ZENER EFFECT

Zener effect is a term often used, in the band theory of solids in external fields, to refer to a general class of quantum mechanical phenomena also known as the Landau–Zener–Stückelberg (*LZS*) effect. Zener effect, as the name applies to solid-state phenomena, is usually caused by external dc electric fields, such as those occurring in degeneratively-doped *p–n* junctions of *Esaki diodes* or in less-heavily doped *p–n* junctions of *Zener diodes,* as well as in intermediately doped *backward tunnel diodes.* Zener effect involves the passage of an electron through a region in the energy-parameter space where the electron energy is forbidden (energy gap). The physical process is analogous to a train passing through a tunnel across a mountain barrier. Without the tunnel, the energy of the train is not enough to go over the barrier; this is circumvented by passing through a tunnel across the physical barrier. The mountain barrier is a potential barrier, defined here as a spatially forbidden region in classical mechanics. An energy gap is a forbidden region in energy-parameter space obtained from the quantum mechanical treatment of time-independent systems.

Whereas the absence of a tunnel negates the passage of a train across a mountain barrier, quantum mechanical tunneling deals with the probability of a particle tunneling through potential barriers and energy gaps. Hence the quantum mechanical tunneling probability is between 0 and 1. The quantum mechanical wavefunction belonging to the tunneling particle is responsible for the nonzero probability of tunneling through the forbidden region. This is illustrated in Fig. 1 for an electron tunneling through a potential barrier. The Zener effect may be approximately viewed as the tunneling of electrons through a potential barrier, as in the electron emission from cold metals in high electric fields.

BLOCH ELECTRONS IN EXTERNAL ELECTRIC FIELDS

In crystalline semiconductor materials, the electron energy levels form bands of very closely spaced energy levels, which arise from the splitting of the degeneracies of the energy levels of the individual atoms making up the crystal. These electrons are known as Bloch electrons; their wavefunctions are the Bloch functions. The corresponding atomic (i.e., localized) wavefunctions become Wannier functions. The energy level in each band is a function of a parameter $\mathcal{K}$, often referred to as the crystal momentum. These energy levels are periodic function of $\mathcal{K}$, with a “period in 3-D” defined by the so-called first Brillouin zone. The length of the first Brillouin zone along the direction of the real-space lattice vector a_1 is $2\pi\hbar/a_1$. Thus, knowing the energy levels in the first Brillouin zone is all that is needed. Figure 2 shows a simple model of the energy bands as a function of $\mathcal{K}$, and as a function of the position inside the crystal. Realistic energy bands in semiconductors are more complicated.

When a dc electric field, F, is applied to the crystal, to a first approximation the energy bands become tilted, since all the energy levels acquire an extra term given by $-eFx$, as shown in Fig. 3(a). Together with this, the crystal momentum $\mathcal{K}$ acquires a time dependence given by $\mathcal{K}(t) = \mathcal{K}_o + eFt$. Therefore, for a certain value of the energy, the electron executes oscillatory motion within the allowed energy band; this is known as Bloch oscillations. These oscillations are hardly observable in bulk materials, due to scattering effects, which destroy the coherence of the particle wavefunction. In the absence of scattering effects within the allowed energy band (this can be realized in superlattice structures where the band is broken into minibands of allowed energies), these oscillations are quantized in a more exact treatment, leading to an energy level structure known as the Stark levels, first introduced by Wannier (1). The discrete Stark energy levels are displaced from each other by one lattice constant and the separation between neighboring levels is given by $|eFa_1|$, where a_1 is the lattice vector along the direction of the electric field, as shown in Fig. 3(b). Indeed, Bloch oscillations, Stark levels, as well as Zener tunneling, have been experimentally observed in superlattice structures (2). Note that, for very narrow bands, the Stark levels of Fig. 3(b) resemble a ladder; hence the name Stark ladder is also used in the literature. The localized Stark level wavefunctions from within one band are generalization of the Wannier functions in the absence of the electric field. The corresponding generalization of the Bloch functions are the Houston wavefunctions, which are electric field dependent.

The influence of an external electric field on the optical properties of crystalline solids is a topic of sustained and growing research interest in optoelectronics. Optically induced transitions between two localized states belonging to different bands lead to an optical absorption coefficient of direct band gap semiconductors, which exhibits an exponential tail for photon energies less than the band gap. For photon energies larger than the band gap, the absorption coefficient has an oscillatory behavior. The exponential tail, as well as the periods and amplitudes of the oscillations, increase with the applied external electric field. This phenomenon is often referred to as the Franz–Keldysh effect.

It is more revealing to view the energy versus time-dependent crystal momentum, $E_n = (\mathcal{K}(t))$, in the repeated Brillouin zone scheme, where the subscript n is the band index. The arrows in Fig. 4 indicate the oscillatory motion of the electron within the respective allowed band of energies, corresponding to Bloch oscillations. If the particles stay within their respective energy bands, the time-dependent motion is called adiabatic. The dotted line shows the path in the forbidden region if the electron tunnels to another band of energies. This process, if it occurs, is called diabatic transition or Zener effect. From Fig. 4, one can readily calculate the period of oscillation within the allowed energy band from the relation: $|eF|T = 2\pi\hbar/a_1$, where T is the period.

The Zener effect is often treated as a transmission across a potential barrier (*TPB*), that is, a quantum mechanical tunneling through a classically forbidden region in space. A potential that is often used to calculate the Zener effect is the triangular potential, depicted by the dotted lines in Fig. 3(a). However, the result differs from the exact result derived below, which also accounts for the reduced effective mass entering in the problem. Moreover,

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

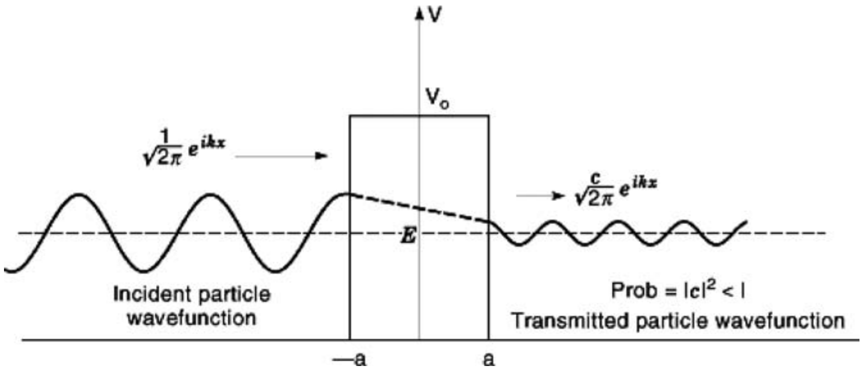
V
V$_o$
$\frac{1}{\sqrt{2\pi}}e^{ikx}$
$\frac{c}{\sqrt{2\pi}}e^{ikx}$
E
Incident particle wavefunction
Prob = $|c|^2 < 1$
Transmitted particle wavefunction
—a
a

Figure 1. Quantum mechanical wavefunction of a particle incident from the left of the potential barrier and transmitted to the right with probability approximately given by $|c|^2 \approx 4\ \exp\{-2\ \int^{a} -a \times 2m[v(x) - E]\ dx\}$, where $V(x)$ is the potential barrier and E is the particle energy.

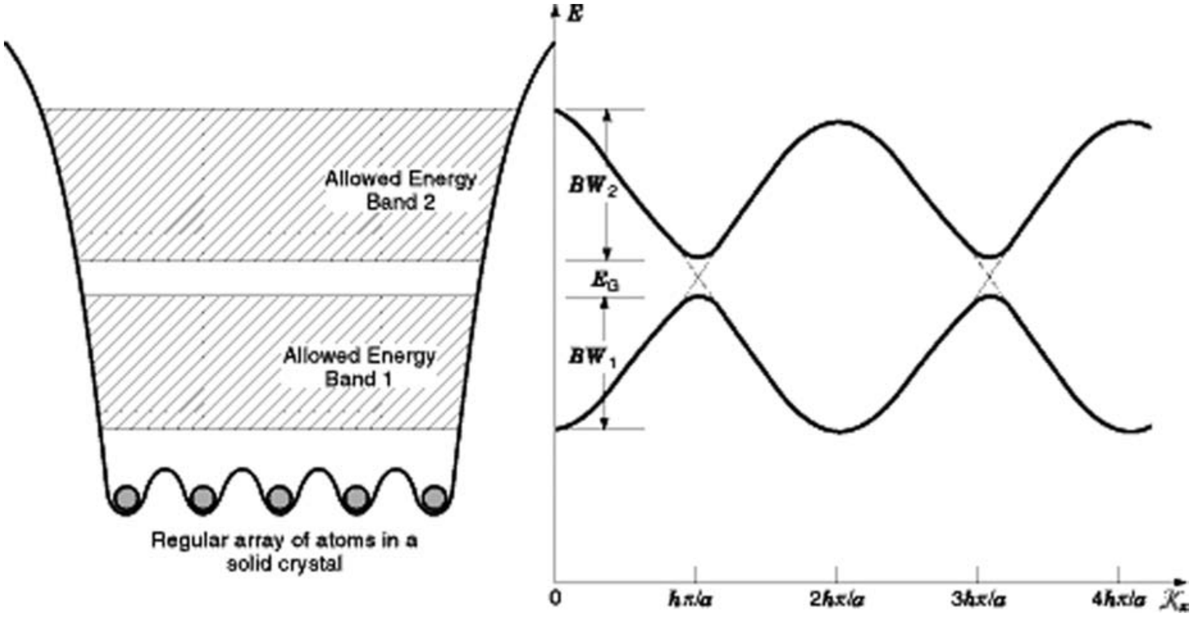

Figure 2. Simple model of Bloch electron energy bands as functions of crystal momentum $\mathcal{K}$ in repeated Brillouin zone scheme. The corresponding energy bands as function of position within the crystal is shown on the left.

the use of a parabolic potential barrier (3) to capture the exact result is physically unjustifiable and is only misleading. Thus, the difficulty with obtaining a physically justifiable answer when viewed as a TPB problem signifies that the Zener effect contains essentially a different and inherently time-dependent physics in the energy-parameter space.

