

where q is the electronic charge. Although μ depends strongly on the concentration of doping atoms and implantation damage, values of R_S have been tabulated utilizing known mobility data.³¹ For a known dose, full electrical activity is reached when the predicted R_S is reached.

Electrical activation of implanted impurities in amorphous layers proceeds differently than in layers with primary crystalline damage. As will be discussed, electrical activation in amorphous layers occurs as the impurities in the layer are incorporated onto lattice sites during recrystallization. Electrical activation in crystalline damaged regions exhibits more complex behavior.

For example, Fig. 10-20 shows the *isochronal* electrical activation behavior of implanted boron (i.e., anneals performed at varying temperatures, but for identical times). In this curve, the measured surface carrier concentration (normalized for different junction depths to the dose, in cm^{-2}) is used to indicate the degree of activation. That is, when $p_{\text{Hall}}/\phi = 1$, full activation is reached. Note that other impurities exhibit similar behavior to that shown in Fig. 10-20, provided the implantation does not cause a continuous amorphous layer to be formed.

The temperature range up to 500°C (Region 1 of Fig. 10-20) shows a monotonic increase in electrical activity. This is due to the removal of trapping defects and the concomitant large increase in free carrier concentration as the traps release the carriers to the valence or conduction bands. In Region 2 (500–600°C), substitutional B concentration actually decreases. This is postulated to occur as a result of the formation of dislocations at these temperatures. Some boron atoms that were already on substitutional sites are believed to precipitate on or near these dislocations. In Region 3 (>600°C), the electrical activity increases until full activation is achieved at temperatures ~800–1000°C. The higher the dose, the more disorder, and the higher the final temperature required for full activation. At such elevated temperatures, Si self-vacancies are generated. They migrate to the B precipitates, allowing boron to dissociate and fill the vacancy (i.e., a substitutional site).

Activation of implanted impurities by rapid thermal processing (RTP, see Chap. 9) has also been studied. The time-temperature cycle to reach minimum sheet resistance for As, P, and B is ~5–10 sec at 1000–1200°C, the exact condition being dependent on implanted species, energy, and dose.^{33,34}

10.3.4.2 Annealing of Primary Crystalline Damage: Isolated point defects and point defect clusters (that predominantly occur during light ion implantation), and locally amorphous zones (that are typically observed from light doses of heavy ions), are both regions of primary crystalline damage that exhibit comparable annealing behavior. At low temperatures (up to ~500°C), vacancies and self-interstitials that are in close proximity undergo recombination, thereby removing trapping defects. At higher temperatures (500–600°C), as described above, dislocations start to form, and these can capture impurity atoms. Temperatures of 900–1000°C are required to dissolve these dislocations. Note that the activation energy of impurity diffusion in Si is always smaller than that of Si self-diffusion (see Chap. 9). Therefore, the ratio of defect annihilation to the rate of impurity diffusion becomes greater as the temperature is raised. This implies that *the higher the anneal temperature the better*, with the upper limit being constrained by the maximum allowable junction depth dictated by the device design.³⁰

It is also important that the steps used to anneal implantation damage be conducted in a neutral ambient, such as Ar or N₂. That is, dislocations which form during annealing³⁵ can serve as nucleation sites for oxidation induced stacking faults (OISF, see Chap. 2) if oxidation is carried out simultaneously with the anneal (i.e., the annealing is performed in an oxygen ambient).

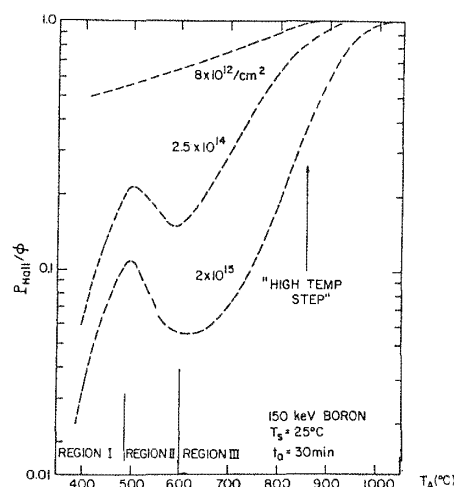


Fig. 10-20 Isochronal annealing behavior of boron. The ratio of free-carrier content, P_{Hall} , to dose ϕ is plotted against anneal temperature T_A for three doses of boron.³² Reprinted with permission of Gordon and Breach Publishing Company.

10.3.4.3 Annealing of Amorphous Layers: The annealing of continuous amorphous layers that extend to the Si surface has been found to occur by solid-phase epitaxy (SPE),³⁶ at temperatures between 500–600°C (Fig. 10-21). That is, a recrystallization process occurs on the underlying crystalline substrate, and regrowth proceeds toward the surface. The amorphous layer regrows at varying rates, depending on the annealing temperature, crystal orientation, and implanted species, although at 600°C regrowth is usually completed in a matter of minutes. Regrowth is faster on (100) than (111) Si, and impurities such as B, P, and As enhance regrowth, while O, C, N, or Ar retard regrowth. In practice, 550°C is favored, since (100) regrowth at this temperature occurs at a controllable and reproducible rate.

The impurity dopant atoms are swept into substitutional lattice sites during the SPE regrowth, and thus full electrical activation within the amorphous layer can be obtained at relatively low temperatures. The impurities that are implanted into the region beyond the amorphous layer, however, must be subjected to the higher temperatures needed to electrically activate impurities in regions of primary crystalline damage (Fig. 10-22). Therefore, to fully activate an amorphous layer, and the region of heavy primary crystalline damage behind it, higher temperature anneals than 600°C must be used (normally 800–1000°C). Note that specific *minimum* times and temperatures depend on the particular implanted species and dose.

If the amorphous layer does not extend all the way to the surface (i.e., a buried amorphous layer), the annealing proceeds differently.³⁸ That is, SPE occurs at both amorphous-single crystal interfaces, and the regrowing interfaces meet below the surface. This meeting point has been found to be a heavily damaged layer, with properties likely to cause device degradation. Thus, it is prudent to select implantation conditions that avoid the formation of buried amorphous layers. Note that if a wafer exhibits color bands after implantation and anneal, they are probably symptomatic of a subsurface damage layer left behind by a buried amorphous

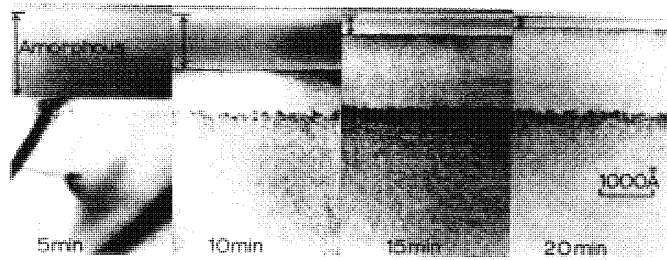


Fig. 10-21 Solid phase regrowth of a 200 keV, $6 \times 10^{15}/\text{cm}^2$ antimony implantation at 525°C. TEM cross section micrograph. Courtesy of Institute of Physics, Conference Series.³⁶

layer. The gives rise to optical interference effects from the light reflected off the subsurface damage layers and the surface.

* Some of the crystalline defects in the region beyond the amorphous layer are annealed out during subsequent thermal cycles, but others give rise to extended defects (such as dislocation loops and stacking faults), which then grow and interact. Under some implantation and annealing conditions, these defects move to the surface and eventually disappear, while in others they grow into larger structures which intersect the surface or remain in the bulk.

In a large number of cases, the total number of "excess" Si atoms found in dislocation loops after thermal annealing fits a remarkably simple model. Even though hundreds to thousands of atoms are displaced from their positions in the Si lattice by each ion impact, the number of residual Si atoms left out of the lattice after the completion of thermal annealing increases with ion dose and is closely proportional to the number of ions. This is known as the "+1" model, where the number of excess Si interstitial atoms is equal to the number of implanted dopant atoms that occupy lattice sites after thermal annealing.^{38,39} These Si interstitials arise from the dopant atom replacing the Si-atom in the lattice.

10.3.4.4 Dynamic Annealing Effects: The heating of the wafer during implantation can impact the implantation damage and the effects of subsequent annealing. A rise in temperature increases the mobility of the point defects caused by the damage, and this gives rise to healing of damage even as the implant process is occurring, hence the name *dynamic annealing*.⁴⁰ In the case of light ions, sufficient damage healing may occur to prevent the formation of amorphous layers, even at very high implantation doses. In the case of heavy ion implantations, dynamic annealing can cause amorphous layer regrowth during the implantation step.

A study by Prussin *et al.*,⁴¹ showed that the wafer cooling capability of an ion implanter can impact the structure of the damage following implantation because of dynamic annealing effects. That is, if a wafer is prevented from being significantly heated above room temperature by adequate heat sinking during implantation, dynamic annealing is minimized. On the other hand, if no heat sinking is provided and wafers are allowed to rise to temperatures ~150–300°C, dynamic annealing effects can produce changes in implantation damage structures. This typically occurs in non-reproducible and unwanted ways, such as the formation of buried amorphous layers, or crystalline layers containing high densities of dislocation loops.

10.3.4.5 Diffusion of Implanted Impurities: As described in Chap. 9, the diffusion of impurities in single-crystal Si is a complex phenomenon. The diffusion of impurities in implanted Si is

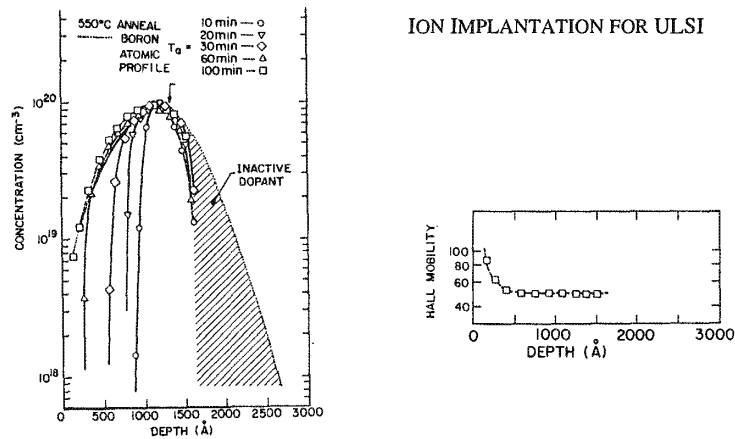


Fig. 10-22 a) Profiles for BF_2^+ implanted into $\langle 100 \rangle$ Si at 150 keV and $10^{15}/\text{cm}^2$ after different isothermal anneals. The dotted curve is the as-implanted atomic profile from SIMS analysis. The original amorphous-crystalline interface is denoted by the arrow, SPE is complete 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 ($\geq 1000^\circ\text{C}$), 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 $4\times$) 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

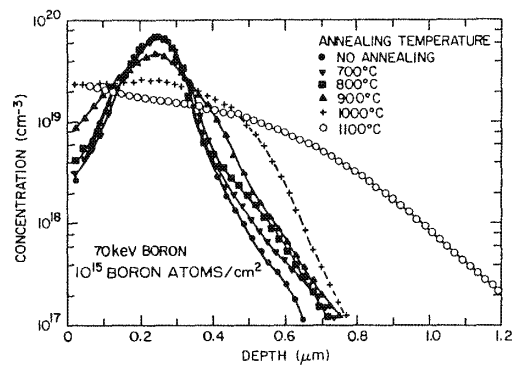


Fig. 10-23 Boron concentration as a function of annealing at various temperatures. The anneal time is 35 minutes.⁵ Reprinted with permission of Philips Research Reports.

phenomenon is called *transient enhanced diffusion* (or *TED* – which is discussed in detail in Chap. 9). It is driven by the extra defects and Si interstitial atoms that are present after the implant, but which are rapidly annihilated during the first stages of thermal anneals.

