. Even at anneal temperatures of 800°C, the value of $\rho_c$ of $TiSi_2$-to-$p^+$-Si is $10^{-5}$-Ω-cm$^2$, while at temperatures of 900°C, it increases to $10^{-3}$-Ω-cm$^2$.

The cause of this problem is the rapid diffusion of the boron from the PMOS source/drain regions into the overlying silicide. Since silicon is the moving species in the course of the silicide reaction, during the formation of silicide, a large number of vacancies are generated in the Si substrate near the Si surface. This enhances the diffusion of dopants in the junction. This enhanced

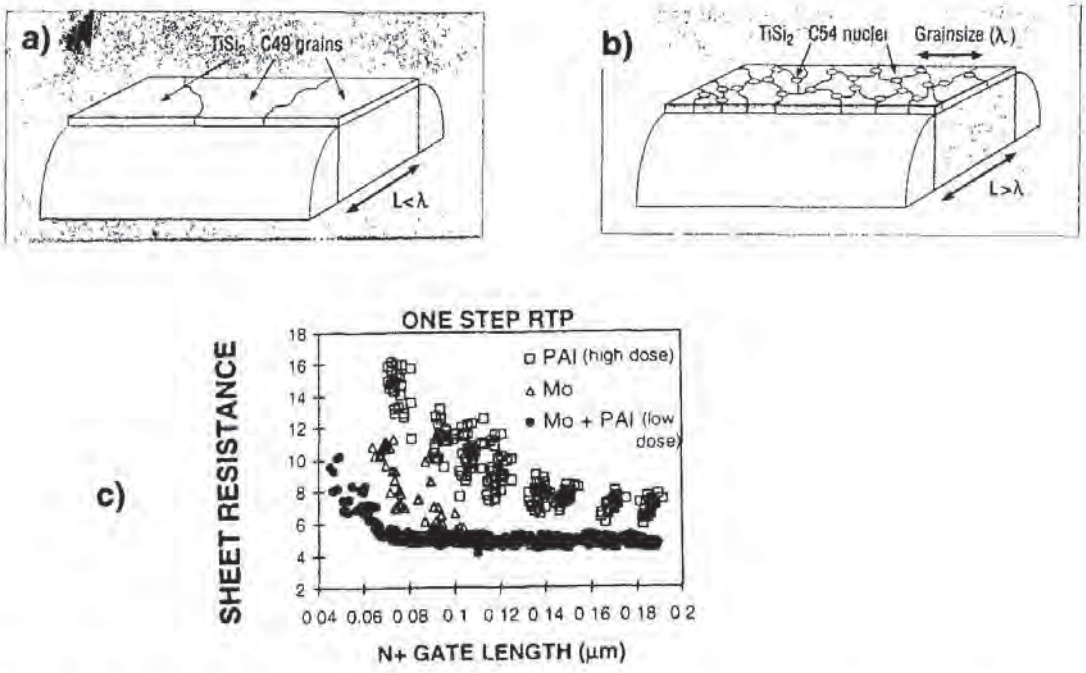



**Fig. 13-9** The $TiSi_2$ C-49 (high resistivity) to C-54 (low resistivity) transformation-rate depends on the density of nucleation sites: a) Without a pre-amorphization implant (PAI) there are few nucleation sites on narrow polysilicon-lines; b) PAI reduces the C-49 grain size, increasing the number of triple-grain-boundaries and C-54 sites; c) By using PAI + MO the sheet–resistance of $TiSi_2$ lines remains low for narrower lines than when only PAI is used.

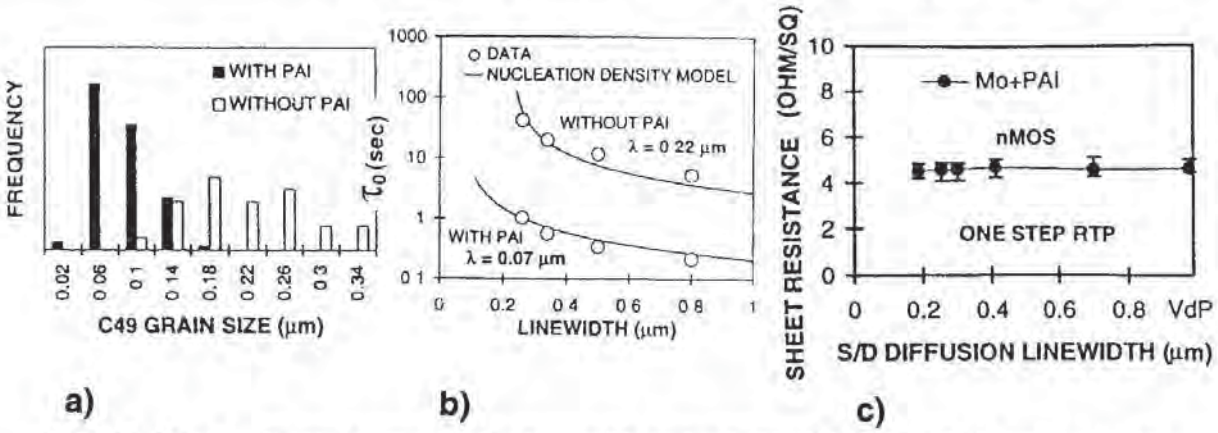



**Fig. 13-10** a) C-49 grain size distribution with and without PAI. The C-49-grain size is significantly reduced by PAI, making it possible to achieve low sheet-resistance on 0.1-$\mu$m gates. b) Results and predictions of the nucleation-density-model for $n^+$ gates with- and without-PAI. The difference in C-49 grain size ($\lambda$) accounts for the large increase in transformation times. Transformation times $\tau_0$ increase dramatically as the linewidth approaches $\lambda$, making the process without PAI incompatible with linewidths below 0.2-$\mu$m: c) Mo + PAI allows low-sheet resistance (4-$\Omega$-sq) to be obtained for 0.2-$\mu$m-wide lines.

dopant diffusion has three deleterious effects. First, some of the dopants in the silicon diffuse deeper into the silicon, creating deeper junctions than would occur under normal diffusion. Second, the silicon containing the highest doping-concentration is consumed by the silicide. Third, the enhanced-diffusion causes more dopants from the silicon to diffuse into the silicide, further depleting the junction of dopants at the silicide/junction interface (and consequently, causing higher contact-resistance between the junction and silicide layer). Temperatures in excess of 800°C must thus be avoided following $TiSi_2$-formation in CMOS technologies. This can represent a problem if doped-glass-reflow is used in a CMOS process.[24] Reports have been published, however, that describe the use of nitrogen implants into the Si surface prior to the Ti-deposition to suppress the diffusion of B and As during the silicide formation.

Note, that one might think that a solution to the dopant-depletion problem would be to implant dopants into the *silicide* after it was formed, and then drive-out the dopants from the silicide to form the junction. Unfortunately, this solution is not effective for $TiSi_2$ because the dopants implanted into the $TiSi_2$ to produce the source/drain regions (e.g., B and As) form stable compounds such as TiB and TiAs.[25] Thus, the B and As species are tightly bound to the Ti and do not easily diffuse out from the silicide. This problem also contributes to the dopant depletion in the Si in contact with $TiSi_2$. That is, the B and As that diffuse into the silicide, do not back-diffuse into the Si, because once inside the silicide, the

dopants bind tightly with the Ti. An extensive study of the effects of $TiSi_2$-formation on dopant redistribution is presented in Refs. 26 and 27.

Another serious issue involving the use of silicides on shallow source/drain-regions is the junction-leakage caused by the silicide. To form the silicide, a substantial portion of the silicon from the junction is consumed. For example, to form 50-nm of $TiSi_2$ requires a 20-nm-thick layer of Ti and the consumption of approximately 50-nm of silicon from the junction region. For a shallow-junction of about 150-nm deep (0.15-$\mu$m), this represents one-third of the junction-depth (assuming the silicide is uniformly produced). However, the silicide film tends to be not totally uniform in depth, and even a minor non-uniformity (say 20-nm), means that in some places 70-nm out of the 150-nm-deep junction are consumed - leaving a very shallow effective junction. As the diffused junction-depth beneath the silicide will continue to decrease with technology scaling, the silicide must also decrease in thickness so that the amount of silicon consumed is not greater than the junction-depth. In fact, it is typically desirable to leave about 50-nm between the bottom of the silicide and the junction, or else leakage current through the junction can become too large.

## 13.6 SALICIDES FORMED WITH COBALT SILICIDE ($CoSi_2$)

Although $TiSi_2$ has been the most widely implemented material for the salicide process, $CoSi_2$ has also received significant attention as an alternative candidate material. In fact, many companies are switching to Co at the 0.25/0.18-$\mu$m nodes. The switch to $CoSi_2$ relieves some of the constraints that exist for many established $TiSi_2$ processes used in production, widening a process window with good margins. On the other hand, it shifts the concerns from issues of lateral scaling (i.e., the narrow-line effect that afflicts $TiSi_2$), to issues related to the vertical scaling of junctions (such as diode leakage in junctions with $CoSi_2$).

### 13.6.1 Advantages and Drawbacks of Cobalt Salicides

$CoSi_2$ exhibits the following attractive characteristics for application in the salicide process:

1. It offers the benefits of low resistivity (16-18-$\mu\Omega$-cm) and high temperature stability. For example, 300-nm-thick $CoSi_2$ films exhibit a $\rho_{sh}$ of 0.5-$\Omega$/sq and remain stable when in contact with Si up to 900°C.[28,29] This demonstrates the compatibility of $CoSi_2$ with a 900°C glass-reflow step.

2. Unlike a conventional $TiSi_2$ process, nucleation does not limit the formation or transformation of Co-silicide phases, even for polysilicon linewidths smaller than 0.1-$\mu$m (see Fig. 13-9). This is due to the fact that

$CoSi_2$ has grain sizes ten times smaller than those of $TiSi_2$ (i.e., ~0.05-$\mu$m versus ~0.5-$\mu$m).

**3.** $CoSi_2$ is normally formed with a two-step RTP process:

$$Co + Si \dashrightarrow CoSi \quad (450°C) \qquad \textbf{(13.15)}$$

$$CoSi + Si \dashrightarrow CoSi_2 \quad (700°C) \qquad \textbf{(13.16)}$$

$CoSi_2$ can also be formed by RTP directly in a single annealing step. A wet-etch is then applied directly to the $CoSi_2$. Note that while Co is the species that moves during several stages of the Co-Si reaction, Si is still the main moving species during the formation of CoSi. Thus, bridging across the sidewall-spacer or isolation can still occur. The use of a Ti or TiN capping layer effectively reduces lateral growth and eliminates bridging for $CoSi_2$. In contrast with Ti, Co does not react with oxide or nitride films, alleviating some of the issues found for Ti salicide processing. An RTP system optimized for the formation of $CoSi_2$ films is described in Ref. 30.

**4.** Under optimized process conditions, the interface of the silicide and silicon can be smooth, and low-resistivity contacts to the silicon can be formed (as low as 15-$\mu\Omega$-cm$^2$ to both $n^+$ and $p^+$Si). The conditions that give rise to such smooth $CoSi_2$/Si interfaces will be described later.

**5.** $CoSi_2$ processes (unlike those for conventional $TiSi_2$) generally do not exhibit the strong dopant interaction and redistribution issues that can lead to interfacial dopant-depletion and higher silicide-to-source/drain contact-resistance. Instead, low-resistance-contacts to shallow As and B junctions can be successfully fabricated, and the dopant profiles in such junctions appear to be essentially unchanged after the silicide has been formed.

