Chapter 20
Generation-Recombination and Mobility

20.1 Introduction

The chapter illustrates the main contributions to the transitions of the inter-band type, that give rise to the generation-recombination terms in the continuity equations for electrons and holes, and to those of the intra-band type, that give rise to the electron and hole mobilities in the current-density equations. The inter-band transitions that are considered are the net thermal recombinations (of the direct and trap-assisted type), Auger recombinations, impact-ionization generations, and net-optical recombinations. The model for each type of event is first given as a closed-form function of the semiconductor-device model's unknowns, like carrier concentrations, electric field, or current densities. Such functions contain a number of coefficients, whose derivation is successively worked out in the complements by means of a microscopic analysis. Some discussion is devoted to the optical-generation and recombination events, to show how the concepts of semiconductor laser, solar cell, and optical sensor may be derived as particular cases of non-equilibrium interactions between the material and an electromagnetic field. The intra-band transitions are treated in a similar manner: two examples, the collisions with acoustic phonons and ionized impurities, are worked out in some detail; the illustration then follows of how the contributions from different scattering mechanisms are combined together in the macroscopic mobility models. The material is supplemented with a brief discussion about advanced modeling methods.

20.2 Net Thermal Recombinations

As anticipated in Sect. 19.5.5, it is customary to separate the net generation rates W_n, W_p into two contributions, namely, those deriving from the phonon collisions and those of the other types (e.g., electron-electron collisions, electron-photon collisions, and so on). The separate contributions are defined in (19.132); this section deals with the net thermal recombination rates U_n, U_p.

M. Rudan, *Physics of Semiconductor Devices,*
DOI 10.1007/978-1-4939-1151-6_20

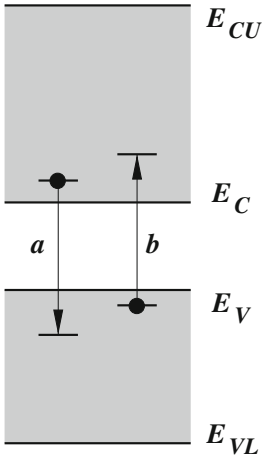

Fig. 20.1 A graphic example of direct thermal recombination (*a*) and generation (*b*). The edges of the conduction and valence bands are indicated with the same symbols used in Sect. 18.2. The same drawing applies also to the description of the direct optical recombinations and generations (Sect. 20.4)

In the calculations carried out below, the non-equilibrium carrier concentrations are derived by integrating over the bands' energy. This is consistent with the general definitions (19.31) and (19.109). In fact, considering the non-equilibrium electron concentration n as defined in (19.31), one introduces the variable transformation illustrated in Sect. B.5 and replaces the quantities appearing in it as follows:

$$(u, v, w) \leftarrow (k_1, k_2, k_3), \qquad \sigma \leftarrow (\mathbf{r}, t), \qquad \eta \leftarrow E, \tag{20.1}$$

$$S \leftarrow n, \qquad s \leftarrow f = Q\,\Phi, \qquad b \leftarrow g, \qquad \bar{s} \leftarrow P, \tag{20.2}$$

where Q, $g(E)$ are the densities of states in the phase space $\mathbf{r}, \mathbf{k}$ and, respectively, in energy, while $\Phi(\mathbf{r}, \mathbf{k}, t)$, $P(\mathbf{r}, E, t)$ are the non-equilibrium occupation probabilities in the phase space and, respectively, in energy; the integration in energy is carried out over the range corresponding to the conduction band's branch. The hole concentration is treated in the same manner. In conclusion,

$$n(\mathbf{r}, t) = \iiint_{-\infty}^{+\infty} Q\,\Phi\,\mathrm{d}^3k = \int_{E_C}^{E_{CU}} g\,P\,\mathrm{d}E, \tag{20.3}$$

$$p(\mathbf{r}, t) = \iiint_{-\infty}^{+\infty} Q\,(1-\Phi)\,\mathrm{d}^3k = \int_{E_{VL}}^{E_V} g\,(1-P)\,\mathrm{d}E. \tag{20.4}$$

20.2.1 Direct Thermal Recombinations

To begin, a graphic example of thermal transitions is shown in Fig. 20.1, where the edges of the conduction and valence bands are indicated with the same symbols

used in Sect. 18.2; the transition marked with a is a recombination event, in which an electron belonging to an energy state of the conduction band transfers to an empty state of the valence band. The energy difference between the initial and final state is released to the lattice in the form of a phonon. The opposite transition, where the electron's energy increases due to phonon absorption, is an electron-hole generation and is marked with b in the figure. The transitions of type a and b are called *direct thermal recombination* and *direct thermal generation*, respectively. Let r_a be the number of direct thermal recombinations per unit volume and time, and r_b the analogue for the generations; considering the conduction band as a reference, the difference $r_a - r_b$ provides the contribution to the net thermal recombination rate U_n due to the direct thermal transitions. When the valence band is considered instead, the rates of electron transitions reverse; however, for the valence band the transitions of holes must be considered: as consequence, the contribution to U_p is again $r_a - r_p$. In conclusion,

$$U_{DT} = U_{DTn} = U_{DTp} = r_a - r_b, \tag{20.5}$$

where D stands for "direct" and T for "thermal". The expressions of r_a, r_b are determined by a reasoning similar to that used in Sect. 19.3.1 to express the collision term of the BTE; here, however, the analysis is carried out directly in the energy space instead of the $\mathbf{k}$ space.[1] Let $P(\mathbf{r}, E, t)$ be the occupation probability of a state at energy E; then, let C be the probability per unit time and volume (in $\mathbf{r}$) of an electron transition from a state of energy E to a state of energy E' belonging to a different band, induced by the interaction with a phonon.[2] Such a probability depends on the phonon energy $\hbar\,\omega$ (Sect. 12.5), and also on the position in $\mathbf{r}$ if the semiconductor is non uniform. Typically, the equilibrium distribution is assumed for the phonons, which makes C independent of time; as the collisions are point-like (Sect. 19.3.2), the spatial positions of the initial and final states coincide, whence $C = C(\mathbf{r}, \hbar\,\omega, E \to E')$.

Indicating with $g(E)$ the density of states of the band where the initial state belongs, the product $g\ \mathrm{d}E\ P$ is the number of electrons within the elementary interval $\mathrm{d}E$ around the initial state; such a product is multiplied by C to find the number of unconditional $E \to E'$ transitions. On the other hand, the transitions take place only if the final states around E' are empty; as the fraction of empty states in that interval is $g'\,\mathrm{d}E'\,(1 - P')$, the number of actual transitions from $\mathrm{d}E$ to $\mathrm{d}E'$ turns out to be $g\ \mathrm{d}E\ P\ C\ g'\,\mathrm{d}E'\,(1 - P')$. Now, to calculate the r_a or r_b rate it is necessary to add up all transitions: for r_a one lets E range over the conduction band and E' over the valence band; the converse is done to calculate r_b. As the calculation of the latter is

[1] A more detailed example of calculations is given below, with reference to collisions with ionized impurities.

[2] The units of C are $[C] = \mathrm{m}^{-3}\,\mathrm{s}^{-1}$. Remembering that the phonon energy equals the change in energy of the electron due to the transition (Sect. 14.8.2), it is $C = 0$ for $\hbar\,\omega < E_C - E_V = E_G$ (refer also to Fig. 20.1).

somewhat easier, it is shown first:

$$r_b = \int_{E_{VL}}^{E_V} g \, \mathrm{d}E \; P \int_{E_C}^{E_{CU}} C \, g' \, \mathrm{d}E' \left(1 - P'\right). \tag{20.6}$$

As in normal operating conditions the majority of the valence-band states are filled, while the majority of the conduction-band states are empty, one lets $P \simeq 1$ and $1 - P' \simeq 1$, whence, using symbol G_{DT} for r_b,

$$G_{DT}(\mathbf{r}, \hbar\,\omega) = \int_{E_{VL}}^{E_V} g \, \mathrm{d}E \int_{E_C}^{E_{CU}} C \, g' \, \mathrm{d}E'. \tag{20.7}$$

Thus, the generation rate is independent of the carrier concentrations. To proceed, one uses the relation $g = \Omega\,\gamma$, with γ the combined density of states in energy and volume, given by (15.65), and the definition (20.4) of the hole concentration. Thus, the recombination rate is found to be

$$r_a = \int_{E_C}^{E_{CU}} g \; \mathrm{d}E \; P \int_{E_{VL}}^{E_V} C \, g' \, \mathrm{d}E' \; (1 - P') = p \int_{E_C}^{E_{CU}} K \, g \, P \; \mathrm{d}E, \tag{20.8}$$

where $K(\mathbf{r}, \hbar\,\omega, E)$, whose units are $[K] = \mathrm{s}^{-1}$, is the average of $\Omega\,C$ over the valence band, weighed by $g'\,(1 - P')$:

$$K = \frac{\int_{E_{VL}}^{E_V} \Omega\, C \, g' \, (1 - P') \, \mathrm{d}E'}{\int_{E_{VL}}^{E_V} g' \, (1 - P') \, \mathrm{d}E'}. \tag{20.9}$$

Strictly speaking, K is a functional of P'; however, the presence of P' in both numerator and denominator of (20.9) makes such a dependence smoother, so that one can approximate K using the equilibrium distribution instead of P'. By the same token one uses the definition of the electron concentration (20.3) to find

$$r_a = \alpha_{DT} \, n \, p, \qquad \alpha_{DT}(\mathbf{r}, \hbar\omega) = \frac{\int_{E_C}^{E_{CU}} \Omega \, K \, g \, P \, \mathrm{d}E}{\int_{E_C}^{E_{CU}} g \, P \, \mathrm{d}E}, \tag{20.10}$$

where the integrals are approximated using the equilibrium probability. In conclusion,

$$U_{DT} = \alpha_{DT} \, n \, p - G_{DT}, \tag{20.11}$$

where α_{DT} is the *transition coefficient* of the direct thermal transitions, with units $[\alpha_{DT}] = \mathrm{m}^3 \; \mathrm{s}^{-1}$, and G_{DT} their *generation rate* ($[G_{DT}] = \mathrm{m}^{-3} \; \mathrm{s}^{-1}$). As in the equilibrium case it is $r_a = r_b$, namely, $G_{DT} = \alpha_{DT} \, n^{\mathrm{eq}} \, p^{\mathrm{eq}}$, it follows $U_{DT} = \alpha_{DT} \, (np - n^{\mathrm{eq}} \, p^{\mathrm{eq}})$.

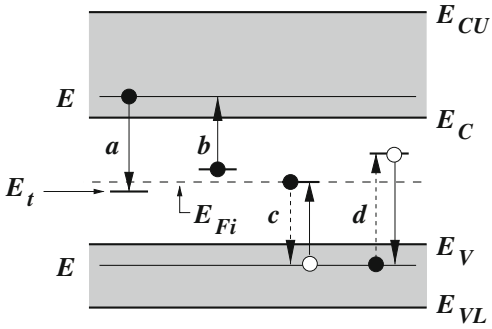

Fig. 20.2 Different types of trap-assisted transitions

20.2.2 Trap-Assisted Thermal Recombinations

An important contribution to the thermal generation and recombination phenomena is due to the so-called *trap-assisted transitions*. As mentioned in Sect. 19.3, among the possible collisions undergone by electrons or holes are those with *lattice defects*. The latter may originate from lattice irregularities (e.g., dislocations of the material's atoms occurring during the fabrication process, Sect. 24.1), or from impurities that were not eliminated during the semiconductor's purification process, or were inadvertently added during a fabrication step. Some defects may introduce energy states localized in the gap; such states, called *traps*, may capture an electron from the conduction band and release it towards the valence band, or vice versa. The phenomena are illustrated in Fig. 20.2, where four traps located in the energy gap are shown in order to distinguish among the different transition events, that are: *a*) capture of a conduction-band electron by a trap, *b*) release of a trapped electron towards the conduction band, *c*) release of a trapped electron towards the valence band (more suitably described as the capture of a valence-band hole by the trap), and *d*) capture of a valence-band electron from the valence band (more suitably described as the release of a hole towards the valence band). Each transition is accompanied by the absorption or emission of a phonon. Thus, transitions of type *a* and *b* contribute to the net thermal recombination U_n of the conduction band, while those of type *c* and *d* contribute to the net thermal recombination U_p of the valence band. Also, a sequence of two transitions, one of type *a* involving a given trap, followed by one of type *c* involving the same trap, produces an electron-hole recombination and is therefore called *trap-assisted thermal recombination*; similarly, a sequence of two transitions, one of type *d* involving a given trap, followed by one of type *b* involving the same trap, produces an electron-hole generation and is therefore called *trap-assisted thermal generation*.

