Chapter

7

CMOS Fabrication

This chapter presents a brief overview of CMOS process integration. Process integration refers to the well-defined collection of semiconductor processes required to fabricate CMOS integrated circuits starting from virgin silicon wafers. This overview is intended to give the reader a fundamental understanding of the processes required in CMOS integrated circuit fabrication. Moreover, there are strong interactions between circuit design and process integration. For instance, the typical design rule set is determined in large part by the limitations in the fabrication processes. Hence, circuit designers, process engineers, and integration engineers are required to communicate effectively. To this end, we first examine the fundamental processes, called unit processes, required for CMOS fabrication. The primary focus is the qualitative understanding of the processes with limited introduction of quantitative expressions. The unit processes are combined in a deliberate sequence to fabricate CMOS. Additionally, the unit processes are typically repeated numerous times in a given process sequence. Here we present a representative modern CMOS process sequence, also called a process flow.

7.1 CMOS Unit Processes

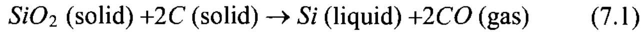
In this section we introduce each of the major processes required in the fabrication of CMOS integrated circuits. We first discuss wafer production. Although wafer production is not a unit process, it is nonetheless important to present the production method which is used by wafer manufacturers. All subsequent discussions are focused on the unit processes incorporated by fabrication facilities to produce integrated circuits. The unit processes are grouped by functionality. Thermal oxidation, doping processes, photolithography, thin-film removal, and thin-film deposition techniques are presented.

7.1.1 Wafer Manufacture

Silicon is the second most abundant element in the Earth's crust; however, it occurs exclusively in compounds. In fact, elemental silicon is a man-made material that is refined from these various compounds. The most common is silica (impure SiO_2). Modern integrated circuits must be fabricated on ultrapure, defect-free slices of single crystalline silicon called wafers, as discussed in Ch. 1.

Metallurgical Grade Silicon (MGS)

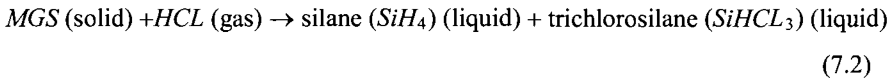
Wafer production requires three general processes: silicon refinement, crystal growth, and wafer formation. Silicon refinement begins with the reduction of silica in an arc furnace at roughly 2000 °C with a carbon source. The carbon effectively "pulls" the oxygen from the SiO_2 molecules, thus chemically reducing the SiO_2 into roughly 98% pure silicon referred to as metallurgical grade silicon (MGS). The overall reduction is governed by the following equation



$$SiO_2 \text{ (solid)} + 2C \text{ (solid)} \rightarrow Si \text{ (liquid)} + 2CO \text{ (gas)} \quad (7.1)$$

Electronic Grade Silicon (EGS)

MGS is not sufficiently pure for microelectronic device applications. The reason is that the electronic properties of a semiconductor such as silicon are extremely sensitive to impurity concentrations. Impurity levels measured at parts per million or less can have dramatic effects on carrier mobilities, lifetimes, etc. It is therefore necessary to further purify the MGS in what is known as electronic grade silicon (EGS). EGS is produced from the chlorination of grounded MGS as



$$MGS \text{ (solid)} + HCL \text{ (gas)} \rightarrow \text{silane } (SiH_4) \text{ (liquid)} + \text{trichlorosilane } (SiHCL_3) \text{ (liquid)} \quad (7.2)$$

Because the reaction products are liquids at room temperature, ultrapure EGS can be obtained from fractional distillation and chemical reduction processes. The resultant EGS is in the form of polycrystalline chunks.

Czochralski (CZ) Growth and Wafer Formation

To achieve a single crystalline form, the EGS must be subjected to a process called Czochralski (CZ) growth. A schematic representation of the CZ growth process is shown in Fig. 7.1. The polycrystalline EGS is melted in a large quartz crucible where a small seed crystal of known orientation is introduced into the surface of the silicon melt. The seed crystal, rotating in one direction, is slowly pulled from the silicon melt, rotating in the opposite direction. Solidification of the silicon onto the seed forms a growing crystal

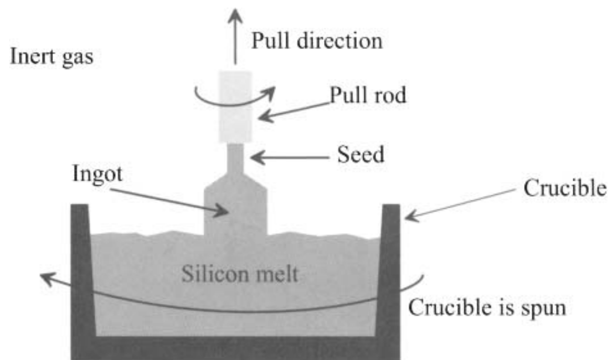


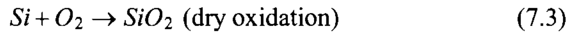
Figure 7.1 Simplified diagram showing Czochralski (CZ) crystal growth.

(called a boule or ingot) that assumes the crystallographic orientation of the seed. In general, the slower the pull-rate (typically mm/hour), the larger the diameter of the silicon crystal. Following CZ growth, the silicon boule is trimmed down to the appropriate diameter. *Flats* or *notches* are ground into the surface of the boule to indicate a precise crystal orientation. Using a special diamond saw, the silicon boule is cut into thin wafers. The wafers are finished by using a chemical mechanical polishing (CMP) process to yield a mirror-like finish on one side of the wafer. Although devices are fabricated entirely within the top couple of micrometers of the wafer, final wafer thicknesses (increasing with wafer diameter) are up to roughly one millimeter for adequate mechanical support.

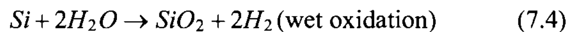
7.1.2 Thermal Oxidation

Silicon, when exposed to an oxidant at elevated temperatures, readily forms a thin layer of oxide at all exposed surfaces. The native oxide of silicon is in the form of silicon dioxide (SiO_2). With respect to CMOS fabrication, SiO_2 can serve as a high quality dielectric in device structures such as gate oxides. Moreover, during processing, thermally grown oxides can be used as implantation, diffusion, and etch masks. The dominance of silicon as a microelectronic material can be attributed to the existence of this high quality native oxide and the resultant near ideal silicon/oxide interface.

Figure 7.2 depicts the basic thermal oxidation process. The silicon wafer is exposed at high temperatures (typically 900 °C–1200 °C) to a gaseous oxidant such as molecular oxygen (O_2) and/or water vapor (H_2O). For obvious reasons, oxidation in O_2 is called dry oxidation, whereas in H_2O it is called wet oxidation, as discussed in Sec. 2.1. The gas/solid interface forms a stagnant layer through which the oxidant must diffuse to reach the surface of the wafer. Once at the surface, the oxidant again must diffuse through the existing oxide layer that is present. As the oxidant species reaches the silicon/oxide interface, one of two reactions occur



$$Si + O_2 \rightarrow SiO_2 \text{ (dry oxidation)} \tag{7.3}$$



$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \text{ (wet oxidation)} \tag{7.4}$$

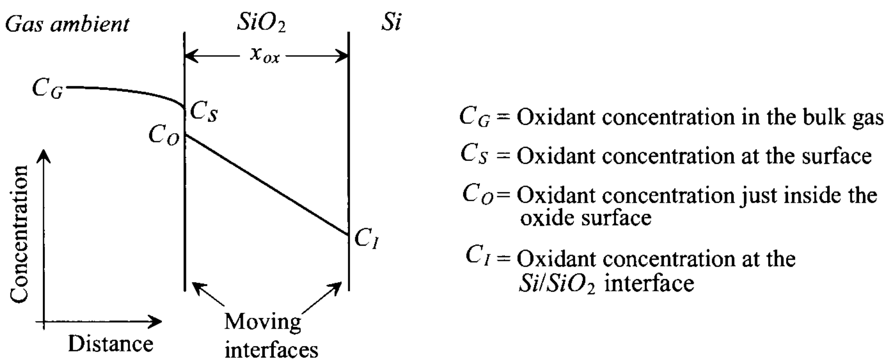


Figure 7.2 A simple model for thermal oxidation of silicon. Notice the oxidant concentrations (boundary conditions) in the gas, oxide, and silicon.

It should be emphasized that reactions specified by Eqs. (7.3) and (7.4) occur at the silicon/oxide interface where silicon is consumed in the reaction. As Fig. 7.3 illustrates, with respect to the original silicon surface, approximately 45% of the oxide thickness is accounted for by consumption of silicon.

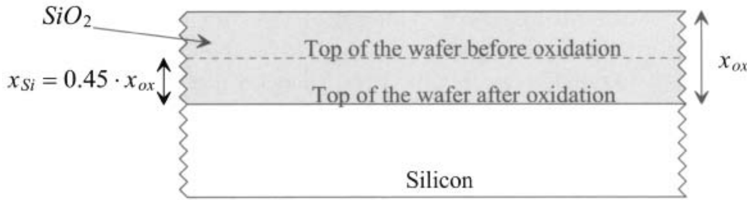


Figure 7.3 Silicon/oxide growth interface. See also Fig. 2.4.

The rate of thermal oxidation is a function of temperature and rate constants. The rate is directly proportional to temperature. The rate constants are, in turn, a function of gas partial pressures, oxidant-type, and silicon wafer characteristics such as doping type, doping concentration, and crystallographic orientation. In general, dry oxidation yields a denser and thus higher quality oxide than does a wet oxidation. However, wet oxidation occurs at a much higher rate compared to dry oxidation. Depending on the temperature and existing thickness of oxide present, the overall oxidation rate can be either diffusion limited (e.g., thick oxides at high temperatures) or reaction rate limited (e.g., thin oxides at low temperatures). Practically, oxide thicknesses are limited to less than a few thousand angstroms and to less than a micron for dry and wet oxidation, respectively.

In a modern fabrication facility, oxidation occurs in either a tube furnace or in a rapid thermal processing (RTP) tool, as schematically shown in Fig. 7.4. The tube furnaces consist of quartz tubes surrounded by heating element coils. The wafers are loaded in the heated tubes where oxidants can be introduced through inlets. The function of the RTP is similar to the tube furnace with the exception that the thermal source is heating lamps.

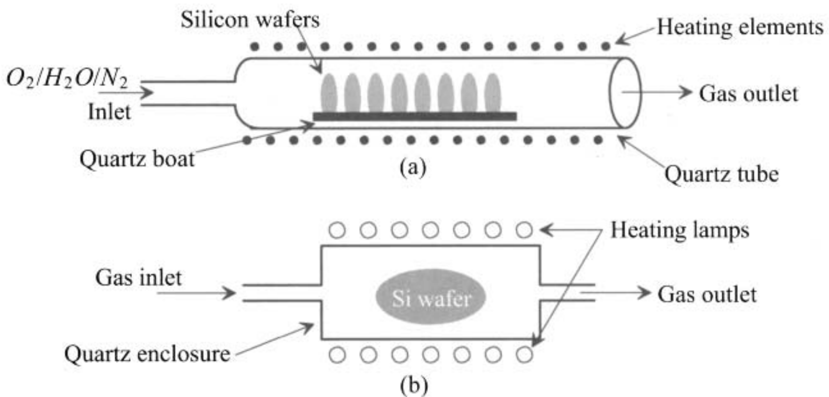


Figure 7.4 (a) Simplified representation of an oxidation tube furnace and (b) simplified diagram for rapid thermal processing.

7.1.3 Doping Processes

The controlled introduction of dopant impurities into silicon is necessary to affect majority carrier type, carrier concentration, carrier mobility, carrier lifetime, and internal electric fields. The two primary methods of dopant introduction are solid state diffusion and ion implantation. Historically, solid state diffusion has been an important doping process; however, ion implantation is the preferred method in modern CMOS fabrication.

Ion Implantation

The workhorse method of introducing dopants into the near-surface region of wafers is a process called ion implantation. In ion implantation, dopant atoms (or molecules) are ionized and then accelerated through a large electric potential (a few kV to MV) towards a wafer. The highly energetic ions bombard and thus implant into the surface. Obviously, this process leads to a high degree of lattice damage, which is generally repaired by annealing at high temperatures. Moreover, the ions do not necessarily come to rest at a lattice site, hence an anneal is required to electrically activate (i.e., thermally agitate the impurities into lattice sites) the dopant impurities.

Figure 7.5 shows a schematic diagram of an ion implanter. The ions are generated by an RF field in the ion source where they are subsequently extracted to a mass spectrometer. The spectrometer only allows ions with a user-selected mass to enter the accelerator, where the ions are passed through a large potential field. The ions are then scanned via electrostatic lens across the surface of the wafer.

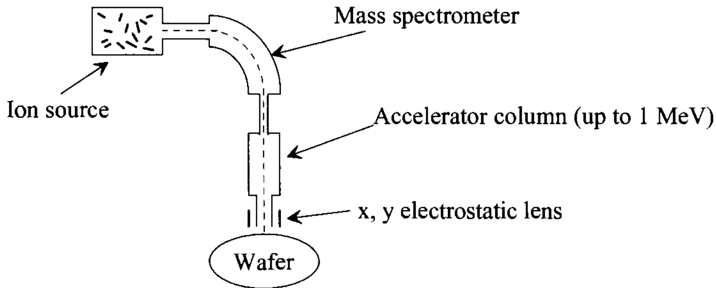


Figure 7.5 Simplified diagram of an ion implanter. The ions are created by an RF field where they are extracted into a mass spectrometer. An electrostatic lens scans the ion beam on the surface of a wafer to achieve the appropriate dose. Electrostatically, the ions can be counted to provide the real-time dose.

A first-order model for an implant doping profile is given by a Gaussian distribution described mathematically as

$$N(x) = N_p \exp\left[-(x - R_p)^2 / 2\Delta R_p^2\right] \quad (7.5)$$

where N_p is the peak concentration, R_p is the projected range, and ΔR_p is called the straggle. By inspection, R_p should be identified as the mean distance the ions travel into the silicon and ΔR_p as the associated standard deviation. Figure 7.6 illustrates a typical ion implant profile. Obviously, N_p occurs at a depth of R_p. Moreover, the area under the

implant curve corresponds to what is referred to as the implant dose Q_{imp}, given mathematically as

$$Q_{imp} = \int_0^{\infty} N(x) \cdot dx \tag{7.6}$$

Localized implantation is achieved by masking off regions of the wafer with an appropriately thick material such as oxide, silicon nitride, polysilicon, or photoresist. Since implantation occurs in the masking layer, the thickness must be of sufficient magnitude to stop the ions prior to reaching the silicon substrate. In comparison to solid state diffusion, ion implantation has the advantage of being a low temperature and highly controlled process.

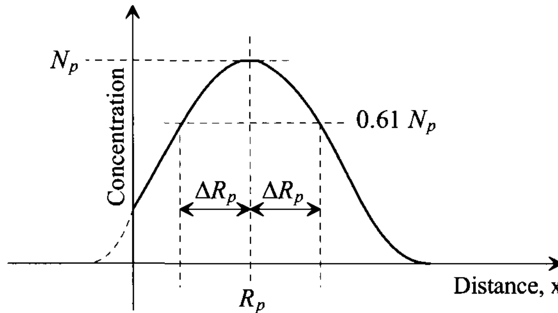


Figure 7.6 Ideal implant profile representing Eq. (7.5). Notice that the peak concentration occurs below the surface and depends on the implant energy.

Solid State Diffusion

Solid state diffusion is a method for introducing and/or redistributing dopants. In this section, we study solid state diffusion primarily to gain insight into "parasitic" dopant redistribution during thermal processes. In typical CMOS process flows, dopants are introduced into localized regions via ion implantation where the subsequent processing often consists of high temperature processing. Solid state diffusion inherently occurs in these high temperature steps, thus spreading out the implant profile in three dimensions. The net effect is to shift the boundary of the implant from its original implant-defined position, both laterally and vertically. This thermal smearing of the implant profiles must be accounted for during CMOS process flow development. If not, the final device characteristics can differ significantly from what was expected.

Solid state diffusion (or simply diffusion) requires two conditions: 1) a dopant concentration gradient, and 2) thermal energy. Diffusion is directly proportional to both. An implanted profile (approximated by a delta function at the surface of the wafer) diffuses to first order as

$$N(x,t) = \frac{Q_{imp}}{\sqrt{\pi \cdot D \cdot t}} \cdot \exp(-x^2/4Dt) \tag{7.7}$$

where Q_{imp} is the implant dose, D is the diffusivity of the dopant, and t is the diffusion time. Figure 7.7 illustrates limited-source diffusion of a one-dimensional implant profile. Notice that the areas under the respective curves for a given time are equal.

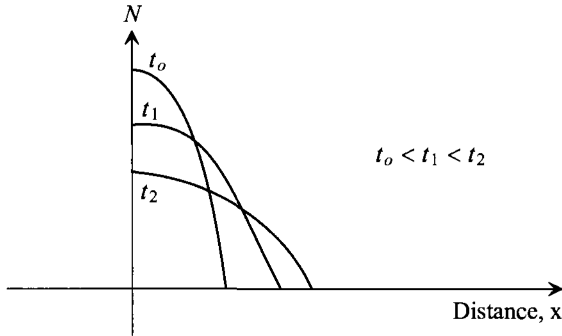


Figure 7.7 Idealized limited-source diffusion profile showing the effects of drive-in time on the profile. Notice that the peak concentration occurs at the surface of the substrate (x = 0) and that the area under the curves is constant.

7.1.4 Photolithography

In the fabrication of CMOS, it is necessary to localize processing effects to form a multitude of features simultaneously on the surface of the wafer. The collection of processes that accomplish this important task using an ultraviolet light, a mask, and a light-sensitive chemical resistant polymer is called *photolithography*. Although there are many different categories of photolithography, they all share the same basic processing steps that result in micron-to-submicron features generated in the light-sensitive polymer called photoresist. The photoresist patterns can then serve as ion implantation masks and etch masks during subsequent processing steps.

Figure 7.8 outlines the major steps required to implement photolithography patterning of a thermally grown oxide. Photoresist, a viscous liquid polymer, is applied to the top surface of the oxidized wafer. The application typically occurs by dropping (or spraying) a small volume of photoresist to a rapidly rotating wafer yielding a uniform thin film on the surface. Following spinning, the coated wafer is softbaked on a hot plate which drives out most of the solvents from the photoresist and improves the adhesion to the underlying substrate. Next, the wafers are exposed to ultraviolet light through a mask (or reticle) that contains the layout patterns for a given drawn layer. Unless the first layer is being printed, the exposure must be preceded by a careful alignment of mask features to existing patterns on the wafer. There are three general methods of exposing (patterning) the photoresist: contact, proximity, and projection photolithography. In both contact and proximity photolithography, the mask and the wafers are in contact and in close proximity, respectively, to the surface of the photoresist. Here the mask features are of the same scale as the features to be exposed on the surface.

