DEPARTMENT OF ELECTRONICS & COMMUNICATION

Notes

IC TECHNOLOGY

(Subject Code: NEC603)

Prepared By: R R TRIPATHI

Class: B. Tech.III Year, VI Semester

SYLLABUS

UNIT 2 : DIFFUSION & ION IMPLANTATION- Ficks diffusion Equation in One Dimension, Atomic model, Analytic Solution of Ficks Law, correction to simple theory , Diffusion in SiO_2. Ion Implantation and Ion Implantation Systems Oxidation. Growth mechanism and Deal-Grove Model of oxidation, Linear and Parabolic Rate co-efficient, Structure of SiO_2, Oxidation techniques and system, Oxide properties

Beyond the Syllabus

SPUTTERING

Learning Objectives

To lay the foundation to understand the various growth and fabrication processes in IC Technology

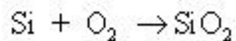
UNIT-2

Thermal Oxidation of Silicon

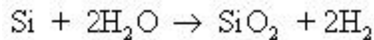
The oxide of silicon, or silicon dioxide (SiO_2), is one of the most important ingredients in integrated circuits. Thermal SiO_2 is amorphous. It has a density of 2.2 gm/cm^3 and molecular Density of 2.3x10^{22} molecules/cm^3 . The crystalline SiO_2 is also possible and known as Quartz has a density of 2.65 gm/cm^3. SiO_2 has excellent properties which makes them necessary in every part of the integrated circuits. It is an excellent electrical insulator having energy gap ~ 9 eV with a resistivity greater than 10^{20} ohm-cm and breakdown electric field greater than 10MV/cm. Si technology became popular because of the stable and reproducible Si/SiO_2 interface. Conformal oxide growth on exposed Si surface is easily possible. SiO_2 is a good diffusion mask for common dopants such as. B, P, As, Sb. In addition there exists good etching selectivity between Si and SiO_2.

The formation of SiO_2 on a silicon surface is most often accomplished through a process called thermal oxidation. As its name implies, is a technique that uses extremely high temperatures (usually between 700-1200°C) to promote the growth rate of oxide layers whose thicknesses range from 20 to 10000 nm. During the process, silicon substrate is exposed to a high purity oxidizing species like oxygen gas (dry oxidation) or water vapour (wet oxidation).

The chemical reaction at the silicon surface for dry and wet oxidation is given as



$$Si + O_2 \rightarrow SiO_2$$



$$Si + 2H_2O \rightarrow SiO_2 + 2H_2$$

Oxidation of silicon is not difficult, since silicon has a natural inclination to form a stable oxide even at room temperature, as long as an oxidizing ambient is present. In both cases the oxidizing species diffuses through the growing oxide and reacts with the silicon surface. These oxidation reactions occur at the Si-SiO_2 interface, i.e., silicon at the interface is consumed as oxidation takes place. As the oxide grows the Si-SiO_2 interface moves into the silicon substrate. As a result, the Si-SiO_2 interface will always be below the original Si wafer surface. The SiO_2 surface, on the other hand, is always above the original Si surface. SiO_2 formation therefore proceeds in two directions relative to the original wafer surface as shown in Fig.3.12.

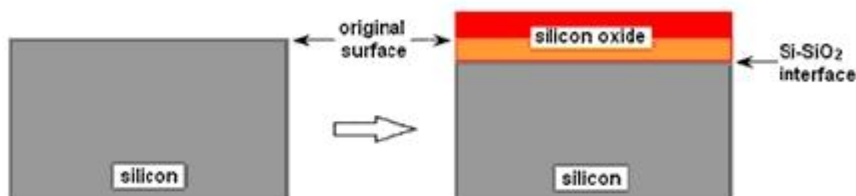


Fig2.1 : The silicon-silicon dioxide interface in thermal oxides

The amount of silicon consumed by the formation of silicon dioxide is also fairly predictable from the relative densities and molecular weights of Si and SiO_2, i.e., the thickness of silicon consumed is 44% of the final thickness of the oxide formed. Thus, an oxide that is 1000 angstroms thick will consume about 440 angstroms of silicon from the substrate. In another words, 1µm thick Si oxidizes to 2.17 µm thick SiO_2.

Thermal oxidation is accomplished using an oxidation furnace which provides the heat needed to elevate the oxidizing ambient temperature. A furnace typically consists of a temperature controlled heating system, fused quartz process tubes, arrange for controlled flow of various gases. The heating system usually consists of several heating coils that control the temperature around the furnace tubes. The wafers are placed in quartz glassware known as boats. A boat can contain many wafers. The oxidizing agent (oxygen or steam) then enters the process tube through its source end, subsequently diffusing to the wafers where the oxidation occurs.

During dry oxidation, the silicon wafer reacts with the ambient oxygen, forming a layer of silicon dioxide on its surface. In wet oxidation, the water is heated in the 40-80°C range and oxygen or nitrogen carrier gases are used for the flow of water vapors to the chamber. Alternatively hydrogen and oxygen gases are introduced into a torch chamber where they react to form water molecules, which are then made to enter the reactor where they diffuse toward the wafers. The water molecules react with the silicon to produce the oxide and another byproduct, i.e., hydrogen gas.

Kinetics of SiO_2 Growth

The Linear and Parabolic growth laws were developed by Deal and Grove, and are known as the Linear Parabolic Model. This oxide growth model has been empirically proven to be accurate over a wide range of temperatures (700-1300°C), oxide thicknesses (300-20,000 angstroms), and oxidant partial pressures (0.2-25 atmospheres).

Fig.2.2 and 2.3 pictures various diffusions possible and the concentration of species during thermal oxidation and is the basis for Deal and Grove model.

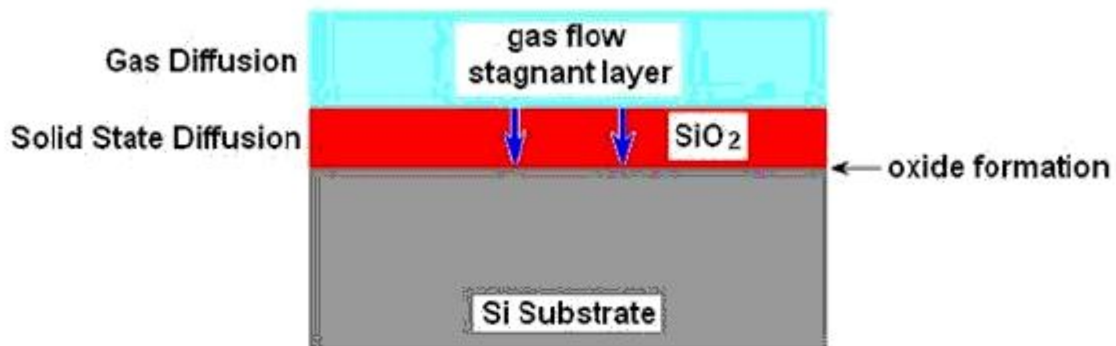


Fig. 2.2 : A Model for thermal oxidation of silicon indicating various diffusions possible

