P. J. Price
J. M. Radcliffe*

# Esaki Tunneling

**Abstract: Tunneling, between propagating electron states, at a semiconductor junction is discussed in terms of customary quantum transition theory for the matrix elements of the hamiltonian between the states representing reflection of an electron (in either band) from the junction. The coordinate representation for the wavefunctions of these states is investigated, and tunneling probabilities (ratios of transmitted to incident current) are found for the "elastic" process proposed by Esaki and for the "phonon-assisted" processes. It appears that the tunneling may be described as taking place in a central region of the junction thinner than the space charge region. Current-voltage characteristics are calculated both for elastic and for phonon-assisted tunneling.**

## 1. Introduction

Esaki[1] discovered that *p-n* junctions made from material doped to degeneracy have a current-voltage characteristic with a negative-conductance portion, in the forward direction, of the kind shown in the accompanying illustration,[2] and that (in contrast to the second rising portion of the characteristic) the "hump" part of the characteristic is not grossly temperature dependent. This phenomenon (which has been verified, with similar details, in at least four other laboratories) is believed to represent majority carrier tunneling, between the bands,[3] inside the junction region. Esaki pointed out that the qualitative features of the characteristic could be simply accounted for by writing the junction current per unit area as

$$J=q\int[f_1(\varepsilon)-f_2(\varepsilon)]\rho_1(\varepsilon)\rho_2(\varepsilon)Z d\varepsilon \qquad (1)$$

where (see Fig. 1) the $f$'s are the Fermi functions (relative to quasi-Fermi levels $\phi_1$, $\phi_2$ differing by $qV$, where $V$ is the applied voltage), the $\rho$'s are the densities of states (reckoned relative to appropriately aligned band edges), and $Z$ represents the quantum-mechanical tunneling rate. The first positive conductance then correlates with the form of $f_1-f_2$, the negative with the form of $\rho_1\rho_2$.

One may formulate the problem of calculating the tunneling rate in the following terms: For a given energy $\varepsilon=\varepsilon_1$, let the normalized wavefunctions $\psi_1(\mathbf{r})$ have only the wavefunctions of the conduction band of the bulk crystal (i.e., without the "step" in electric potential at the junction) as components, and be such that $(\psi_1|H|\psi_1)$ (where $H$ is the electron hamiltonian *with the junction included*) is stationary subject to this constraint and equal to $\varepsilon_1$. The set of $\psi_1$ in the range of $\varepsilon_1$ which is of interest will represent electrons reflected at the junction. Similarly let a set $\psi_2$ be formed for the valence band. Then the matrix elements

$$M_{12}\equiv(\psi_1|H|\psi_2) \qquad (2)$$

give the tunneling rate per unit time as

$$(2\pi/\hbar)\,|M_{12}|^2\,\delta(\varepsilon_1-\varepsilon_2) \qquad (3)$$

in the usual way.

Kane[4,6] has treated the Zener phenomenon[3] for a homogeneous crystal in a uniform electric field, $F/q$, extending over a substantially larger distance than $(1/F)$ times the energy gap plus the band widths. Then the wavefunctions corresponding to the $\psi_1$ and $\psi_2$ above are spaced at discrete intervals, $aF$, in respect of one of their quantum numbers, and are each appreciable over a finite distance only. ($a$ is the lattice constant in the field direction.) Kane calculates a transition probability, in terms of the matrix elements he obtains, combined in the usual way with a density of states, $\rho$, equal to $1/aF$. This procedure is evidently not justifiable, at least by the standard proof, since the latter requires the transition rate out of the initial state to be large compared to the level spacing

*Summer Employee in the IBM Semiconductor Research Department, Poughkeepsie, N. Y.
Permanent address: Physics Department, Carnegie Institute of Technology, Pittsburgh, Pa.

divided by Planck's constant. It is not clear to us, in fact, that the Zener phenomenon may be described by a definite transition rate (unless lattice-scattering transitions between the electronic states are occurring at the same time).

This difficulty does not arise for the formulation outlined above, for a finite junction region in a large crystal, since the level spacing tends to zero in the usual way as the size of the crystal tends to infinity. For the case treated by Kane there is a selection rule on the matrix element: it is zero unless the component of wavevector parallel to the junction plane is the same for $\psi_1$ and $\psi_2$. For the corresponding case with the electric field localized as in the Esaki junction, the same selection rule holds. The density of states for a transition to the valence band is therefore $2l_2/hv_2$, where $l_1$ and $l_2$ are the lengths of the crystal on each side of the junction, $v_1$ and $v_2$ the components, normal to the junction plane, of the electron velocities corresponding to $\psi_1$ and $\psi_2$. The transition rate is then $|M_{12}|^2\, 2l_2/v_2\hbar^2$. Since the electron may be considered to return to the junction at intervals $2l_1/v_1$, there is a definite *tunneling probability*

