Competition between isotropic and strongly anisotropic terms in the impact ionization rate of narrow- and middle-gap cubic semiconductors

© A.N. Afanasiev, A.A. Greshnov, G.G. Zegrya

Ioffe Institute,
194021 St. Petersburg, Russia

E-mail: afanasiev.an@mail.ru

Received August 27, 2024
Revised September 7, 2024
Accepted November 26, 2024

We report on the strong anisotropy of the inter-band process of impact ionization in direct-gap cubic semiconductors with either weak or strong spin-orbit coupling at low effective temperatures of electron distribution T, and the crossover to isotropic behavior with increasing T. Such anisotropy is related to specific mechanism of the impact ionization involving coupling of the electron and heavy hole states via remote bands, which is vanishing for some high-symmetry propagation directions of an initial electron, namely [100] and [111]. At room temperature impact ionization rate in narrow-gap semiconductors InSb, InAs, GaSb and $In_{0.53}Ga_{0.47}As$ is isotropic while in middle-gap InP, GaAs and CdTe both terms are comparable. We propose simple and justified analytic generalization of Keldysh formula for the impact ionization rate, which is suitable for incorporation into modelling software.

Keywords: Impact ionization, direct gap semiconductor, **kp** model, hot carriers, device modelling.

DOI: 10.61011/SC.2024.11.59957.5877

1. Introduction

Interband impact ionization involving the creation of an electron-hole pair as a result of Coulomb interaction between a hot conduction electron and valence band electrons (Figure 1, *a*), plays an important role in the operation of many advanced electronics devices. In some of them, such as semiconductor diodes and field-effect transistors, avalanche breakdown caused by impact ionization limits the operating voltage range. Therefore, impact ionization has traditionally been perceived as a negative effect. At the same time, carrier multiplication due to impact ionization is the basis for the operation of impact avalanche transit time diodes (IMPATT), avalanche photodiodes (APD) [1] and the transistor with field-effect controlled impact ionization (I-MOS) [2], which has an extremely steep drain-gate characteristic. In particular, the operation of I-MOS has been experimentally demonstrated with the slope of the subthreshold part of the current-voltage characteristic $\sim 5\,\mathrm{mV/dec}$. at $T = 400\,\mathrm{K}$, which allows a significant reduction in the switching speed of the device compared to conventional MOSFET devices.

Numerical modelling of the physical processes occurring within semiconductor devices has become an inherent part of device design. However, often the physical models that are used in simulation programs are phenomenological and contain a large number of fitting parameters. Since in practice the output characteristics of devices using impact ionization depend on many details, such as the specific type of band structure of the material and the features of the scattering processes that determine the type of non-equilibrium distribution function, the calculation of the I-V characteristic based on a realistic model of the band structure is conceptually and technically difficult. Therefore, the most popular way of modelling is the Monte Carlo [3–13] method, but its results depend on a particular kind of dependence of the microscopic impact ionization rate $W(E)$ on the energy of the hot electron that initiates it. Phenomenologically, the ionization rate grows like a power of the excess energy above a threshold:

$$W(E) = C(E - E_{\mathrm{th}})^n. \qquad (1)$$

The specific form of the parameters n and C, as well as the limits of applicability (1), are established on the basis of quantum mechanical calculations. The most popular [14] is the quadratic dependence ($n = 2$) first obtained by Keldysh [15] more than half a century ago on the basis of considerations of the phase space volume corresponding to the final states in the elementary act of impact ionization (Figure 1, *a*). Existing estimates of C based on the sum rule [16] give values significantly higher than the results of numerical calculations based on the 30-band **kp**-model [17]. In some references (see [18], p. 511), it is recommended to adjust the prefactor in (1) to agree with experiment at fixed $n = 2$. Analytically, the coefficient C for the quadratic dependence of the form (1) has been calculated only for the case of narrow-bandgap cubic semiconductors [19].

The difficulty in describing the quadratic contribution to the impact ionization rate is due to the fact that in the isotropic band model, the simplest version of which is the 8-band **kp**-model, taking into account the interaction between the *s*-states of the conduction band and the *p*-states

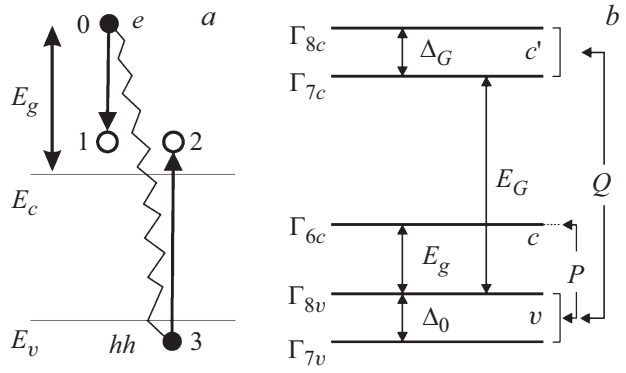

Figure 1. *a* — a diagram of the elementary process of interband impact ionization; *b* — a diagram of the 14-band **kp**-model.

of the valence band, the prefactor C is zero. Therefore, to calculate the Coulomb matrix element that determines prefactor in (1), the interaction with distant bands [19–21] must be taken into account, and the magnitude of the quadratic contribution turns out to be small for the case of narrow-gap semiconductors. This is confirmed by the results of numerical calculations [22], which indicate that the main contribution to the impact ionization rate of narrow-bandgap semiconductors is cubic ($n = 3$) rather than quadratic. The analytical expression for the cubic contribution was first obtained by Gelmont et al. [23]. For direct- and middle-gap semiconductors like GaSb, $In_{0.53}Ga_{0.47}As$, InP, $In_{0.52}Al_{0.48}As$, GaAs and CdTe there is no reasonable analytical form of $W(E)$ and in Monte Carlo simulations of impact ionization processes the expression (1) is used with arbitrary powers n, for example $n = 2.5$ and $n = 4.3$ in [24], $n = 5.2$ [3], $n = 3$ [4,6,25], $n = 3.9$ [26], $n = 1.85$ [27]. Some theoretical papers [28,29] have considered optimising the numerical calculation of the semiconductor band structure for realistic simulation of impact ionization processes in devices, but the integration of band structure calculations with Monte Carlo simulations is too complex for practical use. To date, most device simulation software at best uses (1) with loosely fitted parameters n and C leading to uncontrolled results.