**6.** A selective chemical etch exists that makes it possible to remove the Co without affecting the $CoSi_2$ (e.g., with either a single-etch step using a hot $H_2SO_4$:$H_2O_2$ solution,[31] or with a two-step wet-etch process, consisting of an $NH_4OH$:$H_2O_2$:$H_2O$-etch followed by an HCl:$H_2O_2$:$H_2O$-etch.[32])

7. $CoSi_2$ is far less susceptible to removal by plasma-etching than $TiSi_2$. Hence, there is negligible silicide loss during the overetch time that may be needed when dry-etching the contact-openings through the glass-overlayer.

**8.** No competing reaction to the $CoSi_2$ formation reaction occurs, whereas when $TiSi_2$ is formed in nitrogen, TiN is simultaneously formed. In addition, lower shear forces (important to device integrity) are present in $CoSi_2$ than in $TiSi_2$ of the same thickness.[33]

**9.** It may be possible to form shallow junctions by implanting dopants into $CoSi_2$ after it has been formed, and then drive-out the dopants to form the shallow-junction. Unlike for the case of $TiSi_2$, good junctions have been reported by using this technique. This process is described in more detail in Sect. 13.9.

The use of $CoSi_2$, however, also has some drawbacks:

**1.** Cobalt silicide is sensitive to the surface condition of the Si regions prior to the Co deposition. That is, the presence of a native oxide or other surface contaminants can impact the silicide-formation-process and the resultant $CoSi_2$ film-characteristics. For example, both the $CoSi_2$ sheet-resistance and the leakage of the *pn*-junctions existing beneath the $CoSi_2$ salicide-film are influenced by previous processing steps that alter Si surface conditions, such as the spacer-etch, the screen-oxide-etch, and the subsequent surface-clean-and-preparation steps.

Unlike Ti, Co cannot reduce native oxides on Si, and therefore the formation of $CoSi_2$ is more sensitive than Ti to residual contaminants or to a native oxide on the Si-surface. However, several methods have been devised to reduce the sensitivity of $CoSi_2$ to surface conditions. In one of them, advantage is taken of the reactivity of Ti (and its ability to reduce native oxides), by incorporating Ti at the deposition stage. That is, by applying a capping layer of Ti onto the Co film (i.e., deposited *in situ* prior to the silicide formation), the sensitivity of the $CoSi_2$ formation process to interfacial contaminants such as oxygen is significantly reduced.[34] This allows the formation of $CoSi_2$ through native oxide layers. A similar effect was reported for a Co-Ti (10-20 atom.%) alloy deposition.[35]

Further discussion of the effect of surface-contaminants present on the Si surface prior to Co-deposition and silicide-formation on the leakage-current of $CoSi_2$ salicided *pn*-junctions is given in Sect. 13.7.3.

**2.** The silicon consumption during the $CoSi_2$ formation is about 25% higher than for the formation of $TiSi_2$.

**3.** Gate-oxide integrity is a concern when Co salicides are used because Co contamination will degrade GOI, and both the diffusivity and solubility of cobalt are high (significantly higher than those of Ti) at high temperatures. Degradation of film and interface morphology at high temperatures, as well as process conditions that induce rough, non-uniform silicide-Si interfaces, can effect GOI. These issues could limit the compatibility of Co salicide

processes with aggressive scaling of polysilicon gate-thickness. GOI degradation can be avoided by a careful control of the thermal history.

**4.** $CoSi_2$ is sensitive to oxidation and contamination from the ambient. Measures must be taken to prevent this from occurring, and these are discussed in Sect. 13.7.4.

### 13.6.2 Details of the Cobalt Salicide Fabrication Process

The details of the process sequence for forming the $CoSi_2$ polycide and the self-aligned contacts to Si are as follows:

**1.** Since the formation of $CoSi_2$ is sensitive to the Si surface conditions, procedures must used to ensure that a clean Si surface exists prior to the Co deposition. One technique used to clean the silicon and polysilicon surfaces before the Co sputtering step is to dip the wafers in a dilute (100:1) solution of $H_2O$:HF just prior to loading them into the sputtering chamber. This is often followed by a short *in situ* rf-backsputtering procedure to remove ~10-nm of the silicon surface.[28,29] A prolonged rf-backsputtering process has the risk of allowing contaminant material sputtered from other parts of the wafer to redeposit onto the Si surfaces that are being cleaned. Nevertheless, the cleaning procedure is important, because cobalt is not able to reduce $SiO_2$ and because the presence of a native oxide can inhibit the silicidation reaction.

**2.** The cobalt is sputter-deposited to the desired thickness. The thickness is selected based on how much of the silicon in the source/drain regions can be consumed, as well as what value of silicide sheet-resistance is needed. For example, 35.9-nm of Si is consumed by 10-nm of Co to form 35.9-nm of $CoSi_2$. To form a $CoSi_2$ layer with $\rho_{sh}$ = 1.5-Ω/sq, a 30-nm-thick layer of Co must be deposited. As noted earlier, the Si consumption during $CoSi_2$ formation is about 25% greater than that during $TiSi_2$ formation, constituting a disadvantage of $CoSi_2$ compared to $TiSi_2$.

**3.** The $CoSi_2$ is formed by reacting the Co with the Si and polysilicon. Although the reaction can be accomplished through either a furnace anneal or RTP, the latter has emerged as the more attractive approach.[36] Either a two-step anneal, or a single-step anneal can be employed, but the former is more commonly employed. An 80-nm-thick Co film will be completely converted to $CoSi_2$ at 700°C for 30 seconds. The gas ambient must be kept free of oxygen to ensure that oxide-free silicide films are formed.

4. The unreacted Co is etched away using the etchant mentioned earlier.

5. Finally, the wafer is covered with a CVD glass layer. Note that when Al is contacted to the $CoSi_2$, the maximum anneal temperature must thereafter be limited to 400°C (since Al will react with $CoSi_2$ at temperatures in excess of 400°C). To allow a higher anneal temperature to be used, a diffusion barrier layer (such as Ti:W), must be used between the $CoSi_2$ and the Al layers.[37]

### 13.6.3 Diode Leakage Issues Involving Cobalt Salicides

It was noted earlier that one of the major difficulties of using $CoSi_2$ is the issue of increased leakage of shallow source/drain-junctions when $CoSi_2$ is the salicide. A primary cause of such leakage is the formation of a rough interface between the $CoSi_2$ and the Si-substrate. Such interfacial roughness can arise if a native oxide or other contaminants are on the Si surface prior to the Co deposition. In that case, the silicidation reaction can be less uniform, proceeding faster in clean spots where the Co can diffuse faster, resulting in a rougher interface. As a result, an effective cleaning of the Si surface before sputtering the Co is mandatory for formation of smooth interfaces (and thus also low-leakage). As mentioned earlier, a dip in dilute-HF just before loading the wafer into the sputtering tool is thus normally performed. In addition, a brief *in situ* sputter-cleaning pre-Co-deposition-step is also used to remove contaminants from the Si surface.

The silicide interface is also rough if the silicidation is carried out at low temperatures. This, however, is the usual way of forming $CoSi_2$ films (i.e., a 1st RTP step is carried out at 450°C). At this temperature the formation of long 'needle-like' protrusions, or 'spikes' of some form of cobalt-silicide in the Si-substrate has been observed (with a length that is possibly great enough to penetrate shallow-junctions).[38] In contrast, high-temperature-processing promotes the formation of a smooth interface. At such higher temperatures the spikes formed at the lower temperatures are found to dissolve, and the $CoSi_2$/Si-interface becomes smoother. Such smoothing occurs because of the enhancement of diffusion kinetics and the increased solubility of Co in Si. Therefore, a very effective way to reduce diode leakage is the use of an optimized higher-temperature RTP, in the range of 800-850°C (Fig. 13-11). Because the growth of the low-resistivity Co-salicide-phase is easily achieved in deep-submicron structures, the second RTP step of $CoSi_2$ formation can be optimized in terms of its effects on diode-leakage, agglomeration, and source/drain-resistance, rather

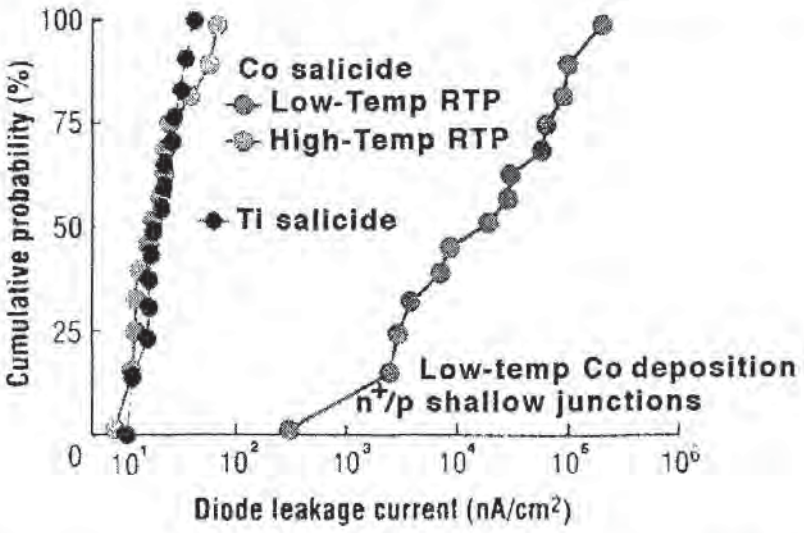



**Fig. 13-11** High-temperature RTP is effective in improving diode-leakage in MOSFETs fabricated with $CoSi_2$ by eliminating high-leakage outliers.

than simply considering the growth of the low-resistivity phase (as must be done when using Ti- instead of Co-salicides). An upper-limit on the temperature that can be used during the second RTP-step is set by silicide agglomeration and the consequent degradation of the source/drain series-resistance. (An alternative technique for the reduction of diode-leakage is high-temperature cobalt deposition [i.e., the Co is sputter-deposited at 450°C].) In this process, the initial stages of the Co/Si reaction take place during the high-temperature cobalt-deposition-process.[39,40] $CoSi_2$-films deposited with such a high-temperature sputter-deposition-process were also reported to exhibit increased thermal stability, as compared to those sputter-deposited at room temperature.[41]

### 13.6.4 Sensitivity of Cobalt Salicide Films to Degradation by Oxygen Ambients: TiN-Capping of Cobalt-Salicide Films

$CoSi_2$ is also sensitive to oxidation and contamination due to the presence of oxygen or other contaminants in the ambient during the silicidation thermal steps. Such oxidation can cause the phase-transformation temperature from Co to $CoSi_2$ to change, and to produce films that exhibit higher sheet-resistance. A very effective approach to control ambient-contamination is the use of a capping layer, consisting of either Ti,[42] or TiN,[43] and deposited *in-situ* immediately following cobalt-deposition (Fig. 13-12). $CoSi_2$ processes with *in-situ* capping-layers have been shown to produce linewidth-independent, low-resistance films well below 0.1-$\mu$m dimensions. As also noted earlier, the use of a Ti or TiN capping layer

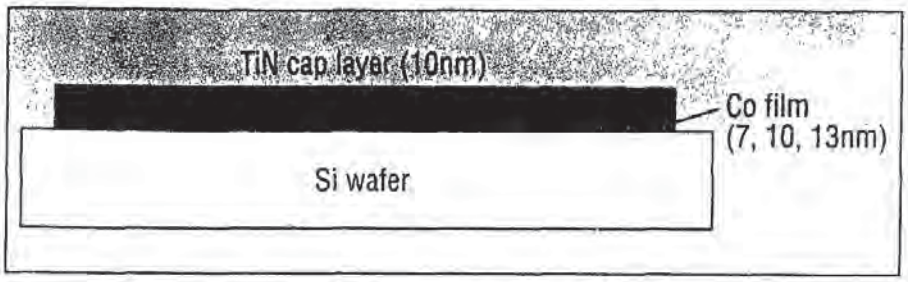



**Fig. 13-12** Schematic cross-section of a $CoSi_2$ film with a TiN capping layer on a Si wafer.

also effectively reduces lateral growth of the $CoSi_2$ and thus eliminates bridging when carrying out the $CoSi_2$ salicide process.