$$P_{12}=\frac{4}{\hbar^2 v_1 v_2}\,(l_1 l_2 |M_{12}|^2)\,. \tag{4}$$

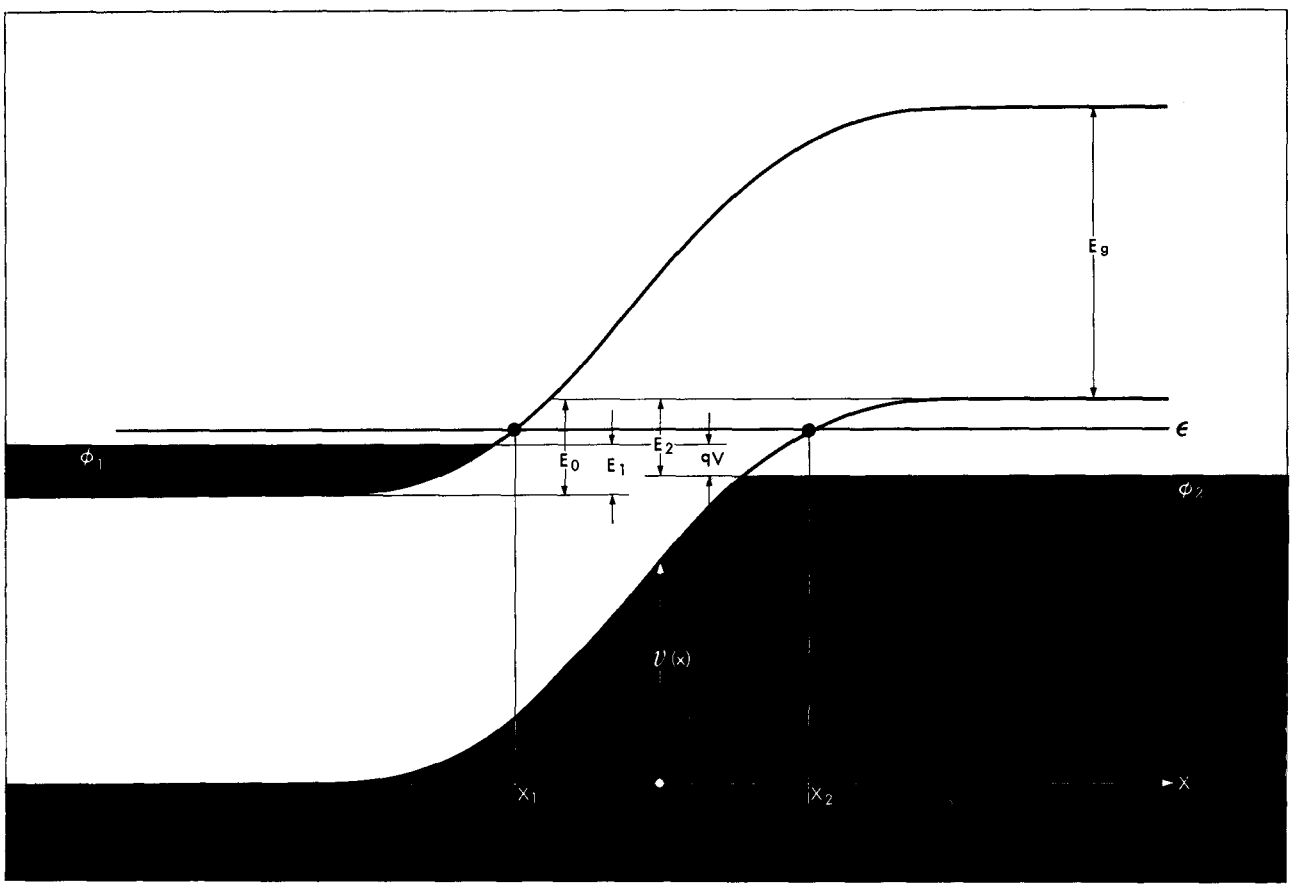
The expression in parentheses is in fact independent of $l_1$ and $l_2$. Then $P$ is the ratio of transmitted current to incident current. A calculation of the tunneling probability by direct solution (using the Green's function) of the wave equation has been found to give just the same result as is obtained, in Section 2, by using (4).

Kane's matrix element may be applied to Esaki tunneling when the junction field is constant (equal to $F/q$) in an interval somewhat longer than the region between the classical turning points for the energy level in question (i.e., from the left of $x_1$ to the right of $x_2$ in Fig. 1). The conversion from his normalization to ours (with $\int|\psi|^2dx=1$ in both) gives

$$|M_{12}|^2=\frac{v_1 v_2}{4l_1 l_2}\left(\frac{h}{aF}\right)^2 |M_{\text{Kane}}|^2\,, \tag{5}$$

$$P=\left(\frac{\pi}{aF}\right)^2 |M_{\text{Kane}}|^2\,. \tag{6}$$

*Figure 1* **Conventional junction diagram for an Esaki diode.**



In the present paper we shall attempt to discuss the character of the wavefunctions $\psi_1$ and $\psi_2$ in the coordinate representation, and to calculate the value of the tunneling probability, idealizing the actual electronic structure, as Kane does, by taking the bulk wavefunctions and energy levels to be those of a lightly doped crystal (i.e., Bloch states and energies, and their forbidden-range analogs) and the surfaces of constant junction potential to be parallel planes. This model deviates from the true situation for tunnel diodes in several obvious, and perhaps some obscure, ways.[5] It has yet to be determined how far the conclusions of the idealized treatments[6] are merely somewhat modified quantitatively, and how far new possibilities enter, in practice. We are hopeful that the formulation described here, if the uncertainties explained below are resolved, will prove to be flexible and general enough for the needed extensions of the theory. It is used below to examine the "elastic process" (corresponding to the Zener phenomenon) and the "phonon-assisted process" (in which there is absorption or emission of a lattice quantum during tunneling[7]). Presumably it could be used to treat tunneling involving exchange of energy with other electrons, if the latter could be described as an Auger process or as absorption or emission of a plasmon.[8]