In early semiconductor device fabrication processes, diffusion of impurities during annealing was often used to drive them to depths beyond the range of implantation damage. This resulted in the production of junctions which were not degraded by lattice damage. Present demands for shallow junctions in CMOS (as well as for narrow-base and shallow-emitter regions in bipolar devices), no longer allow extensive dopant redistribution during anneal. Therefore rapid thermal processing (see Chap. 8) is used to anneal implantation with minimal impurity redistribution. RTP cycles of $\sim 1000^\circ\text{C}$ for 10 sec can activate implanted layers as effectively as 30 minute furnace anneals at 1000°C (Fig. 10-24), but with impurity redistribution distances of only a few hundred angstroms (compared to several thousand angstroms for furnace anneals). RTP cycles for shallow junction annealing often use "spike" anneal temperature profiles, where a fast temperature ramp up is followed immediately by a cooling down after the anneal temperature is reached. Shallow junctions have also been fabricated by "pre-amorphizing," in which a combination implantation is performed as a way to reduce the depth of the tail from channeling. That is, an implant of Si or Ge is first carried out to amorphize the Si surface, and then the dopants (such as B) are implanted. Annealing with RTP is conducted after implanting the desired impurity (see also Sect. 11.6.4 which deals with *Shallow Junctions*).

The redistribution of impurities in polysilicon should also be considered. It is observed that dopants redistribute themselves much more rapidly in polysilicon than in single-crystal Si (as a result of grain boundary diffusion). Thus, even short RTP cycles that anneal implantations in single-crystal Si without appreciable redistribution are likely to uniformly distribute impurities throughout a thin film of polysilicon. This is useful for producing polysilicon gates with reduced boron depletion effects. For some impurities (e.g., As), a capping oxide must be present to prevent significant As outdiffusion from the polysilicon by such cycles.

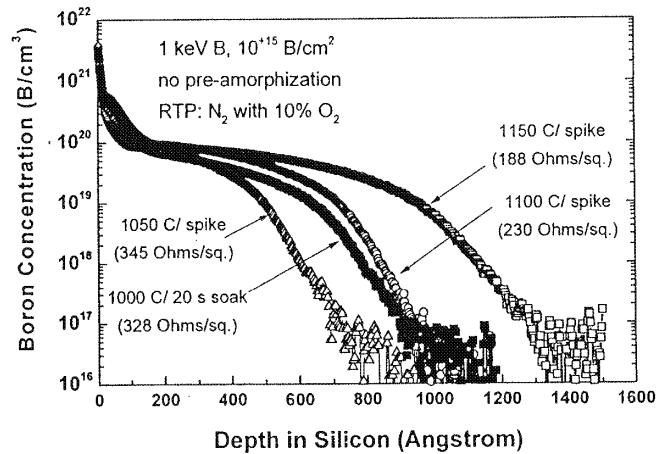


Fig. 10-24 Atomic profiles for 1 keV boron implants after a soak anneal at 1000°C for 20 sec or spike anneals at 1050°C , 1100°C , and 1150°C .¹⁰³ Reprinted with permission of Elsevier.

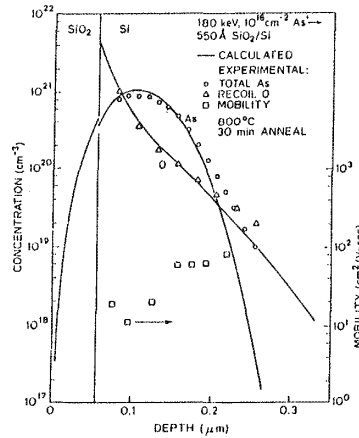


Fig. 10-38 Comparison of a Boltzmann transport equation calculation and experimental SIMS results for the recoil oxygen distribution resulting from an implantation of As into 55 nm of SiO₂ from a dose of 10¹⁶ cm⁻². Reprinted with permission of the American Physical Society.¹⁹

and

$$\frac{R^2}{\sigma_p^2} = \left[\sum_i \gamma_i \frac{C_i}{R_i} \frac{R_i^2}{\sigma_{pi}^2} \right] \left[\sum_i \gamma_i \frac{C_i}{R_i} \right]^{-1} \quad (10-20)$$

where

$$\gamma_i = 4 M_1 M_i / (M_1 + M_i)^2 \quad (10-21)$$

and C_i represents the fraction of the total mass of the i^{th} component, and R_i and σ_i , the range and projected range of the i^{th} component in mass units per cm². Monte Carlo methods should be used when higher accuracy is required.

10.6.3 Threshold-Voltage Control in MOS Devices

One of the most important applications of ion implantation is the *control of the threshold voltage of MOS devices*. When boron atoms are implanted into the channel region below the gate oxide of MOS devices, the threshold voltage V_T changes by $\Delta V_T = -\Delta Q_B / C_{ox}$, where ΔQ_B is the change of the sheet ionized dopant charge in the channel (the ion dose), and C_{ox} is the gate oxide capacitance per unit area. To achieve adequate control over the transistor threshold voltage, the ion dose and energy need to be controlled to within a few percent at a dose of $\approx 10^{12}$ ions/cm² (about 0.1% of a monolayer of atoms). The ability of ion implantation to allow this level of control was a critical factor in its initial introduction into IC fabrication process.

As channel regions are scaled to lengths of 0.1 μm (100 nm), a new problem occurs because the total number of dopant atoms in the channel becomes a countable and statistically fluctuating quantity. For a dose of 10¹² ions/cm², there are approximately 100 atoms implanted into a channel region that is 100 $\times$ 100 nm² (0.1 $\times$ 0.1 μm) in area. This means although the average dose uniformity can be controlled to less than 1% over a wafer (and over a large number of lots), the variation in the number of dopant atoms in individual, adjacent channel regions will be $\approx \pm 10\%$ ($[100]^{1/2}/100$). This variation leads to a large spread in the threshold voltage over a device and to large sub-threshold leakage currents.

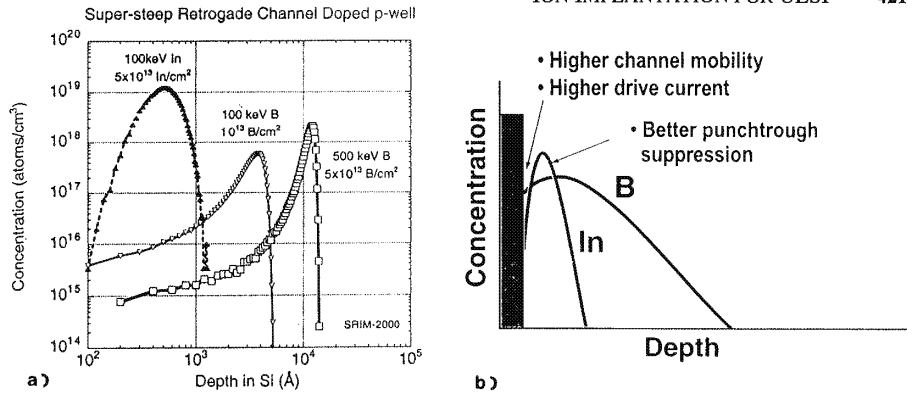


Fig. 10-39 (a) Calculated profiles for a "super-steep retrograde doped" channel implant with 100 keV at a dose of 5×10^{13} indium ions/cm² in a multi-implanted *p*-well background.⁹¹ © IEEE (1995). (b) Schematic comparison of channel profiles using In and B. The super-steep In retrograde profile improves punchthrough suppression, and electron mobility is increased by lowering interface dopant levels.¹⁰⁴

An alternative approach to threshold control is to leave the channel region just below the gate oxide relatively undoped and to confine the source to drain current flow by the use of a rapidly rising well doping profile, a "super-steep, retrograde channel doping" implant.⁹¹ Since the channel implants are done relatively early in the fabrication process (i.e., before the gate oxidation, gate formation and source/drain doping steps), it is important to use dopants which do not diffuse far from their as-implanted profiles during the later thermal processes. For this reason, indium (for *p*-wells) and antimony (for *n*-wells) are the leading candidates for this process. An example of the implant profiles for an In-implanted channel is shown in Fig. 10-39.

10.6.4. Shallow Junction Formation by Ion Implantation

As transistor lateral dimensions are reduced to achieve higher speed performance and higher density of functional components on a die (most notably by shrinking the size of the MOSFET gate), the junction depths and doping profiles are also restricted to shallower locations (i.e., closer to the contact plane at the base of the interconnect network). There are two main regions where this "junction scaling" is important: 1) the "channel region" which encompasses the contacts and doped areas around the CMOS source and drain (including the channel under the gate); and 2) the "well" regions (which determine the background doping for the transistor, see Fig. 10-40). The channel region is extremely shallow, with channel depths for MOSFETs with 100 nm gate sizes in the range of ≈ 30 nm. The well doping extends to much greater depths (typically to $\approx 1 \mu\text{m}$), and well formation is discussed in the next section. As the junctions are scaled to shallower depths, the ion energies used to implant dopants into both the channel and well regions decrease in rough proportion to the gate size (see Fig. 10-41).⁹²

To obtain the shallow junctions required for source/drain extensions and channels, low energy ions must be used (e.g., below 1 keV in the case of boron implantation). Furthermore, all other physical mechanisms which allow dopants to penetrate deeper into Si than the as-implanted profile must be strictly controlled or eliminated. The principal factors to be controlled are ion channeling and dopant diffusion. To accomplish these ends, shallow-junction doping pro-

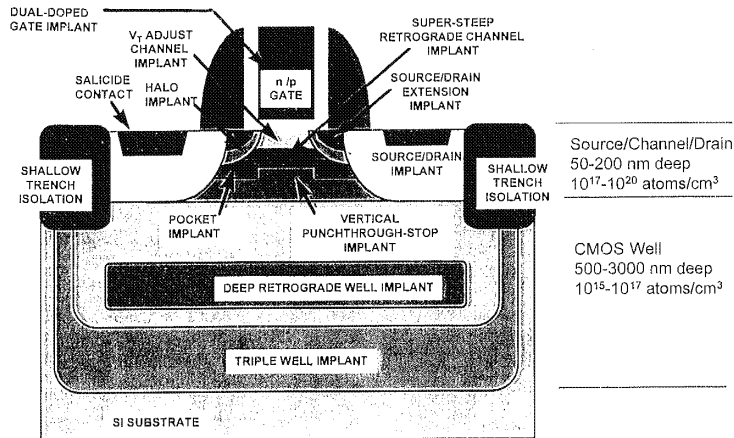


Fig. 10-40 CMOS transistor doping regions.

cesses often use combinations of: 1) high-dose, "pre-amorphization" implantation steps before the actual dopant implantation; and 2) rapid thermal annealing cycles. For example, a 5-keV pre-amorphization Ge implant (at a dose of 10^{15} Ge/cm²) may be used to control channeling. This is combined with a short RTA step (≈ 1 sec) at a temperature of $\approx 1050^\circ\text{C}$ (to anneal the lattice damage created by the implant and to electrically activate the dopant atoms, without allowing appreciable diffusion of the implanted species). In addition to the peak anneal temperature, such factors as the temperature ramp rates and gas ambient (trace levels of oxygen, for example) are important process-control factors.

