Impurity Diffusion in Semiconductors

In the practical fabrication of solid-state electronic devices, it is generally necessary to introduce controlled amounts of various shallow level impurities, *i.e.*, dopants (B, P, or As), into particular regions within the silicon crystal. Indeed, boundaries between regions inside the volume of the wafer for which extrinsic doping changes from *p*-type to *n*-type or vice-versa form electrically active structures called *pn-junctions*. (Along with MOS capacitors, *pn*-junctions are the most important fundamental components of solid-state devices.) In general, although the wafer may have some uniform background doping added to the original melt during manufacture of the substrate itself, it is usual for additional dopants to be introduced through the surface of wafer. These are commonly restricted to specific laterally defined regions of the wafer surface by some type of mask, *i.e.*, one type of dopant might be introduced into some particular area (or areas) and other types of dopants introduced elsewhere. In any case, the vertical and lateral distribution of these dopant atoms may be precisely manipulated by carefully controlled diffusion. Such diffusion processes are thermally activated and, thus, are carried out in quartz tube furnaces very similar to those used for thermal oxidation. (However, the atmosphere inside the furnace generally will be inert or reducing rather than oxidizing.)

Linear Transport Processes

Diffusion of shallow level dopants in semiconductors, *e.g.*, silicon, is a specific example of a broad class of physical processes called *transport processes*. Other examples are conduction of heat and electricity and viscous fluid flow. Physically, transport processes are characteristic of physical systems which are not in thermodynamic equilibrium. Indeed, from a theoretical point of view, transport processes are dissipative in nature, which when occurring within some physical system, proceed to establish the system in equilibrium at which time any net transport comes to a halt. (The general study of transport processes and the approach to equilibrium is called *non-equilibrium thermodynamics*.) Conventionally, transport processes are considered within the context of a *linear phenomenology*, which means that they are described by expressions of the generic form:

$$J_a = \mathrm{L}_{ab} X_b$$

Here, J_a is defined as *flux* (or, more generally, a *flux vector*) identified with transport of some physical property, *a*, *e.g.*, mass, momentum, energy, charge, *etc.* Similarly, X_b is defined as *driving* or *thermodynamic force*, identified with a disequilibrium in some physical property, *b*, *e.g.*, gradients of concentration, fluid velocity, temperature, electrical potential, *etc.* Physically, a thermodynamic force quantifies the magnitude of any disequilibrium driving net transport processes. Thus, fluxes and forces are related by the parameter, L_{ab}, called a *phenomenological transport coefficient*. In the most general formulation (as above), fluxes and forces formally appear as column vector components and transport coefficients as square matrix elements. This allows for the possibility that a force in one physical property, *b*, may drive a flux in some different physical property, *a*. Indeed, such "cross effects" are commonly observed. Representative examples are

provided by thermal diffusion or thermoelectric effects in which case a temperature gradient drives material or electrical transport respectively. Clearly, cross effect transport coefficients correspond to off-diagonal matrix elements as defined by the preceding general expression. Of course, "direct effects" for which a force in a physical property drives a flux in the same property correspond to diagonal matrix elements and, therefore, are generally more important than cross effects. Obviously, ordinary diffusion, heat and electrical transport, viscous fluid flow, *etc.* provide elementary examples of just such processes. Therefore, to describe impurity diffusion in semiconductors, it is only necessary to consider direct effects, *i.e.*, only diffusive forces and fluxes. In such a case, the general matrix expression can be simplified to a simple linear proportionality:

$$J_a = \mathrm{L}_a X_a$$

In this expression, the direct effect transport coefficient, L_a, represents an ordinary numerical quantity rather than a matrix element. Clearly, for impurity diffusion, *a* corresponds to some impurity species, hence, L_a is identified as diffusivity of *a* (usually symbolized as D_a). In passing, it is, again, useful to observe that this same linear phenomenology can be applied to various specific physical situations that might superficially appear unrelated. Therefore, several specific cases are summarized as follows:

Ohm's Law of electrical conduction: $j = \sigma E = E/\rho$		
J = electric current density, j (units: A/cm^2)	X = electric field, $E = -\nabla V$ (units: V/cm) V = electrical potential	L = conductivity, $\sigma = 1/\rho$ (units: mho/cm) ρ = resistivity (Ω cm)
Fourier's Law of heat transport: $q = -\kappa \nabla T$		
J = heat flux, q (units: W/cm^2)	X = thermal force, $-\nabla T$ (units: °K/cm) T = temperature	L = thermal conductivity, κ (units: W/°K cm)
Fick's Law of diffusion: $F = -D\nabla C$		
J = material flux, F (units: /sec cm^2)	X = diffusion force, $-\nabla C$ (units: /cm^4) C = concentration	L = diffusivity, D (units: cm^2/sec)
Newton's Law of viscous fluid flow: $F_u = -\eta \nabla u$		
J = fluid velocity flux, F_u (units: /sec^2 cm)	X = viscous force, $-\nabla u$ (units: /sec) u = fluid velocity	L=viscosity, η (units: /sec cm)

