Chapter 23
Thermal Diffusion

23.1 Introduction

The fabrication of integrated circuits requires the introduction into the semiconductor material of atoms belonging to specifically-selected chemical species. Such atoms are called *impurities* or *dopants*. As shown in Chap. 18, the inclusion of dopants into the semiconductor lattice attains the important goals of fixing the concentration of mobile charges in the material and making it practically independent of temperature.

Dopants are divided into two classes, termed *n-type* and *p-type*. With reference to silicon (Si), the typical *n*-type dopants are phosphorus (P), arsenic (As), and antimony (Sb), while the typical *p*-type dopants are boron (B), aluminum (Al), gallium (Ga), and Indium (In). When a dopant atom is introduced into the semiconductor lattice, in order to properly act as a dopant it must replace an atom of the semiconductor, namely, it must occupy a lattice position. When this happens, the dopant atom is also called *substitutional impurity*. An impurity atom that does not occupy a lattice position is called *interstitial*. Interstitials can not properly act as dopants, however, they degrade the conductivity and other electrical properties of the semiconductor.

The concentration of the dopant atoms that are introduced into a semiconductor is smaller by orders of magnitude than the concentration of the semiconductor atoms themselves. As a consequence, the average distance between dopant atoms within the lattice is much larger than that between the semiconductor atoms. Thus, the material resulting from a doping process is not a chemical compound: it is still the semiconductor in which some of the electrical properties are modified by the presence of the dopant atoms. In fact, while the presence and type of dopants are easily revealed by suitable electrical measurements, they may remain undetectable by chemical analyses.

As a high-temperature condition is necessary to let the dopant atoms occupy the lattice position, during the fabrication of the integrated circuit the semiconductor wafer undergoes several high-temperature processes. This, in turn, activates the dopant diffusion.

M. Rudan, *Physics of Semiconductor Devices*,
DOI 10.1007/978-1-4939-1151-6_23

The chapter illustrates the diffusive transport with reference to the processes that are used for introducing impurities into a semiconductor in a controlled way. First, the expressions of the continuity equation and of the diffusive flux density are derived. These expressions are combined to yield the diffusion equation, whose form is reduced to a one-dimensional model problem. The model problem allows for an analytical solution, based on the Fourier-transform method, that expresses the diffused profile at each instant of time as the convolution of the initial condition and an auxiliary function.

Then, the solution of the model problem is used to calculate the impurity profiles resulting from two important processes of semiconductor technology, namely, the predeposition and the drive-in diffusion. In the last part of the chapter the solution of the model problem is extended to more general situations. Specific data about the parameters governing the diffusion processes in semiconductors are in [46, Chap. 3], [104, Chap. 10], [105, Chap. 7], [71, Chap. 12]. Many carefully-drawn illustrations of the diffusion process are found in [75, Sect. 1.5]. The properties of the Fourier transform are illustrated in [72, 118].

23.2 Continuity Equation

The continuity equation described in this section is a balance relation for the number of particles. Here it is not necessary to specify the type of particles that are being considered: they may be material particles, like molecules or electrons, particles associated to the electromagnetic field (photons), those associated to the vibrational modes of atoms (phonons), and so on. Although the type of particles is not specified, it is assumed that all particles considered in the calculations are of the same type.

The balance relation is obtained by considering the space where the particles belong and selecting an arbitrary volume V in it, whose boundary surface is denoted with S. The position of the volume is fixed. Let $\mathcal{N}(t)$ be the number of particles that are inside S at time t. Due to the motion of the particles, in a given time interval some of them move across S in the outward direction, namely, from the interior to the exterior of S. In the same interval of time, other particles move across S in the inward direction. Let $\mathcal{F}_{\text{out}}(t)$ and $\mathcal{F}_{\text{in}}(t)$ be the number of particles per unit time that cross S in the outward or inward direction, respectively, and let $\mathcal{F} = \mathcal{F}_{\text{out}} - \mathcal{F}_{\text{in}}$. The quantity $\mathcal{F}$, whose units are s^{-1}, is the *flux* of the particles across the surface S. If the only reason that makes $\mathcal{N}$ to change is the crossing of S by some particles, the balance relation takes the form of the first equation in (23.1). The minus sign at the right hand side is due to the definition of $\mathcal{F}$; in fact, $\mathcal{N}$ decreases with time when $\mathcal{F} > 0$, and vice versa.

Besides the crossing of the boundary S by some particles, there is another mechanism able to contribute to the time variation of $\mathcal{N}$, namely, the generation or destruction of particles inside the volume V. This possibility seems to violate some commonly-accepted conservation principle. However it is not so, as some examples given in Sect. 23.7.1 will show. As a consequence, the description of the particle generation or destruction must be included. This is accomplished by letting $\mathcal{W}_{\rm ge}(t)$ and $\mathcal{W}_{\rm de}(t)$ be the number of particles per unit time that are generated or, respectively, destroyed within the volume V. Defining $\mathcal{W} = \mathcal{W}_{\rm ge} - \mathcal{W}_{\rm de}$, the balance relation that holds when generation or destruction are present takes the form of the second equation in (23.1):

$$\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}t} = -\mathcal{F}, \qquad \frac{\mathrm{d}\mathcal{N}}{\mathrm{d}t} = -\mathcal{F} + \mathcal{W}. \tag{23.1}$$

The quantity $\mathcal{W}$, whose units are s^{-1}, is the *net generation rate* within volume V.