In projection photolithography, the dominant type of patterning technology, the mask features are on a larger scale (e.g., 5X or 10X) relative to the features exposed on the surface. This is accomplished with a projection *stepper*. Using reduction with optics, the stepper projects an image through the mask to the photoresist on the surface. For positive-tone photoresist, the ultraviolet light breaks molecular bonds, making the exposed regions more soluble in the developer. In contrast, for negative-tone photoresist, the exposure causes polymerization and thus less solubility. The exposed resist-coated wafer is developed in an alkaline solution. Depending on the formulation of the photoresist, a positive or negative image relative to the mask patterns can be generated.

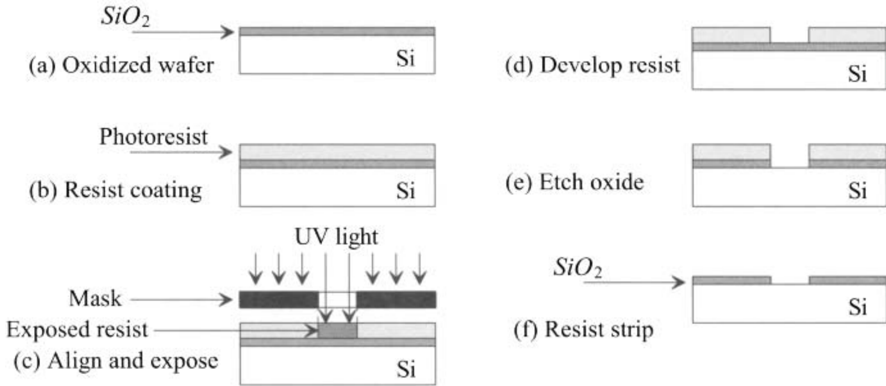


Figure 7.8 Simplified representation of the primary steps required for the implementation of photolighography and pattern transfer.

To harden the photoresist for improved etch-resistance and to improve adhesion, the newly developed wafers are often hardbaked. At this point, the wafer can be etched to transfer the photoresist pattern into the underlying oxide film.

Resolution

In general, there are three critical parameters associated with a given projection stepper: resolution, depth of focus, and pattern registration. The diffraction of light caused by the various interfaces in its path limits the minimum printable feature size as depicted in Fig. 7.9. Resolution is defined as the minimum feature size, M, that can be printed on the surface of the wafer given by

$$M = \frac{c_1 \cdot \lambda}{NA} \tag{7.8}$$

where lambda, λ, is the wavelength of the ultraviolet light source, NA is the numerical aperture of the projection lens, and c_1 is a constant whose value ranges from 0.5 to 1. The NA of a lens is illustrated in Figure 7.10 and is given mathematically as

$$NA = n \cdot \sin\theta \tag{7.9}$$

where n is the index of refraction of the space between the wafer and the lens and θ is the acute angle between the focal point on the surface of the wafer and the edge of the lens radii. Notice that M is directly proportional to wavelength, hence diffraction effects are the primary limitation in printable feature size. To a limit, the NA of the projection lens can be increased to help combat the diffraction effects because large NA optics have an increased ability to capture diffracted light. At first glance, the minimum feature size cannot be less than the wavelength of the light, however, advance techniques such as optical proximity correction (OPC) and wavefront engineering of the photomasks have been developed to push the resolution limits below the wavelength.

Depth of Focus (DOF)

The depth of focus (DOF) of the projection optics limits one's ability to pattern features at different heights, as illustrated in Fig. 7.11. Mathematically, DOF is given by

$$\text{DOF} = \frac{c_2 \lambda}{NA^2} \tag{7.10}$$

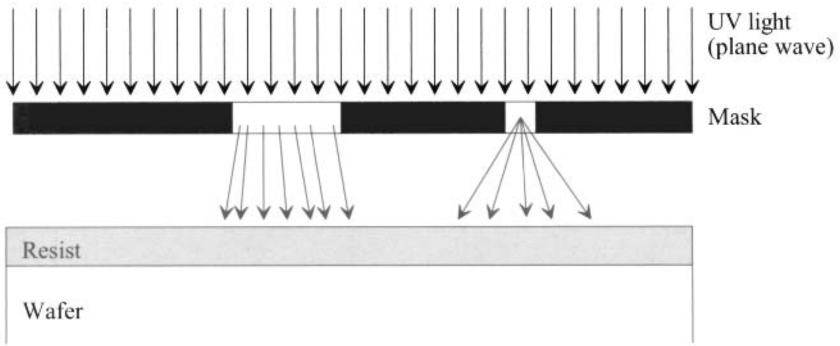


Figure 7.9 The diffraction effects become significant as the mask feature dimensions approach the wavelength of UV light. Notice that the diffraction angle is larger for the smaller opening.

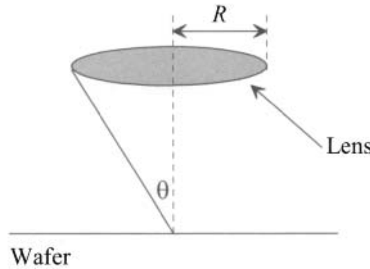


Figure 7.10 The relationship of the lens radii to the angle used to compute NA.

where c_2 is a constant ranging in value from 0.5 to 1. As apparent from Eq. (7.8) and Eq. (7.10), there exists a fundamental trade-off between minimum feature size and DOF. In other words, to print the smallest possible features, the surface topography must be minimized.

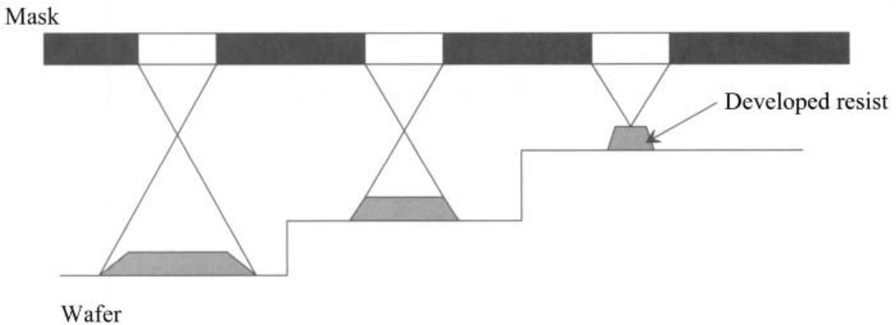


Figure 7.11 Depth of focus diagram illustrating the need to have planar surfaces (minimized topography) during high resolution photopatterning.

Aligning Masks

During CMOS fabrication, numerous mask levels (e.g., active, poly, contacts, etc.) are printed on the wafer. Each of these levels must be accurately aligned to one another. Registration is a measure of the level-to-level alignment error. Registration errors occur in x, y, and z-rotations, as illustrated in Fig. 7.12.

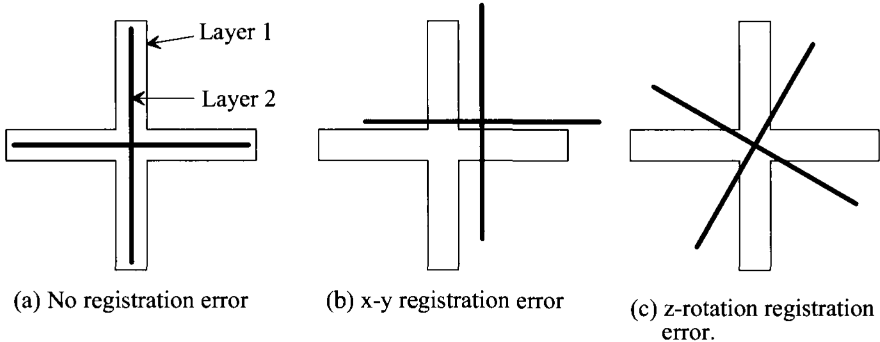


Figure 7.12 Simple registration errors that can occur during wafer-to-mask alignment in photolithography. Other registration errors exist but are not discussed here.

7.1.5 Thin Film Removal

Typically one of two processes are performed following photolithography. One is thin film etching used to transfer the photoresist patterns to the underlying thin film(s). The other is ion implantation using the photoresist patterns to block the dopants from select regions on the surface of the wafer. In this section, we discuss thin-film etching processes based on wet chemical etching and dry etching techniques. Additionally, we discuss a process used to remove unpatterned thin-films called chemical mechanical polishing (CMP).

Thin Film Etching

Once a photoresist pattern is generated there are two commonly employed approaches, wet etching and dry etching, to transfer patterns into underlying films. Etch rate (thickness removed per unit time), selectivity, and degree of anisotropy are key parameters for both wet and dry etching. Etch rates are typically strong functions of solution concentration and temperature. Selectivity, S, is defined as the etch rate ratio of one material to another given by the selectivity equation

$$S=\frac{R_2}{R_1} \tag{7.11}$$

where R_2 is the etch rate of the material intended to be removed and R_1 is the etch rate of the underlying, masking, or adjacent material not intended to be removed. The degree of anisotropy, A_f, is a measure of how rapidly an etchant removes material in different directions, mathematically given by

$$A_f=1-\frac{R_l}{R_v} \tag{7.12}$$

where R_l is the lateral etch rate and R_v is the vertical etch rate. Notice that if $A_f = 1$, then the etchant is completely anisotropic. However, if $A_f = 0$, then the etchant is completely isotropic. In conjunction with photolithography, the degree of anisotropy is a major factor in the achievable resolution. Figure 7.13 illustrates the effects of etch bias (i.e., $d_{film} - d_{mask}$) on the final feature size. For the submicron features that are required in CMOS, dry etch techniques are preferred over wet etch processes. This is due to the fact that dry etch techniques can, in general, have a higher degree of anisotropy. Both wet and dry etching are applied to the removal of metals, semiconductors, and insulators.

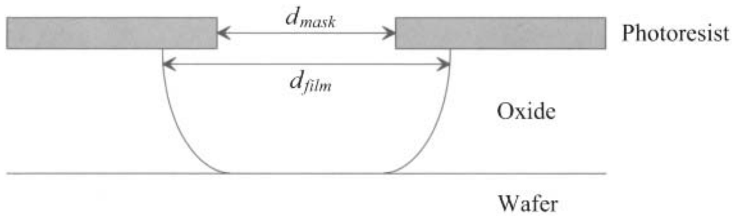


Figure 7.13 Diagram showing a post-etch profile. Notice that because of isotropy in the etch process the mask opening does not match the fabricated opening in the underlying oxide film. The difference between these dimensions is called etch-bias.

Wet Etching

Wet etching consists of using a chemical solution to remove material. In CMOS fabrication, wet processes are used both for cleaning of wafers and for thin-film removal. Wet cleaning processes are repeated numerous times throughout a process flow. Some cleaning processes are targeted to particulate removal, while others are for organic and/or inorganic surface contaminants. Wet etchants can be isotropic (i.e., etch rate is the same in all directions) or anisotropic (i.e., etch rate differs in different directions) although most of the wet etchants used in CMOS fabrication are isotropic. In general, wet etchants tend to be highly selective compared to dry etch processes. A schematic diagram of a wet etch tank is shown in Fig. 7.14. To improve the etch uniformity and to aid particulate removal, it is common to ultrasonically vibrate the etchant, as shown in the figure. Furthermore, microcontrollers accurately control the temperature of the bath. Once the etch is completed, the wafers are rinsed in deionized (DI) water, then spun dried.

Dry Etching

In CMOS fabrication, there are three general categories of dry etch techniques: sputter etching, plasma etching, and reactive ion etching (RIE). Figure 7.15 schematically illustrates a sputter etch process. An inert gas (e.g., argon) is ionized where the ions are accelerated through an electric field established between two conductive electrodes, called the anode and the cathode. A vacuum in the range of millitorr must exist between the plates to allow the appropriate ionization and transfer of ions. Under these conditions, a glow discharge, or plasma, is formed between the electrodes. In simple terms, the plasma consists of positively charged ions and electrons, which respond oppositely to the electric field. The wafer sits on the cathode where it is bombarded by the positively charged ions, causing material to be ejected off of the surface. Essentially, sputter etching

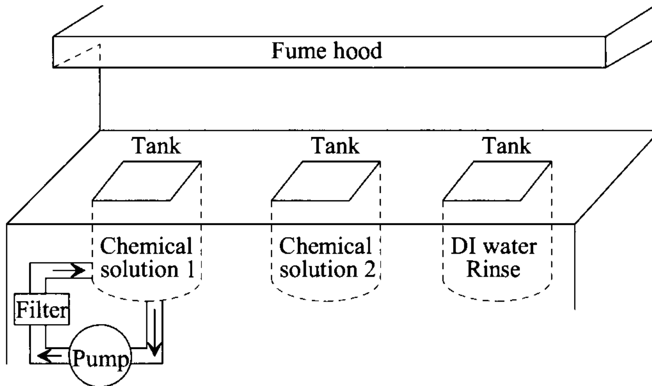


Figure 7.14 Simplfied diagram of a wet bench used for wet chemical cleaning and etching

is atomic-scale sandblasting. A DC power supply can be used for sputter etching conductive substrates, while an RF supply is required through capacitive coupling for etching non-conductive substrates. Sputter etching tends not to be selective, but it is very anisotropic.

A simplified diagram of a plasma etch system is shown in Figure 7.16. A gas or mixture of gases (e.g., halogens) are ionized, producing reactive species called radicals. A glow discharge or plasma is formed between the electrodes. The radicals chemically react with the surface material forming reaction products in the gas phase which are pumped away through a vacuum system. Plasma etching can be very selective, but is typically highly isotropic.

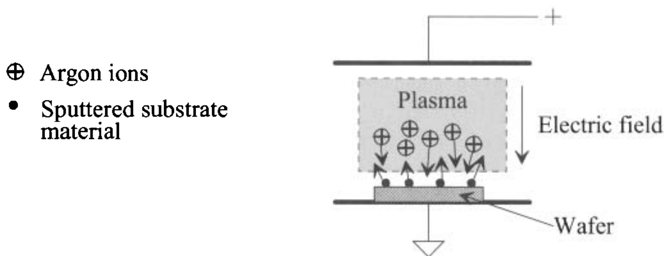


Figure 7.15 Simplified schematic diagram of the sputter etch process. This process is dominated by the physical bombardment of ions on a substrate.

While sputter etching is a purely mechanical process and plasma etching is purely chemical, RIE is a combination of sputter etching and plasma etching, as schematically shown in Figure 7.17. In RIE, a gas or mixture of gases (e.g., fluorocarbons) are ionized where radicals and ionized species are generated, both of which interact with the surface of the wafer. RIE is the dominant etch process because it can provide the benefits of both sputter etching and plasma etching. In other words, RIE can be highly selective and highly anisotropic.

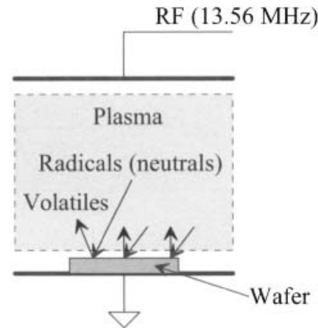


Figure 7.16 Simplified schematic diagram of a plasma etch process. This process is dominated by the chemical reactions of radicals at the surface of the substrate.

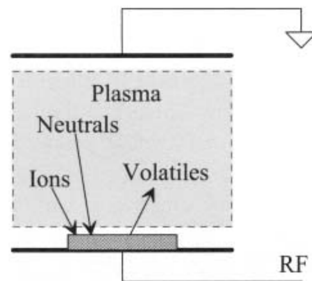


Figure 7.17 Simplified schematic diagram of an RIE etch process. This process has both physical (ion bombardment) and chemical (reaction of radicals) components.

Chemical Mechanical Polishing

Figure 7.18 depicts the key features of chemical mechanical polishing (CMP). In CMP, an abrasive chemical solution, called a slurry, is introduced between a polishing pad and the wafer. Material on the surface of the wafer is removed by both a mechanical polishing component and a chemical reaction component. In modern CMOS fabrication, CMP is a critical process that is used to planarize the surface of the wafer prior to photolithography. The planar surface allows the printed feature size to be decreased. CMP can be used to remove metals, semiconductors, and insulators.

7.1.6 Thin Film Deposition

Insulators, conductors, and semiconductors are all required for CMOS integrated circuits. Semiconductors, such as crystalline silicon for the active areas and polycrystalline silicon for the gate electrodes/local area interconnects, are generally required. Insulators such as Si_3N_4, SiO_2 and doped glasses are used for gate dielectrics, device isolation, metal-to-substrate isolation, metal-to-metal isolation, passivation, etch masks, implantation masks, diffusion barriers, and sidewall spacers. Conductors such as aluminum, copper, cobalt, titanium, tungsten, and titanium nitride are used for local interconnects, contacts, vias, diffusion barriers, global interconnects, and bond pads. In this section, we discuss the various methods to deposit thin films of insulators, conductors, and semiconductors. We

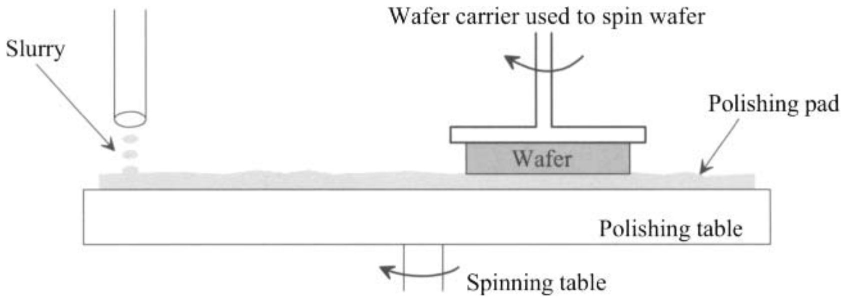


Figure 7.18 Simplified representation of a chemical mechanical polishing process used in the fabrication process.

present two primary categories of thin film deposition: physical vapor deposition and chemical vapor deposition. A third less common category, electrodeposition, for depositing copper for backend interconnects will not be addressed here.

Deposited films are often characterized by several factors. Inherent film quality related to the compositional control, low contamination levels, low defect density, and predictable and stable electrical and mechanical properties are of prime importance. Moreover, film thickness uniformity must be understood and controlled to high levels. To achieve highly uniform CMOS parameters across a wafer, it is common to control the film thickness uniformity to less then +/– 5 nm across the wafer diameter. In addition, film uniformity over topographical features is of critical importance. A measure of this is called step coverage, as depicted in Fig. 7.19. As illustrated, good step coverage results in uniform thickness over all surfaces. In contrast, poor step coverage results in significantly reduced thickness on vertical surfaces relative to surfaces parallel with the surface of the wafer. Related to step coverage is what is referred to as gap fill. Gap fill applies to the deposition of a material into a high aspect ratio opening, such as contacts or gaps between adjacent metal lines. Figure 7.20 illustrates a deposition with good gap fill and a deposition that yields a poor gap fill (also called a keyhole or void).

