SILICON PROCESSING FOR THE VLSI ERA

VOLUME 1:

PROCESS TECHNOLOGY

Second Edition

STANLEY WOLF Ph.D.

RICHARD N. TAUBER Ph.D.

LATTICE PRESS

Sunset Beach, California

DISCLAIMER

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

Published by:

LATTICE PRESS
Post Office Box 340
Sunset Beach, California 90742, U.S.A.

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

Library of Congress Cataloging in Publication Data
Wolf, Stanley and Tauber, Richard N.

Silicon Processing for the VLSI Era
Volume 1: Process Technology

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

ISBN 0-9616721-6-1

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

since they are desired in the film. Unwanted dopants, however, can also become incorporated in the film during the growth process. These are called *unintentional dopants*.

7.4.3 Intentional Doping

In order to control the conductivity type and carrier concentration (electrical resistivity) of epitaxial layers, dopants are intentionally added to the reaction gases during the growth process. These dopants are introduced through their hydrides: diborane [B_2H_6] for boron; phosphine [PH_3] for phosphorus; and arsine [ASH_3] for arsenic. Ultrapure hydrides are introduced into the reactor from gas storage cylinders, where the hydrides are diluted in hydrogen to concentrations of 20–100 ppm. When the dopant gases are mixed with the main hydrogen flow in the reactor their concentration is further diluted to the parts per billion range. The dopant concentration that is ultimately incorporated in the epi film depends on the relative amount of dopant in the gas stream. Although hydrides are relatively unstable above room temperature, they are prevented from dissociating in the gas phase by the presence of large quantities of H_2 in the reactor. Caution must be exercised when handling these dopant gases since they are extremely toxic.

Unfortunately, there is no simple rule to relate the concentration of dopant atoms in the gas phase with the number of atoms incorporated into the solid Si film. As it turns out, the dopant incorporation rate depends upon substrate temperature, deposition rate, dopant mole volume relative to the precursor mole volume in the gas stream, and reactor geometry. As a result, the dopant incorporation rate must be empirically determined for each reactor and set of growth conditions. For low concentrations of dopants, the amount of dopant incorporated is generally linearly proportional to the gaseous dopant concentration. For high mole volumes of dopant, the dopant incorporation becomes sub-linear with respect to the gas phase composition. Nevertheless, epitaxial films with well-controlled doping concentrations in the range of 10^{14}–10^{20} atoms/cm^3 can be routinely achieved once the process is characterized for a particular tool. A key factor adding to the control of the dopant concentration is the use of mass flow controllers (MFCs) to accurately meter the gas flows (see Chap. 6).

The silicon deposition rate during the deposition of doped epi layers is found to depend upon the dopant type and concentration. For example, PH_3 is known to reduce the deposition rate, while B_2H_6 increases the rate.[13] PH_3, which is readily adsorbed on the surface of the growing film, dissociates to form free hydrogen, which then bonds with silicon to form stable Si-H bonds on the surface. The hydrogen can be desorbed above 400°C leaving behind phosphorus on the surface. This phosphorus passivates the surface and suppresses the growth rate of the epi layer by hindering the 2-D growth of the film. AsH_3 exhibits similar behavior, and at high concentrations can substantially reduce the epi growth rate. B_2H_6, conversely, decomposes readily on the growing surface and deposits boron. This enhances the heterogeneous decomposition of the reactive silicon species and therefore enhances the growth rate. The amount of doping incorporated in the growing film also depends on the silicon deposition rate. At low rates more doping is incorporated for the same gaseous concentration, while at high rates less is incorporated. Arsenic incorporation (from AsH_3) decreases by a factor of 4, as the growth rate is increased 0.2 to 1.0 μm/min.[14]

7.4.4 Unintentional Doping (Autodoping and Solid State Diffusion)

Most epitaxial usage requires an epitaxial deposition of a lightly-doped layer (10^{14}–10^{17} atoms/cm^3) upon a heavily-doped substrate (10^{19}–10^{21} atoms/cm^3). The substrate may be uniformly doped or consist of buried layers. During the epitaxial deposition unintentional

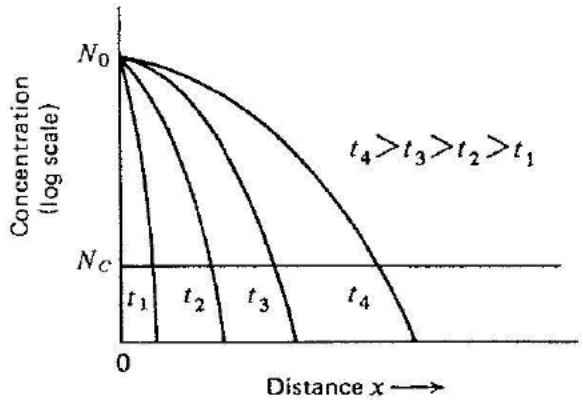

Fig. 9-2 Schematic representation of the diffusion profile for constant source diffusion (pre-deposition). Note the surface concentration remains constant with increasing diffusion time.[72] From S.K. Ghandi, *VLSI Fabrication Principles*. Reprinted with permission of John Wiley & Sons.

The junction depth, x_j, can be calculated by determining the depth (value of x) at which the diffusing dopant concentration just equals the background concentration in the silicon, C_{sub}. This is accomplished by solving Eq. 9-10 for x when $C = C_{sub}$:

$$x_j = (2\sqrt{Dt})\ \text{erfc}^{-1}\left(\frac{C_{sub}}{C_s}\right) \tag{9.12}$$

where erfc^{-1} is the inverse of the complementary error function. Both Eqs. 9-11 and 9-12 contain the "root Dt" term.

