JOURNAL OF APPLIED PHYSICS VOLUME 85, NUMBER 12 15 JUNE 1999

Impact ionization rate calculations in wide band gap semiconductors

D. Harrison, R. A. Abram,[a)] and S. Brand
Department of Physics, University of Durham, South Road, Durham DH1 3LE, United Kingdom

(Received 6 November 1998; accepted for publication 22 February 1999)

An algorithm for calculating impact ionization rates in the semiclassical Fermi's Golden Rule approximation which is efficient close to threshold is presented. Electron and hole initiated rates are calculated for three semiconductors with particular band structure characteristics, as are the distributions of the generated carriers. Simple analytic expressions of the form $R=A(E-E_0)^P$ are fitted to the calculated rates. The role of the matrix elements in influencing the distribution of final states is investigated. In the direct gap materials, they act to significantly enhance the low-**q** transitions, while in the indirect gap case they have a lesser effect on the distribution. Results for GaAs obtained here and by several other workers are compared and possible causes of the discrepancies examined, including differences in band structure and approximations made in evaluation of the matrix element. It is found that these differences do not influence the rate sufficiently to account for the wider variation between authors, and so it is concluded that differences in the implementation of the rate integration algorithm are the main cause. © *1999 American Institute of Physics.* [S0021-8979(99)06311-2]

I. INTRODUCTION

Impact ionization is a process occurring in semiconductors in which a high energy carrier excites a valence band electron across the band gap, thus creating an electron–hole pair. Since the initiating carrier must supply energy at least equal to the band gap, it must have a sufficiently high kinetic energy typically as a result of excitation by a large electric field. Such large fields exist in small high-speed devices such as field effect transistors, in which case impact ionization is usually detrimental to performance,[1–3] or in devices such as IMPATT diodes[4] and avalanche photodiodes,[5,6] whose operation relies on charge multiplication caused by the ionization. Theoretical investigations into the role of impact ionization in carrier transport have been carried out by Wolff,[7] Shockley,[8] Baraff[9] and Ridley,[10] among others. These analytical theories are based on simple carrier transport models, and are concerned mainly with the process by which carriers gain sufficient energy to initiate ionization, assuming that it occurs rapidly once they have done so. To perform more detailed transport modeling, Monte Carlo simulation is commonly used.[11–14] However, the high energy nature of the process requires the use of realistic band structure, and the resulting numerical complexity requires intensive computational effort.

A prerequisite to performing a Monte Carlo transport simulation is to obtain scattering rates due to impact ionization for carriers in the crystal. Keldysh[15] obtained an analytical expression for the rate at which a carrier of given energy initiates ionization, which has been applied in several Monte Carlo simulations.[11,16] However the derivation of his expression is based on the assumption that the valence and conduction bands are spherical and parabolic, the band gap is direct and the matrix elements can be taken to be constant. In semiconductors whose band gap is sufficiently wide that the impact ionization threshold (the minimum kinetic energy a carrier must have in order to initiate ionization) lies at energies well above the band edge, these conditions do not apply. Kane[17] performed a more thorough calculation for Si in which he numerically integrated the rate obtained through Fermi's Golden Rule over all energy and momentum conserving transitions, obtaining a significantly different rate to that of Keldysh. Several other workers[12,14,18–22] have applied methods similar to Kane's to obtain impact ionization rates including the effect of realistic band structure for several semiconductors.

The numerical rate integration is highly computationally intensive, particularly near threshold where accurate results cannot be achieved in a reasonable amount of cpu time. An integration method is presented here which is similar to Kane's algorithm, but which can efficiently obtain rates near threshold in the Fermi's Golden Rule approximation, and is applied to several semiconductors to determine impact ionization rates and distributions of generated carriers.

II. METHOD

The rate of transition due to impact ionization for two electrons initially in states at $\mathbf{k}_1$ (the impacting electron) and $\mathbf{k}_2$ (the impacted electron) to final states at $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$, given by Fermi's Golden Rule is[23]

$$R_{II}(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_{1'},\mathbf{k}_{2'})=\frac{2\pi}{\hbar}|M_{if}|^2\delta(E_{1'}+E_{2'}-E_1-E_2), \tag{1}$$

where E_1, E_2, $E_{1'}$ and $E_{2'}$ are the energies of the electrons at $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ respectively (and where, for notational convenience, the $\mathbf{k}$ vector will be assumed also to denote a band index). The matrix element M_{if} is given by[24,23]

$$M_{if}=M_d-M_e, \tag{2}$$

[a)] Electronic mail: R.A.Abram@durham.ac.uk where the so called direct matrix element M_d is given by

$$M_d=\int \psi^*_{1'}(\mathbf{r}_1)\psi^*_{2'}(\mathbf{r}_2)V\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2)d^3\mathbf{r}_1\,d^3\mathbf{r}_2, \tag{3}$$

$$V(\mathbf{r}_1,\mathbf{r}_2)=\frac{e^2}{(2\pi)^3\epsilon_0}\int \frac{e^{i\mathbf{q}\cdot(\mathbf{r}_2-\mathbf{r}_1)}}{\epsilon(\mathbf{q},\omega)|\mathbf{q}|^2}d^3\mathbf{q}, \tag{4}$$

$$\mathbf{q}=\mathbf{k}_{1'}-\mathbf{k}_1,\quad \hbar\omega=E_1-E_{1'}, \tag{5}$$

and the so called exchange matrix element M_e is obtained by exchanging the indices $1'$ and $2'$ in Eqs. (3) and (5). In Eq. (4), $\epsilon(\mathbf{q},\omega)$ is the wave vector- and frequency-dependent dielectric function of the semiconductor, calculated here using the method of Walter and Cohen,[25] and $\mathbf{q}$ and $\hbar\omega$ are the wave vector and energy transfer in the transition. The pseudowave functions, obtained using the nonlocal pseudopotential method of Chelikowski and Cohen,[26] are not pure spin-up or spin-down but a linear combination

$$\psi_\alpha=\frac{1}{\sqrt{\Omega}}[u_\alpha|\uparrow\rangle+d_\alpha|\downarrow\rangle]e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{6}$$