DEAL GROOVE MODEL OF OXIDATION

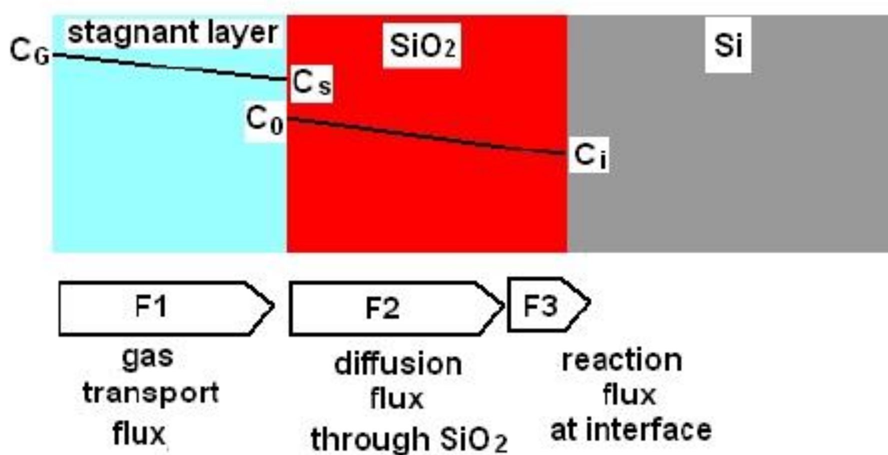


Fig. 2.3 : A Model for thermal oxidation of silicon indicating the concentration of species

The gas-phase flux F_1 is proportional to the difference between the oxidant concentration in the bulk of the gas (C_G) and the oxidant concentration adjacent to the oxide surface (C_S).

$$F_1 = h_G(C_G - C_S)$$

The Fick's law of solid sate diffusion states that

$$F_2 = -D\frac{dC}{dx} \cong D\frac{(C_0 - C_i)}{x}$$

$$F_3 = k_s C_i$$

h_G is the mass transfer constant (cm/s),D is the diffusion coefficient (cm^2/s),k_s is the surface reaction transfer constant (cm/s), Using $C_S = \frac{N}{V}$ and $PV = NkT$, Henry's law to relate C_0 and C_S

$$C_0 = HP_S = H(kT)C_S$$

Where H is the Henry's constant and P_s is the partial pressure of oxidant gases at surface

$$C_S = \frac{C_0}{HkT_0}$$

If the equilibrium concentration of oxidizing species in oxide is $C_A = HkTC_G$

$$F_1 = \frac{h_G}{HkT}(C_A - C_0)$$

Taking $\frac{h_G}{HkT}$ as h, at steady state

$$F_1 = F_2 = F_3$$

Solving these two equations for two unknowns C_0 and C_i

$$C_i = \frac{C_A}{1 + \frac{k_S}{h} + \frac{k_S x}{D}}$$

$$C_0 = C_i\left(1 + \frac{k_S x}{D}\right)$$

$$F = F_1 = F_2 = F_3 = k_S C_i = \frac{k_S C_A}{1 + \frac{k_S}{h} + \frac{k_S x}{D}}$$

To convert *F* into oxide thickness growth rate

$$F = N_1 \frac{dx}{dt} = \frac{k_S C_A}{1 + \frac{k_S}{h} + \frac{k_S x}{D}}$$

Where N_1 is the oxidant molecules per unit volume required to form a unit volume of SiO_2. N_1 is 2.3 x 10^{22} cm^{-3} for dry oxidation and 4.6 x10^{22} cm^{-3} for wet oxidation. Taking boundary conditions as $x = x_0$ at $t=0$. The oxide thickness grown at any point of time t is modeled as $x^2 + Ax = B(t + \tau)$ where A and B are constants.

$$A = 2D\left(\frac{1}{k_S} + \frac{1}{h}\right)$$

$$B = 2D\frac{C_A}{N_1}$$

$$\tau = \frac{x_i^2 + Ax_i}{B}$$

The time displacement τ is included to account for the oxide layer (at $t = 0$) formed by the accelerated growth in the initial phase of oxidation.

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

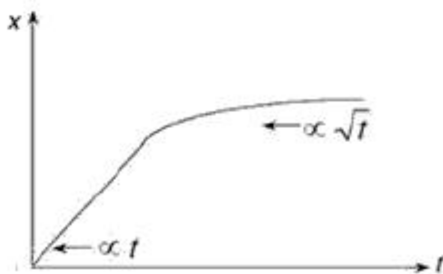


Fig.2.4 : oxide thickness versus time

LINEAR AND PARABOLIC RATE CONSTANTS

Typical oxide thickness versus time dependency is shown in Fig.3.16. For small values of *t* , the oxide growth equation can be approximated as $x = \frac{B}{A}t$ and thus $\frac{B}{A}$ is known as linear rate constant. For large values of *t*, the above equation can be approximated as $x = \sqrt{Bt}$ and thus B is known as parabolic rate constant. That is oxide thickness growth slows down with increase in oxide thickness and can be readily seen from:

$$\frac{dx}{dt} = \frac{B}{A + 2x}$$

Typical oxide thickness versus time dependency is shown in Fig.3.16. For small values of *t* , the oxide growth equation can be approximated as $x = \frac{B}{A}t$ and thus $\frac{B}{A}$ is known as linear rate constant. For large values of *t*, the above equation can be approximated as $x = \sqrt{Bt}$ and thus B is known as parabolic rate constant. That is oxide thickness growth slows down with increase in oxide thickness and can be readily seen from:

$$\frac{dx}{dt} = \frac{B}{A + 2x}$$

Dry oxidation of silicon is typically used to grow a thin, high quality oxide that is used in transistor gates and capacitors. Oxide grown in dry oxygen ambient has excellent insulating properties and is denser, free of defects. Wet oxidation of silicon is typically used to grow thick oxides that are used as diffusion barriers. Silicon dioxide acts as an effective mask against many impurities, allowing dopants into silicon only in regions that are not covered with oxide.

The oxide thickness grown on silicon is dependent on the oxidation time and temperature. Wet oxidation method offers faster growth rate compared to dry oxidation. The linear and parabolic rate constants have larger values in wet oxidation case than in case of dry oxidation. This is because equilibrium concentration of oxidizing species in oxide (C_A) is approximately three orders of magnitude greater for water than in dry oxygen. Therefore, for growing thick oxide within a realistic time, wet oxidation is a better choice.

The linear and parabolic rate constants increase with temperature in both dry and wet oxidation methods. *B* increases in both cases through the diffusion coefficient (*D*) increase due to temperature. The reason for increase in $\frac{B}{A}$ is through k_s.

OXIDATION TECHNIQUES

Dry oxidation of silicon is typically used to grow a thin, high quality oxide that is used in transistor gates and capacitors. Oxide grown in dry oxygen ambient has excellent insulating properties and is denser, free of defects. Wet oxidation of silicon is typically used to grow thick oxides that are used as diffusion barriers. Silicon dioxide acts as an effective mask against many impurities, allowing dopants into silicon only in regions that are not covered with oxide.

The oxide thickness grown on silicon is dependent on the oxidation time and temperature. Wet oxidation method offers faster growth rate compared to dry oxidation. The linear and parabolic rate constants have larger values in wet oxidation case than in case of dry oxidation. This is because equilibrium concentration of oxidizing species in oxide (C_A) is approximately three orders of magnitude greater for water than in dry oxygen. Therefore, for growing thick oxide within a realistic time, wet oxidation is a better choice.

The linear and parabolic rate constants increase with temperature in both dry and wet oxidation methods. *B* increases in both cases through the diffusion coefficient (*D*) increase due to temperature. The reason for increase in $\frac{B}{A}$ is through k_s.

$$A = 2D\left(\frac{1}{k_s} + \frac{1}{h}\right)$$

$$B = 2D\left(\frac{C_A}{N_1}\right)$$

$$\frac{B}{A} = \frac{C_A}{N_1\left(\frac{1}{k_s} + \frac{1}{h}\right)}$$

It can be seen that for small duration, the oxidation is a reaction controlled process and longer duration it is a controlled process.

From the Deal and grove model, we have seen that the oxide growth rate is affected by time, temperature, and Thickness of oxide is raised by an increase in oxidation time, oxidation temperature, or oxidation pressure. Other fa affect thermal oxidation growth rate for SiO_2 include: the crystallographic orientation of the wafer; the wafer's doping presence of halogen impurities in the gas phase and the presence of plasma during growth.