This paper focuses on the study of a reasonable form of $W(E)$ in direct-gap cubic semiconductors with small to medium bandgap under practically relevant conditions when the effective temperature of the nonequilibrium electron distribution is of the order of a few tens of meV. Explicit analytical expressions for the coefficients in quadratic and cubic terms are obtained. An estimate of the T^* crossover temperature is given, at which the carrier generation rates associated with both contributions become equal under the conditions of the model isotropic classical distribution of nonequilibrium electrons. The obtained results provide a qualitative explanation and a quantitative criterion for the dominance of the cubic contribution at room temperature in narrow-gap semiconductors, whereas for middle-gap semiconductors both terms are comparable.

2. Quantum mechanical theory of the rate of interband impact ionization

2.1. General expression for the impact ionization rate

Usually conduction electrons are treated as quasiparticles that do not interact with the valence band electrons. Within the framework of this approach, the Auger-recombination (chcc-type) process inverse to impact ionization can be represented as the result of an interaction between two conduction electrons in which one of them moves to a higher state in the conduction band and the other — to a free state in the valence band. However, as can be shown in the Hartree-Fock approximation, the process of interband impact ionization is determined by the Coulomb interaction between the hot conduction electron and all the valence band electrons. Therefore, in order to obtain the total impact ionization rate for a given hot electron state „0“ (see the scheme in Figure 1, *a*), the partial rates of elementary processes described by the

$$W = \frac{2\pi}{\hbar}\left|\left\langle \alpha_1\alpha_2\left|\frac{e^2}{\varkappa|\mathbf{r}_1-\mathbf{r}_2|}\right|\alpha_0\alpha_3\right\rangle\right|^2\delta(\Delta E), \qquad (2)$$

must be summed over the possible initial states of the „3“ electron in the valence band and the final states of the „1“ and „2“ electrons in the conduction band. Here $\Delta E = E_1 + E_2 - E_0 - E_3$ represents the energy balance, and $\alpha_i = \{\mathbf{k}_i, \xi_i\}$ denotes the set of quantum numbers — the wave vector $\mathbf{k}_i$ and the total angular momentum projection ξ_i for the states in the conduction band ($i = 0$, 1 or 2) and in the valence band ($i = 3$). We consider only processes involving heavy hole states, since light and spin-orbit split hole states lie well below in energy for wave vectors larger than the threshold (for narrow- to middle-gap semiconductors), so that the corresponding impact ionization processes involve very hot and few electrons. Using the Fourier representation of the Coulomb potential and performing integration of the Bloch amplitudes at zero $\mathbf{k}$, expression (2) can be rewritten [16,18] as

$$W = \frac{2\pi}{\hbar}\left(\frac{4\pi e^2}{\varkappa}\right)^2\frac{I_{cc}(\alpha_0\alpha_1)I_{cv}(\alpha_2,\alpha_3)}{|\mathbf{k}_0-\mathbf{k}_1|^4}\delta_{\Delta\mathbf{k},0}\delta(\Delta E), \quad (3)$$

where $\Delta\mathbf{k} = \mathbf{k}_0 + \mathbf{k}_3 - \mathbf{k}_1 - \mathbf{k}_2$ represents the momentum balance, and the squares of the Bloch function overlap integrals I_{cc} and I_{cv} can be written using the state vectors $|F\rangle$ in the Bloch function basis of the point Γ of the first Brillouin zone $u_n^{(0)}(\mathbf{r})$ as $I_{c,c/v}(\alpha_i,\alpha_j) = |\langle F_{\alpha_i}|F_{\alpha_j}\rangle|^2$.

The energy and momentum conservation laws appearing as delta functions in (3) impose constraints on the possible values of the $\mathbf{k}_i$ wave vectors. This leads to the threshold conditions for impact ionization, which, taking into account the nonparabolicity of the initial electron dispersion at $\mu = m_e/m_{hh} \ll 1$ are as follows

$$E_{\text{th}} = E_e(\mathbf{k}_0^{\text{th}}) = E_g(1+2\mu), \qquad (4)$$

The parameters of the band structure (taken from [31–35]) of narrow- to middle-gap semiconductors in the framework of the used 14-band **kp**-model and the corresponding values of dimensionless parameters β (17) and $x = \Delta_0/E_g$ as well as the effective crossover temperature (28)

	E_g, eV**	Δ_0, eV*	P, eV · Å	E_G, eV	Q, eV · Å	β**	x**	T^*, K***
InSb	0.235	0.81	9.63	3.18	7.83	7.53	4.31	2.8
InAs	0.418	0.38	9.01	4.48	7.72	1.13	0.89	10.2
GaSb	0.81	0.76	9.69	3.11	8.25	0.41	0.93	140
$In_{0.53}Ga_{0.47}As$	0.817	0.324	9.81	4.51	8.25	0.22	0.33	75
InP	1.423	0.107	8.45	4.78	7.88	0.01	0.03	464
$In_{0.52}Al_{0.48}As$	1.545	0.295	9.09	4.51	8.25	0.04	0.15	516
GaAs	1.519	0.341	9.88	4.54	8.68	0.06	0.17	440
CdTe	1.61	0.95	9.5	5.4	7.87	0.22	0.56	313

Note. * — at $T = 0$ K; ** — at $T = 300$ K; *** — are calculated taking into account the temperature dependence of the bandgap in accordance with the data from [32,33,36].

