Chapter 4

Wafer Formation and Processing

4.1 Formation of Silicon and Gallium Arsenide Wafers[1]

4.1.1 Introduction

Integrated circuits (ICs) and discrete solid state devices are manufactured on semiconductor wafers. Silicon based devices are made on silicon wafers, while III-V (13-15) semiconductor devices are generally fabricated on GaAs wafers, however, for certain optoelectronic applications InP wafers are also used. The electrical and chemical properties of the wafer surface must be well controlled and therefore the preparation of starting wafers is a crucial portion of IC and device manufacturing. In order to obtain high fabrication yields and good device performance, it is very important that the starting wafers be of reproducibly high quality. For example, the front surface must be smooth and flat on both a macro- and microscale, because high-resolution patterns (lithography) are optically formed on the wafer. In principle, cutting a crystal into thin slices and polishing one side until all saw marks are removed and the surface appears smooth and glossy could produce a suitable wafer. However, due in part to the brittleness of Si and GaAs crystals, as well as the increasing requirements of wafer cleanliness and surface defect reduction with ever decreasing device geometries, a very complex series of processing steps are required to produce analytically clean, flat and damage-free wafer surfaces.

The following focuses on the general principles and methods with regard to wafer formation. Detailed formulas, recipes, and specific process parameters are not given as they vary considerably among different wafer producers. However, in general, techniques for fabrication of Si wafers have generally become standardized within the semiconductor industry. In contrast, GaAs wafer technology is less standardized, possibly due to either (a) the similarity to silicon practices or (b) the lower production volume of GaAs wafers. There are two general classes of processes in the methodology of making wafers: mechanical and chemical. As both Si and GaAs are brittle materials, the mechanical processes for their wafer fabrication are similar. However, the different chemistry of Si and GaAs require that the chemical processes be dealt with separately.

4.1.2 Wafer formation procedures

Each of the processing steps in the conversion of a semiconductor ingot (formed by Czochralski or Bridgeman growth) into a polished wafer ready for device fabrication, results in the removal of material from the original ingot; between $^1/_3$ and $^1/_2$ of the original ingot is sacrificed during processing. Methods for the removal of material from a crystal ingot are classified depending on the size of the particles being removed during the process. If the removed particles are much larger than atomic or molecular dimensions the process is described as being macro-scale. Conversely, if the material is removed atom-by-atom or molecule-by-molecule then the process is termed micro-scale. A further distinction between various types of processes is whether

[1] This content is available online at <http://cnx.org/content/m16627/1.5/>.

the removal occurs as a result of mechanical or chemical processes. The formation of a finished wafer from a semiconductor ingot normally requires six machining (mechanical) operations, two chemical operations, and at least one polishing (chemical-mechanical) operation. Additionally, multiple inspection and evaluation steps are included in the overall process. A summary of the individual steps, and their functions, involved in wafer production is shown in Table 4.1.

Process	**Type**	**Function**
cropping	mechanical	removal of conical shaped ends and impure portions
grinding	mechanical	obtain precise diameter
orientation flatting	mechanical	identification of crystal orientation and dopant type
etching	chemical	removal of surface damage
wafering	mechanical	formation of individual wafers by cutting
heat treatment	thermal	annihilation of undesirable electronic donors
edge contouring	mechanical	provide radius on the edge of the wafer
lapping	mechanical	provides requisite flatness of the wafer
etching	chemical	removal of surface damage
polishing	mechano-chemical	provides a smooth (specular) surface
cleaning	chemical	removal of organics, heavy metals, and particulates

Table 4.1: Summary of the process steps involved in semiconductor wafer production.

4.1.3 Crystal shaping

Although an as-grown crystal ingot is of high purity (99.9999%) and crystallinity, it does not have the sufficiently precise shape required for ready wafer formation. Thus, prior to slicing an ingot into individual wafers, several steps are needed. These operations required to prepare the crystal for slicing are referred to as crystal shaping, and are shown in Figure 4.1.

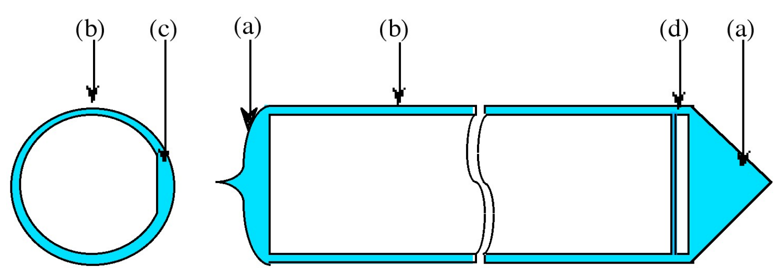


Figure 4.1: Schematic representation of crystal shaping operations: (a) remove crown and taper, (b) grind to required diameter, (c) grind flat, and (d) slice sample for measurements. Shaded area represents material removed.

4.1.3.1 Cropping

The as-grown ingots have conical shaped seed (top) and tang (bottom) ends that are removed using a circular diamond saw for ease of further manipulation of the ingot (Figure 4.1a). The cuttings are sufficiently pure that they are cleaned and the recycled in the crystal growth operation. Portions of the ingot that fail to meet specifications of resistivity are also removed. In the case of silicon ingots these sections may be sold as metallurgical-grade silicon (MGS). Conversely, portions of the crystal that meet desired resistivity specifications may be preferentially selected. A sample slice is also cut to enable oxygen and carbon content to be determined; usually this is accomplished by Fourier transform infrared spectroscopic measurements (FT-IR). Finally, cropping is used to cut crystals to a suitable length to fit the saw capacity.

4.1.3.2 Grinding

The primary purpose of crystal grinding is to obtain wafers of precise diameter because the automatic diameter control systems on crystal growth equipment are not capable of meeting the tight wafer diameter specifications. In addition, crystals are seldom grown perfectly round in cross section. Thus, ingots are usually grown with a 1 - 2 mm allowance and reduced to the proper diameter by grinding Figure 4.1b.

Crystal grinding is a straightforward process using an abrasive grinding wheel, however, it must be well controlled in order to avoid problems in subsequent operations. Exit chipping in wafering and lattice slip in thermal processing are problems often resulting from improper crystal grinding. Two methods are used for crystal grinding: (a) grinding on center and (b) centerless grinding.

Figure 4.2 shows a schematic of the general set-up for grinding a crystal ingot on center. The crystal is supported at each end in a lathe-like machine. The rotating cutting tool, employing a water-based coolant, makes multiple passes down the rotating ingot until the requisite diameter is obtained. The center grinder can also be used for grinding the identification flats as well as providing a uniform ingot diameter. However, grinding the crystal on centers requires that the operator locate the crystal axis in order to obtain the best yield.

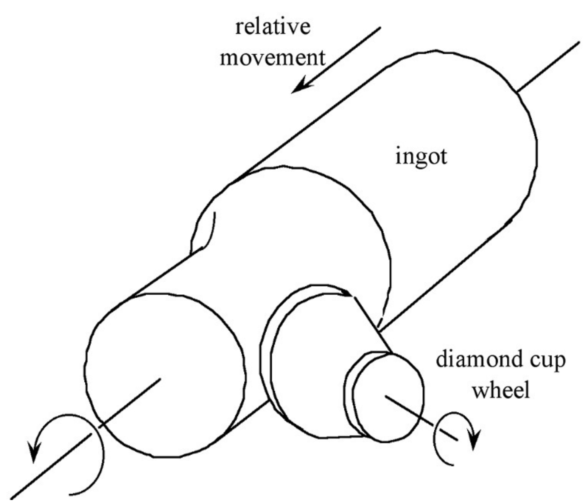


Figure 4.2: Schematic representation of grinding on center.

Centerless grinding eliminates the problems associated with locating the crystal center. The centerless method is superior for long crystals; however, a centerless grinder is much larger than a center grinder of the same diameter capacity. In centerless grinding the ingot is supported between two wheels, a grinding wheel and a drive wheel. A schematic of the centerless grinder is shown in Figure 4.3. The axis of the drive wheel is canted with respect to that of the crystal ingot and the grinding wheel pushing the crystal ingot past the stationary (but rotating) grinding wheel, see Figure 4.3b.

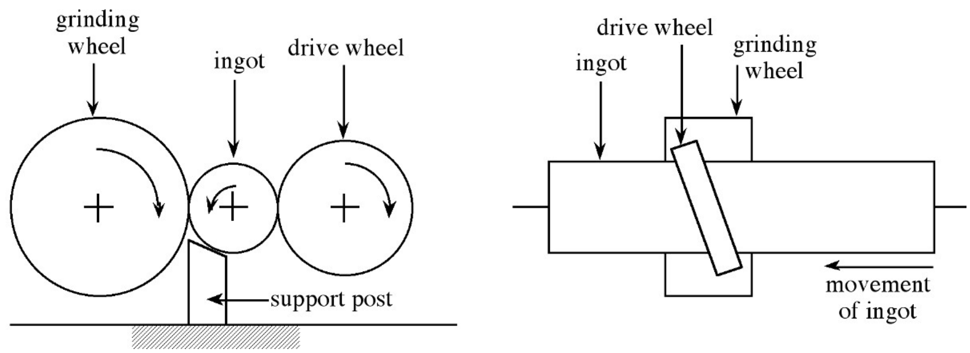


Figure 4.3: Schematic representation of centerless grinding viewed (a) along and (b) perpendicular to the crystal axis.

4.1.3.3 Orientation/identification flats

Following grinding of the ingot to the desired diameter, one or two flats are ground along the length of the ingot. The identification flats (one or two) are ground lengthwise along the crystal according to the orientation and the dopant type. After grinding the crystal on centers the crystal is rotated to the proper orientation, then the wheel is positioned with its axis of rotation perpendicular to the crystal axis and moved along the crystal from end to end until the appropriate flat size is obtained. An optical or X-ray orientation fixture may be used in conjunction with the crystal mounting to facilitate the proper orientation of the crystal on the grinder.

The largest flat is called the primary flat (Figure 4.1c) and is parallel to one of the crystal planes, as determined by X-ray diffraction. The primary flat is used for automated positioning of the wafer during subsequent processing steps, e.g., lithographic patterning and dicing. Other smaller flats are called "secondary flats" and are used to identify the crystal orientation (<111> versus <100>) and the material (n-type versus p-type). Secondary flats provide a quick and easy manner by which unknown wafers can be sorted. The flats shown schematically in Figure 4.4 are located according to a Semiconductor Equipment and Materials Institute (SEMI®) standard and are ground to specific widths, depending upon crystals diameter. Notches are also used in place of the secondary flat; however, the relative orientations of the notch and primary flat with regard to crystal orientation and dopant are maintained.

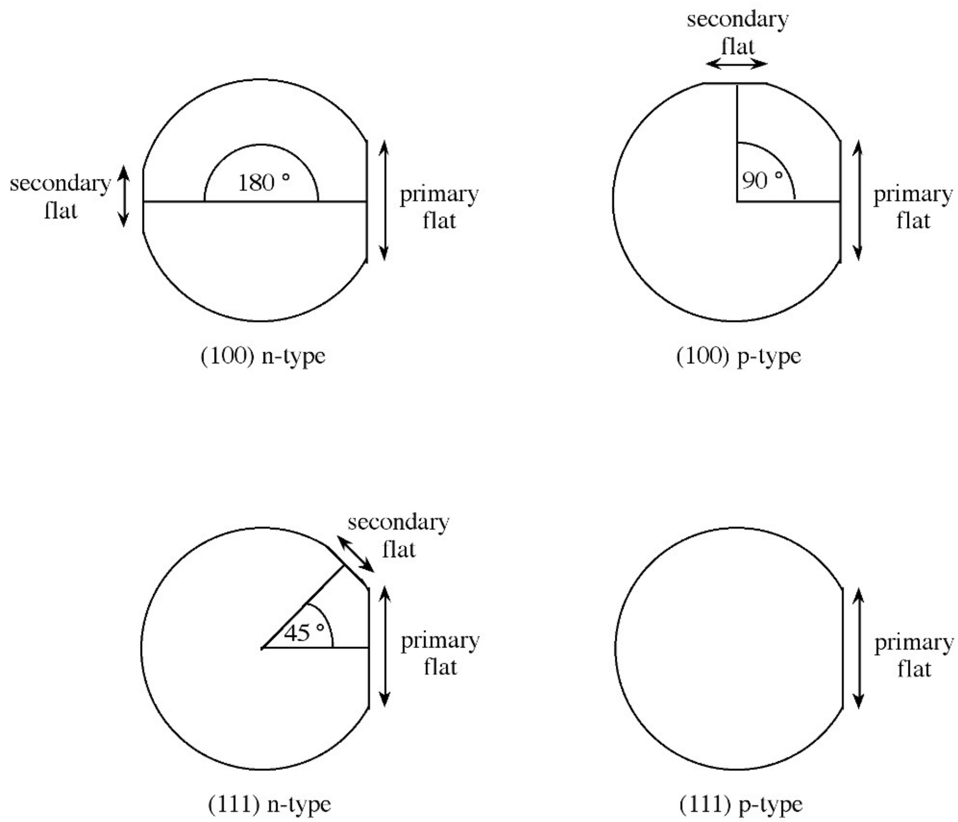


Figure 4.4: SEMI locations for orientation/identification flats.

4.1.3.4 Etching

The cropping and grinding processes are performed with relatively coarse abrasive and consequently a great deal of subsurface damage results. Pits, chips, and cracks all contribute to stress in the cut wafer and provide nuclei for crack propagation at the edges of the finished wafer. If regions of stress are removed then cracks will no longer propagate, reducing exit chipping and wafer breakage during subsequent fabrication steps.

The general method for removing surface damage is to etch the crystal in a hot solution. The most common etchants for Si are based on the HNO_3-HF system, in which etchant modifiers such as acetic acid also commonly used. In the case of GaAs HCl-HNO_3 is the appropriate system. These etchants selectively attack the crystal at the damaged regions. After etching, the crystal is transferred to the slicing preparation area.

4.1.4 Wafering

The purpose of wafering is to saw the crystal into thin slices with precise geometric dimensions. By far, the most common method of wafering semiconductor crystals is the use of an annular, or inner diameter (ID), diamond saw blade. A schematic diagram of ID slicing technology is shown in Figure 4.5.

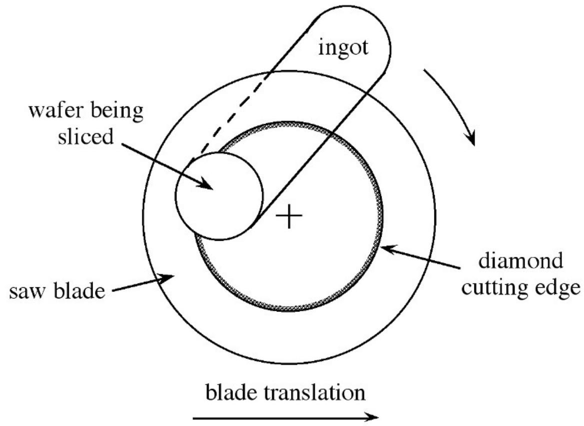


Figure 4.5: Schematic diagrams of ID slicing process.