To calculate the contribution of the trap-assisted transitions to U_n and U_p it is necessary to distinguish between two kinds of traps: those of *donor type*, that are electrically neutral when the electron is present in the trap and become positively charged when the electron is released, and those of *acceptor type*, that are electrically neutral when the electron is absent from the trap and become negatively charged when

the electron is captured. In this respect, the traps are similar to the dopants' atoms. Instead, a strong difference is made by the position of the traps' energy within the gap. Consider, for instance, traps localized near the gap's midpoint (the latter is indicated by the intrinsic Fermi level E_{Fi} in Fig. 20.2); the phonon energy necessary for the transition is about $E_G/2$ in all cases, to be compared with the value E_G necessary for a direct transition. On the other hand, the equilibrium-phonon distribution (Sect. 16.6) is the Bose–Einstein statistics (15.55); it follows that the number $\mathrm{d}N_{\mathrm{ph}}$ of phonons in the interval $\mathrm{d}\omega$ is

$$\mathrm{d}N_{\mathrm{ph}} = \frac{g_{\mathrm{ph}}(\omega)\,\mathrm{d}\omega}{\exp\left[\hbar\,\omega/(k_B\,T)\right] - 1}, \tag{20.12}$$

with $\hbar\,\omega$ the energy and g_{ph} the density of states of the phonons. Due to (20.12), $\mathrm{d}N_{\mathrm{ph}}/\mathrm{d}\omega$ rapidly decreases as the phonon energy increases, this making the probability of an electron-phonon interaction much larger at lower energies. For this reason, even in an electronic-grade semiconductor, where the concentration of defects is very small (Sect. 19.3.2), the traps are able to act as a sort of "preferred path" in energy for the inter-band transitions, to the extent that the contribution to U_n, U_p of the trap-assisted transitions is largely dominant over that of the direct transitions. Therefore, in the continuity Eqs. (20.13) below, and in the subsequent derivation of the trap-assisted, thermal-transition rates, symbols U_n, U_p refer only to the latter transitions, not any more to the sum of the trap-assisted and direct ones.

The net thermal-recombination terms U_n, U_p appear in (19.129) and (19.130) after replacing W_n, W_p with (19.132); this yields

$$\frac{\partial n}{\partial t} + U_n - \frac{1}{q}\,\mathrm{div}\mathbf{J}_n = G_n, \qquad \frac{\partial p}{\partial t} + U_p + \frac{1}{q}\,\mathrm{div}\mathbf{J}_p = G_p. \tag{20.13}$$

To introduce the trap-assisted transitions one formally duplicates (20.13) as if the acceptor and donor traps formed two additional bands; as the acceptor traps are either neutral or negatively charged, the charge and current densities of the band associated to them are thought of as due to electrons; instead, the charge and current densities of the band associated to the donor traps are thought of as due to holes. In summary, the two additional equations read

$$\frac{\partial n_a}{\partial t} + U_{na} - \frac{1}{q}\,\mathrm{div}\mathbf{J}_{na} = G_{na}, \qquad \frac{\partial p_d}{\partial t} + U_{pd} + \frac{1}{q}\,\mathrm{div}\mathbf{J}_{pd} = G_{pd} \tag{20.14}$$

with a and d standing for "acceptor" and "donor", respectively. To ease the calculation it is assumed that the non-thermal phenomena are absent, whence $G_n = G_p = G_{na} = G_{pd} = 0$. Combining (20.13) with (20.14), and observing that $\mathbf{J} = \mathbf{J}_p + \mathbf{J}_{pd} + \mathbf{J}_n + \mathbf{J}_{na}$ is the total current density of the semiconductor, yields

$$\frac{\partial[q\,(p + p_d - n - n_a)]}{\partial t} + \mathrm{div}\mathbf{J} = q\,(U_n + U_{na}) - q\,(U_p + U_{pd}). \tag{20.15}$$

As the net dopant concentration N is independent of time, it is $\partial[q\,(p + p_d - n - n_a)]/\partial t = \partial[q\,(p + p_d - n - n_a + N)]/\partial t = \partial\rho/\partial t$; thus, the left hand side of

(20.15) vanishes due to (4.23), and[3]

$$U_n + U_{na} = U_p + U_{pd}. \tag{20.16}$$

The two continuity Eqs. (20.14) are now simplified by observing that in crystalline semiconductors the current densities $\mathbf{J}_{pd}$, $\mathbf{J}_{na}$ of the traps are negligible. In fact, the trap concentration is so low that inter-trap tunneling is precluded by the large distance from a trap to another; the reasoning is the same as that used in Sect. 18.7.2 with respect to the impurity levels.[4] Letting $\mathbf{J}_{pd} = \mathbf{J}_{na} = 0$ makes the two Eqs. (20.14) local:

$$\frac{\partial n_a}{\partial t} = -U_{na}, \qquad \frac{\partial p_d}{\partial t} = -U_{pd}. \tag{20.17}$$

In steady-state conditions the traps' populations are constant, this yielding $U_{na} = U_{pd} = 0$ and, from (20.16), $U_n = U_p$. In equilibrium all continuity equations reduce to the identity $0 = 0$, whence the net-recombination terms vanish independently, $U_n^{\text{eq}} = U_{na}^{\text{eq}} = U_p^{\text{eq}} = U_{pd}^{\text{eq}} = 0$.

20.2.3 Shockley-Read-Hall Theory

The *Shockley-Read-Hall theory* describes the trap-assisted, net thermal-recombination term in a crystalline semiconductor based upon the steady-state relation $U_n = U_p$. In fact, the outcome of the theory is used also in dynamic conditions; this approximation is acceptable because, due to the smallness of the traps' concentration, the contribution of the charge density stored within the traps is negligible with respect to that of the band and dopant states; the contribution of the time variation of the traps' charge density is similarly negligible. The theory also assumes that only one trap level is present, of energy E_t; with reference to Fig. 20.2, the trap levels must be thought of aligned with each other. If more than one trap level is present, the contributions of the individual levels are added up at a later stage. In the theory it is not important to distinguish between acceptor-type or donor-type traps; however, one must account for the fact that a trap can accommodate one electron at most.

Still with reference to Fig. 20.2, let r_a be the number of type-a transitions per unit volume and time, and similarly for r_b, r_c, r_d. The derivation of these rates is

[3] The result expressed by (20.16) is intuitive if one thinks that adding up all continuity equations amounts to counting all transitions twice, the first time in the forward direction (e.g., using the electrons), the second time in the backward direction (using the holes). The reasoning is similar to that leading to the vanishing of the intra-band contribution in (19.63).

[4] In a polycrystalline semiconductor the traps' current densities are not negligible; in fact, the whole system of equations (20.13) and (20.14) must be used to correcly model the material [16, 17, 18]. The conduction phenomenon associated to these current densities is called *gap conduction*.

similar to that of the direct transitions and is shown in the complements; here the expressions of the net thermal-recombination terms are given, that read

$$U_n = r_a - r_b = \alpha_n\, n\, N_t\, (1 - P_t) - e_n\, N_t\, P_t, \tag{20.18}$$

$$U_p = r_c - r_d = \alpha_p\, p\, N_t\, P_t - e_p\, N_t\, (1 - P_t), \tag{20.19}$$

where N_t is the concentration of traps of energy E_t, P_t the trap-occupation probability, α_n, α_p the *electron- and hole-transition coefficients*, respectively, and e_n, e_p the *electron- and hole-emission coefficients*, respectively.[5] The ratios e_n/α_n, e_p/α_p are assumed to vary little from the equilibrium to the non-equilibrium case. From $U_n^{\rm eq} = U_p^{\rm eq} = 0$ one derives

$$\frac{e_n}{\alpha_n} = n^{\rm eq} \left(\frac{1}{P_t^{\rm eq}} - 1 \right), \qquad \frac{e_p}{\alpha_p} = p^{\rm eq} \left(\frac{1}{P_t^{\rm eq}} - 1 \right)^{-1}. \tag{20.20}$$

The occupation probability at equilibrium is the modified Fermi-Dirac statistics (compare with (18.21) or (18.36))

$$P_t^{\rm eq} = \left[\frac{1}{d_t} \exp\left(\frac{E_t - E_F}{k_B\, T} \right) + 1 \right]^{-1}, \qquad \frac{1}{P_t^{\rm eq}} - 1 = \frac{1}{d_t} \exp\left(\frac{E_t - E_F}{k_B\, T} \right), \tag{20.21}$$

with d_t the degeneracy coefficient of the trap. It follows, after introducing the shorthand notation $n_B = e_n/\alpha_n$, $p_B = e_p/\alpha_p$,

$$n_B = \frac{n^{\rm eq}}{d_t} \exp\left(\frac{E_t - E_F}{k_B\, T} \right), \qquad p_B = p^{\rm eq}\, d_t \exp\left(\frac{E_F - E_t}{k_B\, T} \right). \tag{20.22}$$

Note that $n_B\, p_B = n^{\rm eq}\, p^{\rm eq}$. Replacing (20.22) into (20.18), (20.19), and letting $U_n = U_p$ yields

$$\alpha_n\, n\, (1 - P_t) - \alpha_n\, n_B\, P_t = \alpha_p\, p\, P_t - \alpha_p\, p_B\, (1 - P_t), \tag{20.23}$$

whence

$$P_t = \frac{\alpha_n\, n + \alpha_p\, p_B}{\alpha_n\, (n + n_B) + \alpha_p\, (p + p_B)}, \qquad 1 - P_t = \frac{\alpha_n\, n_B + \alpha_p\, p}{\alpha_n\, (n + n_B) + \alpha_p\, (p + p_B)}. \tag{20.24}$$

In this way one expresses the trap-occupation probability as a function of two of the unknowns of the semiconductor-device model, namely, n and p, and of a few parameters. Among the latter, n_B and p_B are known (given the trap's energy) because

[5] It is $[\alpha_{n,p}] = \mathrm{m}^3\ \mathrm{s}^{-1}$, $[e_{n,p}] = \mathrm{s}^{-1}$.

they are calculated in the equilibrium condition. In conclusion, replacing (20.24) into (20.18) or (20.19) yields, for the common value $U_{\mathrm{SRH}} = U_n = U_p$,

$$U_{\mathrm{SRH}} = \frac{n\,p - n^{\mathrm{eq}}\,p^{\mathrm{eq}}}{(n + n_B)/(N_t\,\alpha_p) + (p + p_B)/(N_t\,\alpha_n)}, \tag{20.25}$$

where the indices stand for "Shockley-Read-Hall". Eventually, the only unknown parameters turn out to be the products $N_t\,\alpha_p$ and $N_t\,\alpha_n$ which, as shown in Sect. 25.2, can be obtained from measurements.