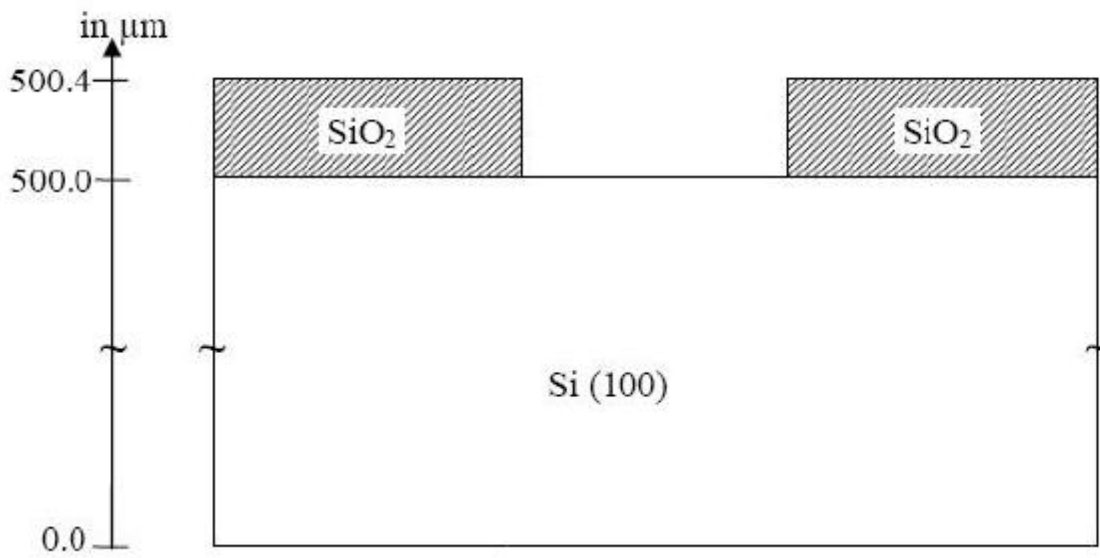


Figure2.5 Silicon oxide growth

The resulting cross section after the wet oxidation is shown in Fig. 3.18 clearly indicating the oxide-silicon inte

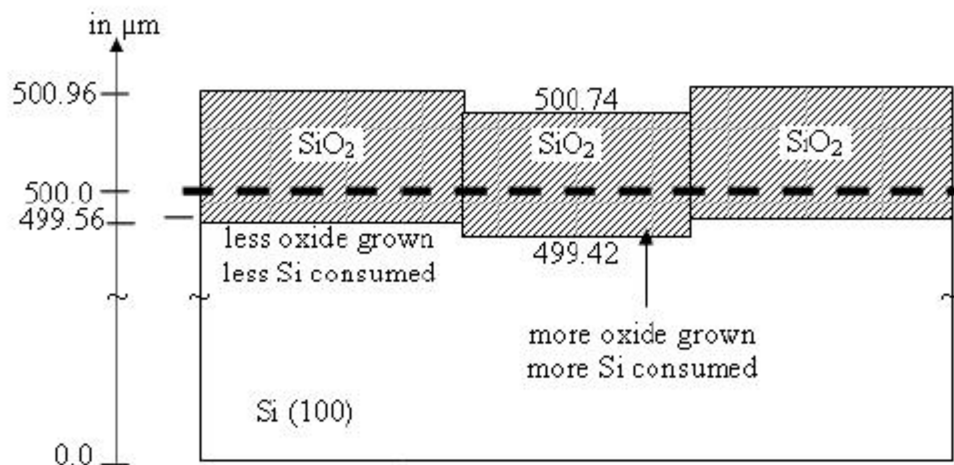


Fig. 2.6 : The Si-SiO_2 interface changes before and after thermal oxide growth

3.5 Ion-Implantation and Diffusion

Diffusion and Ion Implantation are the methods by which impurity is introduced into silicon to change

its resistivity. These processes allow the formation of sources and drains of MOSFETs and active regions of bipolar transistors. In diffusion (in a high temperature environment) the dopant atoms are moving from a highly concentrated region to a low concentrated region. The classic approach to creating regions of silicon with different electrical properties was to deposit a dopant material, on the surface, then diffuse or drive it into the surface of the silicon by exposing it to controlled periods of high temperature. As device geometries have become smaller, the side-ways diffusion has become more difficult to deal with, so the industry has converted to the ion implantation process. In implant, the dopant molecules are implanted vertically into the surface of the silicon by a high--energy ion beam. This penetrates the silicon vertically without any appreciable side-ways diffusion. Thus p-type or n-type regions in silicon are created when they are doped with Boron and Phosphorous ions respectively. The following section will deal with diffusion of dopants in a Si lattice.

Basics of Diffusion

FICKS DIFFUSION EQUATIONS

A drop of color in a glass of water spreads from concentrated region to surroundings. This occurs because of the diffusion of color molecules. Similar situation take place when free charge carriers spread in a semiconductor and dopant atoms diffuse in semiconductor at elevated temperature. In this section we use the term diffusion in the later sense. In general, the diffusion is motion of species in solids in the direction of concentration gradient. The chemical process of diffusion is described by the second order differential equation known as Fick's laws.

Fick's first law is used in steady state diffusion , i.e., when the concentration within the diffusion volume does not change with respect to time ($F_{in}=F_{out}$).

$$F = -D\frac{\partial n}{\partial x}$$

Where F is the diffusion flux (in number of dopants cm^{-2} s^{-1}), D is the diffusion coefficient or diffusivity (in $cm^2 s^{-1}$), n is the concentration of dopants (in cm^{-3}), x is the position (cm).

The diffusion coefficient (D) determines rate with which element moves in a given solid by diffusion. D depends strongly on temperature and varies between elements by orders of magnitude. For example in the case of diffusion in silicon diffusion coefficient for gold (Au), is in the range of 10^{-3} cm^2 /sec (fast diffusant) while for Antimony (Sb) is in the range of 10^{-17} cm^2 /sec. Since D is a function of temperature, the flux (F) is also a function of temperature. Negative sign indicates that the flow is down the concentration gradient.

Fick's second law is used in non-steady or continually changing state diffusion, i.e., when the concentration within the diffusion volume changes with respect to time. Because diffusion in solids is slow, diffusion is almost always transient.

$$\frac{\partial n}{\partial t} = -D\frac{\partial^2 n}{\partial x^2}$$

This can be derived from the First Fick's law and the mass balance assuming the diffusion coefficient *D* to be a constant. This equation state that the dopant will redistribute until it is spread uniformly throughout the material. The time scale of this redistribution is given by the coefficient of diffusion

The tendency toward diffusion is very strong even at room temperature because of the high molecular velocities associated with the thermal energy of the particles. The diffusion coefficient is temperature dependent, and is can be described by the Arrhenius expression:

$$D = D_0 \exp(-E_a/RT)$$

k is Boltzmann's constant. D_0 (is the maximum diffusion coefficient (at infinite temperature) and E_a (the energy of activation) is constants describing the diffusion. All kinds of diffusion require a certain minimum energy to occur and it can be called the activation energy. The higher the activation energy, the harder it is for diffusion to occur..

Typically one use higher doping concentrations and lower thermal budgets for semiconductor fabrication. It was observed that Fick's laws are not accurate enough for the explaining the anomalous diffusion behavior. Therefore it is necessary to focus on atomistic level to get insight on the interaction of dopants with lattice atoms. It is established that diffusion of impurities can only occur within the crystal lattice due to the presence of point-defects.

