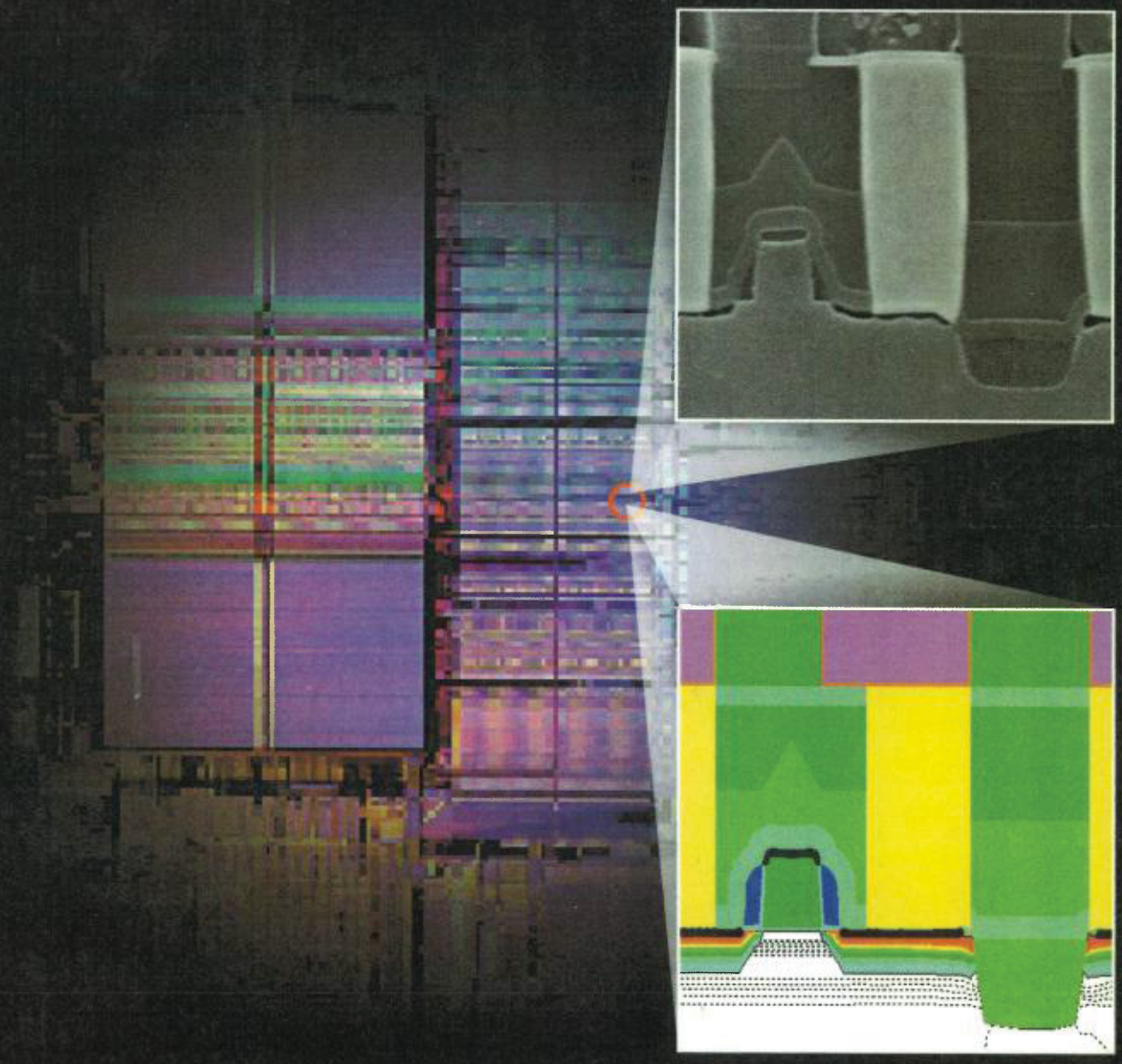
SILICON VLSI TECHNOLOGY

Fundamentals, Practice and Modeling



James D. Plummer • Michael D. Deal • Peter B. Griffin

Prentice Hall Electronics and VLSI Series—Charles Sodini, Series Editor

Silicon VLSI Technology

Fundamentals, Practice and Modeling

James D. Plummer
Michael Deal
Peter B. Griffin
Department of Electrical Engineering
Stanford University

Prentice Hall
Upper Saddle River, NJ 07458

Library of Congress Cataloging-in-Publication Data
Silicon VLSI technology
p. cm.
ISBN 0–13–085037–3
1. Integrated circuits—Very large scale integration—Design and construction. 2. Silicon. 3. Silicon oxide films. 4. Metal oxide semiconductors. 5. Silicon technology.
TK7874.75.S54 2000
621.39'5—dc21 99-42745
CIP

Publisher: Tom Robbins
Associate Editor: Alice Dworkin
Editorial/Production Supervision: Rose Kernan
Vice President and Editorial Director, ECS: Marcia Horton
Vice President of Production and Manufacturing: David W. Riccardi
Executive Managing Editor: Vince O'Brien
Marketing Manager: Danny Hoyt
Managing Editor: David A. George
Manufacturing Buyer: Pat Brown
Manufacturing Manager: Trudy Pisciotti
Art Director: Jayne Conte
Cover Design: Bruce Kenselaar
Editorial Assistant: Jesse Power
Copy Editor: Martha Williams
Composition: D&G Limited, LLC

Prentice Hall

Printed in the United States of America

10 9 8 7 6 5 4 3 2

ISBN 0-13-085037-3

Prentice Hall International (UK) Limited, *London*
Prentice Hall of Australia Pty. Limited, *Sydney*
Prentice Hall Canada Inc., *Toronto*
Prentice Hall Hispanoamericana, S.A., *Mexico*
Prentice Hall of India Private Limited, *New Delhi*
Prentice Hall of Japan, Inc., *Tokyo*
Pearson Education Pte., Ltd. , *Singapore*
Editora Prentice† Hall do Brasil, Ltda., *Rio de Janeiro*

Contents

Modern CMOS Technology

2

2.1 Introduction

In most of the remaining chapters in this book, we will discuss the process technologies used in silicon IC manufacturing individually. Individual technologies are clearly most useful when they are combined in a complete process flow sequence to produce chips. It is often the case that unit process steps are designed the way that they are because of the context in which those steps are used. For example, while a dopant may be diffused into a semiconductor to a desired final junction depth using many combinations of times and temperatures, the fact that the junction being formed might be diffused in the middle of a complex process flow may greatly restrict the possible choices of times and temperatures. In other words, the wafer's past history and the future process steps it may see can greatly influence how one chooses to perform a particular unit process step.

For this reason, and because we believe that understanding how complete process flows are put together aids understanding of individual process steps, we will describe in this chapter a complete modern Very Large Scale Integrated (VLSI) circuit process flow. The example we have chosen is typical of today's state of the art. CMOS technology has dominated silicon integrated circuits for the past 15 years and most people in the industry today believe that its dominance will continue for the foreseeable future for the reasons discussed in Chapter 1 (high performance, low power, supply voltage scalability, and circuit flexibility). In fact the SIA industry roadmap (NTRS) that we discussed in Chapter 1 assumes the continuation of CMOS technology through at least 2012.

For readers who are new to silicon technology, some of the ideas introduced in this chapter may not be fully appreciated until after later chapters on individual process steps have been read. However, we recommend that such people read this chapter before proceeding further because doing so will make the later material more understandable. A second reading of this chapter after the remainder of the book has been studied may also prove useful. In many cases, typical process conditions that might be used in a given step are presented in this chapter without full explanation. This is simply because we have not yet discussed the quantitative models and other tools at our disposal to calculate such parameters. As we do so in later chapters, we will revisit the CMOS process flow described here and discuss in more detail the reasons for particular process conditions used in this chapter.

2.2 CMOS Process Flow

Two typical CMOS circuits are shown in Figure 2–1. The simple inverter circuit on the left was described in Chapter 1. The NOR gate on the right illustrates how additional NMOS and PMOS devices can be added to the inverter circuit to realize more complex logic functions. In the NOR circuit, if either input 1 or input 2 or both of them are high, the output will be pulled to ground through one or both of the NMOS devices which will be turned on. Only if both inputs are low will the output be pulled high through the two series PMOS devices that are both turned on under this condition. The circuit thus implements the NOR function. To build these types of circuits, we need a technology that can integrate NMOS and PMOS devices on the same chip. In fact, many CMOS technologies also implement various types of resistors, capacitors, thin film transistors, and perhaps other types of devices as well. We will limit our discussion here to the two basic devices and describe a technology to build them. Extensions of this technology to include other components are reasonably straightforward and we will see some examples of such extensions in later chapters.

The end result of the process flow we will discuss is shown in Figure 2–2. To fabricate a structure like this, we will find that 16 photolithography steps and well over 100 individual process steps are required. The final integrated circuit may contain millions of components like those shown in the figure, each of which must work correctly.

There are two active device types shown in the figure, corresponding to those required to implement the circuits in Figure 2–1. The individual source, drain, and gate regions of the NMOS and PMOS devices are identifiable in the cross section. In addition to the active devices, there are many other parts to the overall structure. Some of this "overhead" is required to electrically isolate the active devices from each other. Other

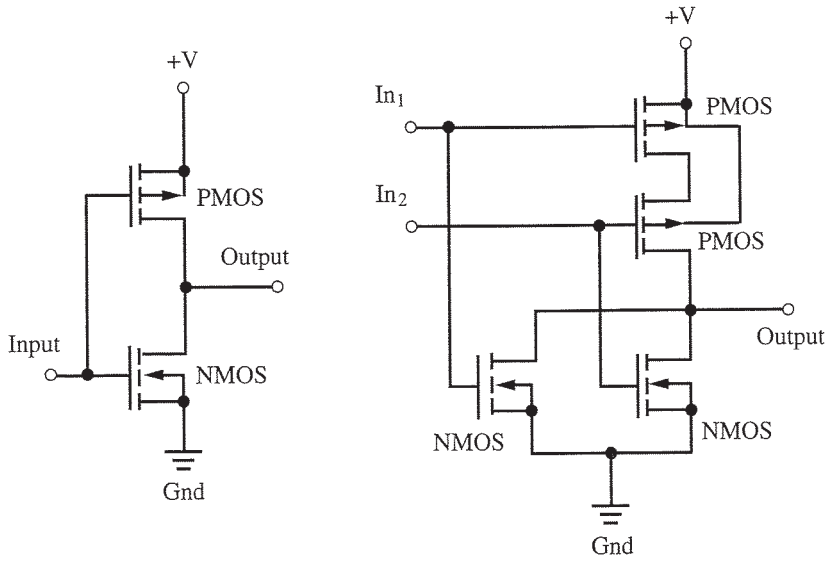


Figure 2–1 Simple CMOS circuits. An inverter is shown on the left and a NOR circuit on the right. The NOR circuit implements the function Output = $\overline{IN_1 + IN_2}$.

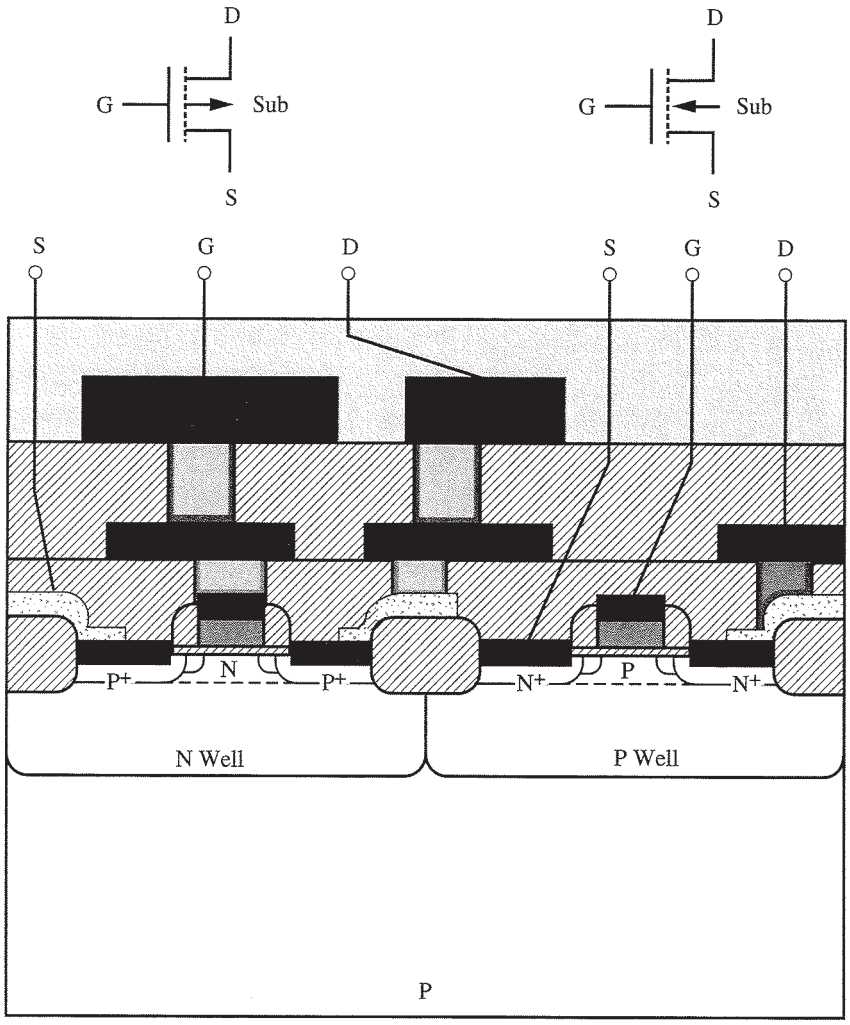


Figure 2–2 Coss section of the final CMOS integrated circit. A PMOS transistor is shown on the left, an NMOS device on the right.

parts of the structure provide multiple wiring levels above the active devices to interconnect them to perform particular circuit functions. Finally, some regions are included simply to improve the performance of the individual devices by decreasing parasitic resistances or improving voltage ratings. As we proceed through the steps required to build this chip, we will discuss each of these points in greater detail.

2.2.1 The Beginning—Choosing a Substrate

Before we begin actual wafer fabrication, we must of course choose the starting wafers. In general this means specifying type (N or P), resistivity (doping level), crystal orientation, wafer size, and a number of other parameters having to do with wafer flatness, trace impurity levels, and so on. The major choices are the type, resistivity, and orientation.

Figure 2–2 indicates that the final structure has a P-type substrate. In most CMOS integrated circuits, the substrate has a moderately high resistivity (25–50 Ωcm) which corresponds to a doping level on the order of 10^{15} cm^{-3} (Figure 1–18). As is apparent from Figure 2–2, the active devices are actually built in wells diffused into the surface of the wafer. The doping levels in these wells are chosen to optimize the electrical properties of the active devices, as we will see later in this chapter. Typically the well doping levels are on the order of $10^{16} - 10^{17}$ cm^{-3} near the wafer surface. In order to reproducibly manufacture such wells, the background doping (the substrate doping in this case) needs to be significantly less than the well doping. Thus the substrate doping is normally chosen to be on the order of 10^{15} cm^{-3}.

The observant reader might notice that the NMOS device could actually be built directly in the P substrate without adding the P well near the surface. In fact this is exactly the way the structure was sketched in Figure 1–34 for simplicity. While some CMOS circuits are actually built this way today, the twin well process illustrated in Figure 2–2 is much more common because the doping process used to produce the P well (ion implantation) is much better controlled in manufacturing than is the substrate doping. Also, since the P well and N well doping concentrations are on the same order, it is easier to start with a much more lightly doped substrate and tailor the wells for the NMOS and PMOS devices individually.

The observant reader might also note in Figure 1–34 that a substrate consisting of a P layer on a P^+ substrate was illustrated. This is one of the technology options we will consider later in Section 2.2.5.

The only other major parameter we need to specify in the starting substrate is the crystal orientation. We will discuss crystal structure in more detail in Chapter 3. However, virtually all modern silicon integrated circuits are manufactured today from wafers with a (100) surface orientation. The principal reason for this is that the properties of the Si/SiO_2 interface are significantly better when a (100) crystal is used. We will discuss the reasons for this in detail in Chapter 6, but the key idea is that the electrical properties of this interface are intimately connected with the atomic bonding between Si and O that takes place when an SiO_2 layer is thermally grown on Si. It is found experimentally that there are fewer imperfections (unsatisfied bonds) on a (100) surface than is the case on other silicon surfaces. Primarily for this reason, we will choose a (100) surface orientation for our starting wafers.

We will also discuss in Chapter 4 some processing which is often done on the starting substrates before any actual device fabrication is begun. This processing is aimed at minimizing the sensitivity of the wafers to trace contaminants that can be introduced to the wafers during the many manufacturing steps they go through to build circuits. These preliminary processing steps are called gettering and the most common process today is known as "intrinsic gettering." Since these steps are not essential to the device fabrication process, we will defer discussion of them to Chapter 4.

2.2.2 Active Region Formation

Modern CMOS chips integrate millions of active devices (NMOS and PMOS) side by side in a common silicon substrate. Circuits are designed with these devices to imple-

ment complex logic or analog functions. In designing such circuits, it is usually assumed that the individual devices do not interact with each other except through their circuit interconnections. In other words, we need to make certain that the individual devices on the chip are electrically isolated from each other. This is accomplished most often by growing a fairly thick layer of SiO_2 in between each of the active devices. SiO_2 is essentially a perfect insulator and provides the needed isolation. This process of locally oxidizing the silicon substrate is known as the LOCOS process (LOCal Oxidation of Silicon). The regions between these thick SiO_2 layers, where transistors will be built, are called the "active" regions of the substrate.

We begin with the steps shown in Figure 2–3. The wafers are first cleaned in a combination of chemical baths that remove any impurities from the surface. A thermal SiO_2 layer is then grown on the Si surface by placing the wafers in a high-temperature furnace. A typical furnace cycle might be 15 minutes at 900°C in an H_2O atmosphere. Although the H_2O ambient could be produced by boiling water, it is more common today to actually react H_2 and O_2 in the back end of the furnace to produce H_2O. This is generally a cleaner method for generating the steam required for the oxidation. The furnace cycle described above would produce an oxide of about 40 nm (400 Å). Such an oxide could also be grown in a pure O_2 ambient using a cycle of about 45 min at 1000°C. The oxide growth rate is much slower in O_2 compared to H_2O (Chapter 6), so higher temperatures and/or longer times are required in O_2 to grow the same oxide thickness.