The expression obtained so far, (20.25), has been derived considering a single trap level E_t. Before adding up over the levels it is convenient to consider how sensitive U_{SRH} is to variations of E_t; in fact, one notes that the numerator of (20.25) is independent of E_t, whereas the denominator D has the form

$$D = c + 2\,\lambda\,\cosh\eta, \qquad \eta = \frac{E_t - E_F}{k_B\,T} + \frac{1}{2}\,\log\mu, \tag{20.26}$$

where

$$c = \frac{1}{N_t}\left(\frac{n}{\alpha_p} + \frac{p}{\alpha_n}\right), \qquad \lambda = \frac{1}{N_t}\sqrt{\frac{n^{\mathrm{eq}}}{\alpha_p}\,\frac{p^{\mathrm{eq}}}{\alpha_n}}, \qquad \mu = \frac{1}{d_t^2}\,\frac{n^{\mathrm{eq}}/\alpha_p}{p^{\mathrm{eq}}/\alpha_n}. \tag{20.27}$$

The denominator has a minimum where $\eta = 0$; thus, U_{SRH} has a maximum there. Moreover, the maximum is rather sharp due to the form of the hyperbolic cosine. It follows that the trap level E_{tM} that most efficiently induces the trap-assisted transitions is found by letting $\eta = 0$. The other traps levels have a much smaller efficiency and can be neglected; in conclusion, it is not necessary to add up over the trap levels.[6] In conclusion, one finds

$$E_{tM} = E_F + \frac{k_B\,T}{2}\,\log\left(d_t^2\,\frac{p^{\mathrm{eq}}/\alpha_n}{n^{\mathrm{eq}}/\alpha_p}\right). \tag{20.28}$$

An estimate of E_{tM} is easily obtained by considering the non-degenerate condition, whence $n^{\mathrm{eq}} = N_C\,\exp\left[(E_F - E_C)/(k_B\,T)\right]$ and $p^{\mathrm{eq}} = N_V\,\exp\left[(E_V - E_F)/(k_B\,T)\right]$ (compare with (18.28)). It follows

$$E_{tM} \simeq \frac{E_C + E_V}{2} + \frac{k_B\,T}{2}\,\log\left(d_t^2\,\frac{N_V\,\alpha_p}{N_C\,\alpha_n}\right). \tag{20.29}$$

Observing that the second term at the right hand side of (20.29) is small, this result shows that the most efficient trap level is near the gap's midpoint which, in turn, is near the intrinsic Fermi level E_{Fi}. In fact, combining (20.29) with (18.16) yields

$$E_{tM} \simeq E_{Fi} + \frac{k_B\,T}{2}\,\log\left(d_t^2\,\frac{\alpha_p}{\alpha_n}\right) \simeq E_{Fi}. \tag{20.30}$$

[6] This simplification is not applicable in a polycrystalline semiconductor.

Defining the *lifetimes*

$$\tau_{p0} = \frac{1}{N_t\,\alpha_p}, \qquad \tau_{n0} = \frac{1}{N_t\,\alpha_n}, \tag{20.31}$$

gives (20.25) the standard form

$$U_{\rm SRH} = \frac{n\,p - n^{\rm eq}\,p^{\rm eq}}{\tau_{p0}\,(n + n_B) + \tau_{n0}\,(p + p_B)}, \tag{20.32}$$

which is also called *Shockley-Read-Hall recombination function*. In equilibrium it is $U_{\rm SRH}^{\rm eq} = 0$; in a non-equilibrium condition, a positive value of $U_{\rm SRH}$, corresponding to an excess of the $n\,p$ product with respect to the equilibrium product $n^{\rm eq}\,p^{\rm eq}$, indicates that recombinations prevail over generations, and vice versa. In a non-equilibrium condition it may happen that $U_{\rm SRH} = 0$; this occurs at the boundary between a region where recombinations prevail and another region where generations prevail.

In a non-degenerate semiconductor (20.22) become, letting $E_t = E_{tM} = E_{Fi}$ and using (18.12),

$$n_B = \frac{n_i}{d_t}, \qquad p_B = d_t\,n_i, \tag{20.33}$$

whence $n_B\,p_B = n_i^2$. This result is useful also in a degenerate semiconductor for discussing possible simplifications in the form of $U_{\rm SRH}$.

20.2.3.1 Limiting Cases of the Shockley-Read-Hall Theory

The operating conditions of semiconductor devices are often such that the SRH recombination function (20.32) can be reduced to simpler forms. The first case is the so-called *full-depletion condition*, where both electron and hole concentrations are negligibly small with respect to n_B and p_B. Remembering that $n^{\rm eq}\,p^{\rm eq} = n_B\,p_B$ one finds

$$U_{\rm SRH} \simeq -\frac{n_B\,p_B}{\tau_{p0}\,n_B + \tau_{n0}\,p_B} = -\frac{\sqrt{n_B\,p_B}}{\tau_g}, \qquad \tau_g = \sqrt{\frac{n_B}{p_B}}\,\tau_{p0} + \sqrt{\frac{n_B}{p_B}}\,\tau_{n0}. \tag{20.34}$$

In a non-degenerate condition n_B, p_B take the simplified form (20.33), whence $\sqrt{n_B\,p_B} = n_i$ and $\tau_g = \tau_{p0}/d_t + \tau_{n0}\,d_t$. In a full-depletion condition $U_{\rm SRH}$ is always negative, namely, generations prevail over recombinations; for this reason, τ_g is called *generation lifetime*.

The second limiting case of interest is the so-called *weak-injection condition*. This condition occurs when both inequalities below are fulfilled:

$$|n - n^{\rm eq}| \ll c^{\rm eq}, \qquad |p - p^{\rm eq}| \ll c^{\rm eq}, \tag{20.35}$$

where c^{eq} is the equilibrium concentration of the majority carriers in the spatial position under consideration. From the above definition it follows that the concept of weak injection is applicable only after specifying which carriers are the majority ones. Expanding the product $n\,p$ to first order in n and p around the equilibrium value yields $n\,p \simeq n^{\mathrm{eq}}\,p^{\mathrm{eq}} + n^{\mathrm{eq}}\,(p - p^{\mathrm{eq}}) + p^{\mathrm{eq}}\,(n - n^{\mathrm{eq}})$. As a consequence, the numerator of (20.32) becomes

$$n\,p - n^{\mathrm{eq}}\,p^{\mathrm{eq}} \simeq n^{\mathrm{eq}}\,(p - p^{\mathrm{eq}}) + p^{\mathrm{eq}}\,(n - n^{\mathrm{eq}}). \tag{20.36}$$

To proceed, it is necessary to distinguish between the n-type and p-type regions.

Weak-Injection Condition, n-Type Semiconductor
The weak-injection condition (20.35) reads $|n - n^{\mathrm{eq}}| \ll n^{\mathrm{eq}}$, $|p - p^{\mathrm{eq}}| \ll n^{\mathrm{eq}}$. As a consequence, one lets $n \simeq n^{\mathrm{eq}}$ in the denominator of (20.32) and neglects n_B with respect to n^{eq}; in fact, in a non-degenerate condition it is $n_B \simeq n_i \ll n^{\mathrm{eq}}$, and the same inequality is also applicable in a degenerate condition. As the lifetimes are similar to each other, the term $\tau_{n0}\,(p + p_B)$ in the denominator is negligible with respect to $\tau_{p0}\,n^{\mathrm{eq}}$, because p is a concentration of minority carriers and p_B is similar to n_B. In conclusion, the denominator of (20.32) simplifies to $\tau_{p0}\,n^{\mathrm{eq}}$, whence

$$U_{\mathrm{SRH}} \simeq \frac{p - p^{\mathrm{eq}}}{\tau_{p0}} + \frac{n - n^{\mathrm{eq}}}{(n^{\mathrm{eq}}/p^{\mathrm{eq}})\,\tau_{p0}}. \tag{20.37}$$

The second term at the right hand side of (20.37) is negligible[7] because $n^{\mathrm{eq}}/p^{\mathrm{eq}} \gg 1$; letting $\tau_p = \tau_{p0}$ finally yields

$$U_{\mathrm{SRH}} \simeq \frac{p - p^{\mathrm{eq}}}{\tau_p}, \tag{20.38}$$

with τ_p the *minority-carrier lifetime* in an n-doped region.

Weak-Injection Condition, p-Type Semiconductor
The weak-injection condition (20.35) reads $|n - n^{\mathrm{eq}}| \ll p^{\mathrm{eq}}$, $|p - p^{\mathrm{eq}}| \ll p^{\mathrm{eq}}$. As a consequence, one lets $p \simeq p^{\mathrm{eq}}$ in the denominator of (20.32) and neglects p_B with respect to p^{eq}; the other term in the denominator of (20.35) is neglected as above, this simplifying the denominator to $\tau_{n0}\,p^{\mathrm{eq}}$. In conclusion,

$$U_{\mathrm{SRH}} \simeq \frac{p - p^{\mathrm{eq}}}{(p^{\mathrm{eq}}/n^{\mathrm{eq}})\,\tau_{n0}} + \frac{n - n^{\mathrm{eq}}}{\tau_{n0}}. \tag{20.39}$$

The first term at the right hand side of (20.39) is negligible because $p^{\mathrm{eq}}/n^{\mathrm{eq}} \gg 1$; letting $\tau_n = \tau_{n0}$ finally yields

$$U_{\mathrm{SRH}} \simeq \frac{n - n^{\mathrm{eq}}}{\tau_n}, \tag{20.40}$$

with τ_n the *minority-carrier lifetime* in a p-doped region.

[7] Considering for instance the example given in Sect. 18.4.1, one has $n^{\mathrm{eq}} \simeq 10^{15}$ cm^{-3}, $p^{\mathrm{eq}} \simeq 10^{5}$ cm^{-3}, whence $n^{\mathrm{eq}}/p^{\mathrm{eq}} \simeq 10^{10}$.

The simplified expressions of U_{SRH} found here are particularly useful; in fact, in contrast to (20.32), the weak-injection limits (20.38) and (20.40) are linear with respect to p or n. Moreover, as (20.38) and (20.40) depend on one unknown only, they decouple the continuity equation of the minority carriers (the first one in (19.129) or in (19.130)) from the other equations of the semiconductor's model; thanks to this it is possible to separate the system of equations. The simplification introduced by the full-depletion condition is even stronger, because (20.34) is independent of the model's unknowns. On the other hand, all simplifications illustrated here are applicable only in the regions where the approximations hold; once the simplified model's equations have been solved locally, it is necessary to match the solutions at the boundaries between adjacent regions.

20.3 Auger Recombination and Impact Ionization

An important, non-thermal recombination mechanism is *Auger recombination*. The phenomenon is due to the electron-electron or hole-hole collision, and is illustrated in Fig. 20.3. With reference to case *a*, two electrons whose initial state is in the conduction band collide and exchange energy. The outcome of the collision is that one of the electrons suffers an energy loss equal or larger than the energy gap and makes a transition to an empty state of the valence band; the other electron absorbs the same amount of energy and makes a transition to a higher-energy state of the conduction band. The phenomenon is also indicated as an Auger recombination *initiated by electrons*. The analogue for holes is shown in case *c* of Fig. 20.3: two holes whose initial state is in the valence band collide and exchange energy. Remembering that hole energy increases in the opposite direction with respect to that of electrons (Sect. 19.2.3), the hole that suffers an energy loss equal or larger than the energy gap makes a transition to a filled state of the conduction band; the other hole absorbs the same amount of energy and makes a transition to a higher-energy state of the valence band. The phenomenon is indicated as an Auger recombination *initiated by holes*.