## 2. Tunneling theory

For the idealized model referred to above, the $\psi_1$ and $\psi_2$ may be obtained by solving the one-dimensional Koster-Slater equation,[9]

$$[\varepsilon-\mathcal{U}(sa)]U(sa)=\sum_{t=-\infty}^{\infty}U[(s+t)a]\bar{\varepsilon}_t\,, \tag{7}$$

where $\mathcal{U}(x)$ is $(-q)$ times the junction electrostatic potential, for the coefficients $U_1(sa)$, $U_2(ta)$ of their expansions as linear combinations of Wannier functions,[10] by an adaptation of the WKB method. The result is

$$U_1(sa)=\frac{1}{2}\left(\frac{2}{l_1}\frac{v_1}{u_1(sa)}\right)^{\frac{1}{2}}\exp[i(s\eta_1-\tfrac{1}{2}\chi_1)-\theta_1(sa)] \tag{8}$$

and similarly for $U_2$. In (8),

$$\theta_1(sa)=\sum_{t<s}\alpha_1(ta) \tag{9}$$

where $\alpha_1$ is the (positive) imaginary part of the solution, $\kappa$, of

$$\varepsilon-\mathcal{U}[(s-\tfrac{1}{2})a]=\sum_{-\infty}^{\infty}\bar{\varepsilon}_t^{c}e^{it a\kappa} \tag{10}$$

and the sum (9) is over the cells to the right of the classical turning point ($x_1$ in Fig. 1),

$$\frac{d\varepsilon(\kappa)}{d\kappa}\equiv\hbar u e^{i\chi} \tag{11}$$

($u$ and $\chi$ real), and $\eta_1$ is zero if the band edge is at the zone center and $\pi/a$ if it is at the zone boundary. The right hand side of (10) is, for $\kappa$ real, the Fourier decomposition of the conduction band energy function $\varepsilon^c(k)$. Obviously, (8) is essentially the same as Eq. (A10) of Kane's paper. (The sums (9) of course will be replaced by integrals over $x$ in evaluating (16).) The solution (8) was obtained on the assumption that the series (10) converges, and the derivation used is in fact valid only where the number of terms before the right hand side of (7) approaches its sum is small compared with $[\hbar u/a^2F(sa)]^{\frac{1}{2}}$ (where $F(x)=d\mathcal{U}/dx$).[11]

We obtain a definite formula for $M$ by assuming that the two series (10), for the valence and conduction bands, converge in regions of the junction (i.e., in corresponding ranges of the forbidden gap) which overlap in the middle.[12] Then $U_1U_2$ has a maximum approximately at the point $x_0$ where

$$\alpha_1(x_0)=\alpha_2(x_0)\equiv\alpha_0\,. \tag{12}$$

(At the corresponding point in the forbidden range of the band energy diagram, $\alpha_1(\varepsilon)$ and $\alpha_2(\varepsilon)$ cross.) About this point, $U_1U_2$ is proportional to

$$\exp[-(sa-x_0)^2/b^2] \tag{13}$$

where

$$\frac{2}{b^2}=\frac{F(x_0)}{\hbar}\left(\frac{1}{u_1}+\frac{1}{u_2}\right)_{x=x_0}\,. \tag{14}$$

When $\eta_1=\eta_2$ (both band edges at the same point in the zone), we find

$$|M_{12}|^2=\frac{\pi}{4}\frac{b^2}{l_1l_2}|X_{cv}F(x_0)|^2\left(\frac{v_1v_2}{u_1u_2}\right)\exp(-2\theta_0) \tag{15}$$

where $X_{cv}$ is the matrix element of $x$ between the two zone center Bloch states and

$$\theta_0=\theta_1(x_0)+\theta_2(x_0)\,. \tag{16}$$

Then

$$P_{12}=\frac{\pi}{u_1u_2}|X_{cv}|^2\left(\frac{F(x_0)b}{\hbar}\right)^2\exp(-2\theta_0)\,. \tag{17}$$

The factor $\exp(-2\theta_0)$ in (17) is of course characteristic, in one form or another, of all treatments of tunneling, and should be normally the predominant factor in expressions such as (17) (see Footnote 5, however). The length $b$, measuring the width of the region in which $\psi_1$ and $\psi_2$ overlap, should be $\sim\sqrt{(ad)}$, where $d$ represents the width of the junction.

When $\eta_1-\eta_2=\pi/a$ then $M_{12}$ is given by a sum, over the overlap region, in which the terms alternate in sign. We find,[13] when $b^2>>a^2$, [5]

$$\frac{P_{12}(\eta_1-\eta_2=\pi/a)}{P_{12}(\eta_1-\eta_2=0)}\equiv Q=2\exp[-2(\pi b/2a)^2]\,. \tag{18}$$

The exponent in (18) should be comparable with that in (17). So according to the foregoing treatment a change in wavevector by $\pi/a$, while not actually forbidden, entails a considerable decrease in tunneling probability, except in junctions which are not many lattice constants

thick.[5] This is the basis for the conclusion in Section 4 that (according to same model as was assumed above) the tunneling process with absorption or emission of a phonon should actually predominate, at least when $b^2 >> a^2$.