$$\mathbf{k}_0^{\text{th}} = \mathbf{k}_g(1 + F_0(\Delta_0/E_g)\mu), \tag{5}$$

$$\mathbf{k}_3^{\text{th}} = -\mathbf{k}_0^{\text{th}}(1 - 2\mu), \tag{6}$$

$$\mathbf{k}_1^{\text{th}} = \mathbf{k}_2^{\text{th}} = \mu\mathbf{k}_g, \tag{7}$$

where

$$F_0(x) = \frac{3/2}{F_1(x)F_2(x)},$$

$$k_g = \frac{2}{\hbar}\sqrt{F_1(\Delta_0/E_g)m_eE_g}$$

— wave vector of electrons with energy $E_0(\mathbf{k}_g) = E_g$ and

$$F_1(x) = \frac{(1 + 2x/3)(1 + x/2)}{(1 + x)(1 + x/3)}, \tag{8}$$

$$F_2(x) = \frac{(1 + x)^2(1 + \frac{x}{3})^3}{(1 + \frac{7}{9}x + \frac{x^2}{6})(1 + \frac{2}{3}x)^2(1 + \frac{x}{2})}. \tag{9}$$

The values of the $F_1(\Delta_0/E_g)$ and $F_2(\Delta_0/E_g)$ functions are equal to unity for both $\Delta_0 \ll E_g$ and $\Delta_0 \gg E_g$ limits. Since **kp**-interaction between the conduction band Γ_{6c} (denoted by „c“ in Figure 1, b) and the valence bands Γ_{8v} and Γ_{7v} (or „v“ in Figure 1, b) does not contribute to the heavy hole dispersion, the smallness of μ is equivalent to $E_g/E_G \ll 1$, where E_G denotes the minimum distance between the v band and the bands contributing to the inverse mass of the heavy hole (the second conduction band c' within the 14-band **kp**-model used in this study (see Figure 1, b and the table). The magnitudes of the spin-orbit splittings of the c' and v bands are also small compared to the distance $c'-v$, $\Delta_{0,G}/E_G \ll 1$ [30] (see table).

In practice, the distribution function of hot electrons capable of initiating impact ionization extends on a much smaller scale (e.g. 25 meV) than the threshold energy E_{th}. Therefore, to describe impact ionization, it is convenient to introduce „above-threshold“ components of the $\mathbf{k}_i = \mathbf{k}_i - \mathbf{k}_i^{\text{th}}$ wave vectors and to consider only the $(E - E_{\text{th}})/E_{\text{th}} \ll 1$ domain. Under these assumptions, the impact ionization rate induced by an electron in the α_0 state reduces to

$$W = \frac{\pi\hbar F_2(\frac{\Delta_0}{E_g})}{12m_eE_g^2}\left(\frac{4\pi e^2}{\varkappa}\right)^2 \int \frac{d^3q_1 d^3q_2}{(2\pi)^6}\,[\tilde{I}_{cv}(\mathbf{q}_1, \mathbf{q}_3) + \tilde{I}_{cv}(\mathbf{q}_2, \mathbf{q}_3)]\delta\left(q_1^2 + q_2^2 - \frac{2m_e(E_0 - E_{\text{th}})}{\hbar^2}\right), \tag{10}$$

where

$$\tilde{I}_{cv}(\mathbf{q}_i, \mathbf{q}_3) = \sum_{\xi_i, \xi_3} I_{cv}(\mathbf{k}_i^{\text{th}} + \mathbf{q}_i, \xi_i; \mathbf{k}_3^{\text{th}} + \mathbf{q}_1 + \mathbf{q}_2 - \mathbf{q}_0, \xi_3) \tag{11}$$

denotes the interband overlap integral summed over the projections of the total angular momentum (on the direction $\mathbf{k}_3$) of the heavy hole states $\xi_3 = \pm\frac{3}{2}$ and the final electrons $\xi_i = \pm\frac{1}{2}$. The excess wave vector of the initial electron above the threshold $|\mathbf{q}_0| = \left(\frac{\partial E_0}{\partial K_0}\right)_{\text{th}}^{-1}(E_0 - E_{\text{th}})$ is assumed to be collinear to $\mathbf{k}_0^{\text{th}}$. Expression (10) shows that the energy (and angular) dependence of the impact ionization rate W is determined by the behaviour of the squared overlap integral $\tilde{I}_{cv}$ near threshold $\mathbf{q}_{1,2} = 0$. Since $\tilde{I}_{cv}$ expresses the degree at which the states of the conduction band and valence bands are overlapping in practice it is strongly dependent on the particular model of the band structure used.

2.2. Cubic contribution to the impact ionization rate

The minimal basis of such a model consists of eight $u_n^{(0)}(\mathbf{r})$ Bloch functions at the Γ-point of the Brillouin band: two of s-type and six of p-type. The interaction to be considered is the direct **kp**-interaction between s- and p-states described by a single Kane matrix element P [37]. This 8-band model describes well the dispersion of electrons and light holes in narrow-gap semiconductors, but the heavy holes remain dispersionless. Near the threshold $k_i \ll k_3$ for the final states ($i = 1, 2$) and the explicit expression for the

square of the matrix element within the 8-band model takes the form of

$$I_{cv}(\alpha_i, \alpha_3) = \frac{P^2|[\mathbf{k}_i \times \mathbf{k}_3]|^2}{2E_g^2 k_3^2}\,\delta_{|\xi_i - \xi_3|,1}. \qquad (12)$$