The phenomenon dual to Auger recombination is illustrated in Fig. 20.4 and is called *impact ionization*. With reference to case *b*, an electron whose initial state is in the conduction band at high energy collides and exchanges energy with an electron whose initial state is in the valence band. The initial energy E of the electron in the conduction band is such that $E - E_C$ is equal or larger than the energy gap, whereas the initial energy of the electron in the valence band is near E_V. The outcome of the collision is that, although the high-energy electron suffers an energy loss equal or larger than the energy gap, its final state is still in the conduction band; the other electron absorbs the same amount of energy and makes a transition to the conduction band. The phenomenon is in fact an electron-hole pair generation and is also indicated as an impact-ionization event *initiated by electrons*. The analogue for holes is shown in case *d* of Fig. 20.4: a hole whose initial state is in the valence band at high energy collides and exchanges energy with a hole whose initial state is in the conduction band. The initial energy E of the hole in the valence band is such that $|E - E_V|$

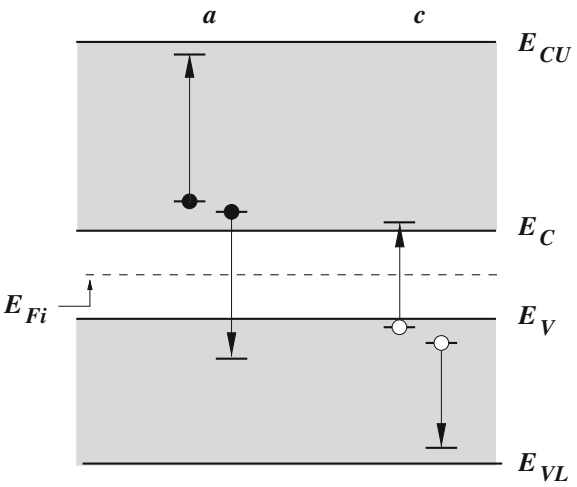

Fig. 20.3 Auger recombinations initiated by electrons (*a*) and holes (*c*)

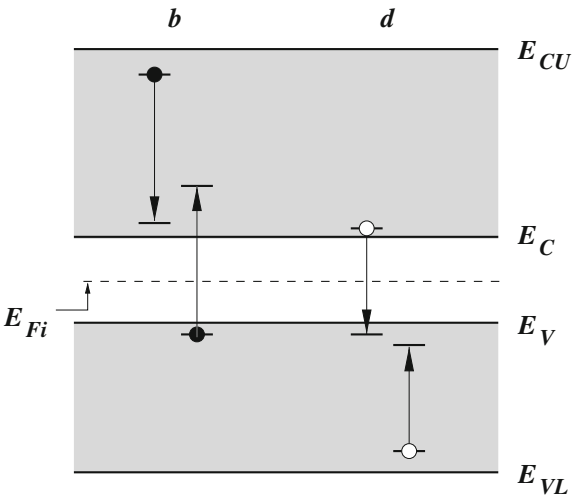

Fig. 20.4 Impact-ionization transitions initiated by electrons (*b*) and holes (*d*)

is equal or larger than the energy gap, whereas the initial energy of the hole in the conduction band is near E_C. The outcome of the collision is that, although the high-energy hole suffers an energy loss equal or larger than the energy gap, its final state is still in the valence band; the other hole absorbs the same amount of energy and makes a transition to the valence band. The phenomenon is in fact an electron-hole pair generation and is also indicated as an impact-ionization event *initiated by holes*.

The derivation of the Auger and impact-ionization rates is shown in the complements; here the expressions of the net recombinations due to the Auger and impact-ionization events are given, that read

$$U_n^{AI} = r_a - r_b = c_n\, n^2\, p - I_n\, n, \qquad U_p^{AI} = r_c - r_d = c_p\, p^2\, n - I_p\, p, \quad (20.41)$$

where U_n^{AI} refers to the electron-initiated transitions and U_p^{AI} to the hole-initiated ones. As usual, r_a indicates the number of transitions of type a per unit time and volume; the same holds for r_b, r_c, and r_d. In (20.41), c_n, I_n are the transition coefficients for the Auger recombination and impact ionization initiated by electrons, and c_p, I_p the analogue for holes; c_n, c_p are also called *Auger coefficients*.[8] In equilibrium it is $U_n^{AI} = U_p^{AI} = 0$, whence $I_n = c_n\, n^{\mathrm{eq}}\, p^{\mathrm{eq}}$, $I_p = c_p\, n^{\mathrm{eq}}\, p^{\mathrm{eq}}$. The above hold also in a non-equilibrium case as long as the operating conditions are not too far from equilibrium; with these premises it follows

$$U_n^{AI} = c_n\, n\ (n\ p - n^{\mathrm{eq}}\, p^{\mathrm{eq}})\,, \qquad U_p^{AI} = c_p\, p\ (np - n^{\mathrm{eq}}\, p^{\mathrm{eq}})\,, \tag{20.42}$$

When the operating condition departs strongly from equilibrium, the simplification leading to (20.42) is no longer applicable and the general expressions (20.41) must be used. Referring to all recombinations as due to transitions of electrons, their rate is easily found to be $r_a + r_c$; similarly, the total generation rate is $r_b + r_d$. In conclusion, the net recombination rate due to the Auger and impact-ionization phenomena is given by

$$U_{AI} = U_n^{AI} + U_p^{AI}. \tag{20.43}$$

For Auger recombination to occur it is necessary that an electron collides with another electron, or a hole collides with another hole. The probability of such an event is relatively small because in normal operating conditions and at room temperature there is a high probability that a carrier collides with a phonon; as a consequence, for the collisionless motion of an electron to be interrupted by a collision with another electron it is necessary that the electron concentration be very high. This situation occurs only in a heavily-doped, n-type region; similarly, an Auger recombination initiated by holes can be significant only in a heavily-doped, p-type region.[9]

Considering now the case of impact-ionization, for this phenomenon to occur it is necessary that an electron, or a hole, acquires a kinetic energy larger than the energy gap. This is a rare event as well,[10] because in general the carrier undergoes a phonon collision when its kinetic energy is still significantly lower than the energy gap. The impact-ionization event occurs only if the carrier acquires a substantial energy over a distance much shorter than the average collisionless path, which happens only in presence of a strong electric field.[11]

The qualitative reasoning outlined above explains why the conditions for a strong Auger recombination are incompatible with those that make impact-ionization

[8] The units are $[c_{n,p}] = \mathrm{cm}^6\ \mathrm{s}^{-1}$ and $[I_{n,p}] = \mathrm{s}^{-1}$.

[9] In fact, Auger recombination becomes significant in the source and drain regions of MOSFETs and in the emitter regions of BJTs, where the dopant concentration is the highest.

[10] In principle, high-energy electrons or hole exist also in the equilibrium condition; however, their number is negligible because of the exponentially-vanishing tail of the Fermi-Dirac statistics.

[11] The high-field conditions able to produce a significant impact ionization typically occur in the reverse-biased p-n juctions like, e.g., the drain junction in MOSFETs and the collector junction in BJTs.

dominant; in fact, a large charge density, like that imposed by a heavy dopant concentration, prevents the electric field from becoming strong. Vice versa, a strong electric field prevents a large carrier concentration from building up. It is therefore sensible to investigate situations where only one term dominates within U_{AI}.

20.3.1 Strong Impact Ionization

As indicated in Sect. 20.3, far from equilibrium the approximations $I_n = c_n\, n^{\rm eq}\, p^{\rm eq}$, $I_p = c_p\, n^{\rm eq}\, p^{\rm eq}$ are not valid, and the general expressions (20.41) must be used. Here the situation where impact ionization dominates over the other generation-recombination mechanisms is considered, using the steady-state case. If impact ionization is dominant, it is $U_n - G_n = U_p - G_p \simeq U_{AI} \simeq -I_n\, n - I_p\, p$. The continuity equations (the first ones in (19.129) and (19.130)) then become

$$\mathrm{div}\mathbf{J}_n = -q\, I_n\, n - q\, I_p\, p, \qquad \mathrm{div}\mathbf{J}_p = q\, I_n\, n + q\, I_p\, p. \tag{20.44}$$

As outlined in Sect. 20.3, impact-ionization dominates if the electric field is high. For this reason, the transport equations in (19.129) and (19.130) are simplified by keeping the ohmic term only, to yield $\mathbf{J}_n \simeq q\, \mu_n\, n\, \mathbf{E}$ and $\mathbf{J}_p \simeq q\, \mu_p\, p\, \mathbf{E}$. As a consequence, the electron and hole current densities are parallel to the electric field. Let $\mathbf{e}(\mathbf{r})$ be the unit vector of the electric field, oriented in the direction of increasing field, $\mathbf{E} = |\mathbf{E}|\, \mathbf{e}$; it follows $\mathbf{J}_n = J_n\, \mathbf{e}$ and $\mathbf{J}_p = J_p\, \mathbf{e}$, with J_n and J_p strictly positive. Extracting n, p from the above and replacing them into (20.44) yields

$$-\mathrm{div}\mathbf{J}_n = k_n\, J_n + k_p\, J_p, \qquad \mathrm{div}\mathbf{J}_p = k_n\, J_n + k_p\, J_p, \tag{20.45}$$

where the ratios

$$k_n = \frac{I_n}{\mu_n\, |\mathbf{E}|}, \qquad k_p = \frac{I_p}{\mu_p\, |\mathbf{E}|}, \tag{20.46}$$

whose units are $[k_{n,p}] = \mathrm{m}^{-1}$, are the *impact-ionization coefficients* for electrons and holes, respectively. Equations (20.45) form a system of differential equations of the first order, whose solution in the one-dimensional case is relatively simple if the dependence of the coefficients on position is given (Sect. 21.5).

20.4 Optical Transitions

The description of the optical transitions is similar to that of the direct thermal transitions given in Sect. 20.2.1; still with reference to Fig. 20.1, the transition marked with a can be thought of as an optical-recombination event if the energy difference between the initial and final state is released to the environment in the form of a photon. The opposite transition (b), where the electron's energy increases due to

photon absorption from the environment, is an optical electron-hole generation. The expression of the net optical-recombination rate is similar to (20.11) and reads

$$U_O = \alpha_O \, n \, p - G_O, \tag{20.47}$$

whose coefficients are derived in the same manner as those of U_{DT} (Sect. 20.2.1).

In normal operating conditions the similarity between the direct-thermal and optical generation-recombination events extends also to the external agent that induces the transitions. In fact, the distribution of the phonon energies is typically the equilibrium one, given by the Bose–Einstein statistics (15.55) at the lattice temperature; as for the photons, the environment radiation in which the semiconductor is immersed can also be assimilated to the equilibrium one, again given by the Bose-Einstein statistics at the same temperature. The conditions of the optical generation-recombination events drastically change if the device is kept far from equilibrium. Consider for instance the case where the electron concentration of the conduction band is artificially increased with respect to the equilibrium value at the expense of the electron population of the valence band, so that both n and p in (20.47) increase. This brings about an excess of recombinations; if the probability of radiative-type generation-recombination events is high,[12] the emission of a large number of photons follows. The angular frequencies of the emitted photons are close to $(E_C - E_V)/\hbar$, because the majority of the electrons in the conduction band concentrate near E_C, and the final states of the radiative transitions concentrate near E_V. In this way, the energy spent to keep the artificially-high concentration of electron-hole pairs is transformed into that of a nearly-monochromatic optical emission. In essence, this is the description of the operating principle of a *laser*.[13] Another method for keeping the device far from equilibrium is that of artificially decreasing both the concentration of electrons of the conduction band and the concentration of holes of the valence band. The outcome is opposite with respect to that described earlier: the decrease of both n and p in (20.47) brings about an excess of generations, which in turn corresponds to the absorption of photons from the environment. The absorption may be exploited to accumulate energy (this leading to the concept of *solar cell*), or to provide an electrical signal whose amplitude depends on the number of absorbed photons (this leading to the concept of *optical sensor*).