The wafers are then transferred to a second furnace, which is used to deposit a thin layer of Si_3N_4 (typically 80 nm). This deposition occurs when reactants like NH_3 and SiH_4 are introduced into the furnace at a temperature of about 800°C, forming Si_3N_4 through a simple chemical reaction such as

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2 \tag{2.1}$$

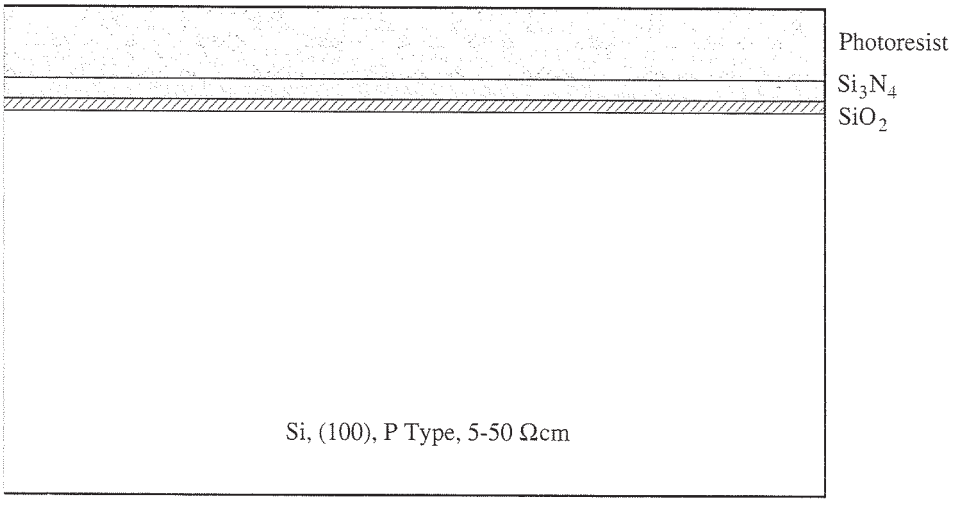


Figure 2–3 Following initial cleaning, an SiO_2 layer is thermally grown on the silicon substrate. A Si_3N_4 layer is then deposited by LPCVD. Photoresist is spun on the wafer to prepare for the first masking operation.

Generally this deposition is done below atmospheric pressure because this produces better uniformity over larger wafer lots in the deposited films. Pumps are normally used on the furnace exhaust to reduce the pressure. Systems in which such depositions are done are usually called Low-Pressure Chemical Vapor Deposition (LPCVD) systems. We will discuss them in more detail in Chapter 9.

The nitride layers deposited by such machines are normally highly stressed, with the Si_3N_4 under tensile stress. This produces a large compressive stress in the underlying Si substrate which can lead to defect generation if it is not carefully controlled. In fact, the major purpose of the SiO_2 layer under the Si_3N_4 is to help relieve this stress. SiO_2 layers are under compressive stress when they are thermally grown on Si and if the thicknesses of the SiO_2 and Si_3N_4 layers are properly chosen, the stresses in the two layers can partially compensate each other, reducing the stresses in the Si substrate. The thicknesses chosen above do this.

The final step in Figure 2–3 is the deposition of a photoresist layer in preparation for masking. Since photoresists are liquids at room temperature, they are normally simply spun onto the wafers. The resist viscosity and the spin speed determine the final resist thickness, which is typically about 1 μm. (Note that the dimensions in all the drawings in this book are not exactly to scale, since the photoresist layer in Figure 2–3 is really more than 10 times the thickness of the oxide or nitride layers, and the substrate is typically 500 times as thick as the photoresist layer. The liberties we take with scale in these drawings are intended to improve clarity.)

After the photoresist is spun onto the wafer, it is usually baked at about 100°C in order to drive off solvents from the layer. The resist is then exposed using a mask, which defines the pattern for the LOCOS regions. The photolithography process is both complex and expensive and was illustrated conceptually in Figure 1–9. We will describe it in much greater detail in Chapter 5. The machines which accomplish the exposure are often called "steppers" because they usually expose only a small area of the wafer during each exposure and then "step" to the next adjacent field to expose. Such machines must be capable today of printing lines on the order of 250 nm (0.25 μm) and placing these patterns on the wafer with an accuracy which is < 100 nm. They typically cost several million dollars.

The photoresists themselves are complex hydrocarbon mixtures. The actual ultra violet (UV) light-sensitive part of the resist is only a portion of the total mixture. In the case of a positive resist, which is the most common type today, the molecule in the resist which is sensitive to light, absorbs UV photons and changes its chemical structure in response to the light. The result is that the molecule and the resist itself then dissolve in the developing solution. Negative resists also respond to UV light but become insoluble in the regions in which they are exposed. Figure 2–4 shows our CMOS wafer after the resist has been exposed and developed.

An additional step is also illustrated in Figure 2–4. After the pattern is defined in the resist, the Si_3N_4 is etched using dry etching, with the resist as a mask. This is usually accomplished in a fluorine plasma. We will discuss dry etching in Chapter 10, but a typical reaction might involve the generation of F atoms in a plasma, using a CF_4 or NF_3 gas source and a reaction of the following type:

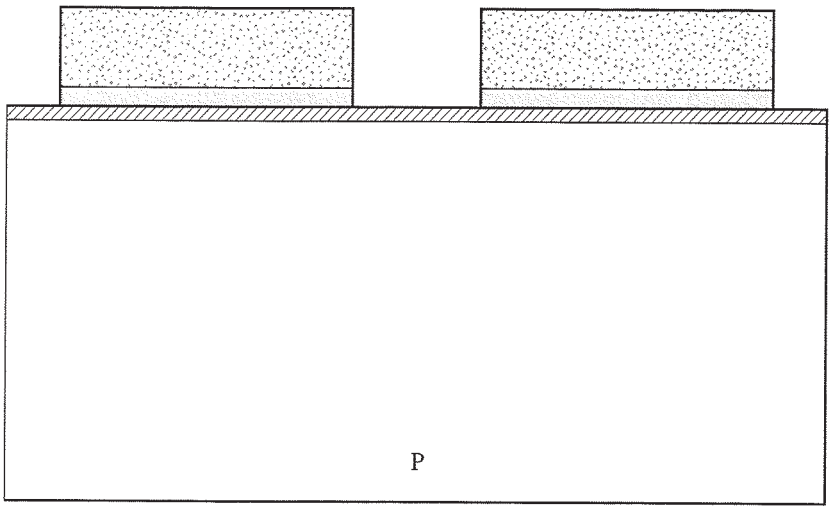


Figure 2–4 Mask 1 patterns the photoresist. The Si_3N_4 layer is removed where it is not protected by the photoresist by dry etching.

$$Si_3N_4 + 12F \rightarrow 3SiF_4 + 2N_2 \tag{2.2}$$

The most common type of etching system uses two parallel plates to confine the gas reactants. An RF voltage (usually 13.56 MHz) is applied across the electrodes to create a plasma and with it many neutral and charged molecules and atoms. It is important that the byproducts of the etching reaction be volatile at the etching temperature (usually room temperature), so that they can easily be pumped out of the reaction chamber.

Once the Si_3N_4 etching is completed, we are through with the resist and it can be chemically removed in sulfuric acid, or stripped in an O_2 plasma, neither of which significantly attacks the underlying Si_3N_4 and SiO_2 layers. Following cleaning, the wafers are then placed into a furnace in an oxidizing ambient. This grows a thick SiO_2 layer locally on the wafer surface. The Si_3N_4 layer on the surface prevents oxidation where it is present because Si_3N_4 is a very dense material and prevents the H_2O or O_2 from diffusing to the Si surface where oxidation takes place. This local oxidation or LOCOS process might be done at 1000°C for 90 min in H_2O to locally grow about 500 nm (0.5 μm) of SiO_2. The structure at this point is illustrated in Figure 2–5.

After the furnace operation, the Si_3N_4 layer can then be stripped. This is conveniently done in hot phosphoric acid, which is highly selective between Si_3N_4 and SiO_2. The Si_3N_4 could also be removed using dry (plasma) etching using a reaction like Eq. (2.2). However a process that gives good selectivity to SiO_2 would be required so that not very much of the LOCOS oxide is etched away during the stripping of the Si_3N_4. Selectivity is often a very important issue in etch steps throughout the wafer fabrication process. We will discuss this issue more carefully in Chapter 10.

An alternative to the SiO_2/Si_3N_4 stack used in Figure 2–3 is to use a three-layer stack of SiO_2, polysilicon, and Si_3N_4 to mask the oxidation process. This process is called the poly-buffered LOCOS process because of the incorporation of the polysilicon layer.

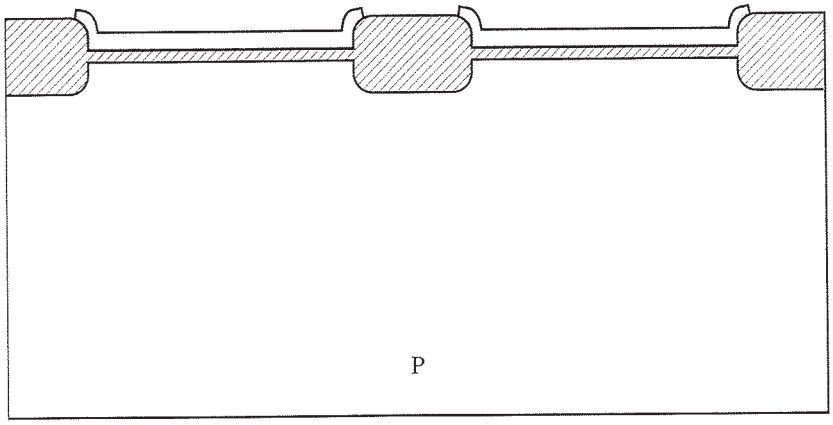


Figure 2–5 After photoresist stripping, the field oxide is grown in an oxidizing ambient.

As was the case in the LOCOS process, the function of the nitride layer is to block the oxidation from occurring wherever the Si_3N_4 is present. The underlying oxide and poly layers are both designed to help with the stress relief problem. Poly-buffered LOCOS uses a thicker Si_3N_4 layer than was the case in the LOCOS process (about 200 nm versus about 80 nm) and a thinner oxide layer (about 20 nm versus about 40 nm). The polysilicon layer permits these changes because it helps to relieve the large stresses, which would otherwise cause defects to form in the silicon substrate during the LOCOS oxidation. The polysilicon is deposited in an LPCVD machine similar to the one described for Si_3N_4 in connection with Eq. (2.1), except that only one reactant gas containing silicon is used (SiH_4 or SiH_2Cl_2, for example). A poly thickness of about 100 nm (0.1 μm) would be typical.

Why would we want to use a thicker nitride layer and a thinner pad oxide in the LOCOS process? The answer lies in a subtlety of LOCOS that we have not discussed to this point. Consider Figure 2–5 for example. The oxidation which takes place during LOCOS extends for some distance under the Si_3N_4 edge. The characteristic shape that this two-dimensional (2D) oxidation process produces is often called a bird's head or a bird's beak. This shape is only shown qualitatively in Figure 2–5; we will study this process in more detail in Chapter 6 and use numerical simulation tools to study the exact shape more carefully. The oxidation extends under the nitride edge because the oxidant (H_2O) can diffuse sideways as well as vertically through the pad oxide layer, to reach the silicon surface where it reacts to grow SiO_2. In fact, the nitride layer will bend up as oxide grows underneath it as Figure 2–5 qualitatively illustrates. This means that the oxide actually grows over a larger surface region than the mask pattern used to define the Si_3N_4. This is a major concern when we are defining very small active devices, because surface area that is lost to this encroachment of the oxide significantly decreases device density.

The answer to the question we posed then is that the combination of a thicker nitride, a thinner pad oxide which provides less of a pathway for lateral oxidant diffusion, and a polysilicon layer which itself can oxidize along its edges during LOCOS produces a much sharper transition between the oxidized and unoxidized regions. This allows for tighter design rules and higher device density.

2.2.3 Process Option for Device Isolation—Shallow Trench Isolation

We digress in our process flow description at this point to consider an alternative method for forming isolation regions between active devices. This alternative—Shallow Trench Isolation or STI—is beginning to be used in manufacturing today and will likely be the method of choice in the future. As we will see, STI actually etches trenches in the silicon substrate between active devices and then refills them with SiO_2. Such a process completely eliminates the bird's beak shape characteristic of LOCOS isolation and thus allows physically smaller isolation regions to be formed.

The process begins the same way as the LOCOS process. SiO_2 and Si_3N_4 layers are thermally grown and deposited respectively, as shown in Figure 2–3. The thicknesses of these layers are approximately the same as in the LOCOS process. However the stress-related issues, which tightly constrained these thicknesses in the LOCOS case, are relaxed somewhat in the STI process because there is no long high-temperature oxidation in STI, during which stresses can generate defects in the silicon substrate. Nevertheless, an SiO_2 thickness of about 10 – 20 nm and an Si_3N_4 thickness of about 50 – 100 nm would be typical. Photoresist is then applied, exposed, and developed as in Figure 2–4.

Figure 2–6 illustrates the next steps. The nitride and oxide layers are etched using the photoresist as a mask. This would typically be done as described above in Eq. (2.2), using a fluorine-based plasma chemistry for both materials. The next step is to etch the trenches in the silicon which are typically on the order of 0.5 μm deep. This can be done again using the photoresist as a mask, or the photoresists can be stripped and the nitride layer can then serve as the mask. The trench etching is a relatively critical step. It is important that the trench walls be relatively vertical so that there is little undercutting into the adjacent active device regions. However, the trench will later be filled with a deposited oxide and this filling process needs to completely fill the trenches without leaving voids. This often means that the trench walls should not be perfectly vertical but rather have a small slope. In addition, the top and bottom corners of the trenches ideally need to be slightly rounded in order to avoid problems later in oxidizing very sharp corners and to avoid electrical effects associated with very sharp corners. Thus the etch

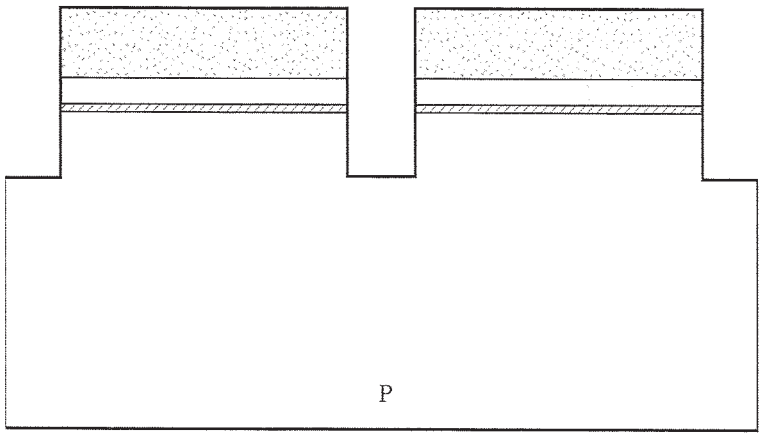


Figure 2–6 After mask 1 defines the photoresist, the Si_3N_4, SiO_2, and Si trenches are successively plasma etched to create the shallow trenches for isolation.

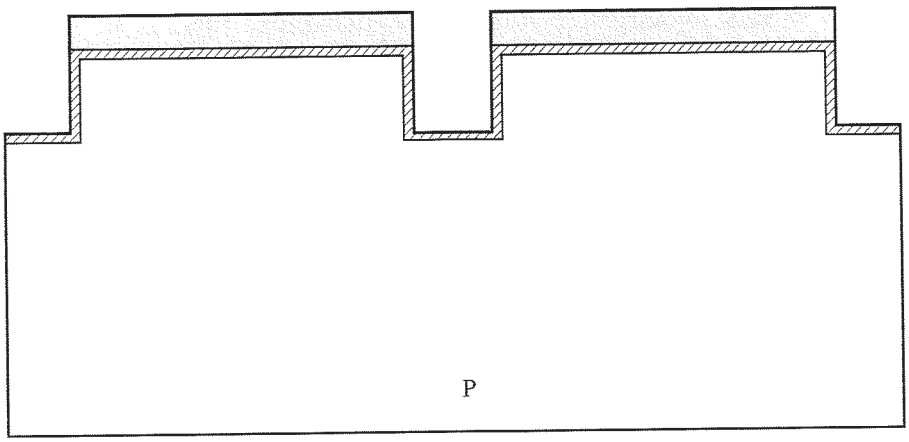


Figure 2–7 A thin "liner" oxide is thermally grown in the trenches. The nitride prevents any additional oxidation on the top surface of the wafer.

chemistry for this silicon etch must be very carefully chosen. Often a bromine-based plasma chemistry is used in this application, as discussed in more detail in Chapter 10.

The next step in the process is to thermally grow a thin (10 – 20-nm) "liner" oxide on the trench sidewalls and bottoms as shown in Figure 2–7. While most of the trench will be filled with a deposited oxide, thermally growing the first part produces a better Si/SiO_2 interface and if the oxidation is done at relatively high temperature ($\approx 1100°C$), the process will also help to round the corners of the trenches as well. The better Si/SiO_2 interface results from the lower electrical charge densities that thermal oxidations can produce. The corner rounding results from the viscoelastic flow properties of SiO_2 at high temperatures. We will discuss these issues in Chapter 6.

The next step, illustrated in Figure 2–8, is the deposition of a thick SiO_2 layer by chemical vapor deposition. It is important here that the filling process not leave gaps or voids in the trenches, which could happen if the deposition closed the top part of the trench before the bottom parts were completely filled. A number of deposition systems exist which do a good job of filling structures like this. One example is a High-Density Plasma or HDP system, which could be used in this application. We will discuss these systems in Chapter 10.

The final step in the STI process is illustrated in Figure 2–9. This involves literally polishing the excess SiO_2 off the top surface of the wafer, leaving a planar substrate with SiO_2 filled trenches. This polishing process uses a technique known as Chemical-Mechanical Polishing or CMP, which we will discuss in Chapter 11 since the most common use of CMP today is in back-end processing. In this process, the wafer is placed face down in a polishing machine and the upper surface is literally polished flat using a high-pH silica slurry. While this process sounds crude compared to the sophisticated processing techniques generally used to fabricate chips, CMP has been found to work extremely well and it has found widespread application. The nitride layer serves as a polishing stop and once the CMP operation is complete, the Si_3N_4 can be chemically removed as in the LOCOS process described earlier.