Before Zener introduced the Zener effect in quantum transport, the tunneling of electrons through a triangular potential barrier was known to Fowler and Nordheim in the early days of quantum mechanics and was used to explain the phenomenon of electron emission from cold metals under the action of high electric fields. The proposed physical model is described in Fig. 5(a). The electrons in metal are confined by a potential wall, whose height is given by the work function φ plus the Fermi energy, ε_F. Upon the action of a high electric field, F, the potential barrier wall thickness is substantially decreased in a triangular fashion, allowing the electrons to tunnel across the forbidden region, as shown in Fig. 5(b). The potential energy inside the metal is shown undisturbed, since the electric field penetration into the metal is negligible. They derived the well-known Fowler–Nordheim tunneling current formula (C is a constant)

$$J = CE^2 \exp\left\{-\frac{4\sqrt{2m}\varphi^{3/2}}{3\hbar|eF|}\right\} \tag{1}$$

The dependence of the tunneling current on the exponential factor $\exp\{-4\sqrt{2m}\ \ \varphi^{3/2}/3\hbar|eF|\}$, where e is the unit charge, is noteworthy and characterizes the behavior of the transition probability across a triangular potential barrier. Note, however, that the Fowler–Nordheim problem is essentially a one ‘single-particle’ state problem; later it will be seen that the Zener tunneling effect is essentially a two-state time-dependent-parameter problem, as evidenced by the presence of reduced effective mass in the exact Zener tunneling expressions. The Zener effect describes nonadiabatic transitions in quantum mechanics, and embodies the violation of the Ehrenfest adiabatic principle, by virtue of

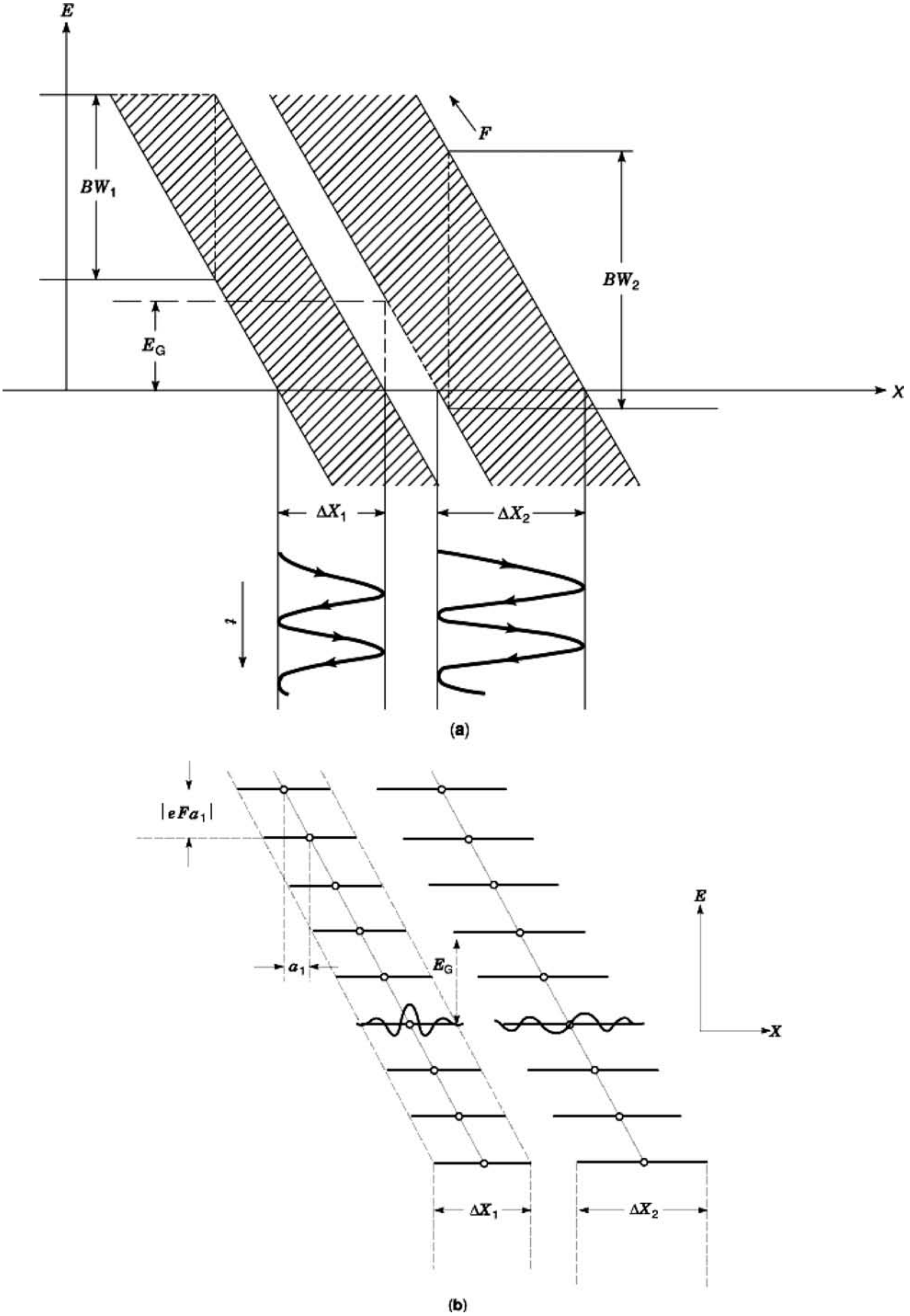

Figure 3. (a) Electron energy bands in the presence of an external electric field. The electron oscillates between all $\mathcal{K}$ values of the allowed energy bands in direction of the field. (b) Corresponding discrete energy level spectrum, called *Stark levels,* and schematic presentation of localized wavefunctions with localization ≈ bandwidth (BW)/$|eF|$.

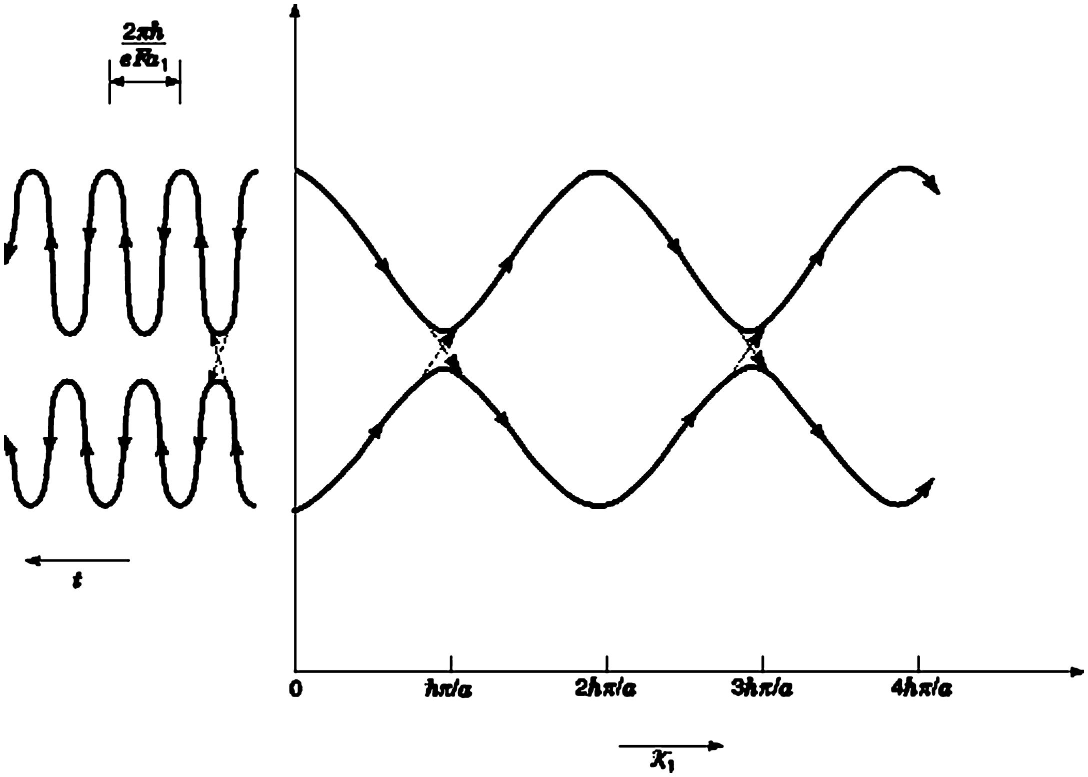

Figure 4. Bloch oscillations in the $\varepsilon_n(\mathcal{K})$ diagram and as a function of time. Through the time-dependence of $\mathcal{K}$, nonadiabatic transition is possible.

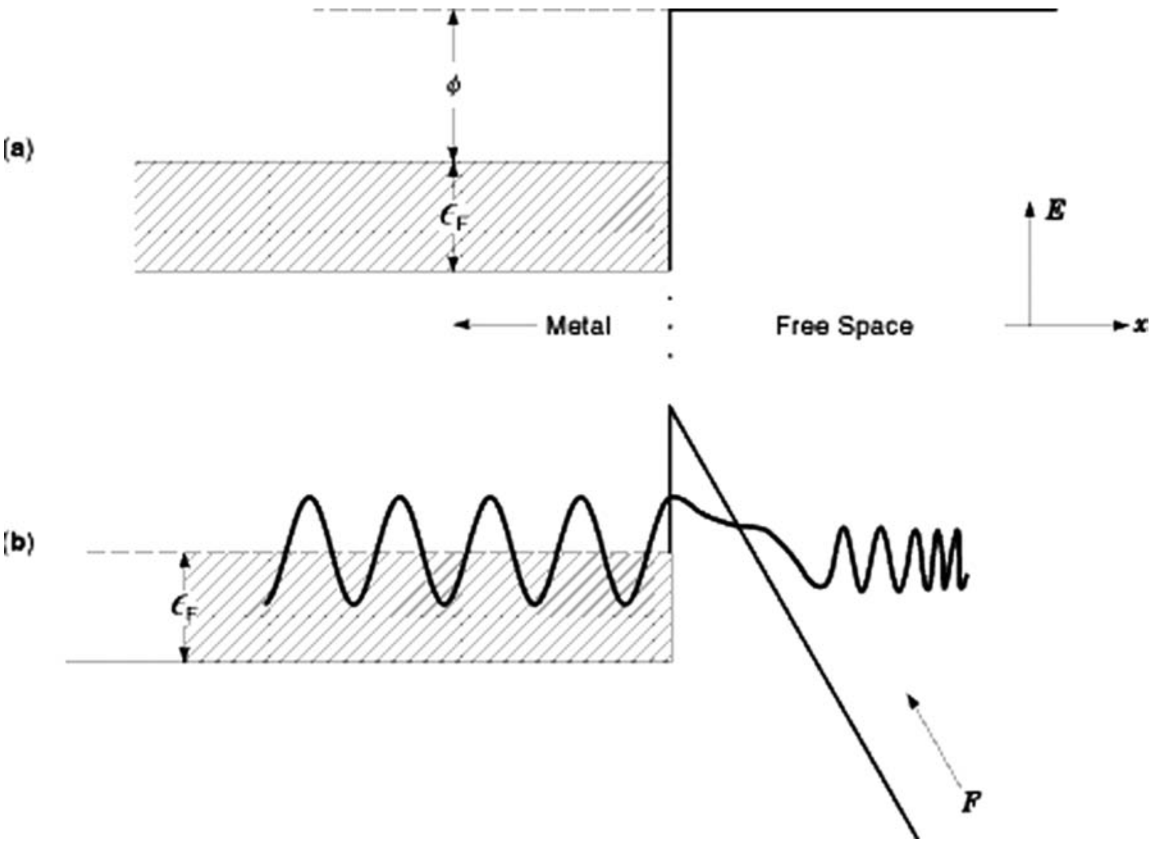

Figure 5. (a) Energy band-edge diagram of a metal surface, with the electrons bound to the metal interior; (b) Electron emission under high electric field is via tunneling through a triangular potential barrier.

the fast passage through almost degenerate states in the avoided crossing region, as described by Wannier (4)

Zener Effect in p–n Junction Diodes

In p–n junction diodes, the disturbance of the band energies is localized in position space and in the parameter $\mathcal{K}$ space. In Fig. 6, note that only for a time interval $2\tau_c$ the conduction electron acquires time dependence, $\mathcal{K}(t) = \mathcal{K}_o + eFt$ and hence traverses the states down to the bottom of the conduction band edge and back. Likewise, only for the time interval $2\tau_v$ does the valence electron acquire the time dependent $\mathcal{K}(t)$ and traverse the states up to the top of the valence band edge and back. This problem has been approximated as a TPB problem with a triangular potential barrier, and as a scattering problem by Fredkin and Wannier (5). This will be considered in the exact treatment below, by appealing to numerical results.