Table 4: Summary of common linear transport phenomena

Here, the subscript *a* has been dropped and *J*, *X*, and L correspond to the simple linear transport relation:

$$J = \mathrm{L}X$$

Clearly, the preceding correspondences are useful because once one type of transport process is considered, *e.g.*, impurity diffusion, any results obtained can be immediately applied to other types of transport by the simple expedient of redefinition and/or substitution of the appropriate phenomenological parameters.

Limiting further consideration specifically to the case of impurity diffusion, one considers transport of impurity species through a hypothetical bar of some solid material, *e.g.*, semiconductor, having a uniform cross section. Furthermore, for additional simplicity, one assumes that the concentration of impurity species varies only along the length of the bar and is constant over any given cross section. Thus, assuming that transport fluxes remain constant over some small time interval, Δt, then the net change per unit time in the number of impurity atoms in a small volume element of width, Δx, located at a distance, x, from the end of the bar is given by the simple difference expression:

$$\frac{\Delta N}{\Delta t} = [F(x) - F(x + \Delta x)]A$$

Here, N is the number of impurity atoms found within the volume element, $F(x)$ is impurity flux, and A is the cross sectional area of the bar. This is illustrated by the following figure:

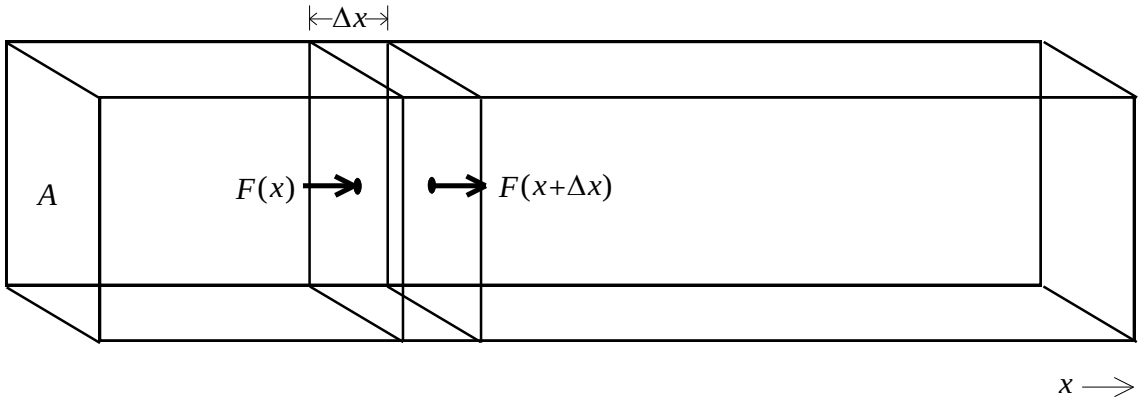

Fig. 49: Diffusion in a rectangular bar of constant cross section

If impurity concentration, C, is defined as usual as the quotient of N divided with volume, then one obtains:

$$\frac{\Delta C}{\Delta t} = -\frac{F(x + \Delta x) - F(x)}{\Delta x}$$

In the limit that the material volume element is allowed to become arbitrarily thin, the right hand side of this expression can just be identified as the negative of the derivative of

the material flux with respect to x. Similarly, the left hand side can be identified as the derivative of concentration with respect to time, hence:

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x}$$

Here, partial derivatives are written since C and F are functions of both position and time. Of course, material flux and concentration are related by Fick's Law, which in a single dimension has the form:

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

Here, D is the diffusivity of the impurity species. If the two preceding equations are formally combined, one obtains a single second order linear partial differential equation as follows:

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

This equation is conventionally called Fick's equation, Fick's Second Law, or just the diffusion equation. Clearly, it is a closed form expression for concentration as a function of position and time. (By convention, Fick's First Law is just the linear transport relation defined previously.) Furthermore, within the present context, the diffusion equation has been derived in a one dimensional form. For an elementary description of impurity diffusion in semiconductors this is adequate. However, in more complicated situations diffusion in more than one dimension must be considered. Obviously, Fick's Second Law can be generalized to all three dimensions just by replacing the second order partial derivative with the Laplacian:

$$\frac{\partial C}{\partial t} = D\nabla^2 C$$

In summary, Fick's Laws are useful for the description of diffusion of relatively dilute solutes. If the concentration becomes sufficiently high, due to interactions between solute atoms D may become dependent on the concentration, C. In this case, diffusion becomes non-linear and is much more difficult to treat mathematically.