where $\alpha=1$, 2, $1'$ or $2'$, and the term in square brackets is the Bloch periodic part of the wave function consisting of spin-up and -down parts u_α and d_α, which are each stored in the computer as expansions of plane waves. Using Eq. (6) the direct matrix element of Eq. (3) becomes[27,28]

$$\begin{aligned}M_d=&\frac{e^2}{\epsilon_0\Omega}\delta_{\mathbf{G}_u,\mathbf{k}_{1'}+\mathbf{k}_{2'}-\mathbf{k}_1-\mathbf{k}_2}\\ &\times\sum_{\mathbf{G}}\frac{1}{\epsilon(\mathbf{q}_d,\omega_d)|\mathbf{q}_d|^2}\{U^{(-\mathbf{G})}_{1',1}U^{(\mathbf{G}+\mathbf{G}_u)}_{2',2}\\ &+U^{(-\mathbf{G})}_{1',1}D^{(\mathbf{G}+\mathbf{G}_u)}_{2',2}+D^{(-\mathbf{G})}_{1',1}U^{(\mathbf{G}+\mathbf{G}_u)}_{2',2}\\ &+D^{(-\mathbf{G})}_{1',1}D^{(\mathbf{G}+\mathbf{G}_u)}_{2',2}\},\end{aligned} \tag{7}$$

where Ω is the crystal volume and U and D are the overlap integrals

$$U^{(\mathbf{G}')}_{\alpha,\beta}=\frac{1}{\Omega}\int u^*_\alpha e^{i\mathbf{G}'\cdot\mathbf{r}}u_\beta\,d^3\mathbf{r}, \tag{8}$$

$$D^{(\mathbf{G}')}_{\alpha,\beta}=\frac{1}{\Omega}\int d^*_\alpha e^{i\mathbf{G}'\cdot\mathbf{r}}d_\beta\,d^3\mathbf{r}, \tag{9}$$

and where $\mathbf{q}_d=\mathbf{G}+\mathbf{k}_1-\mathbf{k}_{1'}$, $\hbar\omega_d=E_1-E_{1'}$, and the Kronecker delta function ensures crystal momentum is conserved to within a reciprocal lattice vector $\mathbf{G}_u$. The exchange matrix element M_e is obtained by swapping the indices $1'$ and $2'$ in Eq. (7)

To obtain the total rate associated with an impacting carrier in state $\mathbf{k}_1$, Eq. (1) must be summed over all possible transitions from that state, i.e., a nine-dimensional integral over the variables $\mathbf{k}_{1'}$, $\mathbf{k}_{2'}$ and $\mathbf{k}_2$ if they are treated as continuous. The Kronecker delta function of Eq. (7) ensures that $\mathbf{k}_2=\mathbf{k}_{1'}+\mathbf{k}_{2'}-\mathbf{k}_1-\mathbf{G}_u$ and so the nine-dimensional integral is reduced to a six-dimensional one over $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$:

$$R_{II}(\mathbf{k}_1)=\frac{\Omega^2}{(2\pi)^2}\int\frac{2\pi}{\hbar}|M_{if}|^2\delta(\Delta E)d^3\mathbf{k}_{1'}\,d^3\mathbf{k}_{2'}, \tag{10}$$

where

$$\Delta E=E(\mathbf{k}_{1'})+E(\mathbf{k}_{2'})-E(\mathbf{k}_1)-E(\mathbf{k}_{1'}+\mathbf{k}_{2'}-\mathbf{k}_1). \tag{11}$$

The Dirac delta function requires that only energy conserving transitions contribute to the rate, and defines a surface within the six-dimensional volume over which the integration is performed.

In order to perform the integral of Eq. (10), the energies and wave functions of single electron states must be obtainable at $\mathbf{k}$ points throughout the Brillouin zone. Here, they are obtained using the empirical pseudopotential method.[26] The overlap integrals [Eqs. (8) and (9)] are calculated using the pseudowave functions, expanded as plane waves. Because of the complicated nature of the band structure, the integration of Eq. (10) must be performed numerically. Following the method of Kane,[17] the Dirac delta function $\delta(\Delta E)$ is approximated by a top-hat function of finite width $2\delta e$. The six-dimensional volume integral can then be performed by the Monte Carlo method in which final state pairs $\mathbf{k}_{1'},\mathbf{k}_{2'}$ are chosen randomly throughout the Brillouin zone and the integrand evaluated only for those which satisfy the approximate energy conservation condition imposed by the top-hat function, i.e., $|\Delta E|\leqslant\delta e$. However, the integral is typically slow to converge due to the fact that only a very small fraction of randomly sampled final state pairs satisfy the approximate energy conservation condition, especially when the impacting vector $\mathbf{k}_1$ lies near threshold. The rate of convergence can be increased by increasing the energy width δe of the top-hat function, but this then leads to errors due to the poor approximation of the Dirac delta function. Below, an integration method is described which avoids this problem.

A. Numerical rate integration

Defining the total volume of final state phase space to be sampled as Ω_0 and the volume in which the top-hat function is nonzero as $\Omega_{\delta e}$, the problem to be overcome concerns the fact that $\Omega_0\gg\Omega_{\delta e}$. Here, the problem is solved by restricting the sampling points to some volume Ω_B which is much smaller than Ω_0, but nevertheless completely encloses $\Omega_{\delta e}$. The method of reducing Ω_0 to Ω_B is an iterative procedure, where the volume over which the integration is to be performed is divided into subvolumes, and then those which do not contain $\Omega_{\delta e}$ are discarded. The algorithm is represented schematically in Fig. 1. In diagram A of the figure, the total volume of final state phase space Ω_0 is represented as the square, with the volume selected by the top-hat function $\Omega_{\delta e}$ lying between the ellipses. The space enclosed by the ellipses is small in comparison to the space contained in the square, and so random sampling within the square will lead to few ''hits'' lying within the ellipses. In diagrams B–D the sampling volume is iteratively refined so as to remove regions not containing the volume of interest between the ellipses. In diagram B, the original volume of phase space has been bisected in each direction to form a number of subvolumes. In the two-dimensional representation, four subvolumes are formed, whereas in the actual six-dimensional phase space, 64 subvolumes are created. Diagrams C and D show the situation after two and four iterations, respectively,

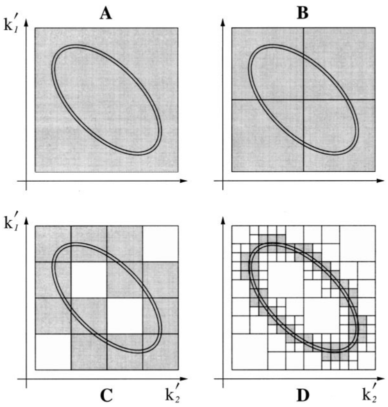

FIG. 1. Schematic representation of the method of reducing the volume of phase space to be sampled.

with regions not containing $\Omega_{\delta e}$ having been removed. It can be seen that after four iterations, the volume Ω_B remaining undiscarded (shaded gray) is considerably reduced from the original Ω_0, but nevertheless completely encloses $\Omega_{\delta e}$.