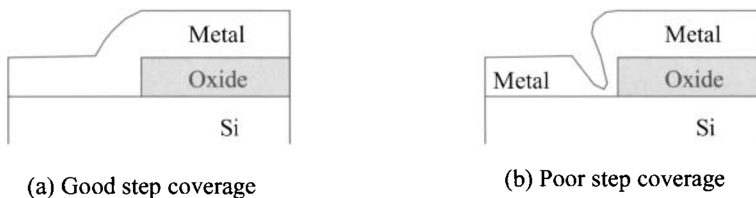


Figure 7.19 Extremes in thin-film deposition coverage over a pre-existing oxide step.

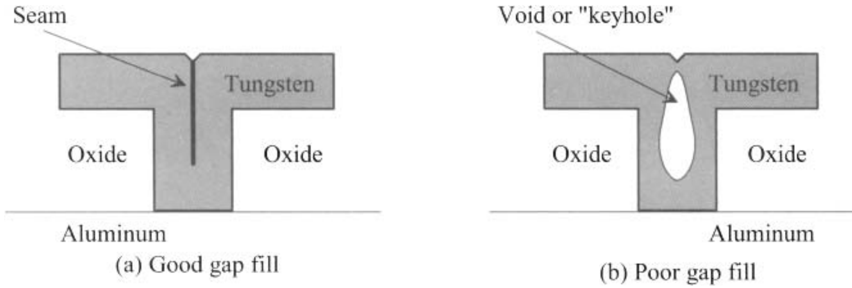


Figure 7.20 Gap-fill profiles (good and bad) of a high aspect ratio opening filled with a deposited film.

Physical Vapor Deposition (PVD)

In physical vapor deposition (PVD), physical processes produce the constituent atoms (or molecules), which pass through a low-pressure gas phase and subsequently condense on the surface of the substrate. The common PVD processes are evaporation and sputter deposition, both of which can be used to deposit a wide range of insulating, conductive, and semiconductive materials. One of the drawbacks of PVD is that the resultant films often have poor step coverage.

Evaporation is one of the oldest methods of depositing thin-films of metals, insulators, and semiconductors. The basic process of evaporation is shown in Fig. 7.21. The material to be deposited is heated past its melting point in a high vacuum chamber where the vapor form of the material coats all exposed surfaces within the mean free path of the evaporant. The heat source can be of one of two types: heating filament or focused electron beam.

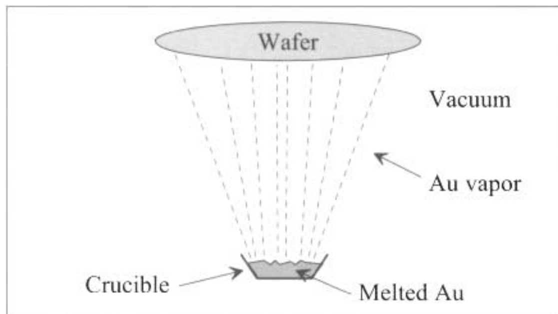


Figure 7.21 Simplified diagram of an evaporation deposition process.

In simple terms, sputter deposition is similar to sputter etching, as discussed in Scc. 7.1.5, with the exception that the wafer serves as the anode, and the cathode is the target material to be deposited. Figure 7.22 outlines a simplified sputter deposition process. An inert gas such as argon is ionized in a low pressure ambient where the positively charged ions are accelerated through the electric field towards the target. The

target is an ultra-high purity disk of material to be deposited. The bombardment of ions with the target sputter (or eject) target atoms (or molecules), where they transit to the surface of the wafer forming a thin-film. Similar to sputter etching, a DC supply can be used for sputtering conductors; however, a capacitively coupled RF supply must be used for depositing non-conductive materials.

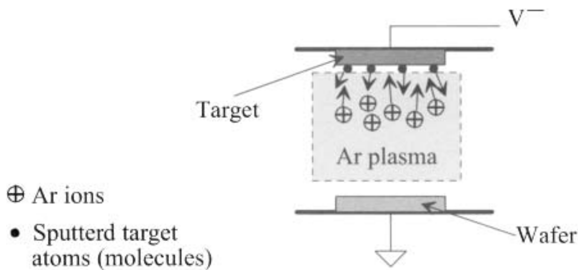


Figure 7.22 Simplified diagram of a sputter deposition process.

Chemical Vapor Deposition (CVD)

In chemical vapor deposition (CVD), reactant gases are introduced into a chamber where chemical reactions between the gases at the surface of the substrate produce a desired film. The common CVD processes are atmospheric pressure (APCVD), low pressure (LPCVD), and plasma enhanced (PECVD). Again, there are a wide variety of insulators, conductors, and semiconductors that can be deposited by CVD. Most importantly, the resultant films tend to have good step coverage compared to PVD processes. APCVD occurs in an apparatus similar to an oxidation tube furnace (see Fig 7.4); however, an appropriate reactive gas is flowed over the wafers. APCVD is performed at relatively low temperatures. As depicted in Fig. 7.23, LPCVD occurs in a reactor in the pressure range of milliTorr to a few Torr. Compared to APCVD, the low pressure process can yield

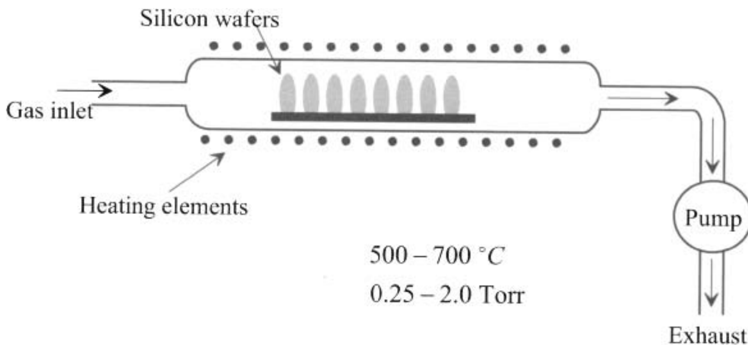


Figure 7.23 Simplified schematic diagram of a LPCVD.

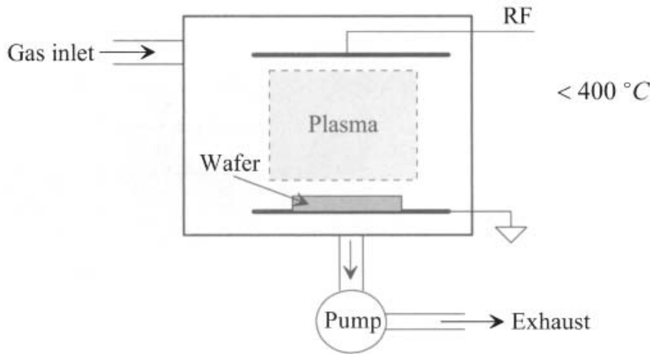


Figure 7.24 Simplified schematic diagram of a PECVD reactor.

highly conformal films, but at the expense of a higher deposition temperature. Fig. 7.24 shows a schematic diagram of a PECVD reactor. In PECVD, a plasma imparts energy for the surface reactions, allowing for lower temperature deposition. By comparison, PECVD has the advantage of being low temperature and highly conformal.

7.2 CMOS Process Integration

Process integration is the task of combining a deliberate sequence of unit processes to fabricate integrated microelectronic circuits (e.g., MOSFETs, resistors, capacitors, etc.) A typical CMOS technology consists of a complex arrangement of unit processes where several hundred steps are required to manufacture integrated circuits on a silicon wafer. Groups of unit processes are combined to form integration modules. For example, the gate module would include a specific sequence of unit processes for yielding a gate electrode on a thin, gate dielectric. The modules could then be combined to yield the overall process flow (or process sequence). The process flow can be divided into frontend-of-the-line (FEOL) and backend-of-the-line (BEOL) processes. A typical process flow consisting of numerous modules is shown in block diagram form in Fig. 7.25.

FEOL

Generally, FEOL refers to all processes preceding salicidation (i.e., silicide formation, see Fig. 4.4). FEOL includes all processes required to fully form isolated CMOS transistors. In Fig. 7.25 we see that the FEOL begins with the selection of the starting material (i.e., type of silicon wafer to be used). Then, the shallow trench isolation (STI) module is implemented to form the regions of dielectric between regions of active area. Next, the wells (or tubs) are formed followed by the gate module, which includes all processes to properly define gate electrodes on a thin oxide. Finally, the FEOL concludes with the source/drain module, which includes the processes required for the formation of the low-doped drain extensions and the source/drain regions themselves.

BEOL

BEOL refers to all processes subsequent to source/drain formation. Hence, BEOL processes are used to “wire” the transistors together using multiple layers of dielectrics

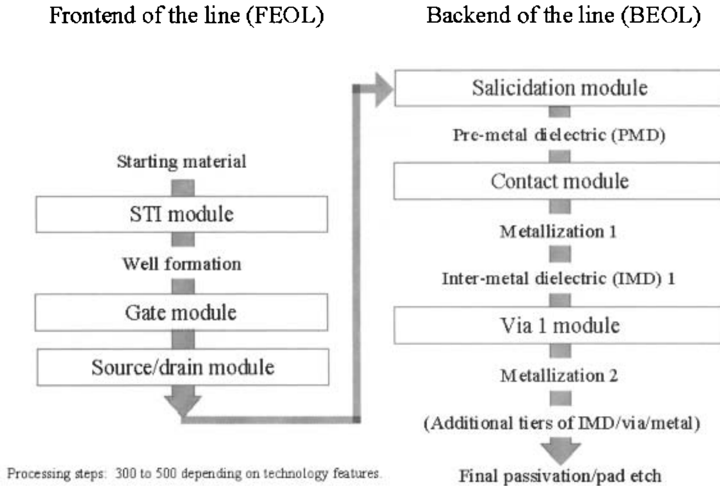


Figure 7.25 A typical CMOS process flow illustrating the difference between FEOL and

and metals. The BEOL begins with the salicidation of the polysilicon and source/drain regions. The remaining BEOL processes proceed in repetitive sets of modules to yield lateral and vertical interconnects isolated from one another with dielectrics. It is important to understand that there is a high degree of inter-relationship between unit processes within each module and between modules themselves. A seemingly "trivial" change in one unit process in a given module can have dramatic effects on processes in other modules. In other words, **there is no trivial process change**.

CMOS Process Description

Even with a single device type, there are numerous variations in process schemes for achieving similar structures. Hence, it would be virtually impossible to outline all schemes. Therefore, we will describe a generic (but representative) deep-submicron CMOS process flow (deep indicating that a deep, or short wavelength, ultraviolet light source is used when patterning the wafers). Our CMOS technology will have the following features:

1. Frontend of the line (FEOL)

   (a) Shallow trench isolation (STI)

   (b) Twin-tubs

   (c) Single-level polysilicon

   (d) Low-doped drain extensions

2. Backend of the line (BEOL)

   (a) Fully planarized dielectrics

   (b) Planarized tungsten contacts and via plugs

   (c) Aluminum metallization

Following each major process step, cross sections will be shown. The cross sections were generated with a technology computer-assisted design (TCAD) package called Tsuprem-4 and Taurus-Visual (2D) released by Technology Modeling Associates, Inc. These tools simulate a defined sequence of unit processes, thus allowing one to model a process flow prior to its actual implementation. In the interest of brevity, there are several omissions and consolidations in the process description. These include:

1. Wafer cleaning performed immediately prior to all thermal processes, metal depositions, and photoresist removals.
2. Individual photolithographic process steps such as dehydration bake, wafer priming, photoresist application, softbake, alignment, exposure, photoresist development, hardbake, inspection, registration measurement, and critical dimension (CD) measurement.
3. Backside film removal following select CVD processes.
4. Metrology to measure particle levels, film thickness, and post-etch CDs.

To implement our CMOS technology, we employ a reticle set as outlined in Table 7.1. The masks are labeled as having either a clear or a dark field. Clear field masks are masks with opaque features totaling less than 50% of the mask area. In contrast, dark field refers to a mask that has opaque features that account for greater than 50% of the total area. Using this mask set, we assume the exclusive use of positive tone photoresist processing. When appropriate, representative mask features that yield isolated, complementary transistors adjacent to one another are shown.

Table 7.1 Masks used in our generic CMOS process.

Layer name	Mask	Aligns to level	Times used	Purpose
1 (active)	Clear	aligns to notch	1	Defines active areas
2 (p-well)	Clear	1	2	Defines NMOS sidewall implants and p-well
3 (n-well)	Dark	1	2	Defines PMOS sidewall implants and n-well
4 (poly1)	Clear	1	1	Defines polysilicon
5 (n-select)	Dark	1	2	Defines nLDD and n+
6 (p-select)	Dark	1	2	Defines pLDD and P+
7 (contact)	Dark	4	1	Defines contact to poly and actives areas
8 (metal1)	Clear	7	1	Defines metal1
9 (via1)	Dark	8	1	Defines via1 (connects M1 to M2)
10 (metal2)	Clear	9	1	Defines metal2
passivation	Dark	Top-level metal	1	Defines bond pad opening in passivation

7.2.1 Frontend-of-the-Line Integration

As previously stated, FEOL encompasses all processing required to fabricate the fully-formed, isolated CMOS transistors. In this subsection we discuss the modules and unit processes required for a representative CMOS process flow.

Starting Material

The choice of substrate is strongly influenced by the application and characteristics of the CMOS ICs to be fabricated. Bulk silicon is the least costly but may not be the optimal choice in high performance or harsh environment CMOS applications. Epitaxial (Epi) wafers are heavily doped bulk wafers with a thin, moderately to lightly doped epitaxial silicon layer grown on the surface. The primary advantage of Epi wafers is for immunity to latch-up. Silicon-on-insulator wafers increase performance and eliminate latch-up. However, SOI CMOS is more costly to implement than bulk or epi technologies. The three general types of silicon substrates are shown in Fig. 7.26. In current manufacturing, wafer diameters typically range from 100 to 300 mm. The wafer thickness correspondingly increases with diameter to allow for greater rigidity. The actual CMOS is constructed in the top one micron or less of the wafer, whereas the remaining hundreds of microns are used solely for mechanical support during device fabrication.

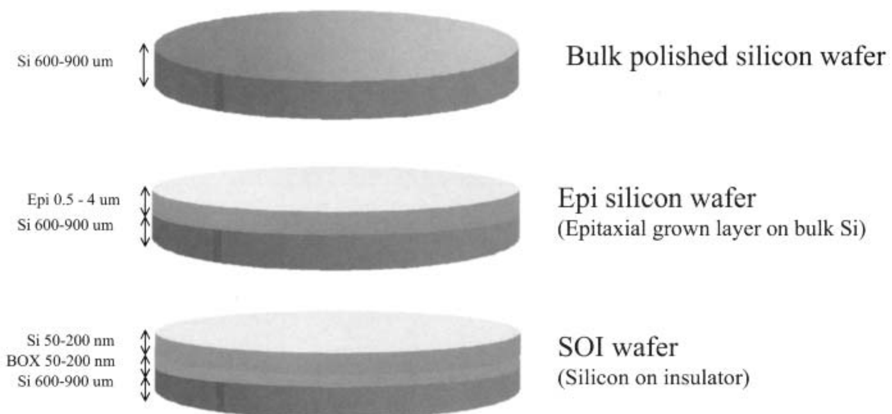


Figure 7.26 The three general types of silicon wafers used for CMOS fabrication.

We use bulk silicon wafers for our CMOS technology in this section. It should be noted that with relatively minor process and integration adjustments our technology could be applicable to epi or SOI CMOS processes. The first of many simulated cross sections are shown in Fig. 7.27. Here only the top two microns of silicon are shown. At the beginning of the fabrication process, the wafer characteristics such as resistivity, sheet resistance, crystallographic orientation, and bow and warp are measured and/or recorded. Furthermore, the wafers are scribed, usually with a laser, with a number which identifies the wafer's lot and number.

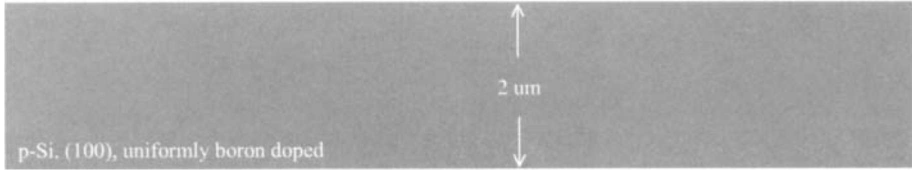


Figure 7.27 Simulated cross-sectional view (the top 2 μm) of the bulk wafer in Fig. 7.26.

Shallow Trench Isolation Module

Devices (e.g., PMOS and NMOS) must be electrically isolated from one another. This isolation is of primary importance for suppressing leakage current between both like and dissimilar devices.

One of the simplest methods of isolation is to fabricate the CMOS such that a reversed-bias pn-junction is formed between the transistors. Oppositely doped regions (e.g., n-well adjacent to a p-well) can be electrically isolated by tying the n-region to the most positive potential in the circuit and the p-region to the most negative. As long as the reverse-bias is maintained and the breakdown voltage is not exceeded for all operating conditions, a small diode reverse saturation current accounts for the leakage current. This junction leakage current is directly proportional to the junction area; hence, for the large p- and n-regions in modern devices, junction isolation alone is not adequate.

The second general method of isolation is related to the formation of thick dielectric regions, called field regions, between transistors. The region without the thick dielectric is where the transistors reside and is known as the active area. The relatively thick oxide that forms between the active areas is called field oxide (FOX). Interconnections of polysilicon are formed over the field regions to provide localized electrical continuity between transistors. This arrangement inherently leads to the formation of parasitic field effect transistors. Effectively, the FOX increases the parasitic transistor's threshold voltage such that the device always remains in the off state. Further, this threshold voltage can be increased by increasing the surface doping concentration, called channel stops, under the FOX. There are two general approaches to forming field oxide regions: LOCOS and STI.

LOCal Oxidation of Silicon (LOCOS) has been used extensively for half-micron or larger minimum linewidth CMOS technologies. In LOCOS, a diffusion barrier of silicon nitride blocks the thermal oxidation of specific regions on the surface of a wafer. Both oxygen and water diffuse slowly through silicon nitride. Hence, nitride can be deposited and patterned to define active and field oxide areas. The primary limitation to LOCOS is bird's beak encroachment, where the lateral diffusion of the oxidant forms an oxide feature that in cross-section resembles a bird's beak. The bird's beak encroaches into the active area, thereby reducing the achievable circuit packing density. Moreover, LOCOS requires a long, high-temperature process, which can result in significant diffusion of previously introduced dopants.