Indirect-Gap Zener Tunneling: The Keldysh Effect

When the conduction and valence bands in a p–n junction are centered at different values of $\mathcal{K}$ in the Brillouin zone (Fig. 7), in the so-called “indirect-gap” semiconductor, then the Zener “inertial” effect cannot occur. However, in the presence of quantized lattice vibrations (phonons), the total system is capable of creating the necessary allowed ‘intermediate’ electronic states for the Zener effect to occur. This can happen at some critical values of the bias voltage for forward and reverse bias. For example, in Ge-based p–n junction diodes, the valence band minimum is located at $\mathcal{K} = 0$ and the conduction band minimum is located at $\mathcal{K} = (\hbar\pi/a_1)(1, 1, 1,)$, separated by an energy gap, E_G. Under forward bias, in order to create the necessary electronic states from the states at the conduction band, the conduction electrons must create phonons in such a way as to give up crystal momentum given by $\mathcal{K} = (\hbar\pi/a_1)(1, 1, 1)$.

However, an electron cannot give up momentum without giving up energy equal to the energy of the phonon created. This process is illustrated in Fig. 7 under forward bias. Indeed, the resulting expression for the diabatic transition probability was shown by Keldysh (6) to be identical to the one obtained with “direct-gap” semiconductors, but with E_G replaced by $E_G - E_{ph}$, which is the reduced effective gap of the electronic states of the total system of electrons and phonons, E_{ph} is the energy of the phonon created. At low temperature in the absence of bias, and for small bias less then the critical value equal to the energy of phonon created, these “displaced states” of the conduction band are not available to assist the Zener tunneling. Thus, in the Esaki diode, which operates by Zener tunneling of conduction electrons to the unoccupied states in the valence band, no Zener tunneling current flows below the critical bias voltage. Above the critical voltage, the ensuing Zener tunneling current is phonon-assisted. There are four types of phonons, belonging to two acoustic and two optical branches of the vibration spectrum, with the same crystal momentum. Thus four current onsets are seen experimentally with increase in forward bias (7).

Under reverse bias, the valence electron tunnels to intermediate states of higher energies before being scattered to the conduction band states, as shown in Fig. 8. The effective energy gap is thus increased under reverse bias. Only the first onset of current due to the first phonon is experimentally resolved; the onset due to the other phonons is masked by the large current of the first onset.

ZENER EFFECT *p–n JUNCTION DEVICES*

Esaki Tunnel Diode

The first Zener effect device was discussed by Esaki, and is now known as the Esaki diode. An Esaki diode consists of a simple p–n junction that is very heavily doped with impurities, thus bringing the p and n sides to degeneracy. This means that the Fermi level is located within the allowed energy bands, resulting in ‘overlapping’ conduction and valence bands across the junction, even at zero bias. The Fermi level, measured from the bottom of the conduction band and from the top of the valence band, $E^{n}{}_{F}$ and $E^{p}{}_{F}$, respectively, is typically a few kT with a depletion layer width of less than 10 nm, of nanometer dimension much narrower than conventional p–n junction diodes. The basic operation of an Esaki diode is depicted in Fig. 9. Zener tunneling occurs from occupied states to the unoccupied overlapping states in the other side of the junction. The situation for different bias is illustrated in this figure. The negative differential resistance property of the Esaki diode can be used in microwave amplification.

‘Backward’ Diode

This is a variant of the Esaki tunnel diode, which exhibits essentially no peak current. This is accomplished by doping the two sides of the junction just to the threshold of degeneracy, so that there is no band overlap for forward bias. The current-voltage characteristics are shown in Fig. 10. A reverse bias causes the bands to overlap and the spatial width of the forbidden region to thin down (resulting in fast passage), so that a large tunneling current flows. Such a diode has a sharper resistance nonlinearity than a normal diode, with the resistance break right at zero bias. These diodes have no significant charge storage effects and can be used for fast switching.

Zener Diode

This is the name given to a heavily doped p–n junction diode with a sharp breakdown voltage, often used in voltage-regulating circuits. Most diodes described by this name actually break down by an avalanche or impact ionization process. A true Zener diode breaks down via a Zener tunneling process, which occurs only below 6 V to 8 V reverse bias. Beyond the breakdown voltage, the voltage across the diode remains approximately constant, independent of the current. Since breakdown occurs at significant reverse bias, the doping need not cause overlapping bands. A typical Zener diode characteristic is shown in Fig. 11. The relation between the Esaki diode, backward diode, and Zener diode lies in the decreasing doping levels or band overlaps from the Esaki diode down to the Zener diode. Another way to characterize the difference is that Zener diodes exhibit tunnel breakdown at reverse bias, the backward diode exhibits this at zero bias, and the Esaki diodes

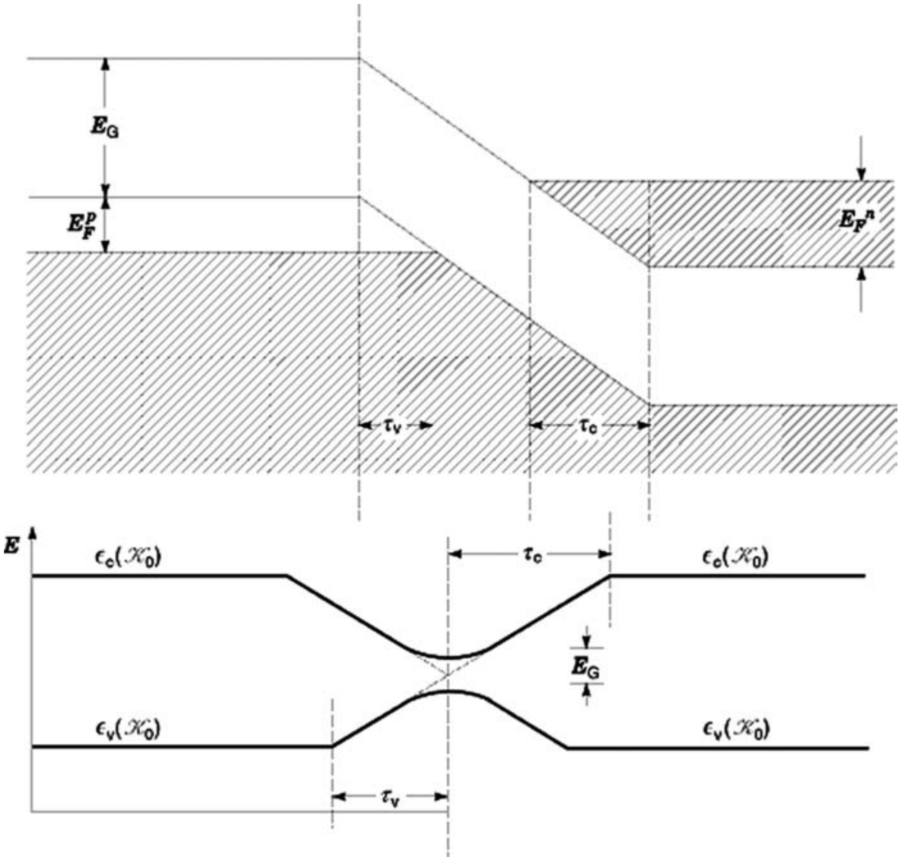
E_G
E_F^p
E_F^n
τ_v
τ_c
E
$\epsilon_c(\mathcal{K}_0)$
τ_c
$\epsilon_c(\mathcal{K}_0')$
E_G
$\epsilon_v(\mathcal{K}_0)$
$\epsilon_v(\mathcal{K}_0')$
τ_v

Figure 6. In a *p–n* junction, the time dependence of $\mathcal{K}$ is localized. Outside the junction region the electrons are field free.

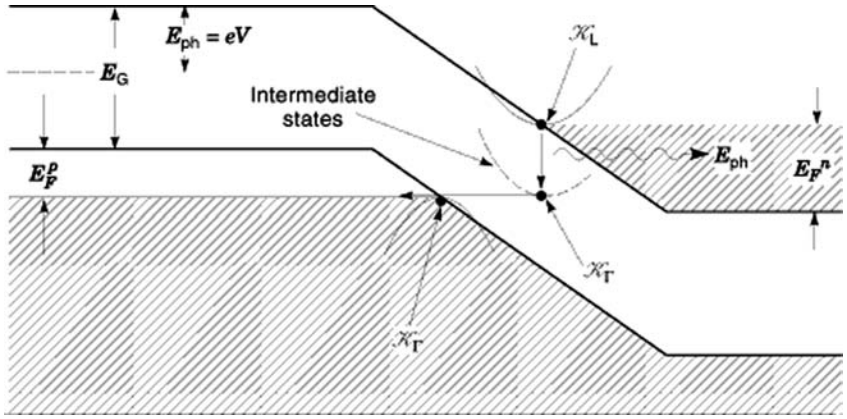

Figure 7. The onset of phonon-assisted Zener tunneling occurs when the applied potential equals the phonon energy with the desired momentum. The effective energy-band gap is reduced under forward bias.

only show less abrupt tunnel breakdown at zero bias, with current peaks at positive bias.

The temperature dependence of the breakdown voltage is opposite in sign for the two different breakdown mechanisms in Zener diodes. In the case of Zener breakdown, the breakdown voltage decreases with increase in temperature because of the increase in the valence-band electrons available for tunneling to the conduction band as the temperature rises. In the case of avalanche breakdown, the breakdown voltage increases with increase in temperature, since the scattering mean free path of the energetic electrons decreases as the temperature rises, thus producing more scattering per unit length at higher temperatures.

MEASUREMENT TECHNIQUES

Quantitative measurements of the electric field and impurity profiles in Zener diodes, solar cells, photodetectors, and metal-semiconductor field-effect transistors (*MESFET*) have recently been successfully demonstrated by Mil'shtein et al. (8) by using the scanning electron microscopy-dark voltage contrast (*SEM-DVC*). This tech-

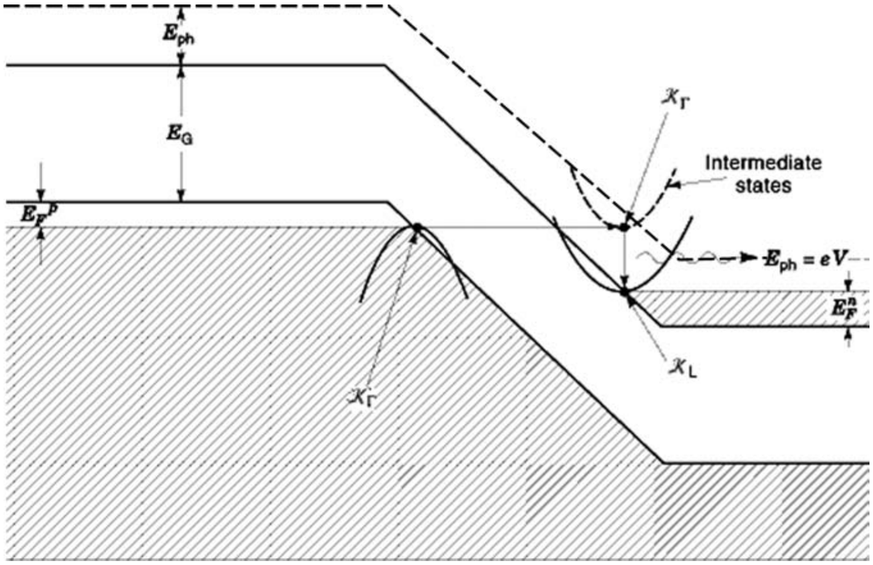

Figure 8. At reverse bias, the onset of phonon-assisted Zener tunneling occurs at applied potential equals to the phonon energy with the desired momentum as shown. The effective energy-band gap is increased under forward bias.

