Chapter 2. The substrate and adding material to it

2.1 Introduction

One of the most important techniques employed in microfabrication is the addition of a thin layer of material to an underlying layer. Added layers may form part of the MEMS structure, serve as a mask for etching, or serve as a sacrificial layer.

Additive techniques include those occurring via chemical reaction with an existing layer, as is the case in the oxidation of a silicon **substrate** to form a silicon dioxide layer, as well as those techniques in which a layer is deposited directly on a surface, such as **physical vapor deposition** (PVD) and **chemical vapor deposition** (CVD). The addition of impurities to a material in order to alter its properties, a practice known as **doping**, is also counted among additive techniques, though it does not result in a new physically distinct layer.

Before exploring the various additive techniques themselves we will first examine the nature of the silicon substrate itself.

2.2 The silicon substrate

2.2.1 Silicon growth

In MEMS and microfabrication we start with a thin, flat piece of material onto which (or into which – or both!) we create structures. This thin, flat piece of material is known as the **substrate**, the most common of which in MEMS is crystalline silicon. Silicon's physical and chemical properties make it a versatile material in accomplishing structural, mechanical and electrical tasks in the fabrication of a MEMS.

Silicon in the form of silicon dioxide in sand is the most abundant material on earth. Sand does not make a good substrate, however. Rather, al-

T.M. Adams, R.A. Layton, *Introductory MEMS: Fabrication and Applications*,
DOI 10.1007/978-0-387-09511-0_2, © Springer Science+Business Media, LLC 2010

most all crystalline silicon substrates are formed using a process call the Czochralski method.

In the Czochralski method ultra pure elemental silicon is melted in a quartz crucible in an inert atmosphere to temperatures of 1200-1414ºC. A small "seed" is introduced to the melt so that as is cools and solidifies, it does so as a crystal rather than amorphously or with a granular structure. This is accomplished by slowly drawing and simultaneously cooling the melt while rotating the seed and the crucible the silicon melt in opposite directions. The size of the resulting silicon ingot is determined by carefully controlling temperature as well as the rotational and vertical withdrawal speeds. Once the crystalline silicon is formed, it is cut into disks called wafers. The thicknesses of silicon wafers vary from 200 to 500 μm thick with diameters of 4 to 12 inches. These are typically polished to within 2 μm tolerance on thickness. Figure 2.1 shows single crystal silicon being formed by the Czochralski method.

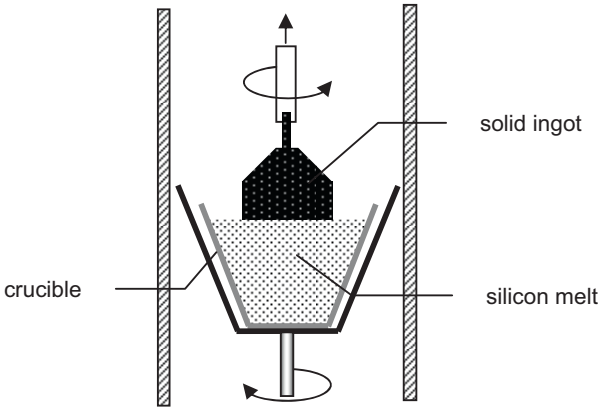


Fig. 2.1. Single crystal silicon formed by the Czochralski method

Sometimes different atmospheres (oxidizing or reducing) are utilized rather than an inert atmosphere to effect crystals with different properties. Furthermore, controlled amounts of impurities are sometimes added during crystal growth in a process known as doping. Typical dopant elements include boron, phosphorous, arsenic and antimony, and can bring about desirable electrical properties in the silicon. We will discuss doping more thoroughly soon.

2.2.2 It's a crystal

Atoms line up in well-ordered patterns in crystalline solids. In such solids, we can think of the atoms as tiny spheres and the different crystal structures as the different ways these spheres are aligned relative to the other spheres. As it turns out, there are only fourteen different possible relative alignments of atoms in crystalline solids. For the semiconductor materials in MEMS, the most important family of crystals are those forming cubic lattices. Figure 2.2 shows the three types of cubic **unit cells,** the building blocks representative of the structure of the entire crystal.

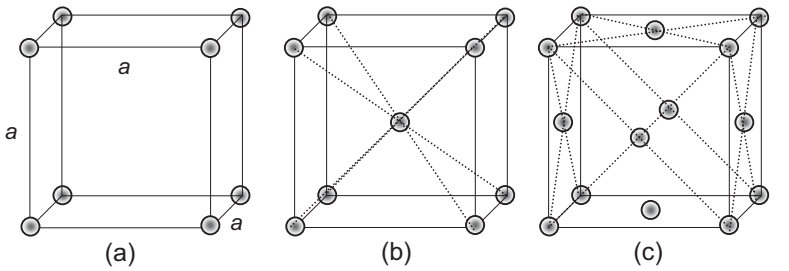


Fig. 2.2 Cubic lattice arrangements: (a) Cubic; (b) Body-centered cubic; (c) Face-centered cubic

Figure 2.2 (a) shows the arrangement of atoms in a simple cubic lattice. The unit cell is simply a cube with an atom positioned at each of the eight corners. Polonium exhibits this structure over a narrow range of temperatures. Note that though there is one atom at each of the eight corners, there is only one atom in this unit cell, as each corner contributes an eighth of its atom to the cell. Figure 2.2 (b) illustrates a body-centered cubic unit cell, which resembles a simple cubic arrangement with an additional atom in the center of the cube. This structure is exhibited by molybdenum, tantalum and tungsten. The cell contains two atoms. Finally, Fig. 2.2 (c) illustrates a face-centered cubic unit cell. This is the cell of most interest for silicon. There are eight corner atoms plus additional atoms centered in the six faces of the cube. This structure is exhibited by copper, gold, nickel, platinum and silver. There are four atoms in this cell. In each unit cell shown, the distance *a* characterizing the side length of the unit cell is called *the lattice constant*.[1]

[1] For non-cubic materials there can be more than one lattice constant, as the sides of the unit cells are not of equal lengths.

Silicon exhibits a special kind of face-centered cubic structure known as the diamond lattice. This lattice structure is a combination of two face-centered cubic unit cells in which one cell has been slid along the main diagonal of the cube one-forth of the distance along the diagonal. Figure 2.3 shows this structure. There are eight atoms in this structure, four from each cell. Each Si atom is surrounded by four nearest neighbors in a tetrahedral configuration with the original Si atom located at the center of the tetrahedron. Since Si has four valence electrons, it shares these electrons with its four nearest neighbors in covalent bonds. Modeling Si atoms as hard spheres, the Si radius is 1.18Å with a lattice constant of 5.43 Å. The distance between nearest neighbors is $d = (3)^{1/2}a/4 = 2.35$Å.

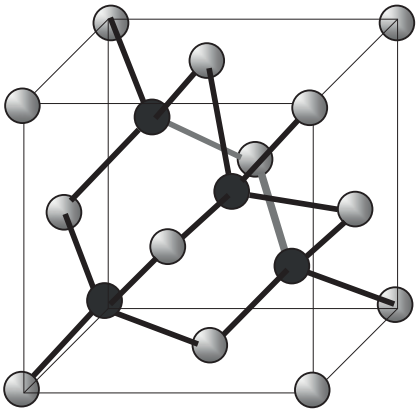


Fig. 2.3 The diamond lattice structure

2.2.3 Miller indices

There are varieties of **crystal planes** defined by the different atoms of a unit cell. The use of **Miller indices** helps us designate particular planes and also directions in crystals. The notation used with Miller indices are the symbols *h*, *k* and *l* with the use of various parentheses and brackets to indicate individual planes, families of planes and so forth. Specifically, the notation (*h k l*) indicates a specific plane; {*h k l*} indicates a family of equivalent planes; [*h k l*] indicates a specific direction in the crystal; and <*h k l*> indicates a family of equivalent directions.

There is a simple three-step method to find the Miller indices of a plane:

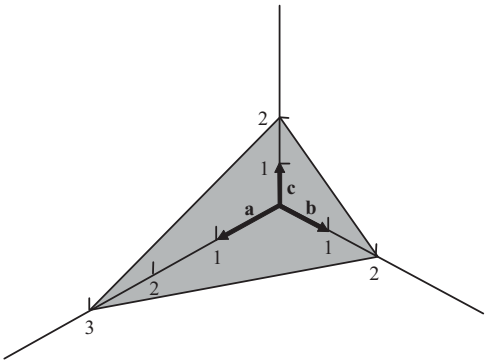
1. Identify where the plane of interest intersects the three axes forming the unit cell. Express this in terms of an integer multiple of the lattice constant for the appropriate axis.
2. Next, take the reciprocal of each quantity. This eliminates infinities.
3. Finally, multiply the set by the least common denominator. Enclose the set with the appropriate brackets. Negative quantities are usually indicated with an over-score above the number.

Example 2.1 illustrates this method.

Example 2.1

Finding the Miller indices of a plane

Find the Miller indices for the plane shown in the figure.



Solution

The plane intersects the axes at 3a, 2b and 2c.
The reciprocals of the theses numbers are 1/3, 1/2 and 1/2.
Multiplying by the least common denominator of 6 gives 2, 3, 3.

Hence the Miller indices of this plane are (2 3 3). ◄

For cubic crystals the Miller indices represent a direction vector perpendicular to a plane with integer components. That is, the Miller indices of a direction are also the Miller indices of the plane normal to it:

$$[h\ k\ l] \perp (h\ k\ l).$$

(This is not necessarily the case with non-cubic crystals.) Figure 2.4 shows the three most important planes for a cubic crystal and the corresponding Miller indices.

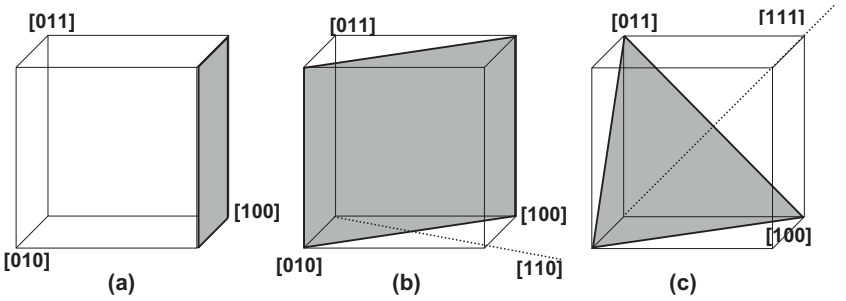


Fig. 2.4. Miller indices of important planes in a cubic crystal (a) (100); (b) (110); (c) (111)

Silicon wafers are classified in part by the orientation of their various crystal planes with relation to the surface plane of the wafer itself. In what is called a (100) wafer, for example, the plane of the wafer corresponds a {100} crystal plane. Likewise, the plane of a (111) wafer coincides with the {111} plane. The addition of straight edges, or flats to the otherwise circular wafers helps us identify the crystalline orientation of the wafers. These flats also tell us if and/or whether the wafers are p-type or n-type. (P-type and n-type refer to wafer's doping, a process affecting the wafer's semiconductor properties. We will explore doping soon.) A few examples are given in Fig. 2.5. Figure 2.6 shows the relative orientations of the three important cubic directions for a {100} wafer.