The source extensions are particularly critical junctions. In order to suppress the roll-off of the threshold voltage as the gate size is decreased, the junction depth of the extensions must also decrease proportionately (i.e., 30-50% as much as the gate length is scaled). To minimize the series resistance from the source contact to the drain, the sheet resistance of the extension region must also be minimized. This is accomplished by achieving a high percentage of dopant activation while still limiting the dopant diffusion (which must be controlled to maintain a shallow junction depth). In addition, the abruptness (change in doping concentration) of the

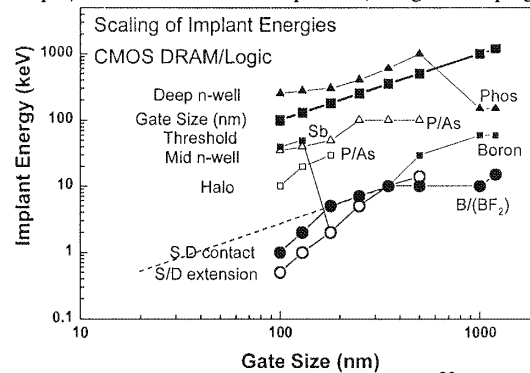


Fig. 10-41 Ion implantation energies as a function of gate size scaling.⁹²

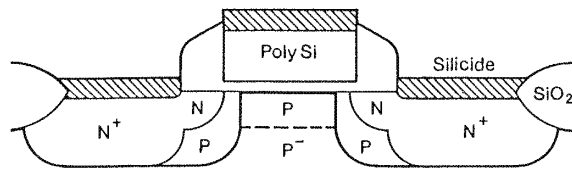


Figure 1. DI-LDD device cross section, vertical doping profile through the LDD region, and lateral doping profile at the surface.

Fig. 10-42 Cross-section of an NMOSFET fabricated with a lightly-doped drain (LDD) and a *p*-halo implant beneath the *n*-drain extensions.⁸⁶

extension boundary into the channel region must also be kept as sharp as possible to minimize resistance effects.

One method for simultaneously increasing the abruptness of the extension junction doping profile, decreasing the extension junction depth, and increasing the punchthrough voltage is to implement a *halo implantation*. In a halo implantation, dopant of the same type as the major well dopant is implanted beneath the drain extension junction (Fig. 10-42).⁸⁶ The “halo” is used because the implantation is done after the formation of the gate, and so the implanted dopant is self-aligned to the gate edge and surrounds the gate with a ring of dopant. For *n*-well transistors, arsenic is often used for the halo implants, while P is used to dope the deeper portions of the well. Arsenic is used for the halo doping since it has a lower diffusion rate and the halo profile can be better controlled during the source/drain activation cycle.

The source/drain contact doping must be deep enough to allow for the formation of low contact resistance interfaces between the junction and the interconnect metal network. The contact junctions must also have a low enough resistance that the current flow from the source contact to the channel region (and the mirror current path at the drain) spreads out evenly though the junction and does not concentrate at the inside edges of the contact pad. If the sheet resistance of the contact junctions is too high, heating of the transistor through “current crowding” effects can not only limit the transistor drive current, but also in the worst case cause circuit failure by contact electromigration effects (see Vol. 2, Sect. 4.7.3.2 of this series).

The contact junctions must also have a rapid change in doping concentration at the location of the bottom of the junction to obtain the lowest sheet resistance and form the shallowest junctions that still have the low-leakage currents of *pn* diodes (see Fig. 10-43).⁹²

10.6.5 High-Energy Implantation

Even as junctions of the source/drain and channel region call for ion energies of a few keV or less, the doping of CMOS well junctions requires ion energies ranging from a few hundred keV to several MeV⁹³ (see Fig. 10-41). The particular ion energy used depends on the application and device type. For example, high-speed logic (microprocessors) and DRAM devices use relatively shallow well structures while SRAM and FLASH memories use deep wells, sometimes in combinations with even deeper “triple wells” (which allow for the *n*- and *p*-well junctions to be biased separately from the substrate).

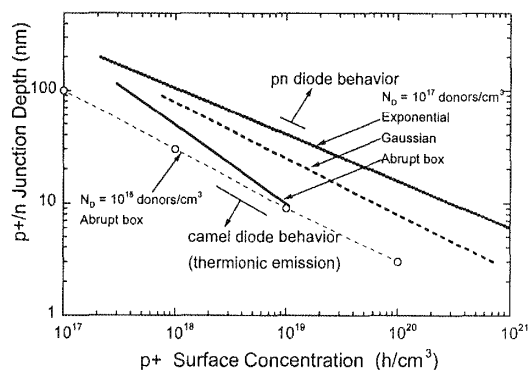


Fig. 10-43 The boundary between *pn* diode behavior and high-leakage currents from thermionic emission of metal contacts through various contact dopant profiles. Note that much shallower junction with *pn* diode behavior can be formed with abrupt (or "box profile") junctions than the more gradual Gaussian or exponential dopant profiles.⁹²

Modern CMOS processes use directly implanted profiles to set doping levels in the wells. The shape of a high-energy ion-implantation profile (which rises from a low concentration at the wafer surface to a peak deep within the Si), is the basis of the name "retrograde" well (see Fig. 16-23). This doping profile contrasts with the doping profile of earlier conventional CMOS processes, in which the well dopants were diffused more deeply into the substrate from their near-surface implanted locations. Such a drive-in diffusion following a near-surface well-implant step results in a maximum concentration at the wafer surface and a monotonic decrease in doping concentration as the well junction is approached. The directly implanted, retrograde doping profile provides much higher levels of protection from accidental well-to-well "latch-up" events in CMOS than is possible with diffused wells (see Vol. 2, Sect. 6.4 of this series for a detailed discussion of latchup in CMOS). In addition, in DRAM devices the risk of bit state loss (also called "soft errors" - see Vol. 2, Sect. 8.3.5 of this series for a description of soft errors in DRAMS) due to transient charges arising from ionizing radiation, is significantly reduced. The directly implanted processes also allow for separate doping of the channel region (for threshold voltage control) and the region beneath the channel (anti-punchthrough implant - for increasing the punch-through voltage - see Fig. 10-39). Details of punchthrough effects in MOSFETs are given in Vol. 3, Chap. 5 of this series. In addition to these device advantages, the use of high-energy implantation to replace the long-cycle well diffusion provides significant process time and wafer stress reduction.

High-energy implantation is also used to program ROM cells by changing the threshold voltage through direct implantation into the channel region. The implantation is done through the gate stack, after the source/drain and gate formation process have been completed. Programming ROMs late in the process cycle not only reduces product cycle time to completion but also provides an added degree of security, since the ROM code cannot be detected in the final devices by standard etching and imaging techniques. This method of programming ROM cells is used to produce high-security, chip specific, "scramblers" for data links between distributed microprocessors and data bases.

Chapter 16

CMOS PROCESS INTEGRATION

In the first fourteen chapters of this book the focus was on the individual process modules employed in IC fabrication. This chapter will consider how these modules can be combined to define a set of manufacturing steps that becomes a recipe for fabricating monolithic silicon integrated circuits. The task of specifying such a sequence of individual process steps is called *process integration*. *Complimentary MOS* (or *CMOS*) technology is used to demonstrate some examples of such *process sequences* (or *process flows*). CMOS was chosen for our examples because it is the dominant silicon IC technology of the 1990's. Figure 16-1 shows that about 80 percent of all semiconductor devices and circuits manufactured in 1995 were CMOS ICs. By the end of the 1990s this percentage grew even larger. Since this text is intended to provide an introduction to *mainstream* silicon IC process technology, it is appropriate to use CMOS as the vehicle for illustrating process flows. Since CMOS technology was preceded by other MOS technologies, their evolution will also be briefly reviewed. How and why CMOS became the dominant semiconductor technology will be presented. Note that CMOS is so-named because it uses both *p*- and *n*-type MOS transistors in its circuits. Figure 16-2 depicts a CMOS inverter.

In the examples of CMOS process integration two CMOS process flows will be described. The first is a generic process flow for the generations of CMOS logic ICs extending from 1.2–0.5 μm . The second pertains to CMOS ICs having minimum feature sizes of 0.25 μm and smaller. For such *deep-submicron CMOS technologies* (i.e., $\leq 0.25 \mu\text{m}$), several major changes occurred in the process flow and device architecture. Hence, the second process flow differs significantly from the first. Some of new processes used are: high-energy implanted retrograde wells, shallow-trench isolation, dual-doped polySi, TiSi_2 and CoSi_2 salicides, chemical-mechanical polishing, and formation of interconnects with damascene processing.

The purpose of this chapter is to present examples of how Si processes are integrated, resulting in the creation of fabrication sequences for CMOS ICs. A discussion of process integration involves the relationships between process technology and device physics. However, the operation of MOS devices is beyond the scope of this text, and is not presented here. Readers interested in more information on device physics should consult such texts as those by Chen,¹ Muller and Kamins,² and Tsividis³ (as well as Vol. 3 of this series).

16.1 INTRODUCTION TO CMOS TECHNOLOGY

Integrated circuits built with MOS transistors were introduced in the early 1960s, and these were fabricated using PMOS technology. However, it was known that since electrons have higher mobility than holes, ICs fabricated with NMOS would have better performance than those made with PMOS. By the 1970s, the technological barriers that blocked NMOS fabrication were overcome (see Vol. 2, Chap. 5 for details), and NMOS became the dominant technology for MOS IC fabrication. NMOS remained the main MOS IC technology until the integration level of devices on a chip became too large (i.e., the late 1980's), when CMOS took over.