The matrix element for absorption or emission of a phonon is found in a similar way, by inserting the electron-lattice interaction terms of $H$ into (2) and including the phonon factor in the $\psi$'s. Transitions between states with differing $k_y$ and $k_z$ are possible: the selection rule is that the $y$ and $z$ components of $\mathbf{k}+\mathbf{f}$ (where $\mathbf{f}$ is the phonon wavevector) are conserved. There is no "conservation of total wavevector" for the $x$ direction, but the matrix elements are proportional to

$$\exp[-b^2|f_x-(k^c_{0x}-k^v_{0x})|^2/4], \tag{19}$$

where $\mathbf{k}^c_0$ and $\mathbf{k}^v_0$ are the band edge wavevectors, and so fall off with increasing "change of total wavevector." This exponential (19) is thus the analog of the exponential in (18).

The assumption on which (15), (17) were derived, that the two series $\Sigma\bar{\varepsilon}^c_s \exp(isa\kappa)$, $\Sigma\bar{\varepsilon}^v_s \exp(isa\kappa)$, converge to real values, $\varepsilon$, with a non-vanishing common range, is certainly untrue for a *one-dimensional* "crystal." It is known[4,6,14] that the eigenvalues $\varepsilon^c(k)$, $\varepsilon^v(k)$ of the one-dimensional "crystal" hamiltonian are two branches of a single analytic function $\varepsilon(\kappa)$ which has a branch point for a real value of $\varepsilon$ in the forbidden gap. The Bloch functions $\psi^c(k,x)$, $\psi^v(k,x)$ also are branches of a single analytic function $\psi(\kappa,x)$ which, for the $\kappa$'s which give real $\varepsilon$'s in the forbidden range, are the unbounded "eigenfunctions" with complex wavevectors. Kohn[14] has shown that the Wannier functions fall off asymptotically like $\exp(-|x-sa|\alpha_0)$, where $\alpha_0$ is the distance of the branch point from the real axis of $\kappa$, so the $\bar{\varepsilon}_s$ must fall off like $\exp(-|s|a\alpha_0)$. From these facts one infers that the right hand side of (10) and its valence-band analog converge respectively for ranges of $\varepsilon$ which touch at the branch point energy, their sums being respectively equal to the two branches of the eigenvalue function $\varepsilon(\kappa)$, and do not converge to any real values, $\varepsilon$, in common.

These facts are proved only for one dimension, however, from properties of the one-dimensional Schrödinger operator, such as the fact that there are just two linearly independent eigenfunctions for each eigenvalue, and it can not be assumed that they hold for a one-dimensional section, $\varepsilon(k_x)$, of $\varepsilon(\mathbf{k})$ for a three-dimensional crystal. The conduction bands of germanium and silicon each have twice the degeneracy of the one-dimensional case, in a given direction in $\mathbf{k}$-space, for part of the energy range. The valence bands also have multiple degeneracy, but in a different way: if the energy on a line in $\mathbf{k}$-space at a distance $k_p$ from the zone center is $\varepsilon(k')$ (where $k'$ is measured along the line from the perpendicular) then $\varepsilon(\kappa')$ has a branch point at $\kappa'=\pm i\alpha(k_p)$, and $\alpha(k_p)\to 0$ when $k_p\to 0$. It follows[15] that, if $\bar{\varepsilon}_s$ are the coefficients of the sum (10) for $\varepsilon(k')$, then if $\Sigma\bar{\varepsilon}_s z^s$ has a radius of convergence $>1$ it must diverge at $z=\exp[a\alpha(k_p)]$: We return to this question briefly at the end of the present Section.

Now Kane shows that for conduction and valence bands with band edges at the zone center (specifically, In Sb), in the approximation that the "$\mathbf{k}\cdot\mathbf{p}$ interaction" operates between these two bands but is not appreciable between them and other bands, a branch point in $\varepsilon(\kappa)$ does occur. Therefore the branch point situation is evidently a possible one, for the three dimensional actuality, and it may conceivably even prevail in general. The discussion below indicates, however, that the conclusions from tunneling theory, for the "overlap situation" treated above and for the branch point situation respectively, are probably very similar, and the results on the voltage-current characteristic for practical purposes the same.

Let the branch point be at $\kappa_0=i\alpha_0$ (we take the real part of $\kappa_0$ as zero, for simplicity) and let

$$\varepsilon(\kappa)=\varepsilon^{cv}_0 \pm E_{cv}[a(\alpha-\alpha_0)]^{\frac{1}{2}}. \tag{20}$$

$E_{cv}$ is a constant of order $E_g$. (In Kane's model, $E_{cv}/E_g = (\hbar/2a)^{\frac{1}{2}}(E_g m_r)^{-\frac{1}{4}}$, where $m_r$ is the reduced mass for the pair of band masses.) The solution (8) prescribes, according to (20), that $U_1$ and $U_2$ tend to zero at $x_0$; but (8) is not valid in the neighborhood of this singular point.[16] In the range of $x$ for which (8) holds, $W_1(\alpha, sa) \equiv U_1(sa)\exp(sa\alpha)$ varies around its maximum (which is very near the position, $x_\alpha$, of the classical turning point for energy $\varepsilon_1-\varepsilon(i\alpha)$) as $\exp[-(x-x_\alpha)^2/c^2]$, where

$$c^2/u=2\hbar/F. \tag{21}$$

Where (20) holds, (21) becomes

$$c^2=(x_0-x)/(\alpha_0-\alpha). \tag{22}$$

Since the number of terms before the right hand of (7) approaches its sum is $t_s \simeq 1/a(\alpha_0-\alpha)$, the derivation of (8) becomes invalid where $c(\alpha_0-\alpha)\sim 1$. By (22), this happens where $x_0-x$, $c$ and $at_s$ are roughly equal, and given by

$$c^3 \sim a(E_{cv}/F)^2 \sim ad^2. \tag{23}$$

For smaller values of $x_0-x$ the values of $u'$ and $c'$, respectively giving the value of $U_1(x)$ according to (8) and the distance over which the $W_1$ having its maximum at $x$ falls off substantially, are no longer those given by (11) and (22). However, it can be shown that

a) Eq. (21) continues to hold, for $c'$ and $u'$;

and made plausible that

b) the order of magnitude of $c'$ continues to be that given by eq. (23).