1. Retaining or discarding subvolumes

To implement this algorithm, a method for rapidly determining whether or not a subvolume contains any of the volume of $\Omega_{\delta e}$ is required. Beattie[29] and Wilson[30] have noted that for an impacting carrier at a given point in **k** space, $\mathbf{k}_1$, the energy difference function ΔE of Eq. (11) will have minimum and maximum values with respect to $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. If

$$\Delta E_{\min}(\mathbf{k}_1) \leqslant 0 \leqslant \Delta E_{\max}(\mathbf{k}_1), \tag{12}$$

then there must be a point in $\mathbf{k}_{1'},\mathbf{k}_{2'}$ space at which $\Delta E = 0$, satisfying the requirement for energy conservation. Crystal momentum conservation has already been satisfied by writing $E(\mathbf{k}_{1'}+\mathbf{k}_{2'}-\mathbf{k}_1)$ in place of $E(\mathbf{k}_2)$ in Eq. (11), and so it follows that there exists a surface of allowed transitions from $\mathbf{k}_1$ and a carrier in this state can initiate impact ionization. The values of $\Delta E_{\min}(\mathbf{k}_1)$ and $\Delta E_{\max}(\mathbf{k}_1)$ for final states lying within a subvolume can be determined. (Note that the minima and maxima within a subvolume will not in general correspond to stationary points.) If

$$\Delta E_{\min} \leqslant +\delta e \quad \text{and} \quad \Delta E_{\max} \geqslant -\delta e, \tag{13}$$

then it follows that somewhere in the subvolume $|\Delta E| \leqslant \delta e$, therefore $\Omega_{\delta e}$ lies within it and it should not be discarded.

In fact, searching subvolumes for maxima and minima in ΔE is very cpu time consuming. Instead the following energies are defined:

$$E^{\min} = E_{1'}^{\min} + E_{2'}^{\min} - E_1 - E_2^{\min}, \tag{14}$$

$$E^{\max} = E_{1'}^{\max} + E_{2'}^{\max} - E_1 - E_2^{\max}, \tag{15}$$

where $E_\alpha^{\min} \ldots E_\alpha^{\max}$ is the range of energies for all states $\mathbf{k}_\alpha$ ($\alpha = 1'$, $2'$ or 2) associated with the subvolume in question. Since it is always the case that $E^{\min} \leqslant \Delta E_{\min} < \Delta E_{\max} \leqslant E^{\max}$, the condition for retaining a subvolume given by Eq. (13) can be replaced with

$$E^{\min} \leqslant +\delta e \quad \text{and} \quad E^{\max} \geqslant -\delta e \tag{16}$$

and all subvolumes containing $\Omega_{\delta e}$ will be retained as required. In fact, Eq. (16) is less efficient at reducing the volume to be sampled than Eq. (13), as it will retain some subvolumes that could have been discarded using Eq. (13). However, as each iteration makes the subvolumes smaller, the efficiency of Eqs. (13) and (16) converges. The advantage of using Eq. (16) is that three independent three-dimensional functions: $E_{1'}(\mathbf{k}_{1'})$, $E_{2'}(\mathbf{k}_{2'})$ and $E_2(\mathbf{k}_2)$, must be searched for minima and maxima instead of one six-dimensional function $\Delta E(\mathbf{k}_{1'},\mathbf{k}_{2'})$. Searching a three-dimensional function for maxima and minima is not significantly easier than searching a six-dimensional one, and in that sense little is gained by using Eq. (16) in place of Eq. (13). However, the advantage of the three-dimensional rule is that all the necessary maxima and minima can be precalculated and stored. It only remains to retrieve their values during the rate integration, which can be done very rapidly.

The three-dimensional minima and maxima are stored as follows. Each six-dimensional hypercubic subvolume within $\mathbf{k}_{1'},\mathbf{k}_{2'}$ space is the ''product'' of two cubic volumes in three-dimensional **k** space, one containing states $\mathbf{k}_{1'}$ and the other containing states $\mathbf{k}_{2'}$. Although it may be impractical to obtain values of $\Delta E_{\min}$ and $\Delta E_{\max}$ for each of N six-dimensional subvolumes, it is feasible to obtain and store values of $E_\alpha^{\min}$ and $E_\alpha^{\max}$ ($\alpha = 1', 2'$) for each of the $\sqrt{N}$ three-dimensional cubes from which the subvolumes are formed. Furthermore, by requiring that impacting vectors be located at the intersections of the stored three-dimensional final state cubes, the corresponding impacted states also lie within these cubes, and values of $E_2^{\min}$ and $E_2^{\max}$ can be stored for these also. During execution of the integration algorithm, values of $E^{\min}$ and $E^{\max}$ can be rapidly obtained from the stored values of $E_\alpha^{\min}$ and $E_\alpha^{\max}$ through Eqs. (14) and (15). Thus by adopting Eq. (16) in favor of Eq. (13) as the rule for determining whether subvolumes should be retained or discarded, it is possible to precalculate and store all the necessary minima and maxima for rapid retrieval during integration.