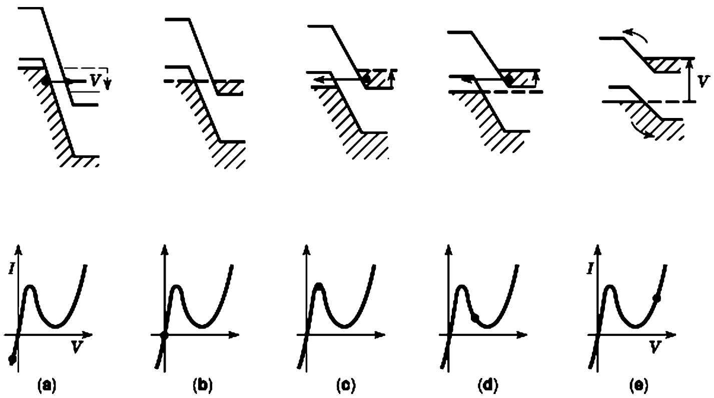

Figure 9. [After (Ref. (3))]. The operation of Esaki diode under different biasing conditions. (a) reverse bias; (b) zero bias; (c) conduction band filled states directly overlapping valence band empty states; (d) bottom of conduction band approaching top of the valence band; and (e) no band overlap with thermal activation of electrons.

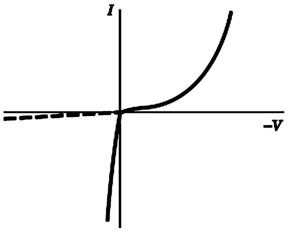

Figure 10. *I*–*V* characteristic of a ‘backward diode’. The band overlap is zero at zero bias. Tunnel breakdown occurs at reverse bias, ‘thermal breakdown’ at forward bias. The dashed curve is for the rectifying-diode characteristic.

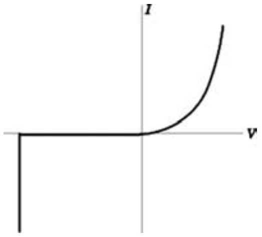

Figure 11. *I*–*V* characteristic of Zener diode. Band overlap is negative at zero bias. Zener breakdown at large reverse bias.

nique also allows one to study the dynamical behavior of the device in response to changes in the terminal voltage. The measurements require the taking of the image of the device with all electrodes at a certain potential (e.g., grounded) and an image of the same device under a changed potential at one of the electrodes, such as the gate in a MESFET, the subtraction of one image from the other, and the calibration of the contrast, in accordance with the voltages at the electrodes. This procedure allows one to visualize the potential distribution inside the device.

QUANTUM THEORY OF LANDAU–ZENER–STÜCKELBERG EFFECT AT AN AVOIDED CROSSING (NONADIABATIC AND ADIABATIC TRANSITIONS)

The quantum level crossing problem involves nonadiabatic processes in quantum mechanics. One is concerned here with systems whose Hamiltonian depends on some variable $\mathcal{K}$, such as the internuclear separation (configuration) between atoms in a molecule, or the wave vector in a semiconductor. It is to be understood that, when an external force or influence is applied, such as an electric field in a semiconductor or a magnetic field in metals, $\mathcal{K}$ becomes time dependent. As a consequence, an explicitly time-dependent situation is obtained, in which the quantum energy levels of the system are brought closer together as time evolves. Such a situation may lead to non-adiabatic transitions between the levels to take place at the region of closest approach or "avoided crossing region."

This level crossing problem is ubiquitous in different contexts, not only in electronics and physics, but also in chemical reaction kinetics and biophysics. Specifically, this problem is also encountered in explaining the conversion of the ν_e neutrinos emitted from the sun into ν_μ neutrinos, when traversing the sun, and thereby rendering the ν_e unobservable. It occurs in numerous situations in atomic and solid-state physics. The LZS problem in nuclear magnetic resonance, laser irradiated atoms, atomic collisions (9), atom-surface scattering, Zener tunneling in dielectric breakdown in solids (10), magnetic breakdown in metals (11), and Zener tunneling in semiconductor p-n junctions (12) are perhaps well-known examples. For other relevant references to these problems see Ref. (13).

Of particular interest are two eigenstate trajectories of the system, which abound in most realistic systems whose paths to first a approximation cross as a function of $\mathcal{K}$, as shown in Fig. 12(a). This situation usually corresponds to the two eigenstates of a 2 × 2 Hamiltonian, when the off-diagonal elements are neglected. However, when the off-diagonal elements (which couple the 'first-order' levels) are taken into account, the degeneracy of the levels at the crossing is avoided. The two levels repel, in accordance with the "no crossing" theorem, as shown in Fig. 12(b). The Landau–Zener–Stückelberg problem can be stated as follows. Consider the system to be initially prepared in state $|1\rangle$ of Fig. 12(b). If $\mathcal{K}$ changes with time (as a result of external influence), traversing the energy levels through the avoided crossing, what is the probability of finding the system in state $|1'\rangle$ and $|2'\rangle$? The transition to $|1'\rangle$ is often called the diabatic transition and that to $|2'\rangle$ is called adiabatic transition. An analytically rigorous solution to this LZS problem can be obtained for the simplest case, where the energy separation between the first-order levels,

$$\hbar\omega(\mathcal{K}) = H_{11}(\mathcal{K}) - H_{22}(\mathcal{K}) \tag{2}$$

varies linearly with $\mathcal{K}$, as indicated in Fig. 12(a) and $\mathcal{K}$ varies linearly with time ($d\mathcal{K}/dt$ is a constant). Then the diabatic transition across the avoided crossing region results in an imaginary phase, acquired by the eigenstate given by

$$\eta = i\pi \frac{|H_{12}|^2}{\hbar\left(\dfrac{d\hbar\omega}{dt}\right)} \tag{3}$$

Therefore, the probability that the system undergoes transition to $|1'\rangle$ is thus given by

$$\mathrm{Prob}_{\mathrm{dia}} = \exp\left\{-2\pi\frac{|H_{12}|^2}{\hbar\left(\dfrac{d\hbar\omega}{dt}\right)}\right\} = \exp\left\{-2\pi\frac{|H_{12}|^2}{\hbar^2\left(\dfrac{d\omega}{d\mathcal{K}}\right)\cdot\dfrac{d\mathcal{K}}{dt}}\right\} \tag{4}$$

This result was first given by Zener (9). The probability depends on the ratio of the square of the energy gap at the avoided crossing to the slew rate, that is, it depends on the energy gap and slew rate, as expected. Hence the name 'inertial effect' (3) is sometimes used. The approximation envisaged here is valid in most physical situations.

It is worthwhile to point out the similarities and differences of the LZS problem with the textbook problem of transmission through potential barriers (*TPB*). The Zener tunneling effect is essentially a two-state time-dependent-parameter problem, as evidenced by the presence of reduced effective mass in the Zener tunneling expressions. In both LZS and TPB, one is dealing with two crossing trajectories as a function of time, in the absence of the off-diagonal matrix elements in LZS or a potential barrier in TPB. The off-diagonal elements lead to avoided crossing (energy gap) in energy-parameter space, whereas the potential barrier is a classically forbidden region in space. Thus, the energy gap is obtained from quantum mechanical treatment of time-independent systems, whereas potential barriers are forbidden position spaces in classical mechanics. The trajectories between LZS and TPB are defined in entirely different spaces; they are defined in the E–$\mathcal{K}$ space in LZS, whereas in TPB, they are defined in position-time (q–t) space at a constant energy. In both LZS and TPB, the trajectory crossings are avoided for an infinitely slow approach in the presence of the off-diagonal matrix elements in LZS or the potential barrier in TPB. In traversing the avoided crossing region, the eigenstates in both LZS and TPB acquire imaginary phases, which determine the transition probability. These phases in both LZS and TPB basically depend on the rate of approach between the two trajectories, compared with the energy gap or height of the potential barrier. Different methods of calculating these phases are usually employed between LZS and TPB.

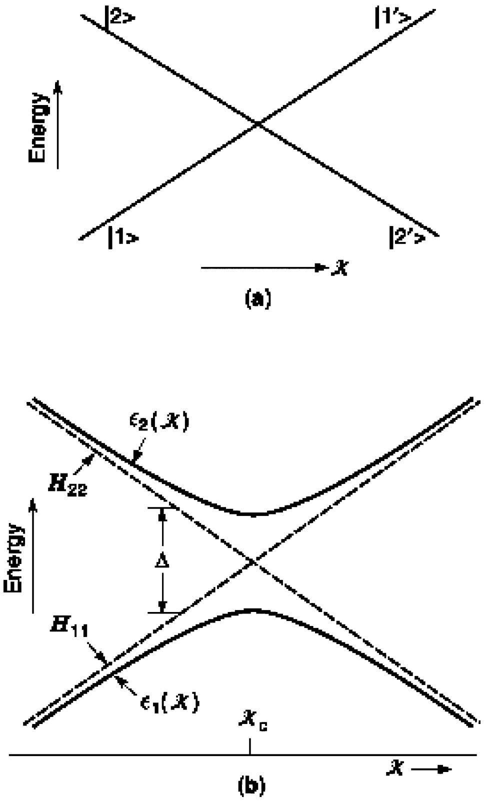

Figure 12. (a) Energy levels as a function of $\mathcal{K}$ in the absence of off-diagonal terms; (b) Solid lines are the energy levels with the off-diagonal elements included, crossing dotted lines correspond to the diagonal elements which may account for diagonal 'renormalization' in the presence of external field. Avoided crossing region results from the off-diagonal elements.

Quantum Mechanics of a Two-State System and the LZS Problem

To derive the diabatic ("inertial") transition probability, consider a two-level system governed by a 2 × 2 Hamiltonian $\mathcal{H}(\mathcal{K})$, which depends explicitly on a parameter $\mathcal{K}$, given by

$$\mathcal{H}(\mathcal{K}) = \begin{pmatrix} H_{11}(\mathcal{K}) & H_{12}(\mathcal{K}) \\ H_{21}(\mathcal{K}) & H_{22}(\mathcal{K}) \end{pmatrix} \equiv \begin{pmatrix} H_{11}(\mathcal{K}) & \frac{\Delta}{2}e^{-i\gamma} \\ \frac{\Delta}{x}e^{i\gamma} & H_{22}(\mathcal{K}) \end{pmatrix} \tag{5}$$

In obtaining an analytically solvable model to the LZS problem, one assumes that the basis states $|1\rangle$ and $|2\rangle$, as well as Δ, are independent of $\mathcal{K}$. Otherwise, one has to resort to numerical methods for solving the LZS problem. In the vicinity of the crossing, $H_{11}(\mathcal{K})$ and $H_{22}(\mathcal{K})$ are taken to vary linearly with $\mathcal{K}$, as shown by the dashed line in Fig. 12(b). In a solid, for example, this 2 × 2 Hamiltonian may describe the $k \cdot p$ Hamiltonian of zinc in the vicinity of the second-zone monster Fermi surface and third-zone needle Fermi surface, used to calculate the magnetic breakdown probability (11). It may also represent a simple $k \cdot p$ Hamiltonian in a semiconductor, used for calculating interband tunneling under high electric fields (14). The adiabatic trajectories, solid lines in Fig. 12(b), can be obtained by diagonalizing $\mathcal{H}(\mathcal{K})$.