The matrix element (12) vanish for collinear wave vectors, so given the threshold conditions (6) and (7) $\tilde{I}_{cv}$ can be represented as

$$\tilde{I}_{cv}(\mathbf{q}_i, \mathbf{q}_3) = \frac{P^2 q_{i\perp}^2}{E_g^2} = \frac{\hbar^2 q_{i\perp}^2}{2m_e}\,\frac{1 + \frac{\Delta_0}{E_g}}{E_g + \frac{2}{3}\Delta_0}, \qquad (13)$$

where $\mathbf{q}_{i\perp}$ denotes the component $\mathbf{q}_{1,2}$ in the plane perpendicular to the wave vector of the initial electron. Taking into account the explicit form $\tilde{I}_{cv}$ (13) the impact ionization rate W (10) reduces to cubic in $E - E_{\text{th}}$ contribution [19]:

$$W_3(E) = B(E - E_{\text{th}})^3, \qquad (14)$$

$$B = \frac{\omega_B^*}{18E_g^3}\,\frac{E_g + \Delta_0}{E_g + \frac{2}{3}\Delta_0}\,F_2\left(\frac{\Delta_0}{E_g}\right), \qquad (15)$$

where $\omega_B^* = \frac{m_e e^4}{2\hbar^3 \varkappa^2}$ denotes the Bohr frequency of the conduction electrons. In the limiting case of infinite Δ_0 spin-orbit splitting (corresponding to the 6-band **kp**-model), this answer reduces to the result (2) of [23], whereas a finite value of Δ_0 leads to an additional multiplier equal to 2/3 at $\Delta_0 \to 0$.

2.3. Quadratic contribution to the impact ionization rate

Thus, the quadratic contribution to the impact ionization rate associated with the magnitude of the interband overlap integral at threshold remains outside the scope of the minimal band structure model. To describe the quadratic contribution, it is necessary to use more complex models that take into account the interaction with distant bands and the reduction of spherical symmetry to cubic symmetry (O_h/T_d groups), in particular, the 14-band **kp**-model (extended Kane model [30]). In this model, in addition to the **kp**-interaction of the valence band and conduction band states, the interaction of the valence band states with six additional Bloch states of symmetry Γ_{7c} and Γ_{8c}, which are a few eV above E_c (the second conduction band, c' in Figure 1, *b*) is explicitly taken into account (first order by k). The strength of the $c'-v$ interaction is described by a matrix element Q having a value of order P (see table). Inversion asymmetry allows the interaction between the c and c' bands described by the matrix elements P' and Δ', whose value is an order of magnitude smaller than P, Q and Δ_0 respectively [38].

To take into account the additional interaction between c- and v-bands by perturbation theory, it is convenient to split the full **kp**-hamiltonian into the $H_0(\mathbf{k})$ main part, which is the Hamiltonian of the minimal 8-band model and the energy of the c' states at $k = 0$, and the $V(\mathbf{k})$, perturbation, which describes the $c'-v$-interaction. The six eigenstates of H_0 at $\mathbf{k} = \mathbf{k}_3^{\text{th}} \simeq -\mathbf{k}_g$ corresponding to the c'-band, are far away in energy from the other eight: electron states with energy $E_e = E_v + 2E_g$ heavy holes at $E_{hh}^{(0)} = E_v$ and light and spin-orbit split hole states with energies

$$E_{lh/so} = E_v - \frac{E_g}{2}\left(1 + x \pm \sqrt{\frac{x^3 + x^2 - x + 3}{x + 3}}\right), \qquad (16)$$

where $x = \Delta_0/E_g$. Expression (16) indicates that the minimum energy distance between heavy holes and other branches of the valence band exceeds $\min(E_g, \Delta_0)/2$ for $x \geq 1$, which corresponds to the case of narrow-bandgap semiconductors (see the table). Hence, the unperturbed state of the heavy hole is nondegenerate in this case and the corresponding perturbation theory can be applied. The specific method for calculating the multiband Bloch functions that we follow is described in *Appendix* I. However, the true parameter, which must not be small for the perturbation theory method outlined in *Appendix* I to be valid, is

$$\beta = \frac{\Delta_0 E_G}{6Q^2 k_g^2} = \frac{\Delta_0 P^2 E_G}{12Q^2 E_g^2}\,\frac{E_g + \Delta_0/3}{E_g + \Delta_0/2}. \qquad (17)$$

The relations (A.I.1)–(A.I.3) show that the **kp**-perturbation $V(-\mathbf{k}_g)$ is applied to the unperturbed state of the $|F_{hh}^{(0)}\rangle$ heavy hole twice, giving rise to an $\propto (Qk_g)^2 \propto E_g^2$ multiplier and two energy denominators associated with the Green's functions. The first denominator is determined by the distance between the c' and v bands equal to E_G, and the second is determined by the distance between the states of heavy holes and electrons or light holes or spin-orbit split holes. When Δ_0 becomes much smaller than E_g, spin-orbit split holes behave similarly to heavy holes and the energy gap between them at the finite wave vector $\mathbf{k} = \mathbf{k}_3^{\text{th}} \simeq -\mathbf{k}_g$ equal to $E_v - E_{so}$, according to (16), tends to the value at $k = 0$ equal to Δ_0. From this we can conclude that the described method of perturbation theory does not work at $\beta \ll 1$ and its application may lead to divergence at $\Delta_0 \to 0$. However, the latter does not occur due to the fact that in this limit the unperturbed state of spin-orbit split holes does not overlap with s-states, since they are transformed into the second branch of heavy holes. Therefore, for the case of middle-gap semiconductors, when $x = \Delta_0/E_g \ll 1$ (or $\beta \ll 1$) becomes a small parameter (see table) in addition to $m_e/m_{hh} \ll 1$, perturbation theory must account for the degeneracy of the heavy hole states. In this case, $|F_{hh}^{(0)}\rangle$ in *Appendix* I will have the meaning of the correct zero approximation wave function, which corresponds to the upper branch of heavy holes split by **kp**-interactions between c- and v-bands.