Note that the use of the iterative procedure represented in Fig. 1 and its use to locate the region of phase space containing $\Omega_{\delta e}$ is essential to avoid the need to consider the very large number of subvolumes that would be created if the algorithm were to begin simply by dividing up the whole of Ω_0 to the final required level of discretization. Note also that the requirement that impacting vectors lie at the nodes of the mesh of three-dimensional cubic volumes is not highly restrictive as the side length of these cubes is typically 1/64th or 1/128th of the length $\Gamma - X$ in the Brillouin zone. Using this algorithm, the width δe of the top-hat function can be set much lower (e.g., 1 meV) and the Brillouin zone discretized much more finely (e.g., into 128 cells from Γ to X) than could be achieved using the more direct Monte Carlo approach without placing unreasonable demands on computational resources.

III. RESULTS

The rate integration algorithm has been applied to the bulk unstrained semiconductors GaAs, $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}$ and $\mathrm{Si}_{0.5}\mathrm{Ge}_{0.5}$ (henceforth referred to as InGaAs and SiGe), all at 300 K. GaAs, an important semiconductor in the fabrication of high-speed devices, is a wide band gap material, in the sense that its impact ionization threshold lies at a sufficiently high energy to invalidate simple analytic band structure approximations. InGaAs and SiGe have applications in the design of devices for optical communications. They have narrower band gaps (direct in InGaAs, indirect in SiGe) than GaAs, but nevertheless are wide gap in the sense that ionization thresholds lie at energies above the applicability of simple analytic band approximations.

Band structure for each material was obtained using a nonlocal pseudopotential method[26] which includes the spin-orbit interaction. Form factors for each material are given in Ref. 31. Pseudowavefunctions were expanded in terms of 65 plane waves (130 expansion terms in all, with spin included). Energy and wave function data were stored on a grid and rapidly retrieved by an interpolation scheme during execution of the rate integration.

A. Ionization rates

As is generally observed when calculating impact ionization rates for semiconductors with real band structure, the calculated rates for both electrons and holes in all the materials are found to depend strongly on the **k** vector of the initiating carrier, and carriers at the same energy but different positions in **k** space can have widely varying rates.[3,14,20,21,32,33] The mean rate at a particular impacting carrier energy is obtained by averaging the rate due to carriers at all **k** vectors at that energy:

$$R_{\mathrm{av}}(E_i)=\frac{\int R(\mathbf{k})\,\delta[E(\mathbf{k})-E_i]d^3\mathbf{k}}{\int \delta[E(\mathbf{k})-E_i]d^3\mathbf{k}}. \tag{17}$$

Mean rates calculated in this way are presented for each material in Fig. 2. Note that in the indirect gap material studied, the electron initiated threshold lies at lower energy than the hole initiated one, with the opposite applying to the direct gap materials. At high carrier energies, rates for both carrier types in all three materials tend towards the same order of magnitude around $10^{14}\,\mathrm{s}^{-1}$, as has been noted elsewhere.[33,34] Allam,[35] however, suggests that this may be coincidental as, although the fundamental band gaps of these materials vary, they all have similar values of $\langle E_{\mathrm{ind}}\rangle$, defined as

$$\langle E_{\mathrm{ind}}\rangle=\tfrac{1}{8}(E_\Gamma+3E_X+4E_L), \tag{18}$$

where E_V is the energy gap between the top of the valence band and the bottom of the conduction band valley at V, leading to similar high energy behavior in the rate. In InP, for example, which has a greater value of $\langle E_{\mathrm{ind}}\rangle$, we should not expect such behavior.

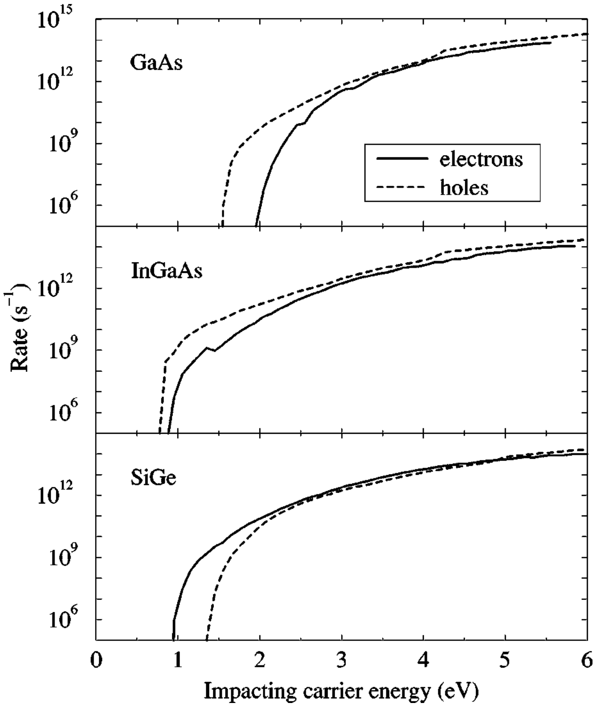

FIG. 2. Mean impact ionization rates of electrons and holes in GaAs, InGaAs, and SiGe.

The mean rates can be fitted by the expression

$$R(E)=A(E-E_0)^P, \tag{19}$$

where A, P and E_0 are the fitted parameters. For a fixed value of E_0, A and P are adjusted to give the best straight line fit by least squares analysis to log(R) vs log($E-E_0$). This fit has an associated rms value, which is itself minimized by adjusting the threshold energy E_0. The results of the fits applied to electron and hole rates in each material are given in Table I. Note that the fitted threshold energy is obtained without reference to the actual threshold energy calculated directly from the energy band structure, and so the discrepancy between the fitted and calculated values is an indication of the reliability of the fit.

Figure 3 compares electronic rates for GaAs obtained here with the results of other workers.[12,14,19,34] Values for the rate obtained by different authors range over more than an order of magnitude at any given energy. Possible sources of these discrepancies include numerical approximations made in the evaluation of the matrix elements, the use of different

TABLE I. Fitting parameters for rates shown in Fig. 2. Fit formula is: $R(E)=A(E-E_0)^P$ (with R in units of s^{-1} and E in eV).