Since the interband matrix elements of $x$ between Wannier functions fall off as $\exp(-|s-t|a\alpha_0)$, the matrix element of $\mathcal{U}$ will be

$$M_{12} \sim (U_1U_2)_{x=x_0} c'(x_0)^2 F(x_0) X^{cv}/a$$

$$\simeq \left(\frac{v_1v_2}{l_1l_2}\right)^{\frac{1}{2}} \frac{c'(x_0)^2}{u'(x_0)} F e^{-\theta_0}. \tag{24}$$

Now, Kane gives

$$M_{\text{Kane}}=\gamma aFe^{-\theta_0}, \tag{25}$$

where $\gamma$ is a dimensionless number $\sim 1$ (and we have, legitimately, written his exponent as $\theta_0$). On converting his matrix element to our normalization by means of (5), we see that (24) and (25) agree according to (a) of the preceding paragraph.

It is not obvious how the branch point type of situation discussed above would occur where the band edges are at different wavevectors.* If it does occur, we may conjecture that it does so in such a way that there are similar regions, of similar widths, within which $U_1 \exp(sa\alpha_0)$ and $U_2 \exp(-sa\alpha_0)$ are appreciable, but with the same additional rapidly oscillating factor as leads to (18). Again, $M_{12}$ and $P_{12}$ should consequently be reduced by a significant factor, and by a very considerable factor like (18) if $c'>>a$. Since the interband matrix elements, between Wannier functions, of the deformation potential operator[17] must also fall off as $\exp(-|s-t|a\alpha_0)$ in the branch point case, the theory for phonon-assisted tunneling in this case should be related to the theory in the overlap case in the same way as the theories for elastic tunneling are related in the two cases. For the branch-point case (with the oscillating factor) the matrix elements for the phonon-assisted process again will be appreciable over a range of $f_x$ which is of order $1/c'$.

The current-voltage characteristic is derived, in Section 4, by taking the matrix elements for phonon-assisted transitions between given states as all equal. This is justified in Section 4 by appeal to the overlap case, which results in (17) and (18). The foregoing discussion indicates, however, that this constancy of the matrix elements should hold (so long as $d/a$ is large enough[5]) for the branch-point case (provided the latter applies as envisaged, with an extra oscillating factor when the band edges are not coincident in the Brillouin zone). Then the same characteristic (apart from the absolute magnitude of the current) is to be expected in either case. On the other hand, Kane's matrix element (25) corresponds to an *absolute magnitude* of tunneling probability (and therefore of current) greater than (17) by a factor $\sim d/a$; and a similar relation may be expected for the tunneling probabilities in the phonon-assisted process.

The foregoing discussion suggests that there may well be some general validity to the concept of a region in the middle of the junction, much thinner than the space charge region when $d>>a$, where $\psi_1$ and $\psi_2$ interact. (One might say that the tunneling takes place in this region.) One may therefore propose to use the analysis at the beginning of this Section (i.e. the case where the series (10) converge to a common range of energies in the forbidden gap) as a theoretical model, as we do in Section 4 to treat phonon-assisted tunneling.

*Added in proof: If there should be a branch point for an intermediate value of the real part of $\kappa$ (at which $\varepsilon$ is complex) then one may expect the solutions (8), for both bands, to apply for values of $\alpha$ (corresponding to real $\varepsilon$) given by the analytic continuations of the series (10). That is, expect the *overlap situation* to obtain.

The question of the branch points near the valence band edges in germanium and silicon must, of course, be dealt with before any satisfactory theory specifically applying to these substances can be developed. One has meanwhile no more than an intuitive expectation that some treatment must exist in which coupled equations for the degenerate valence bands are solved simultaneously and which leads to results such as (8) gives for the maximum value of $\psi_1\psi_2$ —in which results the exponent $\theta_0$ has a similar order of magnitude, and depends on the same physical factors, as in the simple situation without degeneracy.

## 3. Current-voltage characteristics for elastic tunneling

It has been shown earlier that the transfer of charge from one side of the junction to the other can be characterized by a tunneling probability $P$, defined as the ratio of transmitted to incident currents. The current per unit area flowing in the forward direction is

$$J_{12}=q\frac{2}{(2\pi)^3}\int dk_x dk_y dk_z\, P(\varepsilon, k_y, k_z)$$

$$\times\frac{1}{\hbar}\frac{\partial\varepsilon^c(\mathbf{k})}{\partial k_x} f_1[\varepsilon^c(\mathbf{k})]\{1-f_2[\varepsilon^c(\mathbf{k})]\} \tag{26}$$

where $f_1(\varepsilon)$ and $f_2(\varepsilon)$ are the electron distribution functions on the two sides. As indicated in equation (26) the tunneling probability $P$ depends on the variables of integration.