The crystal, when it arrives at the sawing area, has been ground to diameter, flatted, and etched. In order to slice it, the crystal must be firmly mounted in such a way that it can be completely converted to wafers with minimum waste. The crystal is attached with wax or epoxy to a mounting block, which is usually cylindrical in shape and of the same diameter as the ingot. Also, a mounting beam (or strip) is attached along the length of the crystal at the breakout point of the saw blade. This reduces exit chipping (breakage that occurs as the blade exits the crystal at the end of a cut) and also provides support for the sawn wafer until it is retrieved. Graphite or phenolic resins are common materials for the mounting block and beams, although some success has been obtained in mounting ingots using hydraulic pressure. The saw blade is a thin sheet of stainless steel (325 μm), with diamond bonded to its inner edge. This blade is mounted on a drum that rotates at *ca.* 2000 rpm. Saw blades 58 cm ($\approx$23 inches) in diameter with a 20 cm (8 inches) opening are common, however, as wafer sizes increase larger blades are employed: 30 cm (12 inches) wafers are now common for Si. The blade moves relative to the stationary crystal at a speed of 0.05 cm/s, and the cutting process is water-cooled. Thus, considering that wafers are sliced sequentially (one at a time), the overall process is very slow. A further problem is that the kerf loss (loss due to the width of the blade) results in approximately 1/3 of the material being lost as saw dust. Finally, the depth of the drum onto which the blade is attached limits the length of the ingot section that is accessible. In order to overcome this problem, another style of ID blade saw was developed in which the blade is mounted on an air bearing and is rotated by a belt drive. This allows the entire length of the crystal ingot to be sliced.

Both silicon and GaAs crystals are grown with either the crystallographic <100> or <111> direction parallel to the cylindrical axis of the crystal. Wafers may be cut either exactly perpendicular to the crystallographic axis or deliberately off-axis by several degrees. In order to obtain the proper wafer orientation, the crystal must be properly oriented on the saw. All production slicing machines have adjustments for orientation of the crystal; however, it is usually necessary to check the orientation of the first slice in order to assure that all subsequent slices will be properly oriented.

Obvious variables introduced during the wafering process include: cutting rate, wheel speed, and coolant flow rate. However, the condition of the machines, such as alignment and vibration, is the most important variable followed by the condition of the blade. A deviated blade rim may cause taper, bow, or warp. Table 4.2 summarizes the types of deformations that can occur during wafering, their physical appearance and their characteristics.

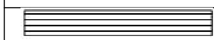
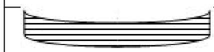
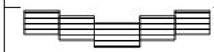
Type of bow and warp	Surface appearance	Lattice curvature	Comments
	flat	flat	ideal
	curved	flat	
	curved	curved	
	flat	curved	
	curved	flat	slips

Table 4.2: Deformed wafers and their characteristics.

4.1.5 Heat treatment

As-produced Czochralski grown crystals often have a level of oxygen impurity that may exceed the concentration of dopant in the semiconductor material (i.e., Si or GaAs). This oxygen impurity has a deleterious effect on the semiconductor properties, especially upon subsequent thermal processing, e.g., thermal oxide growth or epitaxial film growth by metal organic chemical vapor deposition (MOCVD). For example, when silicon crystals are heated to about 450 °C the oxygen undergoes a transformation that causes it to behave as an electron donor, much like an n-type dopant. These oxygen donors, or "thermal donors", mask the true resistivity of the semiconductor because they either add additional carrier electrons to a n-type crystal or compensate for the positive holes in a p-type crystal. Fortunately, these thermal donors can be "annihilated" by heat treating the materials briefly in the range of 500 - 800 °C and then cooling quickly through the 450 °C region before donors can reform. In principle thermal donor annihilation can be performed on wafers at any time during their fabrication; however, it is usually best to perform the heat treatment immediately after wafering since sub-standard wafers may be rejected before additional processing steps are undertaken and thus limiting additional cost. Donor annihilation is a bulk effect, and therefore the thermal treatment can be performed in air, since any surface oxide that may form will be removed in subsequent lapping and polishing steps.

4.1.6 Lapping or grinding

The as-cut wafers vary sufficiently in thickness to require an additional operation, the slicing operation does not consistently produce the required flatness and parallelism required for many wafer specifications, see Table 4.2. Since conventional polishing does not correct variations in flatness or thickness, a mechanical two-sided lapping operation is performed. Lapping is capable of achieving very precise thickness uniformity, flatness and parallelism. Lapping also prepares the surface for polishing by removing the sub-surface sawing damage, replacing it with a more uniform and smaller lapping damage.

The process used for lapping semiconductor wafers evolved from the optical lens manufacturing industry using principles developed over several hundred years. However, as the lens has a curved surface and the wafers are flat, the equipment for lapping wafers is mechanically simpler than lens processing machines. The simplest double-side lapping machine consists of two very flat counter-rotating plates, carriers to hold and move the wafers between the plates, and a device to feed abrasive slurry steadily between the plates. The abrasive is typically a 9 μm Al_2O_3 grit. Commercial abrasives are suspended in water or glycerin with proprietary additives to assist in suspension and dispersion of the particles, to improve the flow properties of the slurry, and to prevent corrosion of the lapping machine. Hydraulics or an air cylinder applies lapping pressure with low starting pressure for 2 to 5 minutes, which is then increased through most of the process. The completion of lapping may be determined by elapsed time or by an external thickness sensing device. The finished process gives a wafer with a surface uniform to within 2 μm. Approximately 20 μm per side is removed during the lapping process.

Although lapping would appear to be simple in concept, the successful implementation of a production lapping operation requires the development of a technique and experience to achieve acceptable quality with good yields. Small adjustments to the rotation rates of the plates and carriers will cause the plates to wear concave, convex or flat.

As lapping is a messy process, various efforts have been made to avoid it or to substitute an alternative process. The most likely approach at present is grinding, in which the wafer is held on a vacuum chuck and a series of progressively finer diamond wheels is moved over the wafer while it is rotated on a turn table. Grinding gives a clearer surface than lapping, however, only one side may be ground at a time and the resulting flatness is not as good as that obtained by lapping.

4.1.7 Edge contouring

The rounding of the edge of the wafer to a specific contour is a fairly recent development in the technology of wafer preparation. It was known by the early seventies that a significant number of device yield problems could be traced to the physical condition of the wafer edge. An acute edge affects the strength of the wafer due to: stress concentration, and a lowering of its resistance to thermal stress, as well as being the source of particle chip, breakage, and lattice damage. In addition, the particles originating from the chipped edges can, if present on the wafer surface, add to the defect density (D_0) of the IC process reducing fabrication yield. Further problems associated with a square edge include the build-up of photoresist at the wafer edge. The solution to these process problems is to provide a contoured edge with a defined radius (r).

Chemical etching of wafers results in a degree of edge rounding, but it is difficult to control. Thus, mechanical edge contouring has been developed and the result has been a dramatic improvement in yields in downstream wafer processing. Losses due to wafer breakage are also reduced. The edge contouring process is usually performed in cassette-fed high speed equipment, in which each wafer is rotated rapidly against a shaped cutting tool (Figure 4.6).

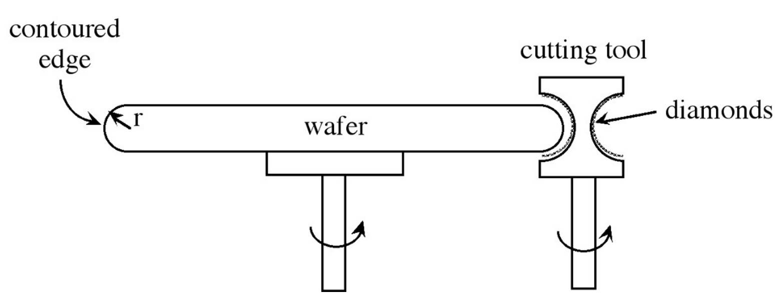


Figure 4.6: Schematic illustration of edge contouring.

4.1.8 Etching

The mechanical processes described above to shape the wafer leave the surface and edges damaged and contaminated. The depth of the work damage depends on the specific process, however, 10 μm is typical. Such damage is readily removed by chemical etching. Etching is used at multiple points during the fabrication of a semiconductor device. The discussion below is limited to etches suitable for wafer fabrication, i.e., non-selective etching of the entire wafer surface.

4.1.8.1 Wet chemical etching

The wet chemical etching of any material can be considered to involve three steps: (a) transportation of the reactants to the surface, (b) reaction at the surface, and (c) movement of the reaction products into the etchant solution (Figure 4.7). Each of these may be the rate limiting step and thus control the etch rate and uniformity. This effect is summarized in Table 4.3.

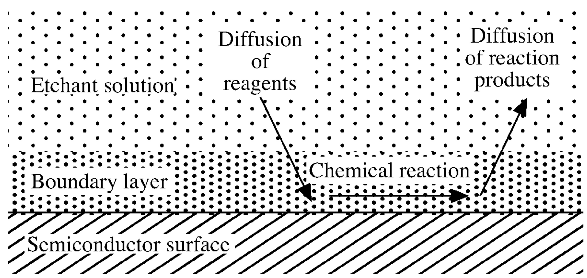


Figure 4.7: Schematic representation of the three steps involved in wet chemical etching: (i) diffusion of the chemical etch reagents through the boundary layer, (ii) chemical reaction at the surface, and (iii) diffusion of the reaction products into the etch solution through the boundary layer.

Rate limiting step	**Etching rate**	**Results**	**Comments**
Diffusion of reagent to the surface	slow	etching(anisotropic)	enhanced surface roughness
Reaction at semiconductor surface	fast	polishing(isotropic)	ideal
Diffusion of reaction products from the surface	slow	polishing(isotropic)	reaction product remains on surface

Table 4.3: Effects of rate limiting step in semiconductor etching.

An etchant that is limited by the rate of reaction at the surface will tend to enhance any surface features and promote surface roughness due to preferential etching at defects (anisotropic). In contrast, if the etch rate is limited by the diffusion of the etchant reagent through a stagnant (dead) boundary layer near the surface, then the etch will result in uniform polishing and the surface will become smooth (isotropic). If removal of the reaction products is rate limiting then the etch rate will be slow because the etch equilibrium will be shifted towards the reactants. In the case of an individual etchant reaction, the rate determining step may be changed by rapid stirring to aid removal of reaction products, or by increasing the temperature of the etch solution, see Figure 4.8. The exact etching conditions are chosen depending on the application. For example, dilute high temperature etches are often employed where the etch damage must be minimized, while cooled etches can be used where precise etch control is required.

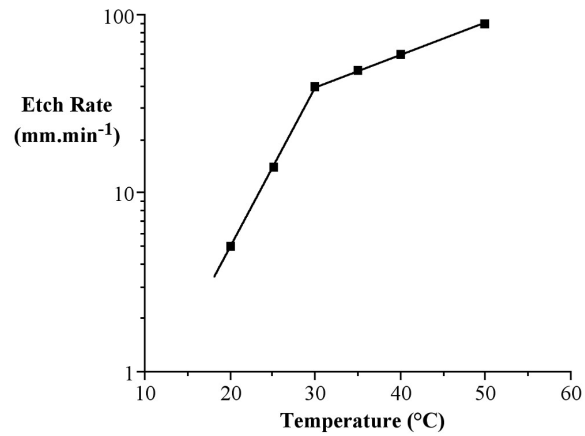


Figure 4.8: Typical etch rate versus temperature plot for a mixture of HF (20%), nitric acid (45%), and acetic acid (35%).

Traditionally mixtures of hydrofluoric acid (HF), nitric acid (HNO_3) and acetic acid ($MeCO_2H$) have been used for silicon, but alkaline etches using potassium hydroxide (KOH) or sodium hydroxide (NaOH) solutions are increasingly common. Similarly, gallium arsenide etches may be either acidic or basic, however, in both cases the etches are oxidative due to the use of hydrogen peroxide. A wide range of chemical reagents are commercially available in "transistor grade" purity and these are employed to minimize contamination of the semiconductor. Deionized water is commonly used as a diluent for each of these reagents and the concentration of commonly used aqueous reagents is given in Table 4.4.

Reagent	Weight %	Reagent	Weight %
HCl	37	HF	49
H_2SO_4	98	H_3PO_4	85
HNO_3	79	$HClO_4$	70
$MeCO_2H$	99	H_2O_2	30
NH_4OH	29		

Table 4.4: Weight percent concentration of commonly used concentrated aqueous reagents.

The equipment used for a typical etchant process includes an acid (or alkaline) resistant tank, which contains the etchant solution and one or more positions for rinsing the wafers with deionized water. The process is batch in nature involving tens of wafers and the best equipment provides a means of rotating the wafers during the etch step to maintain uniformity. In order to assure the removal of all surface damage, substantial over-etching is performed. Thus, the removal of 20 μm from each side of the wafer is typical. Etch times are usually several minutes per batch.

4.1.8.2 Etching silicon

The most commonly used etchants for silicon are mixtures of hydrofluoric acid (HF) and nitric acid (HNO_3) in water or acetic acid ($MeCO_2H$). The etching involves a reduction-oxidation (redox) reaction, followed by dissolution of the reaction products. In the HF-HNO_3 system the HNO_3 oxidizes the silicon and the HF removes the reaction products from the surface. The overall reaction is:

$$Si + HNO_3 + 6\ HF \rightarrow H_2SiF_6 + HNO_2 + H_2O \tag{4.1}$$

The oxidation reaction involves the oxidation of Si^0 to Si^{4+}, and it is auto-catalytic in that the reaction product promotes the reaction itself. The initial step involves trace impurities of HNO_2 in the HNO_3 solution, (4.2), which react to liberate nitrogen dioxide (NO_2), (4.3).

$$HNO_2 + HNO_3 \rightarrow N_2O_4 + H_2O \tag{4.2}$$

$$N_2O_4 \rightarrow 2\ NO_2 \tag{4.3}$$

The nitrogen dioxide oxidizes the silicon surface in the presence of water, resulting in the formation of $Si(OH)_2$ and the reformation of HNO_2, (4.4). The $Si(OH)_2$ decomposes to give SiO_2, (4.5). Since the reaction between HNO_2 and HNO_3, (4.2), is rate limiting, an induction period is observed. However, this is overcome by the addition of NO_2^- ions in the form of $[NH_4][NO_2]$.

$$Si^0 + 2\ NO_2 + 2\ H_2O \rightarrow Si(OH)_2 + 2\ HNO_2 \tag{4.4}$$

$$Si(OH)_2 \rightarrow SiO_2 + H_2 \tag{4.5}$$

The final step of the etch process is the dissolution of the SiO_2 by HF, (4.6). Stirring serves to remove the soluble products from the reaction surface. The role of the HF is to act as a complexing reagent, and thus the reaction shown in (4.6) is known as a complexing reaction. The formation of water as a reaction product requires that acetic acid be used as a diluent (solvent) to ensure better control.

$$SiO_2 + 6\ HF \rightarrow H_2SiF_6 + H_2O \tag{4.6}$$

The etching reaction is highly dependent on the relative ratios of the etchant reagents. Thus, if an HF-rich solution is used, the reaction is limited by the oxidation step, (4.4), and the etching is anisotropic, since the oxidation reaction is sensitive to doping, crystal orientation, and defects. In contrast, the use of a HNO_3-rich solution produces isotropic etching since the dissolution process is rate limiting (Table 4.3). The reaction of HNO_3-rich solutions has been found to be diffusion-controlled over the temperature range 20 - 50 °C (Figure 4.8), and is therefore commonly employed for removing work damage produced during wafer fabrication. The boundary layer thickness (Figure 4.7) and therefore the dimensional control over the wafer is controlled by the rotation rate of the wafers. A common etch formulation is a 4:1:3 mixture of HNO_3 (79%), HF (49%), and $MeCO_2H$ (99%). There are some etchant formulations that are based on alternative (or additional) oxidizing agents, such as: Br_2, I_2, and $KMnO_4$.

Alkaline etching (KOH/H_2O or $NaOH/H_2O$) is by nature anisotropic and the etch rate depends on the number of dangling bonds which in turn are dependent on the surface orientation. Since etching is reaction rate limited no rotation of the wafers is necessary and excellent uniformity over large wafers is obtained. Alkaline etchants are used with large wafers where dimensional uniformity is not maintained during lapping. A typical formulation uses KOH in a 45% weight solution in H_2O at 90 °C.