		A	P	E_0 (fit)	E_0 (calc)
GaAs	e^-	1.4×10^{11}	5.2	1.89	1.85
GaAs	h^+	8.2×10^{10}	5.1	1.43	1.51
InGaAs	e^-	1.6×10^{10}	5.6	0.75	0.87
InGaAs	h^+	1.5×10^{11}	4.2	0.73	0.78
SiGe	e^-	4.6×10^{10}	4.9	0.84	0.91
SiGe	h^+	7.8×10^{10}	4.7	1.23	1.27

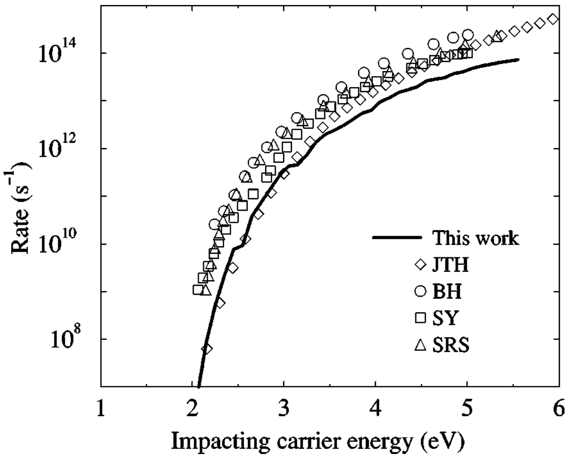

FIG. 3. Comparison of electronic rates in GaAs calculated here and by other workers—JTH (Ref. 14), BH (Ref. 12), SY (Ref. 34), SRS (Ref. 19).

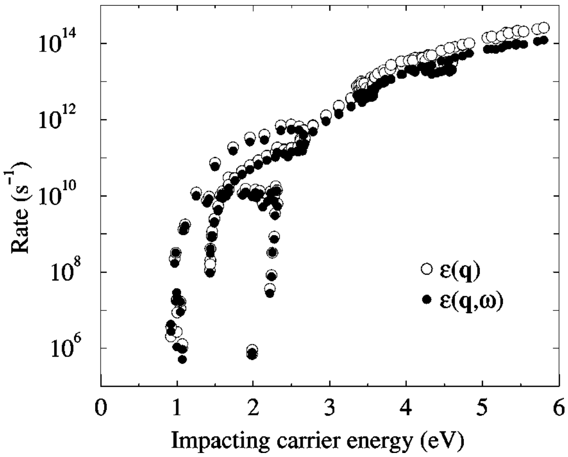

FIG. 5. Comparison of electron initiated rates in InGaAs calculated using a **q**- and ω-dependent expression for the dielectric function (as elsewhere in this work) and using a **q**-dependent expression.

band structure, and different implementations of the rate integration algorithm. The magnitude of the effect of these is investigated here.

The matrix elements are evaluated using the pseudowave functions expanded as plane waves. In Fig. 4, the convergence of the impact ionization matrix elements with respect to the number of plane waves used in the expansion of the wave function is plotted for InGaAs and SiGe. Convergence for InGaAs is very good when using 65 plane waves or more. In SiGe, the rate of convergence is poorer, with 307 plane waves being insufficient to ensure well converged results. Using 65 plane waves, the matrix elements are within ~30% of those obtained using 307. Unfortunately, using more plane waves requires considerably greater numerical effort. As will be examined in Sec. III B, **q** transfer is generally higher for transitions in SiGe than in InGaAs, and this is the likely cause of the slower convergence of the matrix element. In GaAs, mean **q** transfer is similar to that in InGaAs and so it is expected that the rate of convergence of the matrix element will be similar. All the results presented in Fig. 3 were obtained using 65 or more plane waves in the expansion of the pseudowave functions (except Ref. 34 which used a constant matrix element approximation), and so poor convergence of the matrix elements is not the cause of the discrepancies between the calculated rates.

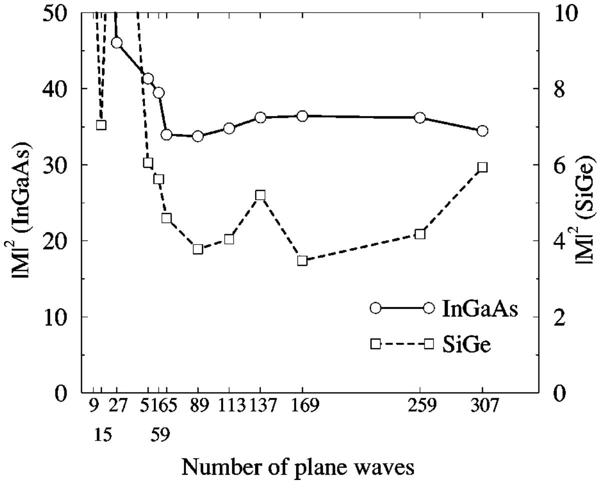

FIG. 4. Convergence of the impact ionization matrix elements with respect to the number of plane waves used in the expansion of the pseudowave functions.

Of the results compared in Fig. 3, those from Ref. 19 used a **q**-dependent expression for the dielectric function while those from Refs. 12 and 14 used **q**- and ω-dependent expressions. (The dielectric function does not enter into the constant matrix element approximation of Ref. 34.) Figure 5 compares rates obtained in InGaAs using **q**-dependent and **q**- and ω-dependent expressions for the dielectric function. As might be expected, the disagreement between the two calculations is greatest at high energy, but the magnitude of the discrepancy is small in comparison to the range of rate values seen in Fig. 3. Thus, the sensitivity of the rate to variations in the dielectric function is not sufficiently great to account for the disagreement between different authors, particularly at low impacting carrier energies.

The sensitivity of the calculated rates to the details of the band structure was tested by comparing rates calculated for GaAs using two different band structure calculations. One was the nonlocal pseudopotential calculation of Chelikowski and Cohen[26] which includes the effect of the spin-orbit interaction, as used elsewhere in this work, and the other was the local pseudopotential of Cohen and Bergstresser,[36] which neglects spin. The results plotted in Fig. 3 were all obtained from local pseudopotential band structure calculations, except for those calculated here. Rates calculated using each method are plotted in Fig. 6. The difference is greatest at low impacting energy where the rates are most sensitive to the details of the energy band structure, but generally the discrepancy is small in comparison to differences seen between the results of different workers in Fig. 3. Stobbe *et al.*[19] have also examined the sensitivity of the impact ionization rate to the choice of band structure method, finding a similar degree of sensitivity as here.