For a moment, allow the basis states and Δ to vary with $\mathcal{K}$. The most general transformation $\mathcal{U}$ that diagonalizes a 2 × 2 Hamiltonian $\mathcal{H}(\mathcal{K})$ is given by a 2 × 2 unitary matrix, with determinant one and having three independent parameters, corresponding to the three Eulerian angles of rotation in space (15), given by

$$\mathcal{U} = \begin{pmatrix} e^{-i(\alpha+\lambda)/2}\cos\frac{\beta}{2} & -e^{-i(\alpha-\lambda)/2}\sin\frac{\beta}{2} \\ e^{i(\alpha-\lambda)/2}\sin\frac{\beta}{2} & e^{i(\alpha+\lambda)/2}\cos\frac{\beta}{2} \end{pmatrix} \tag{6}$$

The three independent parameters are chosen such that $\mathcal{U}\mathcal{H}\mathcal{U}^{-1} = \mathcal{U}_{\text{diagonal}}$. The result for the three independent parameters is

$$\begin{aligned} \alpha &= 0 \\ \lambda &= \gamma \\ \tan\beta &= \frac{\Delta}{(H_{22} - H_{11})} \end{aligned} \tag{7}$$

Note that the angle $\beta = \beta(\mathcal{K})$ is a function of $\mathcal{K}$, as well as the angle $\lambda = \gamma(\mathcal{K})$ if the off-diagonal elements are also function of $\mathcal{K}$. The resulting eigenenergies of $\mathcal{H}$ are

$$H'_{11} \equiv \epsilon_1(\mathcal{K}) = \frac{H_{11}+H_{22}}{2} - \frac{1}{2}\sqrt{(H_{11} - H_{22})^2 + \Delta^2} \tag{8}$$

$$H'_{22} \equiv \epsilon_2(\mathcal{K}) = \frac{H_{11}+H_{22}}{2} + \frac{1}{2}\sqrt{(H_{11} - H_{22})^2 + \Delta^2} \tag{9}$$

The corresponding adiabatic eigenfunctions are given by the matrix

$$\begin{pmatrix} |b_1(\mathcal{K})\rangle \\ |b_2(\mathcal{K})\rangle \end{pmatrix} = \begin{pmatrix} e^{-i\gamma/2}\cos\frac{\beta}{2} & -e^{i\gamma/2}\sin\frac{\beta}{2} \\ e^{-i\gamma/2}\sin\frac{\beta}{2} & e^{i\gamma/2}\cos\frac{\beta}{2} \end{pmatrix} \begin{pmatrix} |1\rangle \\ |2\rangle \end{pmatrix} \tag{10}$$

The above solutions are the adiabatic solutions with the parameter $\mathcal{K}$ independent of time. Note that these solutions acquire the characteristics of the diabatic first-order solutions at values of $\mathcal{K}$, such that $|H_{11} - H_{22}| > \Delta$. When $\mathcal{K}$ varies infinitely slowly with time, the trajectory in $E - \mathcal{K}$ space follows the adiabatic solution of Fig. 12(b) by virtue of the Ehrenfest "adiabatic theorem.' However, if $\mathcal{K}$ changes at a finite rate, it is more convenient to reconsider the problem as a time-dependent problem from the beginning. Thus, for $d\mathcal{K}/dt \neq 0$, one solves the time-dependent Schrödinger equation with the 2 × 2 Hamiltonian $\mathcal{H}$,

$$i\hbar\frac{d}{dt}|\psi(\mathcal{K}(t))\rangle = \mathcal{H}|\psi(\mathcal{K}(t))\rangle \tag{11}$$

We expand the solution in terms of time-dependent "diabatic" basis states,

$$\begin{aligned} |\psi(\mathcal{K})\rangle = {} & A_1(\mathcal{K})\exp\left\{\frac{-i}{\hbar}\int_0^t H_{11}(\mathcal{K}(t'))\,dt'\right\}|1\rangle \\ & + A_2(\mathcal{K})\exp\left\{\frac{-i}{\hbar}\int_0^t H_{22}(\mathcal{K}(t'))\,dt'\right\}|2\rangle \end{aligned} \tag{12}$$

Note that the diagonal elements of the 2 × 2 Hamiltonian may account for the 'renormalization' effects in the presence of the field. It will be seen later that the use of diabatic basis states renders a simple expression for the diabatic transition probability in terms of the expansion coefficients. Substituting the assumed solution in Eq. (11) and using the orthogonality of the basis $|1\rangle$ and $|2\rangle$, one obtains coupled first-order differential equations for the time-dependent expansion coefficients,

$$i\hbar\frac{dA_1}{dt} = \frac{\Delta}{2}e^{-i\gamma}\exp\left\{\frac{i}{\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\}A_2 \tag{13}$$

$$i\hbar\frac{dA_2}{dt} = \frac{\Delta}{2}e^{i\gamma}\exp\left\{-\frac{i}{\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\}A_1 \tag{14}$$

The appropriate boundary conditions correspond to the knowledge that initially the system is in state $|b_1(\mathcal{K})\rangle$ or $|1\rangle$, which are equivalent when $\mathcal{K}$ is far from the avoided crossing region, as shown in Fig. 12(b). Assuming that $\mathcal{K}$ varies monotonically with time, one has the boundary conditions

$$A_1(\mathcal{K}(t = -\infty)) = 1 \tag{15}$$

$$A_2(\mathcal{K}(t = -\infty)) = 0 \tag{16}$$

Note that the use of the diabatic basis states yields a very simple expression for the quantum mechanical probability of nonadiabatic (diabatic) transition. This is given by

$$\text{Prob}_{\text{dia}} = |A_1(t = \infty)|^2 = 1 - |A_2(t = \infty)|^2 \tag{17}$$

The above coupled differential equations can be decoupled to yield two analytically solvable equations. Taking the time derivative and eliminating A_2 in Eq. (13) and A_1 in Eq. (14) yields

$$\frac{d^2}{dt^2}A_1(t) + \left(-i\omega - \frac{\dot{\Delta}}{\Delta}\right)\frac{d}{dt}A_1(t) + \frac{\Delta^2}{4\hbar^2}A_1(t) = 0 \tag{18}$$

$$\frac{d^2}{dt^2}A_2(t) + \left(i\omega - \frac{\dot{\Delta}}{\Delta}\right)\frac{d}{dt}A_2(t) + \frac{\Delta^2}{4\hbar^2}A_2(t) = 0 \tag{19}$$

Assuming a constant Δ, the first-order term can be eliminated by using the following transformations:

$$A_1(t) = \exp\left\{\frac{i}{2\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\}a_1(t) \tag{20}$$

$$A_2(t) = \exp\left\{-\frac{i}{2\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\}a_2(t) \tag{21}$$

to yield, after some rearrangement, the following second-order equations, without first-order terms, for the probability amplitudes:

$$\frac{d^2}{dt^2}a_1 + \left(i\dot{\omega} + \frac{\omega^2}{4} + \frac{\Delta^2}{4\hbar^2}\right)a_1 = 0 \tag{22}$$

$$\frac{d^2}{dt^2}a_2 + \left(-i\dot{\omega} + \frac{\omega^2}{4} + \frac{\Delta^2}{4\hbar^2}\right)a_2 = 0 \tag{23}$$

Consistent with the constant Δ, one can assume that the avoided crossing region is so small that one can regard $\hbar\omega = H_{11} - H_{22} \approx \hbar\Theta t$, where Θ is the constant slew rate, which is taken to be greater than zero, as indicated in Fig. 12(b) ($t = 0$ at the crossing point). The resulting equations are of the form of the Weber equation (16),

$$\frac{d^2}{dt^2}a_1 + \left[n_1 + \frac{1}{2} - \frac{z_1^2}{4}\right]a_1 = 0 \tag{24}$$

$$\frac{d^2}{dt^2}a_2 + \left[n_2 + \frac{1}{2} - \frac{z_2^2}{4}\right]a_2 = 0 \tag{25}$$

where

$$n_1 = -\frac{i\Delta^2}{4\hbar^2|\Theta|}$$

$$z_1 = e^{i\pi/4}\sqrt{\Theta}t \equiv \begin{cases} e^{i\pi/4}R, & t > 0 \\ e^{i5\pi/4}R, & t < 0 \end{cases} \tag{26}$$

$$R \equiv |\sqrt{\Theta}t| \tag{27}$$

$$n_2 = \frac{i\Delta^2}{4\hbar^2|\Theta|} \equiv i\chi$$

$$z_2 = e^{-i\pi/4}\sqrt{\Theta}t \equiv \begin{cases} e^{-i\pi/4}R, & t > 0 \\ e^{-i5\pi/4}R, & t < 0 \end{cases} \tag{28}$$

The solutions are given by the Weber functions: $D_n(z)$, $D_n(-z)$, or $D_{-n-1}(\pm iz)$, or their proper combinations. The solution a_2 is chosen to vanish at $t = -\infty$ by virtue of the boundary condition at the remote past: $|A_2(-\infty)| = |a_2(-\infty)| = 0$. The asymptotic expansion for $D_n(z)$ for $|\arg z| < 3\pi/4$ is given by Gradshteyn and Ryzhik (16)

$$\text{Lim}_{z\to\infty}D_n(z) \Rightarrow z^n e^{-\frac{1}{4}z^2} \tag{29}$$

If one chooses the solution as a $a_2 = \Xi D_{-n_2-1}(-iz_2)$, where Ξ is the normalizing constant, then the requirement has been satisfied on the argument of the variable for the simple asymptotic expression to be valid for $t < 0$, since $|\arg(-iz2)_{t<0}| = \pi/4 < 3\pi/4$. Moreover, the boundary condition is also satisfied as shown by

$$\begin{aligned}
\mathrm{Lim}_{t\to-\infty}|a_2(z_2)| &= \mathrm{Lim}_{t\to-\infty}\left|\Xi D_{-n_2-1}(-iz_2)\right| \\
&\Rightarrow \mathrm{Lim}_{t\to-\infty}\left|\Xi(-iz_2)^{-n_2-1}e^{-\frac{1}{4}(-iz_2)^2}\right| \\
&= \mathrm{Lim}_{t\to-\infty}\left|\Xi e^{(i\pi/2)(n_2+1)}z_2^{-n_2-1}e^{\frac{1}{4}z_2^2}\right| \\
&= \mathrm{Lim}_{R\to\infty}\left|\Xi e^{-(i\pi/4)(n_2+1)}e^{-(i/4)R^2}R^{-n_2-1}\right| \\
&\propto \mathrm{Lim}_{R\to\infty}\frac{1}{R} = 0
\end{aligned} \tag{30}$$