At this stage, the wafers are ready for device fabrication in the active regions. A comparison of Figures 2–5 and 2–9 illustrates both the similarities and the differences

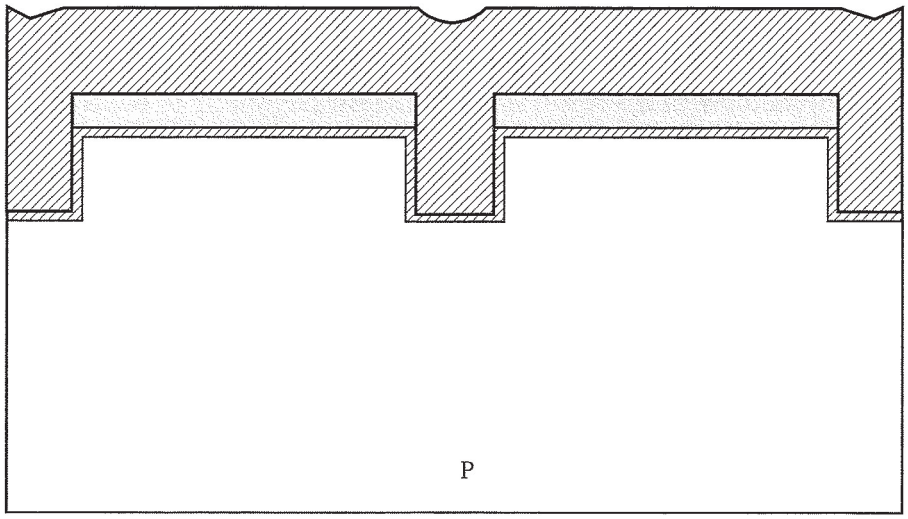


Figure 2–8 SiO_2 is deposited to completely fill the trenches. This would typically require 0.5 – 1 μm of SiO_2 to be deposited, depending on the trench depth and geometry.

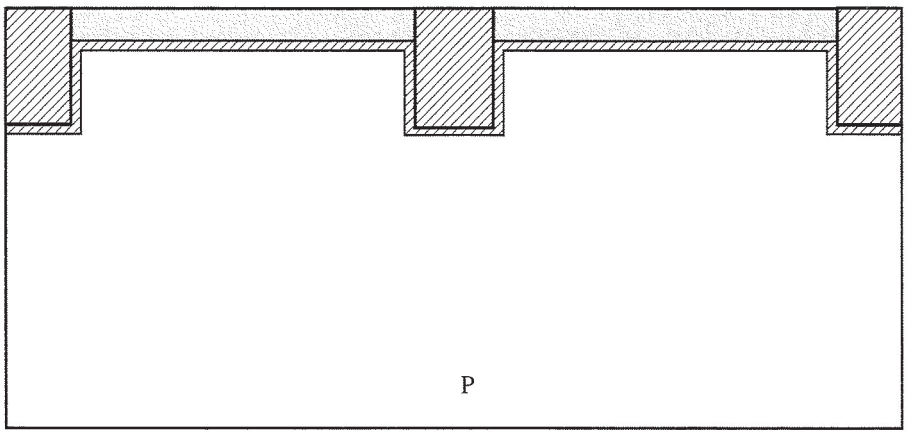


Figure 2–9 The deposited SiO_2 layer is polished back using CMP to produce a planar structure.

in LOCOS and STI. Both processes produce thick SiO_2 regions laterally isolating adjacent device structures. However the STI process produces more compact structures because there is very little lateral encroachment of the isolation structure into adjacent active regions.

One might ask why the STI process has not been used in manufacturing until quite recently since it seems like a simple process and it produces more compact isolation regions compared to LOCOS. There are really two answers to this question. The first is that when device geometries were larger, the small area loss due to lateral encroachment in the LOCOS process was not a significant factor in overall chip density. The second reason is perhaps more important. While STI seems like a simple process, there are actually a number of subtle issues associated with this technology which have proven difficult to solve in a manufacturing environment. These issues include filling the trenches with no gaps, using CMP to planarize the wafer, and avoiding subtle electrical

effects associated with the corners of the trenches which can affect device isolation. These issues have now been largely solved with new processes and new manufacturing equipment, with the result that STI is now beginning to be used in manufacturing. Its inherent density advantage over LOCOS suggests that STI will dominate in the future.

2.2.4 N and P Well Formation

We now return to our CMOS process flow and we will pick up the LOCOS isolation technology where we left it in Figure 2–5. If STI were used as the isolation process, the steps below would be largely unchanged. We would simply use the cross section in Figure 2–9 rather than Figure 2–5 as our starting point to continue the process flow.

In the final device cross section in Figure 2–2, the active devices are shown in P- and N-type wells. These wells tailor the substrate doping locally to provide optimum device characteristics. The well doping affects device characteristics such as the MOS transistor threshold voltage and I-V characteristics and PN junction capacitances. For example, recall in Chapter 1, Eqs. (1.23) – (1.25) which describe the electrical properties of PN diodes and notice the presence of the doping levels N_D and N_A in these expressions. The steps required to form the P and N wells are illustrated in Figures 2–10 to 2–12.

In Figure 2–10, photoresist is spun onto the wafer and mask 2 is used to expose the resist and to define the regions where P wells are to be formed. The P regions are created by a process known as ion implantation, which we will discuss in Chapter 8. The machines which perform this step are really small linear accelerators. A source of the ion to be implanted (boron in this case) is provided, usually from a gas. Positively charged ions (B^+) are formed by exposing the source gas to an arc discharge. The ions

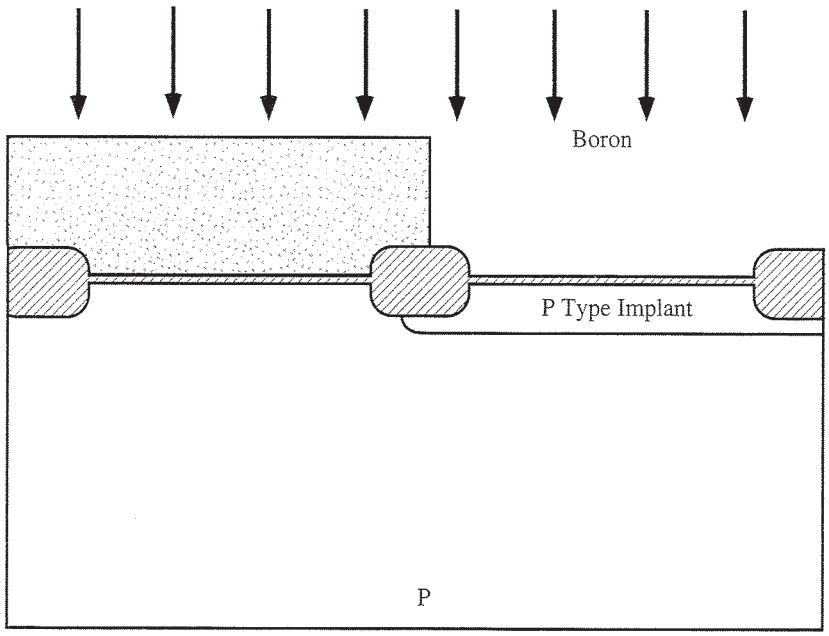


Figure 2–10 Photoresist is used to mask the regions where PMOS devices will be built using mask 2. A boron implant provides the doping for the P wells for the NMOS devices.

are then accelerated in an electric field to some final energy, usually expressed in keV. Since many types of ions may be formed in the source region, all ion implanters select the particular ion to be implanted by bending the ion beam through a magnetic field. Ions of different masses will bend at different rates in the magnetic field, allowing one type of ion to be selected at the output by adjustment of the field strength. Once the selection process is complete, final acceleration of the B^+ takes place along with either electrostatic scanning of the beam or mechanical scanning of the wafer to provide a uniform implant dose across the wafer.

In our case, we would need to pick an implant energy sufficiently large that the B^+ ions penetrated the thin and thick SiO_2 layers on the wafer surface, but not so large that the beam penetrated through the photoresist which must mask against the implant. This is possible in this case because the field oxide is on the order of 0.5 μm and the photoresist is at least twice this thick. The B^+ implant needs to penetrate through the thin SiO_2 layer in order to form the P well. It also needs to penetrate through the field oxide although the reason is not so obvious in this case. As was pointed out earlier, the purpose of the field oxide is to provide lateral isolation between adjacent MOS transistors. If the doping is too light under the field oxide, it is possible that surface inversion can occur in these regions, providing electrical connections between adjacent devices through parasitic MOS devices (field oxide transistors). By ensuring that the well implants penetrate through the field oxide, the doping is increased under the field oxide, preventing this parasitic inversion problem. We will study ion implantation in detail in Chapter 8 and see how to choose the accelerating energy for the B^+ ions, but for the situation described here, an energy of 150 – 200 keV would be typical.

The amount of B^+ we implant (or the dose), is determined by the device requirements. Here we are forming a P well whose concentration is required to be on the order of 5×10^{16} to 10^{17} cm^{-3} in order to provide correct device electrical characteristics. In Chapter 8 we will see how to calculate this dose, but a dose on the order of 10^{13} cm^{-2} would be typical. (Note that implant doses are expressed as an areal dose per cm^2, while doping concentrations are volume concentrations per cm^3.)

An important point about ion implantation has not been made to this point. Implantation of ions into a crystalline substrate causes damage. This is easy to visualize since an incoming ion with an energy of perhaps 100 keV can clearly collide with and dislodge silicon atoms in the substrate which have a binding energy of only 4 Si-Si bonds (about 12 eV). Visualize a billiard-ball-like collision between the incoming 100 keV B^+ ion and a stationary Si atom and you can easily imagine that the silicon atom will likely be recoiled a significant distance from its original lattice site. In fact many such recoils are produced as the B^+ ion gradually comes to rest. This damage must be somehow repaired since the devices we want to end up with require virtually perfect crystalline substrates. Fortunately, the repair process is not as difficult as it might initially seem. A simple furnace step usually suffices. Heating the wafers allows the dislodged silicon atoms to diffuse and find a vacant lattice site, thus repairing the damage. Such a high temperature step will soon occur in our process description and will accomplish this crystal repair function. This can be done in short times at high temperatures (e.g., 10 sec at 1000°C) or in longer times at lower temperatures (e.g., 30 min at 800°C). We will study implantation and damage annealing in greater detail in Chapter 8.

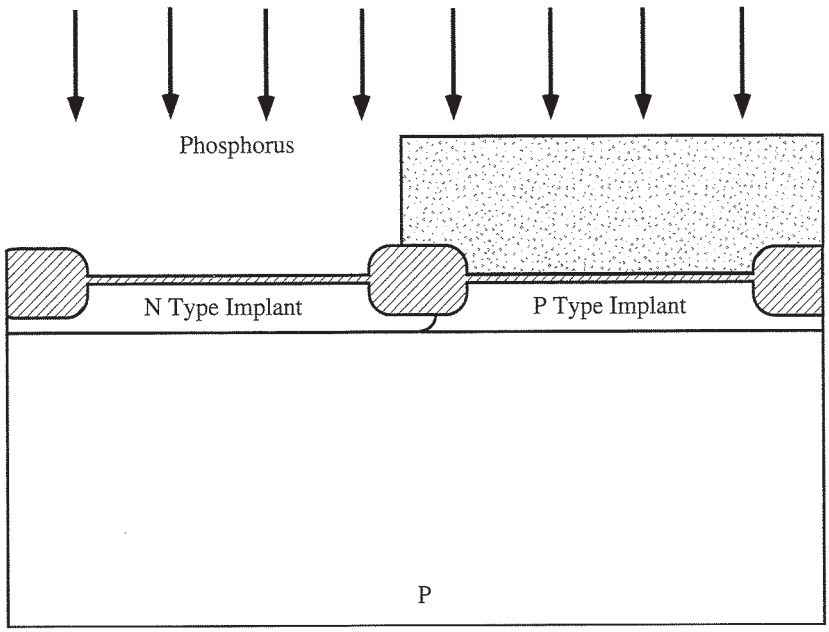


Figure 2–11 Photoresist is used to mask the regions where NMOS devices will be built using mask 3. A phosphorus implant provides the doping for the N wells for the PMOS devices.

Once the boron implant is complete, we are finished with the photoresist and it is then stripped either chemically or in an O_2 plasma. Photoresist and mask 3 are now used as shown in Figure 2–11 to define the regions where N wells will be placed in the silicon. The process is identical to that just described for the P wells except that in this case an N-type dopant, phosphorus, is implanted. The energy of the phosphorus implant is again chosen to penetrate the oxide layers but not the photoresist. Phosphorus is a heavier atom than boron (atomic mass = 31 versus 11), so a higher energy is required to obtain an implant to the same depth into the silicon. In this situation an energy of 300–400 keV would be chosen. The dose of the phosphorus implant would typically be on the same order as the boron P well implant since the purpose is similar in both cases.

There are several common N-type dopants available for use in silicon (phosphorus, arsenic, and antimony) and yet we specifically chose phosphorus in this case. The next step in the process is to diffuse the P and N wells to a junction depth of typically several microns, as illustrated in Figure 2–12. Boron and phosphorus have essentially matched diffusion coefficients and so they will produce wells with about the same junction depth when they are simultaneously diffused. The other N-type dopants, arsenic and antimony, both have much smaller diffusion coefficients and so for the process described here, the N well would be much shallower than the P well, which is not desired if we want matched NMOS and PMOS characteristics. Another issue with arsenic and antimony in this particular step in the process is that they are much heavier atoms than phosphorus and hence would require much higher implant energies.

After the phosphorus implant, the photoresist is removed and the wafers are cleaned. They are next placed in a drive-in furnace, which diffuses the wells to a junction depth of 2–3 microns (Figure 2–12). (Actually the depths they reach in this step will

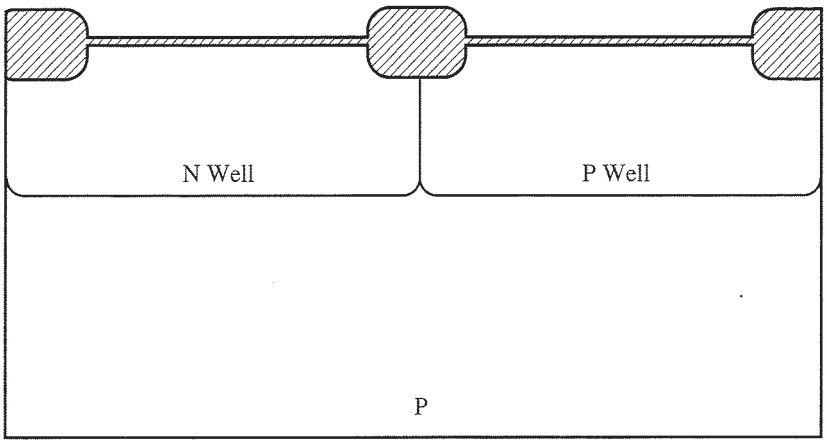


Figure 2–12 A high temperature drive-in completes the formation of the N and P wells.

not be their "final" depths because all subsequent high temperature steps will continue to diffuse the dopants. However, later high temperature steps will generally be either at lower temperatures or for shorter times, so that most of the well diffusion occurs during this drive-in process.) A typical thermal cycle might be 4 to 6 hours at 1000 to 1100°C. Diffusion coefficients increase exponentially with temperature as we will see in Chapter 7, so much shorter times are required at higher temperatures to achieve the same junction depth. This step could be performed in a largely inert ambient because no additional surface oxidation is needed at this point. The well drive-in step also repairs the damage from the implants, restoring the substrate crystallinity.

2.2.5 Process Options for Active Region and Well Formation

At this point we have completed the preparation of the substrate for the fabrication of the active devices. There are many process options which could be considered at almost every stage of our CMOS process. In general we will not consider very many of these options in this chapter, because there are simply too many to consider in a finite chapter. Also, our purpose here is not to explore all options but simply to give the reader a sense of what an integrated process flow looks like. However in addition to the STI option we considered earlier, there are several other options that are very commonly used in industrial manufacturing which will be useful for the reader to understand before reading later chapters. Two such options are explored briefly in this section before we return to our CMOS process flow as it is shown in Figure 2–12.

Option 1: Field Implants under LOCOS Regions

The first process option relates to the field oxide or LOCOS regions which provide lateral electrical isolation between adjacent devices. In the process flow we have described to this point, the implant energies of the P and N wells were carefully chosen to penetrate through the thick field oxide so that the substrate doping was increased under the

LOCOS regions. In practice this is not as simple to do as it might seem. Implanted ions are characterized by a range distribution not by a specific distance they travel. Since the stopping process is statistical in nature, we should expect such a range profile. Chapter 8 explores this idea in more detail, but for our purposes here we need only to understand that the entire implant dose cannot be placed in the silicon under the field oxide in Figures 2–10 and 2–11. If we tried to do this, the implant energy would have to be increased significantly to make certain that the shallowest ions went far enough to get through the field oxide. But the problem would then be that the deepest ions would likely penetrate through the masking photoresist layer. So the process as illustrated in Figures 2–10 and 2–11 is somewhat sensitive to layer thicknesses and to implant energy. This does not mean that it is too sensitive to be used in manufacturing, but it does mean that alternatives are often used.

One common alternative is illustrated in Figures 2–13 to 2–15. In this process flow, the field region doping is accomplished right after the steps shown in Figure 2–4, before the LOCOS oxide is grown. Thus a low-energy boron implant can be used which is easily masked by the photoresist/Si_3N_4/SiO_2 stack. This is illustrated in Figure 2–13. When the field oxide is then grown, most of the boron diffuses ahead of the growing SiO_2, creating the P regions shown in Figure 2–14. Some of the boron is actually incorporated into the growing SiO_2 and is therefore "lost" from the silicon. The fraction of the boron that is lost can be easily calculated, as we will see in later chapters. For obvious reasons, the implant in Figure 2–13 is often called the field implant. It increases the P-type doping in the substrate where we do not want to build active devices. A typical field implant might be 1×10^{13} cm^{-2} B^+ at 50 keV. This implant energy would easily pass through the

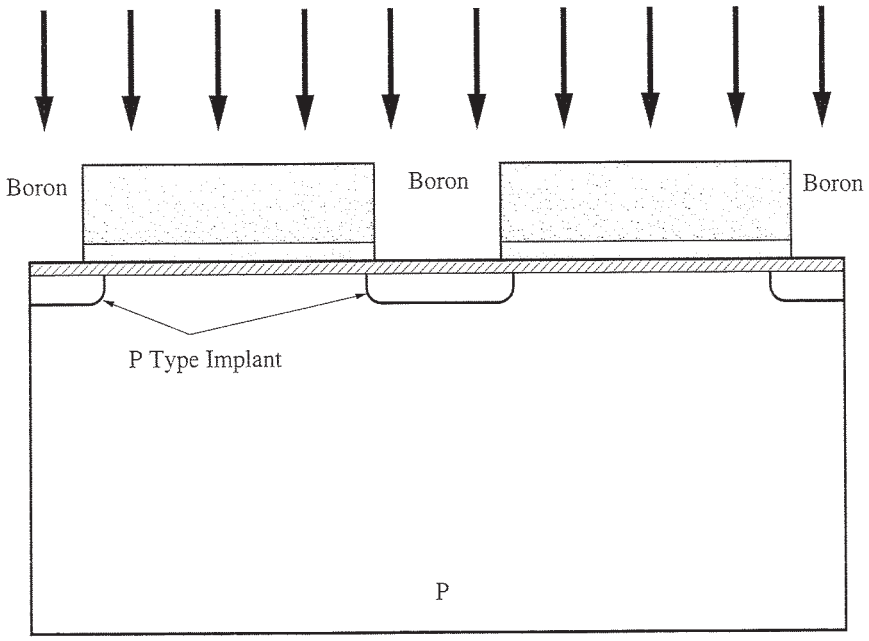


Figure 2–13 Process option for active region formation. A boron implant prior to LOCOS oxidation increases the substrate doping locally under the field oxide to minimize field inversion problems.