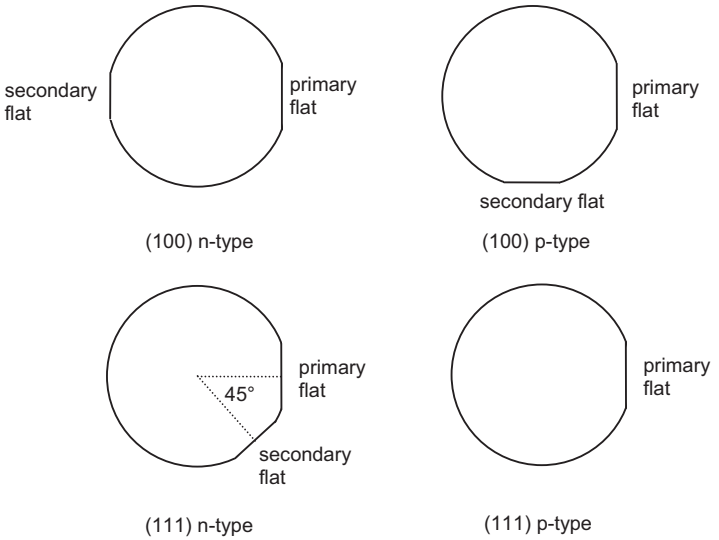


Fig. 2.5. Wafer flats are used to identify wafer crystalline orientation and doping.

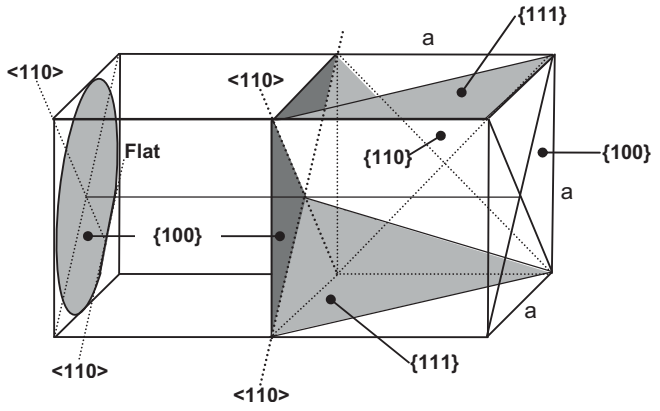


Fig. 2.6. Orientations of various crystal directions and planes in a (100) wafer (Adapted from Peeters, 1994)

2.2.4 It's a semiconductor

In metals there are large numbers of weakly bound electrons that move around freely when an electric field is applied. This migration of electrons is the mechanism by which electric current is conducted in metals, and metals are appropriately called electrical **conductors**. Other materials by contrast have valence electrons that are tightly bound to their atoms, and therefore don't move much when an electric field is applied. Such materials are known as **insulators**. The group IV elements of the periodic table have electric properties somewhere in between conductors and insulators, and are known as **semiconductors**. Silicon is the most widely used semiconductor material. Its various semiconductor properties come in handy in the MEMS fabrication process and also in the operating principles of many MEMS devices.

Figure 2.7 shows the relative electron energy bands in conductors, insulators and semiconductors. In conductors there are large numbers of electrons in the energy level called the conduction band. In insulators the conduction band is empty, and a large energy gap exists between the valence band and the conduction band energy levels, making it difficult for electrons in the valence band to become conduction electrons. Semiconductor materials have only a small energy gap in between the valence band and the otherwise empty conduction band so that when a voltage is applied some of the valence electrons can make the jump to the conduction band, becoming charge carriers. The nature of how this jump is made is affected by both temperature and light in semiconductors, which precipitates their use as sensors and optical switching devices in MEMS.

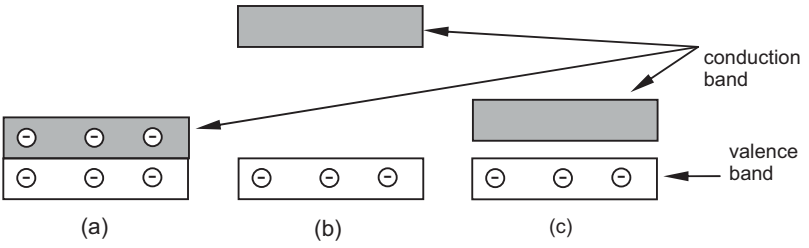


Fig. 2.7. Valence and conduction bands of various materials: (a) Conductor; (b) Insulator; (c) Semiconductor

The **electrical conductivity** or simply conductivity σ of a material is a measure of how easily it conducts electricity. The inverse of conductivity

is **electrical resistivity** ρ. Typical units of conductivity and resistivity are $(\Omega \cdot m)^{-1}$ and $\Omega \cdot m$, respectively. As you might expect, semiconductors have resistivity values somewhere in between those of conductors and insulators, which can be seen in Table 2.1. Resistivity also exhibits a temperature dependence which is exploited in some MEMS temperature sensors.

Table 2.1. Resistivities of selected materials at 20°C

Material	Resistivity ($\Omega \cdot m$)
Silver	1.59×10^{-8}
Copper	1.72×10^{-8}
Germanium	4.6×10^{-1}
Silicon	6.40×10^{2}
Glass	10^{10} to 10^{14}
Quartz	7.5×10^{17}

Conductivity or resistivity can be used to calculate the electrical resistance for a given chuck of that material. For the geometry shown in Fig. 2.8 the electrical resistance in the direction of the current is given by

$$R = \frac{L}{\sigma A} = \rho \frac{L}{A}. \tag{2.1}$$

The trends of Eq. (2.1) should make intuitive sense. Materials with higher resistivities result in higher resistances, as do longer path lengths for electrical current and/or skinnier cross sections. The electrical resistance of a MEMS structure can therefore be tailored via choice of material and geometry.

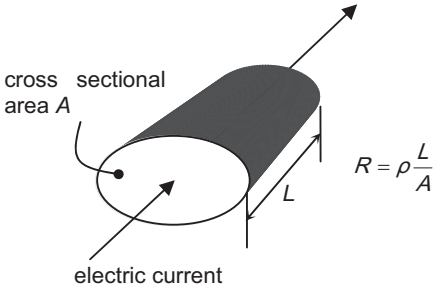


Fig. 2.8. Electrical resistance is determined from resistivity and geometry considerations.

Doping

The properties of semiconductors can be changed significantly by inserting small amounts of group III or group V elements of the periodic table into the crystal lattice. Such a process is called **doping**, and the introduced elements, **dopants**. Careful control of the doping process can also bring about highly localized differences in properties, which has numerous uses both in MEMS device functionality and the microfabrication process itself.

As an example of doping, consider silicon, which has four valence electrons in its valence band. Phosphorous, however, is a group V element and therefore has five valence electrons. If we introduce phosphorous as a dopant into a silicon lattice, one extra electron is floating around and is therefore available as a conduction electron. The dopant material in this case is called a **donor**, as it donates this extra electron. The doped silicon is now an **n-type** semiconductor, indicating that the donated charge carriers are negatively charged. Should the dopant be an element such as boron from group III of the periodic table, however, only three valence band electrons exist in the dopant material. Effectively, a **hole** has been introduced into the lattice, one that is intermittently filled by the more numerous valence electrons of the silicon. This type of dopant is called an **acceptor**, as it accepts electrons from the silicon into its valence band. The electric current in such a semiconductor is the effective movement of these holes from atom to atom. Hence, such semiconductors are called **p-type**, indicating positive charge carriers in the material. By controlling the concentration of the dopant material, the resistivity of silicon can be varied over a range of about 1×10^{-4} to 1×10^{8} $\Omega\cdot$m!

Doping can be achieved in several ways. One way is to build it right into the wafer itself by including the dopant material in the silicon crystal growth process. Such a process results in a uniform distribution of dopant material throughout the wafer, forming what is called the background concentration of the dopant material. Doping already existing wafers is usually achieved by one of two methods, implantation, or thermal diffusion. These methods result in a non-uniform distribution of dopant material in the wafer. Often implantation and/or diffusion are done using an n-type dopant if the wafer is already a p-type, and using a p-type dopant if the wafer is already an n-type. Where the background concentration of dopant in the wafer matches the newly implanted or diffused dopant concentration, a *p-n junction* is formed, the corresponding depth being called the junction depth. P-n junctions play a significant role in microfabrication, often serving as *etch stops*, mechanisms by which a chemical etching process can be halted.

Often implantation and diffusion are done through masks on the wafer surface in order to create p-n junctions at specific locations. Silicon dioxide thin films and photoresist are common masking materials used in this process. Figure 2.9 shows ion implantation occurring though a photoresist mask.

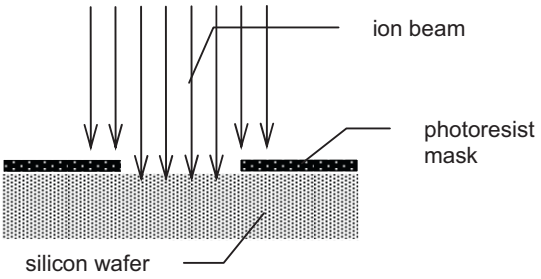


Fig. 2.9. Ion implantation through a photoresist mask

Doping by diffusion

When you pop a helium-filled balloon, the helium doesn't stay in its original location defined by where the balloon was. Rather, it disperses through the surrounding air, eventually filling the entire room at a low concentration. This movement of mass from areas of high concentration to low concentration is appropriately called **mass diffusion**, or simply diffusion. Doping can be achieved by diffusion as well.

When doping by diffusion, the dopant material migrates from regions of high concentration in the wafer to regions of low concentrations in a process called *mass diffusion*. The process is governed by *Fick's law of diffusion*, which in this case simply states that the mass flux of dopant is proportional to the concentration gradient of the dopant material in the wafer, and a material constant characterizing the movement of the dopant material in the wafer material. Flux refers to amount of material moving past a point per unit time and per unit area normal to the flow direction. Flux has dimensions of [amount of substance]/[time][area], typical units for which would be moles/s-m^2 or atoms/s-m^2. Fick's law of diffusion is given by

$$j = -D\frac{\partial C}{\partial x}, \tag{2.2}$$

where j is the mass flux, D is the **diffusion constant** of the dopant in the wafer material , C is the concentration of the dopant in the wafer material, and x is the coordinate direction of interest. In doping, x is usually the direction perpendicular to the surface of the wafer going into the wafer. The negative sign of Eq. (2.2) indicates that the flow of dopant material is in the direction of decreasing concentration.

The diffusion constant D is a temperature dependent constant that characterizes the diffusion of one material in another.[2] It has dimensions of length squared divided by time. For general purposes the constant is well calculated by

$$D = D_0 e^{-\frac{E_a}{k_b T}}, \tag{2.3}$$

where D_o is the **frequency factor**, a parameter related to vibrations of the atoms in a lattice, k_b is Boltzmann's constant (1.38×10^{-23} J/K) and E_a is the **activation energy**, the minimum energy a diffusing atom must overcome in order to migrate. Table 2.2 gives of D_0 and E_a for the diffusion of boron and phosphorus in silicon.

Table 2.2. Frequency factor and activation energy for diffusion of dopants in silicon

Material	D_0 [cm^2/s]	E_a [eV]
Boron	0.76	3.46
Phosphorus	3.85	3.66

[2] The *form* of Ficks's law of diffusion also applies to the transfer of heat and momentum within a material. If you consider only one-dimensional fluxes, heat flux and viscous stress in a flowing fluid are given by $q = -\kappa \cdot dT/dx$ and $\tau = \eta \cdot dV/dx$, respectively. The thermal conductivity κ and the viscosity η play the same role in the transfer of heat and momentum as does the diffusion constant in mass transfer. One difference, however, is that D not only depends on the diffusing species, but also the substance through which it diffuses.