The normalizing constant Ξ can be determined from the other boundary condition for $|A_1(-\infty)| = |a_1(-\infty)| = 1$, since A_1 is related to A_2 through Eq. (14). For convenience, this relation between A_1 and A_2 is given here,

$$\begin{aligned}
A_1 &= \frac{i\hbar\frac{d}{dt}A_2}{\frac{\Delta}{2}e^{i\gamma}}\exp\left\{\frac{i}{\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\} \\
&= \left\{\frac{\frac{\hbar\omega(\mathcal{K}(t))}{2}a_2 + i\hbar\frac{d}{dt}a_2}{\frac{\Delta}{2}e^{i\gamma}}\right\}\exp\left\{\frac{i}{2\hbar}\int_0^t \hbar\omega(\mathcal{K}(t'))\,dt'\right\}
\end{aligned} \tag{31}$$

This yields,

$$\begin{aligned}
&\mathrm{Lim}_{t\to-\infty}|A_1| \\
&\quad = 1 \\
&\quad = \mathrm{Lim}_{t\to-\infty}\left|\frac{\frac{\hbar\omega(\mathcal{K}(t))a_2}{2} + i\hbar\frac{d}{dt}a_2}{\frac{\Delta}{2}e^{i\gamma}}\right| \\
&\quad \Rightarrow \mathrm{Lim}_{R\to\infty}\left|\frac{\frac{R}{2}\left[\Xi e^{-(i\pi/4)(n_2+1)}e^{-(i/4)R^2}R^{-n_2-1}\right] + i\frac{d}{dR}\left[\Xi e^{-(i\pi/4)(n_2+1)}e^{-(i/4)R^2}R^{-n_2-1}\right]}{\frac{\Delta}{2\hbar\sqrt{|\Theta|}}e^{i\gamma}}\right| \\
&\quad = \Xi\left(\frac{e^{(\pi/4)\chi}}{\sqrt{\chi}}\right) = 1
\end{aligned} \tag{32}$$

where

$$\chi = \left(\frac{\Delta}{2\hbar\sqrt{\Theta}}\right)^2 \tag{33}$$

Therefore, the normalizing constant is given by

$$\Xi = \sqrt{\chi}e^{-\pi\chi/4} \tag{34}$$

With $a_2 = \sqrt{\chi}e^{-\pi\chi/4}D_{-n_2-1}(-iz_2)$, calculate the asymptotic value of a_2 at $t \to \infty$, to determine the adiabatic and nonadiabatic transition probabilities. For $t > 0$, one has $-\pi/4 > (\arg(-iz_2) = -3\pi/4) > -5\pi/4$. The corresponding asymptotic expansion is given by Gradsteyn and Ryzhik (16),

$$\begin{aligned}
&\mathrm{Lim}_{t\to\infty}\Xi D_{-n_2-1}(-iz_2) \\
&\qquad \Rightarrow \Xi\left\{e^{-1/4(-iz_2)^2}(-iz_2)^{-n_2-1}\right. \\
&\qquad\qquad \left. - \frac{\sqrt{2\pi}}{\Gamma(1+n_2)}e^{i\pi(1+n_2)}e^{\frac{1}{4}(iz_2)^2}(-iz_2)^{n_2}\right\}
\end{aligned} \tag{35}$$

Substituting the expression for z_2 for $t > 0$, one obtains

$$\begin{aligned}
&\mathrm{Lim}_{t\to\infty}\Xi D_{-n_2-1}(-iz_2) \\
&\quad \Rightarrow \mathrm{Lim}_{R\to\infty}\Xi\left\{\begin{array}{c} e^{-iR^2/4}e^{3\pi i(1+n_2)/4}R^{-n_2-1} \\ +\frac{\sqrt{2\pi}}{\Gamma(1+n_2)}e^{i\pi n_2}e^{iR^2/4}e^{-3\pi n_2/4}R^{n_2}\end{array}\right\} \\
&\quad = \Xi\frac{\sqrt{2\pi}}{\Gamma(1+n_2)}e^{iR^2/4}e^{i\pi n_2/4}R^{n_2} \\
&\quad = \frac{\sqrt{2\pi}}{\Gamma(1+n_2)}e^{iR^2/4}\sqrt{\chi}e^{-\pi\chi/2}R^{n_2}
\end{aligned} \tag{36}$$

Therefore, the probability for an adiabatic transition is given by

$$\begin{aligned}
|a_2|^2 &= \frac{2\pi\chi e^{-\pi\chi}}{\Gamma(1+i\chi)\Gamma(1-i\chi)} = \frac{2\pi\chi e^{-\pi\chi}}{i\chi\Gamma(i\chi)\Gamma(1-i\chi)} \\
&= 2e^{-\pi\chi}\sinh(\pi\chi) = 1 - e^{-2\pi\chi}
\end{aligned} \tag{37}$$

The last line made use of the identity relations for gamma functions. This result yields the exact probability for diabatic transition given by

$$\begin{aligned}
\mathrm{Prob}_{\mathrm{dia}} &= |a_1|^2 = \exp\{-2\pi\chi\} = \exp\left\{-2\pi\frac{\Delta^2}{4\hbar^2\Theta}\right\} \\
&= \exp\left\{-2\pi\frac{(\Delta/2)^2}{\left|\hbar\left(\frac{d}{d\mathcal{K}}\hbar\omega\right)\cdot\frac{d\mathcal{K}}{dt}\right|}\right\} \\
&= \exp\left\{-2\pi\frac{(\Delta/2)^2}{\left|\hbar v_r\cdot\left(-\frac{e}{c}\frac{d\mathcal{A}(t)}{dt}\right)\right|}\right\}
\end{aligned} \tag{38}$$

where, in the last line v_r is the "relative velocity" and $\mathcal{K}(t)$ is written in the form

$$\mathcal{K}(t) = \mathcal{K}_o - \frac{e}{c}\mathcal{A}(t) \tag{39}$$

$\mathcal{A}(t)$ is identified as the 'generalized' gauge field or vector potential of the external influence or "force," with e the elementary charge and c the speed of light in a vacuum. The above time dependence of $\mathcal{K}(t)$ is well known in solids under external electromagnetic fields. In Born–Oppenheimer studies of molecules, however, $\mathcal{K}(t)$ has an entirely different meaning.

APPLICATIONS TO ELECTRONS IN SOLIDS UNDER EXTERNAL ELECTROMAGNETIC FIELDS

Zener Tunneling in Narrow-Gap Direct Semiconductors

The two-band model for a narrow-gap direct semiconductor given by Kane and Blount (14) may be simply written as

$$\mathcal{H} = \begin{pmatrix} s\cdot\mathcal{K} & \frac{\Delta}{2} \\ \frac{\Delta}{2} & -s\cdot\mathcal{K}\end{pmatrix} \tag{40}$$

The adiabatic eigenvalues are

$$\epsilon_1(\mathcal{K}) = \sqrt{(s\cdot\mathcal{K})^2 + \frac{\Delta^2}{4}} \tag{41}$$

$$\epsilon_2(\mathcal{K}) = -\sqrt{(s\cdot\mathcal{K})^2 + \frac{\Delta^2}{4}} \tag{42}$$

In the presence of a dc electric field, F, the time dependence of $\mathcal{K}(t)$ is written as $\mathcal{K} = \mathcal{K}_o + eFt$, with $\mathcal{A}(t) = -cFt$. Therefore, $d\mathcal{K}/dt = eF$. The Zener tunneling probability can thus be immediately written as

$$\mathrm{Prob}_{\mathrm{dia}} = \exp\left\{-2\pi\frac{(\Delta/2)^2}{2|\hbar s\cdot eF|}\right\} \tag{43}$$

One can express the velocity s in terms of the energy gap, $E_G = \Delta$, and reduced effective mass, $m^{*,r}{}_{ii}$, by using the relations

$$m^*_{ii}(1) = \hbar^2(\Delta/2)s_i^2 \tag{44}$$

$$m^*_{ii}(2) = -\hbar^2(\Delta/2)/s_i^2 \tag{45}$$

$$m^{*,r}_{ii} = |m^*_{ii}|/2 \tag{46}$$

where the subscript i indicates the direction of the electric field. Then the exact Zener tunneling probability can be recast as

$$\mathrm{Prob}_{\mathrm{dia}} = \exp\left\{-\pi\frac{\sqrt{m^{*,r}_{ii}}(E_G)^{3/2}}{2\hbar^2|eF|}\right\} \tag{47}$$

This was also obtained by Kane and Blount (14) by a different method. This is an exact result based on the approximate Hamiltonian of a narrow-gap semiconductor. The reduced effective mass is generally present in Zener tunneling expressions, which do not occur if the problem is viewed as electron tunneling through a triangular potential barrier; compare with the exponential factor of Eq. (1).

Magnetic Breakdown in Zinc Metals

In zinc, the magnetic analog of Zener tunneling occurs between the second zone "monster" Fermi surface and the third zone "needle" Fermi surface at magnetic fields of only a few kilogauss (Fig. 13). The second zone "monster" Fermi surface and third zone "needle" Fermi surface of zinc can be described by the use of a $k\cdot p$ Hamiltonian of the form (11)

$$\mathcal{H} = \begin{pmatrix} \nu_x\mathcal{K}_x + \nu_y\mathcal{K}_y + E_F & \dfrac{\Delta}{2} \\ \dfrac{\Delta}{2} & \nu_x\mathcal{K}_x - \nu_y\mathcal{K}_y + E_F \end{pmatrix} \tag{48}$$

where $\mathcal{K}_z$ has been set equal to zero, since it has no effect on the result. In a magnetic field the time dependence of $\mathcal{K}(t)$ is tied to the time dependence of the electron coordinate via the expression for the vector potential,

$$\mathcal{K}(t) = \mathcal{K}_o + \frac{|e|}{c}\mathcal{A}(r) = \mathcal{K}_o + \frac{|e|}{c}(0, Bx, 0) \tag{49}$$

where the vector potential is given in the Landau gauge: $\mathcal{A}(r) = B(0, x, 0)$, B is the magnetic field in the z-direction. Therefore, one obtains

$$\frac{d\mathcal{K}}{dt} = \left(0, \frac{e}{c}B\nu_x, 0\right) \tag{50}$$

Applying Eq. (38), the Zener tunneling probability is thus given as

$$\begin{aligned}\mathrm{Prob}_{\mathrm{dia}} &= \exp\left\{-2\pi\frac{(\Delta/2)^2}{\left|(2\hbar\nu_y)\left(\frac{e}{c}B\nu_x\right)\right|}\right\} \\ &= \exp\left\{-\pi c\frac{\Delta^2}{4B|e\hbar\nu_y\nu_x|}\right\} \equiv \exp\left\{-\frac{B_o}{B}\right\}\end{aligned} \tag{51}$$

where

$$B_o = \pi c\frac{\Delta^2}{4|e\hbar\nu_y\nu_x|} \tag{52}$$

The expression of Eq. (51) is often used in studies of magnetic breakdown in metals in external magnetic fields [see references in (11)].