NMOS is inexpensive to fabricate, very functionally dense and faster than PMOS. The earliest NMOS technologies required only 5 masking steps, including the pad mask.* On the other hand, NMOS logic-gates (e.g., inverters) consume dc power while in one of their two logic states. Therefore, at least some of the logic gates in an NMOS IC are always ON. Such circuits draw a steady current, even when being operated in the standby mode (when no signals are propagating through the circuit). Hence, as the number of logic gates on NMOS chips increased, the current being drawn (and hence the power dissipated) also rose. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a major drawback for most others. Such was the situation at the level of device integration that existed up to the mid-1970s. By the mid-1980's, with the dawning of the VLSI era (i.e., device counts on chips began to exceed 10^5), power consumption in NMOS circuits began to surpass tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.¹

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density and the twenty-fold increase in speed of ICs between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12- μm design rules to create six-transistor SRAM cells; each cell occupied 20,600 μm^2 . By 1998, SRAM cells using 0.25- μm design rules occupied less than 10 μm^2 . (The memory access time in these respective memory circuits decreased from 1 μs down to 10 ns.) By 1998 the decreased device dimensions and the attendant increase in chip size allowed 16-Mbit SRAMs to be built. To take another example: The Intel 4004 4-bit microprocessor (introduced in 1971), had 2300 devices and was built with PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips dissipating in excess of 5 W of power can be housed in ceramic packages. But such packages are expensive. To be able to use the much cheaper plastic packages, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power

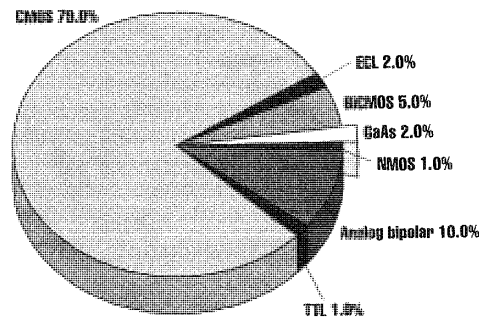


Fig. 16-1 Estimated market share of IC technologies in 1995. Reprinted with permission of Oxford Press.

* Early NMOS logic gates used butted contacts as well as enhancement-mode (E) transistors for both the driver and load. Depletion-mode loads (D) and buried contacts came later. Hence, of the seven masks of the E-D NMOS process, only five were required in early NMOS technology. This resulted not in lower manufacturing costs per wafer and higher yields, and thus ultimately cheaper NMOS circuits.

S. The
On the
eir two
t. Such
signals
hips in-
his was
nic sys-
he level
ning of
NMOS
loit the
t of the
l of ICs
11101)
0 μm^2 .
memory
by 1998
16-Mbit
reduced
rocessor
laced in
but such
aximum
power
both the
masks of
in lower

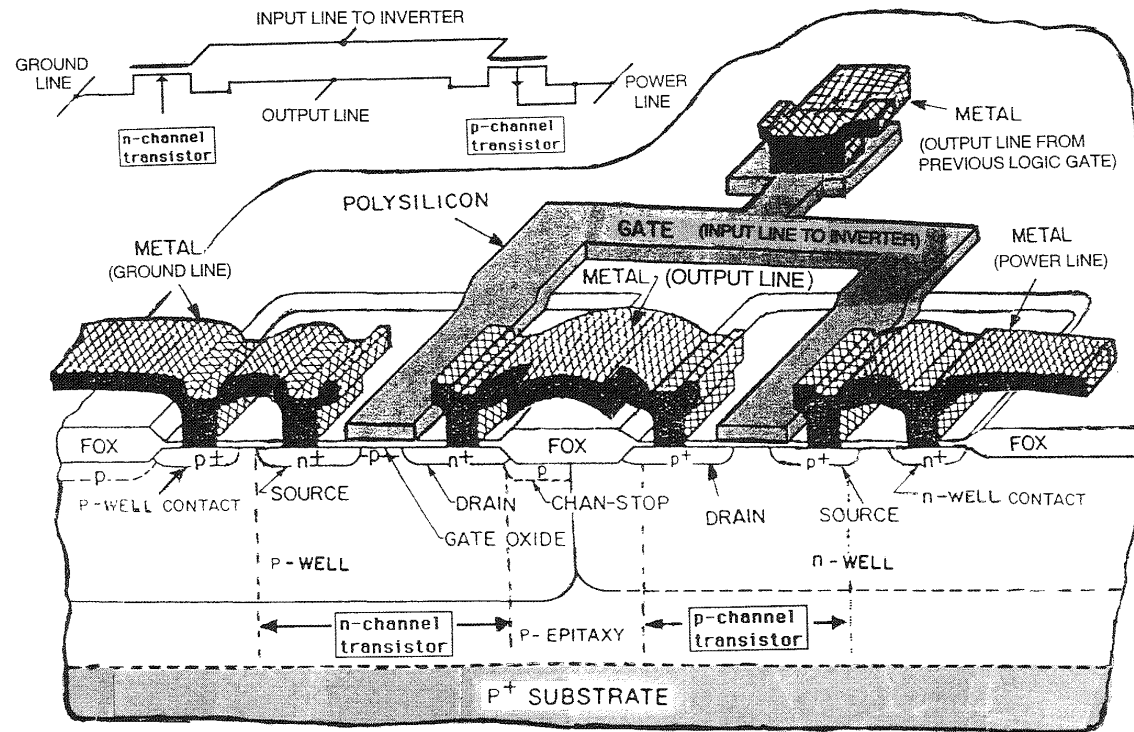


Fig. 16-2 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

when operated at 8 MHz. By the late 1970s it was clear that chips would very soon be manufactured with power dissipation values that were unacceptably high. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 0.25 W.)

Unlike in an NMOS inverter, in a CMOS inverter (Fig. 16-2) only one of the two transistors is driven at any one time (if the inverter is not in a transient switching state). This means that a high-impedance path exists between V_{DD} and ground, regardless of the state the inverter is in. Hence, virtually no current flows between the power and ground lines, and almost no dc power is dissipated by the CMOS circuit. Thus, logic gates consuming only microwatts of standby power can be implemented with CMOS (Fig. 16-3).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, microprocessors of the 64-bit generation are built using CMOS and yet some still dissipate more than 15 W of power. If only NMOS technology was available, the power dissipation would probably be an order of magnitude larger! Now consider memory chips from the system perspective. Although a 1-Mbit DRAM may consume 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be hundreds of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation at the system level are significant. Smaller power-supplies and smaller cooling fans are but two of these.

16.1.1 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, during its first 20-years it was considered to be a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power IC logic gates was originally proposed by

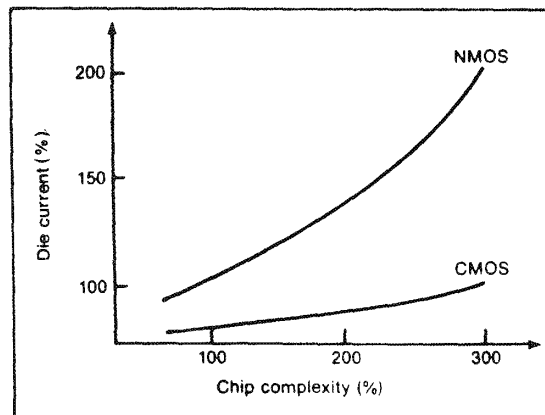


Fig. 16-3 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.

by Sah and Wanlass in 1963.⁴ The first CMOS ICs were fabricated in 1966. Subsequent development of the technology was spearheaded by the RCA Corporation. The earliest volume commercial application of CMOS was its use as a logic-inverter in the frequency divider circuits of digital watches.

CMOS technology at that time had many disadvantages compared even to PMOS and, later, to NMOS. Drawbacks included significantly higher fabrication cost, slower speed, susceptibility to latchup, and much lower packing density. As a result, until the 1980s CMOS was limited to applications that mandated either the lower power dissipation of CMOS (e.g., watches and calculators), or its very-high noise margin (e.g., radiation-hard circuits). Furthermore, the advances made in NMOS fabrication were not rapidly transferred to CMOS. For many years CMOS languished behind advanced Si-gate-NMOS and bipolar technologies. Except for the special applications mentioned above, it lay dormant for nearly two decades.

There were several reasons why this occurred. Recall that both *n*- and *p*-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite to that present in the starting material must be formed. These regions of opposite doping, called *wells* (or sometimes *tubs*) are the first features to be defined on a starting wafer. In the earliest CMOS technologies, a single-well approach was used. A lightly doped *n*-type wafer was the starting material, and only *p*-wells were formed in the wafer. This was called *p*-well CMOS. At the time these circuits were being fabricated (late 1960s), the processes of ion implantation and local oxide isolation (LOCOS) had not yet been developed. Metal gates were still being used in MOS devices and control had not been gained over the large and quite variable positive oxide charges in the gate oxide. These issues made *p*-well technology the only viable technique for fabricating CMOS.

While PMOS enhancement-mode transistors could be successfully fabricated in a lightly doped (e.g., 10^{15} cm^{-3}) *n*-substrate with a V_T of about -2 V, NMOS enhancement-mode transistors could not be fabricated on lightly doped *p*-substrates because the V_T of NMOS metal-gate transistors on such substrates is negative. In addition, it was likely that parasitic channels would be established in the field regions between NMOS devices built on lightly doped substrates. These parasitic channels could not be reliably suppressed because of three other factors: 1) the oxide charge at the silicon/ SiO_2 interface was large, and procedures for lowering it had not yet been discovered; 2) the segregation of boron at the field-oxide/silicon-substrate interface reduced the V_T of parasitic MOSFETs; and 3) the need to use relatively thin field oxides (to permit adequate step coverage of metal layers running over these oxide steps) also reduced the parasitic MOSFET V_T (see Vol. 2). Therefore, the only reliable way to manufacture enhancement-mode NMOS transistors for CMOS inverters without excess leakage was on regions with boron surface concentrations high enough to overcome these problems.

The *p*-well approach to building CMOS provided such regions, since the *p*-well had to be doped about ten times as heavily as the *n*-substrate for adequate control of doping in the well to be achieved. As a result, *p*-well technology became established in the companies that pioneered CMOS technology. Long after the problems of fabricating NMOS on lightly doped substrates had been solved (by controlling V_T with ion implantation, and the interface charge by annealing), most of these companies continued to use *p*-well technology to design new ICs.

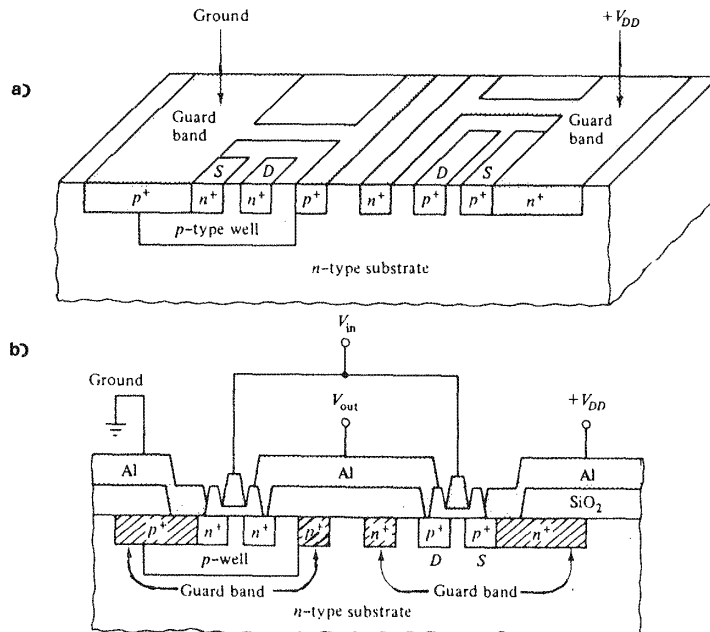


Fig. 16-4 CMOS structure with guard bands. A CMOS inverter circuit is depicted.⁵ From W. C. Till and J. T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication*. Copyright 1982 Prentice-Hall. Reprinted with permission.