We see, then, that mass travels in the direction of decreasing concentration, heat flows in the direction of decreasing temperature and momentum travels in the direction of decreasing velocity. The three areas are therefore sometimes collectively referred to as **transport phenomena**. (The interpretation of force as a momentum transport, however, is not as common. Thus, the usual sign convention for stress is opposite of that which would result in a negative sign in the stress equation.)

In doping by diffusion the dopant is first delivered to the surface of the wafer after which it diffuses into the wafer. The resulting distribution of dopant within the wafer is therefore a function of both time and depth from the wafer surface. In order to determine this distribution, one solves an equation representing the idea that mass is conserved as the dopant diffuses within the wafer. The version of conservation of mass that applies at a point within a substance is called the *continuity equation*, deriving its name from the assumption that the material can be well modeled as a continuum; that is, it makes sense to talk about properties having a value at an infinitesimally small point. This assumption is valid for most MEMS devices, though it often breaks down at the scales encountered in nanotechnology where the length scales are on the order of the size of molecules. In any case, the continuity equation in regards to mass diffusion reduces to

$$\frac{\partial j}{\partial x} = -D\frac{\partial^2 C}{\partial x^2} = \frac{\partial C(x,t)}{\partial t}. \tag{2.4}$$

As this partial differential equation is first order in time and a second order in depth, we require one initial condition and two boundary conditions in order to solve it. The initial condition is that the concentration of dopant at any depth is zero at $t = 0$, or $C(x, t = 0) = 0$. There are several possibilities for the boundary conditions, however.

One common set of boundary conditions is that the surface concentration goes to some constant value for $t > 0$ and remains at that constant value for all times after that. The second boundary condition comes from the wafer thickness being significantly larger than the depths to which the dopant diffuses. In equation form theses two boundary conditions are, respectively

$$C(x = 0, t > 0) = C_s \tag{2.5}$$

$$C(x \to \infty, t > 0) = 0 \tag{2.6}$$

The solution to Eq. (2.4) incorporating these boundary conditions yields

$$C(x,t) = C_s erfc\left(\frac{x}{2\sqrt{Dt}}\right), \tag{2.7}$$

where C_s is the surface concentration and $erfc(\lambda)$ is the *complementary error function* given by

$$erfc(\lambda) \equiv \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-\lambda^2} d\lambda \tag{2.8}$$

The integral of the complementary error function does not have a closed form solution, making it a transcendental function. Appendix C gives values of the complementary error function. Many modern calculators include $erfc(\lambda)$ as a built-in feature, as do many software packages.

Figure 2.10 gives the concentration of boron in silicon as a function of depth with time as a parameter due to thermal diffusion at 1050°C assuming a constant surface concentration boundary condition. Also shown in the figure is the characteristic length called the **diffusion length**, given by

$$x_{diff} \approx \sqrt{4Dt} \,. \tag{2.9}$$

The diffusion length gives a rough estimate of how far the dopant has diffused into the substrate material at a given time. As the concentration profiles are asymptotic, no single length gives a true cut-off at which point the dopant concentration is actually zero.

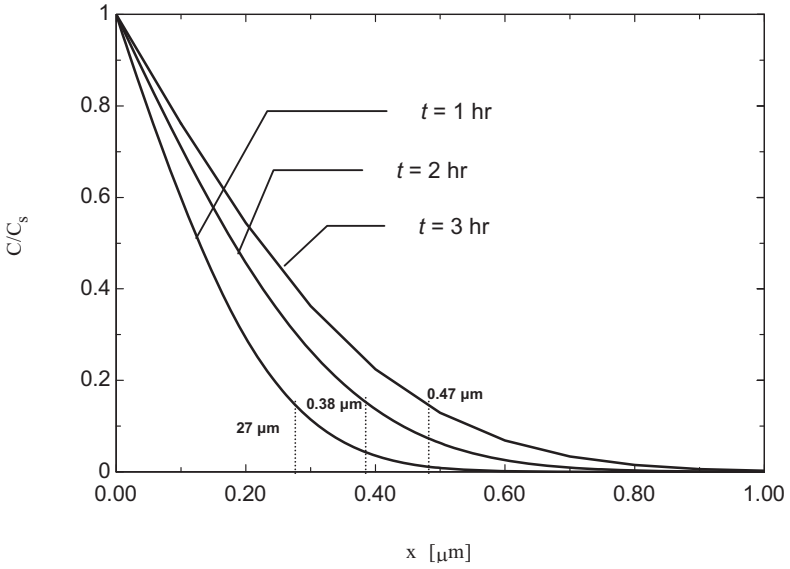


Fig. 2.10. Diffusion of boron in silicon at 1050°C for various times. Diffusion lengths are also shown.

Another quantity of interest in diffusion is the total amount of dopant that has diffused into the substrate per unit area. This can be calculated by integrating Eq. (2.7), the result being

$$Q(t) = \int_0^\infty C_s erfc\left(\frac{x}{2\sqrt{Dt}}\right)dx = \frac{2\sqrt{Dt}}{\sqrt{\pi}}C_s \tag{2.10}$$

where Q is the amount of dopant per area, sometimes called the **ion dose**. Often a more appropriate boundary condition for solving Eq. (2.4) is that the total amount of dopant supplied to the substrate is constant, or that Q is constant. Retaining the same initial condition and the second boundary condition, $C(\infty, t) = 0$, the solution to Eq. (2.4) yields a Gaussian distribution,

$$C(x,t) = \frac{Q}{\sqrt{\pi Dt}}\exp\left(\frac{-x^2}{4Dt}\right) = C_s \exp\left(\frac{-x^2}{4Dt}\right). \tag{2.11}$$

The Gaussian distribution results from the fact that the concentration of ions at the surface is depleted as the diffusion continues. This is reflected in Eq. (2.11) in that the surface concentration is given by $C_s = Q/(\sqrt{(\pi Dt)})$.

Doping by implantation

In implantation a dopant in the form of an ion beam is delivered to the surface of a wafer via a particle accelerator. One may liken the process to throwing rocks into the sand at the beach. The wafer is spun as the accelerator shoots the dopant directly at the surface to ensure more uniform implantation. (Fig. 2.11.) The ions penetrate the surface and are stopped after a short distance, usually within a micron, due to collisions. As a result, the peek concentration of the implanted ions is actually below the surface, with decreasing concentrations both above and below that peek following a normal (Gaussian) distribution. After impact implantation takes only femtoseconds for the ions to stop.

Theoretically this process can occur at room temperature. However, the process is usually followed by a high temperature step (~900°C) in which the ions are "activated," meaning to ensure that they find their way to the spaces in between the atoms in the crystal.[3] Furthermore, the ion bombardment often physically damages the substrate to a degree, and the high temperature anneals the wafer, repairing those defects.

[3] These spaces are called *interstitial sites*.

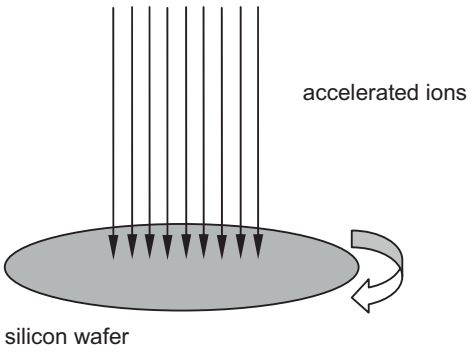


Fig. 2.11. Doping by ion implantation

In ion implantation, the bombardment of ions on the surface results in a Gaussian distribution of ions given by

$$C(x) = C_P \exp\left(-\frac{(x - R_P)^2}{2\Delta R_P^2}\right) \tag{2.12}$$

where C_P is the peak concentration of dopant, R_P is the **projected range** (the depth of peak concentration of dopant in wafer) and ΔR_P is the standard deviation of the distribution. The range is affected by the mass of the dopant, its acceleration energy, and the stopping power of the substrate material. The peak concentration can be found from the total implanted dose from

$$C_P = \frac{Q_i}{\sqrt{2\pi}\Delta R_P} \tag{2.13}$$

where Q_i is the total implanted ion dose. Figure 2.12 gives some typical profiles of the ion implantation of various dopant species.

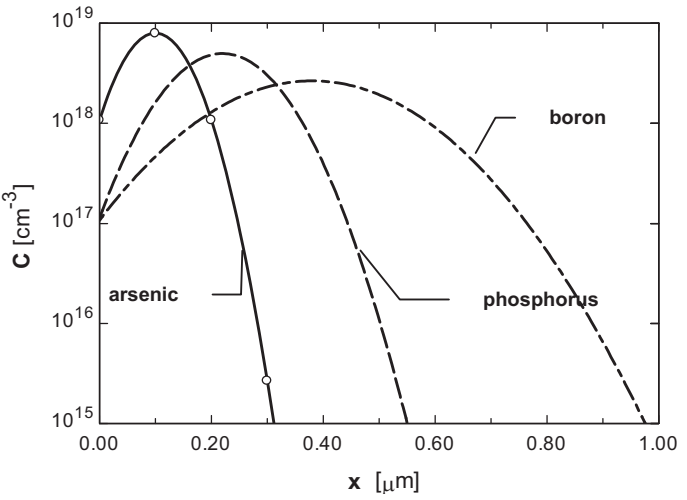


Fig. 2.12 Typical concentration profiles for ion implantation of various dopant species

After ion implantation the dopant material is often thermally diffused even further into the substrate. As such, implantation is often a two step process consisting of the original bombardment of the surface with ions, called *pre-deposition*, and the thermal diffusion step called *drive-in*. The high temperature drive-in step tends to be at a higher temperature than the annealing step that typically follows the implantation process. If the pre-deposition step results in a projected range that is fairly close to the surface, Eq. (2.12) can be used to predict the resulting distribution of dopant after drive-in using $Q = Q_i$.

P-n junctions

One of the major reasons for doping a wafer by thermal diffusion and/or ion implantation is to create a **p-n junction** at specific locations in the wafer. To create a p-n junction, the newly introduced dopant must be of the opposite carrier type than the wafer. That is, if the wafer is already a p-type, the implantation and/or diffusion are done using an n-type dopant, and an n-type wafer would be doped with a p-type dopant. A p-n junction refers to the location at which the diffused or implanted ion concentration matches the existing background concentration of dopant in the wafer. The corresponding depth of the p-n junction called the **junction depth**.

On one side of the junction the wafer behaves as a p-type semiconductor, and on the other it behaves as an n-type. This local variation of semiconductor properties is what gives the junction all its utility. In microelectronic applications p-n junctions are used to create diodes (something like a check valve for electric current) and transistors. In MEMS applications p-n are used to create things such as piezo-resistors, electrical resistors whose resistance changes with applied pressure, enabling them to be used as sensors. P-n junctions can also be used to stop chemical reactions that eat away the silicon substrate, in which case the junction serves as an etch stop. We will learn more about these chemical reactions in Chapter 4 where we discuss bulk micromachining.

For now the important thing to know is how to approximate the location of a p-n junction. Graphically the junction location is at the intersection of the implanted/diffused dopant concentration distribution and the background concentration curves. (Fig. 2.13.) Mathematically the location is found by substituting the background concentration value into the appropriate dopant distribution relation and solving for depth. In the case of a thermally diffused dopant, Eq. (2.11) yields the following for the junction depth:

$$x_j = \sqrt{4Dt \ln\left(\frac{Q}{C_{bg}\sqrt{\pi Dt}}\right)} \tag{2.14}$$

where x_j is the junction depth and C_{bg} is the background concentration of the wafer dopant.