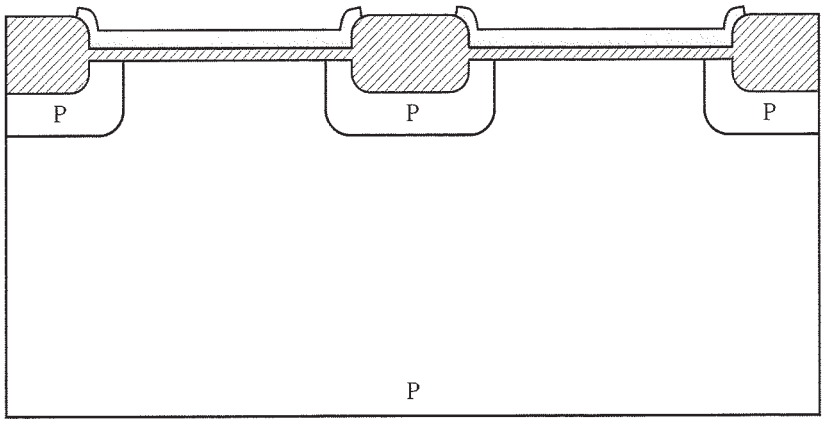


Figure 2–14 Process option for active region formation after LOCOS. The boron implanted regions diffuse ahead of the growing oxide producing the P-doped regions under the field oxide.

thin SiO_2 layer in Figure 2–13 and would also easily be completely masked by the photoresist/Si_3N_4/SiO_2 stack.

The P and N well formation in this process option would essentially follow the steps illustrated in Figures 2–10 and 2–11 except that the implant energies would be considerably lower to make sure that the phosphorus and boron did not penetrate through the field oxide. Implant energies on the order of 50 keV would be typical for the B^+ and P^+. The resulting structure is shown in Figure 2–15. Note that the P-type field implant regions continue to diffuse during the well drive-in. The wells are shown deeper than the field region in this example, but the exact geometry depends on the dopant concentrations in each region. Notice also that this process option does not require any more masking steps than the process we considered through Figure 2–12, since the field implant is done through the same mask that is used for the LOCOS process. As was the case in Figure 2–12, the substrate shown in Figure 2–15 is now ready for active device fabrication.

Option 2: Buried and Epitaxial Layers

Some CMOS circuits are built today in wafers that have buried heavily doped layers incorporated under the active devices. Alternatively, some CMOS circuits are built using heavily doped substrates such as the P on P^+ structure illustrated in Figure 1–34. Such heavily doped regions can help to minimize problems such as "latchup" in operating CMOS circuits. Latchup can occur because CMOS technology inherently contains parasitic PNPN structures. These structures have electrical I-V characteristics which, if triggered, can permit enough current to flow through the circuit that the chip may be destroyed. (Latchup is discussed in more detail in Chapter 8. See, for example, Figure 8–16.) One of the effective ways to prevent this from happening is to incorporate heavily doped buried layers or a heavily doped substrate beneath the active devices. These layers shunt the parasitic PNPN devices with low-value resistances, preventing the PNPN devices from turning on.

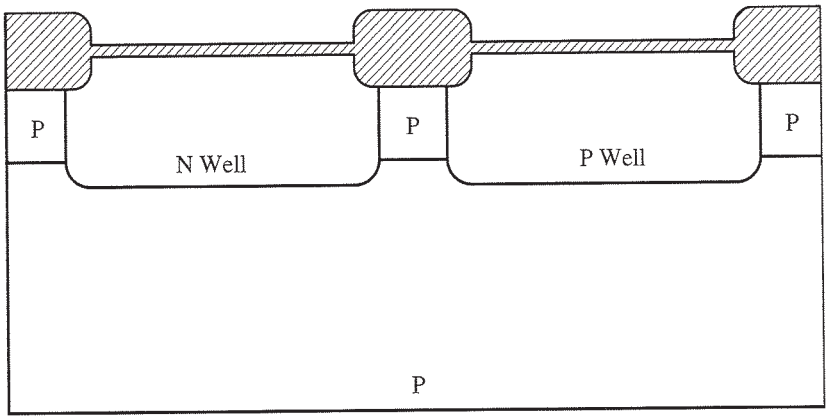


Figure 2–15 Process option for active region formation after the P and N wells are formed.

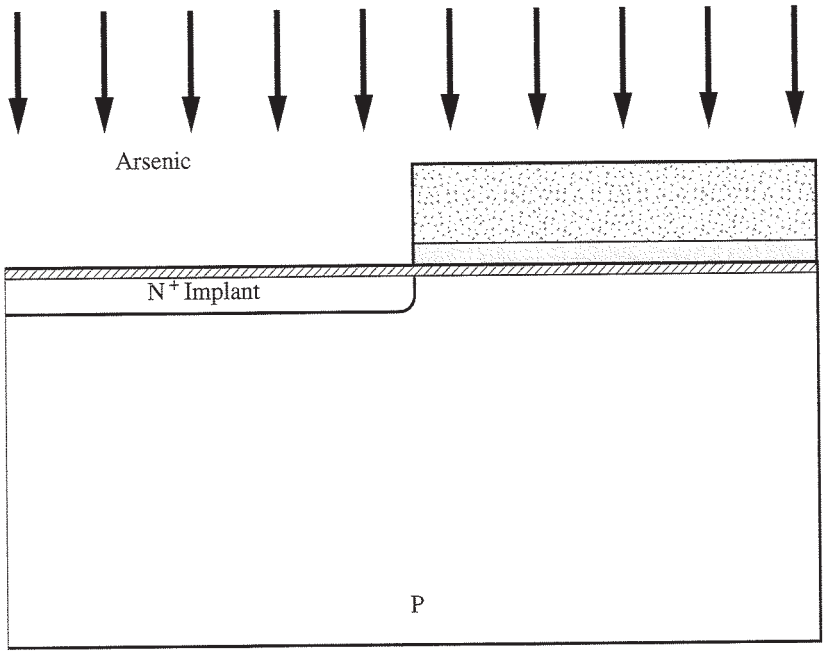


Figure 2–16 Process option incorporating buried and epitaxial layers. Mask 1 defines the regions for N^+ buried layers. An As^+ implant dopes the silicon locally.

The process steps needed to incorporate buried layers are shown in Figures 2–16 to 2–21. These steps are incorporated at the very beginning of the process flow, for reasons that will become obvious shortly.

We begin with the structure in Figure 2–3. The first mask is used as shown in Figure 2–16 to define the regions where an N^+ buried layer will be formed. An As^+ implant is then performed. Since the purpose here is to create a low-resistance region, we would want the N^+ layer to be fairly heavily doped and so a high-dose implant on the order of 10^{15} cm^{-2} would be used. The energy is not critical so long as it is sufficient to accelerate the As^+ through the thin SiO_2 layer. A reasonable energy would be 50 keV. Since

the N^+ region we are forming will be buried below the surface (when we are finished), and we want it to stay there, we would want to pick an N-type dopant that diffuses slowly in silicon. As or Sb would thus be possible choices. Either could be used, but As is more common because of its higher solubility in silicon.

Once the implant is completed, we are through with the resist and it can be chemically removed in acid, or stripped in an O_2 plasma. Following cleaning, the wafers are then placed into a furnace. This high-temperature step accomplishes several things, as illustrated in Figure 2–17. First, it drives in the N^+ buried layer to a depth of about 1–2 μm. Second, part of the drive-in is done in an oxidizing ambient (H_2O). This grows a thick SiO_2 layer on the wafer surface, but only over the N^+ regions because of the masking provided by the Si_3N_4 layer. We will see that using a LOCOS process in this application allows both the N^+ and P^+ buried layers to be defined with a single mask, and it provides self-alignment between these two buried layers. By self-alignment, we mean that the two buried layers are correctly positioned with respect to each other automatically. Finally, the drive-in and oxidation create a step in the silicon surface at the edges of the N^+ buried layers because the oxidation consumes silicon. This step in the surface will be important later after the buried layers are truly buried by growing an epitaxial layer of silicon above them. We will need to know where the buried layers are located in order to align later masks to them and the step in the silicon surface will provide this information. The overall drive-in process might be done at 1000°C for 2 hours, with 60 min of this time in H_2O to locally grow about 0.4 μm of SiO_2. The structure at this point is illustrated in Figure 2–17.

After the furnace operation, the Si_3N_4 layer can then be stripped. The next step is ion implantation of the P^+ buried layer, illustrated in Figure 2–18. Here the self-aligning feature in the two buried layers becomes apparent. The thick oxide over the N^+ buried layers blocks the P^+ (boron) implant, while the original thin oxide layer elsewhere is transparent to the implant. Once again, we would choose the implant energy so that the boron penetrates the thin oxide and is stopped by the thick oxide (about 50 – 75 keV). The boron dose is determined by device constraints, as was the case for the N^+ buried

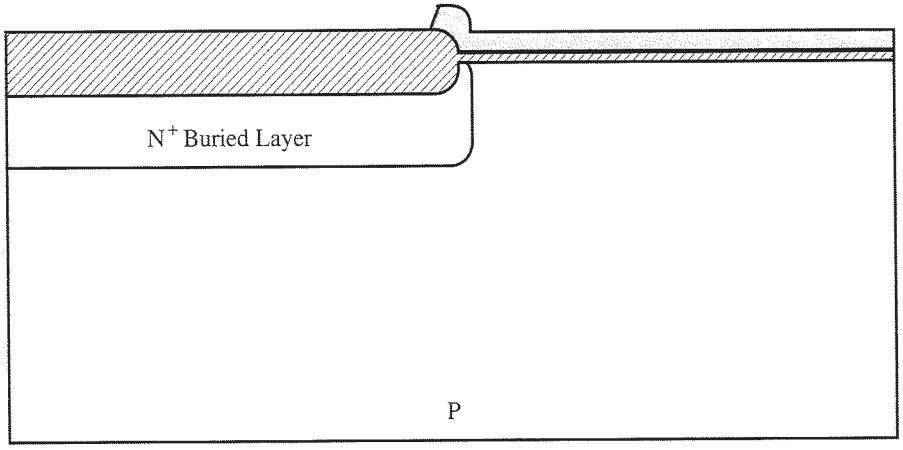


Figure 2–17 Process option incorporating buried and epitaxial layers. The N^+ buried layer is driven in in an oxidizing ambient after the photoresist is stripped. The LOCOS oxide forms only above the N^+ regions.

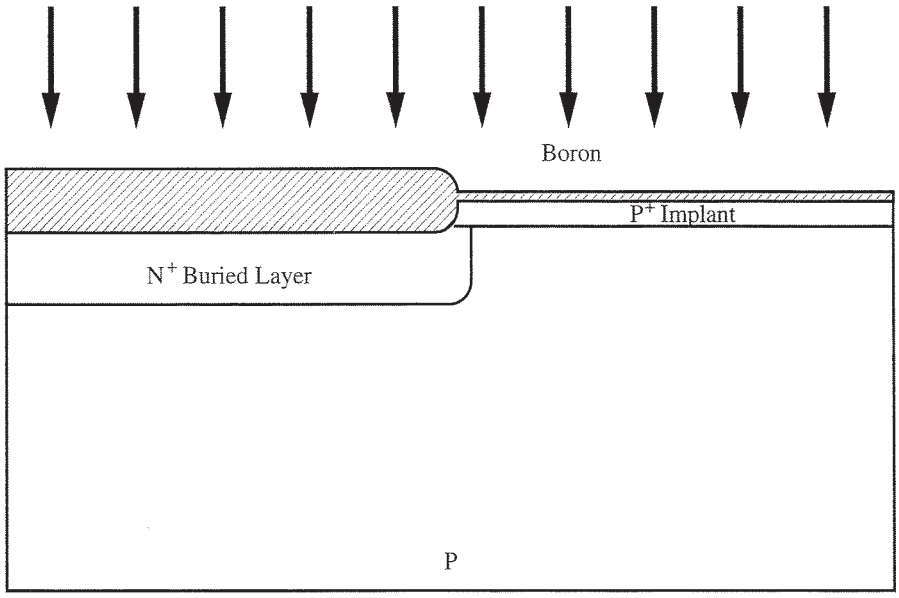


Figure 2–18 Process option incorporating buried and epitaxial layers. The P^+ buried layer is implanted using the thick SiO_2 layer as a mask.

layer. In the P^+ buried layer case, a lower dose is usually used, because boron has a much higher diffusivity than arsenic and in order to keep it from diffusing too far during subsequent processing the boron concentration needs to be kept lower than the arsenic concentration. A dose of about 10^{14} cm^{-2} might be typical. After the P^+ implant, a high-temperature drive-in would be done to diffuse the boron (and the arsenic) deeper into the substrate. No additional oxidation is required at this point so the drive-in could be done in an inert (N_2 or Ar) ambient. A typical furnace cycle at this point might be several hours at 1000 – 1100°C, resulting in the structure shown in Figure 2–19.

The active devices need to be fabricated in much more lightly doped wells than is provided by these buried layer regions in order to have the correct electrical properties. As a result, we require moderately doped P and N regions above the buried layers. These more lightly doped layers reduce the junction capacitances in the active transistors and are also important in setting device parameters such as MOS threshold or turn-on voltage. In principle we could counterdope the surface regions of the buried layers with opposite type dopants to produce these regions, but this is not a manufacturable technique. This is easy to see if we imagine trying to counterdope a 1.0×10^{19} cm^{-3} N^+ buried layer to form a 1.0×10^{16} cm^{-3} N layer. This would require 0.999×10^{19} cm^{-3} P-type counter doping. No doping technique available today provides this degree of precision.

Fortunately there is an alternative, a process called epitaxy. The oxide layer on the surface of the wafer in Figure 2–19 is first stripped in an HF solution. This acid is highly selective to SiO_2 over Si. After cleaning, the wafers are then placed into an epitaxial reactor which heats the wafers to temperatures on the order of 800 – 1000°C and exposes them to a gas ambient containing Si and a small concentration of dopant. SiH_4 or

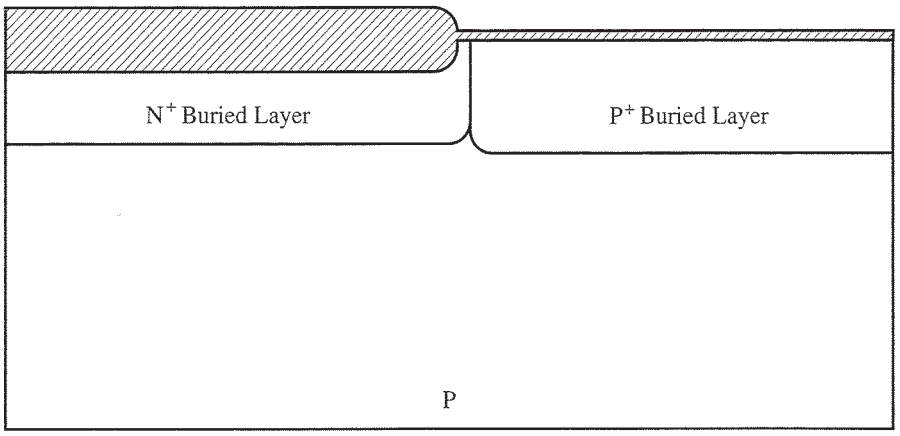


Figure 2–19 Process option incorporating buried and epitaxial layers. The P^+ and N^+ buried layers are driven in together.

SiH_2Cl_2 might be used as the silicon source and B_2H_6 or AsH_3 for the P- or N-type doping, for example. The process of epitaxy is conceptually simple. Si atoms fall on the wafer surface from the gas stream above it. At the high temperature in the reactor, the Si atoms are somewhat mobile on the surface and can diffuse via surface diffusion to a lattice site to which they can bond. Through this process, the crystal structure of the substrate is grown upward, atom by atom, and layer by layer, producing a perfect crystalline or "epitaxial" layer. The doping atoms in the gas stream are incorporated into the growing epitaxial layer in the same manner. This process allows us to grow single-crystal layers on single-crystal substrates at a rate of several tenths of a micron per minute and to dope those layers from 10^{14} cm^{-3} to 10^{20} cm^{-3} N- or P-type. We will study this process in more detail in Chapter 9. In fact, epitaxy is really a special case of CVD which we previously used to deposit Si_3N_4 and SiO_2 layers. In epitaxy, the substrate must be single crystal; in the more general CVD process, arbitrary substrates can be used because the films that are deposited are amorphous or polycrystalline.

Figure 2–20 illustrates our CMOS wafer after an epitaxial layer has been grown. The epilayer for our devices is grown undoped (intrinsic) since we will be doping various parts of it later in the process using ion implantation. The epilayer might typically be a few microns thick. Notice that the step in the silicon surface which we created by oxidation of the N^+ buried layers has propagated upward during epi growth, so that it now appears on the new wafer surface. This step is visible under an optical microscope and is necessary in order to align subsequent masks to the buried layer patterns.

The remaining steps in this process option would follow the LOCOS steps (Figures 2–3 to 2–5) and the P- and N-well steps (Figures 2–10 to 2–12). After the wells are driven in, they link up with the buried layers (which also diffuse upwards during all high temperature steps) as shown in Figure 2–21. Both downward and upward diffusion must be accounted for in order to determine the time and temperature needed to accomplish this linkup. A typical set of conditions might be several hours at 1000 – 1100°C depending on the exact epitaxial layer thickness. This could be performed in a

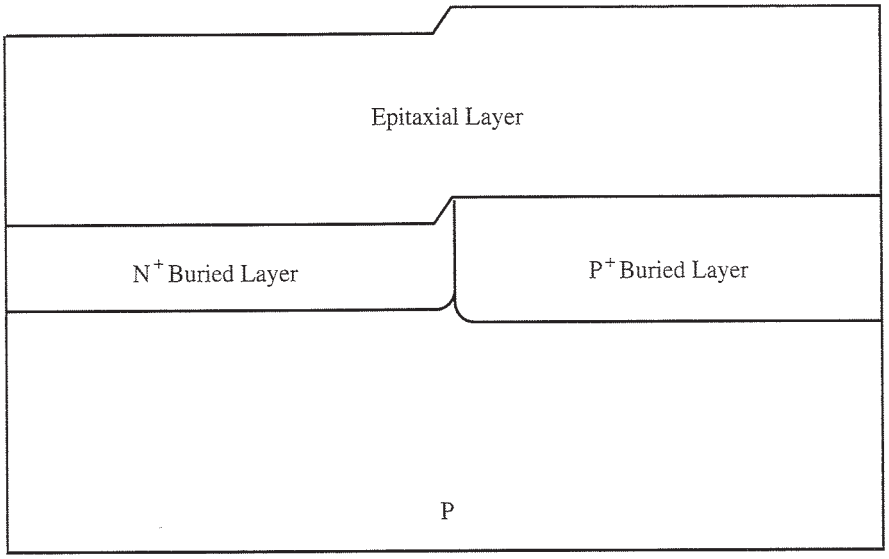


Figure 2–20 Process option incorporating buried and epitaxial layers. The surface SiO_2 layer is stripped off the wafer and an epitaxial layer is then grown.