9.1.3.2 Drive-In Diffusion: Next consider what happens to the dopant atoms that have been introduced during the predeposition if the silicon is subjected to additional thermal treatment. We would expect that the total dose, Q_o, in the silicon should remain constant and the diffusant would move deeper into the silicon. This process is termed the *redistribution* or *drive-in* diffusion. The drive-in is used to move the impurities to the desired junction depth following the predeposition. In order to obtain a closed-form solution of Fick's Second Law, it is necessary to assume that the initial dopant is confined to a thin layer at the surface. As a matter of fact, the solution assumes that the dopant Q_o, exists as a δ-function in a thin rectangular region at the surface. The total quantity of impurity present is fixed at Q_o (atoms/cm^3). For the drive-in case, the initial condition is given by Eq. 9-13:

$$C(x, 0) = 0 \quad \text{for } x > \delta \tag{9.13}$$

and the two boundary conditions are given by Eqs. 9-14 and 9-15:

$$\int_0^{+\infty} C(x,t)dx = Q_o \tag{9.14}$$

and

$$C(\infty,t) = 0 \tag{9.15}$$

The boundary condition in Eq. 9-14, states that the total amount of dopant is fixed at Q_o. The solution to Fick's Second Law, under these conditions is:

$$C(x,t) = \frac{Q_o}{\sqrt{\pi Dt}} \exp\left(\frac{-x^2}{4\sqrt{Dt}}\right) \tag{9.16}$$

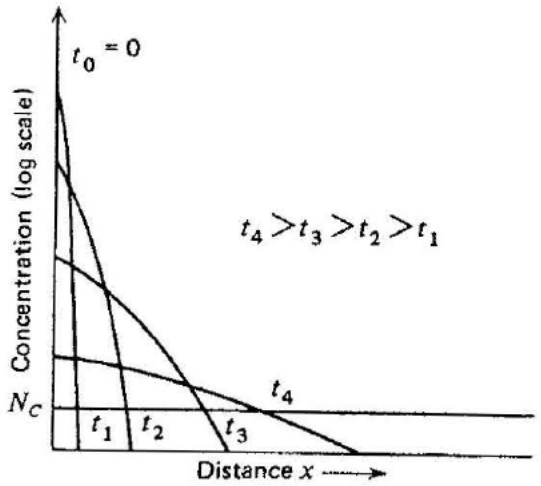

Fig. 9-3 Schematic representation of the diffusion profile for a limited source diffusion (drive-in). Note the surface concentration decreases with increasing diffusion time and the total dopant concentration remains constant.[72] From *VLSI Fabrication Principles*. Reprinted with permission of John Wiley & Sons.

EXAMPLE 9-2: Calculate the diffusion profile and junction depth after the predeposition of Example 9-1 is subjected to a neutral ambient drive-in at 1050°C for 60 min.

SOLUTION: For the drive-in conditions of this example the Gaussian distribution of Eq. 9-16 (the curve labeled *Gaussian* in Fig. 9-4) controls the profile. First calculate the diffusion length ($2\sqrt{Dt}$) for the new diffusion conditions:

$$D_B\ (1050°C) = 0.76\ e^{(-3.46\ eV/k\ 1323\ K)} = 3.3 \times 10^{-14}\ cm^2/sec.$$

Next calculate the value of the surface concentration C(0) from Eq. 9-17:

$$C(0) = Q_0/\sqrt{\pi Dt} = 4.7 \times 10^{14}\ atoms/cm^2/(\pi\ 3.3 \times 10^{-14}\ cm^2/sec \cdot 3600\ sec)^{1/2}$$
$$= 2.44 \times 10^{19}\ atoms/cm^3$$

Set up the following table to assist in the calculation:

x (cm x 10^{-4})	$x/(2\sqrt{Dt})$	$C(x)/C(0)$	$C(x)$ (atoms/cm^3)
0	0	1	2.44×10^{19}
0.1	0.45	0.75	1.83×10^{19}
0.2	0.91	0.60	1.46×10^{19}
0.3	1.36	0.17	4.10×10^{18}
0.4	1.82	0.038	9.20×10^{17}
0.5	2.27	0.006	1.46×10^{17}
0.6	2.73	0.0004	9.80×10^{15}
0.7	3.18	0.00004	9.80×10^{14}

Next calculate x_j in a manner similar to the one used in Example 9-1.

$$C(x_j)/C(0) = 1.5 \times 10^{16}/2.44 \times 10^{19} = 6 \times 10^{-4}.$$

Next find the corresponding value of $x/(2\sqrt{Dt})$ from Fig. 9-4, which is = 2.7. Hence,

$$x_j/\sqrt{Dt} = 2.7$$

or:

$$x_j = 2.7 \cdot 2.2 \times 10^{-5}\ cm = 5.9 \times 10^{-5}\ cm = 0.59\ \mu m.$$

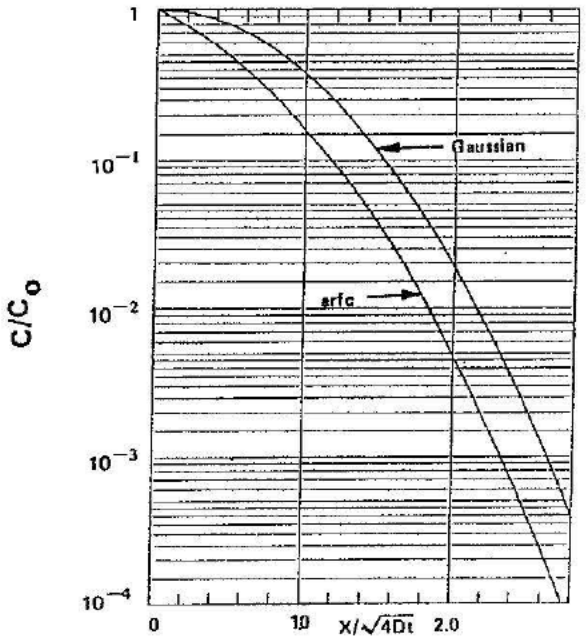

Fig. 9-4 Normalized concentrations [C/C(0)] as a function of normalized distance (dC/dt) for the erfc and Gaussian curves.