The packing-density limits of the early *p*-well CMOS technology, however, were responsible for its poor performance. The packing density was much worse than that of NMOS, primarily because the *n*- and *p*-devices had to be surrounded by guard rings (*n*⁺ or *p*⁺ diffusions that surround the device, as shown in Fig. 16-3)⁵ to prevent the inversion of the field regions and latchup (see Vol. 2).

Before LOCOS isolation and ion implantation became available, guard rings were needed to provide adequately large values of V_T in the field regions. However, their use results in very large-area devices. Until oxide isolation and channel-stop implants were developed, interconnect capacitance and resistance severely degraded the speed performance.* Once guard rings were no longer needed, however, the devices could be brought closer together, and the speed of CMOS circuits improved dramatically.

When it finally became apparent in the late 1970s that the increases in power density and dissipation make it impossible to design future generations of circuits with NMOS,

* It should be noted that the fabrication of NMOS devices in heavily doped *p*-wells also increased the device junction-capacitances and reduced the magnitude of the drive current. This further decreased the performance of the circuits, but to a lesser degree than did the interconnect capacitance.

companies that had been stubbornly continuing to use NMOS for design and fabrication finally began to consider CMOS. It was natural for them to seek a technology that was compatible with the modern high-production-volume Si-gate-NMOS processes that they had successfully perfected. Since n -well CMOS (the other single-well CMOS type) offers near compatibility with such processes, and since it allows NMOS performance to be optimized (through their fabrication in the lightly doped p -substrate regions), it became the technology of choice for many companies that had formerly been manufacturing NMOS integrated circuits.⁶

Some time later, however, it became evident that neither p -well nor n -well would be the optimum choice for submicron CMOS. It was realized that *twin-well CMOS* would be more effective. As a result, many twin-well processes were subsequently developed. Examples of such twin-well CMOS process flows in Sects. 16.2 and 16.3 will be considered.

16.1.2 The Operation of CMOS Inverters

The CMOS inverter (for which the circuit schematic and a sample layout are shown in Fig. 16-2a and 16-5, respectively) uses enhancement-mode transistors for both the NMOS driver and the PMOS load transistors. (In enhancement-mode transistors the devices are *OFF* if the gate [or input] voltage is zero.) It can be seen that the gates of the two transistors in a CMOS inverter are connected, and serve as the input to the inverter. The common drains of each device are also connected to the inverter output. Assume the inverter is driving some load capacitance, C_L ,

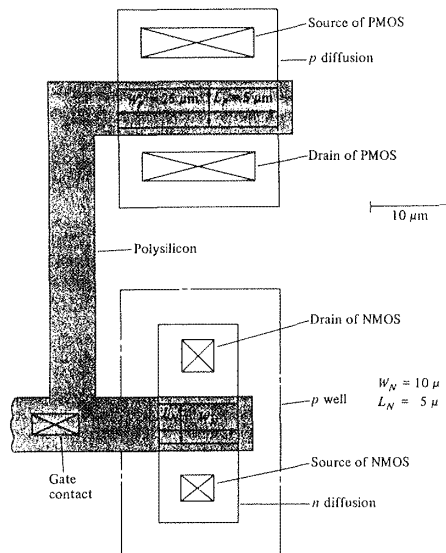


Fig. 16-5 CMOS inverter layout. From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright 1988, 2nd Ed., McGraw-Hill Book Co. Reprinted with permission.

(e.g., the input to another CMOS logic gate). Both the source and body of the NMOS transistor are connected to ground, while those of the PMOS transistor are connected to V_{DD} (e.g., 5 V).

The threshold voltage of the NMOS driver transistor V_{Tn} is positive (e.g., $V_{Tn} = 0.6$ V), while that of the PMOS load transistor V_{Tp} is negative (e.g., $V_{Tp} = -0.6$ V). Figure 16-6a shows the inverter's output voltage V_o as a function of the input voltage V_{in} . This curve is known as the *static voltage input/output characteristic*, or the *transfer characteristic*.

When $V_{in} = 0$, $V_{GSn} = 0$, the NMOS transistor is *OFF* (since V_{GSn} is < 0.6 V). The PMOS transistor, however, is *ON*, since $V_{GSp} = -5$ V, which is much more negative than -0.6 V. Thus, when $V_{in} = 0$, C_L is charged to V_{DD} through the turned-on PMOS load transistor, and $V_o = 5$ V.

When $V_{in} = 5$ V, the NMOS transistor is turned *ON*, since $V_{GSn} = 5$ V (which is > 0.6 V). The PMOS transistor is turned *OFF*, since $V_{GSp} = 0$ V (which is more positive than -0.6 V). Consequently, when $V_{in} = 5$ V (V_{DD}), the output is connected to ground through the turned-on NMOS driver, allowing C_L to be discharged. Since the PMOS device is off, C_L will be completely discharged, and V_o will be 0 V.

The most important property of the CMOS inverter is that when the gate is sitting quiescently in either logic state ($V_o = V_{DD}$ or 0), one of the transistors is *OFF* and the current conducted between V_{DD} and ground is negligible (i.e., it is equal to the leakage current of the *OFF* device). This feature can be seen in Fig. 16-6b, which plots the current through the inverter, I_{DD} , as a function of V_{in} (solid curve). The power dissipated in the static (or standby) mode is then determined by the product of the leakage current and the supply voltage. Since the leakage current of an MOS transistor being operated in cutoff is so small, very little power is consumed in the static mode. Another important feature is that the output voltage V_o swings all the way from V_{DD} to 0 as the inverter changes state. This switching characteristic is referred to as *swinging from rail to rail*.

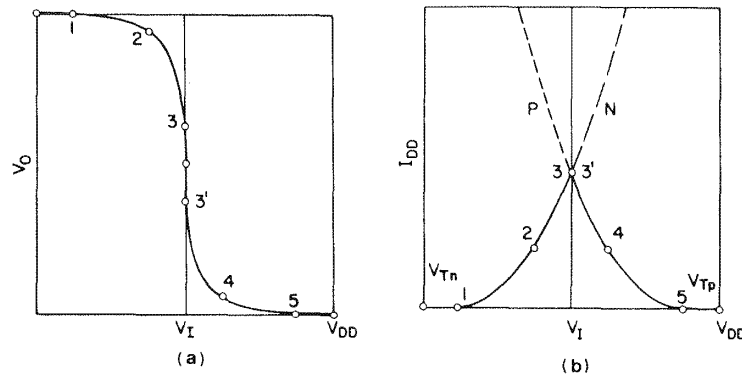


Fig. 16-6 (a) Output (V_o) versus input (V_i) voltage of CMOS inverter. (b) Current through inverter as a function of input voltage (solid curve); I-V characteristics of n - and p -channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.⁸

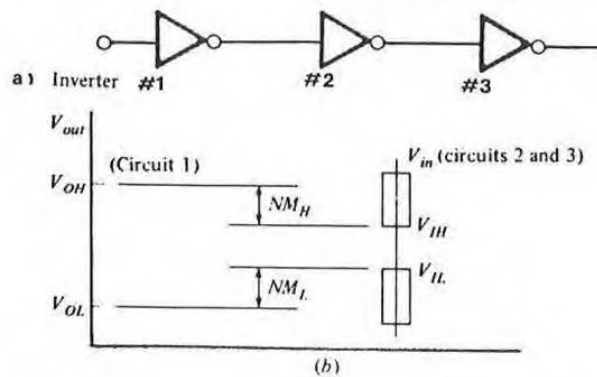


Fig. 16-7 (a) String of inverters connected in series. (b) Definition of noise margins.

Two operational features of the CMOS inverter must also be mentioned. First, as can be seen in Fig. 16-6b, significant current is conducted through the inverter only when both transistors are *ON* at the same time (i.e., during switching). Therefore, most of the power dissipation is due to the charging and discharging of C_L . In fact, it can be shown that the power dissipation is essentially given by $f C_L V_{DD}^2$, where f is the switching frequency. Because only a small fraction of the logic gates in an IC are likely to switch during any given clock cycle, CMOS power dissipation remains much smaller than in NMOS ICs.

Second, because the output voltage swings from rail to rail, the CMOS inverter inherently provides excellent noise margins. Noise margins are usually defined in terms of the *logic-gate threshold voltages* V_{OH} , V_{OL} , V_{IH} , and V_{IL} (which are *not* the same parameters as *device threshold-voltages*, see Fig. 16-7a). The noise margins NM_L and NM_H (Fig. 16-7b) are defined according to the following equations:

$$NM_L = V_{IL} - V_{OL} \quad (16.1a)$$

$$NM_H = V_{OH} - V_{IH} \quad (16.1b)$$

Briefly, the argument for Eq. 16-1a is that the logic gate (e.g., *Inverter 1* in Fig. 16-7a) should provide a maximum "low output" that is less than the maximum "low input" that the subsequent logic gate (e.g., *Inverter 2*) can accept without causing the output voltage of *Inverter 2* to be driven into the ambiguous portion of the transfer characteristic. The difference between V_{IL} and V_{OL} (the noise margin) then specifies how much noise voltage can be tolerated at the input node of *Inverter 2* before its output will be driven to a voltage value that is logically undefined.

In CMOS, V_{OL} approaches zero, and V_{OH} approaches V_{DD} . Because of the steep transition region in the transfer characteristic, V_{IL} and V_{IH} can be designed so that noise margins are on the order of $V_{DD}/4$ for expected process variations and operating temperatures. Thus, for a 5-V CMOS technology, V_{NML} can be 1.25 V, whereas in NMOS, V_{NML} is typically only 0.3 V.

For the performance of the logic gates to be maximized, the threshold voltages of the *p*- and *n*-channel transistors should be comparable (and, ideally, of equal magnitude). In addition, the threshold voltages should be as small as possible. The minimum values selected for V_{Th} and V_{Tp} will be determined by the subthreshold leakage current, I_{Dst} .

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 $\langle 100 \rangle$ -orientation. Typically, in twin-well CMOS the starting material is also either a lightly p -doped wafer (p -bulk wafer, Fig. 16-7a), or a heavily p -doped wafer on which a thin, lightly p -doped epi layer is grown (Fig. 16-7b). The concentration range of the p doping in both the bulk wafer and the p -epi layer is 8×10^{14} – $1.2 \times 10^{15} \text{ cm}^{-3}$, while the p^+ substrate beneath the epi layer is doped to $\sim 10^{20} \text{ cm}^{-3}$. 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 $\sim$ \\$100 US, while that of a 200-mm epi wafer was $\sim$ \\$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 break-down 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 Figs. 16-8 and 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

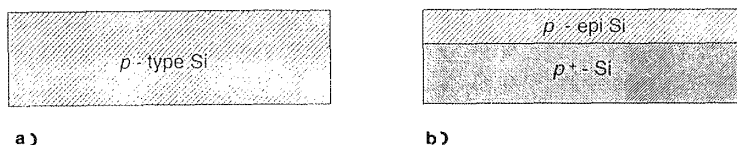


Fig. 16-7 Silicon starting substrates for twin-well CMOS: (a) p -bulk wafer; (b) p -epi-on- p^+ wafer.