In a non-equilibrium condition the amount of energy exchanged between the semiconductor and the electromagnetic field is not necessarily uniform in space. Consider, by way of example, the case of an optical sensor on which an external radiation impinges; as the non-equilibrium conditions are such that the absorption events prevail, the radiation intensity within the material progressively decreases at increasing distances from the sensor's surface. Therefore, it is important to determine the radiation intensity as a function of position.

[12] As indicated in Sect. 17.6.6, among semiconductors this is typical of the direct-gap ones.

[13] In fact, LASER is the acronym of Light Amplification by Stimulated Emission of Radiation.

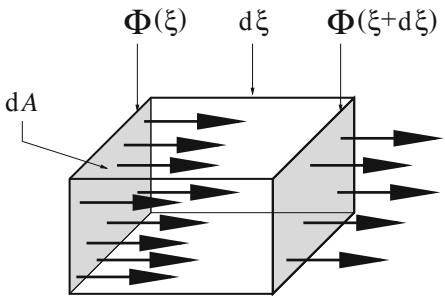

Fig. 20.5 Sketch of photon absorption in a material layer

It is acceptable to assume that the absorption events are uncorrelated from each other. Thus, one can limit the analysis to a monochromatic radiation; the effect of the whole spectrum is recovered at a later stage by adding up over the frequencies. When absorption prevails, (20.47) simplifies to $U_O \simeq -G_O$, where G_O is a function of the radiation's frequency ν and possibly of position. If the radiation's intensity varies with time, G_O depends on time as well.[14] When the radiation interacts with the external surface of the material, part of the energy is reflected; moreover, the radiation is refracted at the boundary, so that the propagation direction outside the material differs in general from that inside. Letting ξ be the propagation direction inside the material, consider an elementary volume with a side $\mathrm{d}\xi$ aligned with ξ and a cross-section $\mathrm{d}A$ normal to it (Fig. 20.5). The monochromatic radiation can be described as a flux of photons of equal energy $h\,\nu$, with h the Planck constant, and a momentum's direction parallel to ξ. Let $\Phi(\xi)$ be the flux density of photons entering the volume from the face corresponding to ξ, and $\Phi(\xi + \mathrm{d}\xi)$ the flux density leaving it at $\xi + \mathrm{d}\xi$; the following holds, $\Phi = K\,u_f$, where $K(\xi)$ is the concentration of the photons and u_f their constant phase velocity. Then,

$$\frac{\partial \Phi}{\partial \xi} = \frac{\partial K}{\partial (\xi/u_f)} = \frac{\partial K}{\partial t}. \tag{20.48}$$

The derivatives in (20.48) are negative because the photon concentration decreases in time due to absorption; as the loss of each photon corresponds to the loss of an energy quantum $h\,\nu$, the loss of electromagnetic energy per unit volume and time is $-h\,\nu\,(\partial\Phi/\partial\xi)$. By a similar token one finds[15] that the energy absorbed by the optical-generation events per unit time and volume is $h\,\nu\,G_O$. The latter is not necessarily equal to $-h\,\nu\,(\partial\Phi/\partial\xi)$; in fact, some photons crossing the elementary volume may be lost due to collisions with nuclei (this, however, is a rare event), or with electrons that are already in the conduction band, so that no electron-hole pair generation

[14] In principle, a time-dependence of the intensity is incompatible with the hypothesis that the radiation is monochromatic. However, the frequency with which the intensity may vary is extremely small with respect to the optical frequencies.

[15] It is implied that $h\,\nu \geq E_C - E_V$, and that the two-particle collisions only are to be considered.

occurs. To account for these events one lets

$$G_O = -\eta \frac{\partial \Phi}{\partial \xi} > 0, \tag{20.49}$$

with $0 < \eta < 1$ the *quantum efficiency*. In moderately-doped semiconductors η is close to unity because the concentration of the conduction-band electrons is small; instead, the efficiency degrades in degenerate semiconductors. The spatial dependence of the generation term can be derived from (20.49) if that of the photon flux is known. To proceed, one defines the *absorption coefficient* as

$$k = -\frac{1}{\Phi} \frac{\partial \Phi}{\partial \xi} > 0, \tag{20.50}$$

with $[k] = \mathrm{m}^{-1}$. In general it is $k = k(\Phi, \xi, \nu)$; however, as the absorption effects are uncorrelated, the flux density lost per unit path $\mathrm{d}\xi$ is proportional to the flux density available at ξ. Then, k is independent of Φ; neglecting momentarily the dependence on ξ as well, one finds

$$\Phi(\xi) = \Phi_B \, \exp[-k(\nu)\,\xi], \tag{20.51}$$

with $\Phi_B = \Phi(\xi = 0^+)$ on account of the fact that, due to the reflection at the interface, the flux density on the inside edge of the boundary is different from that on the outside edge. When k is independent of position, its inverse $1/k$ is called *average penetration length of the radiation*. When k depends on position, (20.50) is still separable and yields

$$\Phi(\xi) = \Phi_B \, \exp(-k_m\,\xi), \qquad k_m = \frac{1}{\xi} \int_0^{\xi} k(\xi'; \nu)\, \mathrm{d}\xi'. \tag{20.52}$$

Combining (20.52) with (20.49), the optical-generation term is found to be

$$G_O = \eta\, \Phi_B\, k(\xi, \nu) \, \exp\left[-\int_0^{\xi} k(\xi', \nu)\, \mathrm{d}\xi' \right]. \tag{20.53}$$

20.5 Macroscopic Mobility Models

It has been shown in Sect. 19.5.2 that the carrier mobilities are defined in terms of the momentum-relaxation times. Specifically, in the parabolic-band approximation it is, for the electrons of the conduction band, $\mu_n = (\mu_l + 2\,\mu_t)/3$, with $\mu_l = q\,\tau_p/m_l$, $\mu_t = q\,\tau_p/m_t$, where τ_p is the electron momentum-relaxation time (19.87); similarly, for the holes of the valence band the carrier mobility is given by inserting (19.118) into the second relation of (19.121), namely, a linear combination of the heavy-hole and light-hole momentum-relaxation times. As, in turn, the inverse momentum-relaxation time is a suitable average of the inverse intra-band relaxation time, the

Matthiessen rule follows (Sect. 19.6.5); in conclusion, the electron and hole mobilities are calculated by combining the effects of the different types of collisions (e.g., phonons, impurities, and so on) suffered by the carrier.[16] In the case of electrons, the application of the Matthiessen rule is straightforward, leading to

$$\frac{1}{\mu_n} = \frac{m_n}{q}\left(\frac{1}{\tau_p^{\mathrm{ph}}} + \frac{1}{\tau_p^{\mathrm{imp}}} + \dots\right), \tag{20.54}$$

where the index refers to the type of collision, and $1/m_n = (1/m_l + 2/m_t)/3$. For holes a little more algebra is necessary, which can be avoided if the approximation $\tau_{ph} \simeq \tau_{pl}$ is applicable.

In the typical operating conditions of semiconductor devices the most important types of collisions are those with phonons and ionized impurities. For devices like surface-channel MOSFETs, where the flow lines of the current density are near the interface between semiconductor and gate insulator, a third type is also very important, namely, the collisions with the interface. The *macroscopic mobility models* are closed-form expressions in which mobility is related to a set of macroscopic parameters (e.g., temperature) and to some of the unknowns of the semiconductor-device model; the concept is similar to that leading to the expressions of the generation-recombination terms shown in earlier sections.

20.5.1 Example of Phonon Collision

By way of example, a simplified analysis of the contribution to mobility of the electron-phonon collision is outlined below, starting from the definition of the ith component of the momentum-relaxation tensor τ_{pi} given by (19.87); the simplifications are such that the first-order expansion $f - f^{\mathrm{eq}} \simeq (\mathrm{d}f/\mathrm{d}\lambda)^{\mathrm{eq}}\,\lambda$ is not used here. Starting from the perturbative form (19.47) one considers the steady-state, uniform case and lets $\mathbf{B} = 0$, $\tau = \tau_v$, to find

$$\frac{q}{\hbar}\,\mathbf{E}\cdot\mathrm{grad}_{\mathbf{k}} f = \frac{f - f^{\mathrm{eq}}}{\tau_v}. \tag{20.55}$$

Replacing f with f^{eq} at the left hand side of (20.55), and using the definition (17.52) of the group velocity, yields $\mathrm{grad}_{\mathbf{k}} f^{\mathrm{eq}} = (\mathrm{d}f^{\mathrm{eq}}/\mathrm{d}H)\,\hbar\,\mathbf{u}$, with H the Hamiltonian function defined in Sect. 19.2.2. Inserting into (19.87) yields

$$\tau_{pi} \iiint_{-\infty}^{+\infty} u_i\,\mathbf{E}\cdot\mathbf{u}\,(\mathrm{d}f^{\mathrm{eq}}/\mathrm{d}H)\,\mathrm{d}^3k = \iiint_{-\infty}^{+\infty} u_i\,\mathbf{E}\cdot\mathbf{u}\,(\mathrm{d}f^{\mathrm{eq}}/\mathrm{d}H)\,\tau_v\,\mathrm{d}^3k. \tag{20.56}$$

As the derivative $\mathrm{d}f^{\mathrm{eq}}/\mathrm{d}H$ is even with respect to $\mathbf{k}$, the integrals involving velocity components different from u_i vanish because the corresponding integrand is odd;

[16] As mentioned in Sect. 19.6.5, it is assumed that the different types of collisions are incorrelated.

as a consequence, only the ith component of the electric field remains, and cancels out. A further simplification is obtained by replacing the Fermi-Dirac statistics with the Maxwell-Boltzmann distribution law, $f^{\mathrm{eq}} \simeq Q \exp[(-E_e + q\,\varphi - E_C + E_F)/(k_B\,T)]$, to find

$$\tau_{pi} \iiint_{-\infty}^{+\infty} u_i^2 \exp[-E_e/(k_B\,T)]\,\mathrm{d}^3k = \iiint_{-\infty}^{+\infty} u_i^2 \exp[-E_e/(k_B\,T)]\,\tau_v\,\mathrm{d}^3k. \tag{20.57}$$

To proceed it is necessary to make an assumption about τ_v. Remembering the definition of the relaxation time given by the first relation in (19.43), it is reasonable to assume that the scattering probability S_0 increases with the kinetic energy E_e of the electron, so that the relaxation time decreases; a somewhat stronger hypothesis is that the relaxation time depends on E_e only, namely, the collision is isotropic.[17]

In this case, (20.57) is readily manipulated by a Herring-Vogt transformation. Following the same procedure as in Sect. 19.6.4, one finds that all numerical factors cancel out; as a consequence, one may replace the auxiliary coordinate η_i^2 with $\eta^2/3 = E_e/3$, this showing that $\tau_{pi} = \tau_p$ is isotropic as well. One eventually finds

$$\tau_p = \frac{\int_0^{+\infty} \tau_v(E_e)\; E_e^{3/2} \exp[-E_e/(k_B\,T)]\,\mathrm{d}E_e}{\int_0^{+\infty} E_e^{3/2} \exp[-E_e/(k_B\,T)]\,\mathrm{d}E_e}. \tag{20.58}$$

A simple approximation for the relaxation time is $\tau_v = \tau_{v0}\,(E_e/E_0)^{-\alpha}$, where τ_{v0}, E_0, and α are positive parameters independent of E_e. From (C.88) it follows

$$\tau_p = \tau_{v0}\,\frac{\Gamma(5/2-\alpha)}{\Gamma(5/2)}\left(\frac{E_0}{k_B\,T}\right)^{\alpha}. \tag{20.59}$$

When the electron-phonon interaction is considered, $\tau_{v0} = \tau_{v0}^{\mathrm{ph}}$ is found to be inversely proportional to $k_B\,T$ and to the concentration N_{sc} of semiconductor's atoms; moreover, for acoustic phonons[18] it is $\alpha = 1/2$ [62, Sects. 61, 62], whence

$$\tau_p^{\mathrm{ap}} = \tau_{v0}(N_{\mathrm{sc}}, T)\,\frac{4}{3\sqrt{\pi}}\left(\frac{E_0}{k_B\,T}\right)^{1/2}, \qquad \mu_n^{\mathrm{ap}} \propto N_{\mathrm{sc}}^{-1}\,(k_B\,T)^{-3/2}, \tag{20.60}$$

where "ap" stands for "acoustic phonon". More elaborate derivations, including also the contribution of optical phonons, still show that carrier-phonon collisions make mobility to decrease when temperature increases.