CORRECTIONS TO LZS TUNNELING IN *p–n* JUNCTION DIODES

The assumption that $\hbar\omega(\mathcal{K})$ varies linearly with time is not satisfied for p–n junction of Esaki or Zener diodes. In p–n diodes $\hbar\omega(\mathcal{K})$ sweeps through the avoided crossing region at a constant rate but is independent of time outside some interval τ, which is the time spent in the avoided crossing region. Figure 7 shows the situation for Zener, backward, or Esaki diodes. As shown in the figure, the energy difference, $\hbar\omega(\mathcal{K}(t))$, only varies linearly as it approaches the p–n junction, but is effectively constant outside some interval $\delta\mathcal{K}$, corresponding to the time interval $\tau = \tau_c + \tau_\nu$. A good approximation to "pulse" behavior of $\hbar\omega(\mathcal{K}(t))$ is the one proposed by Rubbmark et al. (17),

$$\begin{aligned}\hbar\omega(\mathcal{K}(t)) &= \Theta t\left\{\frac{1}{1+\exp\left(-\frac{4t}{\tau}\right)} - \frac{1}{2}\right\} \\ &\Rightarrow \begin{cases} -\frac{1}{2}\Theta\tau & t\to-\infty \\ \Theta t & |t|\ll\tau \\ \frac{1}{2}\Theta\tau & t\to\infty \end{cases}\end{aligned} \tag{53}$$

For arbitrary variation of $\hbar\omega(\mathcal{K}(t))$, the equation has to be solved numerically (18, 19).

The diabatic transition probability was found to depend on three parameters: $\chi = \Delta^2/4\hbar|\Theta|$, $s = (\Delta/2)(\hbar/\tau)^{-1}$, and $d = s/\chi$. For $d\to\infty$, the LZS result for the diabatic transition probability is recovered. However, for $d = 1$ this is almost independent of χ. As a function of s, the Zener tunneling probability exhibits a quasi resonant behavior, exceeding the LZS value at resonance around $s = 1$. For $d > 10$, which is valid in narrow-gap semiconductors, the diabatic transition probability generally agrees with the Zener result.

NOVEL HIGH-FREQUENCY SOURCES

The p–n junction serves as the basic building block of bipolar semiconductor thyristors, transistors, and diodes. The recent development of heterostructure technology has brought the 'hetero' junction as the basic building block of

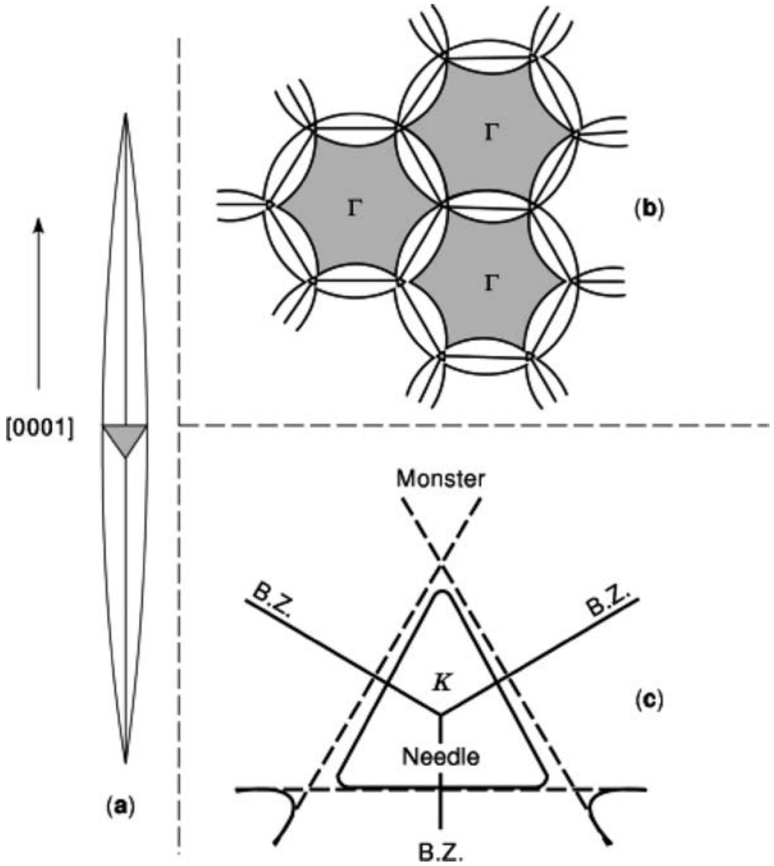

Figure 13. [After (Ref. (11))]. (a) The third zone 'needle' of Fermi surface of zinc; (b) The cross section of the Zinc Fermi surface showing the second zone 'monster' and third zone 'needle'; (c) The three avoided crossing regions of about the K point of the Brillouin zone.

most advanced, high-speed, and high-frequency devices for electronic and optoelectronic applications. A heterojunction is an interface within the semiconductor material, across which the chemical composition changes, in contrast to the *p–n* junction, where only the dopants change. The junction between GaAs and AlGaAs, the junction between AlGaSb and InAs, and the junction between Si and GeSi are some of the examples. A host of different band-edge alignments (20) have generated different Zener effect or interband tunneling devices (21). Heterostructure devices offer ultra-high-frequency sources in the THz range, as found from independent numerical quantum transport simulations (18, 19) of conventional resonant tunneling devices (*RTD*), which are closely spaced double potential-barrier heterostructures. This was also indirectly inferred from experiments (22). This was further discussed in the context of equivalent-circuit models, where the concept of a quantum inductance was introduced (23, 24). The autonomous device oscillation occurs when a RTD is biased in the negative-differential-resistance (*NDR*) region, just after the resonant current peak.

An Interband Tunnel High-Frequency Source

The new oscillation addressed here occurs before the resonant current peak, and is based on interband tunneling in RTD with staggered band-gap alignment. A staggered band-edge alignment can be realized by using InAs/AlSb heterojunctions, as in Fig. 14(a). In a simple implementation of a novel interband tunnel high-frequency source, a deeper quantum-well-for-holes is desirable, which can support a localized hole state; this is obtained by using InAs/AlGaSb heterojunctions; see Fig. 14(b). Unless otherwise specified, quantum well refers to the conduction band edge and conduction-band electrons.

The new mechanism of modulating the resonant energy level in the quantum well with respect to the energy distribution of supply electrons from the emitter can simply be described through the oscillatory build-up and decay of the polarization pairing between electrons in the quantum well and trapped holes in the barrier. This modulation is likely to be more useful, since it is controlled by trapped holes (similar to base charges of a bipolar transistor). For a well-designed emitter and a sufficiently sharp energy level, the modulation of the energy level of the quantum well has a large 'transconductance' as the current peak of RTD is approached, where the trapping of holes in the barrier also occurs. Thus, autonomous control of a significant current by an interband process can be realized for the first time.

The polarization pair is referred to as a *duon,* since this Coulomb-correlated e–h pair, in contrast to an *exciton,* can only be transported in the transverse direction. In what follows, we introduce the physics of the *duon* dynamics. The limit cycle solution leads to an oscillatory voltage drop between the quantum well and the barrier. Since common

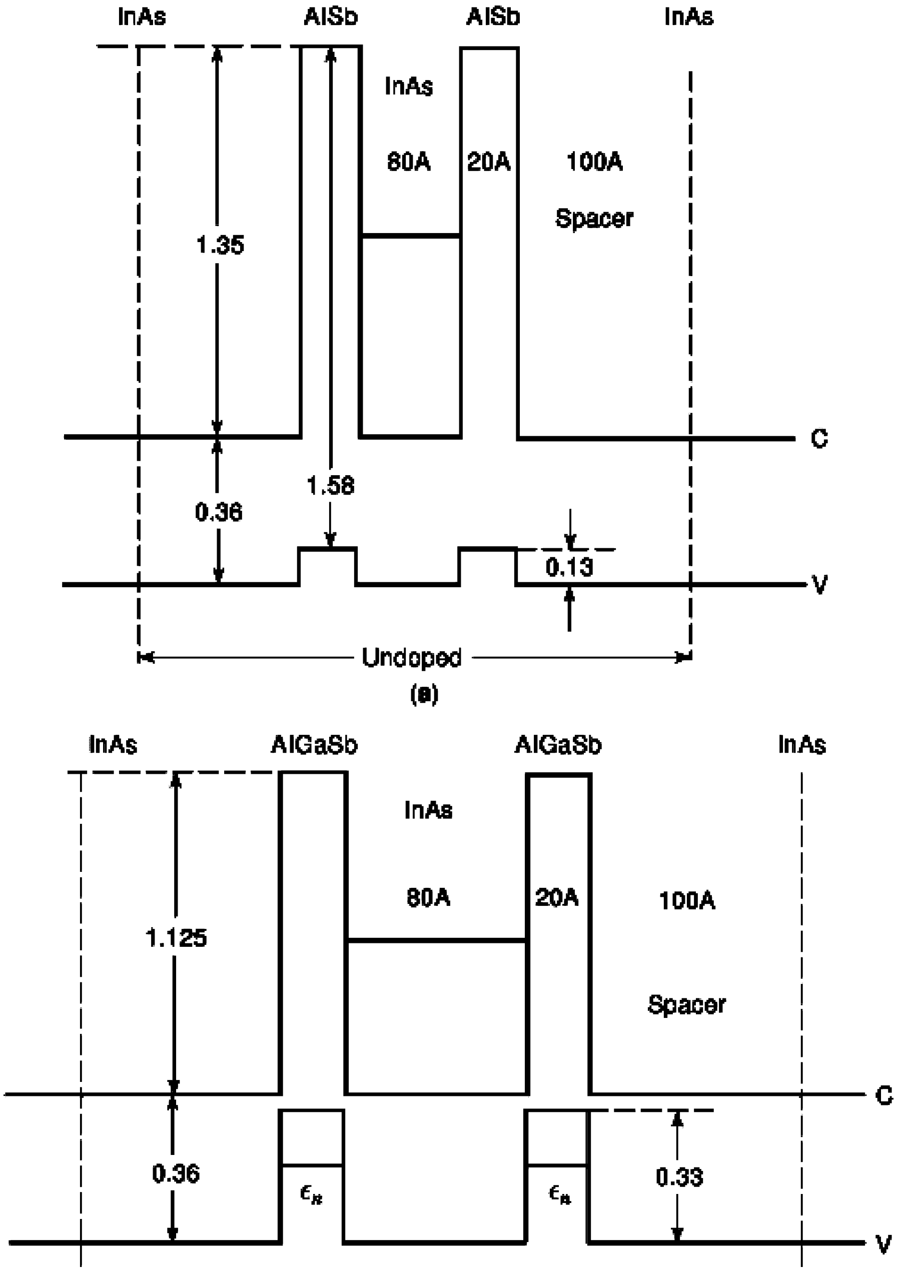

Figure 14. (a) Energy band edge alignment of RTD using InAs/AlSb heterojunction. The shallow potential well in the valence band cannot support a localized hole state; (b) Energy band edge alignment of RTD using InAs/AlGaSb heterojunction. The deeper potential well in the valence band can support localized hole state. Approximated band-edge offsets are indicated in electron volts.

experimental techniques are incapable of investigating terahertz oscillations, the current-voltage *(I–V)* characteristic is also calculated with results in agreement with the experimental *I–V* characteristics (25).

Device Operation

When the localized valence-band electrons, confined in the *AlGaSb* barrier, see the available states in the drain region (refer to Fig. 15), these valence-band electrons tunnel to the drain leaving behind quantized holes. The emerging conduction-band electrons deposit at the spacer layer which is then acted on by the field of the depletion region and eventually recombine at the contact. In effect, this process initiates the "polarization" between the barrier and the quantum well, thus establishing a high-field domain in this region. The result is a consequent redistribution of the voltage drop across the device. The Zener transition is initiated when the discrete level of the right barrier, ε_n, matches with unoccupied conduction-band states in the drain, this first occurs at $(k_z^D)^2 \geq (k_F^D)^2$ in Fig. 15, with the ">" sign holding for indirect band-gap Zener tunneling.