It is convenient to recast (23.1) in local form. This is done basing on the second equation of (23.1), which is more general, and is accomplished by describing the motion of the particles as that of a continuous fluid. Such a description is legitimate if V can be partitioned into equal cells of volume $\Delta V_1, \Delta V_2, \ldots$ having the following properties: (i) the cells can be treated as infinitesimal quantities in the length scale of the problem that is being considered and, (ii) the number of particles within each cell is large enough to make their average properties significant. If the above conditions are fulfilled one lets $\Delta V_k \to \mathrm{d}V$ and introduces the *concentration* $N(\mathbf{r}, t)$, such that $N\,\mathrm{d}V$ is the number of particles that at time t belong to the volume $\mathrm{d}V$ centered at position $\mathbf{r}$. Similarly, one defines the *net generation rate per unit volume* $W(\mathbf{r}, t)$ such that $W\,\mathrm{d}V$ is the net generation rate at time t within $\mathrm{d}V$. The units of N and W are m^{-3} and m^{-3} s^{-1}, respectively. From the definitions of N, W it follows that the number $\mathcal{N}(t)$ of particles that are inside S at time t is found by integrating $N(\mathbf{r}, t)$ over V, and that the net generation rate $\mathcal{W}(t)$ is found by integrating W over V.

To recast in local form the part of (23.1) related to the flux $\mathcal{F}$, one associates a velocity $\mathbf{v}(\mathbf{r}, t)$ to the concentration $N(\mathbf{r}, t)$. In general, such a velocity is different from the velocity of each individual particle that contributes to the concentration N. In fact, $\mathbf{v}$ is a suitable average of the particles' velocities, whose definition (6.6) is given in Sect. 6.2. In the elementary time $\mathrm{d}t$ the concentration originally in $\mathbf{r}$ moves over a distance $\mathbf{v}\,\mathrm{d}t$ in the direction of $\mathbf{v}$. As consequence, if $\mathbf{r}$ belongs to the boundary surface S, a crossing of S by the particles may occur, this contributing to the flux. To calculate the contribution to the flux at a point $\mathbf{r}$ belonging to S, one takes the plane tangent to S at $\mathbf{r}$ and considers an elementary area $\mathrm{d}S$ of this plane centered at $\mathbf{r}$ (Fig. 23.1). The construction implies that the surface S is smooth enough to allow for the definition of the tangent plane at each point of it.

Let $\mathbf{s}$ be the unit vector normal to $\mathrm{d}S$, oriented in the outward direction with respect to S. If $\mathbf{v}$ is normal to $\mathbf{s}$, no crossing of S occurs and the contribution to the flux at point $\mathbf{r}$ is zero. If the scalar product $\mathbf{v} \cdot \mathbf{s}$ is positive, the crossing occurs in the outward direction and contributes to $\mathcal{F}_{\rm out}$. Its contribution is found by observing that the elementary cylinder, whose base area and side are $\mathrm{d}S$ and, respectively, $\mathbf{v}\,\mathrm{d}t$, has a volume equal to $\mathbf{v} \cdot \mathbf{s}\,\mathrm{d}S\,\mathrm{d}t$. Due to the sign of $\mathbf{v} \cdot \mathbf{s}$, the cylinder is outside the surface S. The number of particles in the cylinder is found by multiplying its volume by the concentration $N(\mathbf{r}, t)$. Letting $\mathbf{F} = N\mathbf{v}$, such a number reads $\mathbf{F} \cdot \mathbf{s}\,\mathrm{d}S\,\mathrm{d}t$. As the particles that are in the cylinder at time $t + \mathrm{d}t$ were inside the surface S at time t, dividing the above expression by $\mathrm{d}t$ yields the elementary contribution of point $\mathbf{r}$

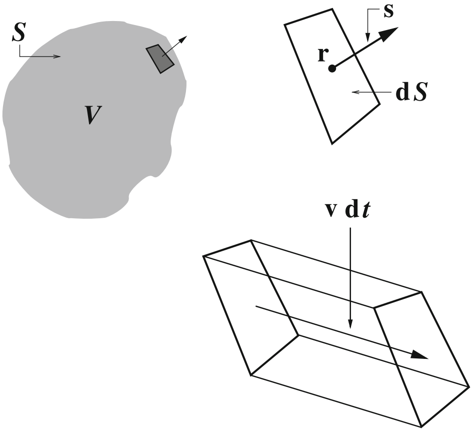

Fig. 23.1 Illustration of the symbols used in the calculation of the flux

to the flux, $\mathrm{d}\mathcal{F} = \mathbf{F} \cdot \mathbf{s}\,\mathrm{d}S > 0$. The contribution from a point $\mathbf{r}$ where $\mathbf{v} \cdot \mathbf{s} < 0$ is calculated in a similar way. The flux $\mathcal{F}$ is then found by integrating $\mathbf{F} \cdot \mathbf{s}$ over the surface S. The quantity $\mathbf{F} \cdot \mathbf{s} = \mathrm{d}\mathcal{F}/\mathrm{d}S$, whose units are $\mathrm{m}^{-2}\mathrm{s}^{-1}$, is the *flux density*.

Introducing the relations found so far into the second form of (23.1), and interchanging the derivative with respect to t with the integral over V, yields

$$\int_V \left(\frac{\partial N}{\partial t} - W \right) \mathrm{d}V = - \int_S \mathbf{F} \cdot \mathbf{s}\,\mathrm{d}S = - \int_V \mathrm{div}\mathbf{F}\,\mathrm{d}V. \tag{23.2}$$

The last equality in (23.2) is due to the divergence theorem (A.23), whereas the use of the partial-derivative symbol is due to the fact the N, in contrast with $\mathcal{N}$, depends also on $\mathbf{r}$. The procedure leading to (23.2) does not prescribe any constraint on the choice of the volume V. As a consequence, the two integrals over V that appear in (23.2) are equal to each other for any V. It follows that the corresponding integrands must be equal to each other, this yielding the *continuity equation*

$$\frac{\partial N}{\partial t} + \mathrm{div}\mathbf{F} = W, \qquad \mathbf{F} = N\mathbf{v}. \tag{23.3}$$

As mentioned above, (23.3) is the local form of the second equation of (23.1), which in turn is a balance relation for the number of particles. In the steady-state condition the quantities appearing in (23.3) do not depend explicitly on time, hence (23.3) reduces to $\mathrm{div}\mathbf{F} = W$. In the equilibrium condition it is $\mathbf{v} = 0$ and (23.3) reduces to the identity $0 = 0$. It is worth noting that in the equilibrium condition the velocity of each particle may differ from zero; however, the distribution of the individual velocities is such that the average velocity $\mathbf{v}$ vanishes. Similarly, in the equilibrium condition the generation or destruction of particles still occurs; however, they balance each other within any $\mathrm{d}V$.