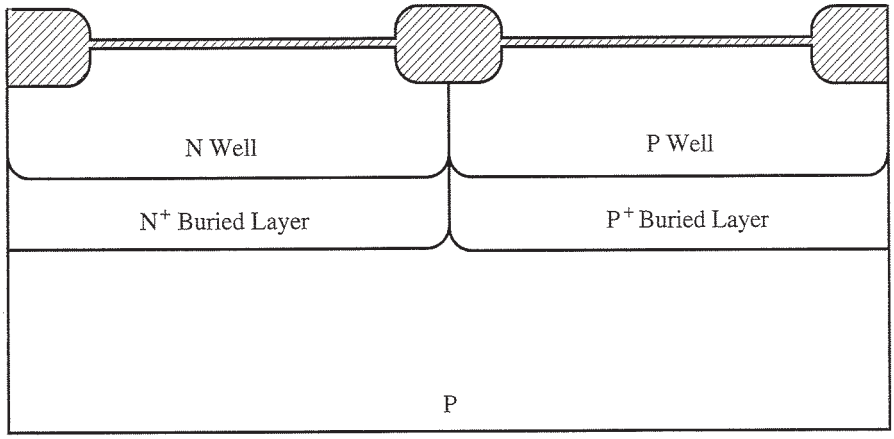


Figure 2–21 Process option incorporating buried and epitaxial layers.

largely inert ambient because no additional surface oxidation is needed at this point. Note that the incorporation of the buried and epitaxial layers into the structure has only required one additional mask (but many process steps!). As was the case in Figure 2–12, the substrate as shown in Figure 2–21 is now ready for active device fabrication. The step in the surface shown in Figure 2–20 would still be present in Figure 2–21 but is not explicitly shown.

Finally, the structure shown in Figure 1–34 is yet another variation on these process steps. This structure incorporates a P^- epitaxial layer and an N well using steps similar to those described above. The process flow in this case is left as an exercise for the reader. (See Problem 2.1.)

2.2.6 Gate Formation

We now return to the main process flow and Figure 2–12. If either of the process options described above were to be used, the substrate would appear as shown in Figure 2–15 or 2–21, but the process flow from this point on would be substantially the same. So for simplicity, we will continue the process description with Figure 2–12.

The next several steps, shown in Figures 2–22 to 2–26 are designed to form critical parts of the MOS devices. Probably the single most important parameter in both the NMOS and PMOS devices is the turn-on or threshold voltage, discussed in Chapter 1 and usually called V_{TH}. V_{TH} in its simplest form is given by

$$V_{TH} = V_{FB} + 2\phi_f + \frac{\sqrt{2\varepsilon_S q N_A (2\phi_f)}}{C_{OX}} \tag{2.3}$$

where V_{FB} is the gate voltage required to compensate for work function differences between the gate and substrate, and for any electrical charges that may be present in the gate oxide. ϕ_f is the position of the Fermi level in the bulk with respect to the intrinsic level and ε_s is the permittivity of silicon. For our present purposes, the two terms that are important are the doping concentration in the silicon N_A and the oxide capacitance C_{ox}. Since C_{ox} is inversely proportional to the gate oxide thickness, it is clear that we must control this thickness in order to control V_{TH}.

In writing the above expression, we have assumed that the doping in the silicon under the MOS gate is constant at N_A. This is usually not the case in modern devices because ion implantation is used to adjust the threshold voltage and this results in a nonuniform doping profile. Again to first order, we can include the effect of the implant on V_{TH} in the following way:

$$V_{TH} = V_{FB} + 2\phi_f + \frac{\sqrt{2\varepsilon_S q N_A (2\phi_f)}}{C_{OX}} + \frac{qQ_I}{C_{OX}} \tag{2.4}$$

where Q_I is the implant dose, in atoms per cm^2. This equation assumes that the entire implant dose is located in the near surface region, inside the MOS channel depletion region. This is often a reasonable approximation.

We are now ready to adjust the threshold voltages of both N- and P-channel MOS devices. In modern CMOS circuits, the target threshold voltage is generally around 0.5–0.8 volts for both the NMOS and PMOS devices. (The threshold voltage is positive for the NMOS devices and negative for the PMOS devices so that both transistors are normally off, enhancement mode devices.) Figure 2–22 illustrates the masking to adjust the NMOS V_{TH}. Photoresist is applied and mask 4 is used to open the areas where NMOS devices are located. After developing, a boron implant is used to adjust V_{TH}. A dose of 1–5 $\times$ 10^{12} cm^{-2} at an energy of 50 – 75 keV might be used. We could estimate the necessary dose using Eq. (2.4) where N_A is the doping in the P well at the surface before the implant, and Q_I is the dose required to achieve a given V_{TH}. The energy is chosen to be high enough to get the implant dose through the thin oxide, but low enough to keep the boron near the silicon surface.

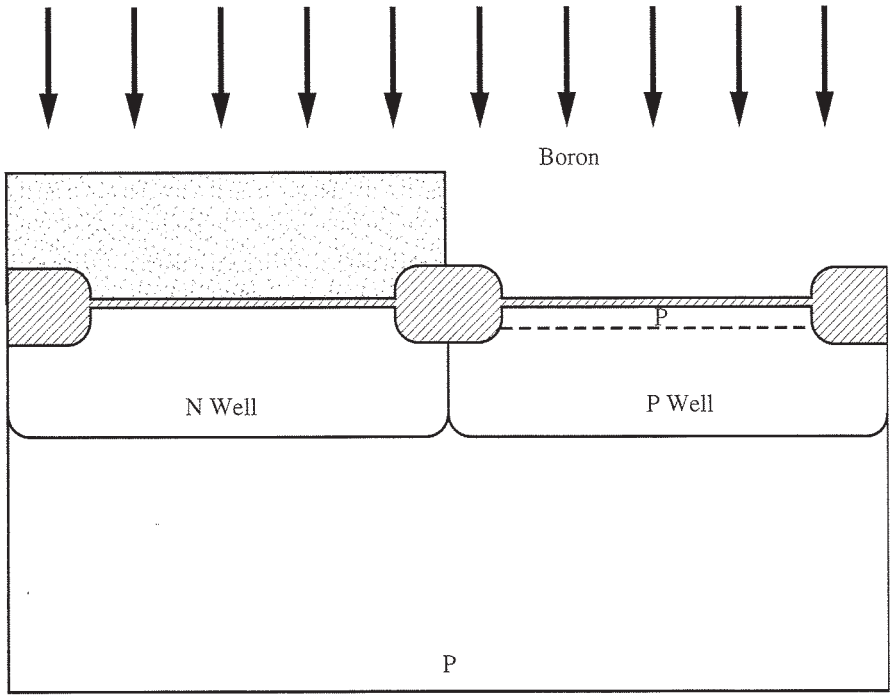


Figure 2–22 After spinning photoresist on the wafer, mask 4 is used to define the NMOS transistors. A boron implant adjusts the N-channel V_{TH}.

Figure 2–23 illustrates the same process sequence now applied to the PMOS device. Mask 5 is used. The required implant could be either N- or P-type depending on the doping level in the N well and the required PMOS V_{TH}. An N-type implant is illustrated in Figure 2–23. This would typically be arsenic with a dose of 1–5 $\times$ 10^{12} cm^{-2}. The energy would be somewhat higher than the NMOS channel implant in Figure 2–22 because of the heavier mass of arsenic.

If a P-type implant were needed, boron would be used with a dose and energy in the same range as for the NMOS device in Figure 2–22. In some cases, it might be possible to use only one mask to adjust both NMOS and PMOS V_{TH} if both require P-type implants. One possible process might be to implant boron unmasked into both devices at the smaller of the two doses required for the MOS devices. A mask would then be used along with a second implant to increase the dose in the device requiring more boron.

Figure 2–24 illustrates the next steps. We are now ready to grow the gate oxides for the MOS transistors. The thin oxide, which is present over the active areas of each transistor, is first stripped in a dilute HF solution. HF is a highly selective etchant and will stop etching when the underlying silicon is reached. Note however that we will etch a small portion of the field oxide during this step because the HF etch is unmasked. Since we are etching only 10 or 20 nm of oxide, this is usually not a problem, although the etch needs to be timed so that it does not etch too much of the field oxide.

The reason that the thin oxide is stripped and then regrown to form the MOS gate oxide is that the oxide on the silicon surface prior to stripping is too thick to serve as the device gate oxide. Stripping and regrowing this oxide results in a well-controlled final

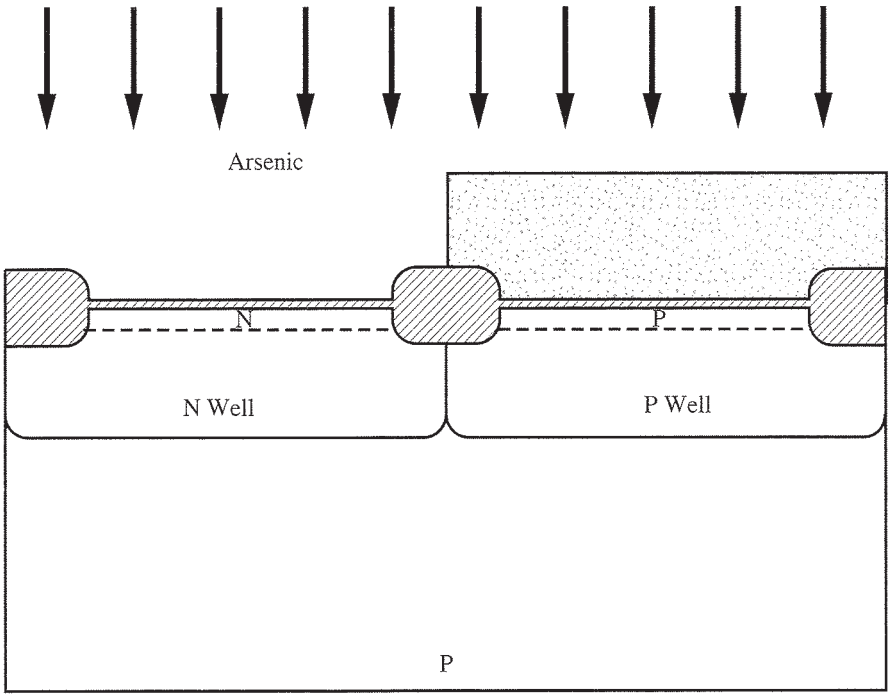


Figure 2–23 After spinning photoresist on the wafer, mask 5 is used to define the PMOS transistors. An arsenic implant adjusts the P-channel V_{TH}.

oxide thickness. The original thin oxide on the wafer surface has also been exposed to several implants at this stage in the process, which create damage in the SiO_2, so stripping and regrowing a new oxide produces a higher quality gate oxide. In state-of-the-art MOS devices today, the gate oxide is typically thinner than 10 nm. This oxide could be formed by a variety of processes (times and temperatures). For example, oxidation in O_2 at 800°C for 2 hours would produce about 10 nm of oxide. Similarly, oxidation in H_2O at 800°C for 25 min would produce a similar thickness oxide. Figure 2–24 illustrates the devices after the gate oxide has been grown.

The next steps deposit and define the polysilicon gate electrodes for the MOS devices. Using LPCVD, a layer of polysilicon is deposited over the entire wafer surface as illustrated in Figure 2–25. This process is similar to that described in Eq. (2.1) except that only a silicon source such as silane is needed. Thermal decomposition of the silane produces a silicon deposition with H_2 as a byproduct, as shown in Eq. (2.5). The deposited layer will be either amorphous or polycrystalline depending on the deposition temperature because it is deposited on an amorphous "substrate," the underlying SiO_2 regions. Typically a polysilicon layer 0.3 – 0.5 μm thick would be deposited in an LPCVD system operating at about 600°C.

$$SiH_4 \rightarrow Si + 2H_2 \tag{2.5}$$

The polysilicon is then doped N-type by an unmasked ion implant. Either phosphorus or arsenic could be used here since both are highly soluble in silicon (and polysilicon) and thus can produce low-sheet-resistance poly layers. The implant energy is not

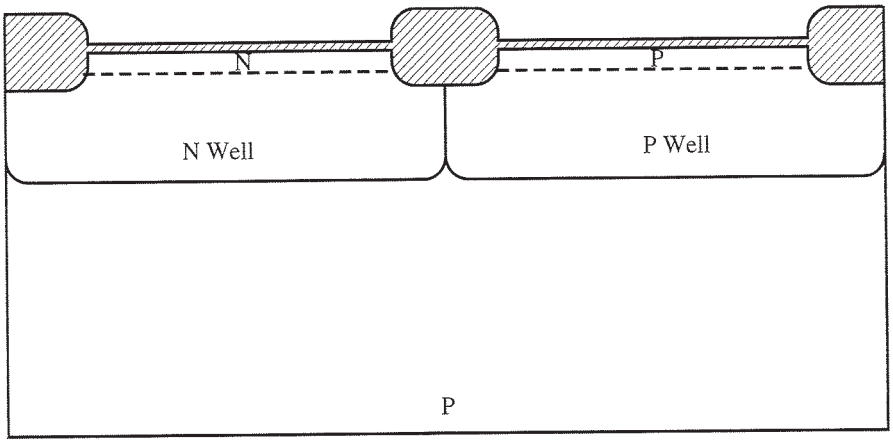


Figure 2–24 After etching back the thin oxide to bare silicon, the gate oxide is grown for the MOS transistors.

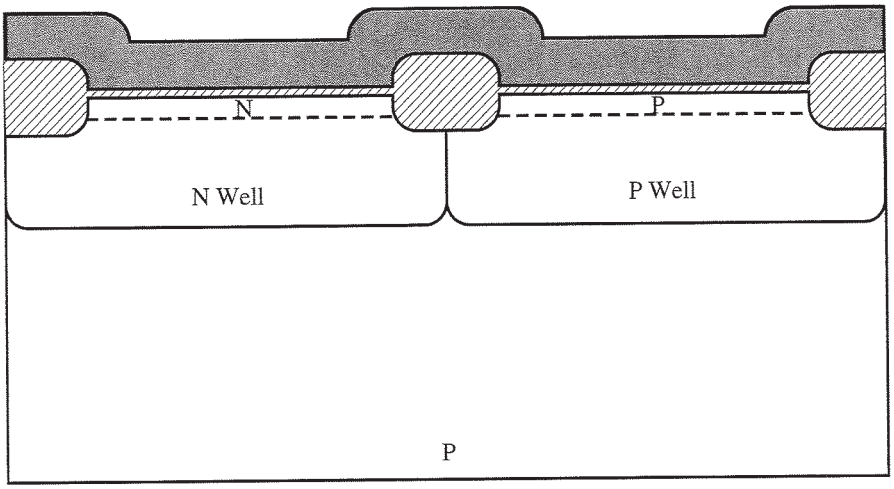


Figure 2–25 A layer of polysilicon is deposited. Ion implantation of phosphorus follows the deposition to heavily dope the poly.

very critical here, provided the phosphorus or arsenic do not penetrate through the poly and into the underlying gate oxide and substrate. Both dopants rapidly redistribute in poly at elevated temperatures because diffusion is rapid along the grain boundaries in poly, so uniform doping of the poly will occur later in the process when the wafers are next heated in a furnace. The N^+ dose is not critical for the MOS gates other than the fact that we would like it to be as high as possible in order to obtain low poly sheet resistivity and hence low gate resistance. A dose of about 5×10^{15} cm^{-2} would be typical. In some polysilicon deposition systems, the poly can be doped while it is being deposited. This is referred to as "in situ" doped poly. In this case, the ion implantation doping step would not be necessary.

The final step, illustrated in Figure 2–26 uses resist and mask 6 to etch the poly away in regions where it is not needed. Photoresist is spun onto the wafer, baked, and then exposed and developed. The poly etching would again be done in a plasma etcher. Typically a chlorine- or bromine-based plasma chemistry would be used in order to achieve

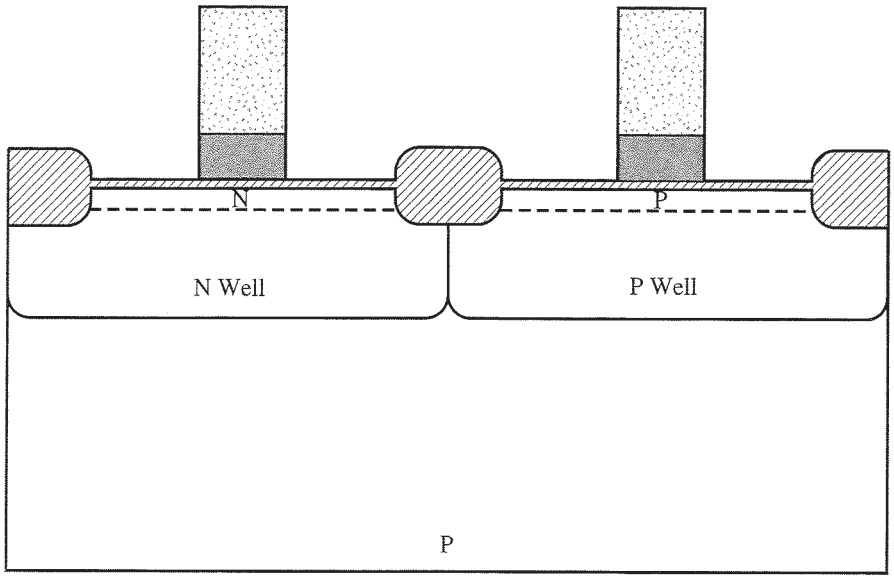


Figure 2–26 Photoresist is applied and mask 6 is used to define the regions where MOS gates are located. The polysilicon layer is then etched using plasma etching.

good selectivity to SiO_2. Although it is not shown in Figure 2–26, the polysilicon layer can also be used to provide wiring between active devices on the chip (for example to connect the NMOS and PMOS poly gates). In this sense it can serve as the first level of interconnect. Since the poly sheet resistance is relatively high compared to later metal layers that will be deposited ($\approx 10\ \Omega$/sq versus $< 0.1\ \Omega$/sq), long interconnects are not made with the polysilicon. The RC delays associated with long poly lines would have a significant effect on circuit performance, hence the term "local interconnects" for these relatively short polysilicon wires.