Diffusion by point defects

Diffusion occurs in silicon primarily at native point-defects and impurity-related point defects. Typically native defects such as point defects, line defects and surface defects are present in an otherwise perfect crystal lattice. At room temperature point defect concentration is low that practically no diffusion occurs. However, an increase of point defects will cause an increase in the diffusion. The concentration of point defects will increase for example if the crystal is heated or irradiated by ion. Both situations are likely to occur during the fabrication of devices. Impurity related defects arise from the introduction of group- III elements or group-V elements into the silicon lattice.

Normally dopants will occupy a lattice site, and the dopant is said to be substitutional. When the dopant is in the substitutional position it is not likely that it will move around. However, when it gets in contact with a point defect, it will be mobile. A point defect in silicon, can be a Si self-interstitial (a silicon atom which does not occupy a lattice site), or a vacancy (a lattice site which does not contain a silicon atom). As illustrated in Fig. 3.37, t here are mainly two possible atomic diffusion mechanisms known as interstitial and vacancy diffusion. The diffusion mechanisms through grain boundaries and surface are also possible. The highest activation energy is for volume diffusion though interstitial and vacancy sites. The grain boundary diffusion requires less energy and surface diffusion requires the least.

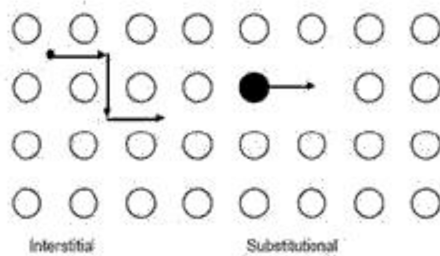


Fig. 2.7 : Diffusion mechanisms within crystal targets

Diffusion through interstitial spaces: Impurities which have small ionic radii can travel directly from one interstitial site to another one (see Fig.3.37). Especially Group-I and Group-VIII elements are diffusing mainly interstitially and are therefore fast diffusers.

Vacancy Diffusion: Vacancies are holes in the matrix and they are always moving. An impurity can move into the vacancy and diffuse through the material. A substitutional dopant can exchange its position with the neighboring vacancy. Similarly when a dopant at a substitutional site is kicked out by a silicon interstitial atom, the dopant occupies the interstitial position while the original self-interstitial has disappeared by occupying the regular lattice site.

Thus dopants can diffuse within a crystal lattice using point-defects as diffusion vehicle either by the vacancy mechanism or the interstitial mechanism or a combination of both. The dopants are selected based on their properties such as solubility in silicon, dissolution into lattice sites and ionization. The most common p-type dopant boron diffuses mainly via interstitials, where the n-type dopant phosphorus shows at intrinsic concentrations interstitial dominated diffusion and for high concentrations a dual mechanism with a strong vacancy component. Arsenic is also diffusing via interstitials and vacancies, but the interstitial diffusion is at least limited due to the relatively large ionic radii. Large dopants like Antimony will find no stable position between lattice sites, so they can only diffuse via vacancies

The transport of dopants within the lattice is not only controlled by the diffusion mechanism, there are also recombination processes between dopants and point-defects occurring, which can disturb or even prevent dopants from diffusion. These recombinations are always possible when dopant-defect pairs are approaching lattice defects of the opposite type. If the dopant concentration level is significantly below the intrinsic carrier concentration, the diffusivities of the dopants show no dependence on the doping level.

In the subsequent section we will focus on two practical circumstances in which the diffusion can be carried out. Diffusion from an unlimited source (Constant source diffusion) occurs when a wafer is exposed to an infinite amount of dopants during the diffusion period. In the case of diffusion from a limited source, a finite quantity of the dopants is first placed on the wafer and diffusion proceeds from this limited source and it is assumed that all the dopants are consumed.

Diffusion from an unlimited source

The surface concentration will be fixed at the solid solubility, and the distribution will be in the form of a complimentary

$$N(x,t) = N_0 \mathrm{erfc}\left(\frac{x}{\sqrt{4Dt}}\right)$$

where :

N = concentration (cm^{-3})
N_0 = solid solubility limit for dopant (cm^{-3})
x = position inside silicon relative to the surface
D = diffusion coefficient for dopant (cm^2/s)
t = time (seconds)

If one sketch of this concentration profile for various diffusion times, one can observe that the surface concentration i s

where as the diffusion depth increases with time. Please note that the Dose beyond x=0 continues to increase with diffu Note that the dose is given by

$$Q = \int_0^\infty N_0 \left[1 - erf\left(\frac{x}{2\sqrt{Dt}}\right)\right] = \int_0^\infty N_0 erfc\left(\frac{x}{2\sqrt{Dt}}\right) = \frac{2N_0}{\sqrt{\pi}}\sqrt{Dt}$$

The error function values for typical values are given in the Table 9.

Table 2.1 : Typical error function values

Z	erf(Z)	Z	erf(Z)
0.00	0	0.60	0.6039
0.10	0.1125	0.70	0.6778
0.20	0.2227	0.80	0.7421
0.30	0.3286	0.90	0.7970
0.40	0.4284	1.00	0.8427
0.50	0.5205	1.50	0.9661

Diffusion from a limited source

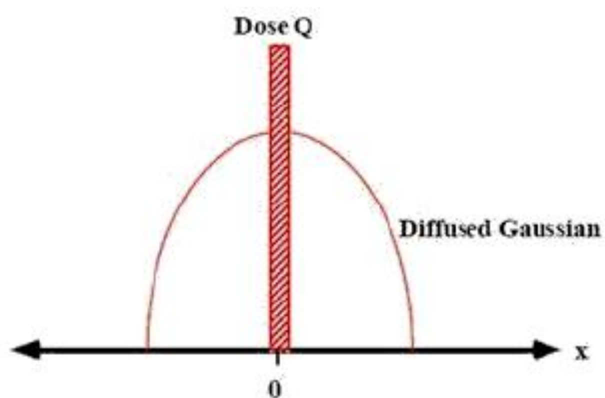


Fig. 2.8: Double sided iffusion from a delta source at origin

The solution that satisfies Fick's second law when a fixed dose Q is introduced as a delta function at the origin is given by

$$N(x,t) = \frac{Q}{2\sqrt{\pi Dt}} \exp\left\{-\frac{x^2}{4Dt}\right\}$$

Please note the following important consequences:

1. Dose Q remains constant before and after diffusion
2. Peak concentration decreases as $\frac{1}{\sqrt{t}}$
3. Diffusion distance from origin increases as $2\sqrt{Dt}$

This equation can be slightly modified to model a practical situation where all the dopants from the constant source are consumed entirely to one side.

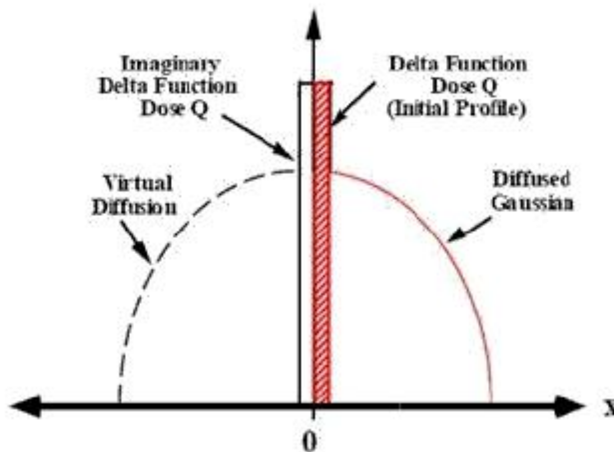


Fig. 2.9 : Single side d iffusion from a delta source at origin

The solution is given by

$$N(x,t) = \frac{Q}{\sqrt{\pi Dt}} \exp\left\{-\frac{x^2}{4Dt}\right\} = N(0,t) \exp\left\{-\frac{x^2}{4Dt}\right\}$$

Also note that the infinite source diffusion case discussed previously can be thought infinite source of dopant made up of small slices each diffusing as a Gaussian. Error function solution is made up of a sum of Gaussian delta function solutions.