To proceed it is assumed that the net generation rate per unit volume W, besides depending explicitly on $\mathbf{r}$ and t, may also depend on N and $\mathbf{F}$, but not on other functions different from them.

23.3 Diffusive Transport

The continuity Eq. (23.3) provides a relation between the two quantities N and $\mathbf{F}$ (or, equivalently, N and $\mathbf{v}$). If both N and $\mathbf{F}$ are unknown it is impossible, even in the simple case $W = 0$, to calculate them from (23.3) alone. However, depending on the specific problem that is being considered, one can introduce additional relations that eventually provide a closed system of differential equations. The important case of the *diffusive transport* is considered in this section.

It is convenient to specify, first, that the term *transport* indicates the condition where an average motion of the particles exists, namely $\mathbf{F} \neq 0$ for some $\mathbf{r}$ and t. The type of transport in which the condition $\mathbf{F} \neq 0$ is caused only by the spatial nonuniformity of the particles' concentration N is called *diffusive*. Simple examples of diffusive transport are those of a liquid within another liquid, or of a gas within another gas. They show that in the diffusive motion of the particles, the flux density is oriented from the regions where the concentration is larger towards the regions where the concentration is smaller.

The analytical description of the diffusion process dates back to 1855 [36]. Here the relation between $\mathbf{F}$ and N in the diffusive case is determined heuristically, basing on the observation that $\mathrm{grad}N$ is a sensible indicator of the spatial nonuniformity of N. Specifically it is assumed, first, that $\mathbf{F}$ depends on N and $\mathrm{grad}N$, but not on higher-order derivatives of N. The dependence on $\mathrm{grad}N$ is taken linear, $\mathbf{F} = \mathbf{F}_0 - D\,\mathrm{grad}N$, with $\mathbf{F}_0 = 0$ because $\mathbf{F}$ must vanish when the concentration is uniform. Finally, one remembers that the particles' flux density is oriented in the direction of a decreasing concentration, namely, opposite to $\mathrm{grad}N$. It follows that $D > 0$, so that the relation takes the form

$$\mathbf{F} = -D\,\mathrm{grad}N, \qquad D > 0. \tag{23.4}$$

The above is called *transport equation of the diffusion type*, or *Fick's first law of diffusion*. Parameter D is the *diffusion coefficient*, whose units are $\mathrm{m^2\ s^{-1}}$. From the derivation leading to (23.4) it follows that, if a dependence of $\mathbf{F}$ on N exists, it must be embedded in D. In the case $D = D(N)$ the relation (23.4) is linear with respect to $\mathrm{grad}N$, but not with respect to N. The diffusion coefficient may also depend explicitly on $\mathbf{r}$ and t. For instance, it depends on position when the medium where the diffusion occurs is nonuniform; it depends on time when an external condition that influences D, e.g., temperature, changes with time.

For the typical dopants used in the silicon technology, and in the temperature range of the thermal-diffusion processes, the experimentally-determined dependence on temperature of the diffusion coefficient can be approximated by the expression

$$D = D_0 \exp\left[-E_a/(k_B T)\right], \tag{23.5}$$

where k_B ($\mathrm{J\ K^{-1}}$) is the Boltzmann constant and T (K) the process temperature. In turn, the *activation energy* E_a and D_0 are parameters whose values depend on the material involved in the diffusion process. The form of (23.5) makes it more convenient to draw it as an *Arrhenius plot*, that displays the logarithm of the function using the inverse temperature as a variable: $\log D = \log(D_0) - (E_a/k_B)(1/T)$. At the diffusion temperatures, E_a and D_0 can often be considered independent of temperature. In this case the Arrhenius plot is a straight line (examples of Arrhenius plots are given in Chap. 24). At room temperature the diffusion coefficient of dopants in silicon is too small to make diffusion significant. In order to activate the diffusion mechanism a high-temperature process is necessary, typically between 900 and 1100°C.

23.4 Diffusion Equation—Model Problem

Inserting (23.4) into (23.3) yields the *diffusion equation*

$$\frac{\partial N}{\partial t} = \mathrm{div}(D\,\mathrm{grad}N) + W, \tag{23.6}$$

where W depends on $\mathbf{r}$, t, N, and grad$N$ at most, while D depends on $\mathbf{r}$, t, and N at most. The above is a differential equation in the only unknown N. It must be supplemented with the initial condition $N_0(\mathbf{r}) = N(\mathbf{r}, t = 0)$ and suitable boundary conditions for $t > 0$. If the diffusion coefficient is constant, or depends on t at most, (23.6) becomes

$$\frac{\partial N}{\partial t} = D\nabla^2 N + W, \qquad D = D(t). \tag{23.7}$$

It is convenient to consider a simplified form of (23.7) to be used as a model problem. For this, one takes the one-dimensional case in the x direction and lets $W = 0$, this yielding

$$\frac{\partial N}{\partial t} = D\,\frac{\partial^2 N}{\partial x^2}. \tag{23.8}$$

Equation (23.8) is also called *Fick's second law of diffusion*. Thanks to the linearity of (23.8), the solution can be tackled by means of the Fourier-transform method, specifically, by transforming both sides of (23.8) with respect to x. Indicating[1] with $G(k, t) = \mathcal{F}_x N$ the transform of N with respect to x, and using some of the properties of the Fourier transform illustrated in Appendix C.2, one finds

$$\mathcal{F}_x \frac{\partial N}{\partial t} = \frac{\mathrm{d}G}{\mathrm{d}t}, \qquad \mathcal{F}_x D\,\frac{\partial^2 N}{\partial x^2} = D\mathcal{F}_x \frac{\partial^2 N}{\partial x^2} = -k^2 DG. \tag{23.9}$$

The symbol of total derivative is used at the right hand side of the first of (23.9) because k is considered as a parameter. The Fourier transform of the initial condition of N provides the initial condition for G, namely, $G_0 = G(k, t = 0) = \mathcal{F}_x N_0$.