Solution of Fick's Equation

Construction of a general solution of the diffusion equation in one dimension is quite straightforward. First of all, one must separate the space and time variables. This can be accomplished by assuming that the concentration, C, is a formal product of a function of position, $g(x)$, and a function of time, $f(t)$:

$$C(x,t) = g(x)\, f(t)$$

Upon substitution of this form, it follows that:

$$\frac{1}{Df}\frac{\partial f}{\partial t} = \frac{1}{g}\frac{\partial^2 g}{\partial x^2}$$

Clearly, all of the *x* dependence appears on the left hand side and all of the *t* dependence appears on the right hand side. Since the variables have now been separated, one can set each side equal to an unknown "separation constant". Therefore, the arbitrary constant, λ, is defined such that:

$$\frac{1}{Df}\frac{\partial f}{\partial t} = -\lambda^2 \quad ; \quad \frac{1}{g}\frac{\partial^2 g}{\partial x^2} = -\lambda^2$$

The form, $-\lambda^2$, is used purely for mathematical convenience. These two ordinary differential equations are easily integrated by elementary methods. In the case of the time equation one has:

$$\ln f = -D\lambda^2 t + \ln \gamma_t$$

$$f = \gamma_t e^{-D\lambda^2 t}$$

Here, γ_t is an unknown constant. In the case of the space equation, one immediately recognizes that the solution can be expressed as either a sine or a cosine:

$$g = \gamma_{x1} \sin \lambda x \quad ; \quad g = \gamma_{x2} \cos \lambda x$$

However, it is more convenient to express this in equivalent form as a complex exponential:

$$g = \gamma_x e^{i\lambda x}$$

Here, γ_x is a second unknown constant. Thus, it follows that a particular solution of the diffusion equation, C_λ, can be written as follows:

$$C_\lambda = \alpha e^{-D\lambda^2 t} e^{i\lambda x}$$

In this expression, α is the product of the arbitrary constants γ_t and γ_x and, thus, is itself an unknown constant. Obviously, the separation constant, λ, labels each particular solution.

It is well known that particular solutions of linear differential equations satisfy the Principle of Superposition. Simply stated, this means that if any two functions are independent solutions of some differential equation, then the sum (or difference) of the two is also a solution. Thus, if one considers λ to be a continuous variable, it follows that a general solution of the diffusion equation can be written as an integral, *i.e.*, a limiting sum, over all particular solutions:

$$C(x,t) = \int_{-\infty}^{\infty} d\lambda \alpha(\lambda) e^{-D\lambda^2 t} e^{i\lambda x}$$

Here, α is now treated as an unknown function of λ. It is desirable to express $\alpha(\lambda)$ in terms of some initial condition, $C_0(x)$, defined such that:

$$C_0(x) = C(x,0) = \int_{-\infty}^{\infty} d\lambda \alpha(\lambda) e^{i\lambda x}$$

Upon inspection, one observes that $C_0(x)$ is just the ordinary Fourier transform of $\alpha(\lambda)$. Furthermore, Fourier transformation is easily inverted, therefore, $\alpha(\lambda)$ can be written explicitly as follows:

$$\alpha(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dx C_0(x) e^{-i\lambda x}$$

Clearly, the "inverse Fourier transform" is identical to the "forward Fourier transform" except that a factor of $1/2\pi$ appears. (These forms can be made exactly identical if one removes a factor of $\sqrt{2\pi}$ from the denominator of the above expression in which case one finds that $C_0(x)$ and $\alpha(\lambda)\sqrt{2\pi}$ define a formal "Fourier transform pair".) Thus, substitution of the above result into the expression for $C(x,t)$ yields:

$$C(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\lambda \int_{-\infty}^{\infty} dx' C_0(x') e^{i\lambda(x-x')} e^{-D\lambda^2 t}$$

This expression can be further simplified by completion of the square in the exponent:

$$C(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dx' C_0(x') e^{-(x-x')^2/4Dt} \int_{-\infty}^{\infty} d\lambda e^{-Dt\left(\lambda - \frac{i(x-x')}{2Dt}\right)^2}$$

The integral over λ can be determined using standard methods (*e.g.*, complex contour integration), however it is essentially an integral over a Gaussian function and is found to be equal to $\sqrt{\pi/Dt}$:

$$C(x,t) = \frac{1}{2\sqrt{\pi Dt}} \int_{-\infty}^{\infty} dx' C_0(x') e^{-(x-x')^2/4Dt}$$

Clearly, impurity concentration at any time, *t*, is completely determined by the initial concentration. Indeed, this is a completely general result and is applicable for any initial concentration, $C_0(x)$.