4.1.8.3 Etching gallium arsenide

Although a wide range of etches have been investigated for GaAs, few are truly isotropic. This is because the surface activity of the (111) Ga and (111) As faces are very different. The As rich face is considerably more reactive than the Ga rich face, thus under identical conditions it will etch faster. As a result most etches give a polished surface on the As face, but the Ga face tends to appear cloudy or frosted due to the highlighting of surface features and crystallographic defects.

As with silicon the etch systems involve oxidation and complexation. However, in the case of GaAs the gallium is already fully oxidized (formally Ga^{3+}), thus, it is the arsenic (formally the arsenide ion, As^{3-} that is oxidized by a suitable oxidizing agent (e.g., H_2O_2) to the soluble oxide, As_2O_3, (4.7). The gallium ions form the oxide Ga_2O_3 via the hydroxide, (4.8). Both oxides are soluble in acid solutions, resulting in their removal from the surface.

$$2\,As^{3-} + 6\,H_2O_2 + H^+ \rightarrow As_2O_3 + 5\,OH^- + 4\,H_2O \tag{4.7}$$

$$2\,Ga^{3+} + 6\,OH^- \rightarrow 2\,Ga(OH)_3 + 3\,H_2O \tag{4.8}$$

The peroxide based oxidative etches for GaAs are divided into acidic and basic etches. The composition and application of some of these systems are summarized in Table 4.5. The most widely used of these is $H_2SO_4/H_2O_2/H_2O$ and is referred to as Caro's acid. The high viscosity of H_2SO_4 results in diffusion-limited etching with high acid concentrations. Etches with low acid concentrations tend to be anisotropic. Phosphoric acid (H_3PO_4) or citric acid (Figure 4.9) may be exchanged for sulfuric acid (H_2SO_4). Replacement of the acid component with bases such as NH_4OH or NaOH can result in near to truly isotropic etchants, although certain combinations can result in strong anisotropy.

Formulation	**Volume ratio**	**(100) etch rate (μm/min)**	**(110) etch rate (μm/min)**	**(111)As etch rate (μm/min)**	**(111)Ga etch rate (μm/min)**
H_2SO_4/ H_2O_2/ H_2O	8:1:1	1.5	1.5	1.5	0.8
H_2SO_4/ H_2O_2/ H_2O	1:8:1	8.0	8.0	12.0	3.0
H_2SO_4/ H_2O_2/ H_2O	3:1:50	0.8	0.8	0.8	0.4
citric acid/ H_2O_2/ H_2O	1:1:1	0.6	0.6	0.6	0.4
NH_4OH/ H_2O_2/ H_2O	1:700	0.3	0.3	0.3	0.3
NaOH/ H_2O_2/ H_2O	1:0.76	0.2	0.2	0.2	0.2

Table 4.5: The composition and application of selected etch systems for GaAs.

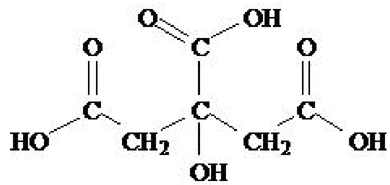


Figure 4.9: Structure of citric acid.

One of the earliest etching systems for GaAs is based on the use of a dilute (ca. 0.05 vol.%) solution of bromine (Br_2) in ethanol. The Br_2 acts as the oxidant, resulting in the formation of soluble bromides. The etch rate of this system is different for different crystallographic planes, i.e., the etch rates for the (111) As, (100), and (111) Ga faces are in the ratio 6:5:1, although more uniform etch rates are observed with high Br_2 concentrations (ca. 10 vol.%). These higher concentration solutions are used for the removal of damage due to cutting with the saw.

4.1.9 Polishing

The purpose of polishing is to produce a smooth, specular surface on which device features can be defined by lithography. In order to allow for very large scale integration (VLSI) or ultra large scale integration (ULSI) fabrication the wafer must have a surface with a high degree of flatness. Variations less than 5 to 10 μm across the wafer diameter are typical flatness specifications. In addition, given the preceding steps, wafer polishing must not leave residual contamination or surface damage. The techniques of wafer polishing are derived from the glass lens industry, with some important modifications that have been developed to meet the special requirements of the microelectronics industry.

4.1.9.1 Differences between polishing and lapping

If the surface of a wafer that has undergone lapping (or grinding) is examined with an electron microscope, cracks, ridges and valleys are observed. The top "relief layer" consists of peaks and valleys. Below this layer is a damaged layer characterized by microcracks, dislocations, slip and stress. Figure 4.10 shows a schematic representation of the abraded surface. Both of these layers must be removed completely prior to further fabrication. Decreasing the particle size of the abrasive during lapping only decreases the scale of the damage, but does not eliminate it entirely. In fact this surface damage is a characteristic of the brittle fracture of single crystal Si and GaAs, and occurs because during lapping the abrasive grains are moved across the surface under a pressure beyond that of the fracture strength of the wafer materials (Si or GaAs). In contrast to the mechanical abrasion employed in lapping, polishing is a mechano-chemical process during which brittle fracture does not occur. A polished wafer does not display any evidence of a relief surface such as that produced by lapping, even at highest resolution electron microscope.

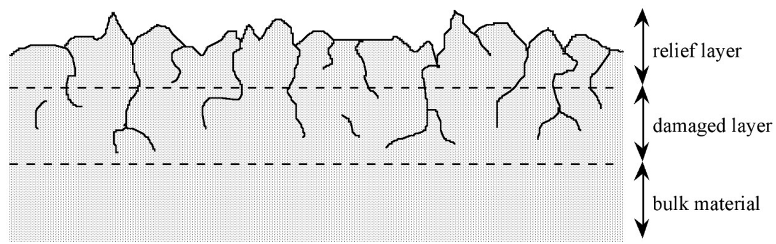


Figure 4.10: Schematic representation of a cross sectional view of an abraded wafer surface prior to polishing.

4.1.9.2 Process of Polishing

Figure 4.11 shows a schematic of the polishing process. Polishing may be conducted on single wafers or as a batch process depending on the equipment employed. Single wafer polishing is preferred for larger wafers and allows for better surface flatness. In both processes, wafers are mounted onto a fixture, by either wax or a composite Felx-Mount™, and pressed against the polishing pad. The polishing pad is usually made from an artificial fabric such as polyester felt-polyurethane laminate. Polishing is accomplished by a mechano-chemical process in which aqueous polishing slurry is dripped onto the polishing pad, see Figure 4.11. The polishing slurry performs both a chemical and mechanical process, and consists of fine silica (SiO_2) particles (100 Å diameter) and an oxidizing agent. Aqueous sodium hydroxide (NaOH) is used for Si, while aqueous sodium chlorate (NaOCl) is preferred for GaAs. Suspending agents are usually added to prevent settling of the silica particles. Under the heat caused by the friction of the wafer on the polishing pad the wafer surface is oxidized, which is the chemical step, while in the mechanical step the silica particles in the slurry abrade the oxidized surface away.

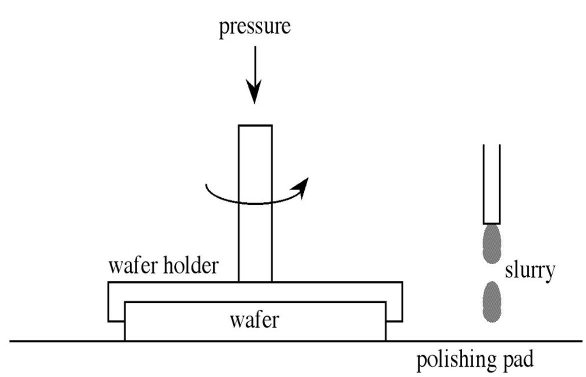


Figure 4.11: Schematic representation of the wafer polishing process.

In order to achieve a reasonable rate of removal of the relief and damaged layers and still obtain the highest quality surface, the polishing is done in two steps, stock removal and haze removal. The former is carried out with a higher concentration slurry and may proceed for about 30 minutes at a removal rate of 1 μm/min. Haze removal is performed with a very dilute slurry, a softer pad with a reaction time of about 5 to 10 minutes, during which the total amount of material removed is only about 1 μm. Due to the active chemical reaction between the wafer and the polishing agent, the wafers must be rinsed in deionized water immediately after polishing to prevent haze or stains from reforming.

There are many variables that will influence the rate and quality of polishing. High pressure results in a higher polishing rate, but excessive pressure may cause non-uniform polishing, excessive heat generation and fast pad wear. The rate of polishing is increased with higher temperatures but this may also lead to haze formation. High wheel speeds accelerate the polishing rate but can raise the temperature and also results in problems in maintaining a uniform flow of slurry across the pad. Dense slurry concentrations increase the polishing rate but are more costly. The pH of the slurry solution can also affect the polishing rate, for example the polishing rate of Si gradually increases with increased pH (higher basicity) until a pH of about 12 where a dramatic decrease is observed. In general, the optimum polishing process for a given facility depends largely upon the interplay of product specification, yields, cost, and quality considerations and must be developed uniquely. The wafer polishing process does not improve the wafer flatness and, at best, polishing will not degrade the wafer flatness achieved in the lapping operation.

4.1.10 Cleaning

During the processes described above, semiconductor wafers are subjected to physical handling that leads to significant contamination. Possible sources of physical contamination include:

a. airborne bacteria,
b. grease and wax from cutting oils and physical handling,
c. abrasive particulates (usually, silica, silicon carbide, alumina, or diamond dust) from lapping, grinding or sawing operations,
d. plasticizers which are derived from containers and wrapping in which the wafers are handled and shipped.

Chemical contamination may also occur as a result of improper cleaning after etch steps. Light-metal (especially sodium and potassium) species may be traced to impurities in etchant solutions and are chemisorbed on to the surface where they are particularly problematical for metal oxide semiconductor (MOS) based devices, although higher levels of such impurities are tolerable for bipolar devices. Heavy metal impurities (e.g., Cu, Au, Fe, and Ag) are usually caused by electrodeposition from etchant solutions during fabrication. While wafers are cleaned prior to shipping, contamination accumulated during shipping and storage necessitates that all wafers be subjected to scrupulous cleaning prior to fabrication. Furthermore, cleaning is required at each step during the fabrication process. Although wafer cleaning is a vital part of each fabrication step, it is convenient to discuss cleaning within the general topic of wafer fabrication.

4.1.10.1 Cleaning silicon

The first step in cleaning a Si wafer is removal of all physical contaminants. These contaminates are removed by rinsing the wafer in hot organic solvents such as 1,1,1-trichloroethane (Cl_3CH_3) or xylene ($C_6H_4Me_2$), accompanied by mechanical scrubbing, ultrasonic agitation, or compressed gas jets. Removal of the majority of light metal contaminants is accomplished by rinsing in hot deionized water, however, complete removal requires a further more aggressive cleaning process. The most widely used cleaning method in the Si semiconductor industry is based on a two step, two solution sequence known as the "RCA Cleaning Method".

The first solution consists of H_2O-H_2O_2-NH_4OH in a volume ratio of 5:1:1 to 7:2:1, which is used to remove organic contaminants and heavy metals. The oxidation of the remaining organic contaminants by the hydrogen peroxide (H_2O_2) produces water soluble products. Similarly, metal contaminants such as cadmium, cobalt, copper, mercury, nickel, and silver are solubilized by the NH_4OH through the formation of soluble amino complexes, e.g., (4.9).

$$2\ Cu^{2+}(s) + 6\ NH_4OH\ (aq) \rightarrow [Cu(NH_3)_6\ (aq) \quad (4.9)$$

The second solution consists of H_2O-H_2O_2-HCl in a 6:1:1 to 8:2:1 volume ratio and removes the Group I(1), II(2) and III(13) metals. In addition, the second solution prevents re-deposition of the metal contaminants. Each of the washing steps is carried out for 10 - 20 min. at 75 - 85 °C with rapid agitation. Finally, the wafers are blown dry under a stream of nitrogen gas.

4.1.10.2 Cleaning GaAs

In principle GaAs wafers may be cleaned in a similar manner to silicon wafers. The first step involves successive cleaning with hot organic solvents such as 1,1,1-trichloroethane, acetone, and methanol, each for 5-10 minutes. GaAs wafers cleaned in this manner may be stored under methanol for short periods of time.

Most cleaning solutions for GaAs are actually etches. A typical solution is similar to the second RCA solution and consists of an 80:10:1 ratio of H_2O-H_2O_2-HCl. This solution is generally used at elevated temperatures (70 °C) with short dip times since it has a very fast etch rate (4.0 μm/min).

4.1.11 Measurements, inspections and packaging

Quality control measurements of the semiconductor crystal and subsequent wafer are performed throughout the process as an essential part of the fabrication of wafers. From crystal and wafer shaping through the final wafer finishing steps, quality control measurements are used to ensure that the materials meets customer specifications, and that problems can be corrected before they create scrap material and thus avoid further processing of reject material. Quality control measurements can be broadly classified into mechanical, electrical, structural, and chemical.

Mechanical measurements are concerned with the physical dimensions of the wafer, including: thickness, flatness, bow, taper and edge contour. Electrical measurements usually include: resistivity and lateral resistivity gradient, carrier type and lifetime. Measurements giving information on the perfection of the semiconductor crystal lattice are classified in the structural category and include: testing for stacking faults, and dislocations. Routine chemical measurements are limited to the measurement of dissolved oxygen and carbon by Fourier transform infrared spectroscopy (FT-IR). Finished wafers are individually marked for the purpose of identification and traceability. Packaging helps protect the finished wafers from contamination during shipping and storage.

Industry standards defining in detail how quality control measurements are to be made and determining the acceptable ranges for measured values have been developed by the American Society of Testing Materials (ASTM) and the Semiconductor Equipment and Materials Institute (SEMI).

4.1.12 Bibliography

- A. C. Bonora, *Silicon Wafer Process Technology: Slicing, Etching, Polishing*, Semiconductor Silicon 1977, Electrochem. Soc., Pennington, NJ (1977).
- L. D. Dyer, in *Proceeding of the low-cost solar array wafering workshop 1981*, DoE-JPL-21012-66, Jet Propulsion Lab., Pasadena CA (1982).
- J. C. Dyment and G. A. Rozgonyi, *J. Electrochem. Soc.*, 1971, **118**, 1346.
- H. Gerischer and W. Mindt, *Electrochem. Acta*, 1968, **13**, 1329.
- P. D. Green, *Solid State Electron.*, 1976, **19**, 815.
- C. A. Harper and R. M. Sompson, *Electronic Materials & Processing Handbook*, McGraw Hill, New York, 2nd Edition.
- S. Iida and K. Ito, *J. Electrochem. Soc.*, 1971, **118**, 768.
- W. Kern, *J. Electrochem. Soc.*, 1990, **137**, 1887.
- Y. Mori and N. Watanabe, *J. Electrochem. Soc.*, 1978, **125**, 1510.

- D. L. Partin, A. G. Milnes, and L. F. Vassamillet, *J. Electrochem. Soc.*, 1979, **126**, 1581.
- D. W. Shaw, *J. Electrochem. Soc.*, 1966, **113**, 958.
- F. Snimura, *Semiconductor Silicon Crystal Technology*, Academic Press, New York (1989).
- D. R. Turner, *J. Electrochem. Soc.*, 1960, **107**, 810.