9.1.3.3 Drive-In From An Ion Implantation Predeposition: If the predeposition is from an ion implanted source (which produces a doping concentration with a near-Gaussian distribution close to the surface), the Gaussian profile will move when subjected to a high temperature anneal (see Chap. 10 on Ion Implantation). The solution to Eq. 9-6 (concentration-independent diffusion) with an initially implanted Gaussian profile has been reported for the cases of drive-ins performed in a neutral ambient,[4] and an oxidizing ambient.[5] For drive-ins in an oxidizing ambient, the solution to Eq. 9-6 is difficult to obtain in closed form, since it involves a moving boundary problem. Consequently, numerical methods must be employed to obtain the solutions (see Chap. 2, Vol. 3 of this series).

9.1.4 Concentration Dependence of the Diffusion Coefficient

The diffusion profiles that were calculated in the previous section are valid for the case when the diffusion constant is not a function of the doping concentration. That is why solutions of Eq. 9-6 were used, rather than those for Eq. 9-5. Constant-value D prevails when the doping concentration is lower than the intrinsic-carrier concentration n_i at the diffusion temperature, and conversely, concentration dependent diffusion occurs when the doping level is greater than the intrinsic carrier concentration. The first case is termed *intrinsic diffusion* while the latter is called *extrinsic diffusion*. To determine where extrinsic diffusion starts one needs to know the intrinsic carrier concentration at the diffusion temperature. Equation 9-19 is the expression used to calculate the intrinsic carrier concentration in silicon:

$$n_i = p_i = \sqrt{N_V N_C}\exp\left(-\frac{E_g(T)}{2kT}\right) \quad (cm^{-3}) \tag{9.19}$$

Here, N_V and N_C are the density of states at the valence and the conduction band edges, respect-

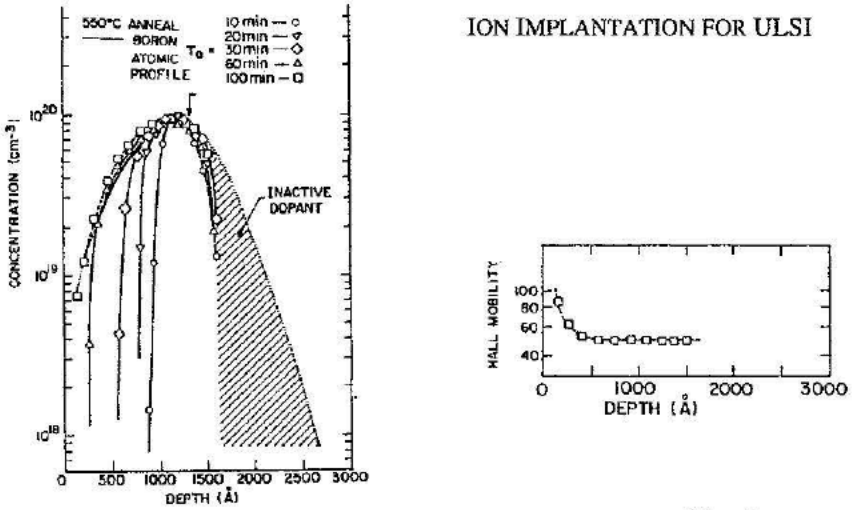

Fig. 10-22 a) Profiles for BF_2^+ implanted into <100> Si at 150 keV and $10^{15}/cm^2$ after different isothermal anneals. The dotted curve is the as-implanted atomic profile from SIMS analysis. The original amorphous-crystalline interface is denoted by the arrow, SPE is complete after ~100 minutes. The hatched region is electrically inactive.[64] b) Free-carrier concentration and mobility for implanted layers which illustrate dopant incorporation by solid phase epitaxy (SPE). Reprinted with permission of the American Physical Society.

even more complicated as a result of the presence of implantation damage. As an example, empirical studies of the *diffusion of B* in implanted single-crystal Si indicate that at high temperatures (≥ 1000°C), the data appears to obey ordinary diffusion theory (Fig. 10-23). At lower temperatures, however, ordinary diffusion theory does not accurately predict the diffusion behavior. That is, at 900°C the boron profile can be closely fit only if a diffusion constant is used that is about three times the value observed under a chemical 900°C diffusion. In addition, at 700–800°C, the depth of profile peak remains fixed, but the concentration in the tail is much deeper than is predicted from normal boron diffusion constants. This increase in the B diffusion (by as much as 4x) in the implanted profile tail occurs during the initial stage of the thermal anneal, but slows to the equilibrium diffusion rates for the rest of the anneal. Consequently, this

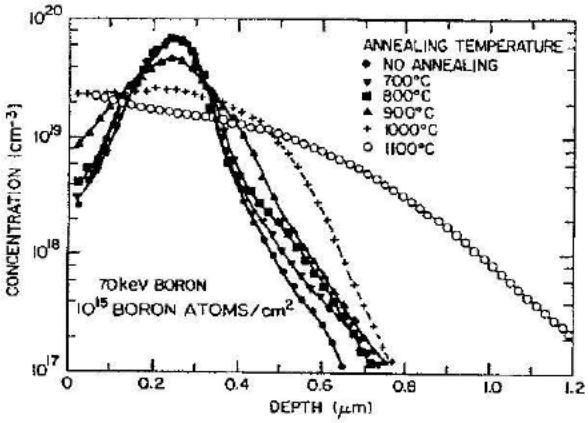

Fig. 10-23 Boron concentration as a function of annealing at various temperatures. The anneal time is 35 minutes.[5] Reprinted with permission of Philips Research Reports.