Shallow trench isolation (STI) is the dominant isolation technology for sub-half micron CMOS technologies. As the name implies, a shallow trench is etched into the surface of the wafer and then filled with a dielectric serving as the FOX. A typical STI process sequence follows. From a processing perspective, STI is complex; however, it

can be implemented with minimal active-area encroachment. Moreover, it has a relatively low thermal budget.

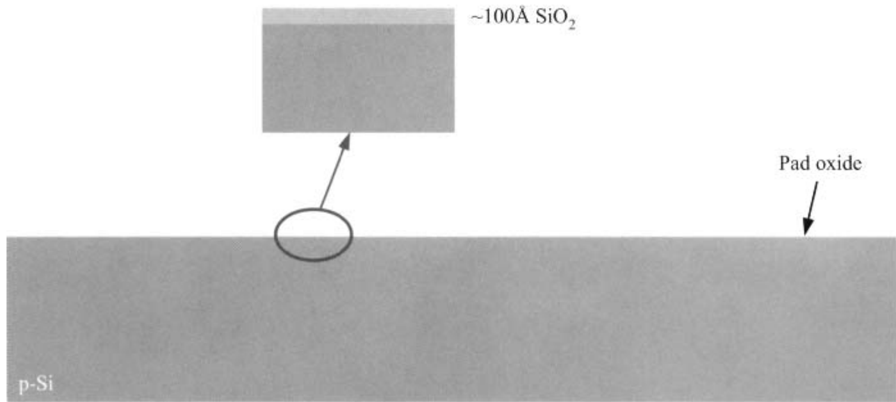


Figure 7.28 STI film stack. The oxide is thermally grown at approximately 900 °C with dry O_2.

The CMOS technology outlined in this chapter uses STI to achieve device isolation. The STI module begins with the thermal oxidation of the wafer surface, as shown in Fig. 7.28. The resultant oxide serves as a film-stress buffer, called a pad oxide, between the silicon and the subsequently deposited Si_3N_4 layer. (In addition, it is also used following the post-CMP nitride strip as an ion implant sacrificial oxide.) Next, silicon nitride is deposited on the oxidized wafer by LPCVD, as shown in Fig. 7.29. Later, this nitride serves as both an implant mask and a CMP stop-layer. As shown in Fig. 7.30, the photolithography (mask layer 1) produces the appropriate patterns in photoresist for defining the active areas. Then, with end-point-detected RIE, the photoresist pattern is transferred into the underlying film stack of nitride and oxide. In Fig. 7.30 notice that the PMOS and NMOS devices will be fabricated on the left side and right side, respectively, under the photoresist. The region cleared of photoresist corresponds to the isolation regions. The 0.4 μm deep silicon trenches are formed by timed RIE with the photoresist

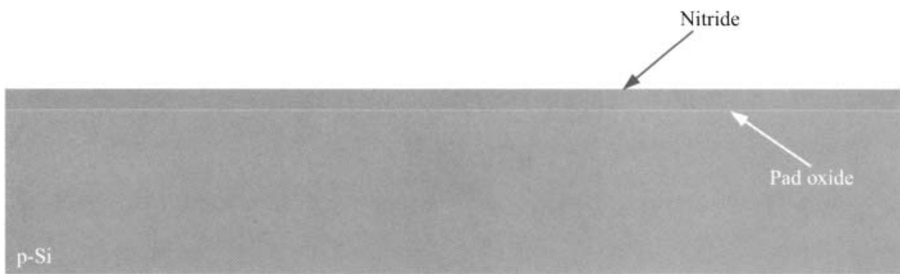


Figure 7.29 STI film stack. Silicon nitride deposited by LPCVD at approximately 800 °C.

Layout (plan) view (mask layer 1)

Cross-section

STI region

PMOS active area

NMOS active area

Photoresist

Photoresist

p-Si

Figure 7.30 STI definition, photolithography and nitride/pad oxide etch with fluorocarbon- based RIE.

softmask present, as shown in Fig 7.31. Although the etching can proceed without the resist, the sidewall profile can be tailored to a specific slope with the presence of the polymer during the etch process. At the cost of reduced packing density, the sloped sidewalls aid in the reduction of leakage current from the parasitic corner transistors.

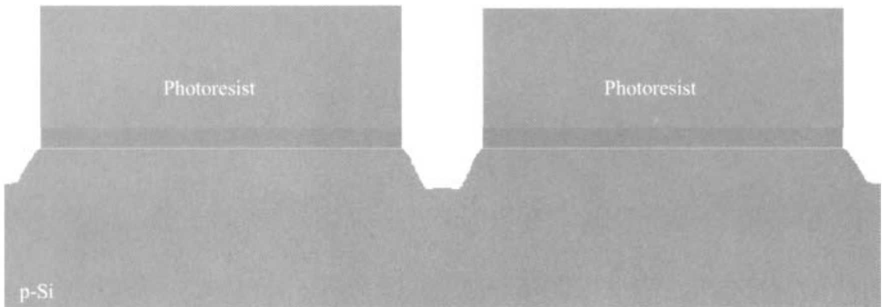


Figure 7.31 Timed silicon trench reactive ion etch.

Following the silicon etch, O_2 plasma and wet processing strip the photoresist and etch by-products from the surface of the wafer. At this point, the general structural form of the STI is finished.

The next series of processes are used to improve effectiveness of the STI to suppress leakage currents. As shown in the expanded view of the trench seen in Fig. 7.32, a thin, sacrificial oxide is thermally grown on the exposed silicon. It should be noted that the nitride provides a barrier to the diffusion of oxygen, hence the oxidation occurs only in the exposed silicon regions. This oxide serves as a sacrificial oxide for subsequent ion implantation and aids in softening the corner of the trench. In general, implant sacrificial oxides are used to (1) suppress ion channeling in the crystal lattice, (2) minimize lattice damage from the ion bombardment, and (3) protect the silicon surface from contamination. Photolithography (mask layer 2) patterns resist to protect the PMOS sides of the trench during the p-wall implant. A shallow BF_2 implant is performed to dope what will eventually become the p-*well* trench sidewalls (called the p-*walls*), as seen in Fig. 7.33. The p-wall implant increases the threshold voltage of the parasitic corner transistor and minimizes leakage under the trench. The BF_2 implanted resist is stripped using O_2 plasma and wet processing, as shown in Fig. 7.34. Again, photolithography (mask layer 3), Fig. 7.35, is used to produce the complementary pattern for the n-wall implant. For the same reasons, this shallow phosphorous implant is introduced into what will become the n-well trench sidewalls. The phosphorous-implanted resist is stripped using O_2 plasma and wet processing, yielding the structure depicted in Fig. 7.36.

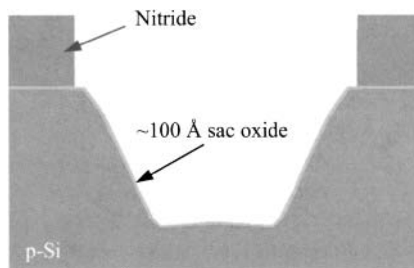


Figure 7.32 Cross-section showing post STI resist strip followed by the dry thermal oxidation (at 900 °C) of a sacrificial (sac) oxide in the trench.

At this point, the sacrificial oxide has been degraded by the implantations and is likewise stripped using a buffered hydrofluoric acid solution. A thin, high-quality thermal oxide is regrown in the trenches to form what is called a *trench liner*. In general, the liner oxide improves the interface quality between the silicon and the subsequent trench fill, thus suppressing the interface leakage current. Specifically, the formation of the trench liner oxide (1) "cleans" the surface prior to trench fill, (2) anneals sidewall implant damage, and (3) passivates interface states to minimize parasitic leakage paths.

Once the liner oxide is grown, CVD is used to overfill the trenches with a dielectric, as shown in Fig. 7.37. The trench fill provides the field isolation required to increase the threshold voltage of the parasitic field transistors. Further, it blocks

Figure 7.33 P-wall sidewall formation via photolithography and BF_2 implantation.

subsequent ion implants. Although not shown, it is common to use a "block-out" pattern to improve the uniformity of the STI CMP. In Fig. 7.38a, CMP is performed to remove the CVD overfill. The nitride is used as a polish-stop layer. Next, a brief buffered oxide etch removes oxide that may have formed on top of the nitride. Then, the nitride is

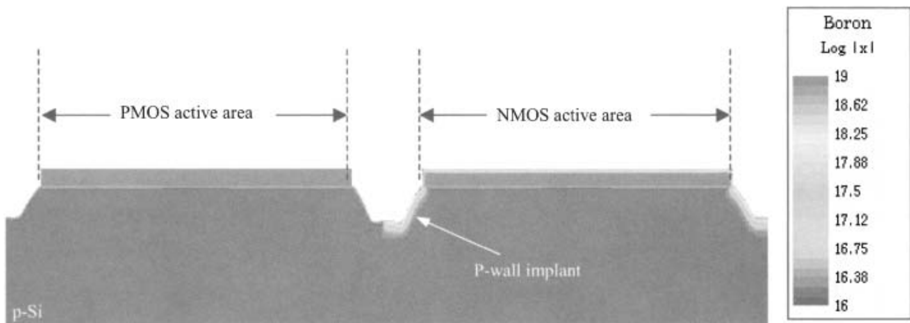


Figure 7.34 Post p-wall photoresist strip using O_2 plasma and wet processing.

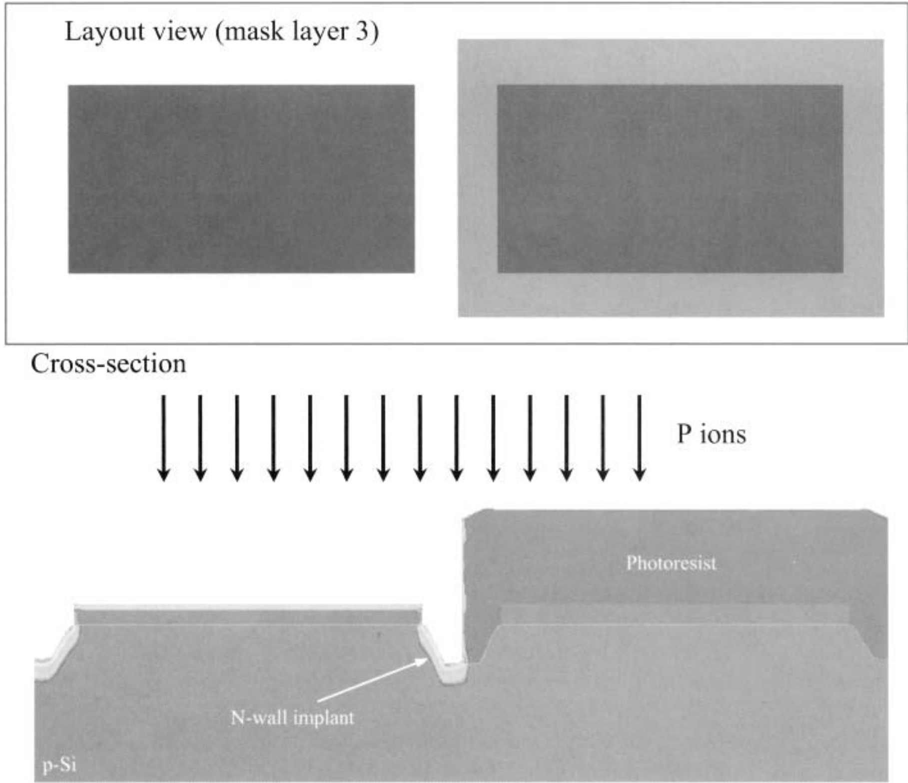


Figure 7.35 N-wall sidewall formation via photolithography and *P* implantation.

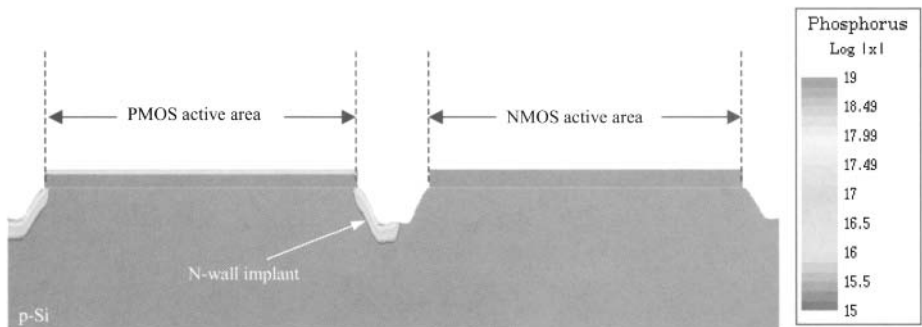


Figure 7.36 Post n-wall photoresist strip using O_2 plasma and wet processing.

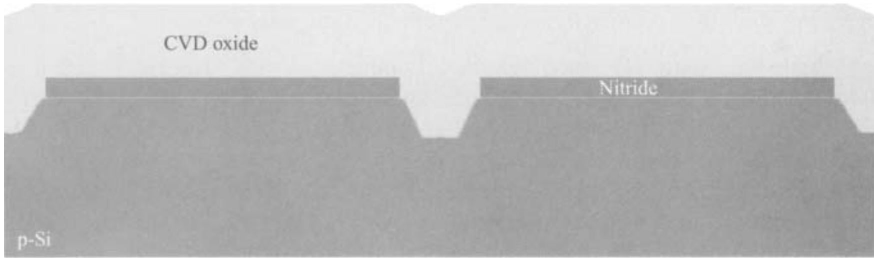


Figure 7.37 High quality, 100 Å thick liner oxide is thermally grown at 900 °C. High density plasma (HDP) CVD trench fill at room temperature. Notice that the trenches are overfilled.

removed from the active areas by using a wet or dry etch process, as illustrated in Fig. 7.38b. Notice that the pad oxide remains after this step. At this point, the STI is fully formed.

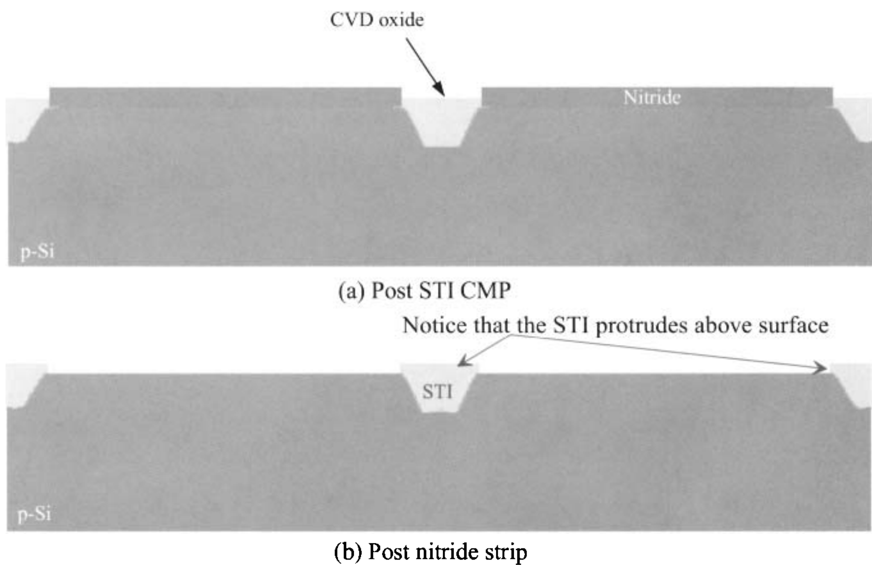


Figure 7.38 STI CMP (a) where the nitride acts like a polish stop and (b) wet nitride etch in hot phosphoric acid and/or dry nitride etch in NF_3/Ar/NO. The remaining oxide will be used as a sacrificial oxide for subsequent implants.

Twin-Tub Module

As explained in Sec. 2.5, CMOS can be implemented in four general forms: n-well, p-well, twin-well (called twin-tub), and triple-well. The CMOS technology discussed in this chapter uses a twin-well approach. The p-well and n-well provide the appropriate

dopants for the NMOS and PMOS, respectively. Modern wells are implanted with retrograde profiles to maximize transistor performance and reliability.

Following the STI module, the twin-tub module begins with p-well photolithography (mask layer 2, second use) to generate a resist pattern that covers the PMOS active regions, but exposes the NMOS active areas, as depicted in Fig. 7.39. A relatively high energy boron implant is performed into the NMOS active areas. Here the implant is blocked from the PMOS active area. The pad oxide that remained from the STI module now serves as the implant sacrificial oxide for the well implants. It should be pointed out that the p-well may be formed by a composition of several implants at different doses and energies to achieve the desired retrograde profile. Following the p-well implant, the resist is removed using O_2 plasma and wet processing, resulting in the structure shown in Fig. 7.40.

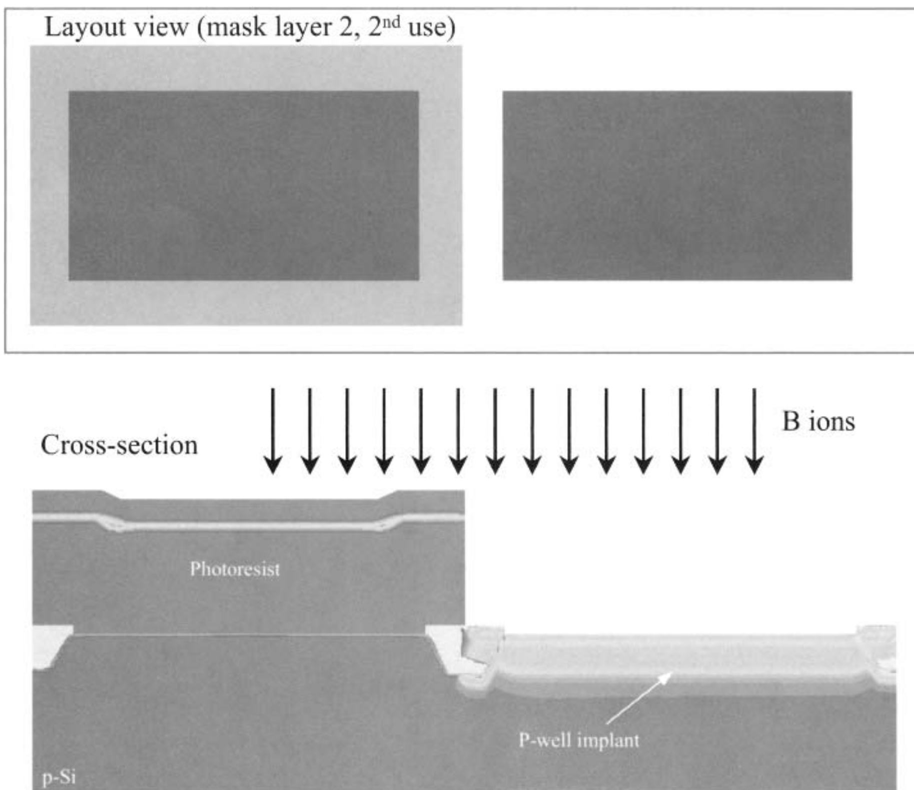


Figure 7.39 P-well formation via photolithography and *B* implantation.