Care must be taken when using the relations in this section, as they are all first order approximations. When second order effects are important, such as diffusion in the lateral direction in a wafer, these relations may not yield sufficiently accurate estimates for design. In such cases numerical modeling schemes using computer software packages are usually employed. Nonetheless, the relations given here are sufficient in many applications and will at least give a “ball park” estimate.

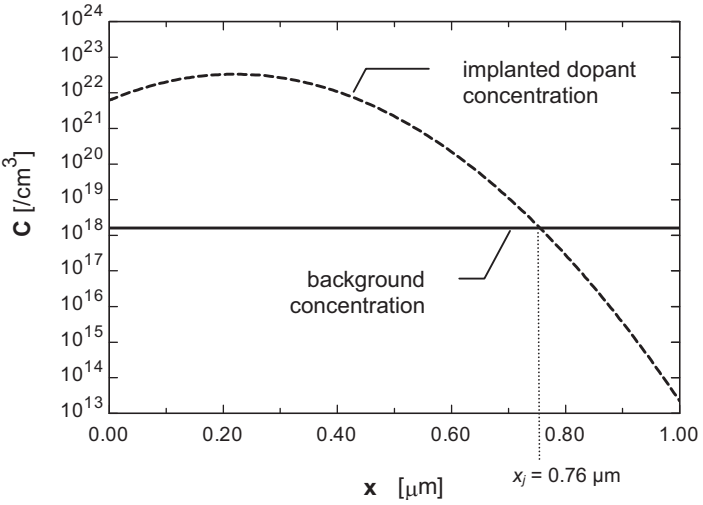


Fig. 2.13. Determination of junction depth

2.3 Additive technique: Oxidation

We have learned much about the nature of the silicon substrate itself and ways of doping it. We next turn our attention to methods of adding physically distinct layers of material on top of the substrate. Sometimes these added materials will serve as masks, at other times as structural layers, and sometimes as sacrificial layers.

2.3.1 Growing an oxide layer

One method of adding a layer of material to the silicon substrate is to "grow" a layer of silicon dioxide (SiO_2) on it. The process is naturally called oxidation. In oxidation, the silicon on the surface of the substrate reacts with oxygen in the environment to form the SiO_2 layer.

The resulting SiO_2 layer is often called an oxide layer for short, or simply oxide. It does a great job as a sacrificial layer or as a hard mask. Oxides can also provide a layer of electrical insulation on the substrate, providing necessary electrical isolation for certain electrical parts in a MEMS.

From simply being in contact with air the silicon substrate will form a thin layer of oxide without any stimulus from the MEMS designer. Such a layer is called a native oxide layer, typical thicknesses being on the order of 20-30 nm. For an oxide layer to be useful as a sacrificial layer or a mask, however, larger thicknesses are typically required. Thus active methods of encouraging the growth of oxide are used.

There are two basic methods used to grow oxide layers: dry oxidation and wet oxidation. Both methods make use of furnaces at elevated temperatures on the order of 800-1200°C. Both methods also allow for careful control of the oxygen flow within the furnace. In dry oxidation, however, only oxygen diluted with nitrogen makes contact with the wafer surface, whereas wet oxidation also includes the presence of water vapor. The chemical reactions for dry and wet oxidation, respectively, are given by

$$\text{(dry) } Si + O_2 \rightarrow SiO_2 \text{ and} \tag{2.15}$$

$$\text{(wet) } Si + 2H_2O \rightarrow SiO_2 + 2H_2 \,. \tag{2.16}$$

Dry oxidation has the advantage of growing a very high quality oxide. The presence of water vapor in wet oxidation has the advantage of greatly increasing the reaction rate, allowing thicker oxides to be grown more quickly. The resulting oxide layers of wet oxidation tend to be of lower quality than the oxides of dry oxidation. Hence, dry oxidation tends to be used when oxides of the highest quality are required, whereas wet oxidation is favored for thick oxide layers. An oxide layer is generally considered to be thick in the 100 nm – 1.5 μm range.

Oxidation is interesting as an additive technique in that the added layer consists both of added material (the oxygen) and material from the original substrate (the silicon). As a result, the thickness added to the substrate is only a fraction of the SiO_2 thickness. In reference to Fig. 2.14, the ratio of the added thickness beyond that of the original substrate (x_{add}) to the oxide thickness itself (x_{ox}) is

$$\frac{x_{add}}{x_{ox}} = 0.54 \,. \tag{2.17}$$

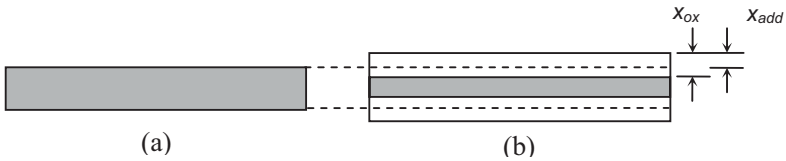


Fig. 2.14 A silicon wafer (a) before oxidation and (b) after oxidation

2.3.2 Oxidation kinetics

As the oxidation reaction continues, oxygen delivered to the surface must diffuse through thicker and thicker layers of silicon dioxide before it can react with the underlying silicon. As such, the time required to grow an additional thickness of oxide becomes longer as the layer thickens. The **Deal-Grove model** of oxidation kinetics is the most widely used model to relate oxide thickness to reaction time. The model comes from the solution to an ordinary differential equation that describes the mass diffusion of oxygen through the oxide layer. The solution is given by

$$x_{ox} = \frac{A}{2}\left\{-1+\sqrt{\frac{4B}{A^2}(t+\tau)+1}\right\}, \tag{2.18}$$

where t is the oxidation time, A and B are temperature dependent constants, and τ is a parameter that depends on the initial oxide thickness, x_i. The value of τ is found by solving Eq. (2.18) with t=0:

$$\tau = \frac{x_i^2}{B} + \frac{A}{B}x_i\,. \tag{2.19}$$

Table 2.3 and Table 2.4 gives the values of A and B for both dry and wet oxidation using the Deal-Grove model for (100) and (111) silicon, respectively. The value of τ in the tables is the recommended value when starting with an otherwise bare wafer, and accounts for the presence of a native oxide layer, assumed to be 25 nm. The Deal-Grove model does not predict oxide growth accurately for thicknesses less than 25 nm, in which case oxide growth occurs much more quickly. Hence, the inclusion of τ can also be viewed as a correction term.

Table 2.3. Suggested Deal-Grove rate constants for oxidation of (100) silicon

Temperature (°C)	A (μm)		B (μm²/hr)		τ (hr)	
	Dry	Wet	Dry	Wet	Dry	Wet
800	0.859	3.662	0.00129	0.0839	17.1	1.10
900	0.423	1.136	0.00402	0.172	2.79	0.169
1000	0.232	0.424	0.0104	0.316	0.616	0.0355
1100	0.139	0.182	0.0236	0.530	0.174	0.0098
1200	0.090	0.088	0.0479	0.828	0.060	0.0034

Table 2.4. Suggested Deal-Grove rate constants for oxidation of (111) silicon

Temperature (°C)	A (μm)		B (μm²/hr)		τ (hr)	
	Dry	Wet	Dry	Wet	Dry	Wet
800	0.512	2.18	0.00129	0.0839	10.4	0.657
900	0.252	0.6761	0.00402	0.172	1.72	0.102
1000	0.138	0.252	0.0104	0.316	0.391	0.0220
1100	0.0830	0.1085	0.0236	0.530	0.114	0.0063
1200	0.0534	0.05236	0.0479	0.828	0.0401	0.0023

Equation (2.18) can be simplified for short time or long time approximations. In the case of a short time, keeping the first two terms in a series expansion of the square root gives

$$x_{ox} \approx \frac{B}{A}(t+\tau). \tag{2.20}$$

The growth rate is approximately linear in this case. As such, the ratio B/A is often referred to as the **linear rate constant**. For very long times, $t >> \tau$, and Eq. (2.18) reduces to

$$x_{ox} \approx \sqrt{B(t+\tau)}\,. \tag{2.21}$$

The oxidation rate is also dependent on the crystalline orientation of the wafer. In situations for which Eq. (2.20) is appropriate, the orientation dependence can be captured by changing the linear rate constant. The ratio of the linear rate constant for (111) silicon to (100) is given by

$$\frac{(B/A)_{(111)}}{(B/A)_{(100)}} \approx 1.68 \tag{2.22}$$

Thus, (111) silicon typically oxidizes 1.7 times faster than does (100) silicon. It is speculated that the increased reaction rate is due to the larger density of Si atoms in the (111) direction.

Example 2.2 illustrates the use of the Deal-Grove model.

Example 2.2

Growth time for wet etching

Find the time required to grow an oxide layer 800 nm thick on a (100) silicon wafer using wet oxidation at 1000°C.

Solution

Assuming a native oxide layer of 25 nm, we can use the suggested constants from Table 2.3. Solving (2.18) for t,

$$t = \frac{A^2}{4B}\left[\left(\frac{2}{A}x_{ox} + 1\right)^2 - 1\right] - \tau$$

$$t = \frac{0.424^2\ \mu\text{m}^2}{4(0.316)\ \mu\text{m}^2/\text{hr}}\left[\left(\frac{2}{0.424\ \mu\text{m}}(0.800\ \mu\text{m}) + 1\right)^2 - 1\right] - 0.0355\ \text{hr}$$

$$t = 3.06\ \text{hr}\ .$$

If we make the long time approximation of Eq. (2.20)

$$x_{ox} \approx \sqrt{B(t+\tau)}$$

Solving for t,

$$t \approx \frac{x_{ox}^{\ 2}}{B} - \tau$$

$$t \approx \frac{(0.800\ \mu\text{m})^2}{0.316\ \mu\text{m}^2/\text{hr}} - 0.0355\ \text{hr}$$

$$t \approx 1.99\ \text{hr}.$$

We see that the long time approximation is not a very good one in this case. And although three hours may seem like forever when waiting by an oxidation furnace, it does not qualify as a long time by the standards of Eq. (2.20). ◄

Oxide layer thicknesses can be measured using optical techniques that measure the spectrum of reflected light from the oxide layer. Due to the

index of refraction of the oxide and its finite thickness, the light reflected from the top and bottom surfaces will be out of phase. Since visible light is in the wavelength range of 0.4-0.7 μm, which is the same order of magnitude as most oxide layers, the constructive and destructive interference of the reflected light will cause it be a different color depending on oxide thickness. Quick estimates of the thickness can be made by visually inspecting the oxide surface and comparing it to one of many available "color charts", such as in Appendix D.

2.4 Additive technique: Physical vapor deposition

Physical vapor deposition (PVD) refers to the vaporization of a purified, solid material and its subsequent condensation onto a substrate in order to form a thin film. Unlike oxidation, no new material is formed via chemical reaction in the PVD process. Rather, the material forming the thin film is physically transferred to the substrate. This material is often referred to as the **source material**.

The mechanism used to vaporize the source in PVD can be supplied by contact heating, by the collision of an electron beam, by the collision of positively charged ions, or even by focusing an intense, pulsed laser beam on the source material. As a result of the added energy, the atoms of the source material enter the gas phase and are transported to the substrate through a reduced pressure environment.

PVD is often called a direct line-of-sight impingement deposition technique. In such a technique we can visualize the source material as if it were being sprayed at the deposition surface much like spray paint. Where the spray "sees" a surface, it will be deposited. If something prohibits the spray from seeing a surface, a concept called *shadowing*, the material will not be deposited there. Sometimes shadowing is a thorn in the side of the microfabricator, but at other times s/he can use shadowing to create structures of predetermined shapes and sizes.