16.2 PROCESS SEQUENCES FOR TWIN-WELL CMOS FOR THE GENERATIONS FROM 1.2 μm TO 0.5 μm

This section describes an 11 mask, twin-well CMOS process flow that is representative of process sequences used to fabricate CMOS ICs from 1.2 μm down to about 0.5 μm. The process flow is nevertheless useful as a vehicle for illustrating the general sequence of steps employed in CMOS fabrication. Twin-well CMOS technology has been generally adopted for CMOS generations below about 1 μm because this approach allows two separate wells to be formed in a lightly doped substrate region. The doping profiles in each well can thereby be tailored independently so that neither device will suffer from excessive doping effects.

16.2.1 Starting Material: All MOS technologies use silicon wafers with a <100>-orientation. Typically, in twin-well CMOS the starting material is also either a lightly p-doped wafer (p-bulk wafer, Fig. 16-8a), or a heavily p-doped wafer on which a thin, lightly p–doped epi layer is grown (Fig. 16-8b). The concentration range of the p doping in both the bulk wafer and the p–epi layer is $8x10^{14}$–$1.2x10^{15}$ cm^{-3}. The p-epi-on-p^+ wafers (see Chap. 7) provide several advantages including: 1) improved latchup protection (see Vol. 2); 2) gate-oxides with better reliability are grown on epi layers than on bulk-Si surfaces; and 3) improved gettering capability (see Chap. 2). Their chief disadvantage is higher cost. (In 1998, the price of a 200-mm bulk starting wafer was ~$100 US, while that of a 200-mm epi wafer was ~$180 US.) For this example process flow, assume p-epi-on-p^+ wafers are used (Fig. 16-8b). The exact doping concentration of the epi layer is chosen to provide the best overall set of the following device characteristics: low source/drain-to-substrate capacitance; high source/drain-to-substrate breakdown voltage; high carrier mobility; and low sensitivity to source-substrate bias effects. As feature sizes shrink, the epi-layer doping concentration value is expected to increase.

16.2.2 Formation of Wells and Channel Stops: The wells are the first features to be formed in these generations of CMOS technology. A number of different procedures have been developed to form twin-wells. The most obvious method is to use two masking steps. Each blocks one of the two implant steps used to introduce the well dopants. A single masking-step procedure, however, was also developed. It is probably the one most commonly used, and it will be described here (see Fig. 16-9).

In this method, a single mask is used to pattern a nitride/oxide film that has been formed on the bare silicon surface (*Mask #1*). A *pad oxide* (which serves as stress-relief layer between the silicon nitride film and the silicon substrate) is first thermally grown to a thickness of 40–50 nm

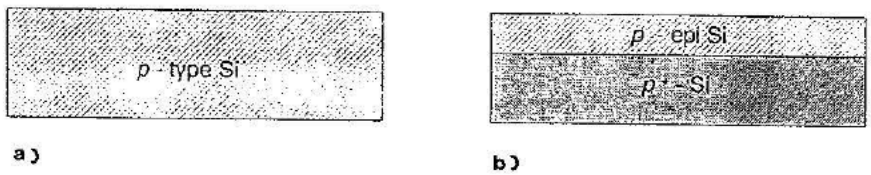

Fig. 16-7 Silicon starting substrates for twin-well CMOS: (a) p-bulk wafer; (b) p-epi-on-p^+ wafer.

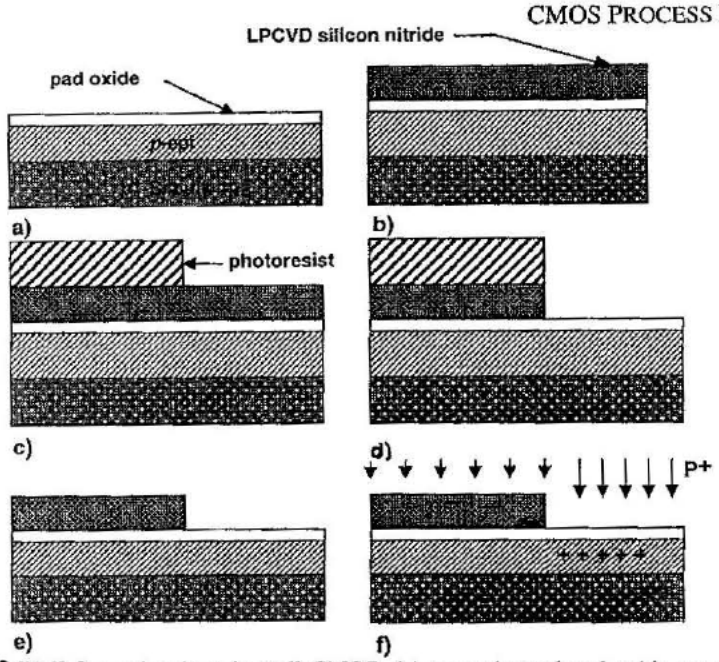

Fig. 16-8 Well formation in twin-well CMOS. (a) grow thermal pad oxide on p-epi-on-p^+-bulk starting wafer; (b) deposit LPCVD silicon-nitride film; (c) use *Mask #1* to define n-well regions; (d) etch nitride; (e) strip resist (f) implant n-well dopants (phosphorus).