For shallow junctions, the thickness of $CoSi_2$ films must also be scaled, to prevent excessive silicon-consumption of the substrate in which the shallow-junction is formed. The thermal-stability of $CoSi_2$ films decreases with thickness, and this becomes a major concern.[44] Deposition of a Ti capping-layer on the Co was found to increase the thermal stability of $CoSi_2$ when a Co-Ti-phase was formed on the surface.[45] According to another report, the Ti capping-layer also produced superior films than those having a TiN capping-layer. That is, the Ti-capped films formed on 0.15-$\mu$m-wide poly lines showed uniformly-low sheet-resistances and a thermal stability up to 900°C for 30 min. The TiN-capped films on the same lines showed a wide distribution of sheet-resistance values after annealing for 30 min at 800°C. The higher thermal-stability of the Ti-capped $CoSi_2$-layers was attributed to the fact that a large amount of Ti-atoms diffused into the $CoSi_2$ film along the grain boundaries, and that the presence of these Ti atoms slowed the agglomeration process.[46]

## 13.7 EPITAXIAL $CoSi_2$

The reaction mechanism that forms $CoSi_2$ layers on silicon-substrates that has been discussed to this point produces polycrystalline films of $CoSi_2$. However, it is also possible to nucleate and grow $CoSi_2$ epitaxially on (100)-Si. A potential advantage of epitaxial cobalt-silicide films is the achievement of smooth silicide-silicon interfaces that could result in lower diode-leakage-currents on ultra-shallow junctions. Thin epitaxial-silicide layers provide a shallow-junction contacting-technology that may open a path for further scaling of junction-depths.

Techniques have been developed for achieving epitaxial-$CoSi_2$-growth on (100)-Si that are compatible with salicide processing for IC manufacturing (Fig. 13-13).[47] One such approach consists of depositing a thin Ti-layer (~5-nm thick) before deposition of the Co film.[48] A different technique involves the growth of an ultrathin oxide layer (~0.5-nm-thick, which can be grown in a peroxide bath), and produces a similar effect.[49] In both cases, the mechanism appears to be solid-state epitaxy through a transport medium, with the main role of the interlayer being that of a diffusion barrier retarding the Co/Si reaction.

Several issues involved with the Ti-interlayer process (such as void formation at the silicide edges in patterned structures), significantly limits its applicability to salicide processing. The ability to control the characteristics of an ultrathin oxide layer may also limit the application of this process in a production environment. Epitaxial growth was also reported for a process that used a Co-Ti (10-20 atom%) alloy,[50] as well one that used a high-temperature sputter-deposition technique to deposit the Co-film.[51] An article comparing these techniques, and the influence of a $BF_2$ implant on the roughness of the $CoSi_2$/Si interfaces produced by each of these techniques, is listed in Ref. 52.

## 13.8 $CoSi_2$ AS A SORCE/DRAIN DIFFUSION-SOURCE

One technique for fabricating shallow-junctions that involves salicides is to implant the junction dopant after the salicide has been formed. To do this, the dopant would have to be implanted either: **1**) entirely into the silicide; or **2**) through the silicide. It is difficult to form very shallow junctions (< 0.15-$\mu$m) by

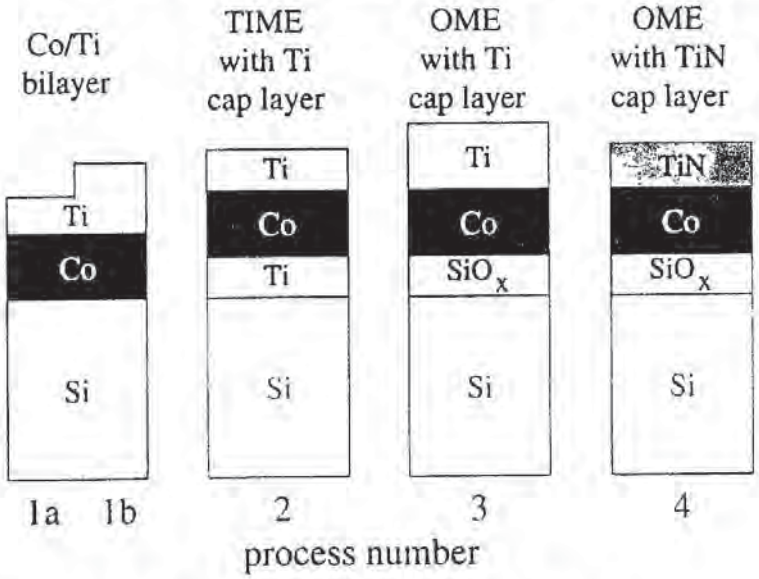



**Fig. 13-13** Different silicide techniques; thickness are not scaled.

implanting through the silicide, since the peak of the implant has to be in the silicide and only the tail portion can be used to form the *pn*-junction in the Si. The junction-depth then becomes a sharp function of the silicide thickness, which is difficult to control precisely.

The second approach (implanting the junction-dopant into the silicide and driving it out) has been shown to successfully form good shallow-junctions with $CoSi_2$ but not with $TiSi_2$. This method has been dubbed the *Silicide as a Diffusion Source* (SADS) process.[53] As noted earlier, dopants such as B and As have high solubilities in Ti and also form stable compounds with Ti (i.e., $TiB_2$ and TiAs). As a result, if these dopants are implanted into Ti, it is very hard to get them to out-diffuse to form a junction.[54]

Forming junctions as shallow as 0.12-$\mu$m by implanting dopants into $CoSi_2$ were first reported in 1987.[55] In the SADS process that involves $CoSi_2$, the conventional S/D-implant-and-activation steps are skipped while a Co-salicide is formed, followed by S/D-implantation into the silicide-layer. A relatively low-temperature anneal is then performed to drive-out the dopants from the silicide film into the underlying silicon to form shallow junctions. Since B and As dopants diffuse rapidly in a $CoSi_2$ film, an ideal junction is expected to follow the silicide/Si interface-contour (i.e., conforming to any existing interface roughness), resulting in low-diode-leakage on the silicided ultra-shallow junctions. Two other benefits of the SADS process are that low source/drain series- resistance is achieved, due to the high dopant-concentrations at the silicide-Si interface, and that SADS is compatible with thin sidewall-spacers.

An example of an $n^+/p$-junction formed with the SADS process is shown in Fig. 13-14.[56] The $p$-doping concentration in the Si was $1x10^{18}$ $cm^{-3}$. It can be seen that a junction-depth of 40-nm beneath the $CoSi_2$/Si interface is obtained. Arsenic was implanted into the $CoSi_2$, followed by a low-temperature RTP-step. (Note that in Fig. 13-14, a baseline-temperature $T_B$ is invoked (i.e., used in the conventional $CoSi_2$ process). Thus, the low-temperature-range referred to here is [$T_B$ - 250°C] to [$T_B$ - 50°C].) Junction-depths were observed to increase as the RTP-temperature was increased in this low-temperature-range. However, as the RTP-temperatures approached the baseline temperature ([$T_B$ - 50°C] in Fig. 13-14), the SADS junction-depth became shallower (due to As evaporation from the $CoSi_2$ film). The minimum possible junction-depths in SADS process will probably thus be limited by the maximum temperature that can be endured by the thin $CoSi_2$ layers before they become thermally unstable (e.g., before they agglomerate). A capping-layer would aid in preventing evaporation at higher temperatures, allowing the formation of deeper junctions.

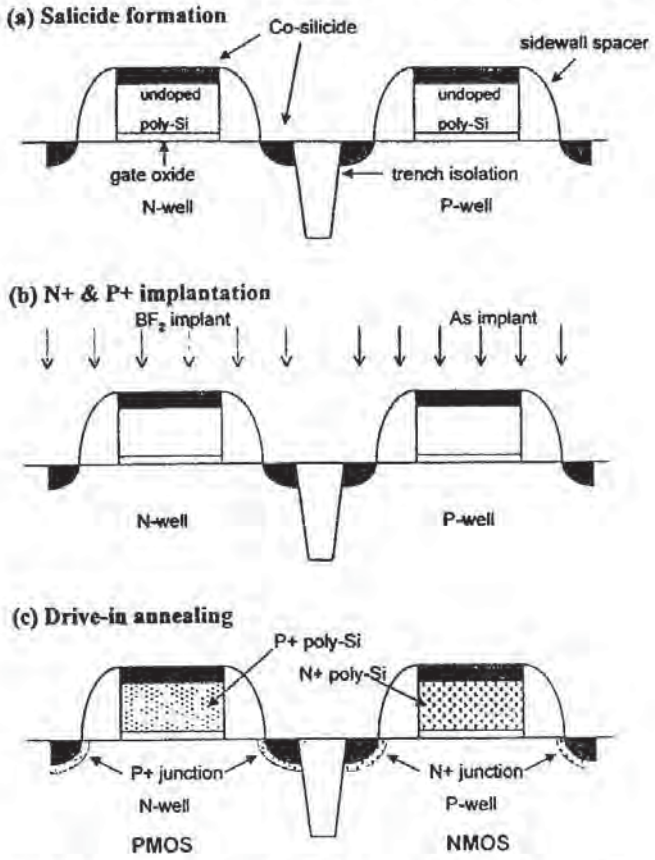



**Fig. 13-14** Schematic of a process flow of 0.18-$\mu$m salicided dual-doped-gate CMOS using Co-silicide as the doping source.

Improvements in short-channel characteristics for thin (35-55-nm) spacers were verified for the SADS process in comparison with the conventional process, as better DIBL-values were observed (see Vol. 3, Chap. 5, *Silicon Processing for the VLSI Era*, for more information on DIBL).[57]

One of the main concerns associated with implementing silicided shallow junctions is diode-leakage. Good leakage characteristics, however, have been demonstrated with $CoSi_2$ as the diffusion source for both $n^+/p$ and $p^+/n$ junctions. However, most of the demonstrations were carried out in devices with low $n$-well and $p$-well concentrations that are not suitable for deep-submicron CMOS applications. For higher background-concentrations the depletion-region of the $n^+/p$-junction extends closer to the $CoSi_2/n^+$-Si interface. If the junction is too shallow, the depletion-region can extend through the whole $n^+$-Si-layer, and the shallow $n^+/p$-junction behaves as a Shannon-contact with high diode-leakage. Applicability of SADS is thus limited by the ability to obtain the appropriate junction-depths needed to achieve low diode-leakage, given the background-concentrations used in specific technologies.

A report that details the use of SADS for a full sub-micron, dual-doped-gate CMOS process was recently published.[58] As shown in Fig. 13-15, $BF_2$ and As were implanted separately into a Ti-capped, 35-nm-thick $CoSi_2$-layer that covered both the polysilicon and source/drain-regions of the MOSFETs. An As implant at 30-40-keV with a dose of $1x10^{16}$ $cm^{-2}$, was used to introduce the *n*-dopants, and a $BF_2$ implant at 30-keV with a dose of $5x10^{15}$ $cm^{-2}$ was used to introduce the *p*-dopants. A single-step anneal at temperatures ranging from 800-1000°C for 30 seconds was used to drive-out the dopants to form the shallow junctions. These shallow-junctions exhibited good leakage-characteristics and low contact-resistance. The process also did not contribute to degradation of the gate-oxide, nor to B-penetration of the gate-oxide from the *p*-doped poly-gates.