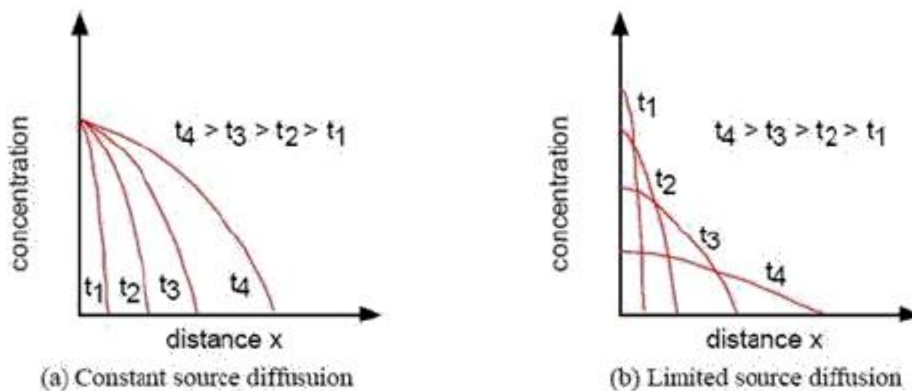


Fig. 2.10 : Diffusion profiles in constant source and limited source cases

Diffusion Systems –ANALYTIC SOLUTION TO FICKS DIFFUSION EQUATIONS

The diffusion process is carried out in systems called diffusion furnaces which may be either sealed or open tube type. The silicon wafer and the dopants are placed into the quartz tube before sealing and subsequently after the high temperature process; the tube is broken to remove the diffused wafers. Thus the sealed tubes are for one time use and can be maintained contamination free. The open tube method is preferred because multiple diffusion runs can be carried out. Separate tubes for each type of dopants are maintained because tubes are contaminated with the respective impurity. The insertion of wafers and solid dopant sources are done from one end of the tube and the other end is used for the flow of gases or impurities in vapour form. The diffusion furnaces are usually operated at temperatures 600C-1200°C and constant temperature zones are maintained with a ± 0.5°C tolerance.

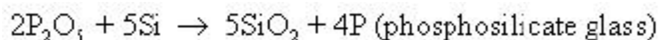
Three main sources of dopants such as gaseous, liquid and solids source of dopant atoms are used in semiconductor doping processes. The gaseous sources are most widely used since they are reliable and convenient. The typical examples are BF_3 (diborane), PH_3 (phosphine) and AsH_3 . Source gas reacts with oxygen at the wafer surface to form a dopant oxide, the dopant then diffuses from the oxide into silicon resulting in a uniform dopant concentration across the surface.

Liquid sources are available in two forms: Bubblers and spin on dopants. Bubblers convert the liquid to vapour, which then reacts with oxygen to form dopant oxides on the wafers. The spins on dopants are solutions which on drying form doped SiO_2 layers. The typical examples of liquid sources are $POCl_3$ and BBr_3 (boron tribromide).

Solid sources are available in two forms; such as tablet/granular and disc/wafer form. BN discs are most commonly used oxidized at 750 - 1100°C to serve as the boron diffusion source. Dopants are introduced into the silicon substrate using a two step, high temperature process. The first diffusion (predeposition) introduces dopants into the wafer. The second diffusion (drive) redistributes the dopants and allows the dopants to diffuse into the wafer more deeply (up to ~3 micrometers). The goal of the dopant predeposition diffusion is to move dopant atoms from a source to the wafer, and later allow the dopants to diffuse into the wafer. In order for the dopants to move into the silicon, they must be given energy, usually in the form of heat. In order for the diffusion to occur in a reasonable time, the temperature must be very high (900ºC <T<1200º). At this temperature the dopant (in the form of an oxide) reacts with the exposed silicon surface to form a highly doped glass. It is from this glass that the dopants can then diffuse into the wafer.

Arsenic has a low misfit factor ($\xi \cong 0$) and can be incorporated to a high concentration without causing strain the lattice. Arsenic's low diffusivity (about one-tenth of boron or phosphorous) and its abrupt doping profile makes it ideally suited for shallow diffused structures where more precise control of junction depth is needed. Phosphorus also has a high solid solubility and active carrier concentration close to arsenic. Its misfit factor also is small ($\xi \cong 0.068$). these characteristics makes its very useful for the fabrication of n+ regions in MOS and bipolar circuits.

Phosphorus, in the form of P_2O_5 , diffuses from the source to the wafer. When it reaches the silicon, it will undergo the following chemical reaction:



$$2P_2O_5 + 5Si \rightarrow 5SiO_2 + 4P \text{ (phosphosilicate glass)}$$

The phosphosilicate glass (PSG) is in contact with the silicon surface. A concentration gradient is formed, and since this process occurs at high temperature, diffusion will occur. The surface concentration will be fixed at the solid solubility, and the distribution will be in the form of a complimentary error function :

$$N(x,t) = N_{s1}\text{erfc}\left\{\frac{x}{\sqrt{4Dt}}\right\}$$

where :

N = concentration (cm^{-3})
N_{s1} = solid solubility limit for dopant (cm
x = position inside silicon relative to the surface
D = diffusion coefficient for dopant (cm^2/s)
t = time (seconds)

$$N(x,t) = N_{s1}\text{erfc}\left\{\frac{x}{\sqrt{4Dt}}\right\}$$

where :

N = concentration (cm^{-3})
N_{s1} = solid solubility limit for dopant (cm^{-3})
x = position inside silicon relative to the surface
D = diffusion coefficient for dopant (cm^2/s)
t = time (seconds)

$$N(x,t) = N_{s1}\,\mathrm{erfc}\left\{\frac{x}{\sqrt{4Dt}}\right\}$$

where :

N = concentration (cm^{-3})
N_{s1} = solid solubility limit for dopant (cm
x = position inside silicon relative to the surface
D = diffusion coefficient for dopant (cm^2/s)
t = time (seconds)

Boron Diffusion from a solid source to form a BJT base: a case study

Boron is the most commonly used p-type dopant for silicon, principally because it is the only column III element that can be masked by SiO_2. The first step, called the predeposition, is an open tube diffusion process that involves the gaseous transfer of a compound containing the dopant to the Si wafer. The gas may be supplied in several different ways, but in almost all cases the final chemical reaction when it reaches the silicon is:

Boron, in the form of B_2O_3, diffuses from the source to the wafer. The surface concentration during such constant source diffusion is usually limited by the solid solubility of the dopant in silicon. The boron predeposition transfers boron from a solid source to the wafer. The borosilicate glass (BSG) is in contact with the silicon surface. A concentration gradient is formed, and since this process occurs at high temperature, diffusion will occur. If the boron source used is a solid source of boron nitride wafers, Si wafers are stacked next to oxidized wafers as shown. At the pre-deposition temperature, concentration gradient induced diffusion is established between the source wafers and the silicon wafers.

This concentration gradient is a function of the distance between the source and silicon, the temperature of the pre-deposition, and the composition of the gas ambient (oxidizing or non-oxidizing, presence of H_2O, etc.). Ideally, however, it is not a function of gas flow rate, which is low enough so that the gas between source and wafer is essentially stagnant. The concentration gradient results in the transfer of B_2O_3 to the silicon surface, and so produces a very thin, very highly doped region at the silicon wafer surface. The final level of doping and resulting sheet resistance is determined by time and the size of the concentration gradient. Note, however, that the pre-deposition temperature is so low that very little boron diffusion into the silicon actually takes place.

After the predeposition diffusion the dopants are situated close to the surface of the wafer. However, they must diffuse even farther to lower the overall concentration in order for some of the devices to work properly. In addition, oxygen and water vapor are introduced during the drive diffusion to grow a new oxide over the areas which were exposed to bare silicon during the photolithography process. This new oxide can be patterned again so that other selective diffusion processes can be performed to create other types of devices.