A final point with regard to all these etching steps is worth making at this point, although we will discuss it in much more detail in Chapter 10. In general there are two key parameters that must be understood and controlled in any etching step. The first is selectivity and the second is the degree of anisotropy the etch provides. Selectivity is a key issue because we nearly always find ourselves in the situation where we wish to etch one material but stop etching when we hit an underlying material. For example, in Figure 2–26, we wanted to etch the poly layer, but we of course did not want to etch through the gate oxide and into the underlying silicon substrate. This means that we would need to select an etching process during the poly etch that was highly selective between SiO_2 and Si, so that the etch would in essence stop when it reached the SiO_2 layer. Techniques exist to detect the endpoint of an etching process, but usually some amount of overetching is necessary to make certain that all areas on a wafer or wafers have completed etching, so selectivity is usually very important.

Anisotropy is a key issue because we are usually very concerned about the shape of edges on etched regions. We are often required to etch through materials whose thicknesses are on the same order as the lines we are trying to etch or define. Ideally we would like the edges of the etched materials to be nearly vertical to preserve the mask dimensions in the etched layers. Anisotropic etches approach this ideal; isotropic etches

produce about as much undercutting as the vertical depth they etch and so are generally unsuitable for small geometry devices. So in general, we want highly selective, anisotropic etches. Unfortunately, it is often the case that anisotropy comes at the expense of selectivity and vice versa. Plasma etching often means a search for the best trade-off between the two that can be achieved in a given system. We will return to these issues in Chapter 10.

2.2.7 Tip or Extension (LDD) Formation

The next several steps are illustrated in Figures 2–27 to 2–30. Our objective in these steps is twofold. First, we want to introduce the N^- and P^- implants shown in the NMOS and PMOS devices in Figures 2–27 and 2–28 and second, we want to place along the edges of the polysilicon gates, a thin oxide layer usually called a "sidewall spacer." Both of these steps are required because of scaling trends that have taken place in the semiconductor industry over the past decade.

Ten years ago, MOS devices used in ICs were built with minimum dimensions well above one micron and were operated in circuits with supply voltages of 5 volts. Today, device dimensions have been reduced to 0.25 μm or smaller in order to improve performance. However, supply voltages in circuits have not been proportionally reduced. Many ICs still use 5-volt power supplies, although many chips are now being designed with 3.3- and 2.5-volt supplies. There is great benefit at the system level to maintaining a standard power supply level because then new ICs are compatible with older parts,

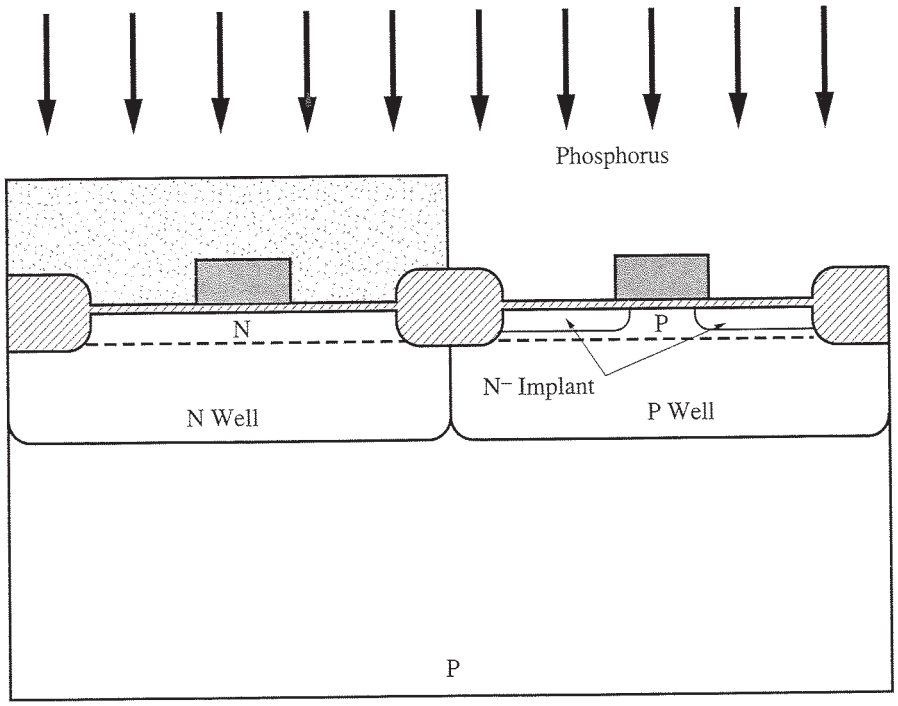


Figure 2–27 Mask 7 is used to cover the PMOS devices. A phosphorus implant is used to form the tip or extension (LDD) regions in the NMOS devices.

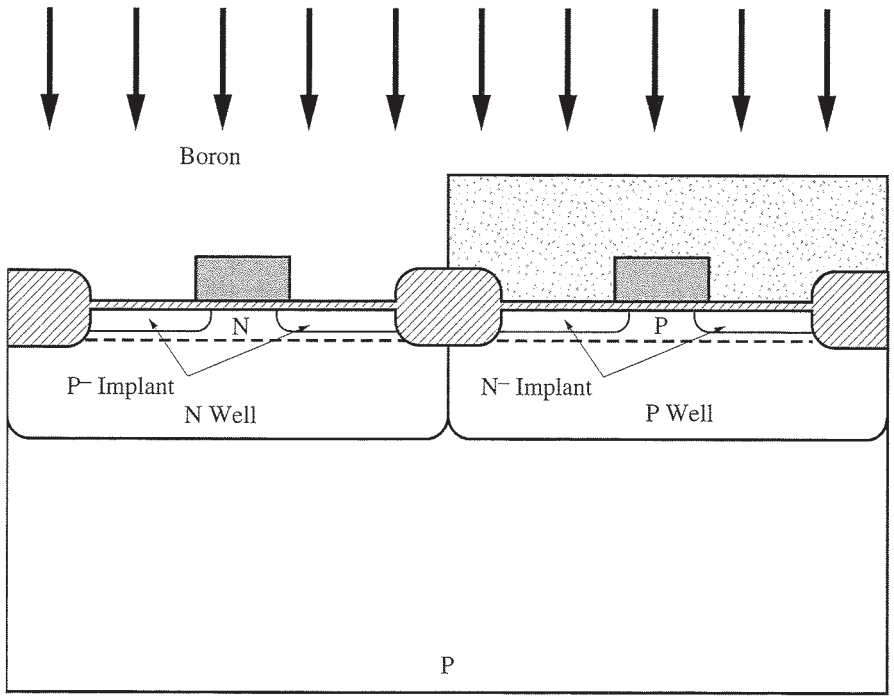


Figure 2–28 Mask 8 is used to cover the NMOS devices. A boron implant is used to form the tip or extension (LDD) regions in the PMOS devices.

system power supplies do not have to be redesigned and circuit noise margins remain adequate. However, if device dimensions are reduced and voltage levels are not correspondingly scaled, electric fields inside the devices necessarily rise. Five volts applied across a 2-μm channel length MOS device implies an average electric field in the channel of about 2.5×10^4 V cm^{-1}. Decreasing the channel length in the device to 0.5 μm without reducing the supply voltage increases this average field to about 10^5 V cm^{-1}. Fields this high are large enough to cause problems in semiconductor devices. Such problems are often called "hot electron" problems because most of them are due to the high energies that electrons (or holes) can reach in high electric fields. Carriers at high energies can cause impact ionization which creates additional hole-electron pairs by breaking Si-Si bonds. Such carriers can also sometimes gain sufficient energy to surmount large energy barriers such as the 3.2 eV barrier between the Si conduction band and the SiO_2 conduction band. The result can be carriers injected into gate dielectrics which may become trapped and cause device reliability problems.

Because of these larger fields in scaled devices, considerable effort has gone into designing MOS device structures that can withstand high electric fields. One of the innovations that is almost universally used is the Lightly Doped Drain or LDD device. The idea behind this structure is to grade the doping in the drain region to produce an N^+N^-P profile between the drain and channel in the NMOS devices and a corresponding P^+P^-N profile in the PMOS devices. This allows the drain voltage to be dropped over a larger distance than would be the case if an abrupt N^+P junction were formed. This reduces the peak value of the electric field in the near drain region. Since many of the deleterious effects of high electric fields in modern MOS devices depend

exponentially on the electric field, modest reductions in the field strength obtained through the LDD structure can make a significant difference in device reliability.

A final point regarding these N^- and P^- implants is also important to make. As device geometries have become smaller, "short channel effects" have become very important in MOS transistors. These effects result when the drain electric field penetrates through the channel region and begins to affect the potential barrier between the source and channel regions. The result is drain current that is not controlled effectively by the gate. An important strategy for minimizing these effects is the use of shallow junctions. Such junctions are less susceptible to short channel effects essentially because their geometry minimizes the junction areas adjacent to the channel. The LDD structure also provides these shallow junctions which in this context are often called the "tip" or "extension" regions since they must be combined with deeper source and drain junctions away from the channel in order to make reliable contacts to the device. The N^- and P^- implants in Figures 2–27 and 2–28 and the sidewall spacers in Figure 2–30 are used to construct these tip or extension or LDD regions.

In Figure 2–27, photoresist is spun on the wafer and mask 7 is then used to protect all the devices except the NMOS transistors. A phosphorus implant is done to form the N^- region. The dose and the energy are carefully controlled in this implant to ultimately produce the desired graded drain junction. Typically, a dose of about 5×10^{13} to 5×10^{14} cm^{-2} at a low energy might be used. A similar sequence of steps is used for the PMOS devices in Figure 2–28 to produce the LDD regions in these devices. A similar implant would be performed although boron would be used in this case. In some modern MOS device structures, the "LDD" implants may actually consist of several implants at different energies and doses. Some of these implants may even be done at angles tilted with respect to the wafer surface in order to get the implant further under the edge of the gate. The objective is to carefully tailor the doping profile near the drain junction (and the source junction) in order to minimize short channel effects. The exact 2D shape of the resulting profile is crucial in obtaining the correct device characteristics and computer simulation is often very useful in understanding and predicting the structure and properties of these critical regions. We will consider these issues more carefully in later chapters.

The next step is the LPCVD deposition of a conformal spacer dielectric layer (SiO_2 or Si_3N_4) on the wafer surface, shown in Figure 2–29. The thickness of this layer will determine the width of the sidewall spacer region and would be chosen to optimize device characteristics. Typically this might be a few hundred nm. If SiO_2 is used, it could be deposited by a $SiH_4 + O_2$ reaction at about 400°C or by a $SiH_2Cl_2 + N_2O$ reaction at about 900°C or other similar reactions, in a standard furnace configured as a CVD or LPCVD system.

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \tag{2.6}$$

$$SiH_2Cl_2 + N_2O \rightarrow SiO_2 + 2N_2 + 2HCl \tag{2.7}$$

We now make use of a technique to form the sidewall spacers that really is a result of the efforts that have gone into developing anisotropic plasma etching capabilities in

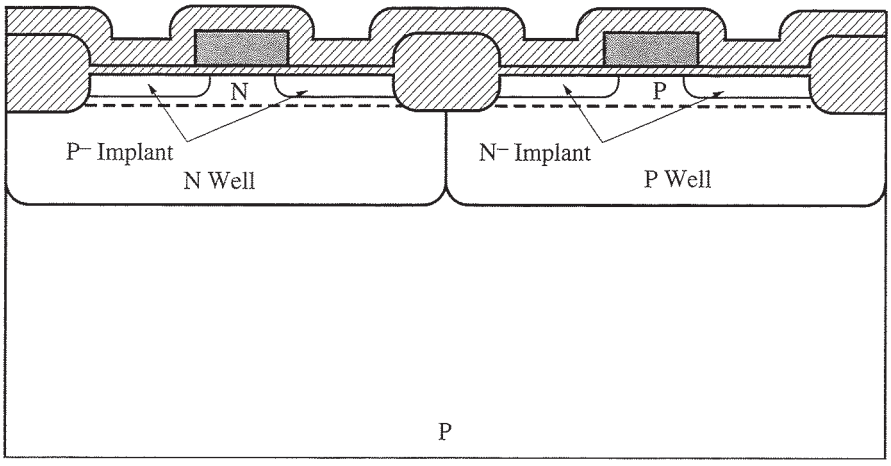


Figure 2–29 A conformal layer of SiO_2 is deposited on the wafer in preparation for sidewall spacer formation.

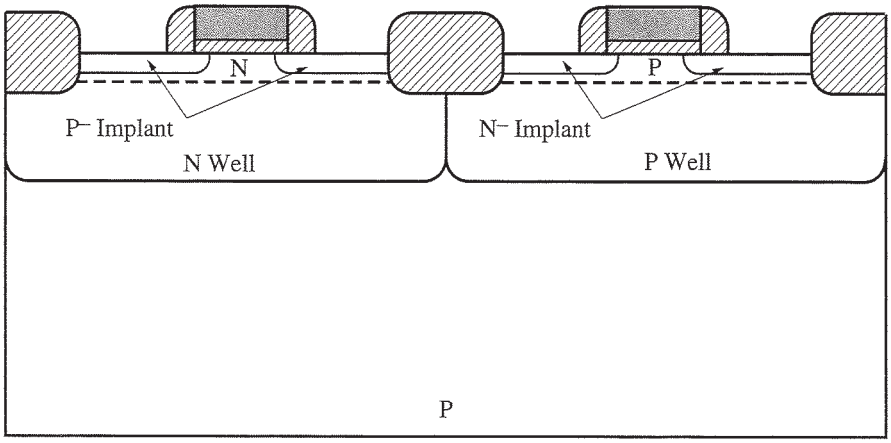


Figure 2–30 The deposited SiO_2 layer is etched back anisotropically, leaving sidewall spacers along the edges of the polysilicon.

recent years. Notice in Figure 2–29 that the deposited SiO_2 layer is much thicker along the edges of the polysilicon than it is above flat regions of the wafer surface. This is because of the vertical edges of the polysilicon regions, and the conformal deposition of the oxide. If we now etch back the deposited SiO_2 layer using an etching technique that is highly anisotropic (etches vertically but not horizontally) then we will be left with the structure shown in Figure 2–30. Typically this would be done in a fluorine-based plasma.

The deposited oxide is removed everywhere except along the edges of vertical steps in the underlying structures. Simply by a deposition and then an etchback, we have formed the sidewall spacers. We have also created lateral features on the chip surface that are smaller than the minimum feature size of the lithographic process. In fact the width of the spacers is determined largely by the thickness of the deposited oxide. It

should also be noted that there is nothing magic about the use of SiO_2 in this application. We could form such spacers using almost any deposited thin film. In this particular case we need an insulating material for reasons that will become apparent shortly, so SiO_2 is a convenient choice.

2.2.8 Source/Drain Formation

At this point, most of the doped regions in the structure have been formed except for the MOS transistor source and drain regions. Note however that in Figure 2–30 the oxide was etched off the source and drain regions as part of the sidewall spacer formation process. Generally implants are done through a thin "screen" oxide, whose purpose is both to help avoid channeling and to minimize the incorporation of trace impurities into the silicon from the implanter. Channeling is a result of the fact the silicon is crystalline. If the implanted ions have a velocity vector that lines up with the crystal structure of the substrate, ions can go down "channels" between lattice sites for long distances without encountering silicon atoms which slow the implanted atoms down via collisions. If this happens, the range of the implanted ions can be significantly larger than expected. This is generally undesirable and the thin screen oxide (which is amorphous) helps to randomize the directions of the implanted ions and therefore minimize channeling. We will discuss these effects in more detail in Chapter 8. Prior to doing the source drain implants then, a thin screen oxide of perhaps 10 nm is grown. Note that this also produces a thin oxide on the top exposed surface of the polysilicon.

The first source drain implant is illustrated in Figure 2–31. Photoresist and mask 9 are used to define the regions where NMOS source/drain implants will be done. Arsenic

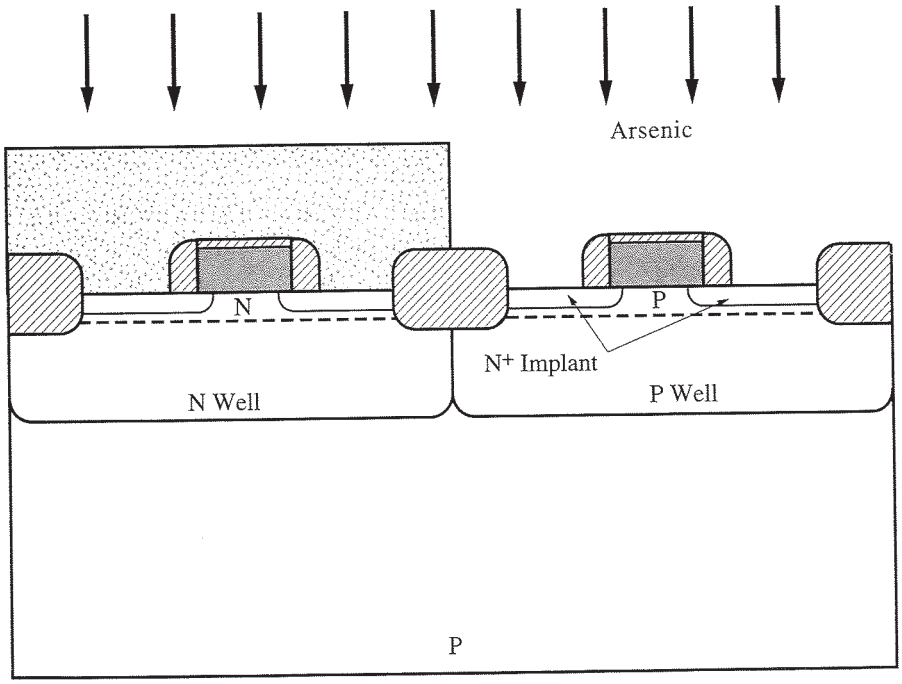


Figure 2–31 After growing a thin "screen" oxide, photoresist is applied and mask 9 is used to protect the PMOS transistors. An arsenic implant then forms the NMOS source and drain regions.