As a result of squaring the overlap integral and summing over ξ the explicit expression for the main contribution to $\tilde{I}_{cv}$, which determines the quadratic term in the impact

ionization rate, takes the form

$$\tilde{I}_{cv}(0,0) = \frac{8E_g^2}{E_G^2}\frac{Q^4}{P^4}K(\mathbf{u},\beta)\frac{1+x/2}{1+x/3}, \tag{18}$$

where $K(\mathbf{u},\beta)$ denotes the cubic invariant, which can be written in terms of the parameter β and invariant polynomials of order 4 and 6

$$I(\mathbf{u}) = u_x^2u_y^2 + u_x^2u_z^2 + u_y^2u_z^2, \tag{19}$$

$$J(\mathbf{u}) = u_x^2u_y^2u_z^2. \tag{20}$$

Here $\mathbf{u} = \mathbf{k}_0/k_0$ characterises the direction of propagation of the initial electron with respect to the crystallographic axes. The explicit form of the cubic invariant $K(\mathbf{u},\beta)$ is given in *Appendix* II. For large ($\beta \to \infty$) and small ($\beta \to 0$) values of β the anisotropy of the quadratic contribution is described by the

$$K_\infty(\mathbf{u}) = I(1-3I) \tag{21}$$

and

$$K_0(\mathbf{u}) = K_\infty(\mathbf{u}) - I^2 + 3J + \frac{I^2(1-4I) - J(2-9I)}{\sqrt{I^2-3J}}, \tag{22}$$

respectively. Previously, a similar (18) (for the $\beta \gg 1$ limit) result was obtained in [20] in terms of the Luttinger parameters γ_2 and γ_3 and in [19] in the framework of the 14-band **kp**-model at $\Delta_0 \to \infty$ and used for the analysis of Auger recombination and impact ionization, respectively. By substituting $\tilde{I}_{cv}(\mathbf{q}_i, \mathbf{q}_3)$ of the form (18) into (10), an analytical expression for the quadratic contribution to the impact ionization rate can be obtained:

$$W_2(E,\mathbf{u}) = A(E-E_{\text{th}})^2, \tag{23}$$

$$A = \frac{3}{4}\frac{\omega_B^*}{E_G^2}\frac{Q^4}{P^4}K(\mathbf{u},\beta)\frac{E_g+\frac{1}{2}\Delta_0}{E_g+\frac{1}{3}\Delta_0}F_2\left(\frac{\Delta_0}{E_g}\right). \tag{24}$$

Taking into account the presence of E_G^2 in the denominator, the quadratic contribution (24) has the 2nd order of smallness in the parameter $\mu = m_e/m_{hh}$, which leads, as shown below, to its competition with the cubic contribution (15) in semiconductors with small and medium bandgap for the characteristic excess of the hot electron energy over the $E - E_{\text{th}}$ threshold of the order of several tens of meV. The ionization rate described by W_2 strongly depends on the orientation of the motion direction of the hot electron relative to the crystallographic axes. In both cases of strong $\beta \gg 1$ and weak $\beta \ll 1$ spin-orbit splitting of the valence band, the quadratic contribution vanishes in highly symmetric directions [100] and [111]. However, the anisotropy W_2, described by $K_\infty(\mathbf{u})$ and $K_0(\mathbf{u})$ is different: in the latter case, the quadratic contribution additionally vanishes in the [110] direction. In planes (111), the quadratic contribution becomes isotropic at strong spin-orbit coupling since $I(\mathbf{u}_{(111)}) = 1/4$, whereas $K_0(\mathbf{u}_{(111)})$ reproduces the non-trivial angular dependence of $J(\mathbf{u})$. In semiconductors of the T_d group, the absence of an inversion centre and the spin-orbit interaction lead to an additional contribution to W_2 that does not vanish in the main crystallographic directions. However, such a contribution is small in $\Delta_{c'v}/\Delta_0$, where $\Delta_{c/v}$ — the magnitude of the off-diagonal spin-orbit $c'-v$-interaction [30], so from a practical point of view this effect is not significant.

3. Results and discussion

In order to compare the significance of the two contributions to the total impact ionization rate

$$W_{\text{tot}}(E,\mathbf{u}) = A(\mathbf{u})(E-E_{\text{th}})^2 + B(E-E_{\text{th}})^3, \tag{25}$$

we consider an ensemble of non-degenerate electrons taken out of equilibrium by the application of an electric field and calculate the carrier generation rates R_2 and R_3 [corresponding to the impact ionization rates (24) and (15)] averaged over the electric field directions. Since this averaging is equivalent to averaging over the directions of the initial electron $\mathbf{u}$ under the assumed isotropy of the distribution, the carrier generation rate can be written as

$$R_i = \overline{W}_i N_0, \tag{26}$$

$$\overline{W}_i = \int_{E_{\text{th}}}^{+\infty}\frac{dE}{T}\frac{d\mathbf{u}}{4\pi}W(E,\mathbf{u})\exp\left(-\frac{E-E_{\text{th}}}{T}\right), \tag{27}$$

where $N_0 = D(E_{\text{th}})\overline{\delta f}(E_{\text{th}})T$ — nonequilibrium concentration of hot electrons above the impact ionization threshold, $\overline{W}_i$ — impact ionization rate averaged over the directions of initial electron propagation and distribution, $D(E)$ — density of states in the conduction band, $\delta f(E)$ — the nonequilibrium part of the distribution function, T — the effective temperature of the distribution, which is determined by the electron energy acquired at the mean free path eEl or its combination with the optical phonon energy $\hbar\omega_{\text{opt}}$ [16]. After integrating $(E-E_{\text{th}})^n$ with the Boltzmann distribution, we arrive at the following expression for the crossover temperature at which $\overline{W}_2 = \overline{W}_3$ (or $R_2 = R_3$):

$$T^* = \frac{\overline{A}}{3B} = 8\frac{Q^4}{P^4}\frac{E_g^3}{E_G^2}\overline{K}(\beta)F_1\left(\frac{\Delta_0}{E_g}\right). \tag{28}$$

The value of the averaged (over the $\mathbf{u}$ directions) cubic invariant at arbitrary β is between the two limits at infinite and zero spin-orbit coupling $\overline{K}_\infty < \overline{K}(\beta) < \overline{K}_0$, with $\overline{K}_\infty = 2/35$, and $\overline{K}_0 = 0.069$. The behaviour of $\overline{K}(\beta)$, calculated based on (A.II.1)–(A.II.6), is shown in Figure 2.