[1] Symbol $\mathcal{F}_x$ indicating the Fourier transform should not be confused with the symbol $\mathcal{F}$ used for the particles' flux in Sect. 23.2.

Equating the right hand sides of (23.9) and rearranging yields $\mathrm{d}G/G = -k^2 D(t)\,\mathrm{d}t$. Integrating the latter from 0 to t,

$$\log(G/G_0) = -k^2 a(t), \qquad a(t) = \int_0^t D(t')\,\mathrm{d}t', \tag{23.10}$$

with a an area. The concentration N is now found by antitransforming the expression of G extracted from the first of (23.10):

$$N(x,t) = \mathcal{F}_k^{-1} G = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} G_0\, \exp(\mathrm{i}kx - ak^2)\,\mathrm{d}k. \tag{23.11}$$

In turn, G_0 within the integral of (23.11) is expressed as the transform of N_0. After rearranging the integrals one finds

$$N(x,t) = \int_{-\infty}^{+\infty} N_0(\xi) \left\{ \int_{-\infty}^{+\infty} \frac{1}{2\pi} \exp[\mathrm{i}k(x-\xi) - ak^2]\,\mathrm{d}k \right\} \mathrm{d}\xi. \tag{23.12}$$

As shown in Appendix C.7, the expression in braces in (23.12) is the integral form of the function $\Delta(x-\xi,t)$ defined by (C.75). As a consequence, the solution of the simplified form (23.8) of the diffusion equation is the convolution between Δ and the initial condition N_0, namely,

$$N(x,t) = \int_{-\infty}^{+\infty} N_0(\xi)\,\Delta(x-\xi,t)\,\mathrm{d}\xi. \tag{23.13}$$

A straightforward calculation shows that Δ fulfills (23.8) for all ξ. As a consequence, (23.13) is a solution as well. In addition, due to (C.79), (23.13) also fulfills the initial condition N_0.

23.5 Predeposition and Drive-in Diffusion

Basing on the model problem worked out in Sect. 23.4 it is possible to describe the thermal diffusion of dopants in silicon. The modification induced in the electrical properties of the silicon lattice by the inclusion of atoms belonging to different chemical species (e.g., phosphorus or boron) are described elsewhere (Sects. 18.4.1 and 18.4.2). Here the analysis deals with the diffusion process in itself.

The formation of a diffused profile in silicon is typically obtained in a two-step process [46, 75, 105]. In the first step, called *predeposition*, a shallow layer of dopants is introduced into the semiconductor. The most common predeposition methods are the diffusion from a chemical source in a vapor form or the diffusion from a solid source (e.g., polycrystalline silicon) having a high concentration of dopants in it. In both methods the silicon wafers are placed in contact with the source of dopant within a furnace kept at a high temperature.

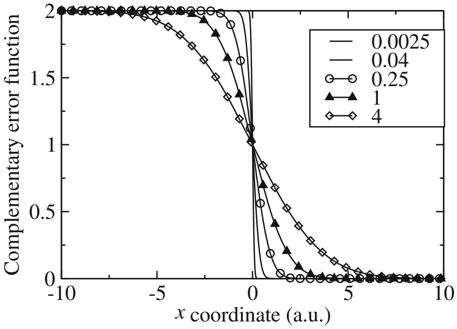

Fig. 23.2 Normalized profiles N/C produced at different instants by a predeposition, using the first of (23.15) as initial condition with arbitrary units for the x coordinate. The outcome is a set of complementary error functions whose expression is the first of (23.16). The legends show the value of $4\,a$ for each curve, also in arbitrary units, with $a = a(t)$ given by the second of (23.10)

During a predeposition step, new dopant atoms are continuously supplied by the source to the silicon region. As a consequence, the number of dopant atoms in the silicon region increases with time. When the desired amount of atoms is reached, the supply of dopants is blocked, whereas the diffusion process is continued. During this new step, called *drive-in diffusion*, the number of dopant atoms in the silicon region remains constant. The drive-in diffusion is continued until a suitable profile is reached.

Typically, the blocking of the flow of dopant atoms from the source to the silicon region is achieved by introducing oxidizing molecules into the furnace atmosphere, this resulting in the growth of a silicon-dioxide layer at the silicon surface (the details of the oxidation process are given in Chap. 24).

It is worth anticipating that in some processes the predeposition step is skipped, and the dopant atoms are introduced into the silicon wafers at low temperature by means of an ion-implantation process. The implanted wafers are then placed into the high-temperature furnace to activate the drive-in diffusion.

23.5.1 Predeposition

Figure 23.2 provides a schematic picture of the source–wafer structure during a predeposition step. The interface between wafer and source is assumed to coincide with the y, z plane, with the x axis oriented towards the wafer's bulk, and the initial condition N_0 is assumed constant in the source region. The diffusion coefficients in the source and wafer regions are provisionally taken equal to each other. Thanks to these assumptions the problem has no dependencies on the y, z variables, and the one-dimensional form (23.8) of the diffusion equation holds. In the practical cases the extent of the source region in the x direction is large and the concentration of the dopant atoms in it is high. As a consequence, the source is not depleted when the atoms diffuse into the wafer. The spatial form of the concentration N at a given time $t = t'$ is called *diffused profile*. Its integral over the semiconductor region,

$$Q(t') = \int_0^{+\infty} N(x, t = t')\, \mathrm{d}x \qquad (\mathrm{m}^{-2}), \tag{23.14}$$

is called *dose*.