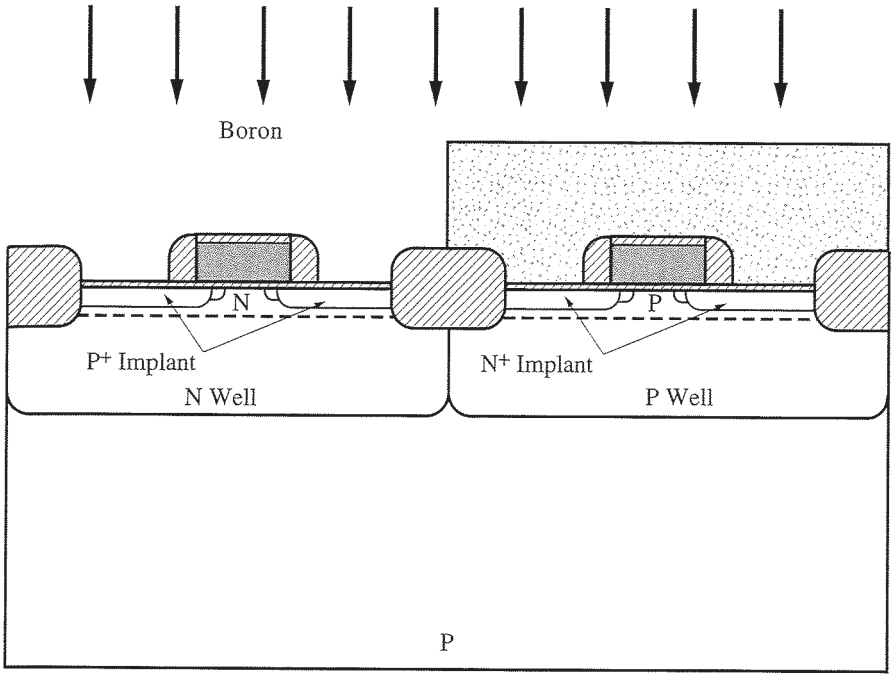


Figure 2–32 After applying photoresist, mask 10 is used to protect the NMOS transistors. A boron implant then forms the PMOS source and drain regions.

would be the dopant of choice in modern processes, because of the need to keep junctions shallow in small geometry devices. An implant of $2-4 \times 10^{15}$ cm^{-2} at an energy of 75 keV might be typical. This would allow the arsenic to penetrate the screen oxide in the implanted areas but still be easily masked by the photoresist.

The final mask used for doping is illustrated in Figure 2–32. Photoresist and mask 10 allow a boron implant to form the PMOS source/drain regions. This implant would also be a high-dose implant, on the order of $1–3 \times 10^{15}$ cm^{-2}, but at a lower energy of about 50 – 75 keV because boron is much lighter than arsenic and therefore requires less energy to reach the same range. High-dose implants minimize the parasitic resistances associated with the source and drain regions in the MOS transistors. It is also interesting to notice that the polysilicon gate regions receive at least two high-dose implants in the process flow we have described. The first, which was N^+, occurred in connection with Figure 2–27 and initially heavily doped the poly N type. In the NMOS devices, a second N^+ implant goes into the poly in Figure 2–31. In the PMOS device a P^+ implant dopes the poly in Figure 2–32. In most processes today both the PMOS and NMOS gates are N type. If this is the case, then the P^+ dose in Figure 2–32 needs to be smaller than the N^+ dose implanted in Figure 2–27. This is the case for the numbers we have used in this process flow.

The final step in active device formation is illustrated in Figure 2–33. A furnace anneal, typically at $\approx$ 900°C for 30 min, or perhaps a rapid thermal anneal for $\approx$ 1 min at 1000 – 1050°C activates all the implants, anneals implant damage, and drives the junctions to their final depths. The many implants that have been done to form the N and P regions in the transistors create significant damage to the crystal structure of the silicon. This damage is generally repairable as was pointed out earlier. However the repair

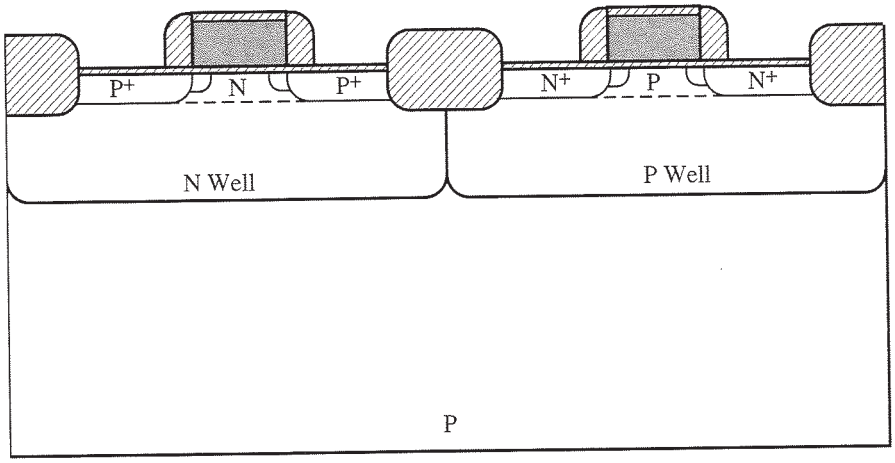


Figure 2–33 A final high-temperature drive-in activates all the implanted dopants and diffuses junctions to their final depth.

process takes some time (< 1 min at 1000°C or several hours at very low temperatures like 700°C). While this repair is occurring, dopants diffuse with anomalously high diffusivities because the damage enhances their diffusion coefficients. This phenomenon in known as Transient Enhanced Diffusion, or TED, and is a very important issue in keeping junctions shallow in scaled devices. We will discuss this in more detail in Chapter 8.

2.2.9 Contact and Local Interconnect Formation

All of the steps needed to form the active devices have now been completed. However, we obviously need to provide a means of interconnecting them on the wafer to form circuits, and a means to bring the input and output connections off the chip for packaging. The CMOS process we are describing will actually provide three levels of wiring to accomplish these objectives. The first or lowest level is often called the "local interconnect" and the steps needed to form it are shown in Figures 2–34 to 2–37. (Actually as we pointed out earlier, the polysilicon gate level itself can also be used as a local interconnect.)

The first step, illustrated in Figure 2–34, removes the oxide from the areas the interconnect is to contact. Since this is the bottom level of the interconnect structure, it will provide the connections to essentially all doped regions in the silicon and to all polysilicon regions. This oxide etch can actually be unmasked because the oxide is quite thin over the regions we wish to contact (the ≈ 10 nm screen oxide we grew just prior to source drain formation). A short dip in a buffered HF etching solution will remove these oxide layers, without significantly reducing the thickness of the oxide layers elsewhere.

The next step (Figure 2–35) involves the deposition of a thin layer (50 – 100 nm) of Ti on the wafer surface. This is usually done by sputtering from a Ti target. In a sputtering system, atoms of the desired material (Ti in this case) are physically knocked off a solid target by bombarding the Ti target with Ar^+ ions. The Ti atoms then deposit on

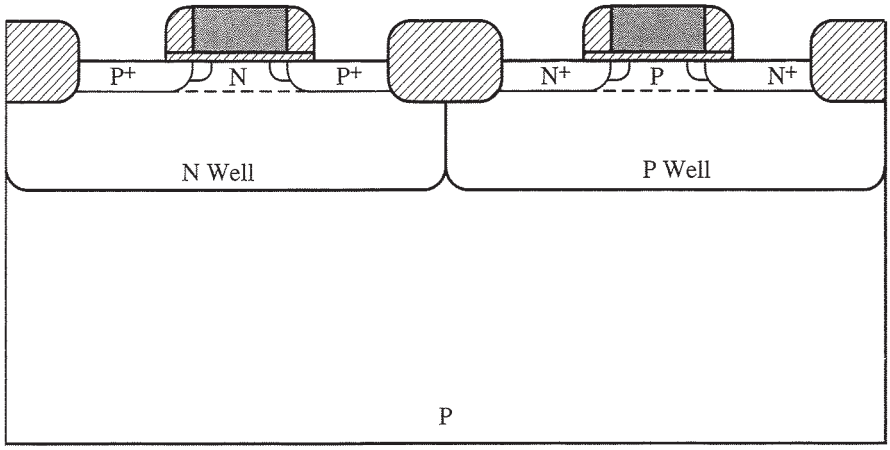


Figure 2–34 An unmasked oxide etch removes the SiO_2 from the device source drain regions and from the top surface of the polysilicon.

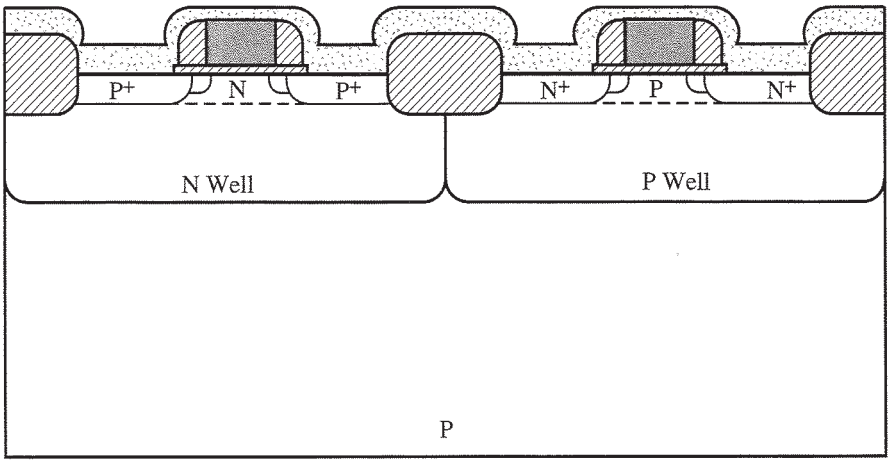


Figure 2–35 Titanium is deposited on the wafer surface by sputtering.

any substrates that are located nearby. This produces a continuous coating of Ti on the wafer as shown in Figure 2–35.

The next step, shown in Figure 2–36, makes use of two chemical reactions. The wafers are heated in an N_2 ambient at about 600°C for a short time (about 1 minute). At this temperature, the Ti reacts with Si where they are in contact to form $TiSi_2$, consuming some silicon in the process. This is why deeper source and drain junctions are required outside the tip or extension regions. $TiSi_2$ is an excellent conductor and forms low resistance contacts to both N^+ and P^+ silicon or polysilicon. This material is shown in black in Figure 2–36. The Ti also reacts with N_2 to form TiN (the dotted top layer in Figure 2–36). This material is also a conductor, although its conductivity is not as high as most metals. For this reason, it is used only for "local" or short-distance interconnects.

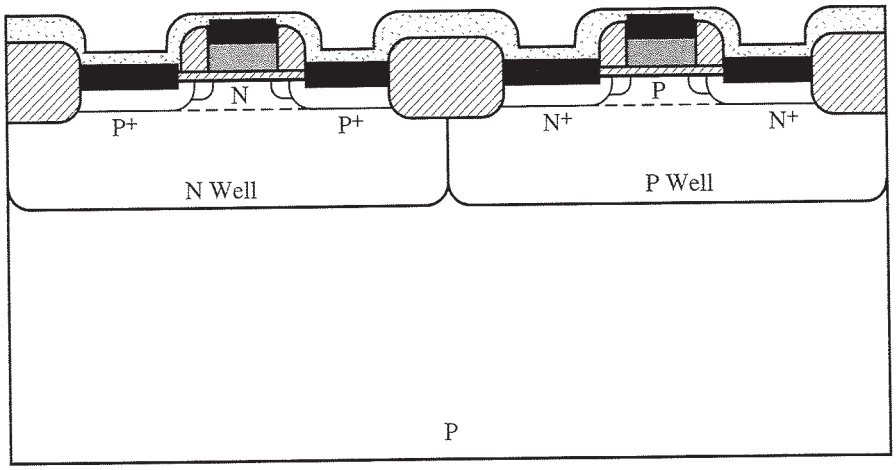


Figure 2–36 The titanium is reacted in an N_2 ambient, forming $TiSi_2$ where it contacts silicon or polysilicon (black regions in the figure) and TiN elsewhere.

The resistance of long lines made from TiN would cause unacceptable RC delays in most circuits.

Figure 2–37 illustrates the patterning of the TiN layer. Photoresist is applied and mask 11 protects the TiN where we want it to remain on the wafer. The remaining TiN is etched in $NH_4OH:H_2O_2:H_2O$ (1:1:5) to remove it. It is interesting to note at this point that the sidewall spacers we used earlier to provide graded N^+N^- or P^+P^- drain junctions also serve the function here of separating the $TiSi_2$ on the poly gates from contacting the silicon doped regions. After photoresist removal, the wafer would typically be heated in a furnace in an Ar ambient at about 800°C for about 1 minute to reduce the resistivity of the TiN and $TiSi_2$ layers to their final values (about 10 Ω/sq and 1 Ω/sq, respectively). The photoresist would be removed prior to this last high temperature anneal in an O_2 plasma or through chemical stripping, since photoresist cannot tolerate temperatures much above 100°C.

2.2.10 Multilevel Metal Formation

The final steps in our CMOS process involve the deposition and patterning of the two layers of metal interconnect. These steps are illustrated in Figures 2–38 to 2–44.

At this stage in the process, the surface of the wafer is highly nonplanar. We have grown and deposited many thin films on the surface and after these films are patterned, they leave numerous hills and valleys on the surface. It is not desirable to deposit the metal interconnect layers directly on such topography because there are potential problems with metal discontinuities (opens) at steps on the surface. There are also potential reliability problems even if the metal does not break at such steps, because it will likely be thinner where it crosses the steps. In addition, and perhaps most importantly, photolithography is very difficult with highly nonplanar substrates, especially when metal patterning is involved.

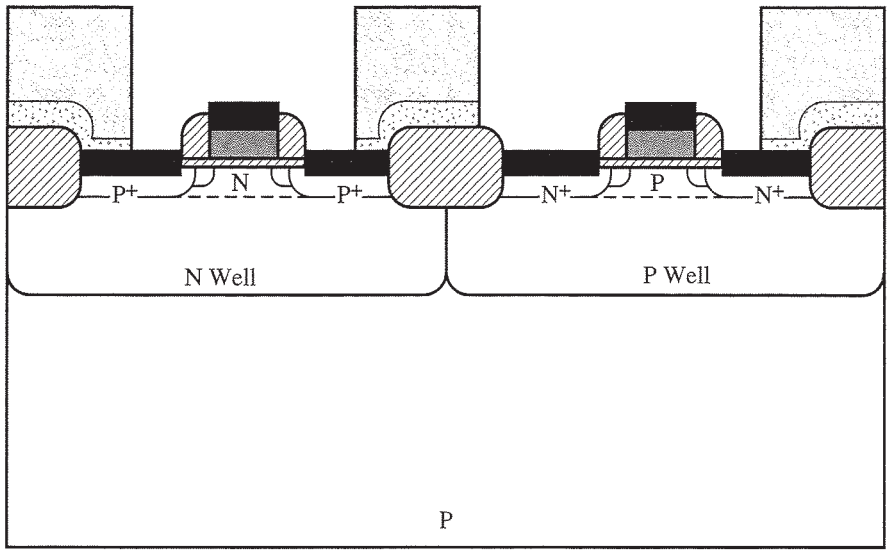


Figure 2–37 Photoresist is applied and mask 11 is used to define the regions where TiN local interconnects will be used. The TiN is then etched.

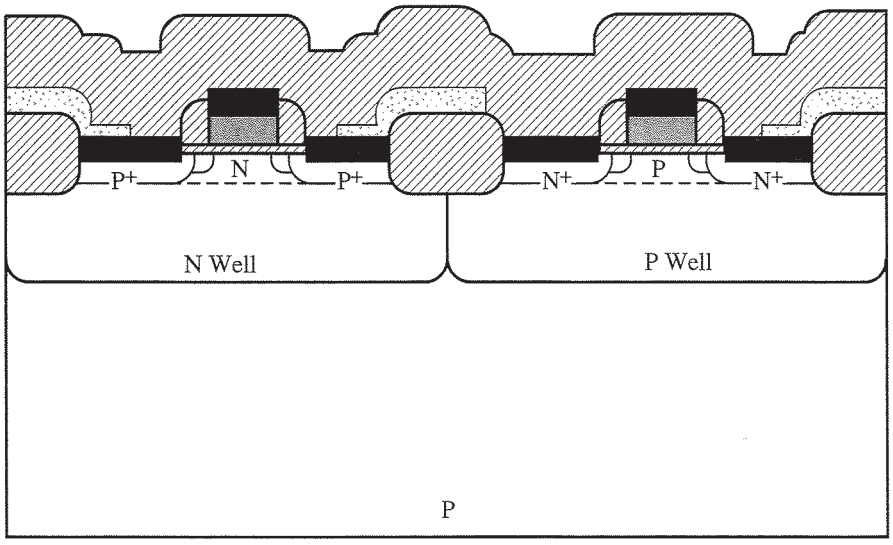


Figure 2–38 After stripping the photoresist, a conformal SiO_2 layer is deposited by LPCVD.

In an effort to circumvent these problems, many techniques have been devised to "planarize" or flatten the surface topography. One such method that is widely used is illustrated in Figures 2–38 and 2–39. A fairly thick SiO_2 layer is first deposited on the wafer surface by CVD or LPCVD. This layer is deposited thicker than the largest steps which exist on the surface and would typically be about 1 μm. This SiO_2 layer is often doped with phosphorus and sometimes with boron as well, in which cases the deposited oxide is known as PSG (phosphosilcate glass) or BPSG (borophosphosilicate glass), respectively. In some cases an undoped SiO_2 layer is added on top of the PSG or BPSG

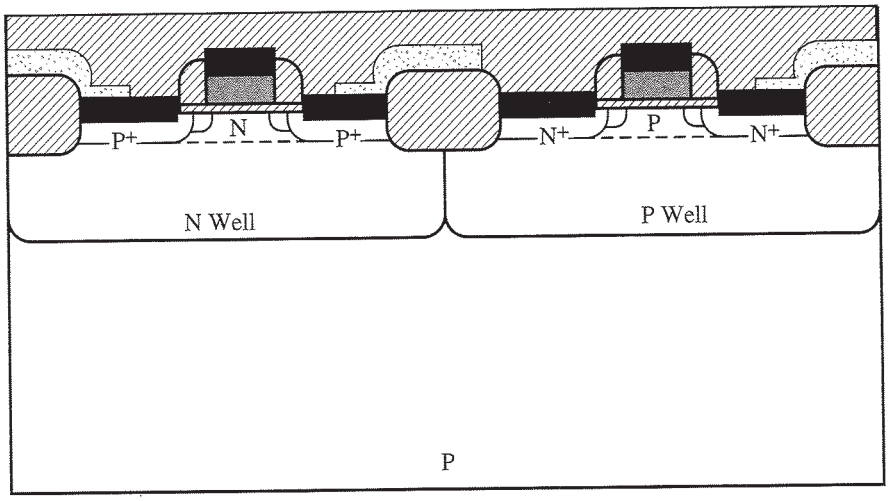


Figure 2–39 Chemical-Mechanical Polishing (CMP) or resist etchback is used to polish or etchback the deposited SiO_2 layer. This planarizes the wafer surface.

layer. The phosphorus provides some protection against mobile ions like Na^+ which can cause instabilities in MOS devices as was briefly mentioned in Chapter 1. The addition of the boron reduces the temperature at which the deposited glass layer "flows." This is important because following the deposition, the wafer is often heated to a temperature of 800 – 900°C which allows the glass to flow and to smooth the surface topography. Adding the boron minimizes the heat treatment required to accomplish this reflow, which is an issue because of the limited temperature tolerance of some of the underlying films at this point in the process. We will discuss these issues in more detail in Chapter 4 (Na^+ problems) and in Chapter 11 (glass reflow issues).