Next, a complementary resist pattern is formed using the n-well mask and photolithography (mask layer 3, second use), as seen in Fig. 7.41. Again, a relatively high energy implant, this time using phosphorus, is performed to generate the n-well. Similar to the p-well, a multitude of implants may be used to achieve the desired retrograde profile. Following the n-well implant, the resist is stripped using O_2 plasma and wet

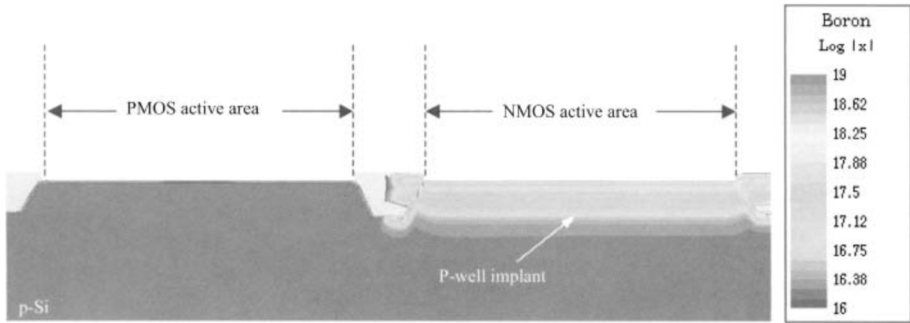


Figure 7.40 Post p-well photoresist strip using O_2 plasma and wet processing.

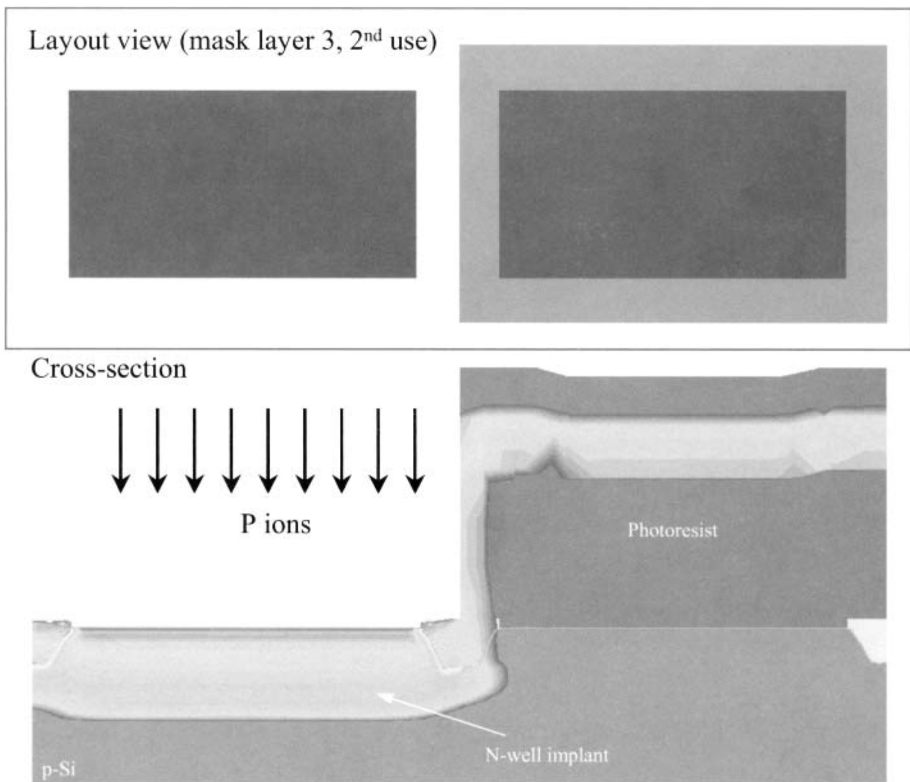


Figure 7.41 N-well formation via photolithography and *P* implantation.

processing, yielding the structure in Fig. 7.42. At this point, both the isolation and the wells are fully formed. Figure 7.43 shows the cross section of the substrate following the twin-tub module. Notice that the net doping profile is given, thus highlighting both well and wall implants simultaneously. It should be emphasized that the PMOS are fabricated in the n-wells; the NMOS, in the p-wells.

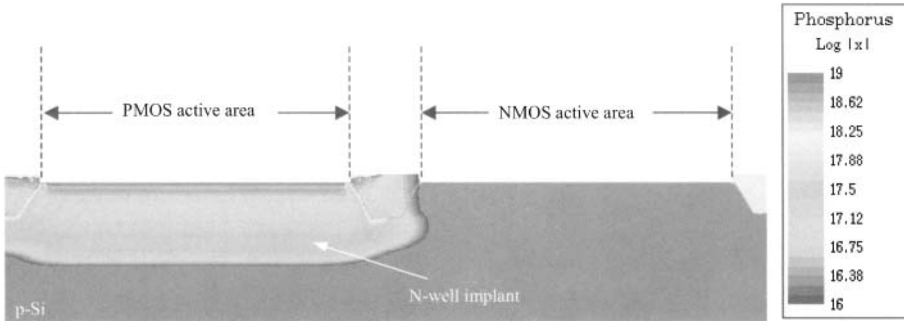


Figure 7.42 Post n-well photoresist strip using O_2 plasma and wet processing.

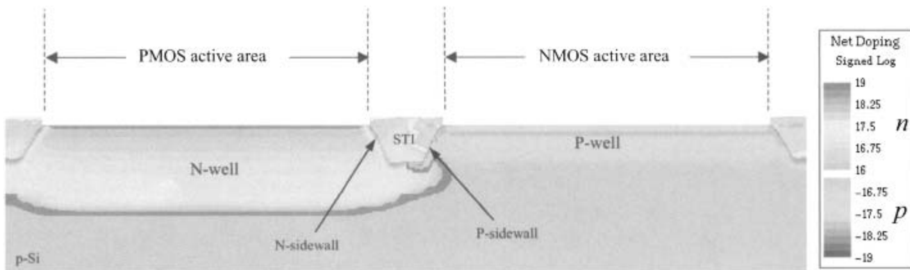


Figure 7.43 Net doping profile of both the n-well and the p-well.

Gate Module

As depicted in Fig. 7.44, we begin the gate module with the buffered oxide etching of the remaining thin oxide in the active areas from the twin-tub module. Then, a sacrificial oxide is thermally grown. This oxide serves as a threshold adjust implant oxide and a pre-gate oxidation "clean-up." Next, a blanket (unpatterned), low energy BF_2 threshold adjust implant is performed, as shown in Fig. 7.45. This implant allows for the "tuning" of both the PMOS and NMOS threshold voltages. The single boron implant is common

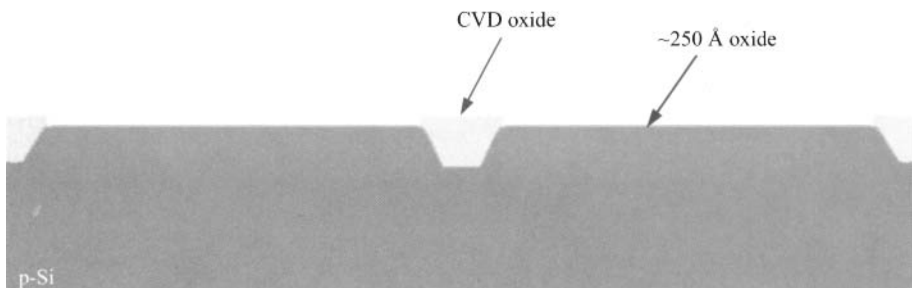


Figure 7.44 Wet etch the remaining trench stack oxide using buffered HF. Sacrificial oxide formation using dry thermal oxidation at approximately 900 °C.

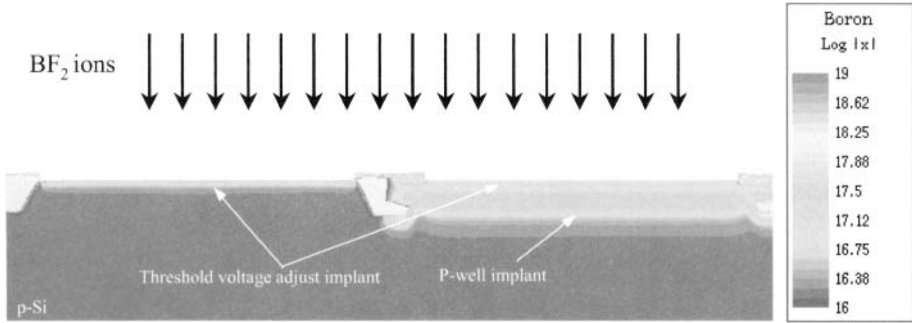


Figure 7.45 Blanket low-energy BF_2 implant for NMOS and PMOS threshold voltage adjust.

for single workfunction gates. However, for dual workfunction gates (common in technologies with minimum gate lengths of 250 nm or less), separate p-type and n-type implants are required for the threshold adjustment in the NMOS and PMOS, respectively.

To form the gate stack (i.e., the gate dielectric and polysilicon gate electrode) the next set of processes are required. Of course, the gate stack provides for the capacitive coupling to the channel. Using wet processing, the sacrificial oxide is stripped from the active areas. As shown in Fig. 7.46, a high quality, thin oxide is thermally grown, which serves as the gate dielectric. In modern CMOS, it is common to use nitrided gate oxide by performing the oxidation in O_2 and NO or N_2O. It can be argued that the gate oxidation is the most critical step in the entire process sequence, as the characteristic of the resultant film greatly determines the behavior of the CMOS transistors. The gate oxidation is immediately followed by an LPCVD polysilicon deposition, as depicted in Fig. 7.47. For single workfunction gates, the polysilicon can be doped with phosphorous during poly deposition or subsequently implanted. For dual workfunction gates, the NMOS and PMOS can be doped during the n+ and p+ source/drain implants, respectively.

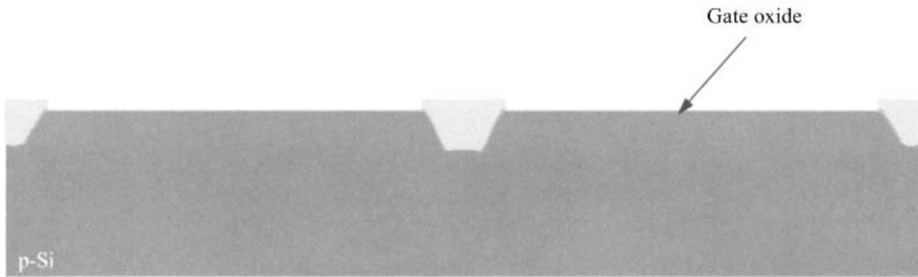


Figure 7.46 Removal of sacrificial oxide using buffered HF followed by gate dielectric formation using dry oxidation in an ambient of O_2, NO and/or N_2O.

Once the gate stack is formed, the transistor gates and local interconnects are patterned using photolithography (mask layer 4) to generate the appropriate patterns in photoresist, as seen in Fig. 7.48. The gate patterning must be precisely controlled as it

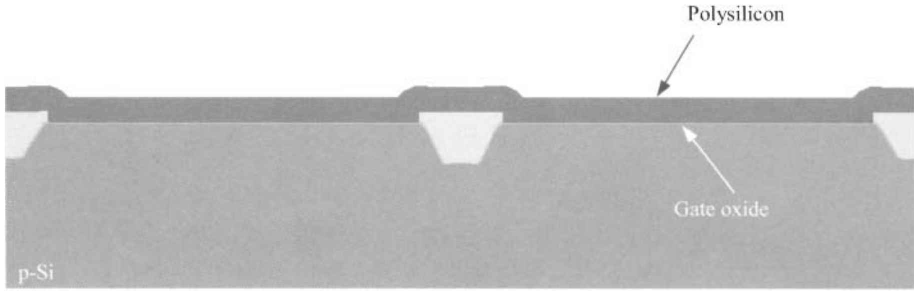


Figure 7.47 Polysilicon deposition via LPCVD at approximately 550 °C. Note that the polysilicon deposition must occur immediately following gate oxidation.

determines the gate lengths. Deviations in the resultant physical gate lengths can cause severe performance issues with the CMOS. Seen in Fig. 7.48 are the ideal gate profiles following the RIE of polysilicon and subsequent resist strip.

The gate module concludes with the poly reoxidation as shown in Fig. 7.49. Here the thermal oxidation of the polysilicon and active silicon is performed to (1) grow a buffer pad oxide for the subsequent nitride spacer deposition and (2) electrically activate the implanted dopants in the polysilicon. Notice that since the polysilicon oxidizes at a faster rate than the crystalline silicon, the resultant oxide thickness is greater on the polysilicon than the active silicon.

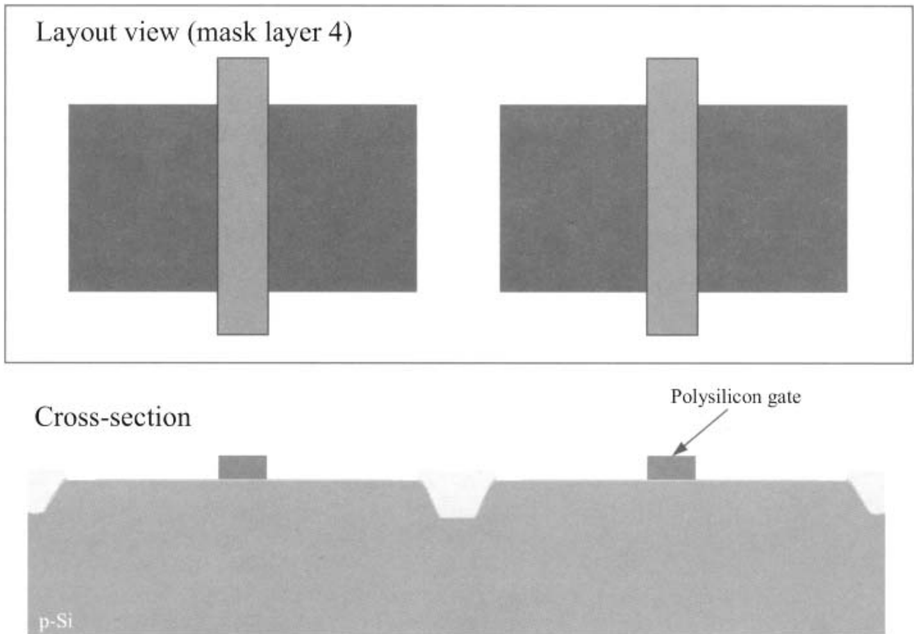


Figure 7.48 Gate electrode and local interconnect photolithography and polysilicon reactive ion etching.

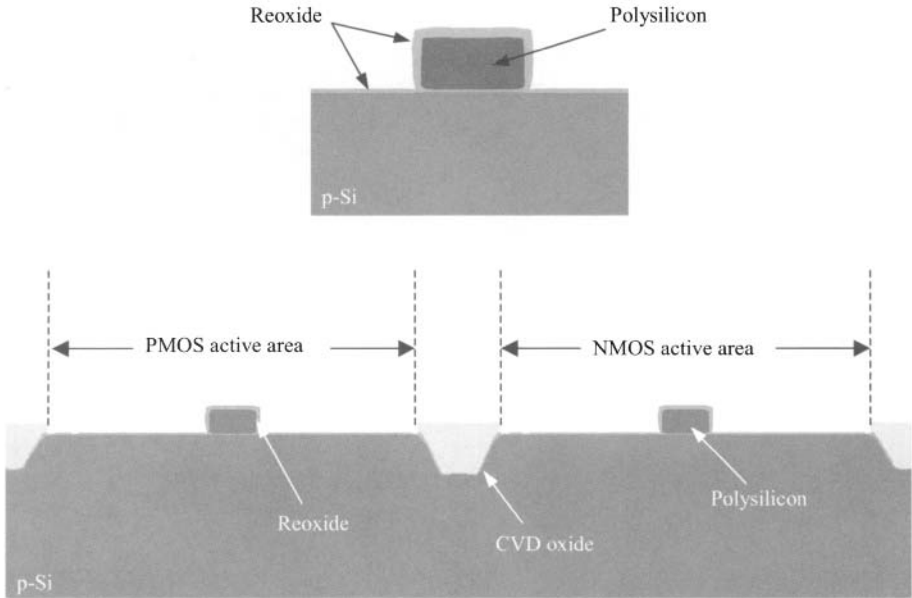


Figure 7.49 Polysilcion reoxidation using dry O_2 at approximately 900 °C. Notice that the resulting oxide is thicker on the polysilicon than on the active silicon.

Source/Drain Module

At the onset of the source/drain module, the source/drain extensions are formed by a series of processes. Photolithography (mask layer 5) is used to pattern resist such that the NMOS devices are exposed, as shown in Fig. 7.50. Then, a low energy phosphorus implant is performed to form the n-channel, low doped drain (nLDD) extensions. Notice that the presence of the polysilicon gate inherently leads to the self-alignment of the extensions with respect to the gate. The nLDD suppresses hot carrier injection into the gate and reduces short-channel effects in the NMOS. At this point in the process sequence, a deep boron pocket implant is often used to prevent source/drain punchthrough in the NMOS. The photoresist is stripped. The resultant structure is shown in Fig. 7.51.

In a similar manner, the p-channel source/drain extensions are formed. Photolithography (mask layer 6) is used to protect NMOS devices with resist, as shown in Fig. 7.52. Boron is implanted at low energy to form the p-channel low doped drain (pLDD) extensions. Again, the polysilicon serves to self-align the implant with respect to the gate electrodes. As was the case with the NMOS, it is common to use a deep phosphorus pocket implant to suppress PMOS punchthrough. Once the photoresist is stripped, the cross-section shown in Fig. 7.53 is achieved.

To complete the source/drain extensions, the gate sidewall spacers must be formed prior to the actual source/drain implants. As shown in Fig. 7.54, conformal silicon nitride is deposited using LPCVD. A CVD oxide may be used in lieu of the nitride. Following the subsequent nitride etch, this nitride will form the LDD sidewall spacers.