For convenience the constant value of the initial condition in the source region is indicated with $2\,C$ (m^{-3}). It follows that the initial condition of the predeposition step is given by the first of (23.15). In turn, the general expression (23.13) of the dopant concentration reduces to the second of (23.15):

$$N_0(\xi) = \begin{cases} 2C & \xi < 0 \\ 0 & \xi > 0 \end{cases} \quad ; \qquad N(x, t) = 2C \int_{-\infty}^{0} \Delta(x - \xi, t)\, \mathrm{d}\xi. \tag{23.15}$$

Using (C.78) and (C.71) one finds the following expressions for the diffused profile and dose of the predeposition step,

$$N(x, t) = C\,\mathrm{erfc}\left(\frac{x}{\sqrt{4a}}\right), \qquad Q(t) = C\sqrt{\frac{4a}{\pi}}, \tag{23.16}$$

where the dependence on t derives from the second of (23.10). As parameter a increases with time, the dose increases with time as well, consistently with the qualitative description of predeposition given earlier in this section. In most cases the diffusion coefficient is independent of time, $a = D\,t$, this yielding $Q \propto \sqrt{t}$.

Still from the second of (23.10) one finds $a(0) = 0$. Combining the latter with the properties (C.69) of the complementary error function shows that $\lim_{t \to 0^+} N(x, t)$ coincides with the initial condition given by the first of (23.15). Also, the solution (23.16) fulfills the boundary conditions $N(-\infty, t) = 2C$, $N(+\infty, t) = 0$ at any $t > 0$. Finally it is $N(0, t) = C$ at any $t > 0$. This explains the term *constant-source diffusion* that is also used to indicate this type of process. In fact, the concentration at the wafer's surface is constant in time. Figure 23.2 shows the normalized concentration N/C calculated from the first of (23.16) at different values of a.

The analysis of the diffusion process carried out so far was based on the assumption of a position-independent diffusion coefficient D. In the actual cases this assumption is not fulfilled because the dopant source and the wafer are made of different materials. As a consequence, the solution of (23.8) must be reworked. In the case of predeposition this is accomplished with little extra work, which is based on the first of (23.16) as shown below.

One assumes, first, that the diffusion coefficient in either region is independent of time, as is the standard condition of the typical processes. In each region the diffusion coefficient takes a spatially-constant value, say, D_S in the source and $D_W \neq D_S$ in the wafer. Now, observe that (23.8) is homogeneous and contains the derivatives of N, but not N itself. It follows that, if $C\,\mathrm{erfc}[x/(4a)^{1/2}]$ is the solution of (23.8) fulfilling some initial and boundary conditions, then $A\,\mathrm{erfc}[x/(4a)^{1/2}] + B$ is also a solution of (23.8), fulfilling some other conditions that depend on the constants A and B. One then lets, with $t > 0$,

$$N_S = A_S\,\mathrm{erfc}\left(\frac{x}{\sqrt{4D_S t}}\right) + B_S, \qquad x < 0, \tag{23.17}$$

$$N_W = A_W\,\mathrm{erfc}\left(\frac{x}{\sqrt{4D_W t}}\right) + B_W, \qquad x > 0, \tag{23.18}$$

and fixes two relations among the constants in order to fulfill the initial conditions (23.15):

$$\lim_{t=0^+} N_S = 2A_S + B_S = 2C, \qquad \lim_{t=0^+} N_W = B_W = 0. \tag{23.19}$$

In order to fix the remaining constants one must consider the matching conditions of the two regional solutions (23.17, 23.18) at the source–wafer interface. The concentrations across an interface between two different media are related by the *segregation coefficient* k [105, Sect. 1.3.2]. Also, given that no generation or destruction of dopant atoms occurs at the interface, the flux density $-D\,\partial N/\partial x$ must be continuous there. In summary, the matching conditions at the source–wafer interface are

$$N_W(0^+, t) = k\,N_S(0^-, t), \qquad D_W\left(\frac{\partial N_W}{\partial x}\right)_{0^+} = D_S\left(\frac{\partial N_S}{\partial x}\right)_{0^-}. \tag{23.20}$$

Using (23.17, 23.18, 23.19) transforms (23.20) into $A_W = k\,(2C - A_S)$ and, respectively, $\sqrt{D_W}\,A_W = \sqrt{D_S}\,A_S$ whence, remembering the first of (23.19) and letting $\eta = D_W/D_S$,

$$A_S = \frac{k\sqrt{\eta}}{1 + k\sqrt{\eta}}\,2C, \qquad B_S = \frac{1 - k\sqrt{\eta}}{1 + k\sqrt{\eta}}\,2C, \qquad A_W = \frac{k}{1 + k\sqrt{\eta}}\,2C. \tag{23.21}$$

Thanks to (23.21), the concentration of the dopant atoms in the source region at the source–wafer interface at $t > 0$ turns out to be $N_S(0^-, t) = A_S + B_S = 2C/(1 + k\sqrt{\eta})$. If, in particular, the source of dopant is in the gaseous phase, it is $\eta \ll 1$. As k is of order unity, one finds for the gaseous source $N_S(0^-, t) \simeq 2C$, namely, the interface concentration of the source region is practically equal to the asymptotic one. Figure 23.3 shows the diffused profile N calculated from (23.17, 23.18) at two different instants t_1 and $t_2 = 16\,t_1$, with $D_S = 400\,D_W$. The coefficients are, in arbitrary units, $A_S = 2$, $B_S = 78$, $A_W = 40$, $B_W = 0$. From the first of (23.19) it follows $C = 41$. Letting $(4\,D_W\,t_1)^{1/2} = 1$ (a.u.) one has $(4\,D_S\,t_1)^{1/2} = 20$, $(4\,D_W\,t_2)^{1/2} = 4$, and $(4\,D_W\,t_2)^{1/2} = 80$. These values are used to calculate the four curves shown in the figure.