Physical vapor deposition includes a number of different techniques including thermal evaporation (resistive or electron beam), sputtering (DC, RF, magnetron, or reactive), molecular beam epitaxy, laser ablation, ion plating, and cluster beam technology. In this section we will focus on the two most common types of PVD, evaporation and sputtering.

2.4.1 Vacuum fundamentals

Physical vapor deposition must be accomplished in very low pressure environments so that the vaporized atoms encounter very few intermolecular collisions with other gas atoms while traveling towards the substrate. Also, without a vacuum, it is very difficult to create a vapor out of the source material in the first place. What's more, the vacuum helps keep contaminants from being deposited on the substrate. We see, then, that the requirement of a very low pressure environment is one of the most important aspects of PVD. As such, it is a good idea to spend some time learning about vacuums and how to create them.

A vacuum refers to a region of space that is at less than atmospheric pressure, usually significantly less than atmospheric pressure. When dealing with vacuum pressures, the customary unit is the torr. There are 760 torr in a standard atmosphere:

$$1 \text{ atm} = 1.01325 \times 10^5 \text{ Pa} = 760 \text{ torr}. \tag{2.23}$$

Vacuum pressures are typically divided into different regions of increasingly small pressure called **low vacuum** (LV), **high vacuum** (HV) and **ultra high vacuum** (UHV) regions. Table 2.5 gives the pressure ranges of these regions.

Table 2.5. Pressure ranges for various vacuum regions

Region	Pressure (torr)
Atmospheric	760
Low vacuum (LV)	Up to 10^{-3}
High vacuum (HV)	10^{-5} to 10^{-8}
Ultra-high vacuum (UHV)	10^{-9} to 10^{-12}

Vacuum pumps

Vacuums are created using pumps that either transfer the gas from the lower pressure vacuum space to some higher pressure region (gas transfer pump), or pumps that actually capture the gas from the vacuum space (gas capture pump). In general, transfer pumps are used for high gas loads, whereas capture pumps are used to achieve ultra-high vacuums.

A common mechanical pump used to create a vacuum is the rotary sliding vane pump. Such a pump operates by trapping gas between the rotary vanes and the pump body. This gas is then compressed by an eccentrically

mounted rotor. (Fig. 2.15.) The pressure of the compressed gas is higher than atmospheric pressure so that the gas is released into the surroundings.

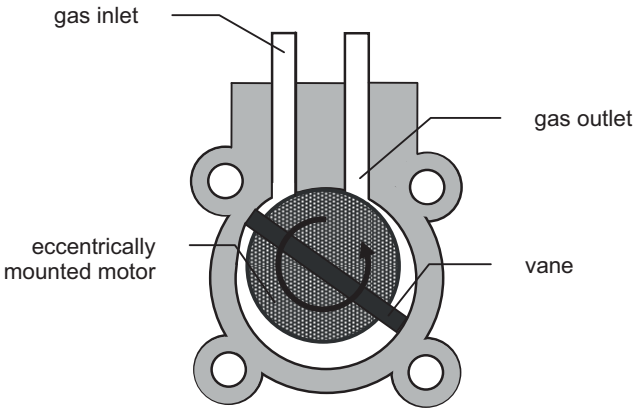


Fig. 2.15. A rotary vane pump

Rotary vane pumps are most commonly used as rough pumps, pumps used to initially lower the pressure of a vacuum chamber and to back (connected to the outlet of) other pumps. In addition to rotary vane pumps, other rough pump types include diaphragm, reciprocating piston, scroll, screw, claw, rotary piston, and rotary lobe pumps. Rough pumps are not capable of creating high vacuums and are only effective from atmospheric pressure down to 10^{-3} torr. Once a vacuum space is in this range, specialized pumps classified as high vacuum pumps are needed to achieve lower pressures.

High vacuum pumps generally work using completely different operating principles than rough pumps, which tend to be mechanical in nature. Common high vacuum pumps types include turbomolecular pumps (turbopumps), diffusion pumps and cryogenic pumps (cryopumps). Turbopumps operate by imparting momentum towards the pump outlet to trapped gas molecules. In such a pump a gas molecule will randomly enter the turbo pump and be trapped between a rotor and a stator. When the gas molecule eventually hits the spinning underside of the rotor, the rotor imparts momentum to the gas molecule, which then heads towards the exhaust. Diffusion pumps entrain gas molecules using a jet stream of hot oil. In diffusion pumps the downward moving oil jet basically knocks air molecules away from the vacuum chamber. Cryopumps trap gas molecules by condensing them on cryogenically cooled arrays.

Vacuum systems

An enclosed chamber called a vacuum chamber constitutes the environment in which PVD occurs. Such chambers are usually made of stainless steel or, in older systems, glass. To create and/or maintain a high vacuum in the vacuum chamber, a two-pump system of a rough pump and a high vacuum pump must be used, as no single pump of any type can both create and maintain a high vacuum space while operating between a high vacuum at its inlet port and atmospheric pressure at its exhaust. A generic vacuum system that may be employed in PVD is shown in Fig. 2.16.

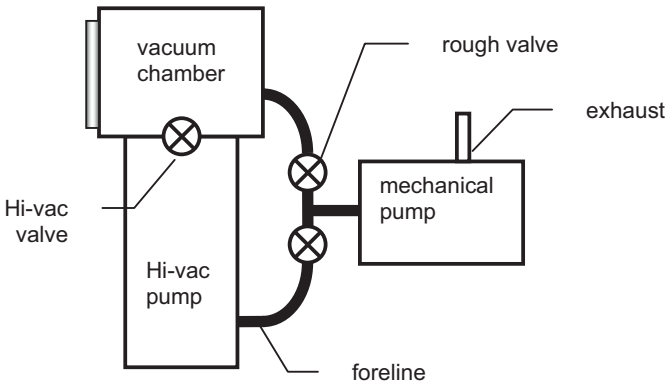


Fig. 2.16. Typical vacuum system setup in a PVD system

In Fig. 2.16 notice that although the exhaust of the rough pump is always the atmosphere, the inlet to the rough pump can be either the vacuum chamber or the exhaust of the high vacuum pump. The line connecting the exhaust of the high vacuum pump to the rough pump is called the foreline. The inlet to the high vacuum pump is the vacuum chamber itself. Operation of a vacuum system in order to first create a vacuum for PVD, perform PVD in the vacuum chamber, and then power down the system essentially involves alternately opening and closing the various valves in the correct order while running the appropriate pump(s).

Vacuum theory and relationships

Using a vacuum deposition process has many advantages essentially stemming from the fact that the number of molecules in a gas is directly related to pressure in a given volume as shown by the ideal gas equation,

$$PV = Nk_bT \tag{2.24}$$

where P is pressure, V is volume, N is the number of *molecules*, k_b is Boltzmann's constant and T is gas absolute temperature (K in the SI system).[4] And so, at the greatly reduced pressures of high vacuums there are correspondingly fewer molecules.

Another big advantage of using a vacuum process in order to deposit a thin film is the long *mean free path* of the desired material atoms. The mean free path represents the average distance a molecule travels before colliding with another molecule. In PVD the mean free path is of the same order of magnitude as the distance from the source to the substrate, and sometimes even longer. This means that the source atoms are unlikely to collide with other atoms on the way to the substrate, causing PVD's line of sight deposition characteristic. As such, a substrate that is placed in the line of sight of the source will receive most of the source atoms.

The kinetic theory of gases gives a very good estimate of the mean free path in a vacuum.[5] The mean free path of a molecule can be determined by:

$$\lambda = \frac{V}{\sqrt{2}N\sigma} = \frac{k_bT}{\sqrt{2}\sigma P} \tag{2.25}$$

where λ is the mean free path of the molecule and σ is the interaction cross section. The interaction cross section represents the likelihood of interaction between particles, and has dimensions of area.

Example 2.3 illustrates these relations.

---

[4] You may be familiar with a different form of the ideal gas equation: $PV = nR_uT$, in which n is the number of *moles* and R_u is the universal gas constant equal to 8.314 J/mol-K. In the form we use here we prefer actual number of molecules rather than moles, and Boltzmann's constant takes the place of R_u. In fact, one interpretation of Boltzmann's constant is the ideal gas constant on a per unit molecule basis.

[5] The kinetic theory models the molecules of a gas as infinitesimally small masses, therefore dictating that the only form of energy they can have is kinetic energy. Furthermore, the theory assumes that the collisions between molecules are all elastic, so that kinetic energy is conserved during the collision. The most well-known result of the kinetic theory is the ideal gas equation.

Example 2.3

Number of molecules and mean free path in air at high vacuum pressure

Estimate the number of air molecules in 1 cm^3 and the mean free path of air at room temperature at

(a) atmospheric pressure and

(b) 1×10^{-7} torr.

Take the interaction cross section to be $\sigma = 0.43$ nm^2.

Solution

(a) Let's take atmospheric pressure to be 760 torr and room temperature to be 20°C. Using the ideal gas equation and solving for N we have

$$N = \frac{PV}{k_b T} = \frac{(760\text{ torr})(1\times10^{-6}\text{ m}^3)}{(1.38\times10^{-23}\frac{\text{J}}{\text{K}})(20^o C + 273)\text{K}} \times \frac{133\text{ Pa}}{\text{torr}} \times \frac{\text{J}}{\text{Pa}\cdot\text{m}^3}$$

$$= 2.50\times10^{19}$$

For the mean free path,

$$\lambda = \frac{k_b T}{\sqrt{2}\sigma P} = \frac{(1.38\times10^{-23}\frac{\text{J}}{\text{K}})(20°\text{C} + 273)\text{K}}{\sqrt{2}(0.43\times10^{-18}\text{m}^2)(760\text{ torr})} \times \frac{\text{torr}}{133\text{ Pa}} \times \frac{\text{Pa}\cdot\text{m}^3}{\text{J}}$$

$$= 6.58\times10^{-8}\text{ m} = 65.8\text{ nm}$$

At atmospheric pressure there are about 2.50×10^{19} molecules of air in 1 cm^3, and they tend to travel an average of a mere 66 nm before colliding with other molecules.

(b) Repeating for a pressure 1×10^{-7} torr (high vacuum region)

$$N = \frac{PV}{k_b T} = \frac{(1\times10^{-7}\text{ torr})(1\times10^{-6}\text{ m}^3)}{(1.38\times10^{-23}\frac{\text{J}}{\text{K}})(20^o C + 273)\text{K}} \times \frac{133\text{ Pa}}{\text{torr}} \times \frac{\text{J}}{\text{Pa}\cdot\text{m}^3}$$

$$= 3.29\times10^{9}$$

Mean free path:

$$\lambda = \frac{k_b T}{\sqrt{2}\sigma P} = \frac{(1.38\times10^{-23}\ \frac{\text{J}}{\text{K}})(20^\circ\text{C}+273)\text{K}}{\sqrt{2}(0.43\times10^{-18}\ \text{m}^2)(1\times10^{-7}\ \text{torr})} \times \frac{\text{torr}}{133\ \text{Pa}} \times \frac{\text{Pa}\cdot\text{m}^3}{\text{J}}$$
$$= 500\ \text{m}$$

We see that the number of molecules in 1 cm^3 of air has decreased by a factor of 10^{10} in the high vacuum compared to atmospheric pressure. If this high vacuum were to be used as the environment for PVD, the likelihood of impurities in the sample, especially reactive species such as oxygen, would also decrease by the same factor. Furthermore, with a mean free path of 500 m there is virtually no chance that a source atom will collide with an air molecule en route to the substrate. ◄

2.4.2 Thermal evaporation

As a physical vapor deposition process, thermal evaporation refers to the boiling off or sublimation of heated solid material in a vacuum and the subsequent condensation of that vaporized material onto a substrate. Evaporation is capable of producing very pure films at relatively fast rates on the order of μm/min.