---

[17] The first-principle derivation of the scattering probabilities is carried out by applying Fermi's Golden Rule (Sect. 14.8.3) to each type of perturbation, using the Bloch functions for the unperturbed states [57]. An example is given later in the case of ionized-impurity scattering.

[18] Acoustic phonons are those whose momentum and energy belong to the acoustic branch of the lattice-dispersion relation (Sect. 17.8.5); a similar definition applies to optical phonons (Sect. 17.8.6).

20.5.2 Example of Ionized-Impurity Collision

As a second example one considers the collisions with ionized impurities. The interaction with a single ionized impurity is a perturbation of the Coulomb type; due to the presence of the crystal, the more suitable approach is the screened Coulomb perturbation, an example of which is shown in Sect. 14.7, leading to the perturbation-matrix element (14.34):

$$h^{(0)}_{\mathbf{kg}} = \frac{A/(2\,\pi)^3}{q_c^2 + q^2}, \qquad A = \frac{\kappa\, Z\, e^2}{\varepsilon_0}. \tag{20.61}$$

In (20.61), $e > 0$ is the elementary electric charge, Z a positive integer, ε_0 the vacuum permittivity, $q_c > 0$ the inverse screening length,[19] $q = |\mathbf{q}| = |\mathbf{k} - \mathbf{g}|$ and, finally, $\kappa = 1\,(-1)$ in the repulsive (attractive) case. The wave vectors $\mathbf{k}$ and $\mathbf{g}$ correspond to the initial and final state of the transition, respectively. In principle, (20.61) should not be used as is because it holds *in vacuo*; in fact, the eigenfunctions of the unperturbed Hamiltonian operator used to derive (20.61) are plane waves. Inside a crystal, instead, one should define the perturbation matrix $h_{\mathbf{kg}}(t)$ using the Bloch functions $w_{\mathbf{k}} = u_{\mathbf{k}}\, \exp(\mathrm{i}\,\mathbf{k}\cdot\mathbf{r})$ in an integral of the form (14.24). However, it can be shown that the contribution of the periodic part $u_{\mathbf{k}}$ can suitably be averaged and extracted from the integral, in the form of a dimensionless coefficient, whose square modulus G is called *overlap factor*. For this reason, the collisions with ionized impurities is treated starting from the definition (20.61) to calculate the perturbation matrix, with the provision that the result is to be multiplied by G and the permittivity $\varepsilon_{\mathrm{sc}}$ of the semiconductor replaces ε_0 in the second relation of (20.61).

Like in Sect. 14.6, a Gaussian wave packet (14.27) centered on some wave vector $\mathbf{b} \neq \mathbf{g}$ is used as initial condition. In this case the perturbation is independent of time, $h_{\mathbf{bg}} = h^{(0)}_{\mathbf{bg}} = \text{const} \neq 0$; as a consequence, the infinitesimal probability $\mathrm{d}P_{\mathbf{b}}$ that such a perturbation induces a transition, from the initial condition (14.27), to a final state whose energy belongs to the range $\mathrm{d}E_{\mathbf{g}}$, is given by (14.32). In turn, the integral (14.32) providing $H^{(0)}_{\mathbf{b}}(E_{\mathbf{g}})$ is calculated in Problem 14.1. Assuming that the duration t_P of the interaction is large enough to make Fermi's Golden Rule (14.44) applicable, and inserting the overlap factor, one finally obtains

$$\mathrm{d}P_{\mathbf{b}} \approx G \left(\frac{2\,\pi\, m}{\hbar^2}\right)^{3/2} \frac{8\,\pi\, t_P\, \delta(E_{\mathbf{b}} - E_{\mathbf{g}})\, A^2}{\lambda^3\, \hbar\, (2\,\pi)^5\, q_c^2\, (q_c^2 + 8\, m\, E_{\mathbf{g}}/\hbar^2)} \sqrt{E_{\mathbf{g}}}\; \mathrm{d}E_{\mathbf{g}}, \tag{20.62}$$

where the relation $E_{\mathbf{g}} = \hbar^2\, g^2/(2\, m)$ has been used. Integrating over $E_{\mathbf{g}}$ and dividing by t_P provides the probability per unit time of a transition from the initial energy $E_{\mathbf{b}}$ to any final energy; letting $E_c = \hbar^2\, q_c^2/(2\, m)$, one finds

$$\dot{P}(E_{\mathbf{b}}) = \frac{1}{\tau_{vc}} \frac{\sqrt{4\, E_{\mathbf{b}}/E_c}}{1 + 4\, E_{\mathbf{b}}/E_c}. \qquad \frac{1}{\tau_{vc}} = \frac{G\, A^2/\sqrt{2\,\pi\, m}}{8\,\pi^2\, (\lambda^2\, E_c)^{3/2}}. \tag{20.63}$$

[19] An example of derivation of the screening length is given in Sect. 20.6.4.

The above expression provides the contribution to the intra-band relaxation time of the scattering due to a single impurity. One notes that, since A is squared, the effect onto (20.63) of a positive impurity is the same as that of a negative one. If the effect of each impurity is uncorrelated with that of the others,[20] the probabilities add up; letting $N_I = N_D^+ + N_A^-$ be the total concentration of ionized impurities, the product $N_I \ \mathrm{d}^3 r$ is the total number ionized impurities in the elementary volume $\mathrm{d}^3 r$; it follows that the probability per unit time and volume is given by $\dot{P}(E_{\mathbf{b}})\, N_I \ \mathrm{d}^3 r$. Considering that N_I depends on position only, mobility inherits the inverse proportionality with N_I; letting "ii" indicate "ionized impurity", one finds $\mu_n^{\mathrm{ii}} \propto 1/N_I$.

The derivation of the dependence on N_I shown above is in fact oversimplified, and the resulting model does not reproduce the experimental results with sufficient precision. One of the reasons for this discrepancy is that the inverse screening length q_c depends on the dopant concentration as well. In order to improve the model, while still keeping an analytical form, the model is modified by letting $1/\mu_n^{\mathrm{ii}} \propto N_I^{\alpha}$, with α a dimensionless parameter to be extracted from the comparison with experiments. One then lets

$$\frac{1}{\mu_n^{\mathrm{ii}}(N_I)} = \frac{1}{\mu_n^{\mathrm{ii}}(N_R)} \left(\frac{N_I}{N_R}\right)^{\alpha}, \tag{20.64}$$

with N_R a reference concentration.

20.5.3 Bulk and Surface Mobilities

Combining the phonon and ionized-impurity contributions using the Matthiessen rule yields $1/\mu_n^{\mathrm{B}}(T, N_I) = 1/\mu_n^{\mathrm{ph}}(T) + 1/\mu_n^{\mathrm{ii}}(N_I)$, namely,

$$\mu_n^{\mathrm{B}}(T, N_I) = \frac{\mu_n^{\mathrm{ph}}(T)}{1 + c(T)\,(N_I/N_R)^{\alpha}}, \tag{20.65}$$

with $c(T) = \mu_n^{\mathrm{ph}}(T)/\mu_n^{\mathrm{ii}}(N_R)$. In practical cases the doping concentration ranges over many orders of magnitude; for this reason, (20.65) is usually represented in a semilogarithmic scale: letting $r = \log_{10}(N_I/N_R)$, $b = \alpha \log_e 10$, and $b_0 = \log_e c$, (20.65) becomes

$$\mu_n^{\mathrm{B}}(T, N_I) = \frac{\mu_n^{\mathrm{ap}}(T)}{1 + \exp(br + b_0)}. \tag{20.66}$$

The curves corresponding to $b = 1,\ 1.5,\ 3$ and $b_0 = 0$ are drawn in Fig. 20.6, using r as independent variable at a fixed T. Index "B" in the mobility defined in (20.65)

[20] In silicon, this assumption is fulfilled for values of the concentration up to about 10^{19} cm^{-3} [64, 84].

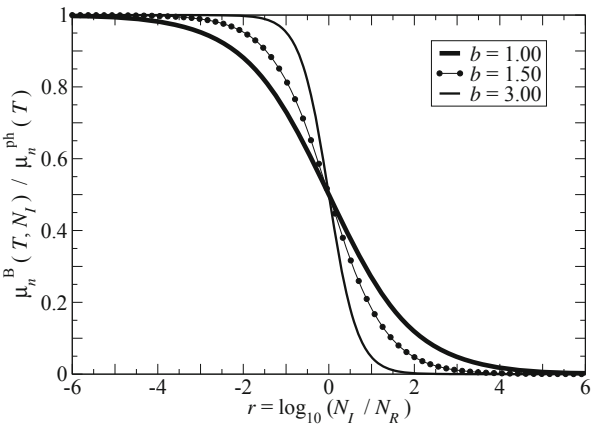

Fig. 20.6 Graph of the theoretical mobility curve (20.66), normalized to its maximum, for different values of b, with $b_0 = 0$. Each curve has a flex at $r = r_{\text{flex}} = -b_0/b$ and takes the value 0.5 there. The slope at the flex is $-b/4$

or (20.66) stands for "bulk". More generally, the term *bulk mobility* is ascribed to the combination of all contributions to mobility different from surface collisions.

As mentioned at the beginning of this section, in surface-channel devices the degradation of mobility produced by the interaction of the carriers with the interface between channel and gate insulator is also very important. The macroscopic models of this effect are built up by considering that the carrier-surface interaction is more likely to occur if the flow lines of the current density are closer to the interface itself; such a closeness is in turn controlled by the intensity of the electric field's component normal to the interface, $E_\perp$. In conclusion, the model describes the contribution to mobility due to surface scattering as a decreasing function of $E_\perp$, e.g., for electrons,

$$\frac{1}{\mu_n^{\text{s}}(E_\perp)} = \frac{1}{\mu_n^{\text{s}}(E_R)} \left(\frac{E_\perp}{E_R}\right)^\beta , \tag{20.67}$$

with E_R a reference field and β a dimensionless parameter to be extracted from experiments. Combining the bulk and surface contributions using the Matthiessen rule yields $1/\mu_n(T, N_I, E_\perp) = 1/\mu_n^{\text{B}}(T, N_I) + 1/\mu_n^{\text{s}}(E_\perp)$, namely,

$$\mu_n(T, N_I, E_\perp) = \frac{\mu_n^{\text{B}}(T, N_I)}{1 + d(T, N_I)\,(E_\perp/E_R)^\beta} , \tag{20.68}$$

with $d(T, N_I) = \mu_n^{\text{B}}(T, N_I)/\mu_n^{\text{s}}(E_R)$.