The temperature dependence of the bandgap $E_g(T)$ leads to a nonlinear dependence of the $\overline{W}_3/\overline{W}_2$ ratio on the effective temperature T, and (28) becomes a transcendental equation. Using the band structure parameters from the table and empirical temperature dependences of the bandgaps (in particular, for CdTe from [32] (Manoogian–Wooley approximation) and from [33,36] (Varshni approximation)

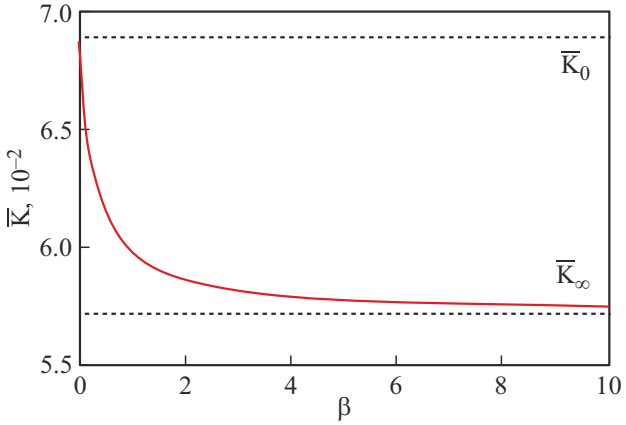

Figure 2. Average value of the cubic invariant (A.II.1) at arbitrary parameters β (17).

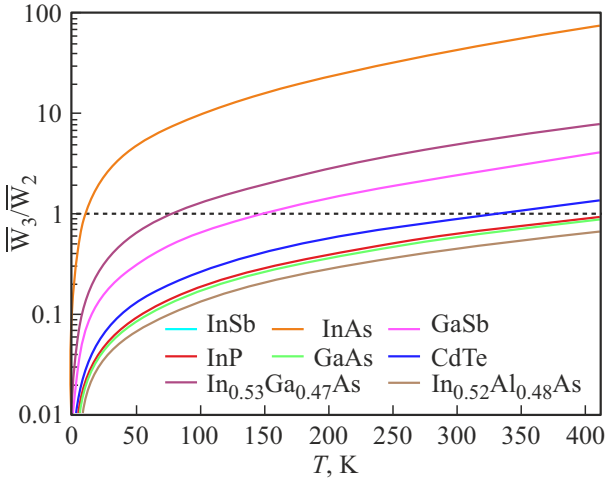

Figure 3. Competition between the averaged cubic and quadratic contributions to the impact ionization rate for the semiconductors presented in the table at different effective temperatures of the hot electron distribution above the threshold. The points where the solid curves intersect the dashed line correspond to the effective crossover temperatures. (A color version of the figure is provided in the online version of the paper).

for other compounds), we calculated the $\overline{W}_3(T)/\overline{W}_2(T)$ dependences for semiconductors with small and medium bandgaps (Figure 3), as well as crossover temperatures (see table). At low effective temperatures, the impact ionization rate of any semiconductor is described by a quadratic contribution and is strongly anisotropic. As T increases, the magnitude of the isotropic cubic contribution grows rapidly and at room temperature it dominates over $W_2(E, \mathbf{u})$ in narrow-bandgap semiconductors such as InSb, InAs, GaSb and $In_{0.53}Ga_{0.47}As$, while for middle-gap, InP, GaAs and CdTe, both contributions are comparable. In $In_{0.52}Al_{0.48}As$ the crossover occurs at a much higher temperature on the order of 500 K.

4. Conclusion

Finally, we will briefly discuss the accuracy of the analytical expressions obtained. Since in describing impact ionization in direct-gap semiconductors the main difficulty, and hence inaccuracy, is to obtain a correct expression for the overlap integrals of the initial and final states, which are determined by a particular type of multiband wave functions, we have carried out calculations of the averaged rate of the impact ionization based on the numerical diagonalisation of the 14-band **kp**-model. To reduce the calculation complexity in (10), we omitted the non-significant dependence of the calculated overlap integrals on $\mathbf{q}_3$, used the main-order approximation by $\mu \ll 1$ for the threshold wave vectors (5)–(7) and the $\tilde{I}_{cv}(\mathbf{q}_{1,2}, \mathbf{q}_3)$ expansion by $q_{1,2}/k_g \ll 1$, and performed analytical integration of the total rate by $\mathbf{q}_{1,2}$ and q_0. As a result, the 9-dimensional integration over $\mathbf{q}_{0,1,2}$ was reduced to averaging the total rate along the directions of the wave vector of the hot electrons. The evolution of the resulting values of the averaged impact ionization rate with E_g (the other parameters of the band structure correspond to InAs) at effective temperatures $T = 300,\ 500,\ 800$ K is shown in Figure 4. The simple analytical expression $\overline{W}_{\text{tot}} = 2\overline{A}T^2 + 6BT^3$ for the impact