4.2 Doping[2]

Starting with a prepared, polished wafer then how do we get an integrated circuit? We will focus on the CMOS process, described in the last chapter. Let's assume we have wafer which was doped during growth so that it has a background concentration of acceptors in it (i.e. it is p-type). Referring back to CMOS Logic[3], you can see that the first thing we need to build is a n-tank or moat. In order to do this, we need some way in which to introduce additional impurities into the semiconductor. There are several ways to do this, but current technology relies almost exclusively on a technique called **ion implantation**. A diagram of an ion-implanter is shown in the figure in the previous section[4]. An ion implanter uses a dopant source gas, ionizes it, and drives the ions into the wafer. The dopant gas is ionized and the resultant charged ions are accelerated through a magnetic field, where they are mass-analyzed. The vertical magnetic field causes the beam of ions to spread out, according to their mass. A thin aperture selects the ions of interest, and lets them pass, blocking all the others. This makes sure we are only implanting the ion we want, and in fact, even selects for the proper isotope! The ionized atoms are then accelerated through several tens to hundreds of kV, and then deflected by an electric field, much like in an oscilloscope CRT. In fact, most of the time the ion beam is "rastered" across the surface of the silicon wafer. The ions strike the silicon wafer and pass into its interior. A measurement of the current flow in the system and its integral, is a measure of how much dopant was deposited into the wafer. This is usually given in terms of the number of dopant $\frac{\text{atoms}}{\text{cm}^2}$ to which the wafer has been exposed.

After the atoms enter the silicon, they interact with the lattice, creating defects, and slowing down until finally they stop. Typical atomic distributions, as a function of implant voltage are show in Figure 4.12 for implantation into amorphous silicon. When implantation is done on single crystal material, channeling, the improved mobility of an ion down the "hallway" of a given lattice direction, can skew the impurity distribution significantly. Just slight changes of less than a degree can make big differences in how the impurity atoms are finally distributed in the wafer. Usually, the operator of the implant machine purposely tilts the wafer a few degrees off normal to the beam in order to arrive at more reproducible results.

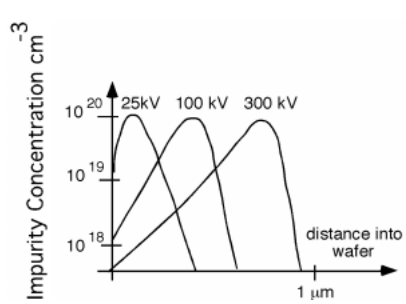


Figure 4.12: Implant distribution with acceleration energy

[2] This content is available online at <http://cnx.org/content/m11364/1.2/>.
[3] "CMOS Logic", Figure 3 <http://cnx.org/content/m11359/latest/#fig46>
[4] "Silicon Growth", Figure 1 <http://cnx.org/content/m11363/latest/#fig05>

As you might expect, shooting 100 kV ions at a silicon wafer probably does quite a bit of damage to the crystal structure. Not only that, but just having, say boron, in your wafer does not mean you are going to have holes. For the boron to become "electrically active" - that is to act as an acceptor - it has to reside on a silicon lattice site. Even if the boron atom does, somehow, end up on an actual lattice site when it stops crashing around in the wafer, the many defects which have been created will act as deep traps. Thus, the hole which is formed will probably be caught at a trap site and will not be able to contribute to electrical conductivity in the wafer anyway. How can we fix this situation? If we carefully heat up the wafer, we can cause the atoms in the crystal to shake around, and if we do it right, they all get back where they belong. Not only that, but the newly added impurities end up on lattice sites as well! This step is called **annealing** and it does just what it is supposed to. Typical temperatures and times for such an anneal are 500 to 1000 °C for 10 to 30 minutes.

Something else occurs during the anneal step however. We have just added, by our implantation step, impurities with a fairly tight distribution as shown in Figure 4.12. There is an obvious gradient in impurity distribution, and if there is a gradient, than things may start moving around by diffusion, especially at elevated temperatures.

4.3 Applications for Silica Thin Films[5]

4.3.1 Introduction

While the physical properties of silica make it suitable for use in protective and optical coating applications, the biggest application of insulating SiO_2 thin films is undoubtedly in semiconductor devices, in which the insulator performs a number of specific tasks, including: surface passivation, field effect transistor (FET) gate layer, isolation layers, planarization and packaging.

The term insulator generally refers to a material that exhibits low thermal or electrical conductivity; electrically insulating materials are also called dielectrics. It is in regard to the high resistance to the flow of an electric current that SiO_2 thin films are of the greatest commercial importance. The dielectric constant (ε) is a measure of a dielectric materials ability to store charge, and is characterized by the electrostatic energy stored per unit volume across a unit potential gradient. The magnitude of ε is an indication of the degree of polarization or charge displacement within a material. The dielectric constant for air is 1, and for ionic solids is generally in the range of 5 - 10. Dielectric constants are defined as the ratio of the material's capacitance to that of air, i.e., (4.10). The dielectric constant for silicon dioxide ranges from 3.9 to 4.9, for thermally and plasma CVD grown films, respectively.

$$\varepsilon = C_{material}/C_{air} \tag{4.10}$$

An insulating layer is a film or deposited layer of dielectric material separating or covering conductive layers. Ideally, in these application an insulating material should have a surface resistivity of greater than 10^{13} Ω/cm^2 or a volume resistivity of greater than 10^{11} Ω.cm. However, for some applications, lower values are acceptable; an electrical insulator is generally accepted to have a resistivity greater than 10^5 Ω.cm. CVD SiO_2 thin films have a resistivity of 10^6 - 10^{16} Ω.cm, depending on the film growth method.

As a consequence of its dielectric properties SiO_2, and related silicas, are used for isolating conducting layers, to facilitate the diffusion of dopants from doped oxides, as diffusion and ion implantation masks, capping doped films to prevent loss of dopant, for gettering impurities, for protection against moisture and oxidation, and for electronic passivation. Of the many methods used for the deposition of thin films, chemical vapor deposition (CVD) is most often used for semiconductor processing. In order to appreciate the unique problems associated with the CVD of insulating SiO_2 thin films it is worth first reviewing some of their applications. Summarized below are three areas of greatest importance to the fabrication of contemporary semiconductor devices: isolation and gate insulation, passivation, and planarization.

[5]This content is available online at <http://cnx.org/content/m24883/1.5/>.

4.3.2 Device isolation and gate insulation

A microcircuit may be described as a collection of devices each consisting of "an assembly of active and passive components, interconnected within a monolithic block of semiconducting material". Each device is required to be isolated from adjacent devices in order to allow for maximum efficiency of the overall circuit. Furthermore within a device, contacts must also be electrically isolated. While there are a number of methods for isolating individual devices within a circuit (reverse-biased junctions, mesa isolation, use of semi-insulating substrates, and oxide isolation), the isolation of the active components in a single device is almost exclusively accomplished by the deposition of an insulator.

In Figure 4.13 is shown a schematic representation of a silicon MOSFET (metal-oxide-semiconductor field effect transistor). The MOSFET is the basic component of silicon-CMOS (complimentary metal-oxide-semiconductor) circuits which, in turn, form the basis for logic circuits, such as those used in the CPU (central processing unit) of a modern personal computer. It can be seen that the MOSFET is isolated from adjacent devices by a reverse-biased junction (p^+-channel stop) and a thick oxide layer. The gate, source and drain contact are electrically isolated from each other by a thin insulating oxide. A similar scheme is used for the isolation of the collector from both the base and the emitter in bipolar transistor devices.

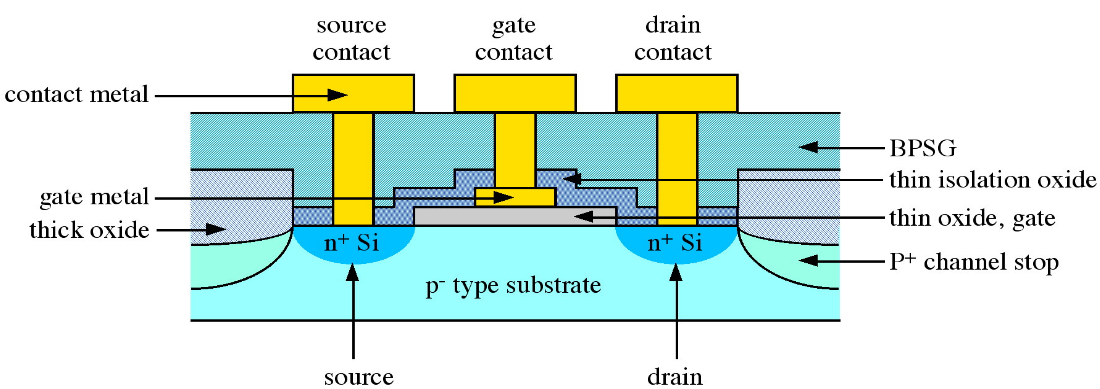


Figure 4.13: Schematic diagrams of a Si-MOSFET (metal-oxide-semiconductor field effect transistor).

As a transistor, a MOSFET has many advantages over alternate designs. The key advantage is low power dissipation resulting from the high impedance of the device. This is a result of the thin insulation layer between the channel (region between source and drain) and the gate contact, see Figure 4.13. The presence of an insulating gate is characteristic of a general class of devices called MISFETs (metal-insulator-semiconductor field effect transistor). MOSFETs are a subset of MISFETs where the insulator is specifically an oxide, e.g., in the case of a silicon MISFET device the insulator is SiO_2, hence MOSFET. It is the fabrication of MOSFET circuits that has allowed silicon technology to dominate digital electronics (logic circuits). However, increases in computing power and speed require a constant reduction in device size and increased complexity in device architecture.

4.3.3 Passivation

Passivation is often defined as a process whereby a film is grown on the surface of a semiconductor to either (a) chemically protect it from the environment, or (b) provide electronic stabilization of the surface.

From the earliest days of solid state electronics it has been recognized that the presence or absence of surface states plays a decisive role in the usefulness of any semiconducting material. On the surface of any solid state material there are sites in which the coordination environment of the atoms is incomplete. These sites, commonly termed "dangling bonds", are the cause of the electronically active states which allow for the recombination of holes and electrons. This recombination occurs at energies below the bulk value, and interferes with the inherent properties of the semiconductor. In order to optimize the properties of a semiconductor device it is desirable to covalently satisfy all these surface bonds, thereby shifting the surface states out of the band gap and into the valence or conduction bands. Electronic passivation may therefore be described as a process which reduces the density of available electronic states present at the surface of a semiconductor, thereby limiting hole and electron recombination possibilities. In the case of silicon both the native oxide and other oxides admirably fulfill these requirements.

Chemical passivation requires a material that inhibits the diffusion of oxygen, water, or other species to the surface of the underlying semiconductor. In addition, the material is ideally hard and resistant to chemical attack. A perfect passivation material would satisfy both electronic and chemical passivation requirements.

4.3.4 Planarization

For the vast majority of electronic devices, the starting point is a substrate consisting of a flat single crystal wafer of semiconducting material. During processing, which includes the growth of both insulating and conducting films, the surface becomes increasingly non-planar. For example, a gate oxide in a typical MOSFET (see Figure 4.13) may be typically 100 - 250 Å thick, while the isolation or field oxide may be 10,000 Å. In order for the successful subsequent deposition of conducting layers (metallization) to occur without breaking metal lines (often due to the difficulty in maintaining step coverage), the surface must be flat and smooth. This process is called planarization, and can be carried out by a technique known as sacrificial etchback. The steps for this process are outlined in Figure 4.14. An abrupt step (Figure 4.14a) is coated with a conformal layer of a low melting dielectric, e.g., borophosphorosilicate glass, BPSG (Figure 4.14b), and subsequently a sacrificial organic resin (Figure 4.14c). The sample is then plasma etched such that the resin and dielectric are removed at the same rate. Since the plasma etch follows the contour of the organic resin, a smooth surface is left behind (Figure 4.14d). The planarization process thus reduces step height differentials significantly. In addition regions or valleys between individual metallization elements (vias) can be completely filled allowing for a route to producing uniformly flat surfaces, e.g., the BPSG film shown in Figure 4.13.

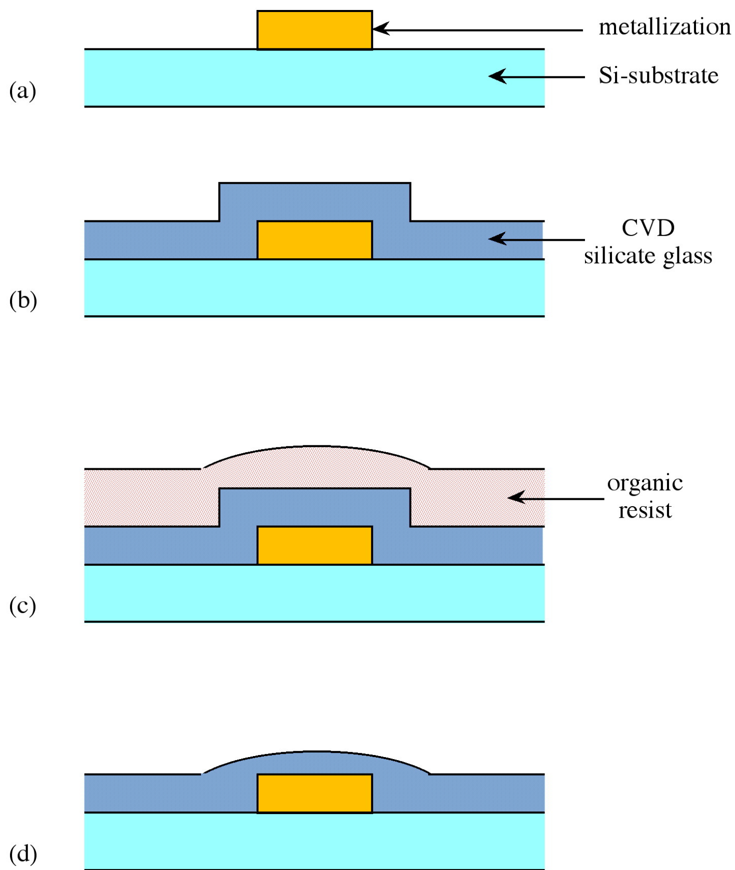


Figure 4.14: Schematic representation of the planarization process. A metallization feature (a) is CVD covered with silicate glass (b), and subsequently coated with an organic resin (c). After etching the resist a smooth silicate surface is produced (d).

The processes of planarization is vital for the development of multilevel structures in VLSI circuits. To minimize interconnection resistance and conserve chip area, multilevel metallization schemes are being developed in which the interconnects run in 3-dimensions.

4.3.5 Bibliography

- J. L. Vossen and W. Kern, *Phys. Today*, 1980, **33**, 26.
- S. K. Ghandhi, *VLSI Fabrication Principles, Silicon and Gallium Arsenide*, Wiley, Chichester, 2nd Ed. (1994).
- S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edition, John Wiley & Sons, New York (1981).
- W. E. Beadle, J. C. C. Tsai, R. D. Plummer, *Quick Reference Manual for Silicon Integrated Cuircuit Technology*, Wiley, Chichester (1985).
- A. C. Adams and C. D. Capio, *J. Electrochem. Soc.*, 1981, **128**, 2630.

4.4 Oxidation of Silicon[6]

NOTE: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Andrea Keys.

4.4.1 Introduction

In the fabrication of integrated circuits (ICs), the oxidation of silicon is essential, and the production of superior ICs requires an understanding of the oxidation process and the ability to form oxides of high quality. Silicon dioxide has several uses:

1. Serves as a mask against implant or diffusion of dopant into silicon.
2. Provides surface passivation.
3. Isolates one device from another (dielectric isolation).
4. Acts as a component in MOS structures.
5. Provides electrical isolation of multi-level metallization systems.

Methods for forming oxide layers on silicon have been developed, including thermal oxidation, wet anodization, chemical vapor deposition (CVD), and plasma anodization or oxidation. Generally, CVD is used when putting the oxide layer on top of a metal surface, and thermal oxidation is used when a low-charge density level is required for the interface between the oxide and the silicon surface.