Instantaneous Source

A specific form for $C_0(x)$ that is of particular interest can be represented mathematically as follows:

$$C_0(x) = 2N\delta(x - x_0)$$

Here, $\delta(x - x_0)$ is a Dirac delta function, which is defined to be zero everywhere except in the case that $x = x_0$ where it becomes infinite. Furthermore, the integral of $\delta(x - x_0)$ over *x* is finite and equal to one. This is called an *instantaneous source*. An instructive way to view a delta function is as a normalized Gaussian function, which has a standard deviation of zero. Therefore, this initial concentration corresponds to an infinitely thin sheet of impurity located at a position, x_0. Within this context, the coefficient, *N*, is the number of dopant atoms per unit area of the sheet and is called *dose*. (The factor of 2 included in the definition of $C_0(x)$ is a geometrical factor which accounts for the fact that a wafer is, perhaps, better regarded as a semi-infinite diffusion domain rather than an infinite domain.) Thus, $C(x,t)$ can be trivially determined if one substitutes the preceding form for $C_0(x)$:

$$C(x,t) = \frac{1}{2\sqrt{\pi Dt}} \int_{-\infty}^{\infty} dx' (2N\delta(x' - x_0)) e^{-(x-x')^2/4Dt} = \frac{N}{\sqrt{\pi Dt}} \int_{-\infty}^{\infty} dx' \delta(x' - x_0) e^{-(x-x')^2/4Dt}$$

Integration over x' is trivial due to the delta function, hence:

$$C(x,t) = \frac{N}{\sqrt{\pi Dt}} e^{-(x-x_0)^2/4Dt}$$

Clearly, an instantaneous source results in a *Gaussian concentration profile*. (The terminology "concentration profile" is generally used to describe dependence of impurity concentration in a one dimensional sense.)

In practice, a Gaussian concentration profile describes an impurity diffusion process for which an "infinitely thin" initial layer of dopant, *i.e.*, shallow level impurity, is deposited on the wafer surface. This surface deposition is followed by diffusion at elevated temperature for some time, *t*. Obviously, since x_0 is zero by definition, the concentration profile takes the simplified form:

$$C(x,t) = \frac{N}{\sqrt{\pi Dt}} e^{-x^2/4Dt} = C_s e^{-x^2/4Dt}$$

Here, C_s is *surface concentration* and is equal to $N/\sqrt{\pi Dt}$. Obviously, for t equal to zero, C_s is infinite just as one expects from the original delta function concentration profile.

Constant Source

A second initial concentration profile, which is generally useful for the description of impurity diffusion has the explicit form:

$$C_0(x) = 2C_s(1 - H(x - x_0))$$

Here, $H(x - x_0)$ is a Heaviside or unit step function and is formally defined to be equal to zero if $x<x_0$ and equal to one if $x>x_0$. Thus, $C_0(x)$ is equal to $2C_s$ if $x<x_0$ and equal to zero if $x>x_0$. This is called a *constant source*. Clearly, upon substitution the step function "cuts off" the integral above a value of x_0 and one obtains the result:

$$C(x,t) = \frac{C_s}{\sqrt{\pi Dt}} \int_{-\infty}^{x_0} dx' e^{-(x-x')^2/4Dt}$$

This integral is modified by defining a new integration variable x'' equal to $x - x'$, hence:

$$C(x,t) = \frac{C_s}{\sqrt{\pi Dt}} \int_{x-x_0}^{\infty} dx'' e^{-x''^2/4Dt}$$

The integral cannot be constructed in closed form but has a standard definition in terms of the error function, erf(x) or complementary error function, erfc(x):

$$C(x,t) = C_s\left(1 - \mathrm{erf}\left(\frac{x - x_0}{2\sqrt{Dt}}\right)\right) = C_s \mathrm{erfc}\left(\frac{x - x_0}{2\sqrt{Dt}}\right)$$

Hence, a constant source results in a *complementary error function concentration profile*.

In contrast to the previous case, this type of concentration profile describes impurity diffusion processes for which the surface of the wafer remains in equilibrium with some dopant source (solid, liquid, or gaseous) during exposure to elevated temperature, *i.e.*, during the diffusion process. Therefore, the surface concentration can generally be identified as the maximum solid solubility of the dopant in silicon. Hence, it is usual to set x_0 to zero and, thus, the concentration profile takes the form:

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

Clearly, the surface impurity concentration remains constant in the case of a complementary error function profile. In contrast, for a Gaussian profile, C_s decreases as $1/\sqrt{t}$. Furthermore, the total integrated amount of impurity present within the wafer is constant in the case of an instantaneous source; however, it continues to increase in the case of a constant source. Physically, this is easily understood if one observes that all impurity species diffused into the wafer are initially present in an instantaneous source. However, a constant source continues to introduce impurity atoms into the wafer surface during the diffusion process.