The second stage of the diffusion process, called the drive-in, has as its objective the redistribution of the dopant deeper into the silicon, and is referred to as a constant source diffusion. For this process we first remove our silicon wafer from the pre-deposition furnace and strip off the SiO_2 - B_2O_3 glass using HF. The HF actually attacks the SiO_2, not the B_2O_3, so care must be used in the pre-deposition step to prevent the deposition of excessive amounts of B_2O_3. This is done mainly by assuring no H_2O is present in the furnace, which would greatly increase the rate of transfer of B_2O_3 to the silicon. Also note the thin SiB_x phase on the wafer is hydrophilic, so the wafer will not de-wet after etching. The actual drive-in is performed at very high temperatures (~1100°C). The drive-in is usually initiated in an oxidizing environment; the initial growth of SiO_2 prevents the out-diffusion of the boron during the rest of the process.

The boron concentration near the surface after predeposition is too high and the junction depth is too shallow to act as a good base. After the BSG is removed from the surface of the wafer, a 'sourceless' diffusion (drive) will lower the surface concentration and simultaneously drive the dopant deeper into the wafer. The distribution can be approximated as a Gaussian profile:

$$N(x,t) = N_0 e^{\frac{x^2}{4Dt}} \qquad \text{and} \qquad N_0 = \frac{2N_{S1}}{\pi}\sqrt{\frac{D_1 t_1}{D_2 t_2}}$$

N = concentration (cm^{-3})
N_0 = surface concentration (cm^{-3})
x = position inside silicon relative to the surface
D = diffusion coefficient for dopant (cm^2/s)
t = time (seconds)

and

N_0 = surface concentration after the drive (cm^{-3})
N_{01} = surface concentration after the drive (cm^{-3})
D_1 = diffusion coefficient f of dopant during the predep (cm^2/s)
t_1 = time of the predep (seconds)
D_2 = diffusion coefficient f of dopant during the drive (cm^2/s)
t_2 = time of the drive (seconds)

In addition to redistribution of the dopant, another oxide layer is grown which may act as a mask for the Further process.

If a dopant is diffused at temperature T_1 for time t_1 and then is diffused at temperature T_2 for time t_2 etc. the total effective Dt is given by the sum of all the individual Dt products.

$$Dt_{effective} = \sum Dt = Dt_1 + Dt_2 + Dt_3 + \ldots\ldots\ldots\ldots$$

Some of the Dt steps may be negligible in a process. The Gaussian solution only holds if the Dt used to introduce the dopant is small compared with the final Dt for the drive-in i.e. if an initial delta function approximation is reasonable.

Example 5 : In a bipolar transistor, if the emitter profile is formed by a predeposition and the base profile by an implant + drive-in, then the junction for the emitter-base occurs when

$$N_0 erfc\left(\frac{x}{\sqrt{4Dt}}\right) = \frac{Q}{\sqrt{\pi Dt}} \exp\left[-\frac{x^2}{4Dt}\right]$$

Emitter Dt

Base - Emitter Dt

And the emitter-base occurs when

$$\frac{Q}{\sqrt{\pi Dt}} \exp\left\{-\frac{x^2}{4Dt}\right\} = N_B$$

Example 6 : Boron Diffusion from a solid source to form well: a case study

We can now consider how to design a boron diffusion process (say for the well or tub of a CMOS process) such that

$\rho_s = 900$

$x_j = 3\mu m$

N_{BC} (substrate concentration) $= 10^{15} cm^{-3}$

The average conductivity of the layer is

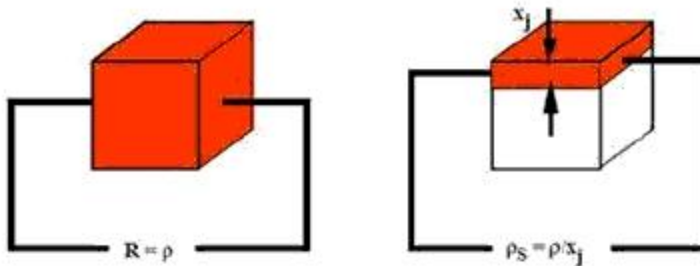


Fig. 2.11 : Resistivity of bulk and surface charges

The resistivity of a cube is given by

$$J = nqv = nq\mu E = \frac{1}{\rho}E$$

$$\rho = \frac{E}{J}\Omega \text{ cm}$$

The sheet resistance of a shallow junction is

$$R = \frac{\rho}{x_j} = \rho_s \left(\text{with units } \frac{\Omega}{\square}\right)$$

For a non-uniformly doped layer

$$\rho_s = \frac{\rho}{x_j} = \frac{1}{q\int_0^{x_j}\left[n(x) - N_B\right]\mu\left[n(x)\right]dx}$$

Sheet resistance can be experimentally measured by a four point probe technique.

This equation has been numerically integrated by Irwin for different analytical profiles (for example P type Gaussian profiles). Such plots with surface concentration (cm^{-3}) verses effective conductivity (Ω - cm)$^{-1}$ are called Irvin's curves

$$\bar{\sigma} = \frac{1}{\rho_s x_j} = \frac{1}{(900\Omega/\square)(3x10^{-4}cm)} = 3.7\ (-cm)^{-1}$$

From the Irvin curve we obtain that $N_s = 4 \times 10^{17}\ cm^{-3}$

Assuming a Gaussian profile after the drive in

$$N_{BC} = \frac{Q}{\sqrt{\pi Dt}} \exp\left\{-\frac{x_j^2}{4Dt}\right\} = N_S \exp\left(-\frac{x_j^2}{4Dt}\right)$$

Ion Implantation

Ion implantation is the introduction of atoms into a solid substrate by bombardment with ions in the keV to MeV energy range. It was early recognized that this technique provides the possibility of introducing a wide selection of atomic species and also external control of both the number of atoms per cm^2 introduced into the semiconductor and the depth distribution of the implanted species. This is in contrast to diffusion processes where temperature governs the surface concentration and time and temperature determine the depth distribution. The initial impact of ion implantation in silicon technology was in charge control. This in combination with the fact that simple masking techniques can be used to define the geometrical area of the implanted region led to extensive to MOS technology. Presently implantation is the most common method of introducing dopant atoms for the fabrication of Si devices.

Advantages

This process provides excellent dose control. Typical 10^{11} to 10^{16} cm^{-2}. It has been possible to control the number of dopant atoms per cm^2 to about 1% over the surface of a wafer, and to control the number within 3% from wafer to wafer.

The ion implantation results in less lateral diffusion i.e. diffusion of dopant atoms in the direction parallel to the surface of semiconductor compared to the vertical diffusion where diffusing species are moving in a direction normal to the wafer surface.

The dopant elemental selection can be easily carried out by mass separation from impure sources and this helps in preventing contamination.

The room temperature process allows resist to be used and it allows doping anytime during process. Modeling is very accurate

Complex doping profiles can be made using multiple energy implants

Highly abrupt junctions can be made

Highly automatic parameters

Limitations

Because ion implantation causes damage to the crystal structure of the target which is often unwanted, ion implantation processing is often followed by a thermal annealing. This can be referred to as damage recovery. As the energetic ions are introduced in the Si lattice, many Si atoms are knocked off their lattice positions resulting in large numbers of Si interstitial atoms and vacant lattice sites (Frenkel pairs). Since the as-implanted dopants are usually electrically non-active and the lattice damage deteriorates the device performance, it is necessary to have the annealing of the implanted sample at temperatures high enough to permit the dopant atoms to incorporate into substitutional sites and thus become electrically active, and for the defects to annihilate and thus repair the damage.