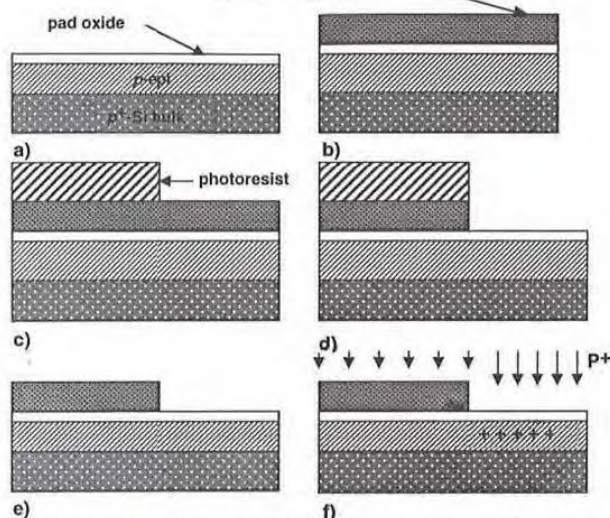


Fig. 16-8 Well formation in twin-well CMOS. (a) grow thermal pad oxide on p -epi-on- p^+ -bulk starting wafer; (b) deposit LPCVD silicon-nitride film; (c) use *Mask #1* to define n -well regions; (d) etch nitride; (e) strip resist (f) implant n -well dopants (phosphorus).

(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 $\sim 1 \times 10^{12} \text{ cm}^{-2}$, 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, $1 \times 10^{12} \text{ cm}^{-2}$, 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 $\sim 1 \times 10^{16}/\text{cm}^3$ for the p -well, and $\sim 3 \times 10^{16}/\text{cm}^3$ 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

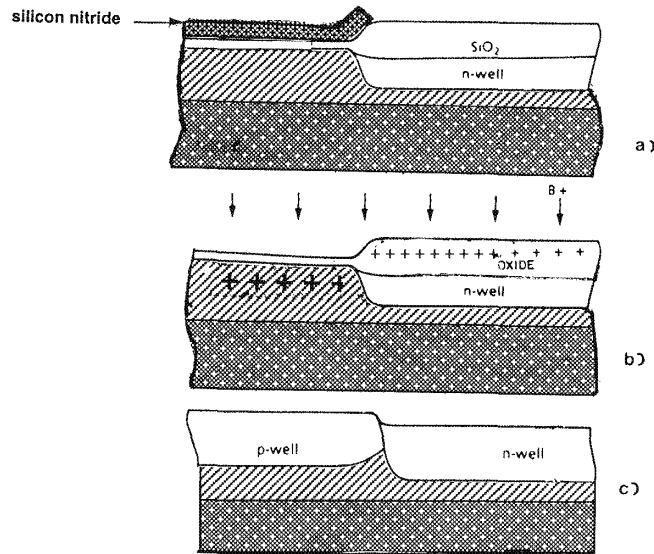


Fig. 16-9 (a) After the phosphorus that creates the n -well has been implanted, a thick oxide is grown, with the nitride preventing oxide from growing beneath it; (b) after stripping the nitride, a boron implant is performed that will form the p -well regions. The thick oxide above the n -wells prevents 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 doping, so the expected variations in epi doping will not significantly affect the well concentration. The junction depth of the wells is typically $\sim 2 \mu\text{m}$ 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-

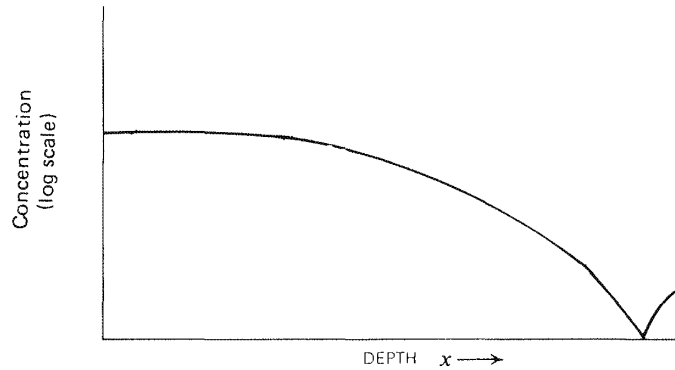


Fig. 16-10 Doping profile in a conventional well formed by a shallow implant and a drive-in step.

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/cm² 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 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

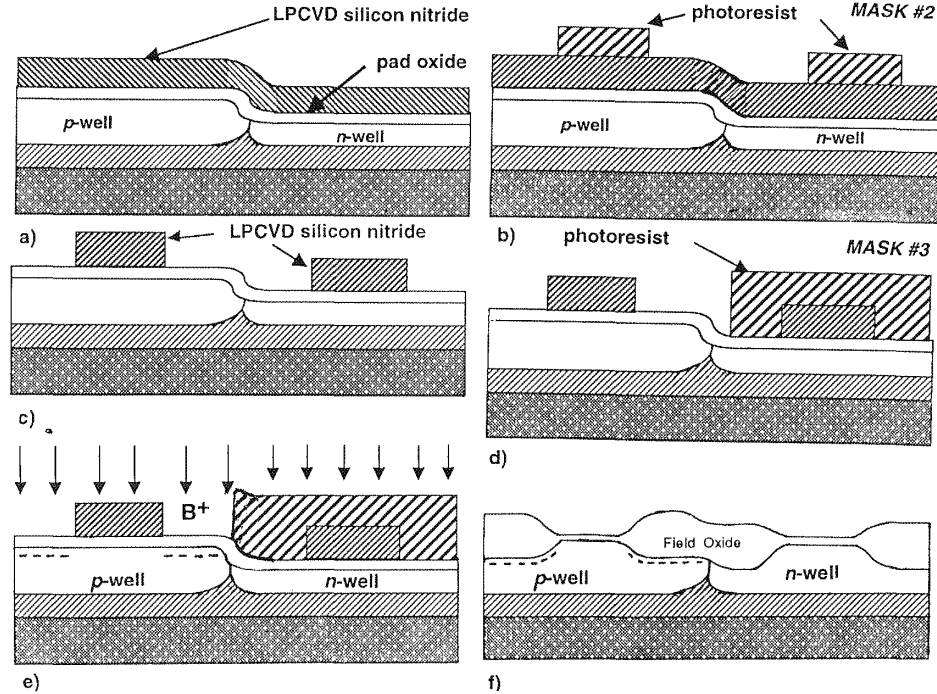
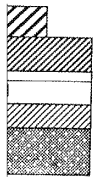


Fig. 16-11 Formation of the active and field regions using LOCOS isolation: (a) grow pad oxide and deposit silicon nitride film by LPCVD; (b) Use *Mask #2* to define active regions; (c) etch the nitride film; (d) Use resist and *Mask #3* to cover *n*-well regions; (e) implant B into the field-regions of the *p*-wells (channel-stop implant); (f) selectively grow field oxide with a wet-oxidation process, strip nitride.

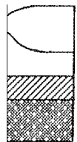
a thin layer of nitride forms on the silicon surface of the active regions (near the edges of the nitride film). Such layers would interfere with the gate oxidation process, insofar as the gate oxide will not grow as rapidly over such regions (and hence it would be thinner there). This problem is called the *Kooi effect*. To overcome it a thin *sacrificial* (or *dummy*) oxide is next grown to a thickness of 50–100 nm. This also oxidizes the thin nitride layer so it can be removed with an HF etch. The V_T -adjust implant is usually performed with this dummy oxide layer in place, as the oxide serves another useful function. That is, it also acts as a screen oxide, preventing contaminants from the implantation process from entering the Si substrate.

The V_T -adjust implant step (Fig. 16-13a) sets the threshold voltage of both the NMOS and PMOS FETs, typically to values of 0.6 V and –0.6 V, respectively. The V_T -adjust implant process in CMOS technologies using only n^+ poly can be accomplished with either a single boron implant step (if the background doping is chosen correctly, see Fig. 16-13b), or with two separate boron implant steps (in which case an additional mask is required). A separate punchthrough-prevention implant for the NMOS devices is also usually needed, especially when the minimum device dimension decreases below about 1 μm (see Vol. 3). The punchthrough

Mask #2



Mask #3



de and
e film;
r-wells

of the
e gate
. This
s next
be re-
oxide,

S and
nplant
single
th two
parate
when
ugh

Original Mask

CMOS PROCESS INTEGRATION 821

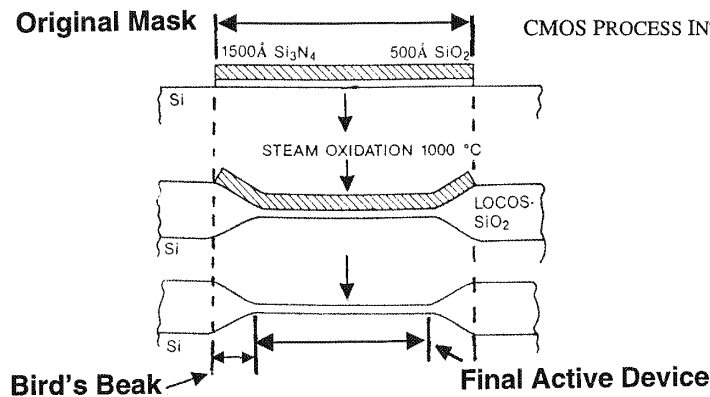


Fig. 16-12 "Bird's beak" in LOCOS isolation structure intrudes into the active regions of the wafer surface. This limits the use of LOCOS in deep-submicron IC process technologies.

implant can be performed sequentially using the same implanter, along with the V_T -adjust implant. Once the V_T -adjust implant step is completed, the dummy oxide is stripped with HF, leaving a residue-free Si surface upon which high-integrity gate oxides can be reliably grown.

16.2.5 Gate Oxide Growth: In the next step the gate oxide is formed. The growth of the gate oxide is a critical step. A defect-free, very thin (6-20-nm), high-quality oxide without contamination is essential for proper device operation. The gate oxide is grown only in the exposed active regions (actually, the field-oxide thickness is also very slightly increased as a result of the gate-oxide growth step). Since the drain current in an MOS transistor is inversely proportional to the gate-oxide thickness (for a given set of terminal voltages), the gate oxide is normally made as thin as possible – commensurate with oxide breakdown and reliability considerations. See Chap. 8 for details on thin gate-oxide growth.