## 13.9 NICKEL SILICIDE FILMS FOR SALICIDE APPLICATIONS

Ni silicides have also been investigated for use as salicide candidates, due to their low resistivity (~15-18-$\mu\Omega$-$cm^2$). A single RTP-step at temperatures in the 400-500°C range forms the NiSi film.[59] Like $CoSi_2$ the sheet resistance of NiSi films are not affected by the strong narrow-line-effects which plague $TiSi_2$ salicides. The slight increase in sheet resistance observed with decreasing linewidth has been attributed to thickening of NiSi films at gate edges. A potential advantage of NiSi over $CoSi_2$ is that silicon consumption is only 80% of that for $CoSi_2$.

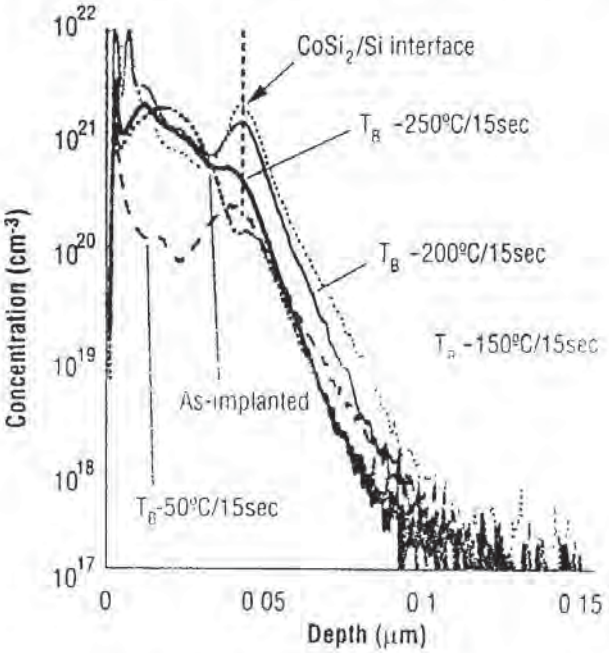



**Fig. 13-15** Shallow junctions obtained using $CoSi_2$ as a diffusion source (SADS) process.

Several challenges, however, exist for implementing NiSi. First, NiSi (like $CoSi_2$) is sensitive to contamination on the Si surface, and this can lead to rough interfaces and higher-diode-leakage on shallow junctions. TiN capping-layers and nitrogen-doping can prevent oxidation from the ambient during silicide formation, and this reduces diode-leakage. High diode-leakage in the source/drain-edges of STI edge-intensive-structures has been reported, correlated to film-thickening and non-uniformity at these edges.[60] Furthermore, transformation into the high-resistivity $NiSi_2$-phase needs to be prevented during later processing steps. That is, NiSi films have poor thermal-stability, with an increase in sheet resistance observed at temperatures as low as 600°C.

## REFERENCES


**1.** S.P. Murarka, *Sillicides for VLSI Applications,* Academic Pr., Orlando, FL 1983.

**2.** M.-A Nicolet and S. Lau, "Silicides," in N.G. Einspruch & G.B. Larabee,eds., *VLSI Microstructure Science*, Vol. 6, Academic Press, N.Y. 1983, p. 329.

**3.** K.C. Saraswat *et al.*, "Properties of Low Pressure CVD Tungsten Silicide for MOS VLSI Interconnection," *IEEE Trans. on Electron Dev.*, **ED-30**, No. 11, November 1983, p. 1497.

**4.** M. Venkatesan and I. Beinglass, "Single-Wafer Deposition of Polycrystalline Silicon," *Solid State Technology*, March 1993, p. 49.

**5.** T. Hara et al., "Composition of Tungsten Silicide Films Deposited by Dichlorosilane Reduction of $WF_6$," *J. Electrochem. Soc.* , **137**, p. 2955 (1990).

**6.** S.G. Telford *et al.*, "CVD $WSi_x$ Films Using Dichlorosilane in a Single Wafer Reactor," *J. Electrochem. Soc.* , **140**, 3689 (1993).

**7.** C. Y. Ting, "Silicide for Contacts and Interconnects," *Tech. Dig. IEDM*, 1984, p. 110.

**8.** R. A. Haken, "Applications of the Self-Aligned $TiSi_2$ Process to VLSI MOS CMOS Technologies," Workshop on Refractory Metals and Silicides for VLSI-IIII, San Juan Batista, CA, May, 1985.

**9.** R. Liu, D. S. Williams, and W. T. Lynch, "Mechanisms for Process-Induced Leakage in Shallow Silicided Junctions," *Tech. Dig. IEDM*, 1986, p. 58.

**10.** S.-T. Chang and K. Y. Chiu, *IEEE Electron Dev. Letts.*, May, 1986, p. 244.

**11.** *ibid. Ref. 7.*

**12.** C. Y. Ting *et al.*, "Interaction Between Ti and $SiO_2$," *J. Electrochem. Soc.* **131**, p. 2934, 1984.

**13.** Y.-B Kim *et al.*, "Characterization of the Post Dry-Etch Cleaning of Silicon Ti-Self-Aligned Silicide Technology," *J. Electrochem. Soc.* April 1999, p. 1549.

**14.** *ibid. Ref. 8,*

**15.** R. M. Vadjikar and R. P. Roberge, *Ext. Abs. Electrochem. Soc. Meeting*, Fall 1988, Abs. No. 467, p. 681, 1988.

**16.** J.B. Lasky *et al.,* "Comparison of Transformation to Low-Resistance Phase and Agglomeration of $TiSi_2$ & $CoSi_2$," *IEEE Trans. Electron Dev.,* 1991, p. 262.

**17.** Z. Ma *J. Appl. Phys.* 77, 4384 (1995).

**18.** J.A. Kittl *et al.,* "Novel One-Step RTP Salicide Process with Low Sheet Resistance 0.06 $\mu$m Gates and High Drive Current," *Tech. Dig. IEDM,* 1997, p. 111.

**19.** J.A. Kittl *et al.,* "A Ti-Salicide Process for 0.10 $\mu$m Gate Length CMOS Technology," *Symp. VLSI Tech* 1996, p. 14..

**20.** J. Kittl *et al.,* "Salicides for 0.10 $\mu$m Gate Lengths: A Comparative Study of One-Step RTP Ti with Mo Doping,, Ti with Pre-Amorphization, and Co Processes," *Symp. VLSI Tech.* 1997, p. 103.

**21.** R.W. Mann *et al., Appl. Phys. Letts.* 67, p. 975 3725, 1996.

**22.** J. Kittl, *Appl. Phys. Lett.,* 73, p. 900, 1998.

**23.** D. B. Scott *et al.,* "Titanium Disilicide Contact Resistivity and Its Impact on 1-$\mu$m CMOS Circuit Performance," *IEEE Trans. Electron Dev.,* March 1987, p. 562.

**24.**R. K. Shukla and J. S. Multani, "Thermal Stabillity of Titanium Silicide Thin Films," *Proceedings of the 4th Intl VMIC Conf.,* Santa Clara, CA, 1987, p. 471.

**25.** V. Probst, *J. Appl. Phys.* 70, 693, p. 1991.

**26.**C. M. Osburn *et al.,*"The Effects of Titanium Silicide Formation on Dopant Redistribution," *J. Electrochem. Soc.,* June 1988, p. 1490.

**27.** M.E. Lunnon, J. Chen, J.E. Baker, "Furnace and Rapid Thermal Annealing of $p^+/n$ Junctions in $BF_2^+$-Implanted Silicon," *J. Electrochem. Soc.,* **132**, 2473 (1985).

**28.** L. C. Van de Hove *et al.,* " A Self-Aligned $CoSi_2$ Interconnection and Contact Technology for VLSI Applications," *IEEE Trans. Electron Dev.,* March, 1987, p. 554.

**29.** S. P. Murarka *et al.,* "Self-Aligned Cobalt Disilicide for Gate and Interconnection and Contacts to Shallow Junctions," *IEEE Trans. Electron Dev.,* Oct. 1987, p. 2108.

**30.** W.S. Yoo *et al.,* "Cobalt Silicide Processing in a Susceptor-Based LR-RTP System," *Solid State Technology,* July 1999, p. 126.

**31.** J. Kittl *et al.,* "Salicides and Alternative Technologies for Future ICs: Part I," *Solid State Technology* June 1999, p. 86.

**32.** D.K. Sohn *et al.,* "Effects of Ti-Capping on Formation and Stability of Co Silicide," *J. Electrochem. Soc.,* January 2000, p. 373

**33.** L. Van den Hove et al., *Electrochem. Soc. Abstracts*, Vol. 88-1, Atlanta, GA, May 15- 20, 1988, p. 312.

**34.** Q.F. Wang, *Dig. Symp.VLSI Tech.,* 1995, p. 17.

**35.** Iinuma *et al.,* "Highly Uniform Heteroepitaxy of Cobalt Silicide by Using Co-Ti Alloy for Sub-Quarter Micron Devices," *Symp. VLSI Tech.*, 1998, p. 188.

**36.** A. R. Sitaram and S. P. Murarka, *Ext. Abs. Electrochem. Soc. Meeting,* Fall 1988, Abs. No. 263, p. 376, 1988.

**37.** E.K. Broadbent *et al.,* "Application of Self-Aligned $CoSi_2$ Interconnection in sub-Micron CMOS Transistors," *Proceedings of the 5th IEEE VMIC Conf.*, Santa Clara, CA, 1988, p. 175.

**38.** K.-I. Goto *et al.,* "A New Leakage Mechanism of Co Salicide and Optimized Process Conditions," *IEEE Trans Electron Devices*, January 1999, p. 117.

**39.** Q.Z. Hong, "$CoSi_2$ With Low Diode Leakage and Low Sheet Resistance at 0.065 $\mu$m Gate Length," *Tech. Dig. IEDM* 1997, p. 107.

**40.** K. Inoue *et al.,* "A New Cobalt Salicide Technology for 0.15-$\mu$m CMOS Devices," *IEEE Trans. Electron Dev.,* November 1998, p. 2312.

**41.** H.Y. Huang, "Thermal Stability of $CoSi_2$ Thin Films on (001)Si by High Temperature Sputtering Deposition," *Mtg. ECS Spring 1999*, Abs 516.

**42.** *ibid. Ref.* 34.

**43.** T. Yamazaki *t al.,* "21 psec Switching 0.1 μm-CMOS at Room Temperature Using High Performance Co Salicide Process," *Tech Dig. IEDM,* 1933, p. 906.

**44.** K.Maex, *Mat. Sci. Eng. Rep.,* R11, p. 53 1993.

**45.** R.T. Tung and F. Schrey, *Appl. Phys. Lett.*, 67, 2164 (1995).

**46.** ***ibid. Ref. 32.***

**47.** J. Hohaus *et al.,* "Influence of $BF_2$-Implantation on Roughness of Cobalt Silicide/Silicon Interfaces," *Electrochem. & Solid State Letts.* March 2000, p. 147.

**48.** L.A Dass, *Appl. Phys. Lett.,* 58, p. 1308, 1991.

**49.** R.T. Tung *Appl. Phys. Lett.*, 61, 3461, (1996

**50.** *ibid.* Ref. 35.

**51.** *ibid.* Ref. 40.

**52.** ***ibid.*** Ref. 47.

**53.** J.C. Hu, *SPIE* Vol. 3506, p. 103, 1998.

**54.** C.Y. Lu *et al.,* "Process Limitation and Device Design Tradeoffs of Self-Aligned $TiSi_2$ Junction Formation in Submicrometer CMOS Devices," *IEEE Trans. Electron Dev.* **ED-38**, p. 246 (1991).

**55.** R. Liu *et al.,* "Formation of $n^+/p$ Junctions with $CoSi_2$," in *Proceedings of the 1st International Symposium on ULSI Science and Technology,* S. Broydo and C.M Osburn, eds. Vol. 87-11, The Electrochemical Society, Pennington, NJ 1987, pp. 446.

**56.** J. A. Kittl *et al.,* "Salicidies and Alternative Technologies for Future ICs,": Part 2," *Solid State Technology*, August 1999, p. 55.