In order to obtain a high deposition rate for the material, the vapor pressure (the pressure at which a substance vaporizes) of the source material must be above the background vacuum pressure. Though the vacuum chamber itself is usually kept at high vacuum, the local source pressure is typically in the range of 10^{-2}–10^{-1} torr. Hence, the materials used most frequently for evaporation are elements or simple oxides of elements whose vapor pressures are in the range from 1 to 10^{-2} torr at temperatures between 600°C and 1200°C. Common examples include aluminum, copper, nickel and zinc oxide. The vapor pressure requirement excludes the evaporation of heavy metals such as platinum, molybdenum, tantalum, and tungsten. In fact, evaporation crucibles, the containers that hold the source material, are often made from these hard-to-melt materials, tungsten being the most common.

The flux, or number of molecules of evaporant leaving a source surface per unit time and per unit area is given by

$$F = N_0 \exp\left(\frac{-\Phi_e}{k_b T}\right) \tag{2.26}$$

where F is the flux, Φ_e is the activation energy[6] (usually expressed in units of eV) and N_0 is a temperature dependant parameter. The utility of Eq. (2.26) is that it will tell you the relative ease or difficulty of evaporating one material versus another. It can be recast into another form:

$$F = \frac{P_v(T)}{\sqrt{2\pi M k_b T}}, \tag{2.27}$$

where $P_v(T)$ is the vapor pressure of the evaporant and M is the molecular weight of the evaporant.

We have already seen that the mean free path of a gas tends to become very large at small pressures. This has a strong correlation to the fraction of evaporant atoms that collide with residual gas atoms during evaporation. The collision rate is inversely proportional to a quantity known as the Knudsen number, given by

$$Kn = \lambda/d \tag{2.28}$$

where Kn is the Knudsen number, d is the source to substrate distance and λ is the mean free path of the residual gas. For Knudsen numbers larger than one, the mean free path of molecules is larger than the distance evaporated molecules must travel to create a thin film, and thus collision is less likely. Ideally this number is much greater than one. Table 2.6 gives the Knudsen number at various pressures for typical source-to-substrate distances of 25–70 cm. As seen in the table, high vacuum pressures are most favorable for evaporation.

Table 2.6. Typical Knudsen numbers for various vacuum pressures

Pressure (torr)	$Kn = \lambda/d$
10^{-1}	~0.01
10^{-4}	~1
10^{-5}	~10
10^{-7}	~1000
10^{-9}	~100,000

[6] We have seen the term “activation energy” once before. In general activation energy refers to the minimum amount of energy required to activate atoms or molecules to a condition in which it is equally likely that they will undergo some change, such as transport or chemical reaction, as it is that they will return to their original state. Here it is the energy required to evaporate a single molecule of the source material.

Types of evaporation

There are two basic methods to create a vapor out of the source material in evaporation. One method is by **resistive heating** in which the source material is evaporated by passing a large electrical current through a highly refractory metal structure that contains the source (such as a tungsten "boat") or through a filament. The size of the boat or filament limits the current that can be used, however, and contaminants within the boat or filament can find their way into the deposited film.

In **electron beam (e-beam) evaporation** an electron beam "gun" emits and accelerates electrons from a filament. The emitted e-beam eventually impacts the center of a crucible containing the evaporant material. The electron beam is directed to the crucible via a magnetic field, usually through an angle of 270°. (Fig. 2.17.) The crucible needs to be contained in a water-cooled hearth so that the electron beam only locally melts the material and not the crucible itself, which could contaminate the deposited film. This process usually occurs at energies less than 10 keV in order to reduce the chances that X-rays are produced, which can also damage the film.

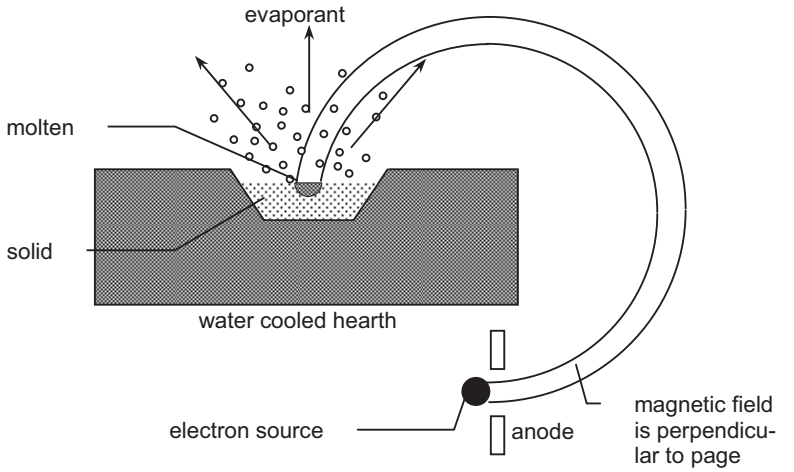


Fig. 2.17. Electron beam evaporation configuration (Adapted from Madou)

Shadowing

When we shine a flashlight, the light beam illuminates certain surfaces, but casts shadows upon others. An evaporant beam behaves much in the same way. In fact, when an evaporant is prevented from being deposited on a certain surface due to the evaporant stream's inability to "see" the surface, the phenomenon is appropriately called **shadowing**.

We have already introduced the flux F of the source material in evaporation to be the amount of material leaving a surface per unit surface area per unit time. Another quantity of interest is the arrival rate, or the amount of material incident on a surface per unit surface area per unit time. Due to the line-of-sight nature of evaporation the arrival rate is dependent on the geometry of the evaporation set-up. In reference to Fig. 2.18 the arrival rate is given by

$$A = \frac{\cos\beta\cos\theta}{d^2} F \tag{2.29}$$

where A is the arrival rate, β is the angle the substrate makes with its normal, θ is the angle the source makes with its normal and d is the distance from the source to the substrate. Equation (2.29) essentially captures how well the deposition surface can "see" the source. From Eq. (2.29) we see that where β and/or θ are large, the source does not see the substrate well and shadowing occurs. When the light-of-sight of the flux is normal to the substrate and source, however, the arrival rate achieves a maximum. Furthermore, we see that the large source to substrate distances result in lower arrival rates. It should be noted that Eq. (2.29) applies to small evaporation sources only.

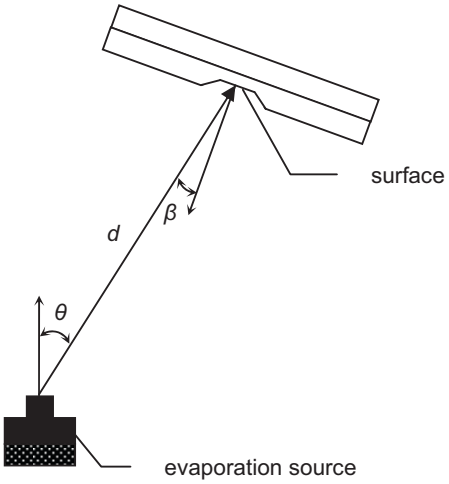


Fig. 2.18. Geometry of small source evaporation (Adapted from Madou)

This directional dependence makes it difficult to obtain uniform coatings over large surfaces areas, and particularly over topographical steps and surface features on a wafer. The ability of a technique to create films of uniform surface features is called *step coverage*. One way to gage step coverage is to compare the thickness of films at different locations on a surface. As the thickness of a film resulting from evaporation is proportional to the arrival rate, Eq. (2.29) can be used to make such a comparison. In reference to Fig. 2.19 (a), Eq. (2.29) predicts that the ratio of film thickness t_1 to film thickness t_2 is given by

$$\frac{t_1}{t_2} = \frac{\cos\beta_1}{\cos\beta_2}. \tag{2.30}$$

In this case $\beta_1 = 0$ and $\beta_2 = 60^o$, resulting in $t_1/t_2 = 2$. That is, the film thickness at location (1) is twice the thickness at location (2). For cases with large angles such as in Fig. 2.19 (b), shadowing can be extreme. We will see in Chapter 5 that this non-uniformity can result in unintended stress in the film, which can lead to the cracks or the film peeling away from the substrate.

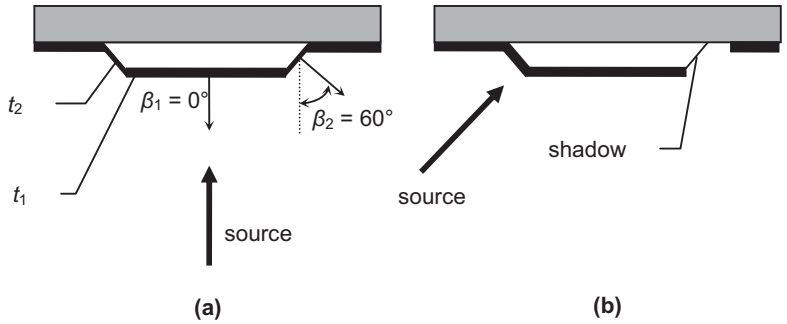


Fig. 2.19. Shadowing: (a) The film thickness on the horizontal surfaces are twice that of the sloped surfaces. (b) An extreme case of shadowing. (Adapted from Madou)

There are two major ways major to reduce shadowing. One way is continuously rotate the substrate in planetary holders during deposition to vary the angle β. Another way is to heat the substrate to 300-400°C in order to increase the ability of the deposited material to move around the surface. In a typical evaporation process both techniques are employed simultaneously.

Though shadowing can often be a problem during evaporation, the clever microfabricator can use shadowing to keep evaporated material from being deposited in undesirable locations.

2.4.3 Sputtering

In evaporation an energy source causes a material to vaporize and subsequently to be deposited on a surface. After the required energy is delivered to the source material, we can visualize the process as a gentle mist of evaporated material en route to a substrate. In sputtering, on the other hand, a target material is vaporized by bombarding it with positively charged ions, usually argon. In sputtering the high energy of the bombarding ions violently knocks the source material into the chamber. In stark contrast to a gentle mist, sputtering can be likened to throwing large massive rocks into a lake with the intent of splashing the water. (Fig. 2.20.)

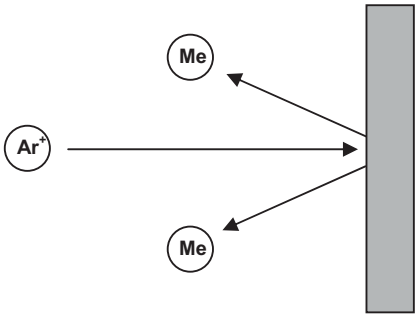


Fig. 2.20. Sputtering of metal by ionized argon

A more technical definition of sputtering is the ejection of material from a solid or molten source (the target) by kinetic energy transfer from an ionized particle. In almost all cases argon atoms are ionized and then accelerated through an electrical potential difference before bombarding the target. The ejected material "splashes" away from the cathode in a linear fashion, subsequently condensing on all surfaces in line-of-sight of the target.