The diffusion of impurities in implanted Si during thermal anneal is complicated as a result of the presence of implantation damage. As an example, the diffusivity of B in implanted crystalline Si is anomalously high compared to equilibrium values. This phenomenon has important consequences for Si processing, since it causes the dopant profile to spread significantly compared to the as-implanted profile during the subsequent annealing steps required for dopant activation. The presence of damage also causes dopant clustering and thus prevents complete electrical activation. The down-scaling of Si CMOS devices requires simultaneously the formation of ultra shallow junctions for the source and drain extensions and a low sheet resistance (high carrier concentration). Conventionally, there is a compromise involved in maximizing dopant electrical activation while minimizing dopant diffusion.

The amount of crystallographic damage can be enough to completely amorphize the surface of the target: i.e. it can become an amorphous solid. However in some cases, complete amorphization of a target is preferable to a highly defective crystal: An amorphized film can be regrown at a lower temperature than required to anneal a highly damaged crystal.

Implantors are very expensive machines, requiring heavy maintainace and upkeep. The Safety hazard precautions are also stringent because of the radiation, high energy beams and high voltages.

Sputtering : Some of the collision events result in atoms being ejected from the surface, and thus ion implantation will slowly etch away a surface. The effect is only appreciable for very large doses.

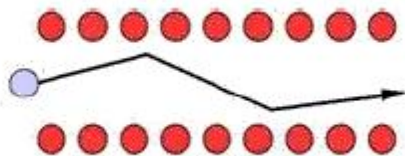


Fig. 2.12 : Basic illustration of channelling

Ion channeling : Since the semiconductor substrates have crystal structure, particular crystallographic directions offer much lower stopping than other directions. The result is that the range of an ion can be

much longer if the ion travels exactly along a particular direction, for example the <110> direction in silicon and other diamond cubic materials as shown in Fig. 3.43. This effect is called ion channelling, and, like all the channelling effects, is highly nonlinear, with small variations from perfect orientation resulting in extreme differences in implantation depth. For this reason, most implantation is carried out a few degrees off-axis, where tiny alignment errors will have more predictable effects. Ion channelling can be used directly in Rutherford backscattering and related techniques as an analytical method to determine the amount and depth profile of damage in crystalline thin film materials.

A comparison of Solid/Gas Phase Diffusion verses Ion Implantation and Annealing is provided in Table 10.

Table 2.2 Diffusion verses Ion Implantation

Diffusion	Ion Implantation
	Room temperature mask
	Precise dose control
No damage created by doping	Wide range of doses
Batch fabrication	Accurate depth control
Usually limited to solid solubility	Implant damage enhances diffusion
Low surface concentration hard to chieve without a long drive-in	Dislocations caused by damage may cause junction leakage
Low dose predeposition very difficul	Implant channeling may affect profile
	Lowthroughput
	Junction depth limited to 1 micron

RANGE THEORY

Impurity Profiles : Range and Depth Distribution

There are two basic mechanisms by which energetic ions are stopped in a substrate. Ions interact with electrons and nuclei of substrate atoms and lose energy during these collisions. The range (R) is the total distance traveled by an ion before resting. Please note that the path is not a straight line. The Nuclear collisions transfer energy away from ion and the coulombic interaction with electrons transfer energy to electrons.

If S_n (E) is energy loss per unit length due to nuclear stopping and S_e (E) is energy loss per unit length due to electronic stopping, the total rate of energy loss ($\frac{dE}{dx}$) is the sum of the two losses:

$$\frac{dE}{dx} = S_n(E) + S_e(E)$$

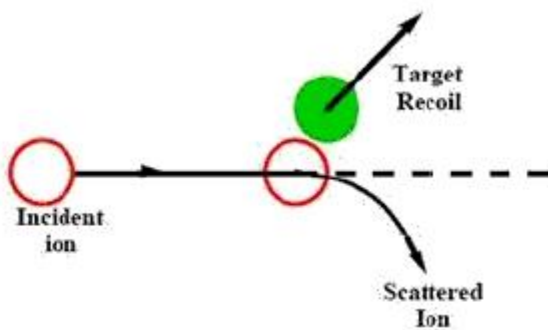


Fig.2.13 : Ion Nuclei interactions

If one consider ions as “balls” with momentum p=mv the nuclear stopping is provided by velocity interaction. If the rate of energy loss is plotted as a function of ion energy, it can be seen that the loss increases linearly at low energy because the ion velocity is smaller, the loss reaches maximum and decreases at higher energy because the higher velocity makes ion move too quickly to interact. S_n (E) increases with ion mass (p=mv) and if ions are light (e.g. Boron), its energy loss/unit length ($\frac{dE}{dx}$) is much lower.

S_e (E) is energy loss per unit length due to electronic stopping. Drag force caused by charged ion in "sea" of electrons (non-local electronic stopping).

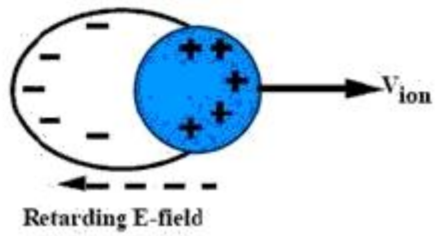


Fig. 2.14 : Ion electron interactions

Collisions with electrons around atoms transfers momentum and results in local electronic stopping .

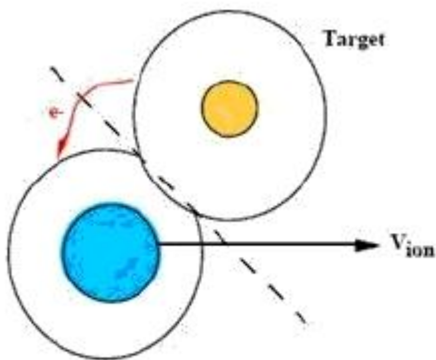


Fig. 2.15 : Electron Electron interactions

This is similar to stopping a projectile in viscous medium:

$$S_e(E) = k\sqrt{E}$$

k depends on ion and target mass and atomic number. The lighter ion, boron, has larger $\frac{dE}{dx}$. The crossover point between electronic and nuclear stopping is an important parameter. The critical energy E_c when the nuclear and electronic stoppings are equal is

B: 17keV
P: 150keV
As and Sb: > 500keV

That is for B, $S_e(E)$ dominates down to 10 keV and for As, nuclear stopping dominates up to 700 keV.

The range (R) is:

$$R = \int_0^R dx = \frac{1}{N}\int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)}$$

A more significant parameter of interest is the projected range (R_P) and it is the distance travelled normal to the surface as shown in Fig. 3.47.

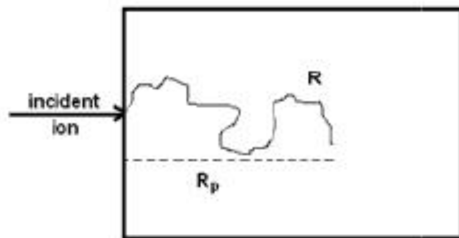


Fig. 2.16 : Range (R) and projected range (R_P) in ion implantation

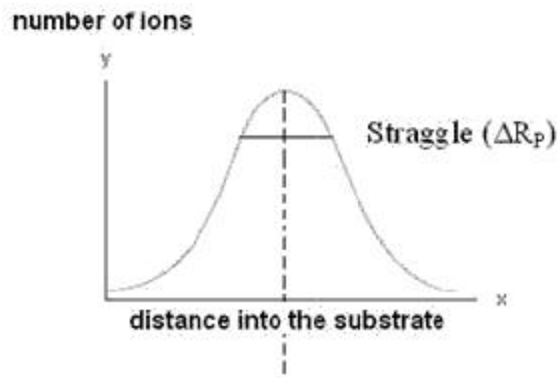


Fig 2.17 guassian curve

The peak concentration occurs where *x*= R_p

$$N(x = R_p) = N_p = \frac{Q_0}{\sqrt{2\pi}\Delta R_p}$$

where Q_0 is the ion dose (cm^{-2}), N_p is the maximum (desired) concentration, R_p is the range or desired depth, and ΔR_p is the straggle or standard deviation. The projected range and straggle can be approximated from range:

$$R_p = \frac{R}{1 + \frac{M_2}{3M_1}}$$

$$\Delta R_p = \frac{2R_p\sqrt{M_1 M_2}}{3(M_1 + M_2)}$$

Where M_1 is the ion mass and M_2 is the substrate mass.