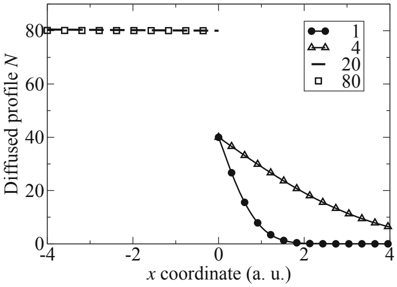

Fig. 23.3 Diffused profiles calculated at t_1 and $t_2 = 16\,t_1$ when two different materials are involved. The calculation is based on (23.17), (23.18) as described at the end of Sect. 23.5.1. The legends show the $(4Dt)^{1/2}$ value for each curve

23.5.2 Drive-in Diffusion

As indicated at the beginning of this section, the drive-in diffusion is started when the desired amount of atoms has been introduced into the silicon lattice, and is continued until a suitable profile is reached.

In principle, the profile to be used as initial condition of a drive-in diffusion is not exactly equal to the final profile of the predeposition step. In fact, the boundary condition $(\partial N_W/\partial x)_{0^+}$ is different from zero during the predeposition step. Instead, during the growth of the silicon-dioxide layer that blocks the supply of dopant atoms from the source region, the boundary condition becomes equal to zero to adapt to the situation of a vanishing flux density of dopants across the interface.

The calculation of the drive-in diffusion is tackled more easily by assuming that the blocking of the supply of dopants atoms is instantaneous, so that the final profile of the predeposition step is "frozen". Then, one considers the full domain $-\infty < x < +\infty$ instead of the wafer domain $0 \le x < +\infty$, with the same diffusion coefficient $D = D_W$ everywhere. In this way one can still use the model problem (23.8). As for the initial condition N_0, one mirrors the final profile of the predeposition step over the negative axis, this making the initial condition even with respect to x. Letting $x \leftarrow -x$ in (23.13) one easily proves that $N(-x,t) = N(x,t)$ if $N_0(-\xi) = N_0(\xi)$; namely, if the initial condition is even, then the solution is even at all times. With the provisions above one finds $(\partial N_W/\partial x)_{0^+} = -(\partial N_W/\partial x)_{0^-}$, which automatically fulfills the condition of a vanishing flux density of dopants across the origin. Then, the application of (23.13) provides the profile of the drive-in diffusion in the wafer region $0 \le x < +\infty$.

The final profile (23.16) of the predeposition step, used as initial condition, does not lend itself to an analytical calculation of the drive-in diffusion. Some examples of calculation are given below, in which profiles of a simpler form than (23.16) are used as approximations. Let Q be the dose present within the wafer region. As a first example one lets

$$N_0(\xi) = 2Q\,\delta(\xi), \qquad N(x,t) = 2Q \int_{-\infty}^{+\infty} \delta(\xi)\,\Delta(x-\xi,t)\,\mathrm{d}\xi. \tag{23.22}$$

From the properties of the Dirac δ (Sect. C.4) it follows

$$N(x,t) = 2Q\, \Delta(x,t) = 2Q\, \frac{\exp[-x^2/(4\,a)]}{\sqrt{4\pi\, a}}, \tag{23.23}$$

showing that, when the initial condition is a Dirac δ, the profile resulting from a diffusion process is Gaussian. Only the portion of (23.23) belonging to the wafer region, that is, $x \geq 0$, must in fact be considered. Integrating (23.23) from 0 to $+\infty$ and using (C.77) yields the expected value Q of the dose at all times. Although rather crude, the approximation of using a Dirac δ as initial condition is acceptable, because the profile obtained from a predeposition or an ion-implantation process is typically very thin.

As a second example one takes a Gaussian profile as the initial condition, specifically, the second of (23.23) where, to better distinguish the symbols, a is replaced with a_1. It is assumed that the drive-in diffusion to be calculated is characterized by another value of the parameter, say, a_2. The difference between a_2 and a_1 may be due to the duration of the diffusion process under investigation, to a temperature-induced difference in the diffusion coefficients, or both. As usual the instant $t = 0$ is set as the initial time of the diffusion process. Applying (23.13) yields

$$N(x,t) = 2Q \int_{-\infty}^{+\infty} \frac{\exp[-\xi^2/(4\,a_1)]}{\sqrt{4\pi\, a_1}}\, \frac{\exp[-(x-\xi)^2/(4\,a_2)]}{\sqrt{4\pi\, a_2}}\, \mathrm{d}\xi. \tag{23.24}$$

Using the auxiliary variable $\eta = \xi - a_1\, x/(a_1 + a_2)$, whence $x - \xi = -\eta + a_2\, x/(a_1 + a_2)$, transforms the exponent of (23.24) as

$$-\frac{\xi^2}{4\,a_1} - \frac{(x-\xi)^2}{4\,a_2} = -\frac{x^2}{4\,(a_1+a_2)} - \frac{a_1+a_2}{4\,a_1 a_2}\,\eta^2. \tag{23.25}$$

Then, integrating with respect to $\sqrt{(a_1+a_2)/(4a_1a_2)}\,\eta$ and using again (C.77) yields

$$N(x,t) = 2Q\, \frac{\exp[-x^2/(4\,a_1 + 4\,a_2)]}{\sqrt{4\pi\,(a_1+a_2)}}. \tag{23.26}$$

As before, the integral of the profile from 0 to $+\infty$ yields the dose Q at all times. The result expressed by (23.26) is important because it shows that a diffusion process whose initial condition is a Gaussian profile yields another Gaussian profile. The parameter of the latter is found by simply adding the parameter $a_2 = \int_0^t D(t')\,\mathrm{d}t'$ of the diffusion process in hand, whose duration is t, to the parameter a_1 of the initial condition. Clearly, the result is also applicable to a sequence of successive diffusion processes. In fact, it is used to calculate the final profiles after the wafers have undergone the several thermal processes that are necessary for the integrated-circuit fabrication.