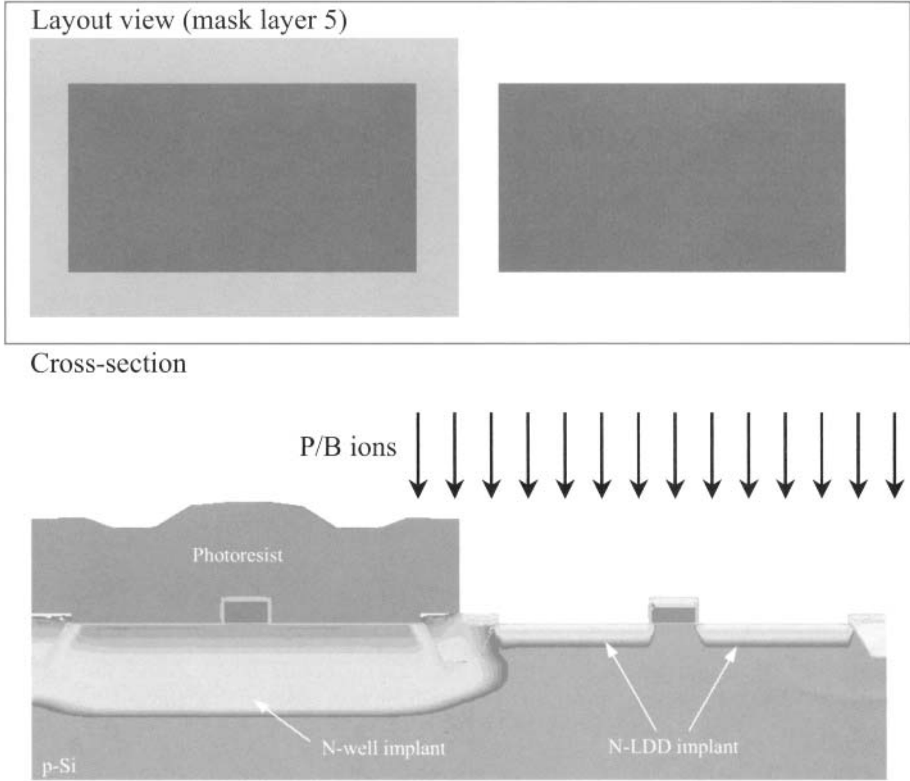


Figure 7.50 N-LDD/n-pocket formation using low energy implantation of *P* and *B*, respectively.

The spacers function as (1) a mask to the source/drain implants and (2) a barrier to the subsequent salicide formation. The actual spacers are formed by an unpatterned anisotropic RIE of nitride, as illustrated in Fig. 7.55. The spacer etch is end-pointed on the underlying oxide. Notice that since the nitride is thickest along the polysilicon

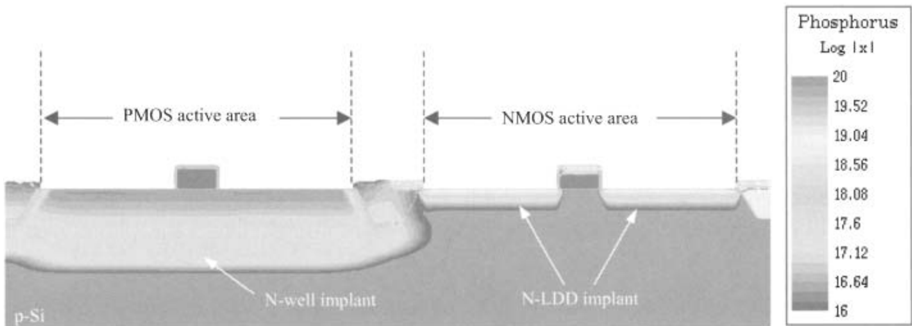


Figure 7.51 Post-n-LDD resist strip using O_2 plasma and wet processing.

Layout view (mask layer 6)

Cross-section

B/P ions

Photoresist

P-LDD implant

P-well implant

p-Si

Figure 7.52 P-LDD/p-pocket formation using low energy implantation of B and P, respectively.

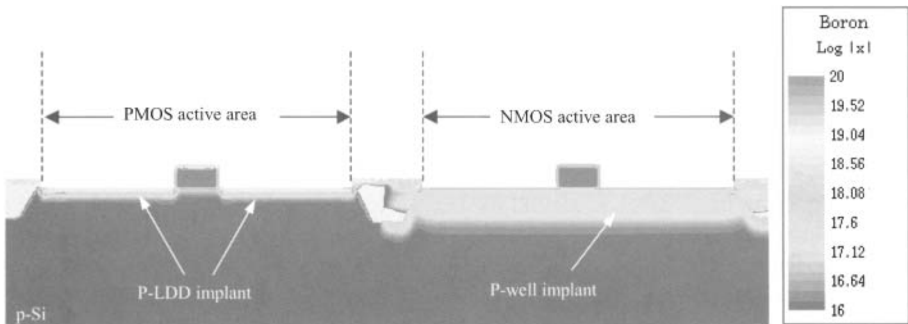


Figure 7.53 Post-p-LDD resist strip using O_2 plasma and wet processing.

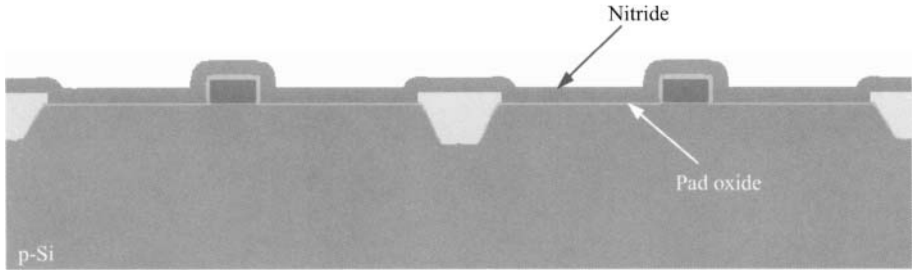


Figure 7.54 Sidewall spacer nitride deposition using LPCVD at approximately 800 °C.

sidewall, a well-formed insulating region remains on both sides of the polysilicon. This structure is called a spacer.

During the source/drain implants, the combination of the polysilicon and spacers block the implantation, thus allowing for self-alignment to not only the gate but also to the LDD extensions. With this stated, the NMOS source/drains are formed, as shown in Fig. 7.56. Photolithography (mask layer 5, second use) protects the PMOS with resist while exposing the NMOS. A relatively low energy, high dose arsenic implant is performed to form the n+ regions. The resist is stripped yielding the structure in Fig. 7.57. In addition to the source/drain formation, this implant forms the necessary n+ ohmic contacts.

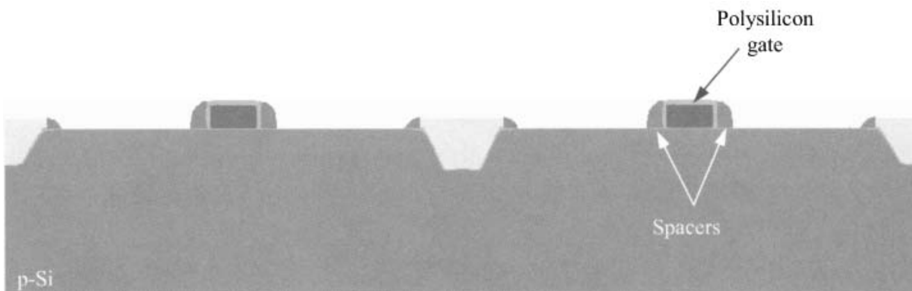
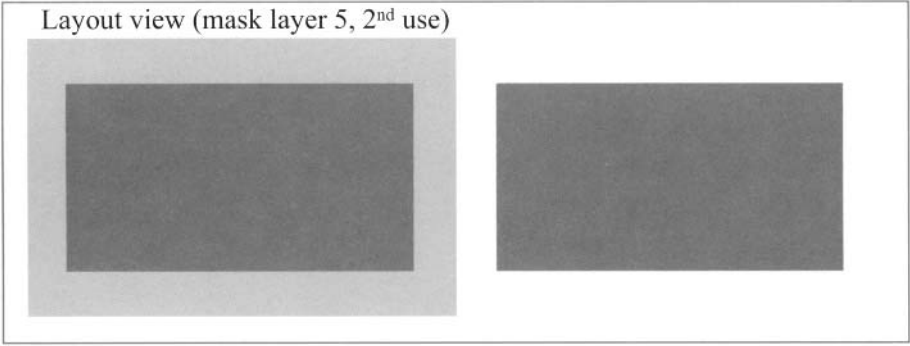


Figure 7.55 Dry, anisotropic, end-pointed reactive ion etch of spacer nitride yielding gate sidewall spacers.

The PMOS source/drains and p+ ohmic contacts are formed in a similar manner. Photolithography (mask layer 6, second use) and a low energy, high dose BF_2 implant is used, as illustrated in Fig. 7.58. The resist is stripped resulting in the structure shown in Fig. 7.59. The source/drain module concludes with a high temperature anneal that electrically activates the implants and re-crystallizes the damaged silicon. In modern CMOS, the primary reason that polysilicon is chosen as the gate electrode material as opposed to metal is that the poly can withstand the high temperatures required to activate the source/drain implants.

At this point in our CMOS process sequence we have fully formed the CMOS transistors and their isolation. This marks the completion of the FEOL. Figure 7.60 provides a summary of the main features generated in the FEOL.



Cross-section

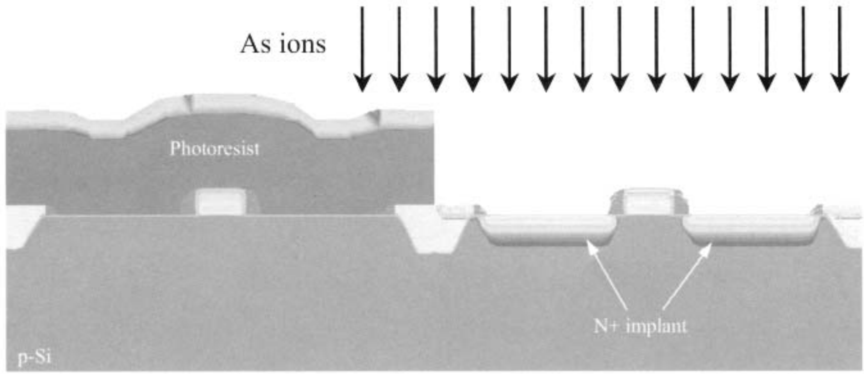


Figure 7.56 N+ source/drain formation using a low energy, high dose implantation of *As*.

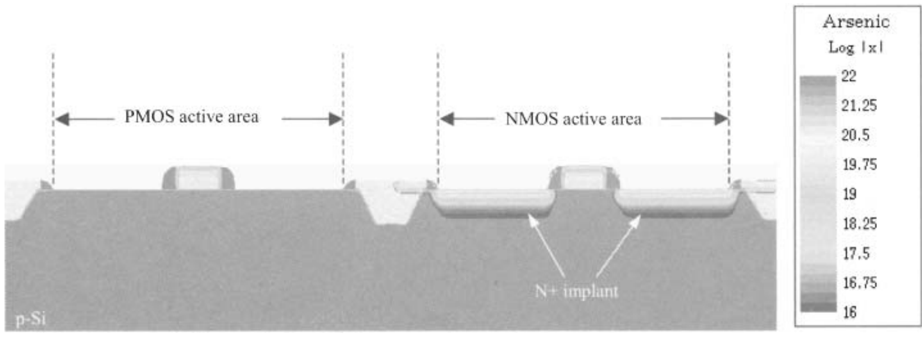


Figure 7.57 Post-n+ resist strip using O_2 plasma and wet processing.

Layout view (mask layer 6, 2nd use)

Cross-section

BF_2 ions

Photoresist

P+ implant

p-Si

Figure 7.58 P+ source/drain formation using a low energy, high dose implantation of BF_2.

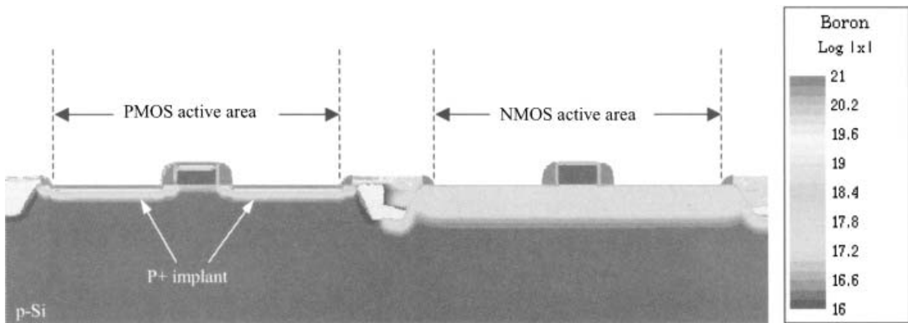


Figure 7.59 Post p+ resist strip using O_2 plasma and wet processing.

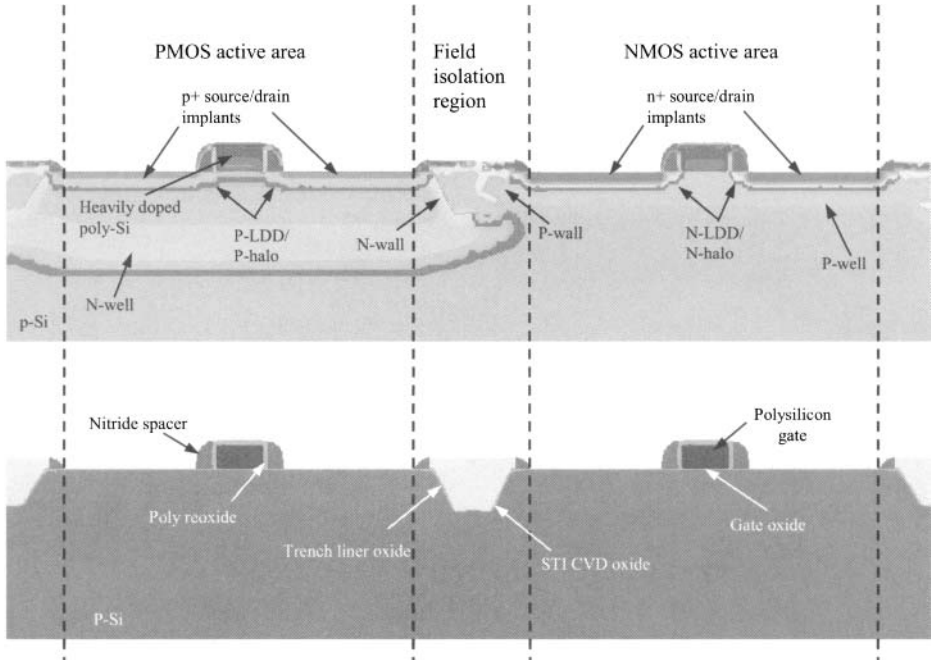


Figure 7.60 Summary of the FEOL features.

7.2.2 Backend-of-the-Line Integration

In this section we continue our CMOS process flow through the BEOL. The BEOL encompasses all processes required to "wire" the transistors to one another and to the bond pads. CMOS requires several metal layers to achieve the interconnects necessary for modern designs. We discuss the processing through the first two metal layers to give an appreciation of the overall BEOL integration.

Self-Aligned Silicide (Salicide) Module

At the boundary of the FEOL and BEOL is the self-aligned silicide, called salicide, formation. Silicide lowers the sheet resistance of the polysilicon and active silicon regions. The self-aligned silicide (salicide) relies on the fact that metal silicide will generally not form over dielectric materials such as silicon nitride. Therefore, a metal such as titanium or cobalt can be deposited over the entire surface of the wafer, then annealed to selectively form silicide over exposed polysilicon and silicon. Because of the presence of the trench fill and sidewall spacers, the silicide become self-aligned without the need for photopatterning.

The salicide module begins with the removal of the thin oxide, present from the FEOL, using buffered *HF*, as shown in Fig. 7.61. Next, a refractory metal (e.g., titanium or cobalt) is deposited by sputtering, as depicted in Fig. 7.62. To minimize contamination, a thin layer of *TiN* is deposited as a cap. A relatively low temperature, nitrogen ambient, rapid thermal annealing (RTA) is used to react titanium (or cobalt) with the silicon, forming $TiSi_2$ (C49 phase) or ($CoSi_x$). The resultant silicide (i.e., C49) is a high resistivity

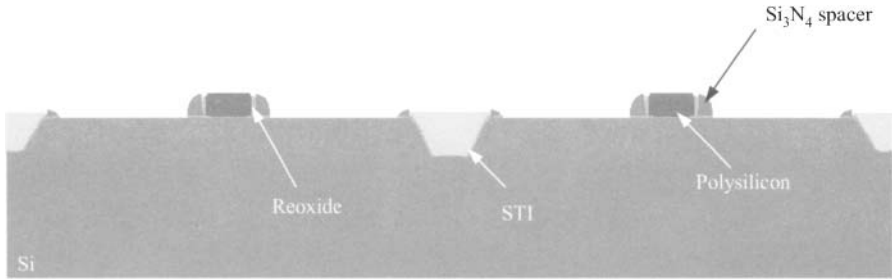


Figure 7.61 Removal of the exposed reoxide (present from the FEOL) using buffered-*HF*.

phase. Also, notice that the underlying nitride and oxide serves to block the formation of the silicide from the sidewalls and trenches, respectively.

To prevent spacer overgrowth of the silicide, the low resistivity phase is achieved by processing with two separate anneals. The first, as described above, forms the high resistivity phase without the risk of silicide formation on the nitride. The second, as described below, occurs following the wet chemical etching of the unreacted titanium (or cobalt) using a higher temperature, which causes a phase change (C49 to C54 for $TiSi_2$) with a much lower resistivity. If one high temperature anneal was originally performed to achieve the low resistivity phase, then significant overgrowth could occur, leading to leakage current from the source and drain to the gate of the transistors.

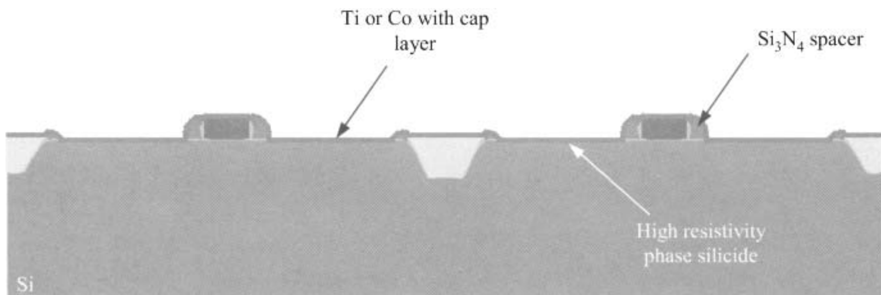


Figure 7.62 Titanium or cobalt deposited by PVD followed by the first salicide rapid thermal anneal.

To continue the salicide module, following the first anneal, the unreacted titanium (or cobalt) is wet chemically etched from the wafer. The second RTA, in argon ambient at a slightly higher temperature, achieves the low resistivity phase, as shown in Fig. 7.63.