20.5.4 Beyond Analytical Modeling of Mobility

In general the analytical approaches outlined above do not attain the precision necessary for applications to realistic devices. For this reason, one must often resort to numerical-simulation methods; in this way, the main scattering mechanisms are

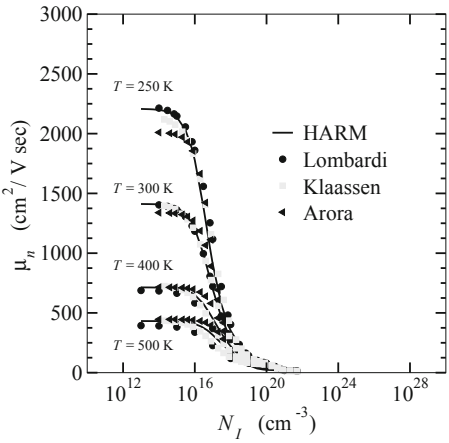

Fig. 20.7 Electron mobility in silicon calculated with the spherical-harmonics expansion method (HARM) as a function of the total ionized-dopant concentration N_I, using the lattice temperature T as parameter. The calculations are compared with measurements by Lombardi [73], Klaassen [64], and Arora [1] (courtesy of S. Reggiani)

incorporated into the analysis (e.g., for silicon: acoustic phonons, optical phonons, ionized impurities, and impact ionization), along with the full-band structure of the semiconductor, which is included in the simulation through the density of states and group velocity defined in the energy space. The latter, in turn, are obtained directly from the corresponding functions in the momentum space by integrating the full-band system over the angles. The energy range considered to date allows for the description of carrier dynamics up to 5 eV.

As mentioned above, the ionized-impurity collisions can be treated as interaction between the carrier and a single impurity as long as the impurity concentration is below some limit. When the limit is exceeded, impurity clustering becomes relevant and must be accounted for [64]. In fact, at high doping densities the carrier scatters with a cluster of K ions, where K is a function of the impurity concentration. Finally, different outcomes are found for majority- or minority-mobility calculations: e.g., minority-hole mobility is found to be about a factor 2 higher than the majority-hole mobility for identical doping levels.

Figures 20.7 and 20.8 show the outcome of electron- and hole-mobility calculations for bulk silicon, obtained from the spherical-harmonics method illustrated in [112]. The method incorporates the models for the scattering mechanisms listed above. The electron and hole mobility have been calculated as a function of the total ionized-dopant concentration N_I, using the lattice temperature T as a parameter; in the figures, they are compared with measurements taken from the literature.

To include the surface effects in the analysis it is necessary to account for the fact that in modern devices the thickness of the charge layer at the interface with the gate insulator is so small that quantum confinement and formation of subbands must be considered. The typical collisions mechanisms to be accounted for at the semiconductor-insulator interface are surface roughness, scattering with ionized impurities trapped at the interface, and surface phonons. Figures 20.9 and 20.10 show the outcome of electron and hole surface-mobility calculations in silicon, also obtained from the spherical-harmonics method [84]. The electron and hole mobility

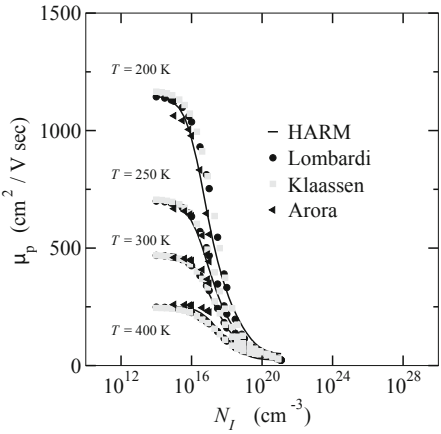

Fig. 20.8 Hole mobility in silicon calculated with the spherical-harmonics expansion method (HARM) as a function of the total ionized-dopant concentration N_I, using the lattice temperature T as parameter. The calculations are compared with measurements by Lombardi [73], Klaassen [64], and Arora [1] (courtesy of S. Reggiani)

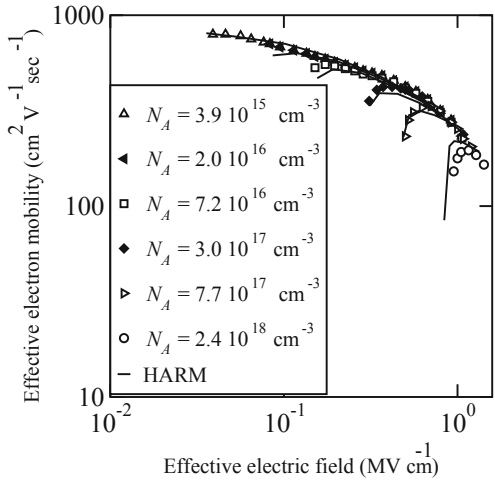

Fig. 20.9 Electron surface mobility in silicon calculated with the spherical-harmonics expansion method (HARM) method at room temperature, using the acceptor concentration N_A as parameter. The calculations are compared with measurements by Takagi [106] (courtesy of S. Reggiani)

have been calculated as functions of the dopant concentration (N_A and N_D, respectively), at room temperature; in the figures, they are compared with measurements taken from the literature.

20.6 Complements

20.6.1 Transition Rates in the SRH Recombination Function

The expressions of the transition rates r_a, r_b, r_c, r_d to be used in the calculation of the Shockley-Read-Hall recombination function (20.32) are determined by the same reasoning as that used in Sect. 20.2.1 for the direct thermal transitions. Let $P(\mathbf{r}, E, t)$

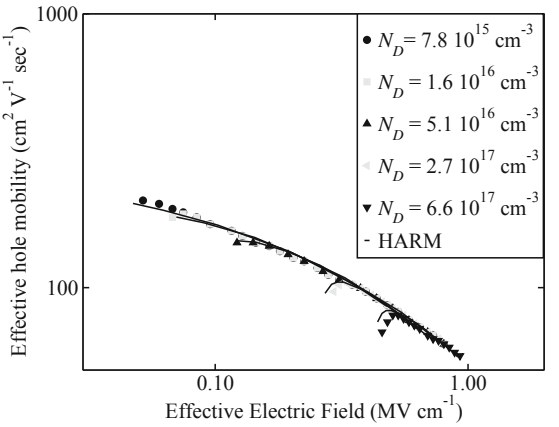

Fig. 20.10 Hole surface mobility in silicon calculated with the spherical-harmonics expansion method (HARM) at room temperature, using the donor concentration N_D as parameter. The calculations are compared with measurements by Takagi [106] (courtesy of S. Reggiani)

be the occupation probability of a state at energy E, and $C(E \to E')$ the probability per unit time and volume (in $\mathbf{r}$) of a transition from a filled state of energy E to an empty state of energy E'. Such a probability is independent of time; it depends on the energy of the phonon involved in the transition, and possibly on position. Then, define $P' = P(\mathbf{r}, E = E', t)$, $P_t = P(\mathbf{r}, E = E_t, t)$, where E_t is the energy of the trap. Finally, let $\gamma(E)$ be the combined density of states in energy and volume of the bands, and $\gamma_t(\mathbf{r}, E)$ the same quantity for the traps (the latter depends on position if the traps' distribution is non uniform). The number of transitions per unit volume and time, from states in the interval $\mathrm{d}E$ belonging to a band, to states in the interval $\mathrm{d}E'$ belonging to the trap distribution, is obtained as the product of the number $\Omega\,\gamma(E)\,\mathrm{d}E\,P$ of filled states in the interval $\mathrm{d}E$, times the transition probability per unit volume and time C, times the number $\Omega\,\gamma_t(\mathbf{r}, E')\,\mathrm{d}E'\,(1 - P')$ of empty states in the interval $\mathrm{d}E'$. Thus, letting ΔE_t be an energy interval belonging to the gap and containing the traps, the transition rate from the conduction band to the traps is given by

$$r_a = \int_{E_C}^{E_{CU}} \int_{\Delta E_t} \Omega\,\gamma(E)\,\mathrm{d}E\,P\,C(E \to E')\,\Omega\gamma_t(\mathbf{r}, E')\,\mathrm{d}E'\,(1 - P'). \qquad (20.69)$$

By the same token, the transition rate from the valence band to the traps is

$$r_d = \int_{E_{VL}}^{E_V} \int_{\Delta E_t} \Omega\,\gamma(E)\,\mathrm{d}E\,P\,C(E \to E')\,\Omega\gamma_t(\mathbf{r}, E')\,\mathrm{d}E'\,(1 - P'). \qquad (20.70)$$

In turn, the number of transitions per unit volume and time, from states in the interval $\mathrm{d}E'$ belonging the trap distribution, to states in the interval $\mathrm{d}E$ belonging to a band, is obtained as the product of the number $\Omega\,\gamma_t(\mathbf{r}, E')\,\mathrm{d}E'\,P'$ of filled states in the interval $\mathrm{d}E'$, times $C(\mathbf{r}, E' \to E)$, times the number $\Omega\,\gamma(E)\,\mathrm{d}E\,(1 - P)$ of empty states in the interval $\mathrm{d}E$. Thus, the transition rates from the traps to conduction or valence band are respectively given by

$$r_b = \int_{E_C}^{E_{CU}} \int_{\Delta E_t} \Omega\,\gamma_t(\mathbf{r}, E')\,\mathrm{d}E'\,P'\,C(\mathbf{r}, E' \to E)\,\Omega\,\gamma(E)\,\mathrm{d}E\,(1 - P), \qquad (20.71)$$

$$r_c = \int_{E_{VL}}^{E_V} \int_{\Delta E_t} \Omega\, \gamma_t(\mathbf{r}, E')\, \mathrm{d}E'\, P'\, C(\mathbf{r}, E' \to E)\, \Omega\, \gamma(E)\, \mathrm{d}E\, (1 - P). \quad (20.72)$$

The combined density of states of the traps is treated in the same manner as that of the dopant atoms (compare with (18.20) and (18.35)) by letting

$$\gamma_t(\mathbf{r}, E') = N_t(\mathbf{r})\, \delta(E' - E_t), \quad (20.73)$$

where $N_t(\mathbf{r})$ is the trap concentration. Thanks to this, the integrals over ΔE_t are easily evaluated, to yield

$$r_a = N_t\, (1 - P_t)\, \Omega^2 \int_{E_C}^{E_{CU}} \gamma\, P\, C(\mathbf{r}, E \to E_t)\, \mathrm{d}E = N_t\, (1 - P_t)\, \alpha_n\, n, \quad (20.74)$$

$$r_c = N_t\, P_t\, \Omega^2 \int_{E_{VL}}^{E_V} \gamma\, (1 - P)\, C(\mathbf{r}, E_t \to E)\, \mathrm{d}E = N_t\, P_t\, \alpha_p\, p, \quad (20.75)$$

where the definitions (20.3), (20.4) of the electron and hole concentrations are used, and the transition coefficients for electrons and holes are defined as the weighed averages

$$\alpha_n = \Omega^2\, \frac{\int_{E_C}^{E_{CU}} \gamma\, P\, C\, \mathrm{d}E}{\int_{E_C}^{E_{CU}} \gamma\, P\, \mathrm{d}E}, \qquad \alpha_p = \Omega^2\, \frac{\int_{E_{VL}}^{E_V} \gamma\, (1 - P)\, C\, \mathrm{d}E}{\int_{E_{VL}}^{E_V} \gamma\, (1 - P)\, \mathrm{d}E}. \quad (20.76)$$