4.4.2 Oxidation of silicon

Silicon's surface has a high affinity for oxygen and thus an oxide layer rapidly forms upon exposure to the atmosphere. The chemical reactions which describe this formation are:

$$Si_{(s)} + O_{2(g)} \rightarrow SiO_{2(s)} \tag{4.11}$$

$$Si_{(s)} + 2\,H_2O_{(g)} \rightarrow SiO_{2(s)} + 2\,H_{2(g)} \tag{4.12}$$

In the first reaction a dry process is utilized involving oxygen gas as the oxygen source and the second reaction describes a wet process which uses steam. The dry process provides a "good" silicon dioxide but is slow and mostly used at the beginning of processing. The wet procedure is problematic in that the purity of the water used cannot be guaranteed to a suitable degree. This problem can be easily solved using a pyrogenic technique which combines hydrogen and oxygen gases to form water vapor of very high purity. Maintaining reagents of high quality is essential to the manufacturing of integrated circuits, and is a concern which plagues each step of this process.

The formation of the oxide layer involves shared valence electrons between silicon and oxygen, which allows the silicon surface to rid itself of "dangling" bonds, such as lone pairs and vacant orbitals, Figure 4.15. These vacancies create mid-gap states between the valence and conduction bands, which prevents the desired band gap of the semiconductor. The Si-O bond strength is covalent (strong), and so can be used to achieve the loss of mid-gap states and passivate the surface of the silicon.

[6]This content is available online at <http://cnx.org/content/m24908/1.3/>.

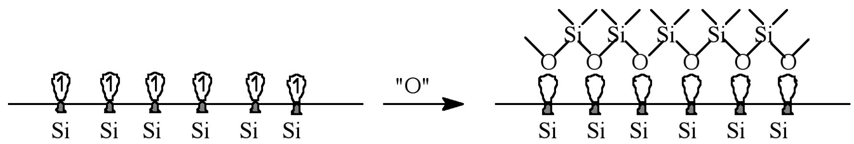


Figure 4.15: Removal of dangling bonds by oxidation of surface.

The oxidation of silicon occurs at the silicon-oxide interface and consists of four steps:

Step 1. Diffusive transport of oxygen across the diffusion layer in the vapor phase adjacent to the silicon oxide-vapor interface.
Step 2. Incorporation of oxygen at the outer surface into the silicon oxide film.
Step 3. Diffusive transport across the silicon oxide film to its interface with the silicon lattice.
Step 4. Reaction of oxygen with silicon at this inner interface.

As the Si-SiO_2 interface moves into the silicon its volume expands, and based upon the densities and molecular weights of Si and SiO_2, 0.44 Å Si is used to obtain 1.0 Å SiO_2.

4.4.2.1 Pre-oxidation cleaning

The first step in oxidizing a surface of silicon is the removal of the native oxide which forms due to exposure to open air. This may seem redundant to remove an oxide only to put on another, but this is necessary since uncertainty exists as to the purity of the oxide which is present. The contamination of the native oxide by both organic and inorganic materials (arising from previous processing steps and handling) must be removed to prevent the degradation of the essential electrical characteristics of the device. A common procedure uses a H_2O-H_2O_2-NH_4OH mixture which removes the organics present, as well as some group I and II metals. Removal of heavy metals can be achieved using a H_2O-H_2O_2-HCl mixture, which complexes with the ions which are formed. After removal of the native oxide, the desired oxide can be grown. This growth is useful because it provides: chemical protection, conditions suitable for lithography, and passivation. The protection prevents unwanted reactions from occurring and the passivation fills vacancies of bonds on the surface not present within the interior of the crystal. Thus the oxidation of the surface of silicon fulfills several functions in one step.

4.4.2.2 Thermal oxidation

The growth of oxides on a silicon surface can be a particularly tedious process, since the growth must be uniform and pure. The thickness wanted usually falls in the range 50 - 500 Å, which can take a long time and must be done on a large scale. This is done by stacking the silicon wafers in a horizontal quartz tube while the oxygen source flows over the wafers, which are situated vertically in a slotted paddle (boat), see Figure 4.16. This procedure is performed at 1 atm pressure, and the temperature ranges from 700 to 1200 °C, being held to within ±1 °C to ensure uniformity. The choice of oxidation technique depends on the thickness and oxide properties required. Oxides that are relatively thin and those that require low charge at the interface are typically grown in dry oxygen. When thick oxides are required (> 0.5 mm) are desired, steam is the source of choice. Steam can be used at wide range of pressures (1 atm to 25 atm), and the higher pressures allow thick oxide growth to be achieved at moderate temperatures in reasonable amounts of time.

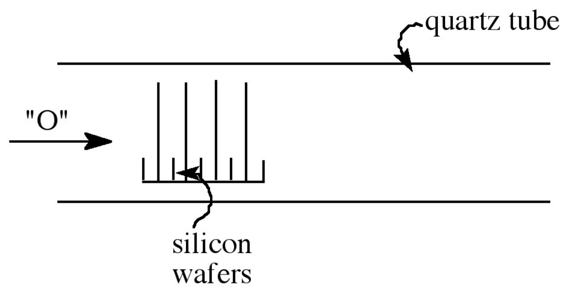


Figure 4.16: Horizontal diffusion tube showing the oxidation of silicon wafers at 1 atm pressure.

The thickness of SiO_2 layers on a Si substrate is readily determined by the color of the film. Table 4.6 provides a guidline for thermal grown oxides.

Film thickness (μm)	**Color**	**Film thickness (μm)**	**Color**
0.05	tan	0.63	violet-red
0.07	brown	0.68	"bluish"
0.10	dark violet to red-violet	0.72	blue-green to gree
0.12	royal blue	0.77	"yellowish"
0.15	light blue to metallic blue	0.80	orange
0.17	metallic to light yellow-green	0.82	salmon
0.20	light gold	0.85	light red-violet
0.22	gold	0.86	violet
0.25	orange to melon	0.87	blue violet
0.27	red-violet	0.89	blue
0.30	blue to violet blue	0.92	blue-green
0.31	blue	0.95	yellow-green
0.32	blue to blue-green	0.97	yellow

continued on next page

0.34	light green	0.99	orange
0.35	green to yellow-green	1.00	carnation pink
0.36	yellow-green	1.02	violet red
0.37	green-yellow	1.05	red-violet
0.39	yellow	1.06	violet
0.41	light orange	1.07	blue-violet
0.42	carnation pink	1.10	green
0.44	violet-red	1.11	yellow-green
0.46	red-violet	1.12	green
0.47	violet	1.18	violet
0.48	blue-violet	1.19	red-violet
0.49	blue	1.21	violet-red
0.50	blue green	1.24	carnation pink to salmon
0.52	green	1.25	orange
0.54	yellow-green	1.28	"yellowish"
0.56	green-yellow	1.32	sky blue to green-blue
0.57	"yellowish"	1.40	orange
0.58	light orange to pink	1.46	blue-violet
0.60	carnation pink	1.50	blue

Table 4.6: Color chart for thermally grown SiO_2 films observed under daylight fluorescent lighting.

4.4.2.3 High pressure oxidation

High pressure oxidation is another method of oxidizing the silicon surface which controls the rate of oxidation. This is possible because the rate is proportional to the concentration of the oxide, which in turn is proportional to the partial pressure of the oxidizing species, according to Henry's law, (4.13), where C is the equilibrium concentration of the oxide, H is Henry's law constant, and p_O is the partial pressure of the oxidizing species.

$$C = H_{(pO)} \qquad (4.13)$$

This approach is fast, with a rate of oxidation ranging from 100 to 1000 mm/h, and also occurs at a relatively low temperature. It is a useful process, preventing dopants from being displaced and also forms a low number of defects, which is most useful at the end of processing.

4.4.2.4 Plasma oxidation

Plasma oxidation and anodization of silicon is readily accomplished by the use of activated oxygen as the oxidizing species. The highly reactive oxygen is formed within an electrical discharge or plasma. The oxidation is carried out in a low pressure (0.05 - 0.5 Torr) chamber, and the the plasma is produced either by a DC electron source or a high-frequency discharge. In simple plasma oxidation the sample (i.e., the silicon wafer) is held at ground potential. In contrast, aniodization systems usually have a DC bias between the sample and an electrode with the sample biased positively with respect to the cathode. Platinum electrodes are commonly used as the cathodes.

There have been at least 34 different reactions reported to occur in an oxygen plasma, however, the vast majority of these are inconsequential with respect to the formation of active species. Furthermore, many of the potentially active species are sufficiently short lived that it is unlikely that they make a significant contribution. The primary active species within the oxygen plasma are undoubtedly O^- and O^{2+}. Both being produced in near equal quantities, although only the former is relevant to plasma aniodization. While these species may be active with respect to surface oxidation, it is more likely that an electron transfer occurs from the semiconductor surface yields activated oxygen species, which are the actual reactants in the oxidation of the silicon.

The significant advatage of plasma processes is that while the electron temperature of the ionized oxygen gas is in excess of 10,000 K, the thermal temperatures required are significantly lower than required for the high pressure method, i.e., < 600 °C. The advantages of the lower reaction temperatures include: the minimization of dopant diffusion and the impediment of the generation of defects. Despite these advantages there are two primary disadvantages of any plasma based process. First, the high electric fields present during the processes cause damage to the resultant oxide, in particular, a high density of interface traps often result. However, post annealing may improve film quality. Second, the growth rates of plasma oxidation are low, typically 1000 Å/h. This growth rate is increased by about a factor of 10 for plasma aniodization, and further improvements are observed if 1 - 3% chlorine is added to the oxygen source.

4.4.2.5 Masking

A selective mask against the diffusion of dopant atoms at high temperatures can be found in a silicon dioxide layer, which can prove to be very useful in integrated circuit processing. A predeposition of dopant by ion implantation, chemical diffusion, or spin-on techniques typically results in a dopant source at or near the surface of the oxide. During the initial high-temperature step, diffusion in the oxide must be slow enough with respect to diffusion in the silicon that the dopants do not diffuse through the oxide in the masked region and reach the silicon surface. The required thickness may be determined by experimentally measuring, at a particular temperature and time, the oxide thickness necessary to prevent the inversion of a lightly doped silicon substrate of opposite conductivity. To this is then added a safety factor, with typical total values ranging from 0.5 to 0.7 mm. The impurity masking properties result when the oxide is partially converted into a silica impurity oxide "glass" phase, and prevents the impurities from reaching the SiO_2-Si interface.

4.4.3 Bibliography

- M. M. Atalla, in *Properties of Elemental and Compound Semiconductors*, Ed. H. Gatos, Interscience: New York (1960).
- S. K. Ghandhi, *VLSI Fabrication Principles, Silicon and Gallium Arsenide*, Wiley, Chichester, 2nd Ed. (1994).
- S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edition, John Wiley & Sons, New York (1981).
- D. L. Lile, *Solid State Electron.*, 1978, **21**, 1199.
- W. E. Spicer, P. W. Chye, P. R. Skeath, and C. Y. Su, I. Lindau, *J. Vac. Sci. Technol.*, 1979, **16**, 1422.
- V. Q. Ho and T. Sugano, *IEEE Trans. Electron Devices*, 1980, **ED-27**, 1436.
- J. R. Hollanhan and A. T. Bells, *Techniques and Applications of Plasma Chemistry*, Wiley, New York (1974).
- R. P. H. Chang and A. K. Sinha, *Appl. Phys. Lett.*, 1976, **29**, 56.

4.5 Photolithography[7]

NOTE: This module is based upon the Connexions module entitled *Photolithography* by Bill Wilson.

[7] This content is available online at <http://cnx.org/content/m33811/1.1/>.

Actually, implants (especially for moats) are usually done at a sufficiently high energy so that the dopant (phosphorus) is already pretty far into the substrate (often several microns or so), even before the diffusion starts. The anneal/diffusion moves the impurities into the wafer a bit more, and as we shall see also makes the n-region grow larger.

"The n-region"! We have not said a thing about how we make our moat in only certain areas of the wafer. From the description we have so far, is seems we have simply built an n-type layer over the whole surface of the wafer. This would be bad! We need to come up with some kind of "window" to only permit the implanting impurities to enter the silicon wafer where we want them and not elsewhere. We will do this by constructing an implantation "barrier".

To do this, the first thing we do is grow a layer of silicon dioxide over the entire surface of the wafer. We talked about oxide growth when we were discussing MOSFETs but let's go into a little more detail. You can grow oxide in either a dry oxygen atmosphere, or in a an atmosphere which contains water vapor, or steam. In Figure 4.17, we show oxide thickness as a function of time for growth with steam. Dry O_2 does not behave too much differently, the rate is just somewhat slower.

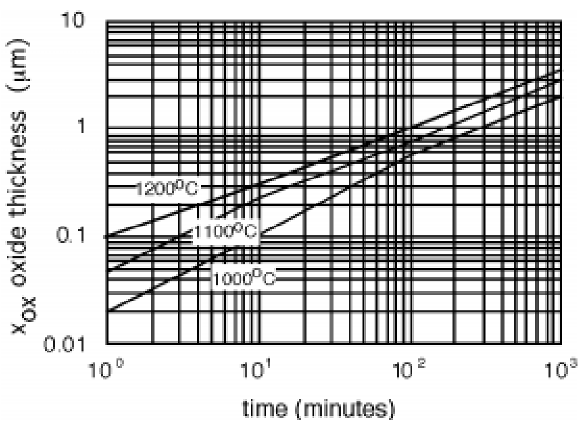


Figure 4.17: A plot of oxide thickness as a function of time.

On top of the oxide, we are now going to deposit yet another material. This is silicon nitride, Si_3N_4 or just plain "nitride" as it is usually called. Silicon nitride is deposited through a method called chemical vapor deposition or "CVD". The usual technique is to react dichlorosilane and ammonia in a hot walled low pressure chemical vapor deposition system (LPCVD). The reaction is:

$$3\,SiH_2Cl_2 + 10\,NH_3 \rightarrow Si_3N_4 + 6\,NH_4Cl + 6\,H_2 \tag{4.14}$$

Silicon nitride is a good barrier for impurities, oxygen and other things which do not want to get into the wafer. Take a look at Figure 4.18 and see what we have so far. A word about scale and dimensions. The silicon wafer is about 250 μm thick (about 0.01") since it has to be strong enough not to break as it is being handled. The two deposited layers are each about 1 μm thick, so they should actually be drawn as lines thinner than the other lines in the figure. This would obviously make the whole idea of a sketch ridiculous, so we will leave things distorted as they are, keeping in mind that the deposited and diffused layers are actually much thinner than the rest of wafer, which really does not do anything except support the active circuits up on top.

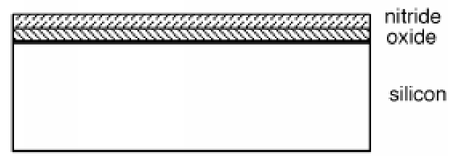


Figure 4.18: Initial wafer configuration.

Now what we want to do is remove part of the nitride, so we can make our n-well, but not put in phosphorous where do not want it. We do this with a processes called *photolithography* and *etching* respectively. First thing we do is coat the wafer with yet another layer of material. This is a liquid called photoresist and it is applied through a process called spin-coating. The wafer is put on a vacuum chuck, and a layer of liquid photoresist is sprayed uncap of the wafer. The chuck is then spun rapidly, getting to several thousand RPM in a small fraction of a second. Centrifugal force causes the resist to spread out uniformly across the wafer surface. The solvent for the photoresist is quite volatile and so the layer of photoresist dries while the wafer is still spinning, resulting in a thin, uniform coating across the wafer Figure 4.19.

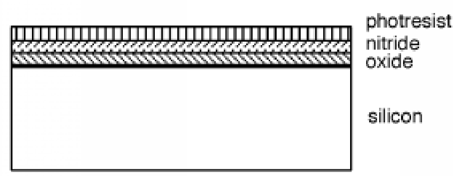


Figure 4.19: After the photoresist is spun on.