A wide variety of masking materials can be used during the implantation and the thickness of the mask can be estimated using the above equations once the range (or stopping) parameters of the mask

material is known. The important goals of damage annealing are:

Remove primary damage created by the implant and activate the dopants.

Restore silicon lattice to its perfect crystalline state.

Restore the electron and hole mobility.

Do this without appreciable dopant redistribution.

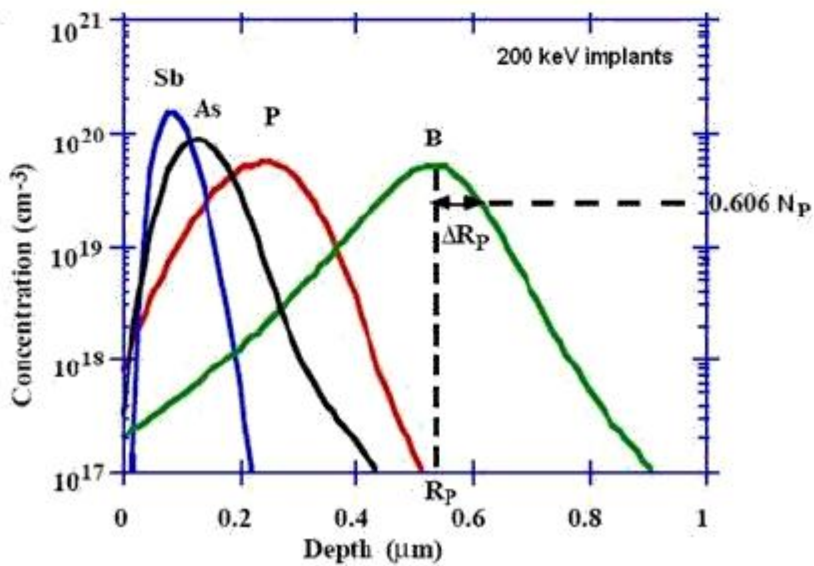


Fig. 2.18 : B, P, As and Sb implanted concentration profile in Si

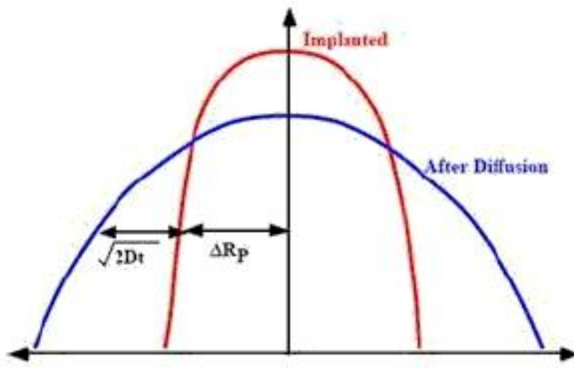


Fig. 2.19: Gaussian profile in the case of implantation followed by thermal anneal

$\sqrt{Dt} = 0.02\mu\text{m}$, using the expression

$$N(x) = \frac{Q_0}{\sqrt{2\pi(\Delta R_p^2 + Dt)}} \exp\left[-\left(\frac{(x - R_p)^2}{2(\Delta R_p^2 + Dt)}\right)\right]$$

,the peak concentration and junction depth can be calculated; $N_P = 6.6x10^{18} cm^{-3}$ and junction depth $(x_j) = 0.435\ \mu m$

Ion Implantation systems

Ion implanters are usually specified in terms of two key parameters; the ion energy which determines the penetration depth and the ion current which decides the dose. The basic building blocks of an implanter are ion sources, ion extractors, accelerators to high energies and beam manipulators as shown in Fig.3.51. Ions are produced by passing the gas vapour through a discharge chamber. The electrons are accelerated towards an anode which is typically at 100V. A magnetic field is provided to force electrons into a spiral trajectory and this will enhance the ionisation efficiency. The positive ions are moved out of the discharge chamber by means of another anode biased at 15-20 kV. The pressure in the remaining part of the system is kept below 10^{-6} Torr to minimize the ion scattering.

The output of an ion discharge chamber consists of many species in addition to the contamination produced by sputtering from its walls. Thus a mass separator is required to make source pure. The mass separation is based on the dynamics of charged particles, of mass (m) and velocity (v), moving at right angles to a uniform magnetic field (a flux density of B). These particles will experience a force (F) such that $F = q(v \times B)$. This force tends ions to move in a circular path of radius r, and experience a

centrifugal force $\frac{mv^2}{r}$. These forces are equal and opposite. The velocity of ions is related to the accelerating potential (V) by $\frac{1}{2}mv^2 = qV$. Combining these equations, the radius of the

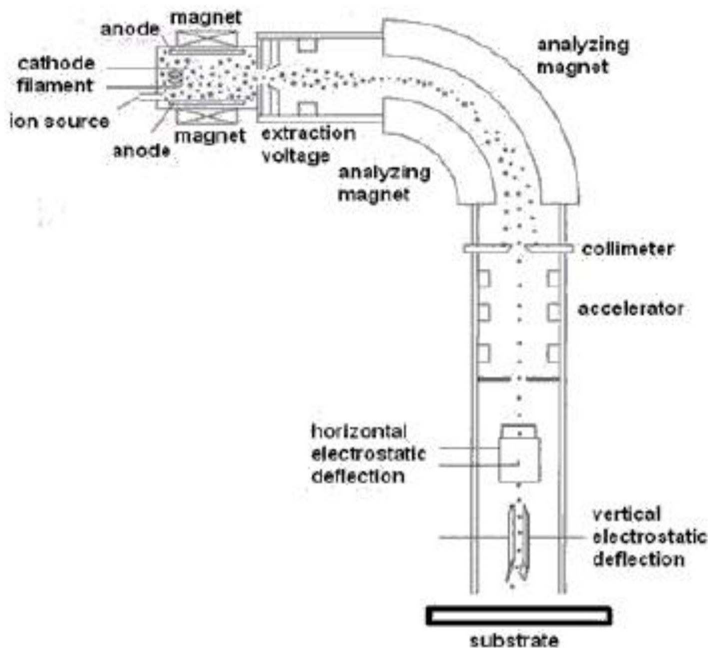


Fig. 2.20 : Schematic diagram of a basic implantation system

ion path is given as $r = \frac{1}{B}\sqrt{\left(\frac{2mV}{q}\right)}$

For a given extraction voltage and magnetic flux density, the r is proportional to square root of mass and thus ions of any particular mass can be selected by appropriately placing an exit slit. The collimators filter a stream of rays so that only those travelling parallel to a specified direction are allowed through. They help to shape the beam of radiation emerging from the machine and can limit the maximum field size of a beam. The surviving ions move into the accelerator tube where the ion beams gets energised under the very high acceleration potential. The output of the tube is usually maintained at ground potential for safety reasons.

In order to implant in the x and y directions the wafer surface, electrostatic and mechanical beam scanning techniques are used. The electrostatic scanning in both x and y directions is the simplest, and has advantage of very high scanning speed. However, when the beam voltages are 50 keV or higher, beam expansion (blow up) takes place and so mechanical scanning is usually used. In the mechanical scanning system the beam is stationary and the substrate wafer is moved.