Pre-Metal Dielectric

The pre-metal dielectric (PMD) provides electrical isolation between metal1 and polysilicon/silicon. To aid the subsequent contact etch process, a thin layer of silicon nitride is deposited as an etch stop. This is followed by a high density plasma deposition of the PMD oxide, as shown in Fig. 7.64. The resultant surface of the PMD must be

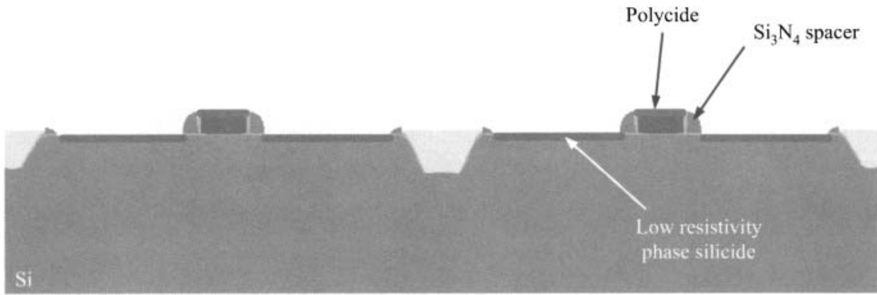


Figure 7.63 Wet chemical etch of the unreacted titanium or cobalt followed by the second salicide rapid thermal anneal.

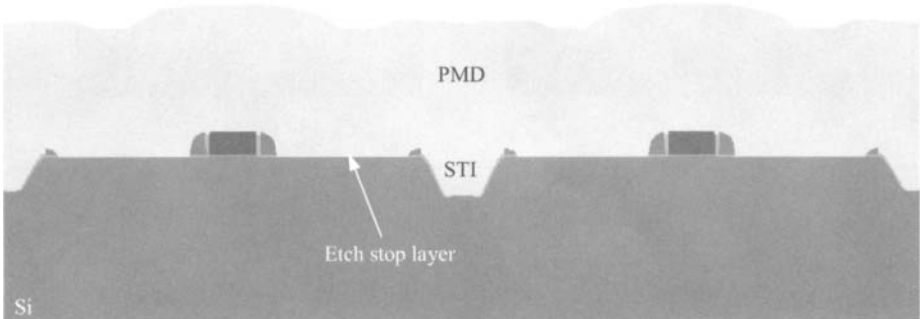


Figure 7.64 Pre-metal dielectric (PMD) deposition using high density plasma

planarized to allow for improved depth-of-focus for the subsequent high resolution photopatterning of metal1. With CMP, the PMD is planarized, producing the cross-section shown in Fig. 7.65.

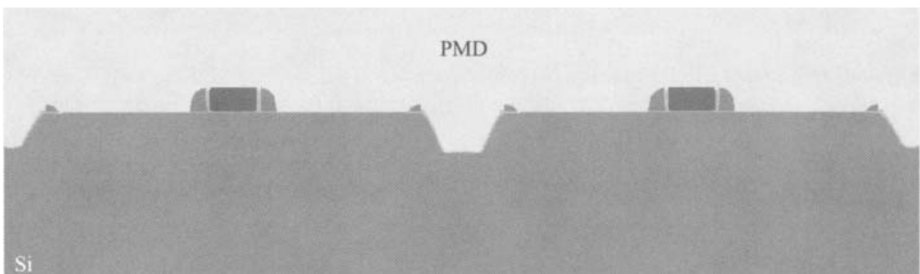


Figure 7.65 Planarizing of the PMD using CMP.

Contact Module

The contacts provide the electrical coupling between metal1 and polysilicon/silicon. The first BEOL photolithography (mask layer 7) step patterns contact openings in the resist. The PMD and nitride is then dry etched using the nitride as an etch-stop layer. The resist is stripped from the wafer, resulting in the structure shown in Fig. 7.66. Contact openings to the source and drains are shown; however, contacts to polysilicon (not shown) over field oxide are simultaneously formed.

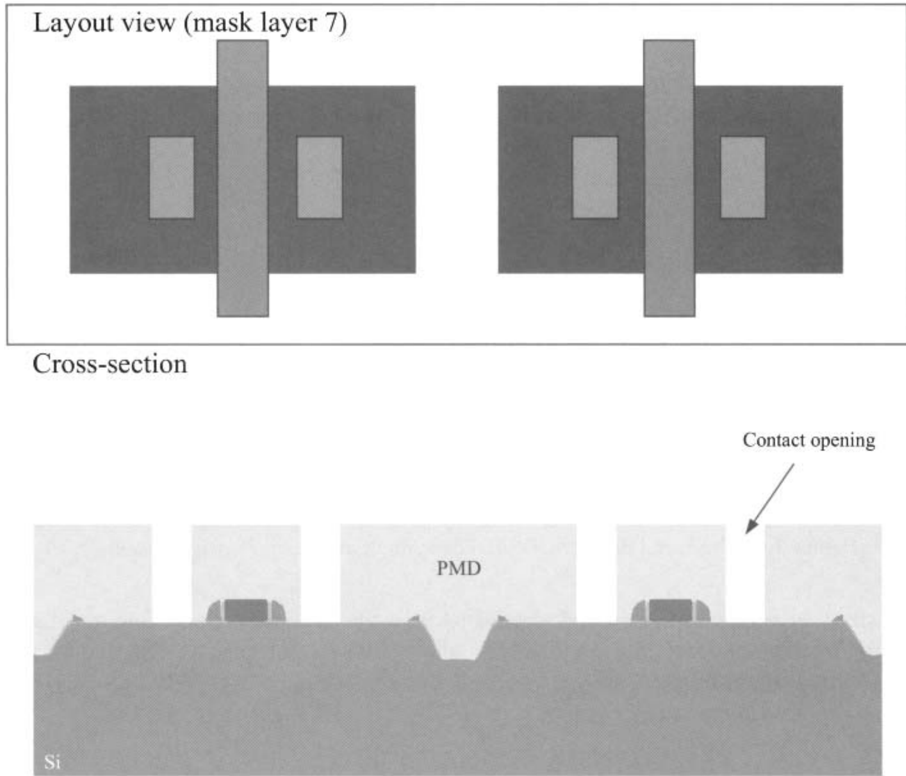


Figure 7.66 Contact definition using photolithography and RIE of the PMD and stop layer. Notice that the contacts to the poly are formed at the same time but not shown.

Next, a thin layer of titanium is deposited by ionized metal plasma (IMP) sputtering, preceded by an in-situ argon sputter etch to clean the bottom of the high aspect ratio contact openings. The titanium is the first component of the contact liner and functions to chemically reduce oxides at the bottom of the contacts. The second component of the liner is a thin CVD *TiN*. The primary purpose of this layer is to act as a diffusion barrier to the fluorine (which readily etches silicon) used in the subsequent tungsten deposition. The contact openings are filled (actually overfilled) with tungsten using a WF_6 CVD process, as shown in Fig. 7.67 As depicted in Fig. 7.68, the overfilled tungsten is polished back to the top of the planarized PMD using CMP. At this point, the surface of the wafer is ultra-smooth and essentially free of all topography. To aid the

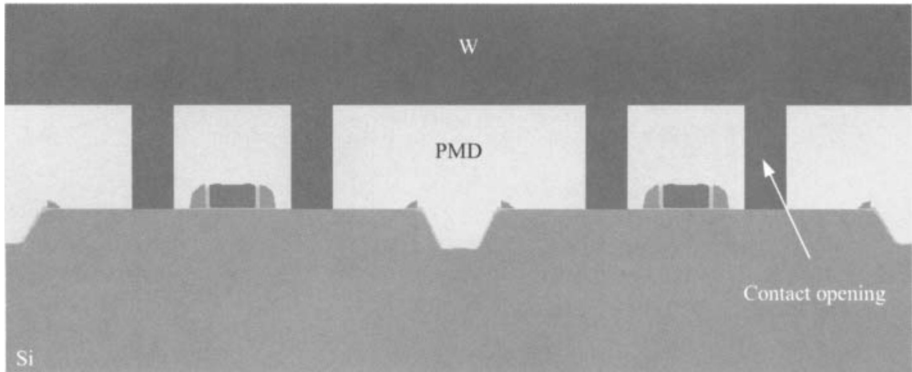


Figure 7.67 *Ti/TiN* liner deposition using IMP and CVD, respectively. *W* contact fill deposition using WF_6 CVD.

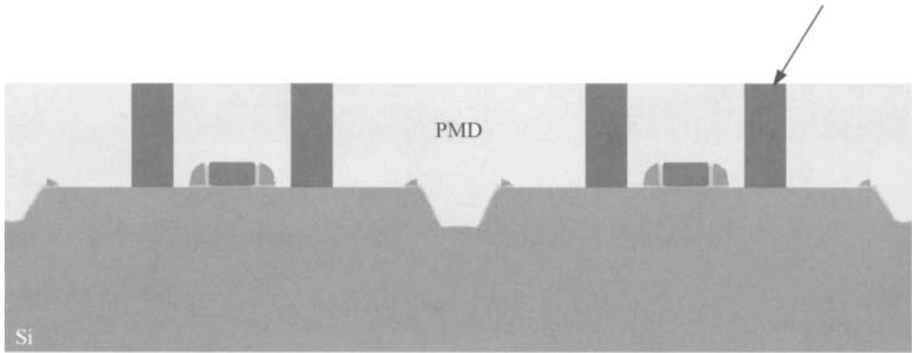


Figure 7.68 *W* CMP to form defined contacts.

photolithographic alignment of the metal layer to the contacts, a tungsten recess etch is performed to provide adequate alignment reference. The recessed contacts are shown in Fig. 7.69.

Metallization 1

To allow for electrical signal transmission from contact-to-contact and from contact-to-via1, defined metallization must be formed. Following the recess etch, sputtering is used to deposit a film stack consisting of *Ti/TiN/Al/TiN*, as seen in Fig. 7.70. The Ti provides adhesion of the TiN and reduction in electromigration problems. The bottom TiN serves primarily as a diffusion barrier to $TiAl_3$ formation. The topmost TiN acts as an anti- reflective coating for the metal photolithography as well as an etch stop for the subsequent via formation.

Using photolithography (mask layer 8), the metal 1 pattern in resist is generated, as depicted in Fig. 7.71. A dry metal etch transfers the pattern into the metal. To prevent metal corrosion, the resist is plasma stripped in an $O_2/N_2/H_2O$ ambient, producing the cross-section shown in Fig. 7.71.

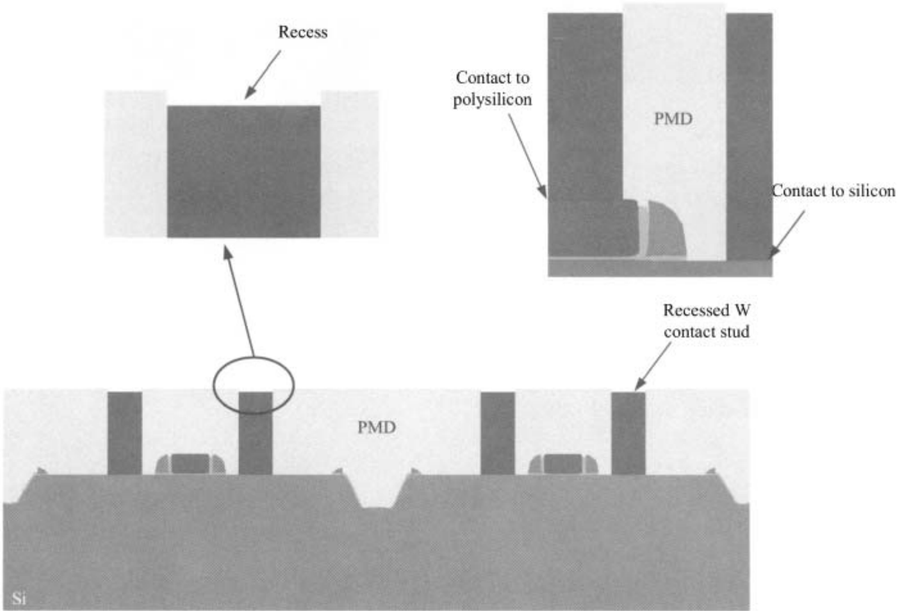


Figure 7.69 *W* recess etch using "buff" polish or dry *W* etch.

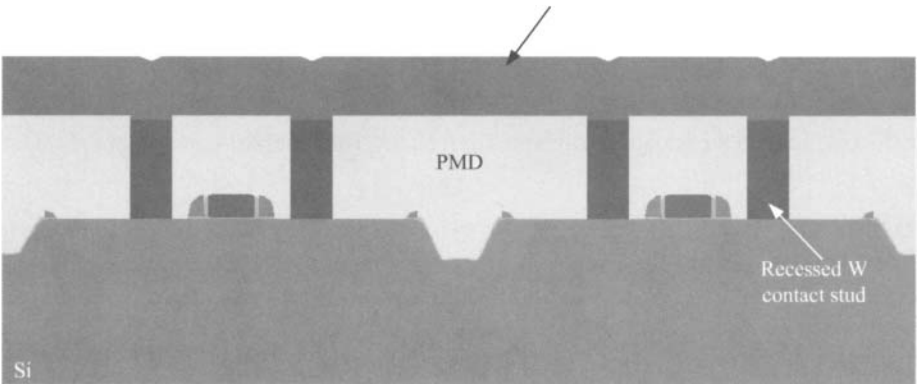


Figure 7.70 Metal1 stack deposition using PVD.

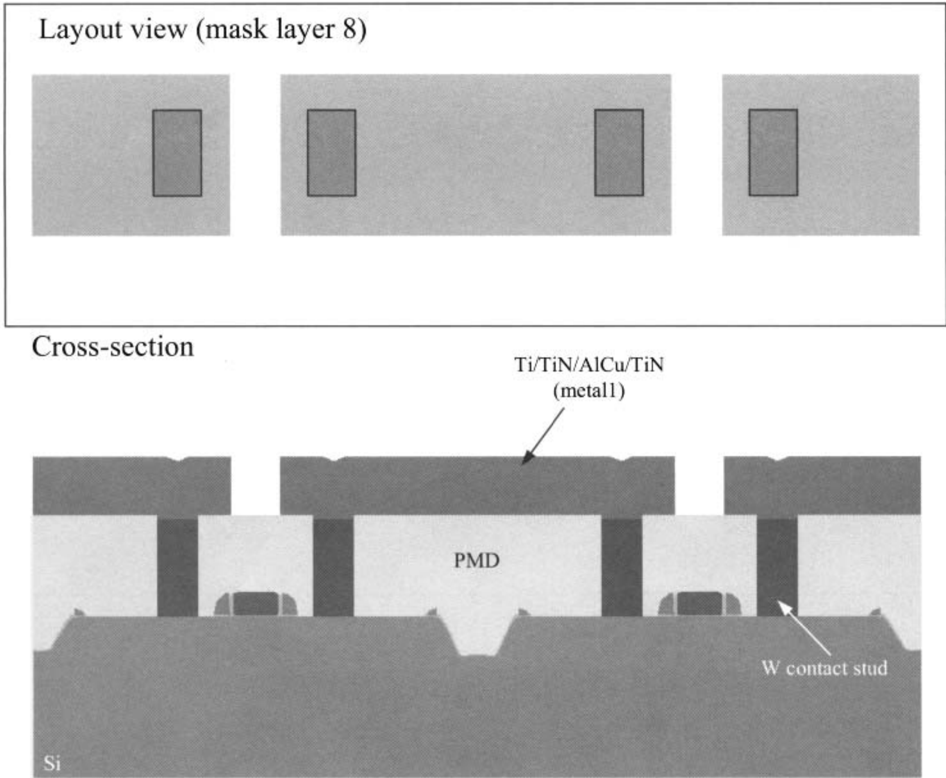


Figure 7.71 Metal1 definition using photolithography and dry metal etch.

Intra-Metal Dielectric 1 Deposition

The intra-metal dielectric 1 (IMD1) provides the electrical isolation between metal1 and metal2. It is common to deposit this film using high density plasma CVD, as shown in Fig. 7.72. As can be seen, the conformal deposition results in a surface topography that must be planarized by CMP. The depth-of-focus is improved for the subsequent photolithographic steps as it is for the PMD planarization.

Via 1 Module

Electrical coupling between metal1 and metal2 is achieved by the via1 module. The planarized IMD is photolithographically defined (mask layer 9), and an RIE opens the vias. The resist is stripped using O_2 plasma and wet processing, producing the cross-section shown in Fig. 7.73. Next, similar to the contact fill, an argon sputter etch is performed followed by the deposition of thin layers of IMP titanium and CVD TiN. Using a WF_6 CVD process, the vias are deposited (overfilled), as seen in Fig. 7.74. The excess tungsten is removed by CMP utilizing the IMD1 as a "polish stop," as depicted in Fig. 7.75. Again, as to provide observable alignment features, a tungsten recess etch is often required (also shown in Fig. 7.75).

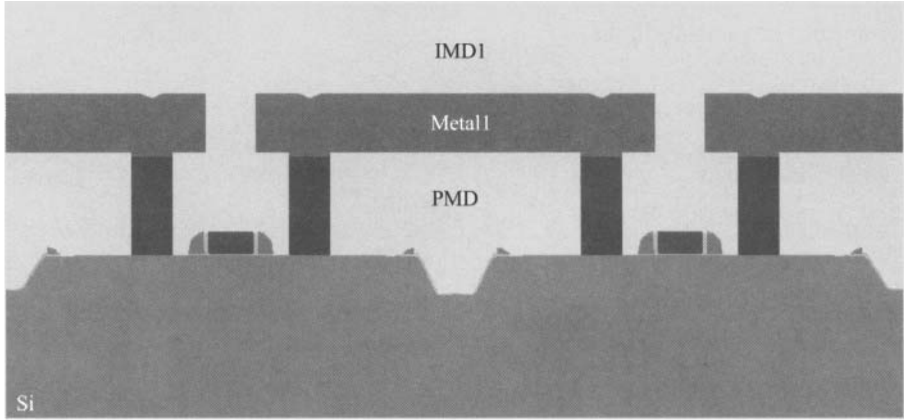


Figure 7.72 Intra-metal dielectric 1 (IMD1) deposition using HDP CVD. This is followed by IMD1 planarization using CMP.