**57.** J.C. Hu, *SPIE* Vol. 3506, p. 103, 1998.

**58**. J.-S. Park *et al.*, "Simultaneous Formation of Shallow Junctions and Gate Doping for Dual Gate Structures Using Cobalt Silicide as a Dopant Source," *J. Electrochem. Soc.*, October 1999, p. 3856.

**59**. T. Ohguro *et al.*, "Analysis of Resistance Behavior in Ti- and Ni-Salicided Polysilicon Films," *IEEE Trans. Electron Dev.*, p. 2305, 1994.

**60.** K. Goto, *Proc. 36th Intl. Reliability Phys. Symp., IEEE* p. 363, 1998.

# Chapter 14

# LOW-*k* DIELECTRICS

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

## 14.1 INTRODUCTION TO LOW-*k* DIELECTRICS

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

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

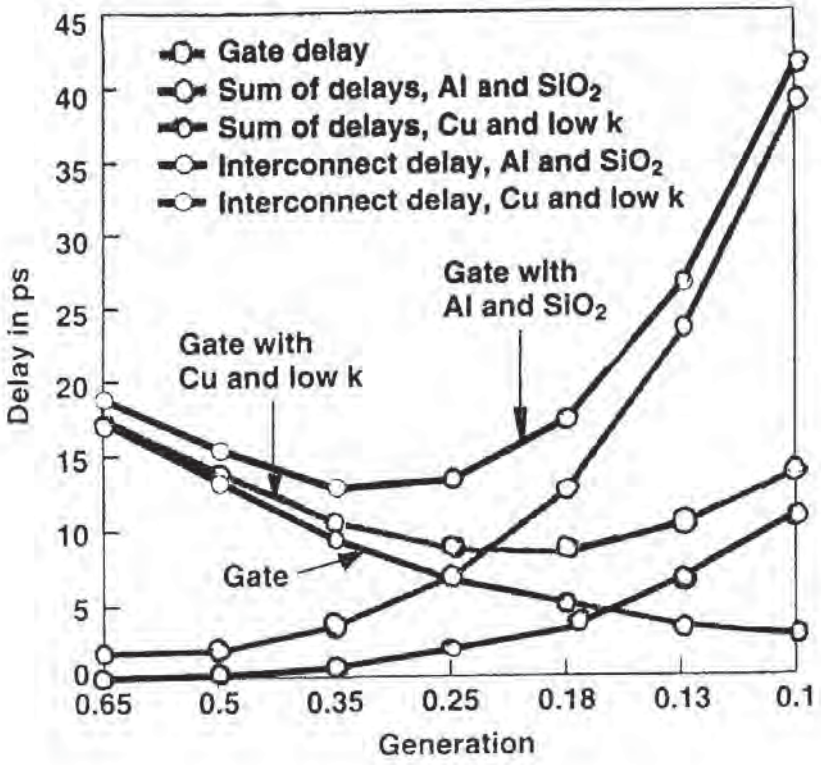



**Fig. 14-1** *RC*-delays caused by interconnects are charted as a function of feature size for both Al/$SiO_2$ and Cu/low-*k* approaches. (© 1997 IEEE)

The conclusions from Fig. 14-1 can be justified by noting that the $SiO_2$ molecule is a pyramid-shaped tetrahedron (see Chap. 3), and its lattice structure thus fits together rather tightly. Although thermal $SiO_2$ does not possess long-range-order, it is nevertheless a relatively dense material (~2.3 g/cm$^3$). Since the goal is to find dielectric films with *k*-values smaller than that of oxide (~4.0), the factors that impact the dielectric constant should first be considered.

In general, no dielectric films as dense as thermal oxide can achieve *k*-values significantly smaller than 3.5. In fact, even CVD-oxides (which are also silica-based films) have about the same *k*-values, even though they are somewhat less dense. While CVD-oxides are more disordered than thermal-$SiO_2$ (which lowers their density), such disorder also makes them contain more silanol (OH) groups (i.e., at sites where Si-O-Si bonds are unable to form). Thus, they are prone to absorb more moisture than thermal oxides, and the absorbed water raises their *k*-value. Furthermore, materials that are less dense than thermal-oxide are also generally mechanically weaker. Nevertheless, to achieve a $k < 2$ in any dielectrics (including silica films), free space must be added, even though the strength of the material diminishes according to some power of the solids content.

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

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

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

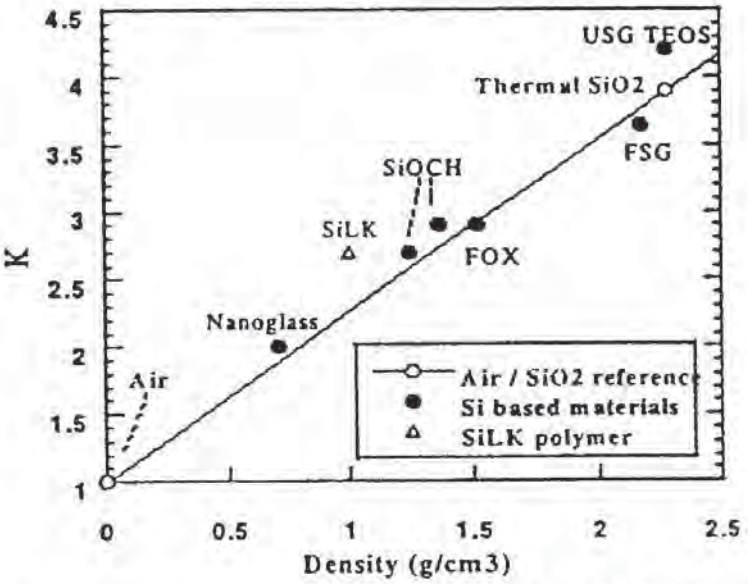



**Fig. 14-2** Value of $k$ is shown vs. density for various dielectrics in ULSI interconnects.

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

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

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

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

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

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

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

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

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

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

of some spin-on formulations to temperature, moisture, and oxygen, materials management could play a critical role in the successful integration of spin-on dielectrics. It should also be mentioned that use of fast heating and cooling rates in RTP reduces the spin-on-deposition processing time. Thus, implementation of rapid thermal processing tools into the spin-on low-$k$ deposition process is being explored.

At present it is too early to tell which technology (spin-on or CVD), and which type(s) of low-$k$ material will prevail. It seems that the whole industry will be characterized by uncertainty regarding this issue over the next few years. As one process integration engineer was said to remark in 1999, "My biggest problem right now is the confusion caused by the variety of material and process options that I face."[4]

### 14.1.2 Silicon-Based versus Carbon-Based Low-$k$ Dielectrics

Low-$k$ dielectrics are also categorized into two other groups: *silicon-based low-k dielectrics* (Si-based materials) and *carbon-based low-k dielectrics* (C-based materials). C-based materials typically have lower $k$-values, but some of their other characteristics may make Si-based materials the better overall choice, depending on the application. For example, Si-based materials usually have higher thermal stability and hardness compared to C-based materials, but Si-based materials tend to be more prone to moisture absorption.

Insofar as manufacturability issues are concerned, the Si-based materials are more compatible with existing device-fabrication flows. Adhesion to silicon, $SiO_2$, and silicon nitride (as well as to Al and Cu metallization) is better with Si-based materials. Also, the Si-based materials are easily etched in fluorine-based plasmas, whereas $O_2$-chemistries are needed for C-based materials. The selectivity of oxygen-based plasmas is not as high over organic photoresist. In addition, the Si-based materials are more compatible with CMP-processing because of their greater hardness.

### 14.1.3 Status of Low-$k$ Dielectrics in 2001

Even by 2001 the news from the industry reflected the state of uncertainty as to which process will prevail. Early in 2000, IBM announced it would use SiLK (a spin-on film from Dow Corning with $k = 2.65$) in its 0.15- and 0.13-$\mu$m Cu/low-$k$ technologies, while Texas Instruments stated it would use Black Diamond (a CVD process from Applied Materials which produces a film with $k = 2.7$) together with Cu, for its high-end DSP-products. Soon after, LSI Logic described how it would push the performance of its aluminum-based interconnects with

**TABLE 14-2 List of Candidate Low-$k$ Materials**

| Low-k Dielectric Candidates | | |
|---|---|---|
| Material | Dielectric constant Low-k materials | Deposition method |
| Polyimides | 3.0-3.6 | Spin-on |
| Spin-on glasses | 2.7-3.1 | Spin-on |
| Fluorinated polyimides | 2.6-2.9 | Spin-on |
| DLC (diamond-like carbon) | 2.8-3.0 | CVD |
| Poly(arylene ethers) | 2.6-2.9 | Spin-on |
| Poly(arylenes) | 2.6-2.8 | Spin-on |
| Cyclotenes | 2.6-2.8 | Spin-on |
| Parylene N | 2.6-2.8 | CVD |
| Poly(norbornenes) | 2.5-2.7 | Spin-on |
| Polyimide-SSQ hybrids | 2.7-3.0 | Spin-on |
| Alkyl-silanes/$N_2O$ | 2.4-2.7 | CVD |
| | Ultralow k | |
| Teflon-AF | 1.9-2.1 | Spin-on |
| Teflon microemulsion | 1.9-2.1 | Spin-on |
| | Porous dielectrics | |
| Polyimide nanofoams | 2.2 | Spin-on |
| Silica aerogels | 1.1-2.2 | Spin-on |
| Silica xerogels | 1.5-2.2 | Spin-on |
| Mesoporous silica | 1.9-2.2 | Spin-on |

low-$k$ FlowFill® CVD films from Trikon products on its G12 0.18-$\mu$m products. SONY said it had integrated FLARE® (a spin-on material from Honeywell) in its advanced devices. Yet, all of these announcements also appear to signal that the time of low-$k$ dielectrics has come. In the discussion here, three ranges of low-$k$ dielectric materials will be described: **1**) first generation ($k$ = 2.8–3.5); **2**) second generation ($k$ = 2.5–2.8); and **3**) ultra low-$k$ ($k$ = < 2.0). Table 14-2 is a list of various low-$k$ dielectric material candidates.

As of mid-2000, the situation was described by Ku (of IBM), as[5] "Though leading candidates have emerged from among the dozen or so being considered a year ago, any choice appears to have downsides. No matter which material is selected, they are all fundamentally flawed in some major characteristic. Therefore, the criterion of low-$k$ material selection for IC manufacturers is not purely technical; the question usually becomes which material shortcomings are we going to choose to integrate around." Table 14-3 shows the main commercial low-$k$ products being used or evaluated on 2001.

## 14.2 DESIRED CHARACTERISTICS OF LOW-*k* DIELECTRIC FILMS

If a dielectric material is to be successfully used in an IC, it must not only exhibit the desired *k*-value, but it must also possess the following characteristics: **1**) It must exhibit adequate material properties (i.e., thermal, electrical, and mechanical); **2**) It must be able to work with the other materials of the interconnect structure; **3**) It must be compatible with the IC processes of cleaning, etching, CMP, and thermal treatments; **4**) It must be available in high purity form, and at low cost; and **5**) It must be able to operate reliably over the life of the product under the specified device operating conditions. Here the general characteristics of groups **1**) and **2**) are considered. Comments are made on how well the various dielectric materials meet these desired goals with respect to each type. The remaining issues are also discussed for each film in particular.