The name "photoresist" gives some clue as to what this stuff is. Basically, photoresist is a polymer mixed with some kind of light sensitizing compound. In positive photoresist, wherever light strikes it, the polymer is weakened, and it can be more easily removed with a solvent during the development process. Conversely, negative photoresist is strengthened when it is illuminated with light, and is more resistant to the solvent than is the unilluminated photoresist. Positive resist is so-called because the image of the developed photoresist on the wafer looks just like the mask that was used to create it. Negative photoresist makes an image which is the opposite of what the mask looks like.

We have to come up with some way of selectively illuminating certain portions of the photoresist. Anyone who has ever seen a projector know how we can do this. But, since we want to make small things, not big ones, we will change around our projector so that it makes a smaller image, instead of a bigger one. The instrument that projects the light onto the photoresist on the wafer is called a projection printer or stepper Figure 4.20.

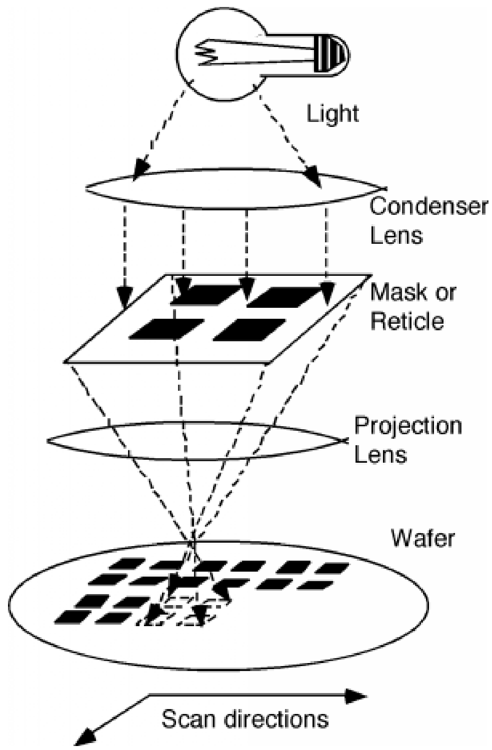


Figure 4.20: A schematic of a stepper configuration.

As shown in Figure 4.20, the stepper consists of several parts. There is a light source (usually a mercury vapor lamp, although ultra-violet excimer lasers are also starting to come into use), a condenser lens to image the light source on the mask or reticle. The mask contains an image of the pattern we are trying the place on the wafer. The projection lens then makes a reduced (usually 5x) image of the mask on the wafer. Because it would be far too costly, if not just plain impossible, to project onto the whole wafer all at once, only a small selected area is printed at one time. Then the wafer is scanned or stepped into a new position, and the image is printed again. If previous patterns have already been formed on the wafer, TV cameras, with artificial intelligence algorithms are used to align the current image with the previously formed features. The stepper moves the whole surface of the wafer under the lens, until the wafer is completely covered with the desired pattern. A stepper is one of the most important pieces of equipment in the whole IC fab however, since it determines what the minimum feature size on the circuit will be.

After exposure, the photoresist is placed in a suitable solvent, and "developed". Suppose for our example the structure shown in Figure 4.21 is what results from the photolithographic step.

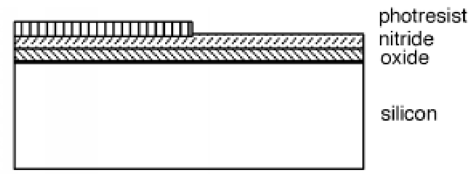


Figure 4.21: After photoresist exposure and development.

The pattern that was used in the photolithographic (PL) step exposed half of our area to light, and so the photoresist (PR) in that region was removed upon development. The wafer is now immersed in a hydrofluoric acid (HF) solution. HF acid etches silicon nitride quite rapidly, but does not etch silicon dioxide nearly as fast, so after the etch we have what we see in Figure 4.22.

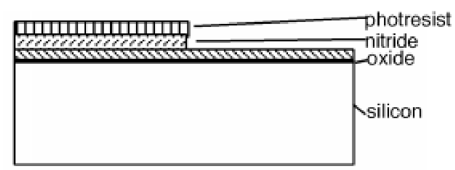


Figure 4.22: After the nitride etch step.

We now take our wafer, put it in the ion implanter and subject it to a "blast" of phosphorus ions Figure 4.23.

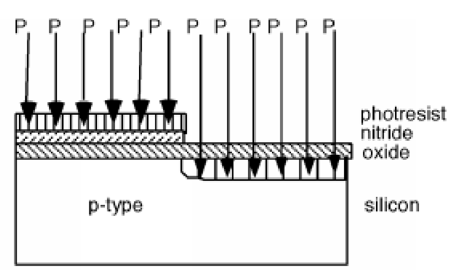


Figure 4.23: Implanting phospohrus.

The ions go right through the oxide layer on the RHS, but stick in the resist/nitride layer on the LHS of our structure.

4.6 Optical Issues in Photolithography[8]

NOTE: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Zane Ball.

4.6.1 Introduction

Photolithography is one of the most important technology in the production of advanced integrated circuits. It is through photolithography that semiconductor surfaces are patterned and the circuits formed. In order to make extremely small features, on the order of the wavelength of the light, advanced optical techniques are used to transfer a pattern from a mask onto the surface. A polymeric film or *resist*, is modified by the light and records the information in a process not dissimilar to ordinary photography.

An illustration of the photolithographic process is shown in Figure 4.24. The process follows the following basic steps:

Step 1. The wafer is spin coated with resist to form a uniform ~1 μm thin film of resist on the surface.
Step 2. The wafer is exposed with ultraviolet light through a mask which contains the desired pattern. In the simplest processes the mask is simply placed over the wafer, but advanced sub-micron technologies require the pattern to imaged through a complex optical system.
Step 3. The photoresist is developed and the irradiated area is washed away (positive resist) or the unirradiated area is washed away (negative resist).
Step 4. Processing (etching, deposition etc.)
Step 5. Remaining resist is stripped.

[8]This content is available online at <http://cnx.org/content/m25448/1.4/>.

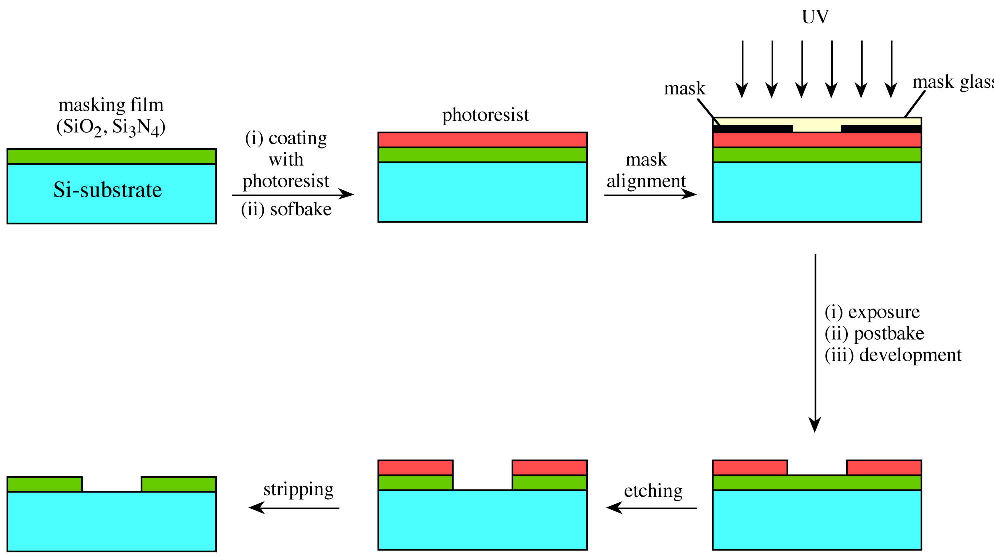


Figure 4.24: Steps in optical printing using photolithography.

In addition to being possibly the most important semiconductor process step, photolithography is also the most expensive technology in semiconductor manufacturing. This expense is the result of two considerations:

1. The optics in photolithography tools are expensive where a single lens can cost a $1 million or more
2. Each chip (often referred to as a "dye") must be exposed individually unlike other semiconductor processes such as CVD where an entire wafer can be processed at a time or oxidation processes where many wafers can be processed simultaneously.

This means that not only are photolithography machines the most expensive of semiconductor processing equipment, but more of them are needed in order to maintain throughput.

4.6.2 Optical issues in photolithography

4.6.2.1 The critical dimension and depth of focus

A semiconductor process technology is often described by a characteristic length known as the critical dimension. The critical dimension (CD) is the smallest feature that needs to be patterned on the surface. The exact definition varies from process to process but is often the channel length of the smallest transistor (typical of a memory chip) or the width of the smallest metal interconnection line (logic chips). This critical dimension is defined by the photolithographic process and is perhaps the most important figure of merit in the manufacture of integrated circuits. Making the critical dimension smaller is the primary focus of improving semiconductor technology for the following reasons:

1. Making the CD smaller dramatically increases the number of devices per unit area and this increase goes with the square of the CD (i.e., a reduction in CD by a factor of 2 generates 4 times the number of devices).

2. Making the CD smaller of a device already in production will make a smaller chip. This means that the number of chips per wafer increases dramatically, and since costs generally scale with the number of wafers and not the number of chips to a wafer, costs are dramatically reduced.
3. Smaller devices are faster.

Therefore, improvements in lithography technology translate directly into better, faster, more complex circuits at lower cost.

Having established the importance of the critical dimension it is important to understand what features of a photolithography system impact. The theory behind projection lithography is very well known, dating from the original analysis of the microscope by Abbe. It is, in fact, the Abbe sine condition that dictates the critical dimension:

$$CD_{Coherent} = 0.82\frac{\lambda}{n\sin(\theta)}$$

$$CD_{Incoherent} = 0.61\frac{\lambda}{n\sin(\theta)} \quad (4.15)$$

where the two expressions refer to the limit of a purely coherent illuminating source and purely incoherent source respectively, and λ is the vacuum wavelength of the illuminating light source, n the index of refraction of the objective lens, and Θ refers to the angle between the axis of the lens and the line from the back focal point to the aperture of the entrance of the lens. The quantity in the denominator, $n\sin(\Theta)$ is referred to as the numerical aperture or NA. As the degree of coherence can be adjusted in a lithography system, the critical dimension is usually written more generally as:

$$CD = k_1\frac{\lambda}{n\sin(\theta)} \quad (4.16)$$

From this equation, we begin to see what can be done to reduce the critical dimension of a lithography system:

1. Change the wavelength of the source.
2. Increase the numerical aperture (NA).
3. Reduce k_1.

Before we discuss how this is accomplished, we must consider one other key quantity, the depth of focus or DOF. The depth of focus is the length along the axis in which a sharp image exists. Naturally a large DOF is desirable for ease of alignment, since the entire dye must with lie within this region. In reality, however, the more meaningful constraint is that the DOF must be thicker than the resist layer so that the entire volume of resist is exposed and can be developed. Also, if the surface morphology of the device dictates that the resist to be exposed is not planar, then the DOF must be large enough so that all features are properly illuminated. Current resists must be 1 μm in thickness in order to have the necessary etch resistance, so this can be considered a minimum value for an acceptable DOF. The depth of focus can also be expressed as a function of numerical aperture and wavelength:

$$DOF = k_2\frac{\lambda}{[n\sin(\theta)]^2} \quad (4.17)$$

If we desire to minimize the critical dimension simply by making optics of large numerical aperture that we will simultaneously reduce the depth of focus and at a much faster rate owing to the dependence on the square of the numerical aperture.

These two quantities, DOF and CD, provide the direction in lithography and semiconductor processing as a whole. For example, a design with an improved surface planarity or a new resist that is effective at smaller thicknesses would allow for a smaller depth of focus which would in turn allow for a larger numerical aperture implying a smaller critical dimension. The resist, the source wavelength, and the optical delivery system all affect the critical dimension and that further refinements require a multifaceted approach to improving lithography systems. What also must be realized is that, as far as the optical system is concerned, virtually all that can be done with conventional optics has been done and that fundamental restraints on k_1 have been reached.

4.6.2.2 Wavefront engineering

One way to get around the fundamental limitations of an imaging system illustrated in (4.15) is through one of a variety of techniques often termed *wavefront engineering*. Here, not only is the amplitude mapped from the object plane to the image plane, but the phase structure of the light going through the mask is manipulated to improve the contrast and allow for effective values of k_1 lower than the theoretical minimum for uniform illumination. The most important example of these techniques is the phase shift mask or PSM. Here the mask consists of two types of areas, those that allow light to pass through unaffected and some regions where the amplitude of the light is unaffected but its phase is shifted. The resulting electric fields will then sum to zero in some places where use of an ordinary mask would have resulted in a positive intensity.

There are many problems with the practical introduction of various phase shifting techniques. Construction of masks with phase shifting elements (usually a thin layer of PMMA) is difficult and expensive. Mask damage, already a key problem in conventional production techniques, becomes an even greater issue as traditional mask repair techniques can no longer be used. Also identifying errors in a mask is made more difficult by the odd design.

4.6.2.3 Interaction with resists

The ultimate resolution of a photolithographic process is not dependent on optics alone, but also on the interaction with the resist. One of the key concerns, particularly as wavelengths of sources become shorter, is the ability of the source light to penetrate the resist film. Many polymers absorb strongly in the UV which can limit the interaction to the surface. In such a case only a thin layer of the polymer is exposed and the pattern may not be fully uncovered during developing. One important property of resist is the presence of *saturable absorption.*. Saturable absorbers are those absorption sites in the polymer that when excited to a higher state remain there for relatively long periods of time and do not continue to absorb into higher states. If only saturable absorption is present in a polymer film, then continued irradiation eventually leads to transparency as all absorption sites will be saturated. This allows light penetration through the resist film with full exposure to the substrate surface.

Full penetration of the film leads to a second problem, multiple reflection interference. This occurs when light which has penetrated the film to the substrate is then reflected back towards the surface. The result is a standing wave interference pattern which causes uneven exposure through the film. The problem becomes more severe as optical limits are approached where feature size is approximately equal to the wavelength of the light source meaning such standing waves are the same size as the irradiated features. In the most advanced lithography techniques such as 248 nm lithography with excimer lasers, a special anti-reflectance coating must be laid down before the resist is deposited. Development of an AR coating that has no adverse effects during the exposure and development process is difficult.

One completely new approach to photolithography resists are top-surface-imaged resists or TSI resists. These processes do not require light penetration through the whole volume of resist. In a TSI resist, a silyl amine is selectively in-diffused from the gas phase into a phenolic polymer in response to the laser irradiation. This diffusion process creates a silyl ether, and development takes place in the form of an oxygen plasma etch, sometimes termed 'dry developing'. Depth of focus limitations are thus avoided as exposure is necessary only at the surface of the resist layer, and the resolution of the etching process determines the final resist profile. Such a technique has tremendous advantages, particularly as source wavelengths become shorter and transparent polymers more rare. Such as resist has a clear optical advantage as well since the image need only be formed at the surface of the resist layer reducing the DOF needed to 100 nm or less, allowing for larger numerical aperture lithography systems with smaller critical dimensions.

4.6.2.4 Light sources

Current photolithography techniques in production utilize ultraviolet lamps as the light source. In the most advanced production facilities, 0.35 μm mercury i-line technology is used. For the next generation of chips such as 64 Mbit DRAMS better performance is necessary and either i-line technology combined with PSM or a new light source is required. Certainly for the 256 Mbit generation using 0.25 μm technology, the i-line source is no longer adequate. The apparent successor is the 248 nm KrF laser, which entered the most advanced production facilities in the late 1990s. KrF technology is often referred to in the literature as Deep UV or DUV lithography. For further shrinkage to 0.18 μm technology, the ArF excimer laser at 193 nm will likely be used with the transition likely to take place in the first few years of the next decade.