16.2.6 Polysilicon Deposition and Patterning: Once the gate oxidation step is completed, a heavily n -doped polysilicon gate structure is fabricated. Polysilicon is the preferred gate material for several reasons. First, it can withstand the high-temperature steps required to form the source/drain junctions. Second, the polysilicon-SiO₂ interface is well understood and electrically stable. The gate formation procedure begins with the deposition of a 0.4–0.5 μm thick, undoped polysilicon film by LPCVD (Fig. 16-14a, and see Chap. 6 for details). This layer is then doped with phosphorus by ion implantation or chemical doping (Fig. 16-14b), producing a film with a sheet resistance of 20–30 Ω/sq . While this sheet-resistance value is small enough for MOS circuits with gate lengths $\geq 2 \mu\text{m}$, lower values are required for submicron ICs. In such applications, composite layers of refractory metal silicides and polysilicon (called *polycides*) are used to reduce the sheet resistance to ~ 1 –3 Ω/sq . This allows the benefits of the polysilicon-SiO₂ interface to be preserved and the gate sheet resistance to be lowered. For the CMOS processes being discussed here, CVD WSi_x became the most popular material used to form such polycide structures (see Chap. 6 for details on CVD WSi_x deposition).

The gate structure and polysilicon interconnect structures are then patterned using *Mask #4* (Fig. 16-14c). Following exposure and development of the resist, the polysilicon (or polycide) film is dry-etched (Fig. 16-14d). This is a critical etch step for several reasons. First, due to the

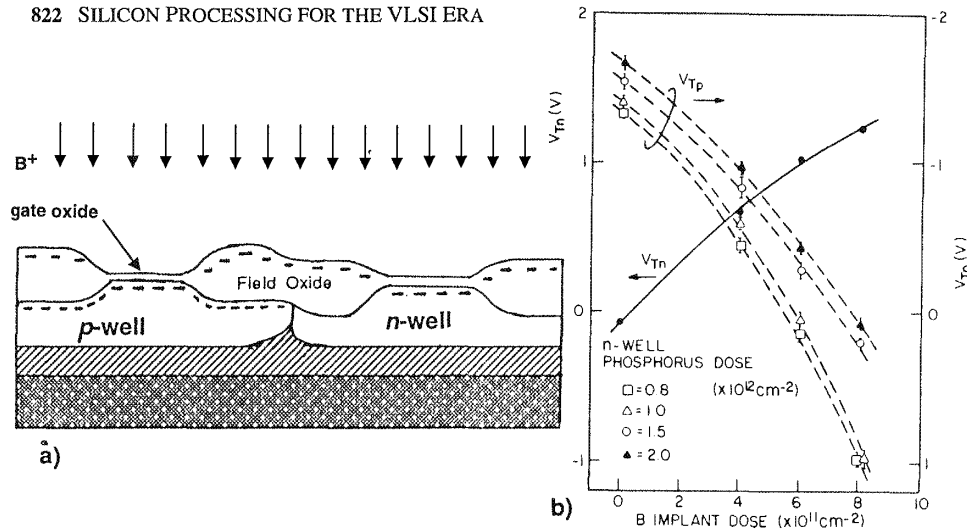


Fig. 16-13 (a) Threshold-voltage adjustment carried out using a single boron implant; (b) threshold voltages of NMOS (V_{Tn}) and PMOS (V_{Tp}) transistors as a function of boron dose. The CMOS structure uses an n -well implanted into a p -substrate (with doping level of 6×10^{14} atoms/cm³). V_{Tp} results are shown for various implant doses of the n -well.¹⁰ (© 1980 IEEE).

self-aligned nature of silicon-gate technology, the channel length of the device depends on the width of the polysilicon line. Hence, the gate-length dimension must be precisely maintained across the entire wafer, and from wafer to wafer. If the gate length is too long, the drain current of the MOSFETs will decrease, slowing down the performance of the IC. If the gate length is too short, the source and drain may punch through. Second, the profile of the etched poly gate structure should be vertical. This will prevent variation of channel lengths by the penetration of the ions of the thinner regions of the gate sidewalls during formation of the source/drain regions by ion implantation (Fig. 16-14e). Third, to achieve the above goals, an anisotropic polysilicon etch process must be employed. This process, however, requires overetching to remove the locally thicker regions of polysilicon that exist wherever it crosses steps on the wafer surface. During the overetch time, areas of the thin gate oxide are exposed to the etchants. Thus, it is necessary to use a polysilicon etch process that is highly selective with respect to SiO_2 .

After the polysilicon is etched, a step called *poly reox* is usually performed (16-14f). This grows a thin thermal SiO_2 layer on the polySi (on the order of 10 nm thick). The reasons for performing this step are discussed in Vol. 3, Chap. 9.

16.2.7 Formation of Source/Drain Regions: The next step in the process flow is the formation of source and drain (S/D) regions of the MOSFETs. Such regions are paths of current flow in the silicon between the metal interconnect lines and the channel of the transistor. As such, it is important that they have the lowest possible resistance. In addition, submicron MOSFETs require S/D junctions to be as shallow as possible to suppress such short-channel effects as *punch-through*, see Vol. 3. To obtain low resistivity, the S/D regions are doped as heavily as possible (typically using ion implantation, with a dose on the order of $\sim 10^{15} \text{ cm}^{-3}$). The NMOS S/D

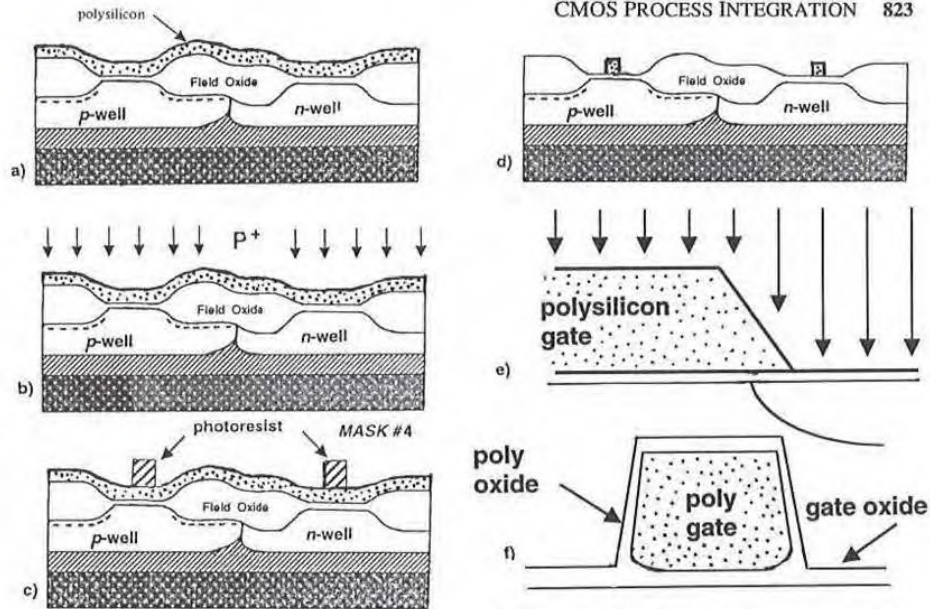


Fig. 16-14 n^+ -polysilicon gate formation: (a) deposition of undoped polysilicon film by LPCVD; (b) *ex-situ* doping of the polysilicon with phosphorus by chemical diffusion or ion implantation; (c) Use Mask #4 to define the polysilicon gates; (d) etch polysilicon and strip resist; (e) the etch process used to pattern the gate must yield poly features with vertical sidewalls (to maintain control of the effective gate length during the implant process that forms the source and drain regions); (f) polysilicon *re-ox* step grows a thin oxide on the poly by a short thermal oxidation process.

regions are doped with As, because it has a high solubility, low diffusivity, and a shallow projected range at the low ion implant energies that are used. Boron or BF_2^+ are used to dope the PMOSFET S/D regions. However, boron has a higher diffusivity than As. Hence, shallow S/D junctions in PMOSFETs are harder to achieve than in NMOSFETs. Since it is necessary to selectively implant source/drains to form n^+ regions for the n -channel FETs and p^+ regions for the p -channel FETs, this usually requires two masks (i.e., Mask #5 and Mask #6, Fig. 16-15).

In submicron-CMOS processes, gate lengths become so small that *lightly doped drain (LDD)* structures must be used to minimize hot-electron effects, especially in NMOS devices. Thus, procedures are integrated into the CMOS process flow to fabricate such LDD structures. If these structures are used only for NMOS transistors, Masks #5 and #6 will still allow the sources and drains of the two transistor types to be selectively implanted.* However, if LDD structures are needed for both PMOS and NMOS devices, two more masking layers may be needed.

* A procedure that requires only one masking step for creating both types of source and drain regions has also been reported. In this approach, the heavy boron implant is carried out non-selectively. The mask is then used to cover the PMOS device active regions. An n -type implant heavy enough to overcompensate for the boron implant in the active regions of the n -channel devices is used to form the sources and drains of the NMOS transistors.

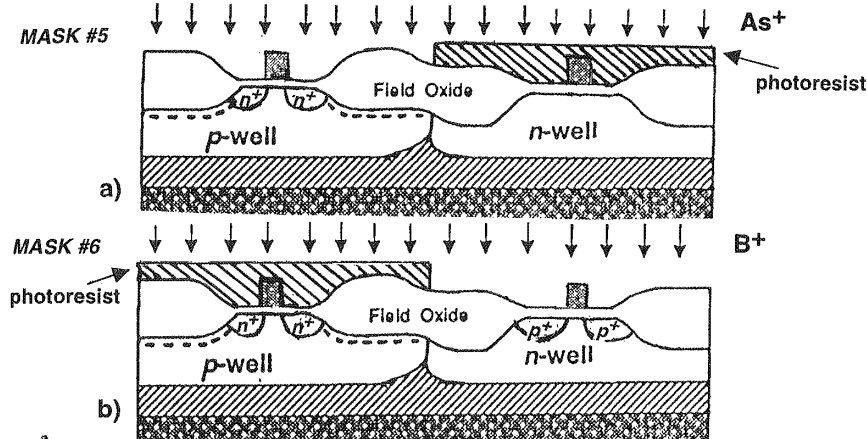


Fig. 16-15 Mask #5 and Mask #6 are used to allow selective implants of PMOS and NMOS source/drains.

The LDD structure is formed in the following way. The lightly doped regions of the source and drain are created with an ion implant step (Fig. 16-16b). A dose of approximately 3×10^{13} – 3×10^{14} cm^{-2} dopant ions is implanted at low energy (30–50 keV). The implantation process causes the edge of these implanted ions to be automatically aligned to the edge of the gate (i.e., it is a *self-aligned process*). A conformal layer of dielectric material (usually SiO_2 or silicon nitride) is then deposited over the entire wafer. An anisotropic etch process is used to clear the oxide in the flat areas while leaving sidewall spacers on the sidewalls of the poly gates (Fig. 16-16c). These spacers cover and protect the regions beneath them from the subsequent high-dose implant that forms the rest of the S/D regions. A dose of 1×10^{15} – 5×10^{15} atoms/cm^2 at 40–80 keV is used for this step (as shown in Fig. 16-16d).