Like in the case of (20.10), the integrals in (20.76) are approximated using the equilibrium probability. The remaining transition rates r_b, r_d are determined in a similar manner, using also the approximation $1 - P \simeq 1$ in (20.71) and $P \simeq 1$ in (20.70). Like in Sect. 20.2.1, the approximation is justified by the fact that in normal operating conditions the majority of the valence-band states are filled, while the majority of the conduction-band states are empty. In conclusion,

$$r_b = N_t\, P_t\, \Omega^2 \int_{E_C}^{E_{CU}} \gamma\, (1 - P)\, C(\mathbf{r}, E_t \to E)\, \mathrm{d}E \simeq N_t\, P_t\, e_n, \quad (20.77)$$

$$r_d = N_t\, (1 - P_t)\, \Omega^2 \int_{E_{VL}}^{E_V} \gamma\, P\, C(\mathbf{r}, E \to E_t)\, \mathrm{d}E \simeq N_t\, (1 - P_t)\, e_p, \quad (20.78)$$

with the emission coefficients defined by

$$e_n = \Omega^2 \int_{E_C}^{E_{CU}} \gamma\, C\, \mathrm{d}E, \qquad e_p = \Omega^2 \int_{E_{VL}}^{E_V} \gamma\, C\, \mathrm{d}E. \quad (20.79)$$

20.6.2 Coefficients of the Auger and Impact-Ionization Events

The expression of the coefficients c_n, c_p and I_n, I_p, to be used in the calculation of the net recombination rates (20.41) due to the Auger and impact-ionization phenomena, are found in the same way as the transition rates of the SRH recombination function (Sect. 20.6.1) or the direct thermal recombinations (Sect. 20.2.1). Let $P(\mathbf{r}, E, t)$ be the occupation probability of a state of energy E, and $C_n(E_1, E_2 \rightarrow E_1', E_2')$ the combined probability per unit time and volume (in $\mathbf{r}$) of an electron transition from a filled state of energy E_1 in the conduction band to an empty state of energy E_1' in the conduction band, and of another electron from a filled state of energy E_2 to an empty state of energy E_2', where E_2 and E_2' belong to different bands.

Auger Coefficients

In an Auger recombination it is $E_1' > E_1$; also, E_2 belongs to the conduction band while E_2' belongs to the valence band. Due to energy conservation it is[21]

$$C_n = C_{n0}\,\delta\left[(E_1 - E_1') + (E_2 - E_2')\right], \tag{20.80}$$

where $E_2 - E_2' \simeq E_G$; it follows $E_1' \simeq E_1 + E_G$. Then, define $P_i = P(\mathbf{r}, E = E_i, t)$, $P_i' = P(\mathbf{r}, E = E_i', t)$, with $i = 1, 2$, and let $\gamma(E)$ be the combined density of states in energy and volume for the bands; in particular, let $g_i = \Omega\,\gamma(E_i)$ and $g_i' = \Omega\,\gamma(E_i')$. From the above definitions one finds, for the rate r_a of the Auger recombinations initiated by electrons,

$$r_a = \int g_1\ \mathrm{d}E_1\ P_1\ g_2\ \mathrm{d}E_2\ P_2\ C_n\ g_1'\ \mathrm{d}E_1'\,(1 - P_1')\ g_2'\ \mathrm{d}E_2'\,(1 - P_2'), \tag{20.81}$$

where $\int$ indicates a fourfold integral that extends thrice over the conduction band and once over the valence band. Observing that $P_1' \ll 1$ and integrating over E_1' with $C_n = C_{n0}\,\delta(E_1 + E_G - E_1')$ yields

$$r_a = \int_{E_C}^{E_{CU}} g_1\ \mathrm{d}E_1\ P_1\ C_{n0}\,g_G \int_{E_C}^{E_{CU}} g_2\ \mathrm{d}E_2\ P_2 \int_{E_{VL}}^{E_V} g_2'\ \mathrm{d}E_2'\,(1 - P_2'), \tag{20.82}$$

where $g_G = g(E_1 + E_G)$ and $[C_{n0}\,g_G] = \mathrm{s}^{-1}\,\mathrm{m}^{-3}$. Thanks to (20.3) and (20.4), the second integral in (20.82) equals $\Omega\,n$ and the third one equals $\Omega\,p$. Letting

$$c_n = \Omega^3\,\frac{\int_{E_C}^{E_{CU}} C_{n0}\,g_G\,g_1\,P_1\,\mathrm{d}E_1}{\int_{E_C}^{E_{CU}} g_1\,P_1\,\mathrm{d}E_1}, \tag{20.83}$$

finally yields $r_a = c_n\,n^2\,p$. The derivation of $r_c = c_p\,p^2\,n$ is similar.

[21] The units of C_{n0} are $[C_{n0} = \mathrm{J\ s^{-1}\ m^{-3}}]$.

Impact Ionization's Transition Coefficients
Using the same symbols introduced at the beginning of Sect. 20.6.2, for an impact-ionization event initiated by an electron it is $E_1 > E_1'$; in turn, E_2 belongs to the valence band and E_2' belongs to the conduction band. It follows

$$r_b = \int g_1 \, \mathrm{d}E_1 \, P_1 \, g_2 \, \mathrm{d}E_2 \, P_2 \, C_n \, g_1' \, \mathrm{d}E_1' \, (1 - P_1') \, g_2' \, \mathrm{d}E_2' \, (1 - P_2'), \tag{20.84}$$

where the fourfold integral extends thrice over the conduction band and once over the valence band. From the energy-conservation relation $E_1 + E_2 = E_1' + E_2'$ and from $E_2' - E_2 \simeq E_G$ it follows $E_1' \simeq E_1 - E_G$. Observing that $P_2 \simeq 1$, $P_1' \ll 1$, $P_2' \ll 1$, and integrating over E_1' with $C_n = C_{n0} \, \delta(E_1 - E_G - E_1')$ yields

$$r_b = \int_{E_C}^{E_{CU}} C_{n0} \, g_G \, g_1 \, P_1 \, \mathrm{d}E_1 \int_{E_{VL}}^{E_V} g_2 \, \mathrm{d}E_2 \int_{E_C}^{E_{CU}} g_2' \, \mathrm{d}E_2', \tag{20.85}$$

where $g_G = g(E_1 - E_G)$, and the product of the second and third integral is a dimensionless quantity that depends only on the semiconductor's structure. Indicating such a quantity with ν_n, and letting

$$I_n = \nu_n \frac{\int_{E_C}^{E_{CU}} C_{n0} \, g_G \, g_1 \, P_1 \, \mathrm{d}E_1}{\int_{E_C}^{E_{CU}} g_1 \, P_1 \, \mathrm{d}E_1}, \tag{20.86}$$

finally yields $r_b = I_n \, n$. The derivation of $r_d = I_p \, p$ is similar.

20.6.3 Total Recombination-Generation Rate

The expressions for the most important generation-recombination terms have been worked out in this chapter. Only one of them, the SRH recombination function U_{SRH}, involves energy states different from those of the conduction and valence bands; in principle, such states would require additional continuity equations to be added to the semiconductor-device model. However, as discussed in Sect. 20.2.3, this is not necessary in crystalline semiconductors. The other mechanisms (direct thermal recombination-generation U_{DT}, Auger recombination and impact ionization U_{AI}, and optical recombination-generation U_O) do not involve intermediate states. As a consequence, with reference to (20.13) the generation-recombination terms of the electron-continuity equation are equal to those of the hole continuity equation. Finally, assuming that the different generation-recombination phenomena are uncorrelated, and neglecting U_{DT} with respect to U_{SRH} (Sect. 20.2.2), yields

$$U_n - G_n = U_p - G_p \simeq U_{SRH} + U_{AI} + U_{DO}. \tag{20.87}$$

20.6.4 Screened Coulomb Potential

In the context of physics, the general meaning of *screening* is the attenuation in the electric field intensity due to the presence of mobile charges; the effect is treated here using the Debye-Hückel theory [29], which is applicable to a non-degenerate semiconductor where the dopants are completely ionized. For a medium of permittivity ε, with charge density ρ, the electric potential in the equilibrium condition is found by solving Poisson's equation

$$-\varepsilon \, \nabla^2 \varphi = \rho. \tag{20.88}$$

One starts by considering a locally-neutral material, to which a perturbation is added due, for instance, to the introduction of a fixed charge $Z_c \, e_c$ placed in the origin; this, in turn, induces a variation in ρ. The corresponding perturbation of φ is calculated to first order by replacing φ with $\varphi + \delta\varphi$ and ρ with $\rho + (\partial\rho/\partial\varphi)\,\delta\varphi$, where the derivative is calculated at $\delta\varphi = 0$; the perturbed form of Poisson's equation reads:

$$-\varepsilon \, \nabla^2 \varphi - \varepsilon \, \nabla^2 \delta\varphi = \rho + \frac{\partial \rho}{\partial \varphi} \, \delta\varphi. \tag{20.89}$$

As the unperturbed terms cancel out due to (20.88), a Poisson equation in the perturbation is obtained,

$$\nabla^2 \delta\varphi = q_c^2 \, \delta\varphi, \qquad q_c^2 = -\frac{\partial\rho/\partial\varphi}{\varepsilon}, \tag{20.90}$$

where $1/q_c$ is the *screening length* or *Debye length*. The definition implies that $\partial\rho/\partial\varphi < 0$; this is in fact true, as shown below with reference to a non-degenerate semiconductor with completely-ionized dopants. Letting $N_D^+ = N_D$, $N_A^- = N_A$ in (19.125), and using the non-degenerate expressions (18.60), (18.61) of the equilibrium concentrations, one finds that $N = N_D - N_A$ is left unaffected by the perturbation, while the electron concentration[22] n transforms into $n \, \exp[e \, \delta\varphi/(k_B \, T)]$ and the hole concentration p transforms into $p \, \exp[-e \, \delta\varphi/(k_B \, T)]$. From $\rho = e\,(p - n + N)$ one obtains, to first order,

$$\frac{\partial \rho}{\partial \varphi} = -\frac{e^2}{k_B \, T} \, (n + p), \qquad q_c^2 = \frac{e^2 \, (n + p)}{\varepsilon \, k_B \, T} > 0. \tag{20.91}$$

The left hand side of the Poisson equation in (20.90) is conveniently recast using a set of spherical coordinates r, θ, ϕ whose origin coincides with the center of symmetry of the perturbation; using (B.25) one finds

$$\nabla^2 \delta\varphi = \frac{1}{r} \frac{\partial^2}{\partial r^2} (r \, \delta\varphi) + \frac{r^{-2}}{\sin\theta} \frac{\partial}{\partial\theta} \left(\sin\theta \, \frac{\partial \delta\varphi}{\partial\theta} \right) + \frac{r^{-2}}{\sin^2\theta} \frac{\partial^2 \delta\varphi}{\partial \phi^2}. \tag{20.92}$$

[22] The electron charge is indicated here with e to avoid confusion with q_c.

Considering a perturbation with a spherical symmetry, only the first term at the right hand side of (20.92) is left, whence (20.90) becomes an equation in the unknown $r\,\delta\varphi$:

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}(r\,\delta\varphi) = q_c^2\,(r\,\delta\varphi). \tag{20.93}$$

The general solution of (20.93) is $r\,\delta\varphi = A_1\,\exp(-q_c\,r) + A_2\,\exp(q_c\,r)$, where it must be set $A_2 = 0$ to prevent the solution from diverging as r becomes large. In conclusion,

$$\delta\varphi = \frac{A_1}{r}\,\exp(-q_c\,r). \tag{20.94}$$

The remaining constant is found by observing that for very small r the pure Coulomb case $\delta\varphi \simeq A_1/r$ is recovered, whence $A_1 = Z_c e_c e/(4\,\pi\,\varepsilon)$. This makes (20.94) to coincide with (14.33).