At critical dimensions lower than 0.18 - 0.1 μm and below, a whole host of technological problems will need to be overcome in every stage of manufacturing including photolithography. One likely scheme for future lithography is to use X-rays where the wavelength of the light is so much smaller than the feature size such that proximity printing can be used. This is where the mask is placed close to the surface and an X-ray source is scanned across using no optics. Common X-ray sources for such techniques include synchrotron radiation and laser produced plasmas. It has also been widely suggested that the cost of implementing X-ray or other post-optical techniques together with the increased cost of every other manufacturing process step will make improvements beyond 0.1 μm cost prohibitive where benefits in increased circuit speed and density will be dwarfed by massive manufacturing cost. It is noted however that such predictions have been made in the past with regard to other technological barriers.

4.6.3 Bibliography

- M. Born and E. Wolf, *Principles of Optics 6th Edition*, Pergamon Press, New York (1980).
- M. Nakase, *IEICE Trans. Electron.*, 1993, **E76-C**, 26.
- M. Rothschild, A. R. Forte, M. W. Horn, R. R. Kunz, S. C. Palmateer, and J. H. C. Sedlacek, *IEEE J. Selected Topics in Quantum Electronics*, 1995, **1**, 916.

4.7 Composition and Photochemical Mechanisms of Photoresists[9]

NOTE: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Angela Cindy Wei.

4.7.1 Photolithography

In photolithography, a pattern may be transferred onto a photoresist film by exposing the photoresist to light through a mask of the pattern. In the semiconductor industry, the photolithographic procedure includes the following steps as illustrated in Figure 4.25: coating a base material with photoresist, exposing the resist through a mask to light, developing the resist, etching the exposed areas of the base, and stripping the remaining resist off.

[9] This content is available online at <http://cnx.org/content/m25525/1.2/>.

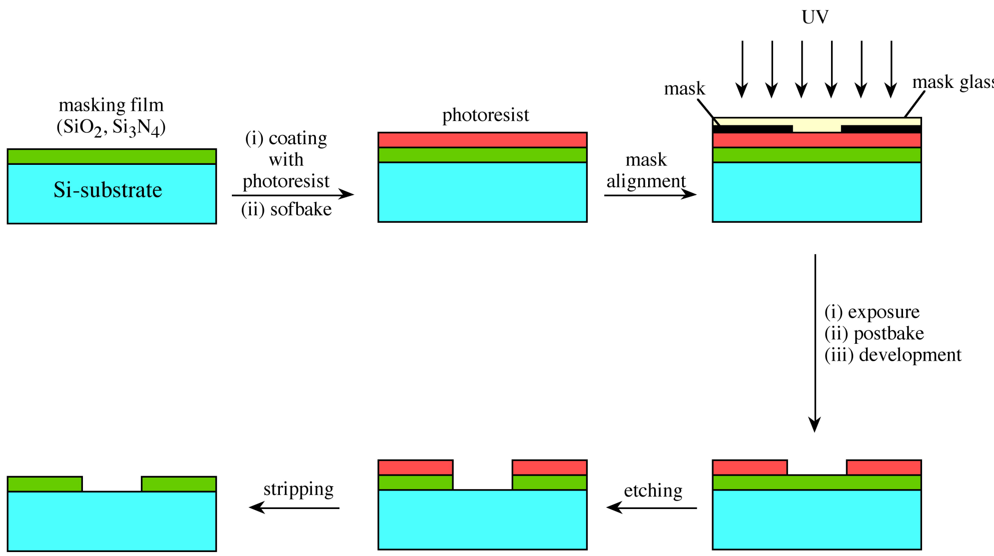


Figure 4.25: Steps in optical printing using photolithography.

Upon exposure to light, the photoresist may become more or less soluble depending on the chemical properties of the particular resist material. The photochemical reactions include chain scission, cross-linking, and the rearrangement of molecules. If the exposed areas of the photoresist become more soluble, then it is a positive resist; conversely, if the exposed resist becomes less soluble, then it is a negative resist. In developing the photoresist, the more soluble material is removed leaving a positive or a negative image of the mask pattern.

4.7.2 Photoresist

Photoresists were initially developed for the printing industry. In the 1920s, the application of photoresists spread to the printed circuit board industry. Photoresists for semiconductor use were first developed in the 1950s; Kodak developed commercial negative photoresists and shortly after, Shipley developed a line of positive resists. Several other companies have entered the market since that time in hopes of manufacturing resist products which meet the increasing demands of the semiconductor industry: narrower line widths, fewer defects, and higher production rates.

4.7.2.1 Photoresist composition

Several functional requirements must be met for a photoresist to be used in the semiconductor industry. Photoresist polymers must be soluble for easy deposition onto a substrate by spin-coating. Good photoresist-substrate adhesion properties are required to minimize undercutting, to maintain edge acuity, and to control the feature sizes. The photoresist must be chemically resistant to whichever etchants are to be used. Sensitivity of the photoresist to a particular light source is essential to the functionality of a photoresist. The speed at which chemical changes occur in a photoresist is its contrast. The contrast of a resist is dependent on the molecular weight distribution of the polymers: a broad molecular weight distribution results in a low contrast resist. High contrast resists produce higher resolution images.

The four basic components of a photoresist are the polymer, the solvent, sensitizers, and other additives. The role of the polymer is to either polymerize or photosolubilize when exposed to light. Solvents allow the photoresist to be applied by spin-coating. The sensitizers control the photochemical reactions and additives may be used to facilitate processing or to enhance material properties. Photochemical changes to polymers are essential to the functionality of a photoresist. Polymers are composed primarily of carbon, hydrogen, and oxygen-based molecules arranged in a repeated pattern. Negative photoresists are based on polyisopreme polymers; negative resist polymers are not chemically bonded to each other, but upon exposure to light, the polymers crosslink, or polymerize. Positive photoresists are formulated from phenol-formaldehyde novolak resins; the positive resist polymers are relatively insoluble, but upon exposure to light, the polymers undergo photosolubilization.

Solvents are required to make the photoresist a liquid, which allows the resist to be spun onto a substrate. The solvents used in negative photoresists are non-polar organic solvents such as toluene, xylene, and halogenated aliphatic hydrocarbons. In positive resists, a variety of organic solvents such as ethyl cellosolve acetate, ethoxyethyl acetate, diglyme, or cyclohexanone may be used.

Photosensitizers are used to control or cause polymer reactions resulting in the photosolubilization or crosslinking of the polymer. The sensitizers may also be used to broaden or narrow the wavelength response of the photoresist. Bisazide sensitizers are used in negative photoresists while positive photoresists utilize diazonaphthoquinones. One measure of photosensitizers is their quantum efficiencies, the fraction of photons which result in photochemical reactions; the quantum efficiency of positive diazonaphthoquinone photoresist sensitizers has been measured to be 0.2 - 0.3 and the quantum efficiency of negative bis-arylazide sensitizers is in the range of 0.5 - 1.0.

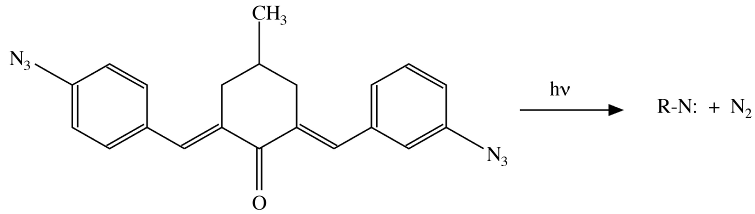
Additives are also introduced into photoresists depending on the specific needs of the application. Additives may be used to increase photon absorption or to control light within the resist film. Adhesion promoters such as hexamethyldisilazane and additives to improve substrate coating are also commonly used.

4.7.2.2 Negative photoresist chemistry

The matrix resin material used in the formulation of these (negative) resists is a synthetic rubber obtained by a Ziegler-Natta polymerization of isoprene which results in the formation of poly(cis-isoprene). Acid-catalyzation of poly(cis-isoprene) produces a partially cyclized polymer material; the cyclized polymer has a higher glass transition temperature, better structural properties, and higher density. On the average, microelectronic resist polyisoprenes contain 1-3 rings per cyclic unit, with 5-20% unreacted isoprene units remaining'. The resultant material is extremely soluble in non-polar, organic solvents including toluene, xylene, and halogenated aliphatic hydrocarbons.

The condensation of para-azidobenzaldehyde with a substituted cyclohexanone produces bis-arylazide sensitizers. To maximize the absorption of a particular light source, the absorbance spectrum of the photoresist may be shifted by making structural modifications to the sensitizers; for example, by using substituted benzaldehydes, the absorption peak may be shifted to longer wavelengths. A typical bisazide-cyclized polyisoprene photoresist formulation may contain 97 parts cyclized polyisoprene to 3 parts bisazide in a (10 wt%) xylene solvent.

All negative photoresists function by cross-linking a chemically reactive polymer via a photosensitive agent that initiates the chemical cross-linking reaction. In the bisazide-cyclized polyisoprene resists, the absorption of photons by the photosensitive bisazide in the photoresist results in an insoluble crosslinked polymer. Upon exposure to light, the bisazide sensitizers decompose into nitrogen and highly reactive chemical intermediates, called nitrenes (4.18). The nitrines react to produce polymer linkages and three-dimensional cross-linked structures that are less soluble in the developer solution.



(4.18)

4.7.2.3 Positive photoresist chemistry

Positive photoresist materials originally developed for the printing industry have found use in the semiconductor industry. The commonly used novolac resins (phenol-formaldehyde copolymer) and (photosensitive) diazoquinone both were products of the printing industry.

The novolak resin is a copolymer of a phenol and formaldehyde (Figure 4.26). Novolak resins are soluble in common organic solvents (including ethyl cellosolve acetate and diglyme) and aqueous base solutions. Commercial resists usually contain meta-cresol resins formed by the acid-catalyzed condensation of meta-cresol and formaldehyde.

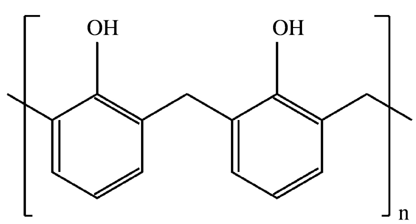
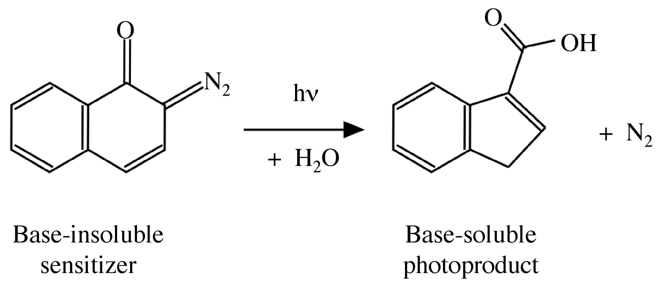


Figure 4.26: Structure of a novolak resin.

The positive photoresist sensitizers are substituted diazonaphthoquinones. The choice of substituents affects the solubility and the absorption characteristics of the sensitizers. Common substituents are aryl sulfonates. The diazoquinones are formed by a reaction of diazonaphthoquinone sulfonyl chloride with an alcohol to form sulfonate ester; the sensitizers are then incorporated into the resist via a carrier or bonded to the resin. The sensitizer acts as a dissolution inhibitor for the novalac resin and is base-insoluble. The positive photoresist is formulated from a novolac resin, a diazonaphthoquinone sensitizer, and additives dissolved in a 20 - 40 wt% organic solvent. In a typical resist, up to 40 wt% of the resist may be the sensitizer.

The photochemical reaction of quinonediazide is illustrated in (4.19). Upon absorption of a photon, the quinonediazide decomposes through Wolff rearrangement, specifically a Sus reaction, and produces gaseous nitrogen as a by-product. In the presence of water, the decomposition product forms an indene carboxylic acid, which is base-soluble. However, the formation of acid may not be the reason for increased solubility; the release of nitrogen gas produces a porous structure through which the developer may readily diffuse, resulting in increased solubility.



(4.19)

4.7.2.4 Image reversal

By introducing an additive to the novolac resins with diazonaphthaquiones sensitizers, the resultant photoresist may be used to form a negative image. A small amount of a basic additive such as monazoline, imidazole, and triethylamine is mixed into a positive novolac resist. Upon exposure to light, the diazonaphthaquiones sensitizer forms an indene carboxylic acid. During the subsequent baking process, the base catalyzes a thermal decarboxylation, resulting in a substituted indene that is insoluble in aqueous base. Then, the resist is flood exposed destroying the dissolution inhibitors remaining in the previously unexposed regions of the resist. The development of the photoresist in aqueous base results in a negative image of the mask.

4.7.3 Comparison of positive and negative photoresists

Into the 1970s, negative photoresist processes dominated. The poor adhesion and the high cost of positive photoresists prevented its widespread use at the time. As device dimensions grew smaller, the advantages of positive photoresists, better resolution and pinhole protection, suited the changing demands of the semiconductor industry and in the 1980s the positive photoresists came into prominence. A comparison of negative and positive photoresists is given in Figure 4.27.

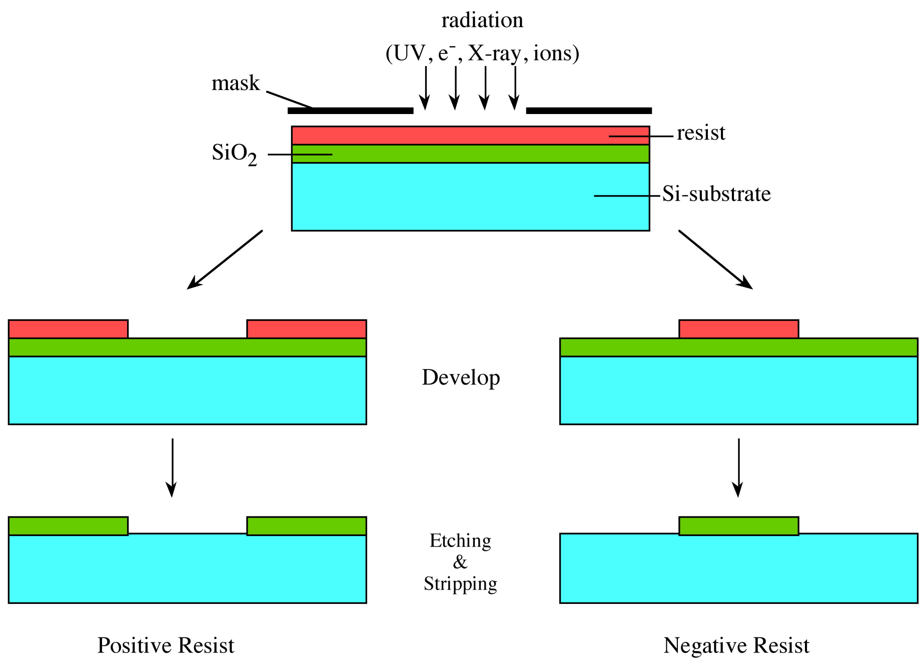


Figure 4.27: A comparison of negative and positive photoresists.

The better resolution of positive resists over negative resists may be attributed to the swelling and image distortion of negative resists during development; this prevents the formation of sharp vertical walls of negative resist. Disadvantages of positive photoresists include a higher cost and lower sensitivity.

Positive photoresists have become the industry choice over negative photoresists. Negative photoresists have much poorer resolution and the positive photoresists exhibit better etch resistance and better thermal stability. As optical masking processes are still preferred in the semiconductor industry, efforts to improve the processes are ongoing. Currently, researchers are studying various forms of chemical amplification to increase the photon absorption of photoresists.