16.2.8 Premetal Oxide Deposition and Contact Formation: Following formation of the source and drain regions, a doped *premetal dielectric* (PMD) is deposited by CVD. (This layer is also known as an *interlevel dielectric* or *ILD*.) Contact windows are opened in this dielectric layer to allow electrical connections to be made between Metal 1 and the following structures: 1) source/drain contact regions; 2) gate contacts; 3) substrate-contact regions; and 4) well-contact regions. A CVD process is used to deposit doped SiO_2 , about $1 \mu\text{m}$ thick (Fig. 16-17a), onto the wafers (see Chap. 6 for details of this process). The dopant in the SiO_2 is either phosphorus (in which case the material is referred to as *phosphosilicate glass* or *PSG*), or both phosphorus and boron (making it *borophosphosilicate glass* or *BPSG*).

The doped CVD glass layer plays several roles in the fabrication and operating aspects of the circuit. First, it acts as an insulating layer between polysilicon and Metal 1. Second, it reduces the parasitic capacitance between Metal 1 and the substrate. Third, the addition of phosphorus to the glass makes the layer an excellent getter of Na ions (contamination by Na can destabilize the V_T of an MOS device). The PSG (or BPSG) binds otherwise mobile Na atoms within the doped-glass layer, preventing them from reaching the gate oxide and altering V_T . Finally, the dopants in the glass make it viscous at elevated temperatures (1000–1100°C for PSG, 850–950°C for BPSG, see Ch. 6). This allows the layer of glass to be flowed after it is deposited (Fig. 16-17b).

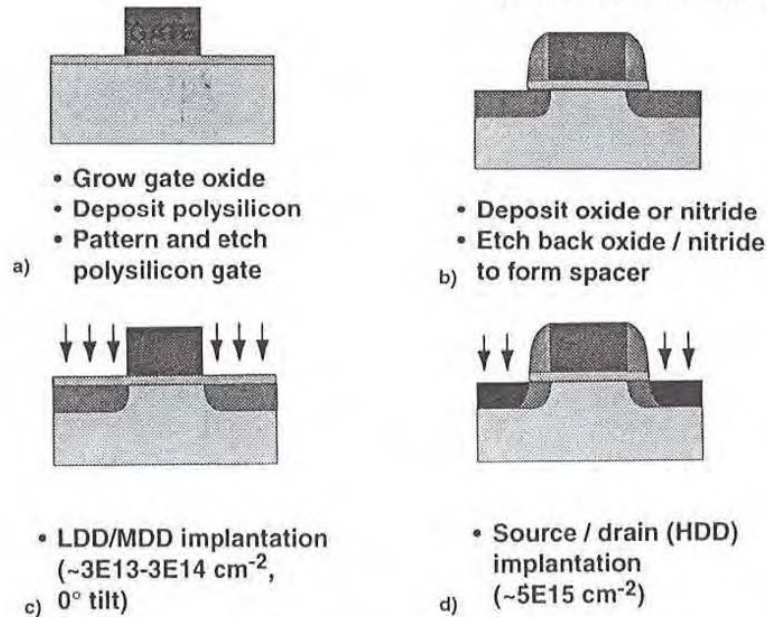


Fig. 16-16 Basic process to form an LDD NMOSFET: (a) after poly etch; (b) LDD implant; (c) after spacer material deposition and etch; (d) n^+ implant.

Through this procedure, a rounding of the glass contours and smoothing of any sharp steps is achieved. This produces better step coverage of the metal (which is deposited next) over the otherwise severe wafer topography. The high-temperature glass-flowing step also serves as an annealing process that activates the dopants implanted to create the source/drain junctions.

Contact openings are next created by a lithography-and-etch step (Fig. 16-17c). *Mask #7* is used to define contact patterns in a photoresist film. A dry-etch process is then used to open the contact windows through the interlevel dielectric to the underlying polysilicon and the source/drain regions in the silicon.

This contact-opening step can be critical, as the contact size and its alignment to underlying patterns limit the minimum size of the device. The source/drain regions must also be large enough for the contact to fit, with an allowance for alignment tolerance. If the contact opening exposes a part of the substrate, the drain or source will be shorted to it. Likewise, if the contact opening overlaps the both the source/drain and the gate, a short will be created between them.

After the contact etch is completed and the resist is stripped, the doped CVD glass is subjected to another flow step. This procedure rounds the corners of the top of the oxide windows so metal step coverage into the contact windows is improved (called *reflow* of the contact windows, see Chap. 6). It may still be too difficult to deposit metal into these contact holes with adequate step coverage. A number of techniques have been studied to improve the coverage of Metal-2 into the vias (Chap. 15). In Fig. 16-17d, the step coverage of Metal-1 into

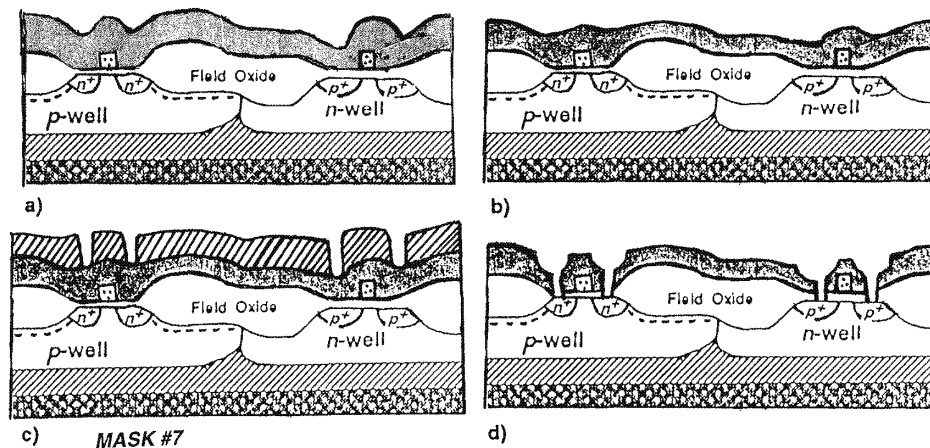


Fig. 16-17 (a) Deposit interlevel dielectric BPSG layer (ILD) by CVD; (b) flow BPSG. (c) use Mask #7 to define contact holes; (d) etch contact holes with "wine glass" etch and strip resist.

the contact hole is facilitated by the use of a *wine-glass* etch process. The contact hole is formed with a two-step via-etch process (i.e., using an isotropic etch process to etch half of the ILD thickness, and an anisotropic etch step to etch the remaining half). Reflow may still be carried out after such a two-step etch.

16.2.9 Metal Deposition and Patterning: After the contacts have been opened, the metallization layer is deposited. Because the metal layer is highly conductive, it is used whenever possible to interconnect circuit elements and to carry large amounts of supply current. The metal interconnect lines that are fabricated must have sufficient thickness, width, and step coverage to keep the current density in each line below the value that could produce electromigration failure (Vol. 2, Chap. 4). In addition, the space between adjacent metal lines must be kept large enough that the lines never touch, even under worst-case process variations.

Although evaporation was the method employed to deposit Al in the early days of MOS, it has generally been replaced by sputtering. The change was made because Al alloys with tightly controlled compositions became the materials of choice for the metal layer. Sputtering allows alloys to be deposited with much better compositional fidelity (see Chap. 11).

The metal alloy eventually chosen for NMOS was Al:1wt% Si. Silicon is added to the aluminum film to prevent spiking of the contacts during subsequent annealing steps (see Chap. 11). In CMOS, however, such Al:Si alloys were replaced with a metallization system that includes a diffusion barrier (typically Ti/TiN, see Chap. 15), and Al:Cu alloys for the main interconnect layer (Fig. 16-18a). Dry etching is used to pattern the interconnect films using Mask #8 (Fig. 16-18b).

If a single level of metal is used in the CMOS process, a sintering step occurs after the metal has been patterned. This step brings the metal and the n^+ and p^+ regions in the silicon into intimate contact. Such intimate contact between the metal and the heavily doped Si regions

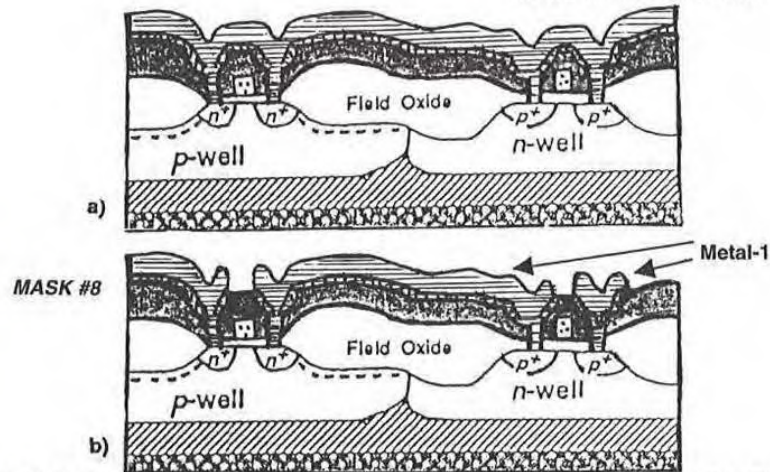


Fig. 16-18 (a) Deposit Ti/TiN/Al:Cu Metal-1 interconnect film. (b) Use Mask #8 to define Metal-1 pattern and then etch the metal with a dry-etch process.

establishes a low-resistance ohmic contact. The annealing process exposes the wafer to a 375–500°C temperature in an H_2 or $N_2 + H_2$ (5%) ambient for about 30 minutes. This step is also used as the annealing process for reducing the interface trap density in the gate oxide that was introduced by earlier processing steps (see Chap. 8).

16.2.10 Intermetal Dielectric Deposition and Via Patterning: If a two-level-metal interconnect structure is used, an *intermetal dielectric (IMD)* must be deposited to electrically isolate the Metal-1 layer from the Metal-2 layer (Fig. 16-19a). A variety of CVD processes for depositing such dielectric films have been developed (see Chap. 6).

Deposition of this layer may make the wafer topography once again too severe to allow Metal-2 films to be deposited with adequate step coverage. One of the planarization processes discussed in Chap. 15 may need to be used to overcome this difficulty.

If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical. It is also necessary to open vias in the intermetal dielectric layer so an electrical connection can be established between Metal-2 and Metal-1 at desired locations (Mask #9). Metal-2 must be deposited with adequate step coverage into these vias, but reflow is not possible because Al is present on the wafer (Metal-1). Thus, in this example, a champagne glass etch process is again used to etch the vias (Fig. 16-19b).

16.2.11 Metal 2 Deposition and Patterning: The processing issues of depositing and patterning (Mask #10) the Metal-2 layer (Fig. 16-19c) are discussed in Chaps. 11, 14, and 15.

16.2.12 Passivation Layer and Pad Mask: Finally, a *passivation* (or *overcoat*) layer, such as CVD PSG or plasma-enhanced CVD silicon nitride (or both), is put down onto the wafer surface (Fig. 16-19d). This layer seals the device structures on the wafer, protecting them from contaminants and moisture. It also serves as a scratch protection layer.