The most important material properties needed by low-*k* dielectric films are listed in order of importance in Table 14-4.[4] First, of course, is the *k*-value (in general, the smaller the better). However, next are the *thermal characteristics.* These include *thermal stability* (i.e., it should be stable up to temperatures of 425°C for at least short periods of time, to make it compatible with the thermal excursions of metal-deposition, assembly, and packaging). Next, it is desirable that these dielectrics have a *thermal conductivity* close to that of $SiO_2$ (12.0 mW/cm°C) for purposes of heat dissipation during device operation. It should be

**TABLE 14-3 Some of the Main Low-*k* Products Being Used in 2001**

| Material | Manufacturer | Deposition | Category | k |
|---|---|---|---|---|
| FOx | Dow Corning | spin-on | inorganic (hydrogen silsesquioxane) | 3.0 |
| XLK | Dow Corning | spin-on | inorganic (porous hydrogen silsesquioxane) | 2.0 |
| SiLK | Dow Chemical | spin-on | organic (aromatic hydrocarbon polymer) | 2.7 |
| 3MS-lowk | Dow Corning | CVD | C-doped oxide (Si-O-C) | 2.7 |
| Aurora | ASM-Japan | CVD | C-doped oxide (Si-O-C) | 2.7 |
| Black Diamond (BDII) | Applied Materials | CVD | C-doped oxide (Si-O-C) | 2.0-2.7 |

noted, however, that the thermal-conductivity of typical low-$k$ polymers is only about 2.4-3.7 mW/cm°C. This low thermal-conductivity is quite undesirable, since high-speed circuits generate significant amounts of heat that must be dissipated. The final thermal property a low-$k$ dielectric should exhibit is a low thermal-expansion coefficient (to prevent cracking during thermal excursions (e.g., < 10 ppm/°C). It should be noted that $k$-value is generally traded off against thermal conductivity. For example, xerogel dielectrics ($k$ = 2.4) possess thermal conductivity that is 10 times smaller than $SiO_2$ (which may therefore give rise to reliability problems from severe increases in circuit temperature).

The low-$k$ films should also exhibit *little tendency to absorb moisture* and should also not exhibit an excessive permeability to moisture. Moisture can adversely impact the other materials of the interconnect structure, and can be outgassed during other processes (which can degrade the process ambient). The films should also have good *adhesion* properties. That is, these films must adhere well to the materials that are deposited on them, and they must also adhere tightly to the materials on which they themselves are deposited.

As far as *electrical characteristics*, such films should have leakage current and breakdown field behavior (at least 2–3 MeV/cm) similar to $SiO_2$ and low charge-trapping characteristics. They should also exhibit high reliability under normal operating voltages for the design life of the IC in which they will be used.

As far as *mechanical properties* are concerned, the films should exhibit low stress (< $10^9$ Pa) and have the capability of being deposited to a thickness greater than 2-$\mu$m without cracking. From a *chemical perspective* the films must also possess chemical compatibility with various reactants (e.g., they must demonstrate good resistance to such reactants as acids, bases, solvents, and resist strippers). They should also exhibit good etch selectivity with respect to the photomasking material (e.g., resist), and with respect to the substrate (e.g., underlying metal). Note that a variety of plasma treatments, thermal treatments, or other steps may be added to stabilize the film during processing. These become a large part of a chip manufacturers expertise in bringing low-$k$ materials into the production environment.

For non-damascene structures (i.e., in which the metal [Al] patterns are etched and then covered with a dielectric layer), the gap-fill characteristics of the low-$k$ material will also be important (e.g., no voids should occur when filling a space that is 0.35-$\mu$m wide and has an aspect-ratio = 2). However, although the gap-fill characteristics of the low-$k$ films are not a consideration for damascene structures (since the dielectric films are deposited on surfaces that have been planarized by CMP), they will need to be deposited in a single pass at a thickness become a

**TABLE 14-4 List of Desirable Properties of Low-*k* Materials**

| | | |
|---|---|---|
| 1 | Dielectric constant (out-of-plane, 1MHz @ room temperature) | <3.2 |
| 2 | Thermal stability (C) | >425 |
| 3 | Moisture absorption | <2.0 |
| 4 | Adhesion (ILD/,metal, metal/ILD, ILD/ILD) | No peel after 450C cycles |
| 5 | Thermal expansion coefficient (ppm/C) | -10 to +50 |
| 6 | Solvent resistance (acid, bases, solvents, PR strippers) | No weight change after dip |
| 7 | Etch rate (wet or dry) | |
| 8 | Etch selectivity (ILD/metal, metal/ILD, PR/ILD) | |
| 9 | Stress (Pa, tensile or compressive) | <1E9 |
| 10 | Gap fill | No void @0.35μm and AR=2 |
| 11 | Tensile strength (MPa) | <200 |
| | Tensile modulus (GPa) | <1 |
| | Elegation at break (%) | >5% |
| 12 | Thickness uniformity over topography (before and after curing) | |
| 13 | Cost | IBO |
| 14 | Corrosion to Al-Cu | |
| 15 | Environmental, health and safety | Pass |

of about twice that used in dielectric films in non-damascene structures. Thus, they will need to exhibit crack-resistant characteristics for such film thicknesses. In addition, they will need to be compatible with CMP processing. Finally, use of spin-on dielectric films that evolve toxic vapors will be avoided.

From an economic perspective, the costs involved with both the materials needed to deposit the films and the CoO of the equipment required to fabricate the films should be low. In addition, the materials used to form the low-*k* films should be commercially available and have a reasonable shelf life.

## 14.3 GENERAL PROCESS INTEGRATION ISSUES INVOLVING LOW-*k* FILMS

There are several processes that need to be performed on the dielectric films after their deposition, and thus these films must be properly integrated into such processing. First, the films will need to undergo either etching or CMP.

As far as etching is concerned, vias (and perhaps grooves in a damascene technology) will have to be etched in the films. Issues of etch chemistry, etch selectivity to oxides, nitrides, and oxynitrides will need to be addressed. After

etching, the resist-mask must be removed. The resist-removal processing (and any resist-residue-removal steps) should not damage the dielectric layers, nor undercut the low-*k* dielectric layer beneath the hard-mask (if one is being used).

When etching *organic* low-*k* films, $O_2$-based plasmas are generally used. However, since an $O_2$-based plasma will also ash the resist masking-layer while the low-*k*-layer is being etched, an oxide or nitride hard-mask must be used to transfer the pattern to the low-*k* film. The etch-process-sequence thus consists of opening the hard-mask (using the PR to define the pattern), and then proceeding with the low-*k*-dielectric etch. Depending on the type of low-*k* film, a wet-chemistry or proprietary dry-chemistry may be used to subsequently remove the resist. Since use of a pure-$O_2$ plasma will tend to undercut the hard-mask, such undercutting must be suppressed through the use of additives to the $O_2$-plasma.

When etching *inorganic* low-*k* dielectrics such as *organosilicate glasses, OSG* (which are oxide-based materials that also contain hydrogen and carbon), a common requirement is that the etch-process be selective to PR and to a silicon-nitride etch-stop-layer. (Silicon-nitride layers are used as embedded hard-masks in dual-damascene processes - see Chap. 15.) In general, oxide-etch-processes that are selective to nitride do not etch the OSG films due to the presence of the organic component in the film. Therefore, alternate process chemistries that include appropriate gases to etch the organic component in the OSG films and that maintain selectivity to photoresist and silicon nitride must be developed.

With respect to films that must be planarized by CMP, such materials will have to be polishing-process-compatible. As an example of the issues involved with CMP, polishing xerogel films is difficult due to the poor mechanical strength of the film. This necessitates the deposition of conformal oxide liners (both beneath the xerogel layer, and also as a capping layer on top of the xerogel). The issues of CMP involving low-*k* films are also covered in Chap. 8.

Another issue involving the use of CVD-deposited organosilicate glasses (OSG) is the possibility that they will poison the DUV-resists. That is, if the OSG contains any nitrogen, it could enter the DUV-resist and cause a "foot" to be produced in the DUV-resist layer that contacts the nitrogen-containing OSG interface. There are a number of "fixes" for this problem, such as: **1**) the use of a bottom anti-reflective coating (BARC); **2**) the deposition of a spin-on or CVD capping-layer on the OSG layer; or **3**) treating the OSG layer in an nitrogen-containing plasma (to form a thin nitride barrier on the OSG surface). More information on the DUV-resist-poisoning problem is given in Chaps. 6 and 15.

Another issue of integrating a low-*k* film into an interconnect process is the

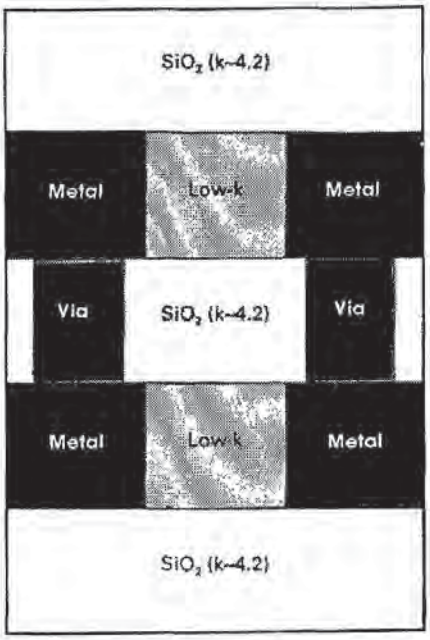



**Fig. 14-3** Using low-$k$ material between metal lines (but not at the via level) improves the mechanical strength and reliability of the stack at the expense of manufacturing complexity.

fact that such films lack the mechanical strength of oxides. One way to circumvent this difficulty is to place the low-$k$ dielectric only *between* the metal lines, but not at the via levels (Fig. 14-3). Beyond strengthening the structure, this "embedded-approach" can reduce moisture-absorption, prevent via poisoning, and allow the fab to keep its conventional dry-etch and cleaning processes. The embedded-structure also makes for a more compatible CMP process, while allowing a company to learn from gradual integration of the low-$k$ material. In addition, roughly 80% of the benefit is gained by using low-$k$ material just between the metal lines. However, this approach requires that two ILD processes must be used. Figure 14-4 is a summary of the process integration issues involved with the use of low-$k$ dielectrics as ILDs.

## 14.4 FIRST GENERATION LOW-$k$ DIELECTRICS ($2.8 < k < 3.5$)

Dielectric materials with $k$-values moderately lower than $SiO_2$ were the first dielectric films to be adopted for this function. Three types of such moderately low-$k$ films had found their way into IC production by 1999. The first two were the *spin-on dielectric* (SOD) films composed of *HSQ* and *MSQ,* with the third being CVD-deposited *fluorinated oxide films,** Their main use was reported to be in the 0.5-$\mu$m to 0.25-$\mu$m CMOS generations.

* These are also referred to as *fluorosilicated glass* (*FSG*), or *silicon oxyfluoride* (*SiOF*).

| Film properties | Manufacturing Integration |
|---|---|
| Dielectric constant:<br>— Bulk: k=2.5-3.0<br>— Effective: k<3.0<br>Thermal stability:<br>— High thermal conductivity<br>— Tg>400°C, stable above 425° for short periods<br>— Low expansion<br>Electrical properties:<br>— High reliability<br>— Leakage current: similar to $SiO_2$<br>— Breakdown field: similar to $SiO_2$<br>— Dissipation factor: <0.01<br>— Low charge trapping<br>Film composition:<br>— Low film stress<br>— >2 μm thick cracking threshold | —Good adhesion to metals (Ta, TaN, TiN, Cu), oxides/ nitrides<br>—CMP compatible<br>—Minimize need for liner/capping films<br>— Etch selectivity to nitrides, oxides, oxynitrides<br>— $O_2$ ash/solvent compatible<br>— Avoid $C_2H_6$, $C_3H_8$ (CVD)<br>— Avoid toxic solvents (spin-on dielectrics) |

**Fig. 14-4** Some of the many challenges of integrating low-$k$ dielectrics in a copper dual-damascene process flow. Courtesy of Semiconductor International.