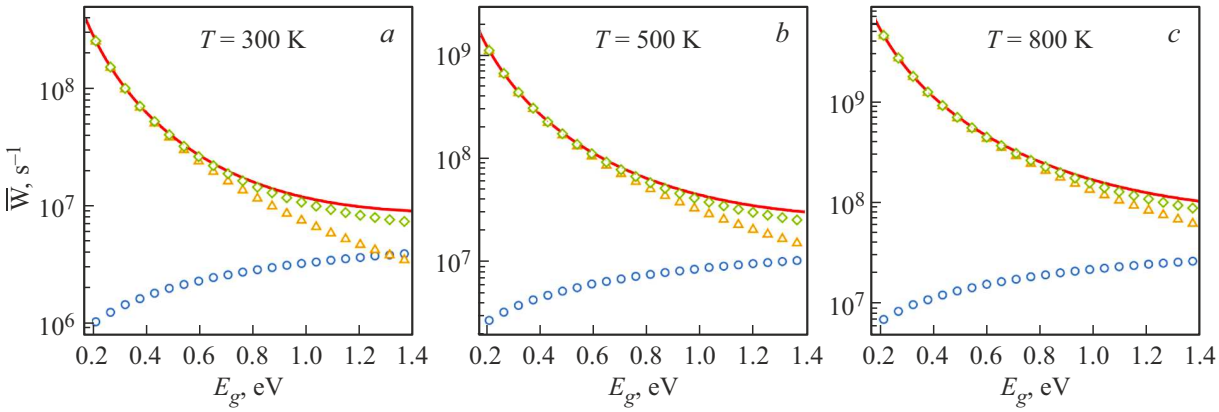

Figure 4. Dependence of the averaged impact ionization rate on the bandgap at different effective carrier temperatures: 300 (*a*), 500 (*b*), and 800 K (*c*). The solid red curves correspond to the full $\overline{W}_{\text{tot}} = \overline{W}_2 + \overline{W}_3$ rates calculated from the analytical expressions (15) and (24); the markers correspond to the numerically calculated total rate (green rhombuses), partial quadratic (blue circles), and cubic (orange triangles) contributions.

ionization rate derived from (25) and (27) agrees well with numerical results over a wide range E_g up to 1.4 eV. In particular, the corresponding value of the mean deviation is 7%, and the maximum discrepancy is limited to 18% for all temperatures. The magnitudes of the analytical and numerical „crossover bandgaps" [when $\overline{W}_2(E_g) = \overline{W}_3(E_g)$] are also close.

The discrepancy between the analytical and numerical results is due to the expression (24), which underestimates the quadratic contribution especially for wide-gap semiconductors when the primary small parameter in our $\mu = m_e/m_{hh}$ theory approaches unity. Nevertheless, a 5% agreement with numerical results for the whole range of the considered bandgaps can be achieved by including in the analytical expressions higher order corrections (up to the 2nd order) by E_g/E_G to (15) and (24). We also expect that in the case of strong anisotropy of the distribution of hot electrons in a strong electric field [39], the angular dependence of the carrier generation rate will follow the anisotropy of the total impact ionization rate (25). Thus, the obtained simple analytical generalization (25) of the conventional Keldysh formula for the impact ionization rate in direct-gap semiconductors is suitable for incorporation into device modelling software.

Conflict of interest

The authors declare that they have no conflict of interest.

Appendix I
The method of perturbation theory for the multiband heavy-hole function

In order to calculate the multiband wave function of the heavy hole state by perturbation theory, it is convenient to introduce „non-interacting" Green's function $G_0(E) = (E - H_0)^{-1}$. Then the 1st order correction to the unperturbed state of the heavy hole $|F_{hh}^{(0)}\rangle$ can be represented as

$$|F_{hh}^{(1)}\rangle = \lim_{E\to\tilde{E}_v} G_0(E, -\mathbf{k}_g)V(-\mathbf{k}_g)|F_{hh}^{(0)}\rangle. \quad \text{(A.I.1)}$$

According to the definition of V and G_0 in section 2, $|F_{hh}^{(1)}\rangle$ belongs to the c' subspace of basis functions and is orthogonal to the eight states belonging to the c and v bands. Therefore, the 1st order approximation does not give a contribution to the overlap integral $\tilde{I}_{cv}(0, 0)$, nor does the 1st order correction to the energy $E_{hh}^{(1)} = 0$. The 2nd order correction to the heavy hole energy E_{hh} is

$$E_{hh}^{(2)} = \lim_{E\to\tilde{E}_v} \langle F_{hh}^{(0)}|V(-\mathbf{k}_g)|F_{hh}^{(1)}(E)\rangle \quad \text{(A.I.2)}$$

and the corresponding correction to the multiband wave function $|F_{hh}\rangle$ can be rewritten as

$$|F_{hh}^{(2)}\rangle = \lim_{E\to\tilde{E}_v} G_0(E, -\mathbf{k}_g) \times \left[V(-\mathbf{k}_g)|F_{hh}^{(1)}(E)\rangle - E_2(E)|F_{hh}^{(0)}\rangle\right]. \quad \text{(A.I.3)}$$

Expression (A.I.2) defines the energy of the heavy hole and the relation between its mass and parameters of the 14-band model Q and E_G, and (A.I.3) leads to the main approximation for the interband $(c-v)$ overlap integral,

$$\langle F_e(\mathbf{k}_i^{\text{th}}, \xi_i)|F_{hh}(\mathbf{k}_3^{\text{th}}, \xi_3)\rangle \simeq \langle F_e^{(0)}(\mu\mathbf{k}_g, \xi_i)|F_{hh}^{(2)}(-\mathbf{k}_g, \xi_3)\rangle, \quad \text{(A.I.4)}$$

where $|F_e^{(0)}\rangle$ denotes the pure s function describing the low-energy states of the final electrons in the single-band approximation, whose wave vector is small compared to the wave vectors of the initial states (0 and 3). Therefore, corrections to $|F_e^{(0)}\rangle$ do not appear in (A.I.4) in the main order on $\mu = m_2/m_{hh}$.

Appendix II
Cubic invariant describing the anisotropy of the quadratic contribution

The expression for the cubic invariant describing the anisotropy of the quadratic contribution to the impact ionization rate at arbitrary β is

$$K(\mathbf{u}, \beta) = K_1(\mathbf{u}, \beta) + K_2(\mathbf{u}, \beta), \quad \text{(A.II.1)}$$

$$K_1(\mathbf{u}, \beta) = \frac{K_1(K_\infty - K_1) + \beta K_2 + \beta^2(K_\infty + K_1)}{\sqrt{K_1 + \beta^2}(\beta + \sqrt{K_1 + \beta^2})}, \quad \text{(A.II.2)}$$

$$K_2(\mathbf{u}, \beta) = \frac{K_2 + \beta(K_\infty - K_1)}{\beta + \sqrt{K_1 + \beta^2}}, \quad \text{(A.II.3)}$$

$$K_\infty(\mathbf{u}) = I - 3I^2, \quad \text{(A.II.4)}$$

$$K_1(\mathbf{u}) = I^2 - 3I, \quad \text{(A.II.5)}$$

$$K_2(\mathbf{u}) = -4I^3 + I^2 + 9I\,J - 2J, \quad \text{(A.II.6)}$$

where $I(\mathbf{u})$ and $J(\mathbf{u})$ denote the invariant polynomials of 4th and 6th orders respectively (Sect. 2.3).