The bombarding ions tend to impart much larger energies to the sputtered material compared to thermally evaporated atoms, resulting in denser film structures, better adhesion and larger grain[7] sizes than with evaporated films. Furthermore, because sputtering is not the result of thermally melting a material, virtually any material can be sputtered, and at lower temperatures than evaporation. As a result, sputtering does not require as high a vacuum as evaporation, typically occurring in the range of 10^{-2} to 10^{-1} torr. This is three to five orders of magnitude higher than evaporation pressures. However, sputtering tends to take a significantly longer period of time to occur than thermal evaporation, a relatively quick process.

The number of target atoms that are emitted per incident ion is called the *sputtering yield*. Sputtering yields range from 0.1 to 20, but tend to be around 0.5-2.5 for the most commonly sputtered materials. Typical sput-

[7] The silicon in a silicon substrate is an exception to most solid structures in that it is composed of one big single-crystal structure. By contrast, most deposited materials are *polycrystalline*; that is, they are made of a large number of single crystals, or *grains*, that are held together by thin layers of amorphous solid. Typical grain sizes vary from a just few nanometers to several millimeters. Indeed, when one sputters a thin film of silicon, it is usually referred to as *polysilicon*.

tering yields for some common materials are given in Table 2.7 for an argon ion kinetic energy of 600 eV.

Table 2.7. Sputtering yields for various materials at 600 eV

Material	Symbol	Sputtering yield
Aluminum	Al	1.2
Carbon	C	0.2
Gold	Au	2.8
Nickel	Ni	1.5
Silicon	Si	0.5
Silver	Ag	3.4
Tungsten	W	0.6

The amount of energy required to liberate one target atom via sputtering is 100 to 1000 times the activation energy needed for thermal evaporation. This, combined with the relatively low yields for many materials, leads to sputtering being a very energy inefficient process. Only about 0.25% of the input energy goes into the actual sputtering while the majority goes into target heating and substrate heating. Hence, deposition rates in sputtering are relatively low.

There are many methods used to accomplish the sputtering of a target. The original method is called DC sputtering (direct current sputtering) because the target is kept at a constant negative electric potential, serving as the cathode. As such, DC sputtering is limited to materials that are electrically conductive. RF sputtering (radio frequency sputtering), on the other hand, employs a time-varying target potential, thereby allowing nonconductive materials to be sputtered as well. In reactive sputtering a gas that chemically reacts with the target is introduced into the system and the products of reaction create the deposited film. To help overcome the relatively low deposition rates associated with sputtering, the use of magnets behind the target is sometimes employed in a process called magnetron sputtering. Magnetron sputtering can be used with either DC or RF sputtering. Each of these methods is described in more detail below.

DC sputtering

In DC sputtering the target material serves as a negatively charged surface, or cathode, used to accelerate positively charged ions towards it. As such,

the two basic requirements for DC sputtering are that the source material be electrically conductive and that is has the ability to emit electrons. Typically the metallic vacuum chamber walls serve as the anode, but sometimes a grounded or positively biased electrode is used.

In the process of DC sputtering, an inert gas, typically argon, is introduced into a vacuum chamber. The gas must first be ionized before sputtering can begin. (An ionized gas is usually referred to as a *plasma*.) At first random events will cause small numbers of positive argon ions to form. The Ar^+ ions sufficiently close to the target will be accelerated towards the cathode, while the electrons near the anode are accelerated towards the anode. As the Ar^+ ions collide with the target, the target material, electrons and X-rays are all ejected. The target itself is heated as well. The ejected electrons are accelerated away from the electrode and into the argon gas, causing more ionization of argon. The sputtering process eventually becomes self-sustaining, with a sputtering plasma containing a near-equilibrium number of positively ionized particles and negatively charged electrons. Figure 2.21 shows a DC sputtering configuration.

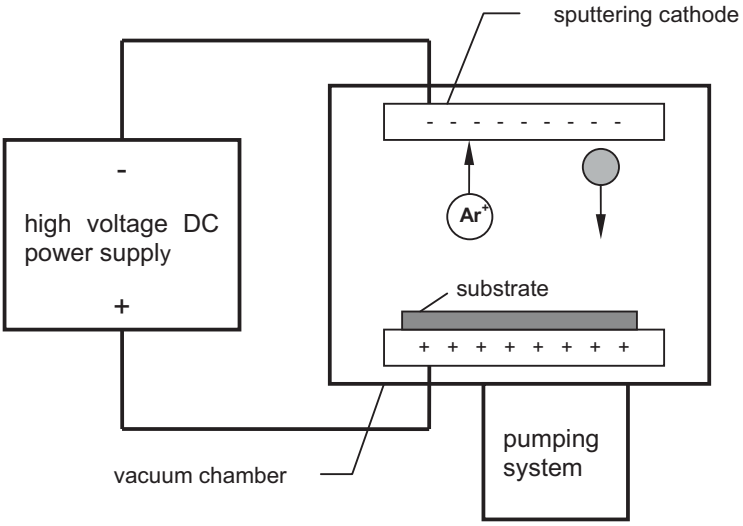


Fig. 2.21. A typical DC sputtering configuration

RF Sputtering

RF sputtering removes the requirement that the target material be electrically conductive. In RF sputtering the electric potential applied to the target alternates from positive to negative at high enough frequencies, typically greater than 50 kHz, so that electrons can directly ionize the gas atoms. This scheme works because the walls of the vacuum chamber and the target form one big electrical capacitor, and the ionized gas in the region next to the target and the target itself form another small capacitor contained in the larger one. The two capacitors thus have some capacitance in common, allowing for the transfer of energy between them in what is known as *capacitive coupling*. A result of this coupling, however, is that sputtering occurs both at the target and at the walls of the chamber. Luckily, the relative amount of sputtering occurring at the walls compared to the target is correlated to the ratio of wall to target areas. Since the area of the target is much smaller than the wall area, the majority of sputtering is of the target material.

Reactive Sputtering

Reactive sputtering is a method to produce a compound film from a metal or metal alloy target. For example, an aluminum oxide film can be deposited in reactive sputtering by making use of an aluminum target. In reactive sputtering a reactive gas, such as oxygen or nitrogen, must also be added to the inert gas (argon). The products of a chemical reaction of the target material and the gas form the deposited material. (The occurrence of a chemical reaction is a trait reactive sputtering has in common with **chemical vapor deposition**, discussed in the next section.). Formation of the reactive compounds may occur not only on the intended surface, but also on the target and within the gas itself.

The amount of reactive gas is very important issue in reactive sputtering, since excessive reactive gas tends to reduce the already low sputtering rates of most materials. A fine balance must be made between reactive gas flow and sputtering rate. Furthermore, the behavior of the plasma in reactive sputtering is quite complicated, as the reactive gas ionizes along with the inert gas. Hence, drastically different deposition rates may result for a metal oxide compared to the metal by itself. A number of approaches used to avoid the reduction in sputtering rate for reactive sputtering are detailed in the literature.

Magnetron sputtering

Sputtering processes do not typically result in high deposition rates. However, the use of magnetic discharge confinement can drastically change this. Magnets carefully arranged behind the target generate fields that trap electrons close to the target. Electrons ejected from the target feel both an electric force due the negative potential of the target and a magnetic force due to the magnets placed behind the target given by the Lorentz force

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) . \tag{2.31}$$

With properly placed magnets of the correct strength, electrons can only go so far away from the target before turning around and hitting the target again, creating an electron "hopping" behavior. The path length of the electrons increases near the cathode improving the ionization of the gas near the cathode. This local increase in ionization produces more ions that can then be accelerated toward the cathode, which in turn provides higher sputtering rates. Figure 2.22 shows a typical magnetron sputtering geometry and the resulting electron path.

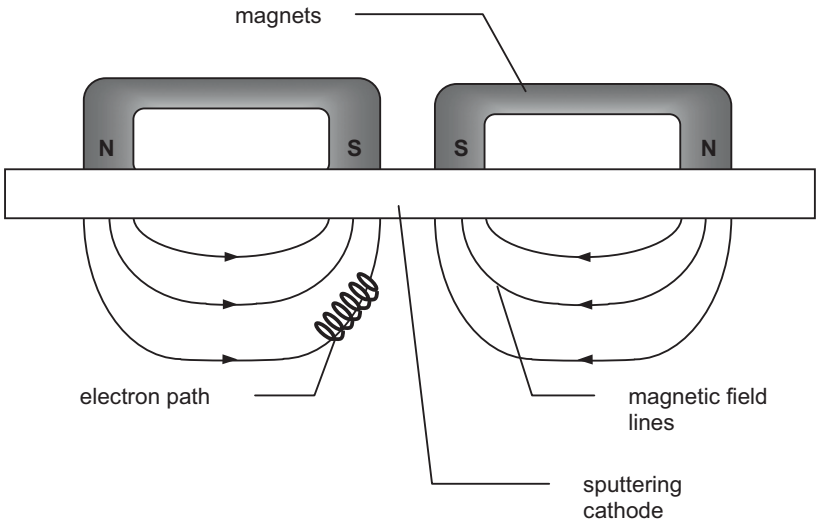


Fig. 2.22. Magnetron principle

2.5 Other additive techniques

2.5.1 Chemical vapor deposition

In chemical vapor deposition, or simply CVD, a chemical reaction takes place in order to deposit high-purity thin films onto a surface. In CVD processes, a surface is first exposed to one or more volatile **precursors**, vapors containing the to-be-deposited material in a different chemical form. A chemical reaction then takes place, depositing the desired material onto the surface, and also creating gaseous byproducts which are then removed via gas flow. (Fig. 2.23.) A common example of CVD is the deposition of silicon films using a silane (SiH_4) precursor. The silane decomposes forming a thin film of silicon on the surface with gaseous hydrogen as a byproduct.

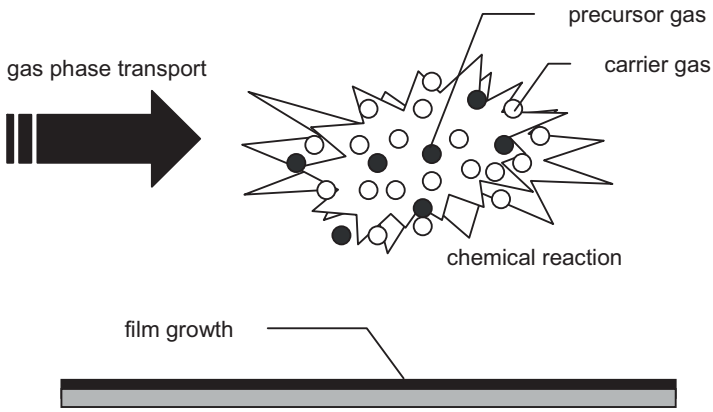


Fig. 2.23. Basic chemical vapor deposition process

Often CVD processes result in the formation of volatile and/or hazardous byproducts, such as HCl in the deposition of silicon nitride (Si_3N_4) using dichlorosilane (SiH_2Cl_2) and ammonia (NH_3) precursors. This hazardous waste is one of the disadvantages of the CVD technique as compared to physical vapor deposition methods. However, CVD is not a line-of-sight deposition method, and thus offers excellent step coverage and greatly improved uniformity over PVD. However, temperatures ranging from 500°-850°C are often required for CVD, making it impossible to use with silicon

surfaces already deposited with certain metals such as aluminum or gold, as eutectics[8] may form.

There are several methods of enhancing the deposition rates during CVD. These include employing low pressures (LPCVD - low pressure CVD) as well as the use of plasmas (PECVD – plasma enhanced CVD).

2.5.2 Electrodeposition

Also called electroplating, **electrodeposition** refers to the electrochemical process of depositing metal ions in solution onto a substrate. It is commonly used to deposit copper and magnetic materials in MEMS devices.