The foregoing theory suggests that $P$ may be taken as a function of $k_y$, $k_z$ only, if $F$ is constant over the tunneling region. It is convenient to use $\varepsilon$ as a variable of integration, replacing $(\partial\varepsilon/\partial k_x)dk_x$ by $d\varepsilon$. Then the net current density is

$$J=J_{12}-J_{21}=\frac{q}{\hbar}\frac{2}{(2\pi)^3}\int d\varepsilon[f_1(\varepsilon)-f_2(\varepsilon)]\int P dk_y dk_z\,. \tag{27}$$

The integrations over $k_y$ and $k_z$ are restricted by the fact that $\varepsilon$, $k_y$ and $k_z$ are all conserved in the transition, and that also the initial and final states belong to the energy bands. If the constant energy surfaces are spherical,

$$0\le k_y^2+k_z^2\le \text{Min.}\left[\frac{2m_c\mathcal{E}_1}{\hbar^2}, \frac{2m_v\mathcal{E}_2}{\hbar^2}\right]\equiv k_r(\varepsilon)^2, \tag{28}$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are electron energies relative to the band edges, and

$$\mathcal{E}_1+\mathcal{E}_2=E_1+E_2-qV\equiv E_0 \tag{29}$$

is the amount by which the bands overlap when a voltage $V$ is applied across the junction. (See Fig. 1.)

Unless $\theta_0$ in (17) is not very large, $P$ should fall off rapidly over the range of the second integral of (27). We approximate this variation by writing

$$P=P_0\exp[-(a_c^2+a_v^2)(k_y^2+k_z^2)], \tag{30}$$

where

$$a_c^2=\alpha_0\hbar^2/m_cF \tag{31}$$

and similarly for $a_v$, and find

$$J=2\pi P_0\frac{qFm_r}{\alpha_0\hbar^3}\int[f_1(\varepsilon)-f_2(\varepsilon)]g(\mathcal{E}_1)g(\mathcal{E}_2)d\varepsilon \tag{32}$$

where $g(\mathcal{E})=1$ for $\mathcal{E}>0$, $g(\mathcal{E})=0$ for $\mathcal{E}<0$, and where $m_r$ is the reduced mass. $P_0$ is the tunneling probability for $k_y=k_z=0$. If the two distributions are Fermi-Dirac ones with Fermi levels $\phi_1-\phi_2=qV$ apart, then $f_2(\varepsilon)=f_1(\varepsilon+qV)$. So long as $KT<<E_1, E_2$ the integral (32) has a simple form, symmetrical for $V>0$. For $qV<$Min. $(E_1, E_2)$, it equals $qV$. It then remains constant while $qV$ increases to Max.$(E_1, E_2)$, and finally decreases to zero with constant slope $-q$. It should not be overlooked that, even with the idealized model and approximations which lead to this simple result, in (32) $P_0$ will decrease significantly while $V$ increases over the range of interest.

## 4. Phonon-assisted tunneling

According to the results of Section 2 the component of the crystal wavevector parallel to the junction plane is conserved when an electron tunnels directly from one side to the other. This would forbid, for instance, tunneling through a junction lying in (100) planes in Ge from states near the conduction band minima. However, tunneling is possible when this change in wavevector can be taken up by a phonon. It was also noted in Section 2 that although there is no strict selection rule for the wavevector component perpendicular to the junction plane, even in the idealized model, one would expect the tunneling rate when there is a large change in this component to be much smaller (in sufficiently thick junctions) than it is when the band edges are at the same point in the zone. This means that if the electron-phonon interaction is sufficiently strong, then the phonon-assisted tunneling rate can well exceed the direct tunneling rate through junctions in (111) planes in Ge and in (100) planes in Si.[18]

The fluctuating field due to lattice vibrations couples Wannier functions from the two bands just as the static junction field does. Perturbation theory then gives a tunneling rate directly. There should be distinct phonon transition processes for the different branches of the lattice mode spectrum. For simplicity we consider here a single branch only. (The total transition rate, and current, will be given by a sum over the branches.) Then the interband matrix elements of $H_2$, the electron-phonon interaction term of $H$, between Wannier functions at different lattice sites is

$$\tilde{D}_{cv}(\mathbf{R}_i-\mathbf{R}_j)N^{-\frac{1}{2}}\sum_{\mathbf{f}}\left(\frac{\hbar}{2M\omega(\mathbf{f})}\right)^{\frac{1}{2}}f$$

$$\times\{b_f\exp[i\mathbf{f}\cdot(\mathbf{R}_i+\mathbf{R}_j)/2]+\text{c.c.}\}. \tag{33}$$

Here $\tilde{D}_{cv}(\mathbf{R}_i-\mathbf{R}_j)$ is the interband matrix element of the deformation potential operator[17] between the Wannier functions, $M$ is the mass of the unit cell and $N$ is the number of unit cells in the entire crystal. The matrix element of $H_2$ between initial and final states on the same assumptions as lead to (15) (i.e. for the overlap case) is

$$(\psi_1|H_2|\psi_2)=\frac{D_{cv}}{N^{\frac{1}{2}}}\sum_{\mathbf{R}}U_1(\mathbf{R})U_2(\mathbf{R})\sum_{\mathbf{f}}\left(\frac{\hbar}{2M\omega}\right)^{\frac{1}{2}}f$$

$$\times\{e^{i\mathbf{f}\cdot\mathbf{R}}\langle 1|b_f|2\rangle+\text{c.c.}\} \tag{34}$$