(Fig. 16-8a). This is followed by an LPCVD of silicon nitride (140–200 nm thick- Fig. 16-8b). After the photoresist is applied, exposed and developed (Fig. 16-8c), the nitride is dry etched (Fig. 16-8d). The openings in the nitride film define the n-well regions. The required n-dopants are next introduced into these regions by ion implanting phosphorus (Fig. 16-8f). This implant is performed at 80 keV with a dose of ~$1x10^{12}$ cm^{-2}, using a medium-current implanter. The resist and unetched nitride regions (which define the p-well regions) act as an implant mask, preventing the phosphorus from entering the substrate in those regions. Next, a thermal oxidation step is carried out, using a wet oxide process (Fig. 16-9a). This grows an oxide to a thickness of 350–400 nm over the n-well regions. (Since nitride acts as a diffusion barrier, oxide grows only over the regions not covered by nitride.) The thickness of the thermal oxide is selected so it will block the subsequent boron implant used to form the p-wells. The nitride is next stripped using hot phosphoric acid (an etchant that does not attack the thermal oxide grown over the n–well regions). This leaves the silicon in the p-well regions covered only with the pad oxide, while the n-well regions remain covered with the thicker oxide. Thus, during the p-well implant (e.g., B at 50 keV, $1x10^{12}$ cm^{-2}, using a medium-current implanter), boron ions enter the silicon substrate only in the areas where the p–wells are to be formed (Fig. 16-9b). Next the well dopants are diffused into the substrate using a long, high-temperature drive-in step (e.g., 1100–1200°C for 3–20 hours). This is performed in an oxidizing ambient so that an oxide is also grown on the entire wafer surface during this step. At the conclusion of the drive-in step, the dopant concentration at the Si surface in the wells of 0.8-μm-CMOS is ~$1x10^{16}/cm^3$ for the p-well, and ~$3x10^{16}/cm^3$ for the n-well. The n-well might have a higher doping concentration to improve the punchthrough performance of the PMOS devices and to eliminate the need for a

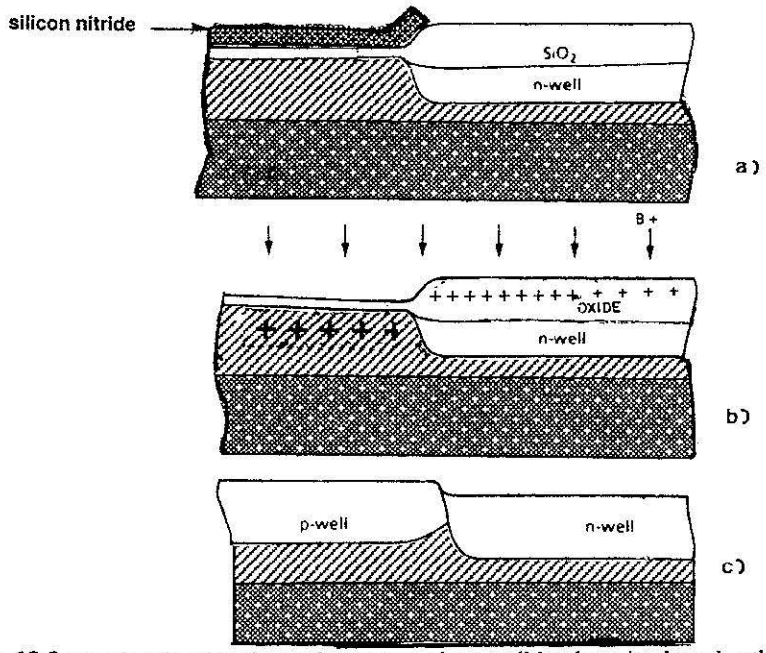

Fig. 16-9 (a) After the phosphorus that creates the *n*-well has been implanted, a thick oxide is grown, with the nitride preventing oxide form growing beneath it; (b) after stripping the nitride, a boron implant is performed that will form the *p*-well regions. The thick oxide above the *n*-wells prevents the boron from penetrating there; (c) wafer-surface topography after wells are driven-in and the oxide is stripped.

separate channel-stop implant step for the *n*-well. The oxide thickness is different over the *n*-well than the *p*-well, creating a step in the oxide film at the well boundaries. This oxide is nevertheless next etched away, but the step in the silicon remains behind (Fig. 16-9c). This step serves as an alignment mark to allow subsequent layers to be aligned to the well boundaries. (An alternative procedure that uses an additional mask can be used to create alignment marks. That is, deep pits can be etched into the wafer prior to the well formation step. Such marks provide improved registration accuracy for lithographic steps, and have become standard in deep-submicron CMOS technologies.)

It is important to observe that when wells are formed in this manner the well doping profile is at a maximum at the surface and then monotonically decreases with depth (see Fig. 16-10). The well doping concentrations are selected to be at least an order of magnitude higher than the epi-layer doping, so the expected variations in epi doping will not significantly affect the well concentration. The junction depth of the wells is typically ~2 μm deep.

16.2.3 Definition of Active and Field Regions: In conventional CMOS processes a standard LOCOS process (described in more detail in Vols. 2 and 3) is next used to define the active and field regions, and to create isolation structures between the active devices (Fig. 16-11). This is done by selectively growing a thick thermal oxide over the field regions, while keeping the active regions devoid of oxide. The steps that accomplish this begin with the growth of a pad oxide by thermal oxidation over the entire wafer to a thickness of 40–50 nm. Next, a blanket layer of silicon nitride (140–200 nm thick) is deposited by LPCVD (Fig. 16-11a). Then photo-

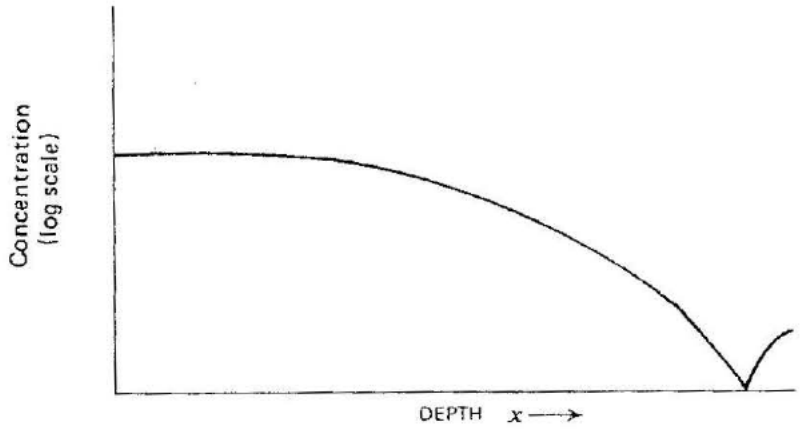

Fig. 16-10 Doping profile in a conventional well formed by a shallow implant and a drive-in step.

photoresist is spun onto the wafer and *Mask #2* is used to create the patterns that define the field and active regions of the circuit (Fig. 16-11b). After exposing and developing this pattern, the nitride is etched and the resist is stripped (Fig. 16-11c).