References

[1] S. Zi. *Fizika poluprovodnikovyk priborov* (M., Mir, 1984). (in Russian).
[2] K. Gopalakrishnan, P.B. Griffin, J.D. Plummer. IEEE Trans. Electron Dev., **52**, 69 (2005).
[3] S. Trumm, M. Betz, F. Sotier, A. Leitenstorfer, A. Schwanhäußer, M. Eckardt, O. Schmidt, S. Malzer, G.H. Döhler, M. Hanson, D. Driscoll, A.C. Gossard. Appl. Phys. Lett., **88**, 132113 (2006).

[4] S. Chen, G. Wang. J. Appl. Phys., **103**, 023703 (2008).
[5] F. Bertazzi, M. Moresco, E. Bellotti. J. Appl. Phys., **106**, 063718 (2009).
[6] C.K. Chia. Appl. Phys. Lett., **97**, 073501 (2010).
[7] E. Bellotti, F. Bertazzi. J. Appl. Phys., **111**, 103711 (2012).
[8] S. Shishehchi, F. Bertazzi, E. Bellotti. J. Appl. Phys., **113**, 203709 (2013).
[9] S. Ašmontas, R. Raguotis, S. Bumelienė. Semicond. Sci. Technol., **28**, 025019 (2013).
[10] K. Kodama, H. Tokuda, M. Kuzuhara. J. Appl. Phys., **114**, 044509 (2013).
[11] K. Ghosh, U. Singisetti. J. Appl. Phys., **124**, 085707 (2018).
[12] S. Ašmontas, S. Bumelienė, J. Gradauskas, R. Raguotis, A. Sužiedėlis. Semicond. Sci. Technol., **34**, 075016 (2019).
[13] S. Ašmontas, S. Bumelienė, J. Gradauskas, R. Raguotis, A. Sužiedėlis. Sci. Rep., **10**, 10580 (2020).
[14] M.V. Fischetti, S.E. Laux. Phys. Rev. B, **38**, 9721 (1988).
[15] L.V. Keldysh. ZhETF, **37**, 713 (1959). (in Russian).
[16] B.K. Ridley. *Quantum Processes in Semiconductors* (Oxford University Press, N.Y., 2013).
[17] M.G. Burt, S. Brand, C. Smith, R.A. Abram. J. Phys. C: Solid State Phys., **17**, 6385 (1984).
[18] K.F. Brennan. *The Physics of Semiconductors: With Applications to Optoelectronic Devices* (Cambridge University Press, Cambridge, 1999).
[19] A.N. Afanasyev, A.A. Greshnov, G.G. Zegrya. Pisma ZhETF **105**, 586 (2017). (in Russian).
[20] B.L. Gelmont. ZhETF, **75**, 536 (1978). (in Russian).
[21] R. Redmer, J.R. Madureira, N. Fitzer, S.M. Goodnick, W. Schattke, E. Schöll. J. Appl. Phys., **87**, 781 (2000).
[22] A.R. Beattie, R.A. Abram, P. Scharoch. Semicond. Sci. Technol., **5**, 738 (1990).
[23] B. Gelmont, K.-S. Kim, M. Shur. Phys. Rev. Lett., **69**, 1280 (1992).
[24] K.Y. Choo, D.S. Ong. J. Appl. Phys., **96**, 5649 (2004).
[25] C.K. Chia, G.K. Dalapati. IEEE Trans. Electron Dev., **60**, 3435 (2013).
[26] D. Dolgos, A. Schenk, B. Witzigmann. J. Appl. Phys., **111**, 073714 (2012).
[27] I.C. Sandall, J.S. Ng, S. Xie, P.J. Ker, C.H. Tan. Opt. Express, **21**, 8630 (2013).
[28] P. Scharoch, R.A. Abram. Semicond. Sci. Technol., **3**, 973 (1988).
[29] S. Brand, R.A. Abram. J. Phys. C: Solid State Phys., **17**, L201 (1984).
[30] R. Winkler. *Spin-Orbit Coupling Effects in Two-Dimensional Electron and Hole Systems* (Springer Verlag, Berlin–Heidelberg, 2003).
[31] M. Cardona, N.E. Christensen, G. Fasol. Phys. Rev. B, **38**, 1806 (1988).
[32] G. Fonthal, L. Tirado-Mejía, J. Marín-Hurtado, H. Ariza-Calderón, J. Mendoza-Alvarez. J. Phys. Chem. Solids, **61**, 579 (2000).
[33] I. Vurgaftman, J.R. Meyer, L.R. Ram-Mohan. J. Appl. Phys., **89**, 5815 (2001).
[34] W.H. Lau, J.T. Olesberg, M.E. Flatté. *Electronic structures and electron spin decoherence in (001)-grown layered zincblende semiconductors* (2004). arXiv:condmat/ 0406201 [cond-mat.mes-hall]
[35] J.-M. Jancu, R. Scholz, E.A. de Andrada e Silva, G.C. La Rocca. Phys. Rev. B, **72**, 193201 (2005).
[36] *New Semiconductor Materials Database. Characteristics and Properties*. Ioffe Institute (http://www.matprop.ru/)
[37] E.O. Kane. J. Phys. Chem. Solids, **1**, 249 (1957).
[38] S. Richard, F. Aniel, G. Fishman. Phys. Rev. B, **70**, 235204 (2004).
[39] A.P. Dmitriev, M.P. Mikhailova, I.N. Yassievich. Phys. Status Solidi B, **113**, 125 (1982).

Translated by J.Savelyeva