4.7.4 Bibliography

- W.M. Alvino, *Plastics For Electronics*, McGraw-Hill, Inc, New York (1995).
- R. W. Blevins, R. C. Daly, and S. R. Turner, in *Encyclopedia of Polymer Science and Engineering*, Ed. J. I. Krocehwitz, Wiley, New York (1985).
- M. J. Bowden, in *Materials for Microlithography: Radiation-Sensitive Polymers*, Ed. L. F. Thompson, C. G. Willson, and J. M. J. Frechet, American Chemical Society Symposium Series No. 266, Washington, D.C. (1984).
- S. J. Moss and A. Ledwith, *The Chemistry of the Semiconductor Industry*, Blackie & Son Limited, Glasgow (1987).
- E. Reichmanis, F. M. Houlihan, O. Nalamasu, and T. X. Neenan, in *Polymers for Microelectronics*, Ed. L. F. Thompson, C. G. Willson, and S. Tagawa, American Chemical Society Symposium Series, No. 537, Washington, D.C. (1994).
- P. van Zant, *Microchip Fabrication*, 2nd ed., McGraw-Hill Publishing Company, New York (1990).

- C. Grant Willson, in *Introduction to Microlithography*, 2nd ed., Ed. L. F. Thompson, C. G. Willson, M. J. Bowden, American Chemical Society, Washington, D.C. (1983).

4.8 Integrated Circuit Well and Gate Creation[10]

> NOTE: This module is based upon the Connexions module entitled *Integrated Circuit Well and Gate Creation* by Bill Wilson.

We then remove the remaining resist, and perform an activation/anneal/diffusion step, also sometimes called the "drive-in". The purpose of this step is two fold. We want to make the n-tank deep enough so that we can use it for our p-channel MOS, and we want to build up an implant barrier so that we can implant into the p-substrate region only. We introduce oxygen into the reactor during the activation, so that we grow a thicker oxide over the region where we implanted the phosphorus. The nitride layer over the p-substrate on the LHS protects that area from any oxide growth. We then end up with the structure shown in Figure 4.28.

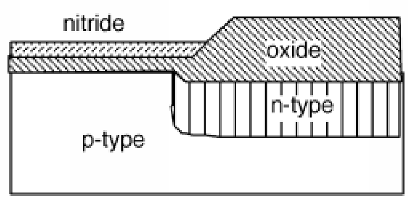


Figure 4.28: After the anneal/drive-in.

Now we strip the remaining nitride. Since the only way we can convert from p to n is to add a donor concentration which is greater than the background acceptor concentration, we had to keep the doping in the substrate fairly light in order to be able to make the n-tank. The lightly doped p-substrate would have too low a threshold voltage for good n-MOS transistor operation, so we will do a V_T adjust implant through the thin oxide on the LHS, with the thick oxide on the RHS blocking the boron from getting into the n-tank. This is shown in Figure 4.29, where boron is implanted into the p-type substrate on the LHS, but is blocked by the thick oxide in the region over the n-well.

[10] This content is available online at <http://cnx.org/content/m33810/1.1/>.

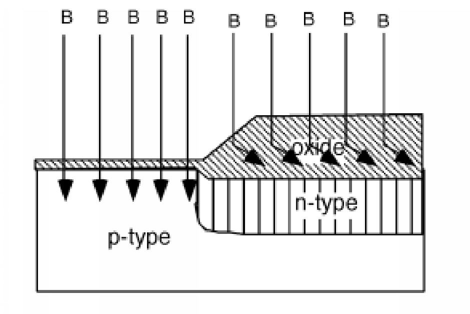


Figure 4.29: V_T adjust implant.

Next, we strip off all the oxide, grow a new thin layer of oxide, and then a layer of nitride Figure 4.30. The oxide layer is grown only because it is bad to grow Si_3N_4 directly on top of silicon, as the different coefficients of thermal expansion between the two materials causes damage to the silicon crystal structure. Also, it turns out to be nearly impossible to remove nitride if it is deposited directly on to silicon.

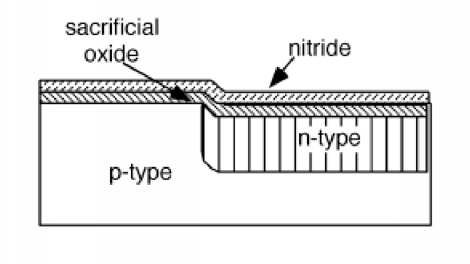


Figure 4.30: Strip of the oxide and grow a new nitride layer.

The nitride is patterned (covered with photoresist, exposed, developed, etched, and removal of photoresist) to make two areas which are called "active" Figure 4.31. The wafer is then subjected to a high-pressure oxidation step which grows a thick oxide wherever the nitride was removed. The nitride is a good barrier for oxygen, so essentially no oxide grows underneath it. The thick oxide is used to isolate individual transistors, and also to make for an insulating layer over which conducting patterns can be run. The thick oxide is called field oxide (or FOX for short) Figure 4.32.

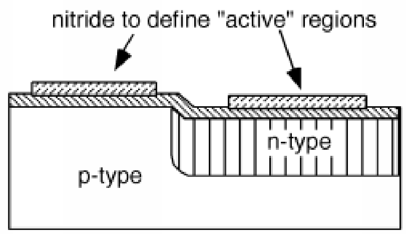


Figure 4.31: Nitride remaining after etching.

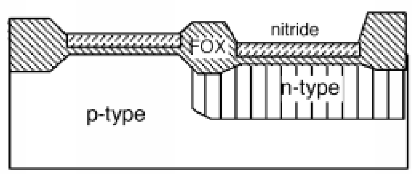


Figure 4.32: After growth of the field oxide (FOX).

Then, the nitride, and some of the oxide are etched off. The oxide is etched enough so that all of the oxide under the nitride regions is removed, which will take a little off the field oxide as well. This is because we now want to grow the gate oxide, which must be very clean and pure Figure 4.33. The oxide under the nitride is sometimes called a *sacrificial oxide*, because it is sacrificed in the name of ultra performance.

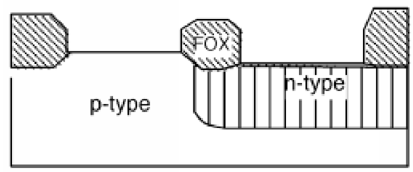


Figure 4.33: Ready to grow gate oxide.

Then the gate oxide is grown, and immediately thereafter, the whole wafer is covered with polysilicon Figure 4.34.

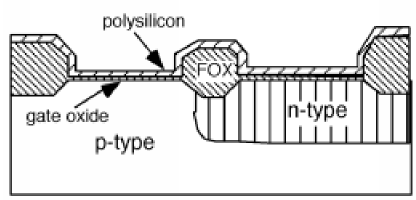


Figure 4.34: Polysilicon deposition over the gate oxide.

The polysilicon is then patterned to form the two regions which will be our gates. The wafer is covered once again with photoresist. The resist is removed over the region that will be the n-channel device, but is left covering the p-channel device. A little area near the edge of the n-tank is also uncovered Figure 4.35. This will allow us to add some additional phosphorus into the n-well, so that we can make a contact there, so that the n-well can be connected to V_{dd}.

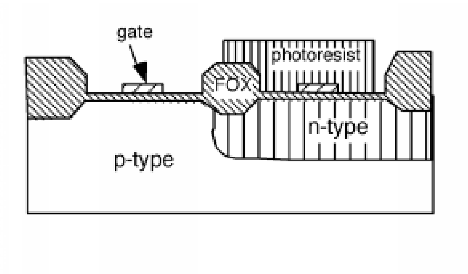


Figure 4.35: Preparing for NMOS channel/drain implant.

Back into the implanter we go, this time exposing the wafer to phosphorus. The poly gate, the FOX and the photoresist all block phosphorus from getting into the wafer, so we make two n-type regions in the p-type substrate, and we have made our n-channel MOS source/drain regions. We also add phosphorous to the V_{dd} contact region in the n-well so as the make sure we get good contact performance there Figure 4.36.

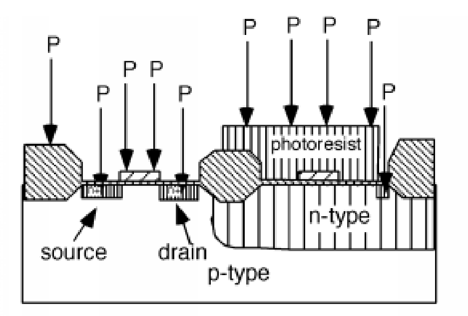


Figure 4.36: Phosphorus S/D implant.

The formation of the source and drain were performed with a *self-aligning technology*. This means that we used the gate structure itself to define where the two inside edges of the source and drain would be for the MOSFET. If we had made the source/drain regions before we defined the gate, and then tried to line the gate up right over the space between them, we might have gotten something that looks like what is shown in Figure 4.37. What's going to be the problem with this transistor? Obviously, if the gate does not extend all the way to both the source and the drain, then the channel will not either, and the transistor will never turn on! We could try making the gate wider, to ensure that it will overlap both active areas, even if it is slightly misaligned, but then you get a lot of extraneous fringing capacitance which will significantly slow down the speed of operation of the transistor Figure 4.38. This is bad! The development of the self-aligned gate technique was one of the big breakthroughs which has propelled us into the VLSI and ULSI era.

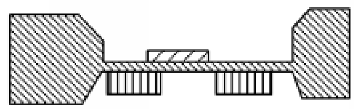


Figure 4.37: A representation of a misaligned gate.

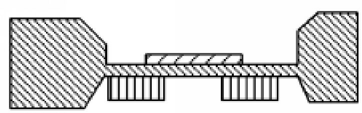


Figure 4.38: A representation of a wide gate.

We pull the wafer out of the implanter, and strip off the photoresist. This is sometimes difficult, because the act of ion implantation can "bake" the photoresist into a very tough film. Sometimes an rf discharge in an O_2 atmosphere is used to "ash" the photoresist, and literally burn it off the wafer! We now apply some more PR, and this time pattern to have the moat area, and a substrate contact exposed, for a boron p^+ implant. This is shown in Figure 4.39.

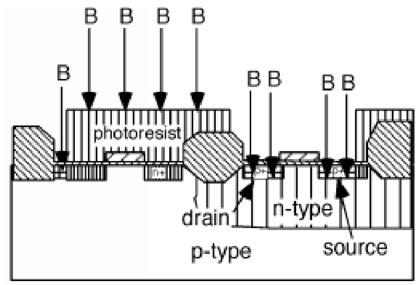


Figure 4.39: Boron p-channel S/D implant.

We are almost done. The next thing we do is remove all the photoresist, and grow one more layer of oxide, which covers everything, as shown in Figure 4.40. We put photoresist over the whole wafer again, and pattern for contact holes to go through the oxide. We will put contacts for the two drains, and for each of the sources, make sure that the holes are big enough to also allow us to connect the source contact to either the p-substrate or the n-moat as is appropriate Figure 4.41.

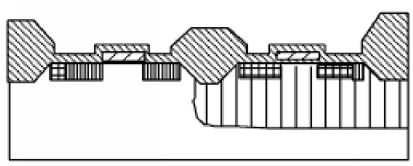


Figure 4.40: Final oxide growth.

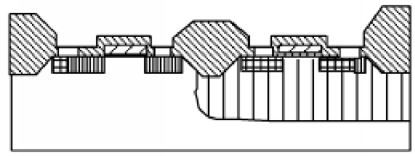


Figure 4.41: After the contact holes are etched.

4.9 Applying Metallization by Sputtering[11]

NOTE: This module is adapted from the Connexions module entitled *Applying Metal/Sputtering* by Bill Wilson.

We now put the wafer in a sputter deposition system. In the sputter system, we coat the entire surface of the wafer with a conductor. An aluminum-silicon alloy is usually used, although other metals are employed as well.

A sputtering system is shown schematically in Figure 4.42. A sputtering system is a vacuum chamber, which after it is pumped out, is re-filled with a low-pressure argon gas. A high voltage ionizes the gas, and creates what is known as the Crookes dark space near the cathode, which in our case, consists of a metal target made out of the metal we want to deposit. Almost all of the potential of the high-voltage supply appears across the dark space. The glow discharge consists of argon ions and electrons which have been stripped off of them. Since there are about equal number of ions and electrons, the net charge density is about zero, and hence by Gauss' law, so is the field.

[11]This content is available online at <http://cnx.org/content/m33800/1.3/>.

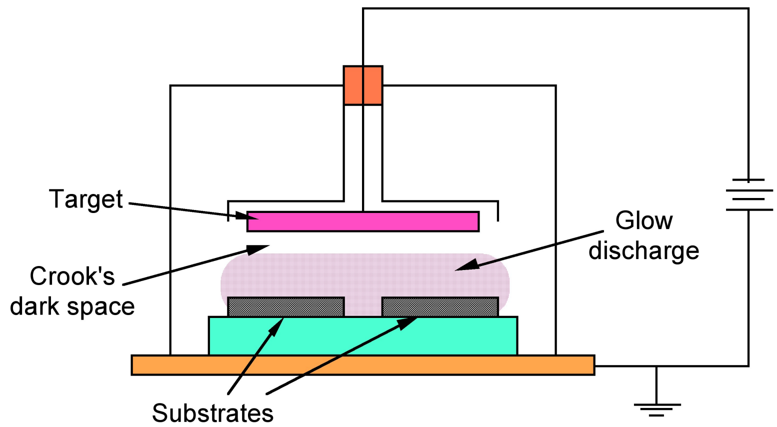


Figure 4.42: A schematic representation of a sputtering apparatus.

The electric field accelerates the argon atoms which slam into the aluminum target. There is an exchange of momentum, and an aluminum atom is ejected from the target (Figure 4.43) and heads to the silicon wafer, where it sticks, and builds up a metal film Figure 4.44.

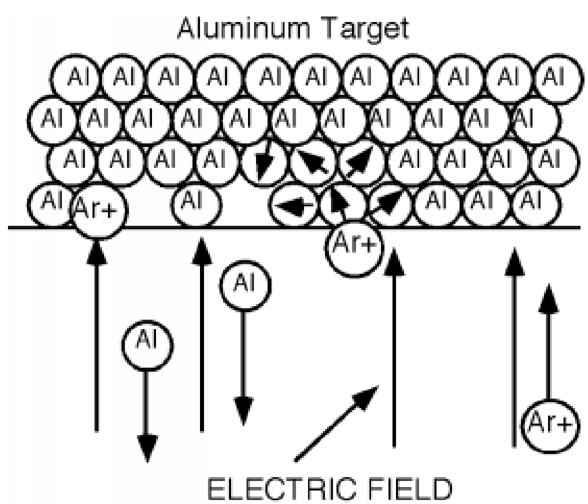


Figure 4.43: The sputtering mechanism.

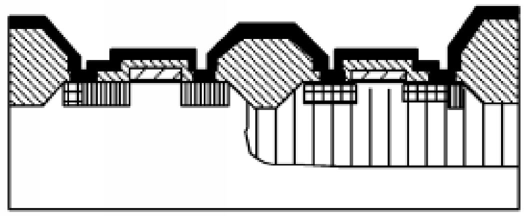


Figure 4.44: Wafer coated with metal.

If you look at Figure 4.44, you will note that we have seemingly done something pretty stupid. We have wired all of the elements of our CMOS inverter together; but all is not lost. We can do one more photolithographic step, and pattern and etch the aluminum, so we only have it where we need it. This is shown in Figure 4.45.

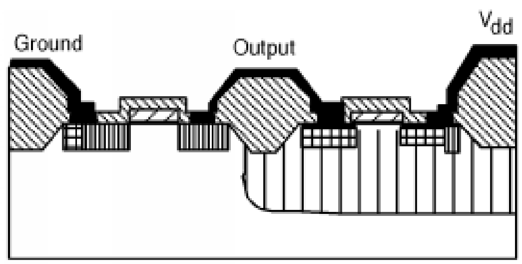


Figure 4.45: After interconnect patterning.