The next step creates the channel-stop dopant layer in the field regions of the *p*-wells. This is done by implanting a layer of boron atoms near the Si surface in the *p*-wells. This layer of implanted B atoms increases the threshold voltage of the parasitic PMOSFETs in the field regions of the *p*-wells. To prevent the boron from being implanted into the field regions of the *n*-wells, another mask is used. A layer of resist is applied prior to the boron channel-stop implant. *Mask #3* is used to form a pattern that covers the *n*-wells (Fig. 16-11d). Boron is then ion implanted into the surface (B with a dose of 10^{12}–10^{13} atoms/cm^2 at 40-80 keV, using a medium-current implanter). Boron only enters the silicon substrate in regions not covered by *Mask #3* or nitride (Fig. 16-11e).

A thick oxide (called a *field oxide*) is next grown selectively to a thickness of 400-700 nm over the field regions using a wet oxidation process (Fig. 16-11f). The combination of a thick field-oxide and channel-stop dopants beneath this oxide forms *parasitic MOSFETs* in the field regions whose threshold voltages that are too high to be turned on by normal voltages used to run the circuits. This keeps *active MOSFETs* in the IC electrically isolated from one another.

No oxide grows over regions covered with nitride (the active regions). Such a process of selectively growing field oxide has been named *local oxidation of silicon* (*LOCOS*). Observe, however, that some oxide does grow under the edges of the nitride film as a result of the oxidizing species diffusing laterally as well vertically. This forms a tapered oxide region at the edge of the active region called a *birds beak* (Fig. 16-12 and Chap. 8). As the active regions get smaller with device scaling, bird's beak begins to limit the use of LOCOS.

16.2.4 Threshold-Adjust Implantation Step: After the field oxide is grown, the nitride is stripped and the circuit is almost ready for carrying out the threshold adjust (or V_T-adjust) implant process and for growing the gate oxide. Prior to this, however, a problem caused by the previous field oxidation step of the LOCOS process must be rectified. During the wet oxidation process,

Generation	0.5um	0.35um	0.25um	0.18um	0.15um	0.13um
Isolation	LOCOS		STI			
Well	Diffused		Retrograde			
Gate	WiSix n+ Polycide		Dual Poly (n+/p+) Salicide			
Gate Oxide	Single Oxide		Dual/Triple Oxide => Oxynitride			
Gate Litho	i-line		248nm, Hi NA		PSM	193nm
Silicide	None	TiSix	TiSix => CoSix			
Contact/Via	Etchback		CMP			
BEOL Dielectric	SOG	Oxide	Low-k			
Metal	Al		Al =>Cu Damascene			

Fig. 16-21 The process modules used for successive submicron-CMOS generations.[13] © IEEE (1998).

boron deep below the surface of a bulk wafer, using high-energy ion implantation. This step creates a heavily doped B buried layer that acts in the same way as the heavily doped substrate of the epi wafer.[16] Good latchup protection and GOI are reportedly achieved by this approach. The cost for carrying out this step is 10–20% that of depositing an epi layer on a bulk wafer.

16.3.2 Formation of the Shallow-Trench-Isolation Structure: The isolation structures are formed next. Note that not only is a different type of isolation structure formed (*shallow trench isolation*, or *STI*) instead of LOCOS, but it is formed prior to the well structures. STI replaces LOCOS because STI has no bird's beak and provides a planar surface for further processing.[15]

The sequence of steps for forming STI structures begins with the growth of a pad oxide, followed by an LPCVD nitride layer. Then resist is applied and *Mask #1* is used to pattern the STI trench openings. The nitride and pad oxide are etched first. Then the trench is anisotropically etched (Fig. 16-22a) to a depth of 400–500 nm (e.g., using an etch gas mixture of $HBr/Cl_2/O_2$). The trench etching 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. 16-22f). After the trenches are etched the resist is stripped and a thin thermal oxide is grown on the trench walls (Fig. 16-22b). Next, a CVD dielectric film is used to fill the trench. This layer also covers the areas of the wafer where the nitride remains (Fig. 16-22c). A CMP step is used to polish back the dielectric layer until the nitride is reached (the nitride acts as a CMP-stop layer, Fig. 16-22d). The dielectric material is densified at about 900°C. Then the nitride is stripped, leaving the STI structure in place (Fig. 16-22e).