The three factors examined above: convergence of the matrix element with respect to the number of plane waves used to expand the wave function, the use of different expressions for the dielectric function, and the use of different

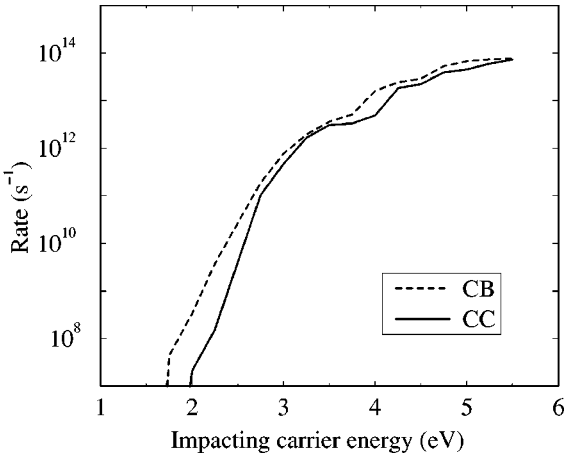

FIG. 6. Comparison of electron initiated rates in GaAs, calculated using local pseudopotential method [Cohen and Bergstresser (Ref. 36)] and nonlocal method with spin [Chelikowski and Cohen (Ref. 26)].

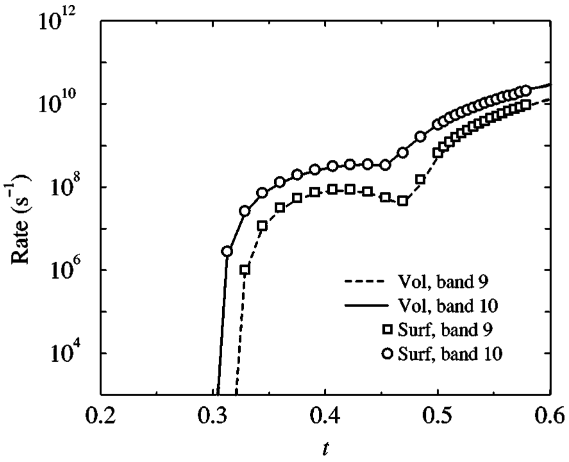

FIG. 7. Comparison of rates in GaAs initiated by electrons in the first conduction band along the line $\mathbf{k}=(t,0.055+t/2,0)$, calculated using the volume integration algorithm developed here and the surface integration algorithm of Beattie (Refs. 22 and 37).

methods of obtaining the crystal band structure, do not have a sufficiently great influence on the calculated rate, either individually or in combination, to account for the magnitude of the disagreement between the various results of Fig. 3. It therefore seems likely that the most significant factor in accounting for the discrepancies is the differences in the implementation of the rate integration scheme, particularly with regard to the degree of discretization of the Brillouin zone and the width of the function used to approximate the energy conserving Dirac delta function. The numerical reliability of the algorithm developed here was tested by comparing results obtained by it with those obtained by a different numerical algorithm developed by Beattie.[22] In Beattie's approach, the energy conserving Dirac delta function of Eq. (10) is treated exactly (as opposed to the top-hat approximation made here) and the integration is carried out as a surface rather than volume integral. The application of Beattie's algorithm to real band structures has been described elsewhere.[37] Here, the two algorithms are used to calculate rates in the first conduction band of GaAs along the line $\mathbf{k}_1=(t,0.055+t/2,0)$. The results are compared in Fig. 7. The two algorithms approach the problem of integrating over allowed transitions in quite different ways, and the similarity in their final results can be taken as a good indication that they are both numerically reliable.

B. Final state distributions

Investigation of the distributions of states of the final and impacted particles is of interest, both in the implementation of transport simulations and in gaining a better understanding of factors influencing the impact ionization rate. Figure 8 presents the mean energy of the final and impacted particles for electron and hole initiated transitions in each material. In all three materials there is an approximately linear dependence of the mean energy of each of the generated carriers on the impacting carrier energy. The scattering of points is due to the dependence of final state energies on the actual $\mathbf{k}$ vector of the initiating particle rather than just its energy, particularly at lower impacting energies. This is similar to the situation with the rates themselves, which are also explicitly $\mathbf{k}$- rather than just energy-dependent, especially at low energy. The direct gap materials show similar behavior with each generated electron generally taking a slightly greater share than the generated hole of the available energy in both electron and hole initiated cases, while in the indirect gap material, the distribution of the available energy between the generated carriers is roughly equal on average. Naturally, a given impacting carrier does not generate secondary carriers

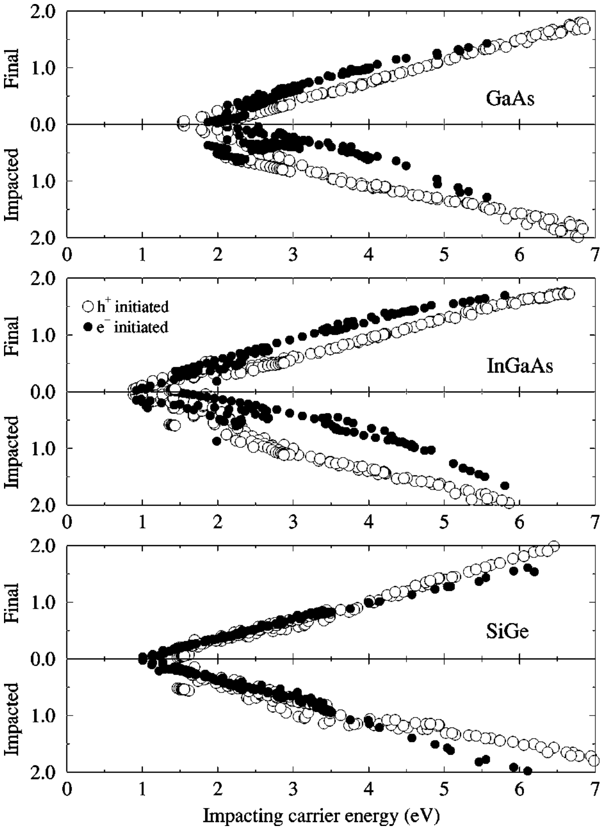

FIG. 8. Mean energies (in eV) of final and impacted states for electron and hole initiated transitions in each material.