### 14.4.1 Hydrogen Silsesquioxanes (HSQ)

Films of the spin-on inorganic polymer, *hydrogen silsesquioxane* (*HSQ*) were initially investigated as replacements for siloxane spin-on-glass (SOG) films in non-etchback SOG planarization processes. When it was found that they also exhibit a lower $k$-value than do $SiO_2$ films (their $k$-value is < 2.9) as well as good gap-filling capability, they became candidates for low-$k$ dielectric applications.[7]

HSQ is a subset of polyhedral oligomeric silsesquioxane macromers, a family of ordered three-dimensional polymers whose structure resembles a cage (Fig. 14-5a). HSQ can be considered a derivative of $SiO_2$, in which one of the four oxygen atoms bonded to every Si atom is replaced by hydrogen. That is, the caged structure of HSQ has an empirical chemical formula of $HSiO_{1.5}$. (The *sesqui* refers to the *one-and-a-half* in the stoichiometric bonding arrangement of $SiO_{1.5}$ to H.) In the $SiO_2$ lattice, all O-atoms act as bridges between the $SiO_4$ tetrahedra of the $SiO_2$ lattice (see Chap. 3 for details on the microscopic structure of $SiO_2$). However, when H (or a methyl group, $CH_3$), replaces an O-atom in the lattice, it no longer "bridges," but instead forms a terminal group. Thus, the tetra-

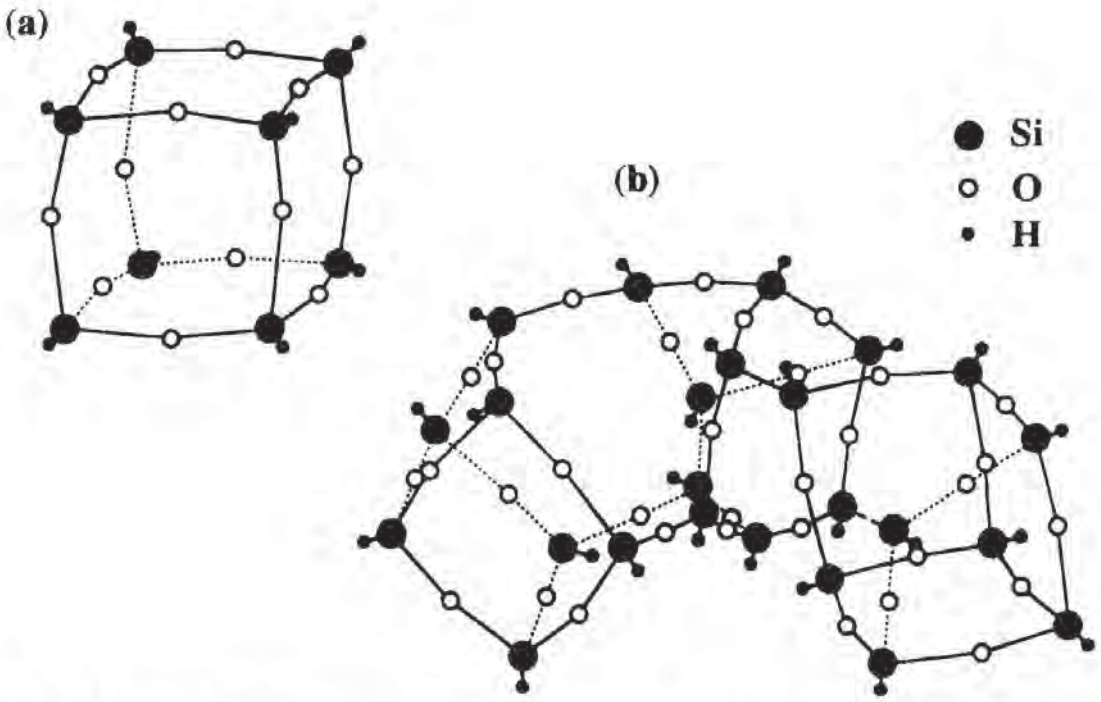



**Fig. 14-5** Schematic representation of the molecular structure of HSQ: a) eight-corner oligomer; and b) resinous material.

hedron that contains the Si-H bond is then bonded only at three of its four corners. This results in a more open lattice structure (i.e.., it exhibits microporosity). Note that in HSQ films some of the Si atoms that would have bonded to O-atoms in $SiO_2$ are instead bonded to H-atoms. The incorporation of such Si-H bonds in spin-on dielectrics (SODs) is inherent to the spin-on material. In low-*k* Si-H bond-containing films formed by CVD, the Si-H bonds are formed by incorporating gases in a CVD-chamber that contain H.

In its uncured state, HSQ resin is composed of a mixture of Si-O networks and cage-like networks (Fig. 14-5b). This random mixture of networks is polymerized during a two-step thermal procedure consisting of : **1**) a bake at 150-250°C, and **2**) a cure at ~400°C. During these thermal treatments, a rearrangement occurs. That is, the cage structures are opened up, and there is an intermolecular redistribution of Si-O and S-H bonds. In other words, HSQ is cured from a near cage-like structure to a three-dimensional network which provides mechanical integrity to the film. The built-in microporosity of the cured HSQ network results in a reduced dielectric constant.

However, in addition to the rearrangement reaction during curing, other reactions which convert Si-H bonds to silica can also occur. If such conversion to silica is excessive, it will result in densification of porous HSQ (i.e., shrinkage of

also increase the dielectric constant.[8] That is, low-$k$ properties in cured HSQ films can only be achieved if the density of Si-H bonding is maintained at a high level and the formation of Si-OH bonds and the absorption or creation of water in the film is minimized. Thus, when curing HSQ films, the process must balance between promoting the favorable rearrangement-reaction and inhibiting the Si-H conversion-reactions.[9] To achieve this balance, the curing-temperature must not exceed 450°C, and an optimum curing-temperature was reported to be in the vicinity of 400°C.[10]

The HSQ material is available in solution from Dow Corning under the trade name of FOx® (Flowable Oxide).[11] HSQ was reported to have been integrated into products by Texas Instruments in the 0.5-$\mu$m generation, and in 1997 an article describing its use in a five-layer Al-interconnect structure was published.[3] An oxide cap was used on the HSQ film to improve the stability of its low-$k$ value during W-CVD and Al-fill. That is, such HSQ-films must be capped with oxide to prevent moisture uptake in the film, and the integration sequence must be optimized to prevent cracking.

The built-in microporosity of the HSQ films, nevertheless, also reduces their mechanical strength. To overcome this drawback, a plasma treatment in $NH_3$ at 350°C for 20 min was performed. This was found to incorporate N atoms into the film. With the incorporated nitrogen, the HSQ films were found to be mechanically stronger. Specifically, scratching that was observed during CMP of non-plasma-treated HSQ films is mitigated.[12] In addition, such films could now be cured at temperatures up to 500°C without triggering the Si-H conversion-reactions.[13] Another benefit of such a plasma-treatment is that the thin nitride film formed on the HSQ-surface (35-nm thick) reduces leakage current through the film and also provides a good barrier against Cu migration. Leakage current was reportedly reduced by three orders of magnitude below that of an as-cured HSQ film, and evidence that Cu-migration was also suppressed was presented.[14]

### 14.4.2 Spin-On Methyl-Silsesquioxanes (MSQ)

Spin-on *methyl-silsesquioxane* (*MSQ*) films are similar to the HSQ materials described in the previous section, except methyl groups ($CH_3$) are used instead of hydrogen to replace some of the Si-O bonds (i.e., Si-$CH_3$ bonds are formed instead of Si-H bonds). Therefore, MSQ is often referred to as an organo-silicate material. With an empirical chemical formula of $CH_3SiO_{1.5}$, there are one-and-one-half (*sesqui*) stoichiometry of oxygen bound to silicon in a caged structure of MSQ. MSQ films have the benefit exhibiting excellent crack-resistance, and can

thus be used to form thicker dielectric films. In addition, they show better moisture absorption properties than HSQ.

One example of a spin-on MSQ material is T-9,™ sold by Tokyo Ohka Kogyo. It's $k$-value is 2.7. The curing procedure is about the same as that for HSQ. Another spin-on film containing $CH_3$ is sold by Allied-Signal ($k = 2.9$). It's commercial name is Accuspin T-23.™ When cured at 380–400°C, it can tolerate temperatures as high as 450°C before decomposing. Note that development is being conducted to develop ultra-low-$k$ dielectrics based on spin-on MSQs. One commercially available MSQ with sub-2.2 $k$ value is Dendriglass.™

### 14.4.3 Fluorinated Silicate Glass (FSG) Films

*FSG films* are silicon oxyfluorides ($F_xSiO_y$) deposited by CVD. That is, FSG is produced by doping an oxide-film with F in an otherwise conventional CVD process. Fluorine is one of the most electronegative elements, and its presence in the oxide makes the oxide less polarizable (which causes the $k$-value to decrease). Specifically, fluorine atoms replace oxygen atoms in the oxide film to create some Si-F bonds instead of Si-O bonds. Because the attraction of fluorine to its valence electrons is greater, these Si-F bonds are less polarizable than the Si-O bonds they displaced. (Figure 14-6 shows the molecular building blocks of FSG.) In addition, fluorine-to-fluorine repulsion is stronger than oxygen-to-oxygen, and this makes oxide films containing fluorine less dense than those without incorporated fluorine. The FSG films used in ICs exhibit a minimum $k$ value of ~3.5. This is because FSG films are formed in PECVD or HDP-CVD tools by adding a fluorine-source gas (e.g., $SiF_4$, $CF_4$, $C_2F_6$ , or $NF_3$, or other organo-silicate F sources), to the process-gas ambient used to deposit $SiO_2$ films by CVD (either silane and oxygen, or TEOS). Usually, $SiF_4$ or organo-silicate F sources are preferred because the incorporation of F into the silicate glass networks does not require breaking the strong Si-F bond (whose bond strength

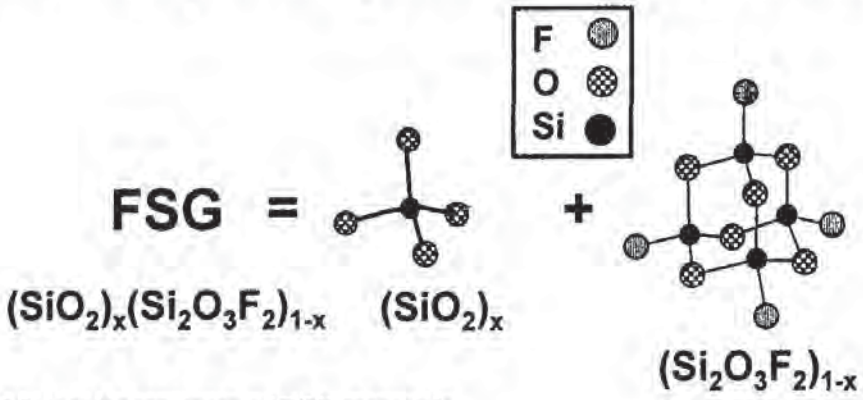



**Fig. 14-6** Molecular structure of FSG materials.

is roughly 135 kcal/mole). By increasing the fluorine-source-gas flow-rate, more F is incorporated into the film. Higher concentrations of F cause the value of $k$ to decrease, for the reasons explained above. However, a maximum of about 6% F is permitted to be incorporated into the FSG films because higher concentrations would make the FSG film unstable. That is, in such unstable films, F is evolved during RIE etching. This evolved F is highly reactive toward water. Aluminum interconnect lines embedded in this material would be corroded by the HF that is formed when the SiOF reacts with water. Because of this limitation, only a relatively small decrease in $k$ values is possible with FSG. Thus, the benefits of replacing non-fluorinated CVD-oxides with FSG are modest.