where $D_{cv}$ is the interband matrix element of the deformation potential between Bloch states at the zone center ($\sim$ a few ev). We now evaluate this matrix element for tunneling from a conduction band minimum at $\mathbf{k}=\mathbf{f}_0$ when the junction plane is perpendicular to $\mathbf{f}_0$, which is taken to lie along the $x$ axis. The transition takes place between a conduction band state at $\mathbf{k}=\mathbf{f}_0+\mathbf{k}'$ and a valence band state at $\mathbf{k}=\mathbf{k}''$. It is clear from the form of Eq. (34) that the matrix element will vanish unless $f_y=k_y''-k_y'$, $f_z=k_z''-k_z'$. There is no such precise selection rule for $f_x$. For the same reason as in Section 2, we expect appreciable contributions to (34) from values of $f_x$ given by $|f_x-f_0|<\sim 1/b$ where $b$ is the distance over which the wavefunctions of our model overlap appreciably. When $b>>\hbar/(m_cv_{F1}+m_vv_{F2})$ it will thus be a good approximation to replace $\omega(\mathbf{f})$ by $\omega_0\equiv\omega(\mathbf{f}_0)$ and $f$ by $f_0$ in (34). Using the amplitudes $U$ given by Eq. (8), squaring the matrix element and summing over $f_x$, we obtain

$$\mathfrak{M}_{12}^2\equiv\Sigma_{f_x}|M(f_x)|^2=$$

$$\frac{\sqrt{2\pi}}{8}\frac{N^{\frac{2}{3}}}{N_1N_2}\left(\frac{v_1v_2}{u_1u_2}\right)e^{-2\theta_0}|D_{cv}|^2\frac{b}{a}\frac{\hbar f_0^2}{2M\omega_0}\begin{cases}n(\omega_0)\\ n(\omega_0)+1\end{cases} \tag{35}$$

where $n(\omega_0)$ is the phonon occupation number. (We ignore the effect on $M(f_x)$ of the energy difference $\varepsilon_1-\varepsilon_2$: This is unimportant so long as $d/a<E_g/\hbar\omega$.) $N_1$ and $N_2$ are the numbers of unit cells in the $n$ and $p$ sides respectively. The tunneling probability is now

$$P_{12}^{\pm}=\frac{2Na}{N^{\frac{2}{3}}v_1}\frac{2\pi}{\hbar}\sum_{\left(\begin{smallmatrix}\text{final}\\ \text{states}\end{smallmatrix}\right)}\mathfrak{M}_{12}^2\delta(\varepsilon_1-\varepsilon_2\pm\hbar\omega_0). \tag{36}$$

In the same way as leads to (32) we obtain, for $f_0=\pi/a$,

$$P_{12}=\frac{\sqrt{2\pi}}{8}\frac{P_0}{|X_{cv}|^2}\frac{|D_{cv}|^2}{\hbar\omega_0\alpha_0F}\frac{a}{b}\frac{m_v}{M}\times$$

$$\exp[-a_c^2(k_{1y}^2+k_{1z}^2)]\{n(\omega_0)g_2^-$$

$$+(1+n(\omega_0))g_2^+\}, \tag{37}$$

where $g_2^{\pm}\equiv g(\mathcal{E}_2\pm\hbar\omega_0)$. The net current density is

$$J=\frac{\sqrt{2\pi}}{4}\frac{qP_0}{\hbar^3\alpha_0^2|X_{cv}|^2}\frac{|D_{cv}|^2}{\hbar\omega_0}\frac{a}{b}\frac{m_cm_v}{M}\times$$

$$\int d\varepsilon\{g_1g_2^-[n(\omega_0)f_1(1-f_2^+)-(1+n(\omega_0))f_2^+(1-f_1)]$$

$$+g_1g_2^+[(1+n(\omega_0))f_1(1-f_2^-)-n(\omega_0)f_2^-(1-f_1)]\}, \tag{38}$$

where $g_1 \equiv g(\mathcal{E}_1)$, $f_1 \equiv f_1(\varepsilon)$, $f_2^+ \equiv f_2(\varepsilon+\hbar\omega_0)$, etc.

The current-voltage characteristic is especially interesting at temperatures such that $KT \ll \hbar\omega_0$. Then $n(\omega_0) \ll 1$ and the only processes possible are those in which a phonon is emitted. But for this to occur, there must be an unoccupied state to which the tunneling electron can go after giving up a quantum of energy to the lattice. It follows that there will be no appreciable tunneling for $-\hbar\omega_0 < qV < \hbar\omega_0$. This prediction has recently been confirmed for Ge and Si diodes at liquid helium temperatures.[7] The corresponding current-voltage relation for voltages near the threshold is obtained by substituting

$$\int f_1(\varepsilon)\,[1-f_1(\varepsilon+qV-\hbar\omega_0)]\,d\varepsilon \tag{39}$$

for the integral in (38), for the forward direction, and similarly for the reverse direction. The integral (39) reduces to $qV-\hbar\omega_0$ when the latter is large compared to $KT$. For $KT \gg \hbar\omega_0$ the integral in (38) is just $(2n(\omega_0)+1)$ times the integral in (32), and the two characteristics have the same form. We may make a rough comparison in general by writing

$$\frac{J(\text{phonon})}{J(\text{elastic},\ \eta_1=\eta_2)} = \pi^2\sqrt{2\pi}\,\frac{|D_{cv}|^2}{\hbar\omega_0\alpha_0|X_{cv}|^2F}\,\frac{a}{b}\,\frac{m_c+m_v}{M}$$

$$\times\,(2n(\omega_0)+1) \sim \frac{|D_{cv}|^2}{\hbar\omega_0 E_g}\,\frac{d}{b}\,\frac{m}{M}. \tag{40}$$

This ratio may be expected to be not $\ll 1$, and is to be compared with (18); so the phonon-assisted tunneling rate should be much higher than the elastic tunneling rate (for band edges not coincident but with the field in the direction of the difference of wave-vectors) so long as $b \gg a$.