23.6 Generalization of the Model Problem

The generalization of the model problem (23.8) to three dimensions, that is, Eq. (23.7) with $W = 0$ and initial condition $N_0(\mathbf{r}) = N(\mathbf{r}, t = 0)$, is still tackled by means of the Fourier transform. For this, it is necessary to define the vectors $\mathbf{r} = (r_1, r_2, r_3)$, $\mathbf{s} = (s_1, s_2, s_3)$, $\mathbf{k} = (k_1, k_2, k_3)$, and the elements $\mathrm{d}^3k = \mathrm{d}k_1\,\mathrm{d}k_2\,\mathrm{d}k_3$, $\mathrm{d}^3s = \mathrm{d}s_1\,\mathrm{d}s_2\,\mathrm{d}s_3$. Using (C.20) and following the procedure of Sect. 23.4, one finds again the relations (23.10). This time, however, it is $k^2 = k_1^2 + k_2^2 + k_3^2$. The solution $N(\mathbf{r}, t)$ is readily found as a generalization of (23.12), namely

$$N(\mathbf{r}, t) = \iiint_{-\infty}^{+\infty} N_0(\mathbf{s}) \left\{ \iiint_{-\infty}^{+\infty} \frac{1}{(2\pi)^3} \exp\left[\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{s}) - ak^2\right] \mathrm{d}^3k \right\} \mathrm{d}^3s. \tag{23.27}$$

The expression in braces in (23.27) is the product of three functions of the same form as (C.75). It follows

$$N(\mathbf{r}, t) = \iiint_{-\infty}^{+\infty} N_0(\mathbf{s}) \frac{\exp\left[-|\mathbf{r}-\mathbf{s}|^2/(4a)\right]}{(4\pi\, a)^{3/2}} \mathrm{d}^3s. \tag{23.28}$$

When the net generation rate per unit volume, W, is different from zero, it is in general impossible to find an analytical solution of (23.7). An important exception is the case where W is linear with respect to N and has no explicit dependence on $\mathbf{r}$ or t. In this case (23.7) reads

$$\frac{\partial N}{\partial t} = D\nabla^2 N - \frac{N - N_a}{\tau}, \qquad D = D(t), \tag{23.29}$$

where the two constants N_a (m^{-3}) and τ (s) are positive. This form of W is such that the particles are generated if $N(\mathbf{r}, t) < N_a$, while they are destroyed if $N(\mathbf{r}, t) > N_a$. Equation (23.29) is easily solved by introducing an auxiliary function N' such that $N = N_a + N' \exp(-t/\tau)$. In fact, N' turns out to be the solution of the three-dimensional model problem, so that, using (23.28), the solution of (23.29) reads

$$N(\mathbf{r}, t) = N_a + \exp(-t/\tau) \iiint_{-\infty}^{+\infty} N_0(\mathbf{s}) \frac{\exp\left[-|\mathbf{r}-\mathbf{s}|^2/(4a)\right]}{(4\pi\, a)^{3/2}} \mathrm{d}^3s. \tag{23.30}$$

23.7 Complements

23.7.1 Generation and Destruction of Particles

The discussion carried out in Sect. 23.2 about the continuity equation implies the possibility that particles may be generated or destroyed. To tackle this issue consider the problem "counting the time variation of students in a classroom". Assuming that the classroom has only one exit, to accomplish the task it suffices to count the students that cross the exit, say, every second. The students that enter (leave) the room are counted as a positive (negative) contribution.

Consider now a slightly modified problem: "counting the time variation of *non-sleeping* students in a classroom". To accomplish the task it does not suffice any more to count the students that cross the exit. In fact, a student that is initially awake inside the classroom may fall asleep where she sits (one assumes that sleeping students do not walk); this provides a negative contribution to the time variation sought, without the need of crossing the exit. Similarly, an initially-sleeping student may wake up, this providing a positive contribution. Falling asleep (waking up) is equivalent to destruction (creation) of a non-sleeping student.

In the two examples above the objects to be counted are the same, however, in the second example they have an extra property that is not considered in the first one. This shows that creation/destruction of a given type of objects may occur or not, depending on the properties that are considered. When particles instead of students are investigated, it is often of interest to set up a continuity equation for describing the time variation, in a given volume, of the particles *whose energy belongs to a specified range*. Due to their motion, the particles undergo collisions that change their energy. As a consequence a particle may enter, or leave, the specified energy range without leaving the spatial volume to which the calculation applies. In this example the origin of the net generation rate per unit volume W introduced in Sect. 23.2 is the extra property about the particles' energy.

23.7.2 Balance Relations

As indicated in Sect. 23.2, and with the provisions illustrated in Sect. 23.7.1, the continuity equation is a balance relation for the number of particles. Due to its intrinsic simplicity and generality, the concept of balance relation is readily extended to physical properties different from the number of particles; for instance, momentum, energy, energy flux, and so on. A detailed illustration of this issue is given in Chap. 19. It is also worth noting, in contrast, that the transport equation of the diffusion type (23.4), being based on a specific assumption about the transport mechanism, is less general than the continuity equation.