16.3.3 Formation of Retrograde Wells and Carrying Out of V_T-Adjust Implants: The well formation sequence occurs next. In addition, the V_T-adjust implants and the punchthrough implants are often carried out on the same ion implanter and during the same pumpdown as the well-formation implants. This process of performing multiple implants during the same pumpdown is termed a *chained implant*. The well formation step employs high-energy ion implantation to locate the well dopant peak ~1–2 μm below the Si surface. The implanted dopants will

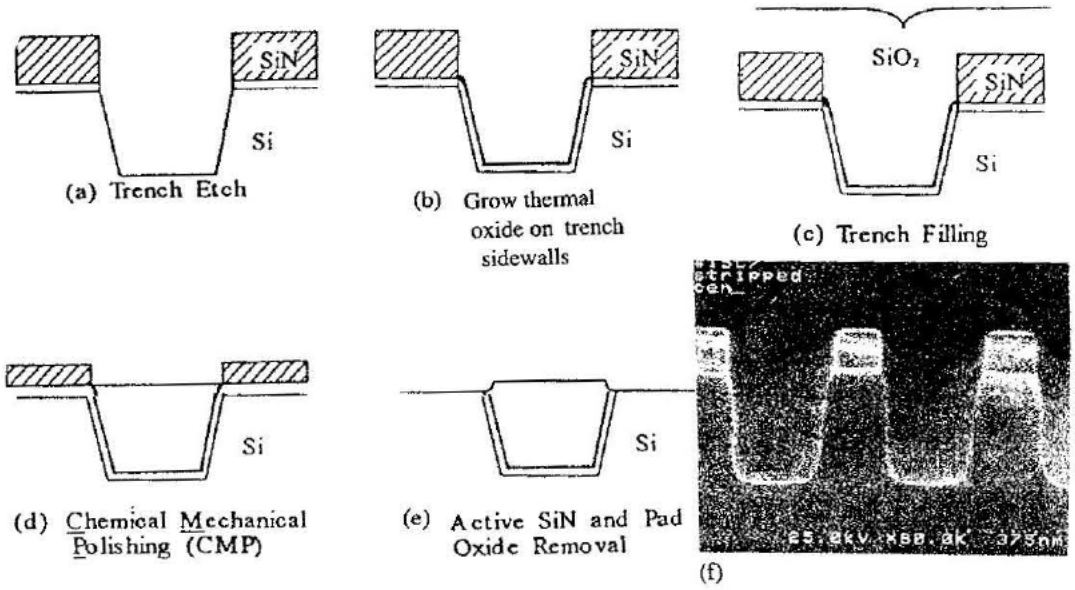

Fig. 16-22 Process sequence used to form shallow-trench isolation structures (STI): (a) etch trench in silicon substrate; (b) grow thermal oxide trench liner to improve trench Si/SiO_2 interface (and for trench corner rounding); (c) fill trench with CVD oxide; (d) use CMP to etch back the oxide to the nitride layer; (e) Strip nitride to leave the STI structure; (f) SEM of trench profile with 87° trench sidewall slope, and top and bottom trench corner rounding. Courtesy of Lam Research Corp.

have a *retrograde* concentration profile toward the surface from this peak (Fig. 16-23). Such a doping profile has several device performance advantages (discussed in Vol. 3). Since the dopants are placed deeply enough by the implant itself, there is no need for the long, high-temperature drive-in step required previously. Instead, only a lower-temperature anneal (carried out in a furnace at 900°C for 30 min), or a very-short, high-temperature RTP anneal (1100°C for 30 sec), is required to activate the implanted dopants. However, the high-energy ion beam used in the implantation process (500 keV–1 MeV) requires the use of a high-energy ion implanter and very thick resist films (~5 μm). Such thick resists are needed to prevent high-energy ions from reaching the Si surface in regions covered by the resist.

The well-formation process is begun by growing a layer of sacrificial oxide on the wafer. After applying the thick resist layer, *Mask #2* is applied to open the patterns for where the *n*-wells will exist. Then a high-energy P implant is carried out (700–850 keV, $1x10^{13}$ cm^{-2}). The peak of the P implant is ~1 μm below the Si surface (Fig. 16-24b). Both the V_T-adjust implant for the PMOSFETs (using P, since the dual-doped poly will allow the use of surface-channel PMOS devices – see Vol. 3), and the punchthrough implant are carried out in this implanter in the same pumpdown. Then the resist is stripped, and a new coat of thick resist is applied. *Mask #3* is then used to pattern the *p*-well openings. A high-energy B implant is then carried out (450–500 keV, $1x10^{13}$ cm^{-2}, Fig. 16-24c). A retrograde profile is again created, with the B peak concentration residing at 1-2 μm below the surface (Fig. 16-25). The V_T-adjust implant and the

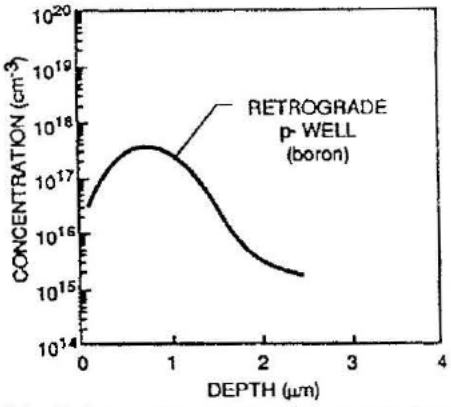

Fig. 16-23 Example of the doping profile of a retrograde *p*-well (boron) implant (400 keV).

punchthrough implant for the NMOSFETs (now using B) are also carried out during this "chained-implantation" procedure.