Reflowing the deposited PSG or BPSG layer is not sufficient to completely planarize the surface topography, so generally additional steps are required. For many years, this was commonly done by next spinning a layer of photoresist on to the wafer. Since the resist is a liquid, it will fill in the hills and valleys on the surface and produce a fairly flat upper surface. (We have actually drawn all the resist layers in earlier drawings this way, although without any explanation until this point.) The "trick" to accomplish planarization now takes place. It is possible to find a set of plasma etch conditions in which both resist and SiO_2 are etched at about the same rate. If we use such a plasma and etch the structure with no mask, then Figure 2–39 will result. We simply etch through the resist and down into the underlying oxide and stop when we have etched into the oxide everywhere. The 1:1 etch rate of resist and oxide preserves the originally flat resist upper surface as we etch.

Within the past few years, a replacement for this resist etchback technique has been commonly adopted in the semiconductor industry. This replacement is Chemical-Mechanical Polishing or CMP which we previously described in connection with the STI process option (Figures 2–8 and 2–9). In this process, the wafer is placed face down in a polishing machine and the upper surface is literally polished flat using a high-pH sil-

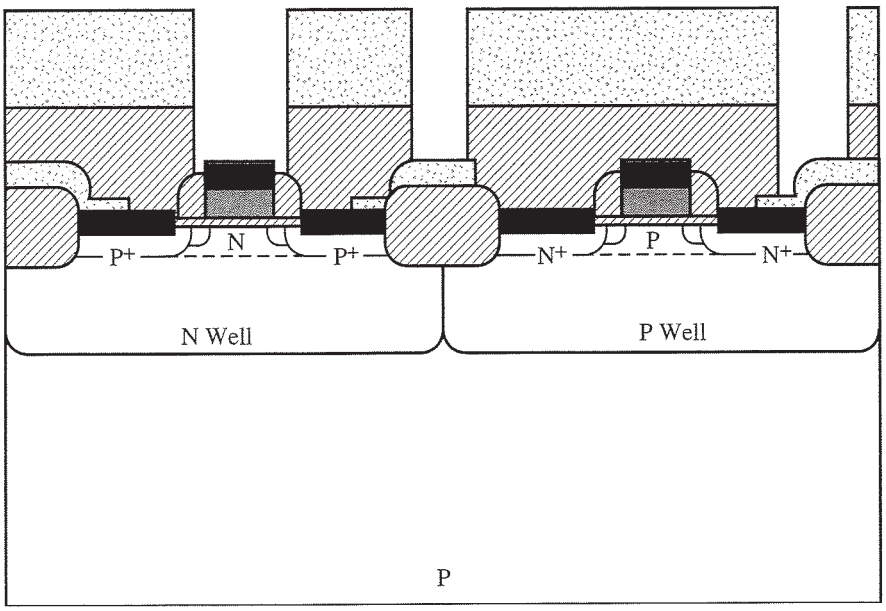


Figure 2–40 Photoresist is spun onto the wafer. Mask 12 is used to define the contact holes. The deposited SiO_2 layer is then etched to allow connections to the silicon, polysilicon and local interconnect regions.

ica slurry. The polishing process also results in the structure shown in Figure 2–39. CMP is discussed in more detail in Chapter 11.

The next step is again application of photoresist. We use mask 12 to define the regions where we want contact to be made between metal level 1 and underlying structures. This is shown in Figure 2–40. The SiO_2 layer would be etched in a plasma. After etching the contact holes, the photoresist would be stripped off the wafer.

We wish to maintain the planar surface as we add metal layers to the structure. While there are a number of process flows which can achieve this, the particular process we will describe here (Figure 2–41) is one of the more common. The first step is a blanket deposition of a thin TiN layer or Ti/TiN bilayer by sputtering or CVD. This layer is typically only a few tens of nm thick. It provides good adhesion to the SiO_2 and other underlying materials present in the structure at this point. The TiN also acts as an effective barrier layer between the upper metal layers and the lower local interconnect layers which connect to the active devices. The next step is deposition of a blanket W layer by CVD as illustrated in Figure 2–41. A typical reaction might be

$$WF_6 + 3H_2 \rightarrow W + 6HF \tag{2.10}$$

The next step, illustrated in Figure 2–42, again involves CMP to planarize the wafer. The polishing in this case removes the W and the TiN everywhere except in the contact holes and provides a planar surface on which the first level metal can be deposited. This process flow we have described, in which contact holes are etched, filled, and planarized, is known as the damascene process. It is very common in back-end processing in silicon chips today and will be discussed in more detail in Chapter 11.

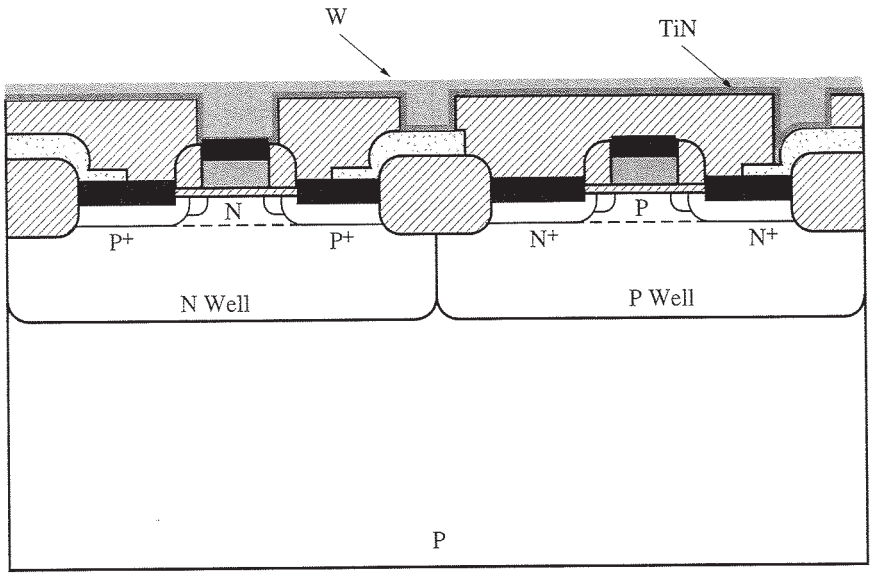


Figure 2–41 A thin TiN barrier/adhesion layer is deposited on the wafer by sputtering, followed by deposition of a W layer by CVD.

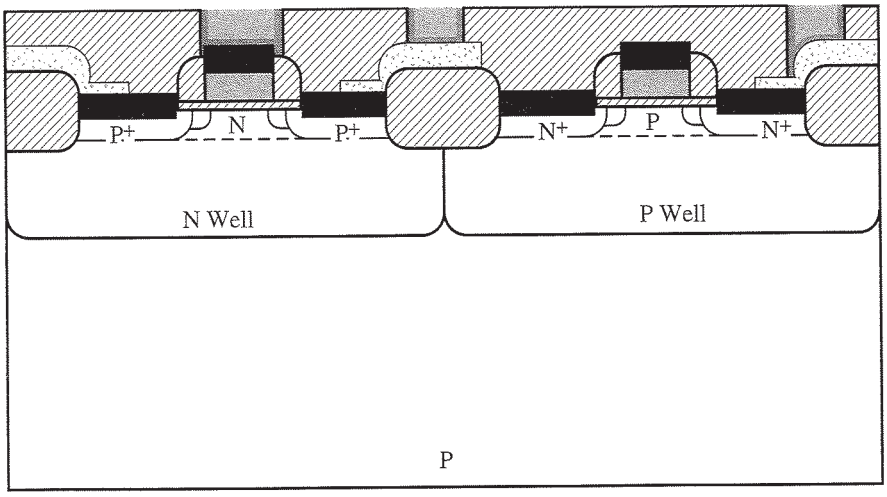


Figure 2–42 CMP is used to polish back the W and TiN layers, leaving a planar surface on which the first level of metal can be deposited.

Metal 1 is then deposited, usually by sputtering, and defined using resist and mask 13, as shown in Figure 2–43. The metal is commonly Al with a small percentage of Si and Cu in it. The Si is used because Si is soluble in Al up to a few percent and if the silicon is not already present in the Al, it may by absorbed by the Al from underlying silicon rich layers. This can cause problems with contact resistance and contact reliability. (Si absorption from underlying layers is less of a problem when barrier layers like TiN are used.) The Cu is added because it helps to prevent a reliability problem known as electromigration in Al thin films. This phenomenon causes open circuits in Al interconnects

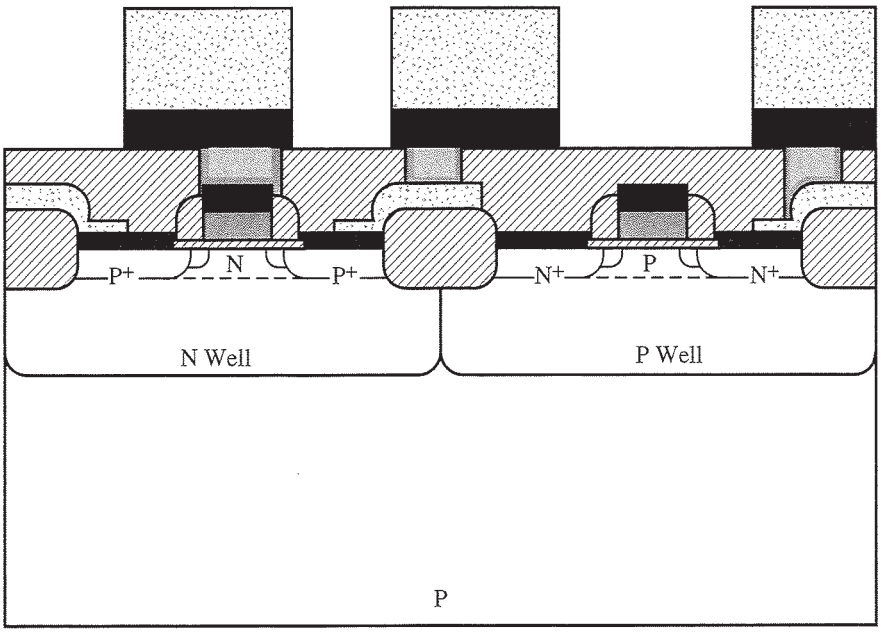


Figure 2–43 Aluminum is deposited on the wafer by sputtering. Photoresist is spun on the wafer and mask 13 is used to define the first level of metal. The A1 is then plasma etched.

after many hours of circuit operation, especially at elevated temperatures and high current densities, and is due to the formation of voids in the Al lines caused by diffusion of Al atoms. The Cu helps to prevent this from occurring.

Because of its better electrical conductivity, Cu is now beginning to replace Al as the interconnect metal. Cu is deposited using electroplating. Because Cu is quite difficult to etch using plasma etching, a somewhat different process flow is required than that described here for Al-based interconnects. Chapter 11 discusses these issues in more detail.

Most modern VLSI processes use more than one level of wiring on the wafer surface because in complex circuits it is usually very difficult to completely interconnect all the devices in the circuit without multiple levels. The processes that are used to deposit and define each level are similar to those we described for level 1 and usually involve a planarization step. Figure 2–44 illustrates this for the dielectric between metal 1 and metal 2. The process would again involve depositing an oxide layer and using CMP to planarize the deposited oxide. Figure 2–44 also illustrates the filling of the via holes between metal 1 and metal 2 with TiN and W, deposition and etching of metal 2, and the final deposition of a top dielectric to protect the finished chip. This top layer could be either SiO_2 or Si_3N_4 and is designed to provide some protection for the chip during the mechanical handling it will receive during packaging, as well as to provide a final passivation layer to protect the chip against ambient contamination (Na^+ or K^+). After the final processing steps are completed, an anneal and alloy step at a relatively low temperature (400 – 450°C) for about 30 minutes in forming gas (10% H_2 in N_2) is used to alloy the metal contacts in the structure and to reduce some of the electrical charges associated with the Si/SiO_2 interface. We will discuss these charges in more detail in

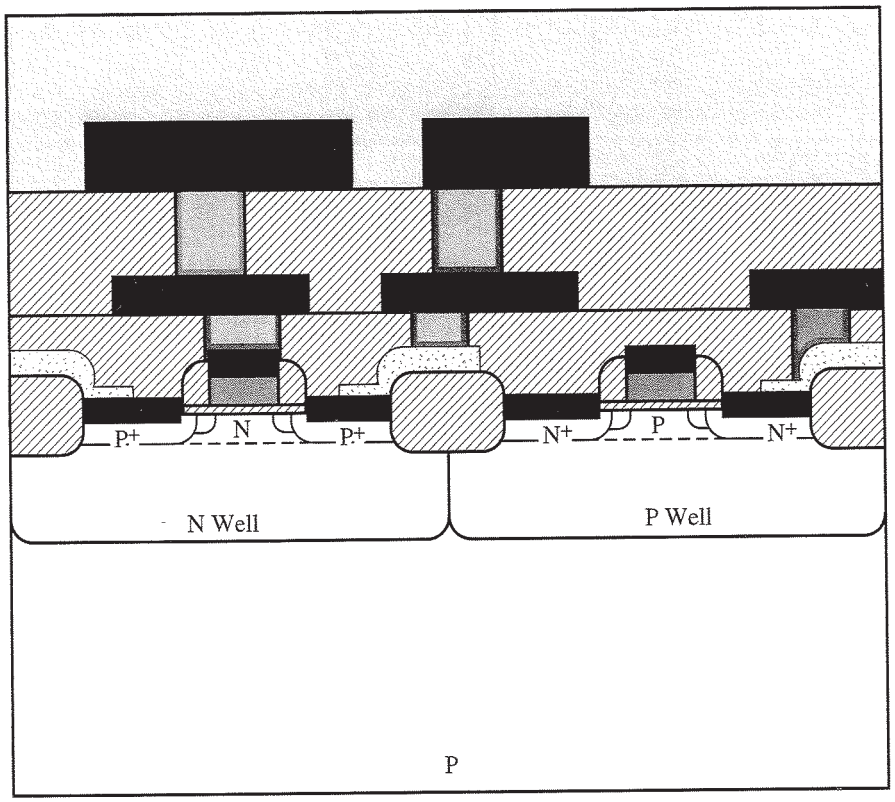


Figure 2–44 The steps to form the second level of Al interconnect follow those in Figures 2–38 to 2–43. Mask 14 is used to define via holes between metal 2 and metal 1. Mask 15 is used to define metal 2. The last step in the process is deposition of a final passivation layer, usually Si_3N_4 deposited by PECVD. The last mask (16) is used to open holes in this mask over the bonding pads.

Chapter 6. This finally brings us back to Figure 2–2 which is the completed CMOS chip that we started out to build.

2.3 Summary of Key Ideas

The purpose of this chapter was to describe in some detail a complete CMOS process flow. For readers relatively new to silicon technology, many new ideas were presented, often without full explanation or justification. All of these will be described in later chapters as we deal with the individual process steps. At this point, the context in which such processes are used should be clear.

A final point which is important to make is that the process we have described is not a unique way of achieving the final result shown in Figure 2–2. Many commercial companies and research laboratories today build chips with final cross sections similar to this figure, with however, quite different process details. As an example of the differences between commercially available CMOS process flows, the process described here required 16 masks through two levels of metal. Commercial processes range from simple CMOS with one level of metal and perhaps 10 masks, to very complex processes

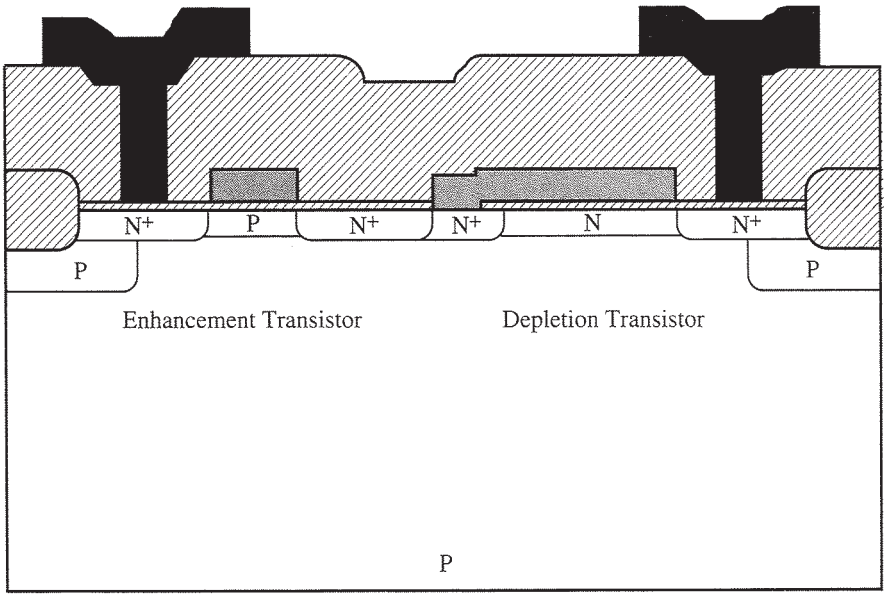
with five to six levels of metal and 20–25 masks. The reasons for these differences from one process to another may have to do with specific types of equipment a particular laboratory or plant has, or they may have to do with the applications targeted for the technology. Trade-offs in technology complexity and device performance may lead an individual company to a process flow quite different than the one we have described. Some of these trade-offs will become clearer as we discuss the individual process steps in later chapters.

Many of the process steps described in this chapter could be simulated with modern Technology Computer-Aided Design (TCAD) tools. We have not chosen to include such simulations in this chapter because the objective here was simply a qualitative description of a CMOS process flow. As we discuss the various technology steps in detail in subsequent chapters, we will also introduce simulation tools for those process steps.

2.4 Problems

2.1 Sketch a process flow that would result in the structure shown in Figure 1–34 by drawing a series of drawings similar to those in this chapter. You only need to describe the flow up through the stage at which active device formation starts since from that point on, the process is similar to that described in this chapter.

2.2. During the 1970s, the dominant logic technology was NMOS as described briefly in Chapter 1. A cross sectional view of this technology is shown below (see also Figure 1–33). The depletion mode device is identical to the enhancement mode device except that a separate channel implant is done to create a negative threshold voltage. Design a plausible process flow to fabricate such a structure, following the ideas of the CMOS process flow in this chapter. You do not have to include any quantitative process parameters (times, temperatures, doses, etc.) Your answer should be given in terms of a series of sketches of the structure after each major process step, like the figures in this chapter. Briefly explain your reasoning for each step and the order you choose to do things.



2.3. The cross section below illustrates a simple bipolar transistor fabricated as part of a silicon IC. (See also Figure 1–32.) Design a plausible process flow to fabricate such a structure, following the ideas of the CMOS process flow in this chapter. You do not have to include any quantitative process parameters (times, temperatures, doses, etc.) Your answer should be given in terms of a series of sketches of the structure after each major process step, like the figures in this chapter. Briefly explain your reasoning for each step and the order you choose to do things.