Surface quality of electrodeposited films tends to be worse than films deposited via physical vapor deposition, exhibiting a higher degree of roughness. Uniformity of the films can also be an issue, but can be improved by careful control of applied electric current.

2.5.3 Spin casting

Some materials, particularly polymers, can be added to a substrate by dissolving them in solution, applying them to a wafer, and then spinning the wafer to distribute the solution across the surface via centrifugal force. Afterwards the wafer is baked to remove the solvent, leaving behind the thin film. The process is also commonly called simply **spinning**.

Gelatinous networks of colloidal suspensions containing solid polymer particles, called sol-gels, are often spin cast. Spin casting is also the standard technique for applying photoresist to wafers. It is discussed in more detail in the next chapter.

2.5.4 Wafer bonding

Perhaps the ultimate additive technique is to add an entire wafer to another one. This technique is used in the fabrication of some MEMS devices, but more commonly it is used to create a protective enclosure for a MEMS, or to **package** it.

---

[8] The term eutectic refers to an alloy of two or more metals whose melting point is lower than that of any other alloy composed of the same constituents in different proportions. In short, by combining silicon with the right proportions of aluminum or gold, the resulting interface can potentially start melting at significantly lower temperatures than any of the materials by themselves.

Silicon wafers can be bonded to one another via a high temperature (~1000°C) anneal, fusing the wafers together. Silicon wafers can also be bonded to 7740 Pyrex glass at temperatures of about 500°C if a voltage is simultaneously applied across the wafer stack. Such a process is called *anodic bonding*, with typical voltages on the order of 400-700 V. The Pyrex is required for the bonding since its coefficient of thermal expansion closely matches that of silicon. Otherwise the glass could shatter upon cooling.

Materials such as adhesives and low melting point solders are sometimes used to bond wafers in cases where higher temperature methods cannot be employed.

Essay: Silicon Ingot Manufacturing

Patrick Ferro
Department of Mechanical Engineering, Gonzaga University

For reliable operation of electronic devices, the chips upon which they are fabricated must be free of atomic-level defects. The precursor to defect-free devices is single-crystal silicon, grown as an ingot or boule.

Single-crystal silicon ingots are grown via a controlled, rotating withdrawal process known as the Czcochralski technique. Sometimes practitioners abbreviate Czochralski-grown silicon as 'CZ silicon'.

The equipment used to grow silicon ingots includes a resistance furnace which supplies radiant heat to a rotating quartz crucible. The resistance furnace and rotating crucible are contained within a vacuum chamber that has at least one interlock chamber and a capability for inert gas backfill.

The process starts by melting chunks of very pure silicon (impurities are measured at parts per million level) in a quartz crucible, along with controlled amounts of dopant elements. The crucible is slowly rotated in one direction at a specific rotational velocity (e.g. 10 rpm) and slowly heated until the silicon within it is molten. The rotating, molten silicon-filled crucible is stabilized at a particular temperature above the liquidus, or temperature above which crystals can no longer coexist with the melt in a homogeneous state. A counter-rotating crystal seed is then slowly lowered to touch the surface of the melt.

As the seed makes contact with the surface of the molten silicon, a temperature gradient is created. At this critical point in the process, the intention is to initiate and sustain the growth of a single grain, or crystal, of silicon. In practical terms, the operator will begin slowly withdrawing the rotating seed away from the counter-rotating melt, to pull a thin 'neck' of

silicon up and out of the melt. The withdrawal rate is process-dependent and is designed to sustain the growth of a neck that is free of defects, called dislocations, within the crystal structure. The neck diameter is similar to that of a pencil, and the withdrawal rate is measured in only millimeters per minute. The neck-growth stage of the process continues until a neck of approximately 10 cm has been grown.

The process continues by temporarily decreasing the withdrawal rate to increase the diameter of the ingot. Once the diameter is at its desired dimension, the withdrawal rate is stabilized and the ingot is allowed to grow. The withdrawal rate is on the order of millimeters per minute and the seed and crucible rotation rates are on the order of 10 to 20 rpm. The transition between the neck and the ingot body is known as the 'shoulder'. Lines, or ridges, at a regular spacing can be seen in the radial direction on the shoulder and on the body of an ingot due to the crystallinity of the silicon. If dislocations develop in the silicon during the body-growth phase of the process, the appearance of these lines is affected. If dislocations are observed during the process, the partially-grown ingot will be scrapped and remelted. In a silicon ingot manufacturing plant, productivity is closely monitored because of the long processing times involved. For example, one good ingot may require more than twenty four hours of furnace time.

Reducing the onset of dislocation in ingots requires careful attention to changes in withdrawal rate and rotational speeds as well as minimizing the effects of vibrations, tramp element levels, vacuum and other process parameters. Because the requirements are so stringent, even secondary parameters such as the bubble morphology in the quartz crucibles may be carefully monitored.

A well-controlled process allows for the sustained growth of a body of single-crystal silicon. The growth of the body continues until the crucible is completely empty of molten silicon. A completed ingot, or boule, will have a tapered point at the bottom indicating that the entire crucible of silicon was pulled out. A partially grown ingot cannot be rapidly pulled out, since the shock of a sudden withdrawal would propagate through the length of the partially-grown ingot, rendering it unusable.

Solidification, in general and for all materials, proceeds in the direction of the highest thermal gradient. It is critically important during the growth of a silicon ingot to control and maintain a high thermal gradient at the solidification front. Furnace design, including the use of baffles, can help in the control of the thermal gradient. Also, the process parameters including withdrawal rate, rotational speed, and chamber pressure affect the thermal gradient.

The CZ process is relatively slow, and is nearly an art form. Some parts of the process may be automated, but still require a relatively high degree

of involvement from skilled operators and process engineers. A silicon wafer manufacturing plant may have twenty or more CZ silicon furnaces to meet the continuous demand.

Dislocation-free and chemically pure silicon ingots allow for defect-free wafers, upon which both electronic devices and MEMS can be reliably fabricated.

References and suggested reading

Alciatore DG, Histand MB (2007) Introduction to Mechatronics and Measurement Systems, 3rd edn. McGraw Hill, Madison, WI

"Substrate Manufacture: Single Crystal Ingot Growth" U.S. Department of Labor Occupational Safety & Health Administration http://www.osha.gov/SLTC/semiconductors/substratemfg/snglcrystlingtgrowth.html

Franssila S (2004) Introduction to Microfabrication. Wiley, Chichester, West Sussex, England

Griffiths D (1999) Electrodynamics. In: Reeves A (ed) Introduction to Electrodynamics, 3rd edn. Prentice Hall, Upper Saddle River, NJ

Jaeger RC (2002) Introduction to Microelectronic Fabrication, 2nd edn. Prentice Hall, Upper Saddle River, NJ

Madou MJ (2002) Fundamentals of Microfabrication, The Art and Science of Miniaturization, 2nd edn. CRC Press, New York

Maluf M, Williams K (2004) An Introduction to Microelectromechanical Systems Engineering, 2nd edn. Artech House, Norwood, MA

Peeters E (1994) Process Development for 3D Silicon Microstructures with Application to Mechanical Sensor Design. Ph.D. thesis, Catholic University of Louvain, Belgium

Senturia S (2001) Microsystem Design. Kluwer Academic Publishers, Boston

Serway RA (1998) Principles of Physics, 2nd edn. Saunders College Pub., Fort Worth, TX

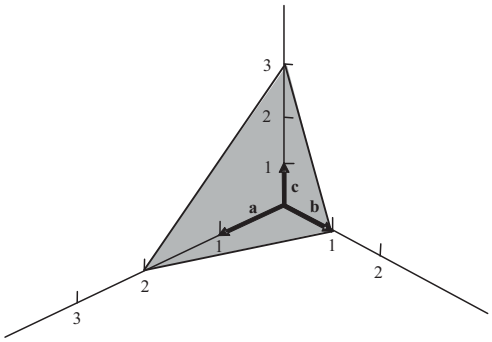
Vossen JL, Kern W (1978) Thin Film Processes, Academic Press, New York

"Sputtering Yields Reference Guide" *Angstrom Sciences* http://www.angstromsciences.com/reference-guides/sputtering-yields/index.html

Wolf S, Tauber RN (2000) Silicon Processing for the VLSI Era, Vol.1: Process Technology, 2nd edn. Lattice Press, Sunset Beach, CA

Questions and problems

2.1 Silicon is the most common material out of which substrates are made. Find at least one other material used for substrates. What are its advantages and/or disadvantages over silicon?
2.2 What are three reasons you might add a layer to a silicon substrate in making a MEMS?
2.3 Would you expect a silicon wafer doped with antimony to be a p-type or n-type wafer? Why? What about a silicon wafer doped with gallium?
2.4 Give three applications of a p-n junction.
2.5 What is shadowing? Is shadowing a good or bad thing? Why?
2.6 Why is a vacuum needed in physical vapor deposition?
2.7 Calculate the mean free path of an air molecule for a pressure of 10^{-6} torr. Also calculate the Knudsen number for an evaporation set-up for which the source to substrate distance is d = 50 cm. Is this an adequate pressure to use for this evaporation? Why or why not?
2.8 Determine the Miller indices for plane shown in the figure.



2.9 The lattice constant for Si, a cubic material, is a = 5.43 Å. Determine the distance between adjacent planes for both the [100] and [111] directions.
2.10 The angle between two crystal directions in a cubic material is given by

$$\cos\theta = \frac{h_1h_2 + k_1k_2 + l_1l_2}{\sqrt{\left(h_1^2 + k_1^2 + l_1^2\right)\left(h_2^2 + k_2^2 + l_2^2\right)}}$$

The direction [h k l] of the line produced by the intersection of planes (h_1 k_1 l_1) and (h_2 k_2 l_2), is given by

$$h = k_1 l_2 - k_2 l_1$$
$$k = h_1 l_2 - h_2 l_1$$
$$l = h_1 k_2 - h_2 k_1$$

Find the angle between the crystal directions [111] and [110]. Find the direction of the line produced by the intersection of these two planes.

2.11 An oxide layer is to be grown on a (111) silicon wafer that is initially 300 μm thick. After one hour of wet oxidation the oxide layer is estimated to be 300 nm thick. What is the new thickness of the wafer?

2.12 A (100) silicon wafer has an initial 100 nm of oxide on its surface.
   a. How long will it take to grow an additional 600 nm of oxide using wet oxygen method at 1100°C?
   b. Graph Eq. (2.18), i.e. plot time vs. oxide thickness and then look up the time required to grow 600 nm of oxide. Comment on the result.
   c. What is the color of the final oxide under vertical illumination with white light? (Hint: Many of the references for this chapter have "color charts" that indicate oxide thickness as a function of apparent color.)

2.13 A p-type Si-wafer with background doping concentration of 1.4×10^{15} cm^{-3} is doped by ion implantation with a dose of phosphorus atoms of 10^{16} cm^{-2}, located on the surface of the wafer. Next thermal diffusion is used for the drive-in of phosphorous atoms into the p-type water (the anneal step). The wafer is then annealed at 1000°C for 3 hours.
   a. What is the diffusion constant of phosphorous atoms at this anneal temperature?
   b. What is the junction depth after the drive-in anneal?
   c. What is the surface concentration after the drive-in anneal?

http://www.springer.com/978-0-387-09510-3

Introductory MEMS
Fabrication and Applications
Adams, Th.M.; Layton, R.A.
2010, XV, 440 p., Hardcover
ISBN: 978-0-387-09510-3