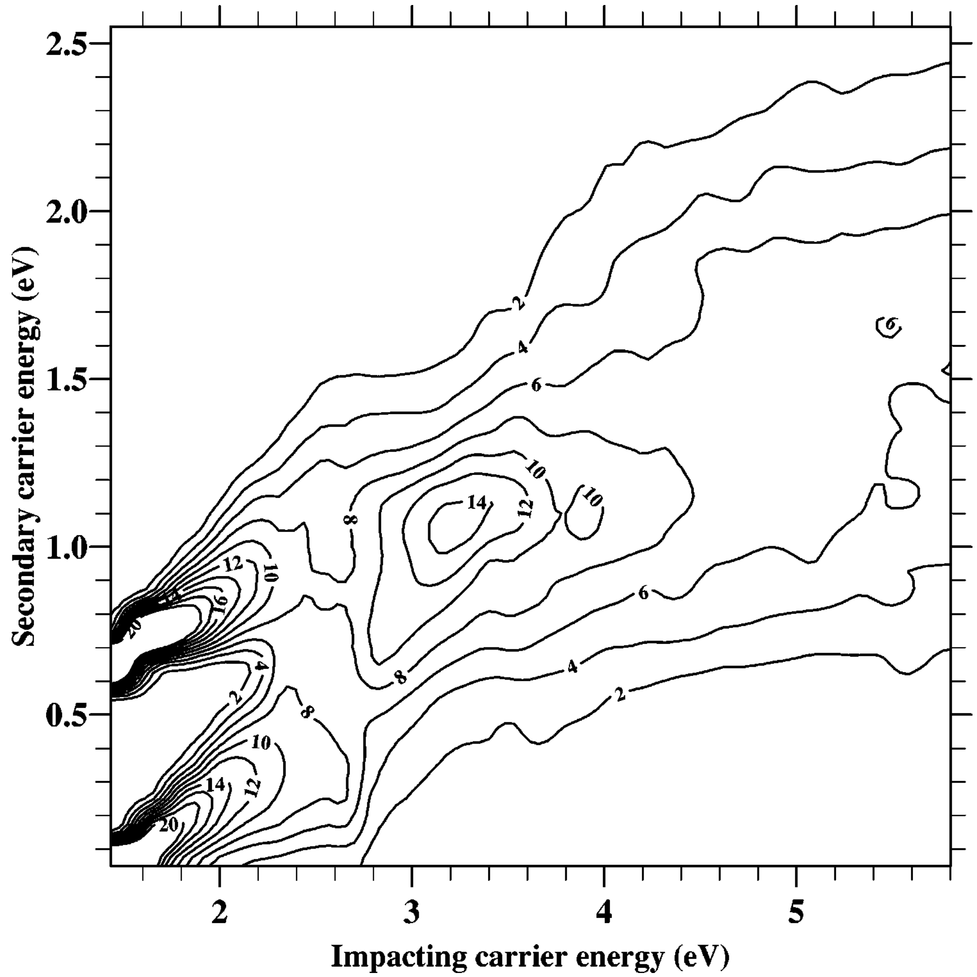

FIG. 9. Energy distribution of generated electrons for transitions initiated by electrons in the second conduction band of InGaAs.

at a specific energy but rather with a distribution of energies. Figure 9 shows the distribution of generated electrons for transitions initiated by electrons in the second conduction band of InGaAs. The influence of the conduction band density of states is clear at low energy, with the two peaks in the distribution corresponding to generated electrons lying in the Γ and *X* valleys, while at higher impacting energies the final state distribution is smoothed out into a single flat peak. The influence of the density of states is not generally as marked as in Fig. 9. Distributions of generated holes or electrons generated by hole initiated ionization, for example, are more featureless single peaks.

Examination of the mean momentum transfer during transitions highlights some interesting differences between the direct and indirect gap materials studied here. Figure 10 compares the mean **q** transfer for transitions initiated by electrons located along the line $\Gamma - K$ in the second conduction band of InGaAs and SiGe. For each material, the line with solid circles indicates the mean value of **q** calculated by weighting all transitions by the squared magnitude of the corresponding matrix element, and the line with empty circles represents the unweighted mean. Two things about the plot are worthy of note. First, the weighted mean **q** transfer is considerably lower in InGaAs (and in GaAs, which shows similar behavior to InGaAs) than in SiGe. It is likely that the higher values of **q** in SiGe are the cause of the slower convergence of the matrix element plotted in Fig. 4. Second, in InGaAs, the weighted line lies at significantly lower **q** values than the unweighted line, indicating that the matrix elements act to favor low **q** transitions in InGaAs (and GaAs). In SiGe however, the weighted and unweighted lines lie close together, indicating that the matrix elements in SiGe do not act in favor of transitions with any particular **q** value. The favoring of low **q** transitions can increase threshold softness in the direct gap materials.[31]

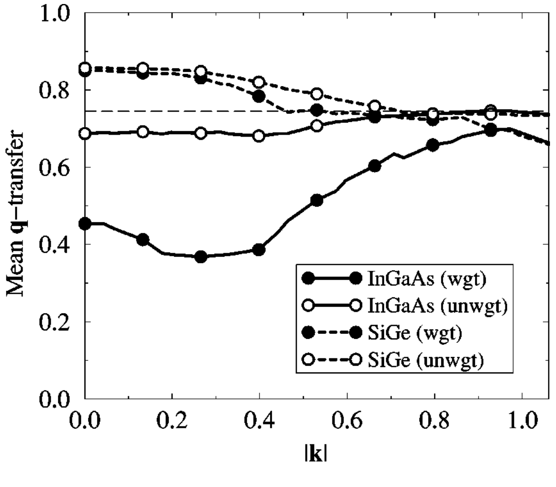

FIG. 10. Mean **q** transfer for transitions initiated by electrons in the second conduction band of InGaAs and SiGe.