However, FSG is relatively simple to integrate because of its similarity to the doped-CVD-oxides currently being used in the IC industry. In fact, FSG-films were reported to be successfully integrated into 0.35,- 0.25,- and 0.18-$\mu$m circuits by IBM.[15] The evolution of even small amounts of F from the FSG oxide has raised concerns about using this material in Cu-interconnect systems. The F readily attacks the Ta-based liners of Cu systems, leading to volatile $TaF_2$ formation, and the consequent loss of adhesion between the low-$k$ film and the Ta. Despite such barrier stabilization problems, use of FSG has permitted extension of PECVD oxide processes to both subtractive-aluminum and copper-damascene interconnects. Note, however, that FSG solutions are different for the subtractive-aluminum process and the Cu-damascene process. That is, to obtain the gap-fill needed for the subtractive aluminum process, the sputter component encountered during the deposition of the FSG film bombards the depositing film. This breaks the bonds in the film and makes it less stable. Hence, even less F can be incorporated. In the damascene process, the dielectric film is deposited on a planar surface, and hence no gap-filling capability is needed. Thus, higher F-concentrations can be used. The $k$-value of FSG-films used with damascene interconnects is reduced by 0.2 to 0.3, compared to the $k$-values of FSG used with Al. As noted above, among the difficulties encountered when using FSG in a damascene process with Cu is the issue of keeping the fluorine tightly bonded in the CVD film (i.e., volatile TaF compounds can form at high temperatures).

Another possible application of FSG is replacing the conventional PECVD layer used to cap SOD low-$k$ films. It has been demonstrated that use of FSG as a capping layer for an SOD resulted in 10% reduction in parasitic capacitance.[16]

## 14.5 SECOND-GENERATION LOW-$k$ DIELECTRICS ($2.5 < k < 2.8$)

The largest variety low-$k$ dielectrics is found in the second generation of such substances, which exhibit $k$ values between 2.5 and 2.8. It is believed that these

materials will begin to be incorporated into ICs at the 0.18-$\mu$m generation. Both *spin-on-* and *CVD-low-k* materials have been developed within this range of permittivities, and each of these types are described here.

As noted earlier, a heated battle at the $k = 2.8$ node between CVD- and spin-on-dielectric (SOD) films is taking place. It is staged on the one hand by the desire of wafer manufacturers to continue to use CVD-tool-sets for one more technology node, versus the potential lure of being able to extend the use of new SOD materials and process integration schemes for multiple generations.

That is, CVD-low-*k* processes can reuse existing tool sets and allow for simpler integration (due to the silicon-oxide-like structure of the CVD dielectrics). From a business perspective, CVD-tool-suppliers can also provide an integrated solution to the customer. For circumstances where there is a tight labor market for engineering talent (e.g., the Taiwan silicon foundries in 2000), the foundries need to rely on low-*k* solutions from their vendors that have been well characterized and can be easily introduced into production. Thus, adoption of CVD-low-*k* dielectrics is likely to be favored in such cases.

However, an alternative case for SOD-low-*k* dielectrics at the $k = 2.8$ node is made by the suppliers of SOD-materials and SOD-track-manufacturers. They claim that SOD processes are extendable to future technology nodes, especially those below 0.25-$\mu$m (when porous low-*k* materials will need to be used). Porous low-*k* SOD materials are available now for evaluation and testing. The SOD porous films may also produce films with smaller pore sizes than those deposited by CVD. This would make them mechanically stronger, allowing better film stability and better integration. The key for SOD processes is to develop an SOD solution that can be extended for multiple generations and can process 200-mm and 300-mm wafers on the same equipment. Some of the issues remaining to be resolved with organic SODs include poor adhesion and via-poisoning due to improper etch and resist removal. Spin-on-dielectrics also remain susceptible to cracking at thicknesses approaching one micron.

There is also a question at the 0.18-$\mu$m technology node as to which process should be implemented first - copper or low-*k*? It appears that the most cost-effective gain occurs with aluminum/low-*k*, particularly if $k < 3.0$ can be implemented. Hybrid structures with aluminum/low-*k* for the lower levels of metal, and copper at the very top levels of interconnect will allow manufacturers the ability to ship 0.15/0.13-$\mu$m products, forestalling the significant challenge of integrating copper and low-*k*. Some companies have elected to take this path, but others have decided to implement copper interconnects first, before incorporating a lower-*k* dielectric material.

### 14.5.1 2nd-Generation Spin-On-Dielectrics: (2.5 < $k$ < 2.8)

Most spin-on *organic* low-$k$-polymers[17] are significantly different from their conventional SOG counterparts, primarily because they do not evolve moisture during curing. Since they exhibit better crack-resistance they can also be applied in a single pass to a greater thickness. Three organic spin-on low-$k$-dielectrics being considered are: **1)** *poly(alylene) ethers* (*PAE*); **2)** *fluoro-polyimides*; and **3)** *bis-benzocyclobutenes* (*BCB*). One other group of spin-on polymers (the *organo-siloxane polymers*), are more like their conventional SOG counterparts.

Spin-on-dielectric films are deposited without the use of an expensive CVD-tool. Instead, they are deposited with a less-costly track cluster-tool, consisting of spinners and hot plates. The cost of the spin-on liquid, however, is typically much higher per wafer than that of the CVD-reactant-gases.[18] In addition, a 100-nm-thick capping $SiO_2$ layer (deposited by CVD) will probably still be needed with spin-on low-$k$-layers. But an added benefit of using spin-on films is that they reduce the topology of any previous steps by more than 50%.

The *PAE materials* are aromatic structures with $k$-values of about 2.5 which can also withstand higher temperatures. A fluorinated PAE-product from Honeywell is named FLARE,™ and it has a $k$ value of ~2.8. It was designed to have low-outgassing characteristics and high thermal and mechanical stability.[19] Dow Chemical has a similar material named SiLK ($k$ = 2.62).™ Despite its name, SiLK is also a hydrocarbon that contains no silicon (and also no fluorine). A report by Fujitsu detailed their experience with integrating FLARE and SiLK into a two-level Cu-damascene-structure. These films showed little change in $k$-value after a five-hour anneal at 425°C.[20] Schumacher has also pioneered the use of non-fluorinated PAE-materials, with their latest such product being called VELOX.™ This material can be deposited without an adhesion promoter and does not require a capping layer. A *fluoro-polyimide* named FPI-136M with $k$ = 2.7 is sold by DuPont.

In April 2000, IBM announced plans to integrate SiLK into its next-generation 0.15-$\mu$m and 0.13-$\mu$m Cu-11 process, intended to go into high-volume production in the first half of 2001. They claimed the interlevel dielectric layers made with SiLK demonstrated excellent resistance to Cu-migration, and the mechanical strength of SiLK films is high enough to make them compatible with CMP (and back-end processes such as wire-bonding). IBM said it chose to use a spin-on-dielectric film for two reasons. First, SiLK meets the critical thermal, mechanical and chemical requirementss. Second, it offers extendability to the next generation. "To get the effective dielectric constant down into the 2.2-2.3

range, there are no materials, and there is no way to get there with PVD or CVD. Our opinion is that the best way to achieve that is by the introduction of porosity. That would say, whether its an organic or an organosilicate, you're going to have to do a spin-on process."[21] One reported difficulty involving the use of SiLK is that if the film is immersed in HF solutions (even as dilute as 0.1%-HF), the HF will diffuse through the film and attack the oxide at the oxide/SiLK interface. This will lead to SiLK liftoff.[22]

*BCB* is a co-polymer of divinyl-siloxane and bis-*benzocyclobutene* (which is often referred to as *DVS-BCB*, or simply *BCB*). It is a thermoset silicon-containing resin that exhibits a *k*-value of 2.7, but which has limited thermal stability, even at temperatures of 350°C. Nevertheless, there are no by-products to the polymerization reaction and no corrosive or volatile by-products are evolved during the cure. BCB demonstrates excellent gap-fill, adhesion, and planarization capabilities. Dow Chemical offers a BCB product called Cyclotene™ for IC ILD-applications. A study in 1998 indicated that Cu drifts much more slowly through BCB films than through PAE.[23] Thus, use of BCB films could relax the requirements of metal diffusion barriers. But its limited thermal stability constrains its use to applications with a reduced thermal ceiling.

As noted above, there is also a class of SOG-like materials available in the 2nd-generation low-*k*-range. These are *hybrid-organo-siloxane polymers* (*HOSPs*) which have a *k*-value of 2.5 after curing. HOSPs have a 12% carbon-content, which allows them to be formed into layers up to 1.2-$\mu$m thick without cracking. Their *k*-value remains stable at temperatures up to 380°C. HOSPs demonstrate good gap-fill and planarization capabilities. However, the quality of HOSP films (and those of other carbon-containing spin-on-dielectrics) is degraded by damage from an oxygen plasma used when stripping resist. The conventional oxygen-based ashing-process can attack the low-*k*-dielectrics that contain any carbon, which will break down the film integrity. To overcome this limitation, a plasma treatment in $H_2/N_2$ or $NH_3$ (as discussed in previous sections), has proved to be effective. Honeywell markets such a low-*k* material under the name HOSP.™

Note that a plasma-treatment of organic SOD layers with an $NH_3$ gas reportedly improves some of the properties of these films (as was also discussed in the previous section the covered HSQ films). That is, such a treatment produces a thin nitride layer on the SOD surface, and this layer acts both as a barrier against Cu migration and a layer that improves the ashing resistance of such films.[24]

### 14.5.2 2nd Generation-CVD-Dielectrics: (2.5 < *k* < 2.8)

Three dielectric materials deposited by CVD and with a value of *k* in the range of 2.5–3.0 have been developed: **1)** *carbon-doped siloxanes,* also termed *organosilicates glasses* (*OSG*), which have a *k* value of 2.5; **2)**, *fluorinated parylene* (*parylene-F*); and **3)** an *inorganic low-k material using organosilane/ozone.* One benefit of CVD deposition of low-*k* films is that the deposition process can be carried out in a multi-chamber tool that allows liner- and cap-layers to be CVD-deposited sequentially in a single tool. Another advantage is that the CVD processes use readily available and inexpensive gaseous precursors (in contrast to the expensive and proprietary spin-on liquid-precursors), which reduces the comparative chemical cost of the CVD approach. Furthermore, CVD does not require the handling of organic-solvent wastes.

**14.5.2.1 2nd Generation Low-*k*-CVD Dielectrics: Organo-Silicate Glasses:** The OSG dielectric films deposited by CVD basically have the same molecular structure as the HSQ- and MSQ-films deposited by spin-on methods. The CVD OSG films can be can be formed by PECVD or HDP-CVD, using a variety of precursors, including methylsilane, dimethylsilane, trimethylsilane (3MS), or tetramethylsilane, together with $N_2O$ (or in some cases ozone). As noted earlier, in the discussion on HSQ- and MSQ-films, the methyl-groups form terminal Si-$CH_3$ bonds on the tetrahedral structures that make up the oxide lattice. These Si-$CH_3$ bonds contribute to reducing the *k*-value by increasing the microporosity of the film and by reducing its polarization tendency. In OSG-low-*k*-films formed by CVD, these methyl-groups come from the precursor gases used in the CVD process. In general, the CVD process must balance the film requirements for low-*k* with other desired film properties. That is, low methyl-content can lead to films that behave merely like standard oxide films (i.e., with *k* ~4.0). On the other hand, high methyl-content can lead to Si-C-Si bonding, which will degrade thermal stability and will hamper the etching of the film (i.e., through the production of carbon polymers during the etch process).

Precursors with more methyl-groups are capable of incorporating more carbon into the film, though the compounds are more difficult to dissociate. As a result, their deposition rates are smaller. For example, tetramethylsilane deposition rates are about 25% slower than trimethylsilane because it takes more energy to dissociate the tetramethyl molecule. Another significant difference lies in the delivery system, since trimethylsilane is a gas and tetramethylsilane is a liquid. (A trimethylsilane precursor gas is offered by Dow-Corning, which they have named Z3MS.$^{TM}$) An advantage of OSG films is that they have thermal