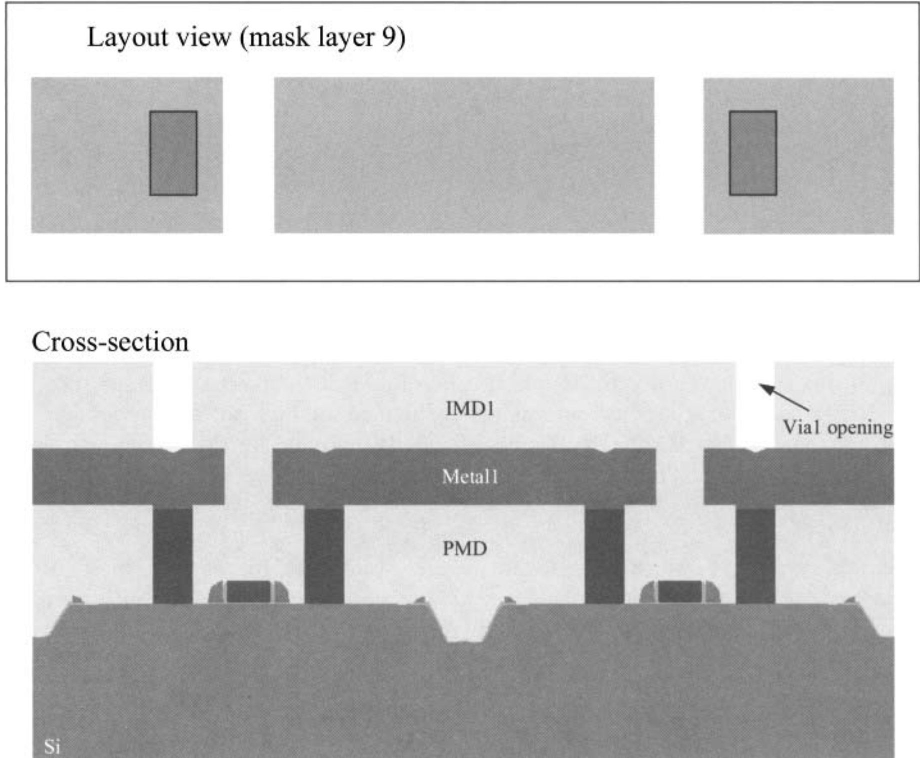


Figure 7.73 Via1 definition using photolithography and dry IMD1 etch.

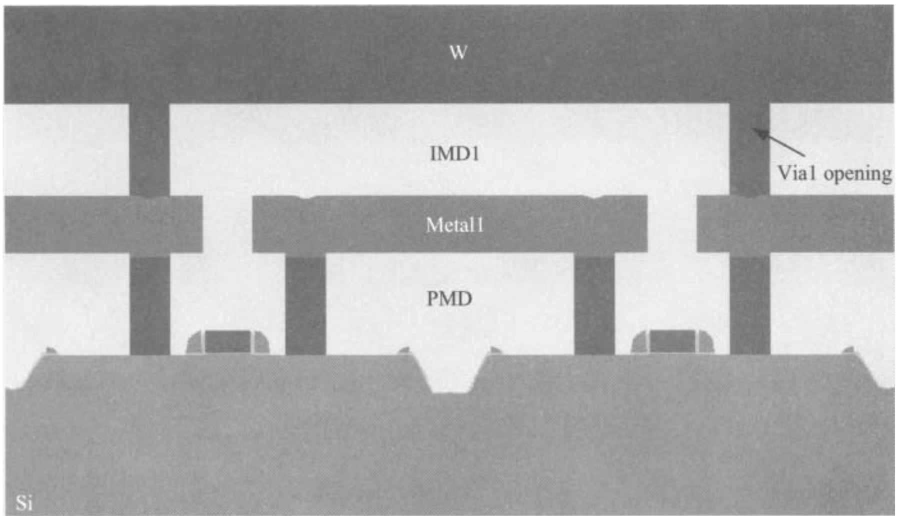


Figure 7.74 *Ti/TiN* liner deposition using IMP and CVD, respectively. *W* via1 fill deposition using WF_6 CVD.

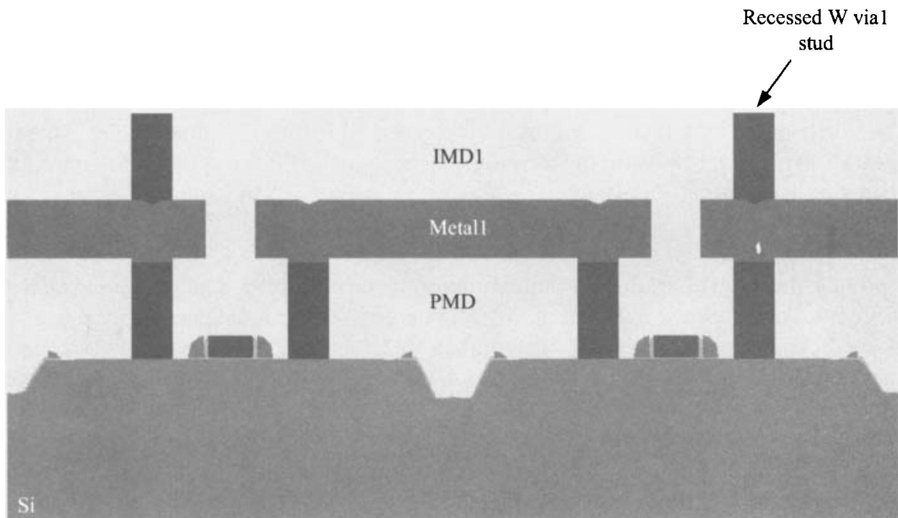


Figure 7.75 *W* CMP to form defined vias. This is followed by *W* recess etch using "buff" polish or dry *W* etch.

Metallization 2

In a similar manner as the metal1 process, the metal 2 stack is deposited (Fig. 7.76) and photolithographically (mask layer 10) defined, as shown in Fig. 7.77. Note that we are not discussing metal implementation using copper. Copper wiring is often implemented with dual-Damascene techniques, see Ch. 4, where both vias and metal layers are simultaneously formed in a series of process steps.

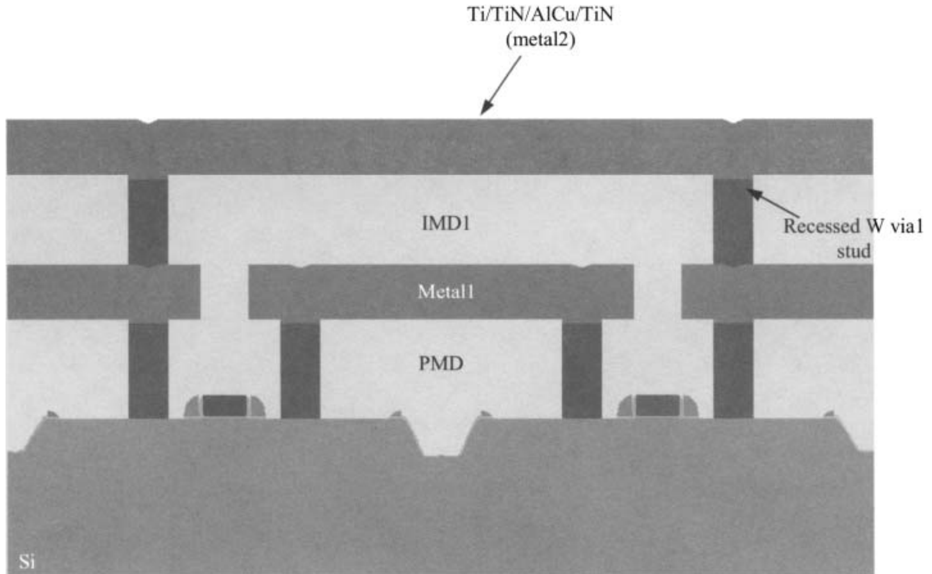


Figure 7.76 Metal2 stack deposition using PVD.

Additional Metal/Dielectric Layers

At this point, additional tiers of dielectric/metal layers can be formed by replicating the aforementioned processes. In modern CMOS there may be more than eight metal layers. It should be noted that as dielectric/metal layers are added, the cumulative film stresses can cause significant bow/warp in the wafers. Hence, great effort is expended to minimize the stresses in the BEOL films.

Final Passivation

To protect the CMOS from mechanical abrasion during probe and packaging and to provide a barrier to contaminants (e.g., H_2O, ionic salts), a final passivation layer must be deposited. The passivation type is determined in large part by the type of package in which the CMOS IC will be placed. Common passivation layers are (1) doped glass and (2) silicon nitride on deposited oxide. Figure 7.78 shows the cross-section of our CMOS process flow following the deposition of the passivation. Finally, the bond pads are opened using photolithography (mask layer n) and by dry etching the passivation. Following photoresist strip, the final CMOS cross-section is shown in Fig. 7.79.

7.3 Backend Processes

Following completion of the final passivation, the wafers are removed from the cleanroom in preparation for a series of backend (i.e., post-fab) processes. These processes include wafer probe, die separation, packaging, and final test/burn-in.

Wafer Probe

Generally, dedicated die with parametric structures and devices are stepped into various positions of the wafer. Alternatively, parametric structures are placed in the dividing

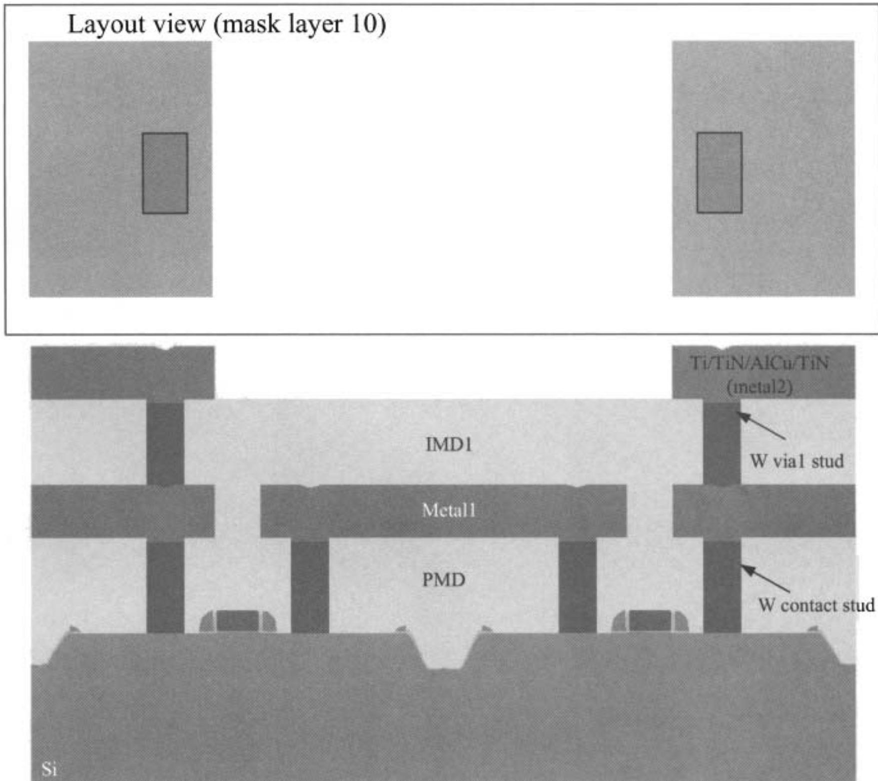


Figure 7.77 Metal2 definition using photolithography and dry metal etch.

regions, called *streets* or *scribe lines*, between die. Electrical characterization of these parametric structures and devices is often performed at select points in the fabrication process flow (such as following metal 1 patterning). Parameters such as contact resistance, sheet resistance, transistor threshold voltage, saturated drain current, off-current, sub-threshold slope, etc., are measured. If problems are observed from the in-line parametric tests, troubleshooting can begin sooner than if the testing were only performed following final passivation. Furthermore, wafers that do not meet the parametric standards can be removed (or "killed") from the fabrication sequence (and thus money is saved).

After the completed wafers are removed from the fab, wafer-level probing is performed to check final device parameters and to check CMOS integrated circuit functionality and performance. Wafer probe is accomplished by using sophisticated testers that can probe individual die (or sets of die) and apply test vectors to determine circuit behavior. Inevitablity, there are a percentage of die that will not pass all the vector sets and thus are considered failed die. The ratio of good die-to-total die represents the wafer *yield* given by

$$Y = \frac{\text{\# of die passing all tests}}{\text{total \# of die}} \tag{7.13}$$

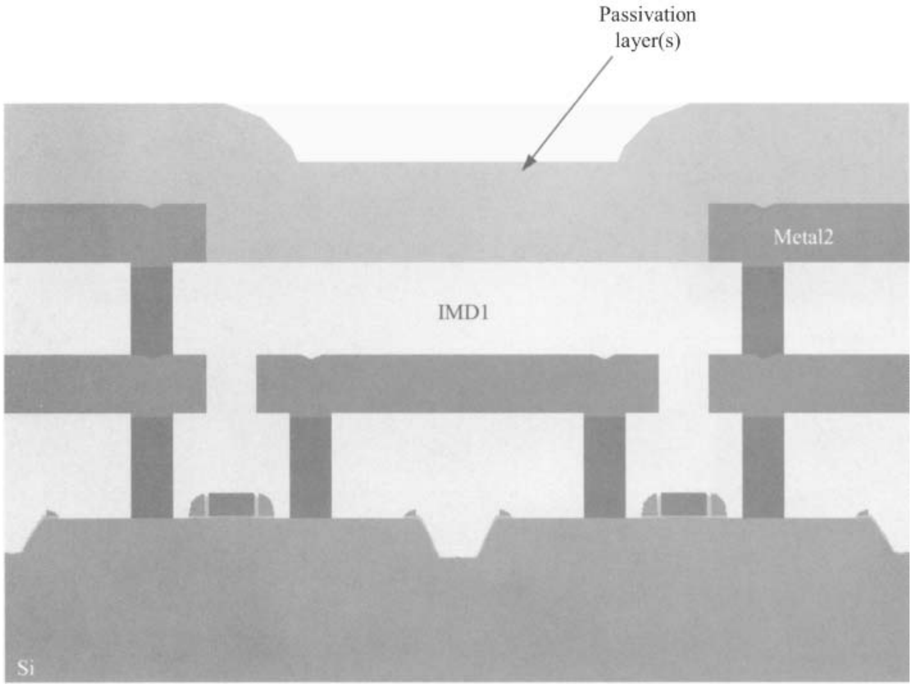


Figure 7.78 Deposition of final

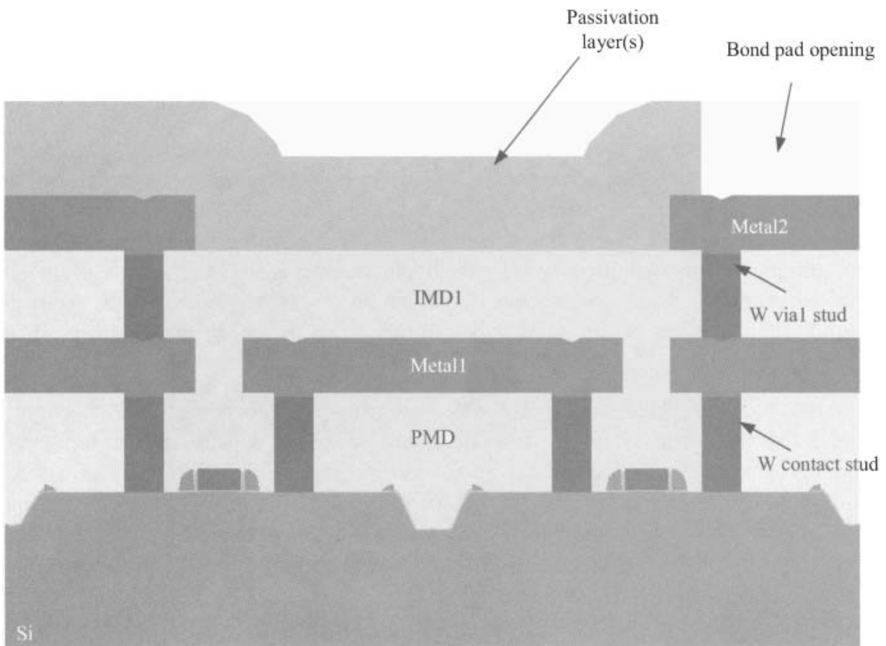


Figure 7.79 Bond pad definition using photolithography and dry etch of passivation.

In these cases, the die are marked, often with an ink dot, to indicate a nonfunctional circuit. Since the processing costs of a wafer are fixed, higher yield equates to higher profit.

Die Separation

Prior to separating the individual die, the backs of the wafers are often thinned using a lapping process similar to the CMP process discussed in Sec. 7.1. This thinning is often required for specific types of packages. Furthermore, thinning can aid in improving the removal of heat from the CMOS circuits.

Next, the die are separated from the wafers using a dicing saw comprised of a diamond-coated blade. The cutting paths are aligned to the *streets* or *scribe lines* on the wafer. Great care is given to minimize damage to the die during this mechanical separation. Along with the inked die, die that are observed to be damaged from the dicing are discarded. Obviously, the separation process can reduce the yield.

Packaging

The good die are now attached to a header in the appropriate package type. The die attach can be accomplished by either eutectic or epoxy attachment. Next, the bond pads are wired to the leads of the package. Common wire bonding techniques include thermocompression, thermosonic, and ultrasonic bonding. At this point, the packaging is completed by a wide range of different processes, greatly dependent on the type of package in which the IC resides. For instance, in plastic dual in-line packages (DIPs), a process similar to injection molding is performed to form a relatively inexpensive package. If ceramic packaging is required, then the attached, wire bonded die will reside in a cavity that is sealed by a metal lid. In general, plastic (or epoxy) packages are inexpensive, but do not provide a hermetic seal. On the contrary, ceramic packages are more costly, but do provide a hermetic seal. For information on other packaging schemes, the reader is referred to the list of additional reading at the end of the chapter. Finally, it should be noted that the packaging process can add to the overall yield loss.

Final Test and Burn-In

Once packaged, the CMOS parts are tested for final functionality and performance. When this is completed, it is common for the parts to go through a *burn-in* step. Here they are operated at extreme temperatures and voltages to weed out infant failures. Additional yield loss can be observed.

7.4 Summary

In this chapter the fundamental unit processes required in the manufacture of CMOS integrated circuits were introduced. These unit processes include thermal oxidation, solid state diffusion, ion implantation, photolithography, wet chemical etching, dry (plasma) etching, chemical mechanical polishing, physical vapor deposition, and chemical vapor deposition. Additionally, a brief overview of substrate preparation was given. With this foundation, a representative, deep-submicron CMOS process flow was provided and the significant issues in both the FEOL and BEOL integration were discussed. Finally, an overview of the backend processes was presented including wafer probe, die separation, packaging, and final test and burn-in.

ADDITIONAL READING

[1] S. A. Campbell, *Fabrication Engineering at the Micro- and Nanoscale*, 3rd ed., Oxford University Press, 2008. ISBN 978-0195320176

[2] M. J. Madou, *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd ed., CRC Publisher, 2002. ISBN 978-0849308260

[3] R. C. Jaeger, *Introduction to Microelectronic Fabrication*, 2nd ed., volume 5 of the Modular Series on Solid State Devices, Prentice-Hall Publishers, 2002. ISBN 0-20-144494-1

[4] S. A. Campbell, *The Science and Engineering of Micrelectronic Fabrication*, 2nd ed., Oxford University Press, 2001. ISBN 0-19-513605-5

[5] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology, Fundamentals, Practice, and Modeling*, Prentice-Hall Publishers, 2000. ISBN 978-0130850379