As discussed in Chap. 10, V_T-adjust implants with retrograde doping profiles have recently been introduced. A high-mass dopant of the same conductivity type is substituted for the ordinary dopant (e.g., indium for boron in NMOSFETs, antimony for phosphorus in PMOSFETs). The higher mass of the dopant reduces its diffusivity, which significantly decreases dopant redistribution during subsequent processes. Thus, the desired as-implanted profile is kept more intact. Such high-mass dopant V_T-adjust implants end up having a reduced

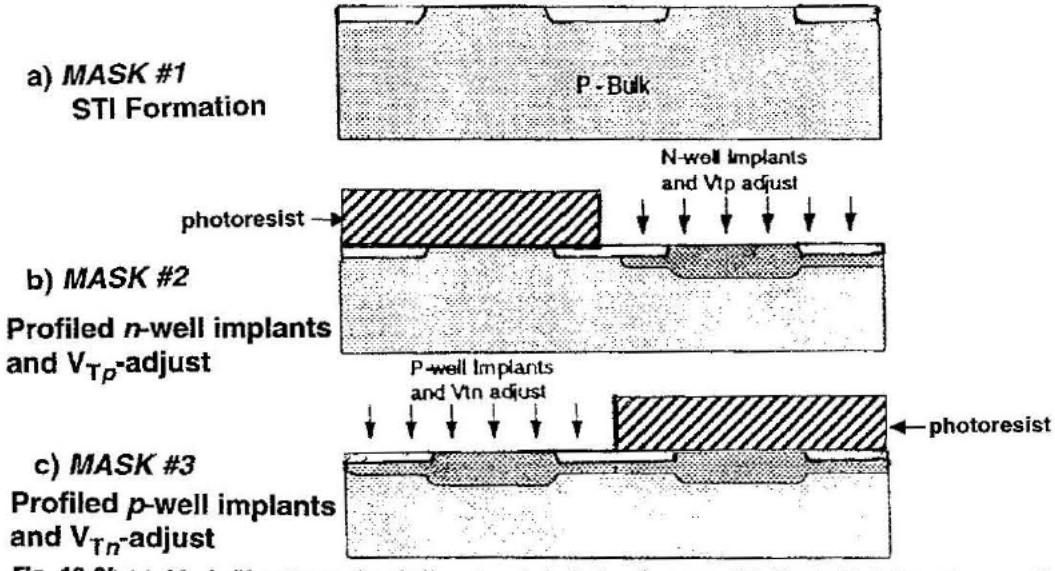

Fig. 16-24 (a) *Mask #1* patterns the shallow-trench isolation features; (b) *Mask #2* defines the *n*-well regions. A high-energy phosphorus (P) implant creates the retrograde *n*-wells, and a lower energy P (or Sb) implant is used to adjust V_{Tp}; (c) *Mask #3* defines the *p*-well regions. A high-energy boron (B) implant creates the retrograde *p*-wells, and a lower energy B (or In) implant is used to adjust V_{Tn}.

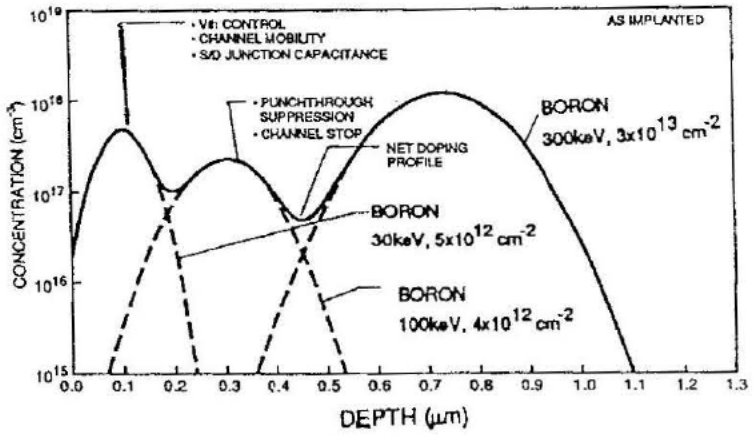

Fig. 16-25 Doping profile in the p-well showing the as-implanted: (a) retrograde p-well implant (B, 300keV); (b) the anti-punchthrough implant (B, 100 keV); and (c) the V_{Tn} implant (B, 30 keV).

doping concentration ($\sim 5x10^{16}$ cm^{-2}) at the interface while at the doping peak (about 80 nm below the oxide interface) the concentration is $\sim 1x10^{18}$ cm^{-2}. These profiles have been named *super-steep retrograde* (SSR) channel profiles (Chap. 10, Fig. 10-39). Such SSR profiles allow deep-submicron MOSFETs to be fabricated in which short-channel V_T-rolloff and drain-induced barrier lowering (DIBL) effects are significantly suppressed (see Vol. 3).

16.3.4 Formation of the Gate Oxide: The gate oxides of deep-submicron CMOS will continue to scale, with gate oxides for logic CMOS predicted to get thinner than 30 nm (see Chap. 8). The gate oxide, however, will have to prevent boron (from the p^+-poly) from penetrating into the Si substrate. The addition of nitrogen to the gate oxide has been reported to combat this problem, and so gate dielectric layers may be more like oxynitrides than SiO_2.[13] Different techniques have been investigated for incorporating the nitrogen into such films (see Chap. 8), including annealing the as-grown SiO_2 film in N_2O[18] or NO,[13] implanting N into Si and then growing SiO_2,[19] or exposing the SiO_2 to a high-density nitrogen discharge and then performing a post-nitridation anneal.[20]

16.3.5 Formation of the Gate-Stack Structures: Dual-doped polysilicon (i.e., n^+-doping of NMOS poly gates and p^+-doping of PMOS poly gates) is needed for 0.25-μm CMOS technologies (and smaller). This permits both the NMOS and PMOS devices to be fabricated as surface channel FETs (see Vol. 3), a necessity for being able to control short-channel effects.

The polysilicon layer is deposited undoped by LPCVD, to a thickness of 300-400 nm. But the selective doping of the appropriate poly structures is not carried out immediately. (Instead, it is performed later, simultaneously with the formation of the heavily doped regions of the source and drain junctions by ion implantation.) *Mask #4* is used to pattern the poly layer, and the gate structure is etched with an anisotropic dry-etch step (Fig. 16-26a).

16.3.6 Formation of the Source/Drain Junctions: After the poly gate structure is etched, the formation of the source/drain regions is undertaken. These regions in the substrate have to provide a low-resistance path from the contact to the channel edge. They are formed in two steps. In the