23.7.3 Lateral Diffusion

The treatment of predeposition and drive-in diffusion carried out in Sect. 23.5 is based on a one-dimensional model. This implies that the concentration of the dopant at the interface between the source and wafer regions is constant along the y and z directions. In the practical cases this is impossible to achieve, because the area over which the source is brought into contact with the wafer is finite. In fact, prior to the predeposition step the surface of the wafer is covered with a protective layer, called *mask*. As indicated in Sect. 24.1, in the current silicon technology the mask is typically made of thermally-grown silicon dioxide. Next, a portion of the mask is removed to expose the silicon surface over a specific area, called *window*, through which the predeposition step takes place.

From the description above it follows that the initial condition N_0 of the predeposition step is constant only within the window, while it is equal to zero in the other parts of the y, z plane. This makes the hypothesis of a one-dimensional phenomenon inappropriate, and calls for the use of the three-dimensional solution (23.28). The subsequent drive-in diffusions must be treated in three dimensions as well, due to the form of their initial conditions. An important effect is the diffusion of the dopant underneath the edges of the mask. This phenomenon, called *lateral diffusion*, makes the area where the doping profile is present larger than the original mask, and must be accounted for in the design of the integrated circuit.

23.7.4 Alternative Expression of the Dose

The definition of the dose Q deriving from the one-dimensional model problem is (23.14). Letting $W = 0$ in (23.3), using its one-dimensional form $\partial N/\partial t = -\partial F/\partial x$, and observing that it is $F(+\infty, t) = 0$ due to the initial condition, gives the following expression for the time derivative of the dose:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\int_0^{+\infty} \frac{\partial F(x,t)}{\partial x}\,\mathrm{d}x = F(0,t). \tag{23.31}$$

Integrating (23.31) and remembering that the dose at $t = 0$ is equal to zero yields

$$Q(t') = \int_0^{t'} F(0,t)\,\mathrm{d}t. \tag{23.32}$$

The procedure leading from the original definition (23.14) of the dose to its alternative expression (23.32) is based solely on (23.3), hence it does not depend on a specific transport model.

23.7.5 The Initial Condition of the Predeposition Step

The initial condition N_0 of the predeposition step is given by the first of (23.15). To carry out the solution of the diffusion equation it is necessary to recast N_0 in an integral representation of the Fourier type. However, (23.15) does not fulfill the condition (C.19) that is sufficient for the existence of such a representation.

Nevertheless the solution procedure leading to (23.13) is still applicable. In fact, remembering the definition (C.8) of the unit step function H, the initial condition can be recast as $N_0(\xi) = 2\,C\,[1 - H(\xi)]$. In turn, as shown in appendix C.4, H can be represented in the required form.

Problems

23.1 A Gaussian doping profile $N = 2Q\ \exp\,(-x^2/c_1)/\sqrt{\pi\,c_1}$ undergoes a thermal-diffusion process at a temperature such that $D = 10^{-11}$ cm^2/s. Assuming $c_1 = 1.6 \times 10^{-7}$ cm^2, calculate the time that is necessary to reduce the peak value of the profile to 2/3 of the initial value.

23.2 A Gaussian doping profile $N = 2Q\ \exp\,(-x^2/c_1)/\sqrt{\pi\,c_1}$, $c_1 = 9 \times 10^{-8}$ cm^2, undergoes a thermal-diffusion process with $c_2 = 16 \times 10^{-8}$ cm^2 yielding another Gaussian profile. Find the value $\bar{x}$ (in microns) where the two profiles cross each other.

23.3 A Gaussian doping profile $N = 2Q\ \exp(\,-\,x^2/c_1)/\sqrt{\pi\,c_1}$, $c_1 = 2.5 \times 10^{-6}$ cm^2, undergoes a 240 min-long thermal-diffusion process at a temperature such that the diffusion coefficient is $D = 2.5 \times 10^{-10}$ cm^2 s^{-1}. Determine the ratio between the peak value of the final profile and that of the initial one.

23.4 A Gaussian doping profile $N = 2Q\ \exp(\,-\,x^2/c_1)/\sqrt{\pi\,c_1}$, $c_1 = 10^{-6}$ cm^2, undergoes a thermal-diffusion process in which $c_2 = 3 \times 10^{-8}$ cm^2. Find the position $\bar{x}$ (in microns) where the value of the initial doping profile equals the value that the final profile has in $x = 0$.

23.5 A Gaussian doping profile $N = 2Q\ \exp(\,-\,x^2/c_1)/\sqrt{\pi\,c_1})$ undergoes a thermal-diffusion process in which $c_2 = 10^{-8}$ cm^2. The value of the final profile in the origin is equal to that of the initial profile at $x_0 = 1.1 \times \sqrt{c_1}$. Find the value of c_1 in cm^2.

23.6 A Gaussian doping profile $N = 2Q\ \exp(\,-\,x^2/c_1)/\sqrt{\pi\,c_1})$, $c_1 = 1.8 \times 10^{-8}$ cm^2, undergoes a thermal-diffusion process whose duration is $t = 10$ min, with $D = 10^{-11}$ cm^2 s^{-1}. At the end of the process the concentration at some point x_0 is $N_1 = 3 \times 10^{16}$ cm^{-3}. If the process duration were 20 min, the concentration at the same point would be $N_2 = 3 \times 10^{17}$. Find the value of x_0 in microns.

23.7 The doping profile resulting from a predeposition process with $D = 10^{-11}$ cm^2 s^{-1} is $N(x) = N_S\,\mathrm{erfc}(x/\sqrt{c})$. The ratio between the dose and surface concentration is $Q/N_S = \lambda/\sqrt{\pi}$, $\lambda = 1095$ nm. Find the duration t of the predeposition process, in minutes.

23.8 The initial condition of a drive-in diffusion is given by $N_0 = 2Q\,(h - x)/h^2$ for $0 \le x \le h$, and by $N_0 = 0$ elsewhere, where $Q > 0$ is the dose. Find the expression of the profile at $t > 0$.