*Added in proof:* G. H. Wannier has drawn our attention to a paper by Keldysh (*Soviet Physics JETP* **7,** 665 (1958)) in which the contribution of the electron-phonon interaction to the Zener phenomenon is calculated.


## Acknowledgments

We are indebted to discussions with E. N. Adams, W. P. Dumke, J. B. Gunn, S. L. Miller and M. I. Nathan on the physics of Esaki diodes and tunneling; to J. C. Marinace, S. L. Miller, and R. F. Rutz for discussions of their experimental results; and to R. L. Anderson and M. J. O'Rourke for making available a transcript of their notes on the Cornell 1959 Solid State Devices Conference. One of us (J.M.R.) would like to thank the members of the Semiconductor Research Department for their hospitality during his stay in Poughkeepsie.


## References and footnotes


1. L. Esaki, *Phys. Rev.* **109,** 603 (1958).
2. See back cover. Data obtained by R. F. Rutz.
3. C. Zener, *Proc. Roy. Soc.* **145,** 523 (1934).
4. E. O. Kane (to be published).
5. Also it should not be overlooked that, as the data on junction capacity shows (S. L. Miller, private communication), the width of real junctions may be not large enough (too small a number of lattice constants) for the kind of theory dealt with here to apply. Especially, the quantity $b \sim \sqrt{(ad)}$ which occurs below will not be large enough compared to $a$.
6. See also L. V. Keldysh, *Soviet Physics JETP* **6,** 763 (1958) and W. Franz, *Halbleiter Probleme* **3,** 1 (1956).
7. Holonyak, Lesk, Hall, Tiemann and Ehrenreich, *Phys. Rev. Letters* **3,** 167 (1959).
8. Some departures from the ideal model should be describable within the ideal model formulation. For example, in germanium or silicon the $\psi_1$ might have small admixtures of Bloch states from the zone center of the conduction band (owing to the donor ion fields), resulting in a contribution to the elastic tunneling probability large enough to compete with the ordinary process which is reduced by the factor (18). (This suggestion is due to W. P. Dumke.)
9. G. F. Koster and J. C. Slater, *Phys. Rev.* **95,** 1167 (1954).
10. Here $s$ and $t$ are integers labeling the lattice cells, and the $\bar{\varepsilon}_t$ are the coefficients appearing in (10) or in the corresponding equation for the valence band. The origins of the Wannier functions (the points $x=sa$) have been chosen so that the expectation of $x-sa$ for the corresponding Wannier function is zero. The matrix elements of $\mathcal{U}(x)$ between Wannier functions at different cells may be, and have been, neglected in (7). The matrix elements between Wannier functions of different *bands* give the interaction between $\psi_1$ and $\psi_2$, by means of (2), (8). For convenience, the solution (8) is normalized so that $\int |\psi|^2\,dx = 1/a$.
11. The solution breaks down near the band edge, of course, being just the ordinary WKB solution, for the effective mass equation, there. We have normalized (8) by using the known solution through the band edge region: see L. D. Landau and E. M. Lifshitz, *Quantum Mechanics, non-relativistic theory* (Addison-Wesley 1958), Sections 22 and 47.
12. The objection which will occur to some readers is discussed below. We suggest the names "weakly forbidden" for the forbidden energies for which (10) converges, and "strongly forbidden" for the energies, if any, for which (10) diverges.

number of neighboring $U$'s. Some important exact relations may be proved for the Wannier function representation (the $U$'s), however, and it appears that the simpler

13. The factor in front of the exponential is the product of a factor 1/2, representing an averaging over the possible "phases" of $x_0$ relative to the lattice points, in (13), and a factor $2^2$ which would be absent if $a\Sigma_s(-)^s \ldots$. were (incorrectly) replaced by $\int dx \cos(\pi x/a) \ldots$.
14. W. Kohn (to be published). More precisely, Kohn shows that a unique choice of phases for the Bloch wavefunctions gives Wannier functions with this property. Since these particular Wannier functions are presumably the ones most localized (each on its proper cell), we assume that they are the set for which the approximation made in (7), of replacing the (exact) sum over the matrix elements of $\mathfrak{V}$ by the single term, is best justified.
15. We are indebted to J. Cocke for a discusion on this point, and to W. Kohn for a discussion of the general situation.
16. In this region the sum of Wannier functions is anyhow a poor representation of the wavefunction in the sense that, because the $U$'s increase there as fast as the values (at this point) of the Wannier functions they multiply fall off, the value of the wavefunction depends on a large form of solution may be in this rather than in the ordinary Schrödinger representation.

    It appears from Kane's (A10) that the wavefunctions tend to a finite limit at $x=x_0$, since[14] the $\psi(\kappa)$ behave like $(\alpha-\alpha_0)^{-\frac{1}{2}}$ near the branch point. The approximations used to derive (A10) may break down in this limit, however (depending on the effect of the pole introduced into (A6) by the last term on the left).
17. G. D. Whitfield, *Phys. Rev. Letters* **2,** 204 (1959), and to be published.
18. The conclusions referred to in this paragraph, for the idealized model of the junction, may well be particularly liable to be modified by the departures which are to be expected, in practice, of the junction equipotentials from parallel planes.

*Original manuscript received September 4, 1959*

*Revised manuscript received September 21, 1959*