IV. CONCLUSIONS

A numerical algorithm for calculating the impact ionization rate in the semiclassical Fermi's Golden Rule approximation, which does not suffer from the instability near threshold of algorithms similar to that of Kane,[17] has been presented. The reliability of this integration method has been demonstrated by comparison with another quite different algorithm (that of Beattie[22,37]). Possible causes of the disagreement between different authors[12,14,19,34] in calculated electron initiated rates in GaAs have been investigated, including the sensitivity of the calculated rate on the choice of method for obtaining the band structure, the number of plane waves used in the expansion of the pseudowave functions and the form of the expression used for the dielectric constant. While variation in each of these was found to lead to changes in the calculated rates, they alone could not account for the larger discrepancies seen between the results of the different authors, and it must be concluded that differences in the implementation of the rate integration algorithms account for much of the discrepancy.

The algorithm developed here was used to obtain impact ionization rates and distributions of generated carriers in the semiconductors GaAs, $In_{0.53}Ga_{0.47}As$ and $Si_{0.5}Ge_{0.5}$. Rates for electron and hole initiated transitions have been presented, and approximated by analytical fit formulas. The mean energies of the generated carriers were found to have an approximately linear dependence on the impacting carrier energy, with generated electrons generally taking a greater share of the available energy than the generated holes in the direct gap materials, for both electron and hole initiated ionization. In the indirect gap material, the generated carriers shared the available energy approximately equally. Mean momentum transfer during transitions was found to differ significantly between the direct and indirect gap materials studied. In the direct gap materials, the mean **q** transfer was lower than in the indirect gap material. The higher **q** values in SiGe are assumed to be the cause of the slower rate of convergence of the impact ionization matrix elements with respect to the number of plane waves used to expand the pseudowave functions that were observed in this material. It was also found that in GaAs and InGaAs, the matrix elements act to favor the low **q** transitions, while in SiGe, this is not the case.

ACKNOWLEDGMENTS

The authors would like to thank A. R. Beattie for several helpful discussions. One of the authors (D.H.) is grateful to the EPSRC(U.K.) for funding this work.

[1] S. M. Sze, *Semiconductor Devices, Physics and Technology* (Wiley, New York, 1985), Chap. 3.

[2] M. H. Somerville, J. A. del Alamo, and W. Hoke, IEEE Electron Device Lett. **17**, 473 (1996).

[3] H. Mizuno, M. Morifuji, K. Taniguchi, and C. Hamaguchi, J. Appl. Phys. **74**, 1100 (1993).

[4] S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (Wiley, New York, 1981).

[5] G. E. Stillman and C. M. Wolf, *Infrared Detectors II*, Semiconductors and Semimetals, Vol. 12 (Academic, New York, 1977), Chap. 5.

[6] F. Capasso, *Lightwave Communications Technology*, Semiconductors and Semimetals, Vol. 22 (Academic, New York, 1985), Chap. 1.

[7] P. A. Wolff, Phys. Rev. **95**, 1415 (1954).

[8] W. Shockley, Solid-State Electron. **2**, 35 (1961).

[9] G. A. Baraff, Phys. Rev. **128**, 2507 (1962).

[10] B. K. Ridley, J. Phys. C **16**, 3373 (1983).

[11] H. Shichijo and K. Hess, Phys. Rev. B **23**, 4197 (1981).

[12] J. Bude and K. Hess, J. Appl. Phys. **72**, 3554 (1992).

[13] I. H. Oğuzman, Y. Wang, J. Kolník, and K. F. Brennan, J. Appl. Phys. **77**, 225 (1995).

[14] H. K. Jung, K. Taniguchi, and C. Hamaguchi, J. Appl. Phys. **79**, 2473 (1996).

[15] L. V. Keldysh, Sov. Phys. JETP **10**, 509 (1960).

[16] J. Kolník, Y. Wang, I. H. Oğuzman, and K. F. Brennan, J. Appl. Phys. **76**, 3542 (1994).

[17] E. O. Kane, Phys. Rev. **159**, 624 (1967).

[18] Y. Wang and K. F. Brennan, J. Appl. Phys. **76**, 974 (1994).

[19] M. Stobbe, R. Redmer, and W. Shattke, Phys. Rev. B **49**, 4494 (1994).

[20] Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo, and M. Takenaka, J. Appl. Phys. **75**, 3500 (1994).

[21] T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji, and C. Hamaguchi, J. Appl. Phys. **75**, 297 (1994).

[22] A. R. Beattie, J. Phys. C **18**, 6501 (1985).

[23] P. T. Landsberg, *Recombination in Semiconductors* (Cambridge University Press, Cambridge, 1991).

[24] A. R. Beattie and P. T. Landsberg, Proc. R. Soc. London, Ser. A **249**, 16 (1959).

[25] J. P. Walter and M. L. Cohen, Phys. Rev. B **5**, 3101 (1972).

[26] J. R. Chelikowski and M. L. Cohen, Phys. Rev. B **14**, 556 (1976).

[27] N. Sano and A. Yoshii, Phys. Rev. B **45**, 4147 (1992).

[28] M. Stobbe, A. Könies, R. Redmer, J. Henk, and W. Schattke, Phys. Rev. B **44**, 11105 (1991).

[29] A. R. Beattie, Semicond. Sci. Technol. **7**, 401 (1992).

[30] S. P. Wilson, S. Brand, A. R. Beattie, and R. A. Abram, Semicond. Sci. Technol. **8**, 1546 (1993).

[31] D. Harrison, R. A. Abram, and S. Brand, J. Appl. Phys. **85**, 8186 (1999), following paper.

[32] N. Sano, M. Tomizawa, and A. Yoshii, Jpn. J. Appl. Phys., Part 1 **30**, 3662 (1991).

[33] T. Kunikiyo, M. Takenaka, M. Morifuji, K. Taniguchi, and C. Hamaguchi, J. Appl. Phys. **79**, 7718 (1996).

[34] N. Sano and A. Yoshii, J. Appl. Phys. **77**, 2020 (1995).

[35] J. Allam, Jpn. J. Appl. Phys., Part 1 **36**, 1529 (1997).

[36] M. L. Cohen and T. K. Bergstresser, Phys. Rev. **141**, 789 (1966).

[37] S. P. Wilson, S. Brand, A. R. Beattie, and R. A. Abram, Semicond. Sci. Technol. **8**, 1944 (1993).