The drain serves as a sink due to unoccupied states above k_F^D that could satisfy the conservation of transverse crystal momentum associated with the discrete "longitudinal energy,"ε_n, of the right barrier. Figure 15 pertains to

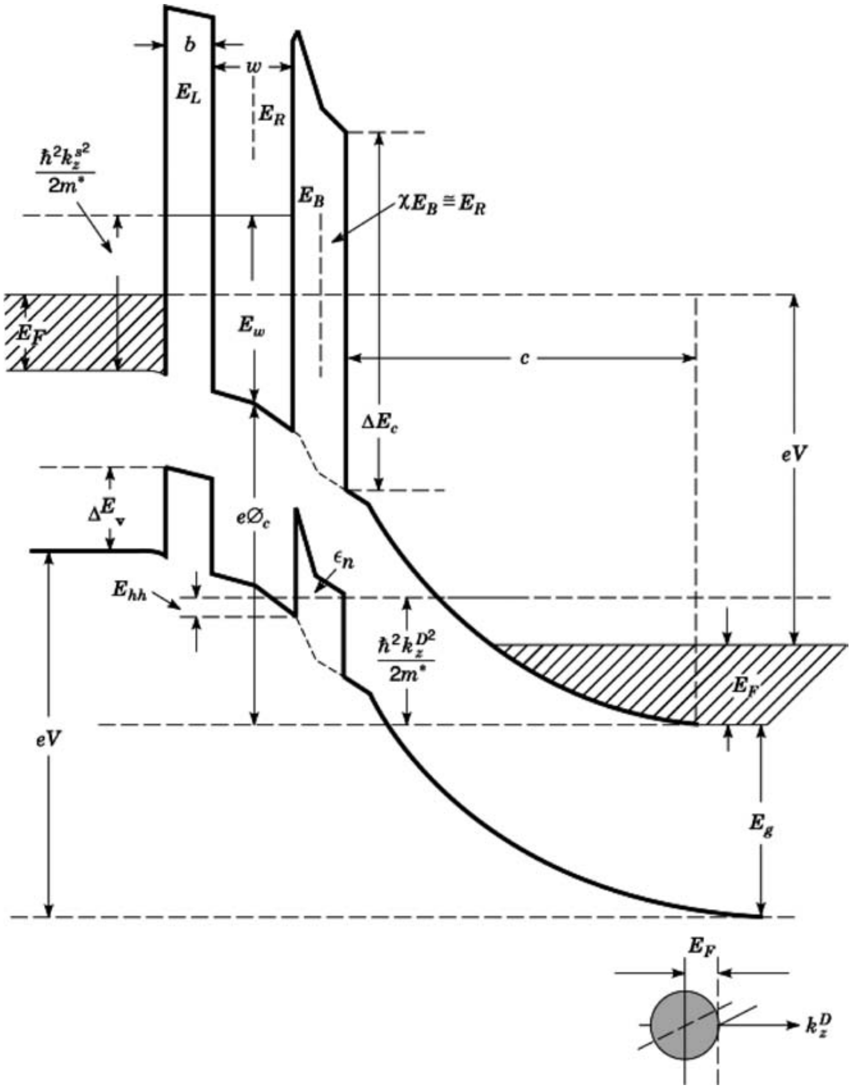

Figure 15. Schematic average EBE profile showing the various quantities used in the calculations. The shaded region in the lower-right-hand corner indicates the occupied transverse and longitudinal momentum states in the drain.

motion in the z direction and therefore E_F, is the Fermi energy for motion in the z direction, with "in-plane energy" (or transverse crystal momentum) and higher values of $(k_F^D)^2$ being unoccupied in the right contact, as schematically shown in the inset of Fig. 15.

The time-dependent dynamics of the hole charging and discharging that follows is dictated by the self-consistency of the potential. At higher voltage bias, this is described by Fig. 16. As hole charging occurs, Fig. 16(a), the polarization between the barrier and the quantum well induced by the trapped hole charge [Fig. 16(b)] creates a high-field domain tending to lower the residual potential drop between the drain contact and the right barrier. This process results in reducing the Zener tunneling probability of the valence-band electrons from the barrier to the right contact. Owing to self consistency of the potential, further polarization leads to the "switching" of the intravalence band tunneling of the trapped holes from the barrier towards the quantum well region [Fig. 16(c)]. When the situation shown in Fig. 16(c) is reached, the hole intravalence band tunneling probability is maximum and the Zener transition probability is minimum. The onset of two other possible mechanisms for hole discharging may also occur at this point, namely. thermal activation of the valence electrons in the continuum to recombine with localized holes, or loss of any bound hole states in the barrier.

The "bound-hole leakage" is by tunneling through a triangular potential barrier, which would likely have a smaller barrier height than for that of Zener tunneling if viewed as tunneling through potential barriers (3). Any of the hole-leakage processes mentioned above will immediately restore the high field between the barrier edge and the right contact. This redistribution of the voltage drop accompanying the discharging process in turn "switches off" the intraband tunneling probability, or reestablishes the localized state in the barrier if this was lost when the situation of Fig. 16(c) was reached, while "switching on" the Zener transition of valence-band electrons towards the right contact, recharging the barrier. The situation shown in Fig. 16(a) is revisited, after which the process repeats.

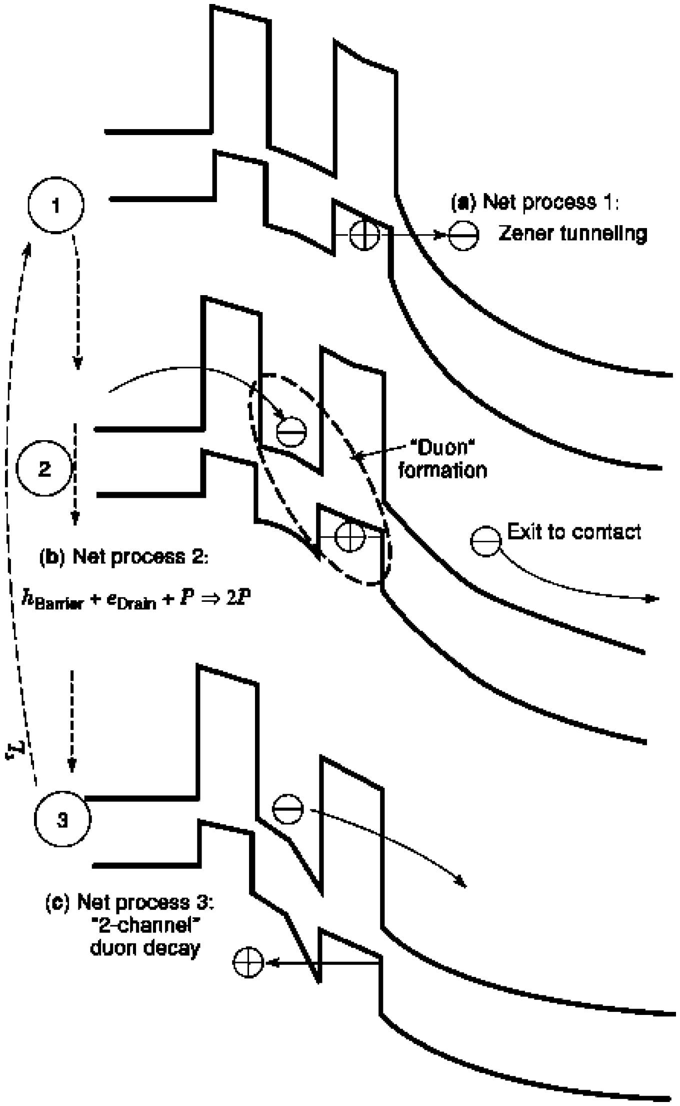

Figure 16. Schematic diagram showing the mechanism of the oscillation of trapped hole charge in the barrier. (a) e–h generation by Zener tunneling; (b) duon generation is through an *autocatalytic* process; (c) three mechanisms for hole discharging are mentioned in the text.

Therefore, oscillations of the hole charging of the *AlGaSb* barrier can occur in the THz range, by virtue of the nanometric dimensional features of the device.

It is important to realize that the whole dynamical process limits the average amount of hole charge that can be trapped in the barrier as a function of drain bias. The direct interband recombination process is not considered since it cannot compete with the conduction-band electron tunneling process. There is not enough time for direct interband recombination to take place, since the velocity of the conduction electrons at the barrier is quite large owing to the small probability of being inside the barrier region.

The discharging and charging processes are in general governed by the two characteristic times. These are the polarization-charge build-up time, τ_B, and the charge-leakage time, τ_L. The build-up time τ_B is through Zener transition and consequent charging of quantum well with conduction electrons. The charge-leakage time τ_L is through intravalence band tunneling, thermal activation of valence electrons from the continuum, and loss of bound-hole state, coupled with the decay of conduction electrons in the quantum well by virtue of the consequent rise of the resonant energy level following the discharge of trapped holes. In general, if $\tau_B > \tau_L$, then oscillatory behavior will occur, the charging process will always be lagging behind the discharging process and oscillations will result. This criterion holds true in conventional RTD and single-electron devices (26). It is estimated that $\tau_B > \tau_L$, by virtue of several possible fast hole-discharge channels mentioned above, and the likelihood of small triangular-potential-barrier height whenever a bound state still exists when the situation of Fig. 16(c) is reached. The oscillation is expected to occur in the THz range. Analogous oscillatory trapped-charge behaviors are not uncommon in nanostructure systems (27,23,26).

Polarization Pairing Dynamics

To describe the physical mechanism described above using quantum transport theory of electrons and holes, including Zener tunneling and pairing dynamics, is a very complex task requiring large computational resources which will make use of the general and fundamental quantum transport equations recently formulated by Buot (28–31).

Our task here is to give a simple phenomenological model for the three physical processes depicted in detail in Fig. 16. A much more rigorous derivation of the physical model is given in Ref. (31). Here, we follow closely the phenomenological modeling of Buot and Krowne (32).

Mathematically, we need to model the respective rate equations for the following processes: (i) the creation of conduction electrons and trapped holes by Zener tunneling, Fig. 16(a), whose rate decreases with the generation of polarization between the barrier and quantum well, (ii) stimulated generation of barrier-well polarization. Fig. 16(b), and (iii) the decay of duons is depicted in Fig. 16(c). The process depicted in Fig. 16(b) generally involves the transport of conduction electrons from the emitter to the quantum well induced by the polarization effect of the preceding hole charging of the barrier, coupled with a succeeding e-h generation by Zener tunneling. In process (2), the Zener-generated electron flows to the metallic contact of the drain, and goes out of the picture; this is substituted by the tunneled conduction electron from the emitter to the quantum well to form a "duon" or polarization pair with the hole in the barrier. We observe here an "autocatalytic" or positive feedback of this process, since duon generation is induced by existing duons, in the presence of e-h through Zener tunneling. The hole process depicted in Fig. 16(c) can take place through one of several mechanisms or their combinations. mentioned before. These are coupled with the decay of conduction electrons in the quantum well by virtue of the consequent rise (note that the device is operating before the resonant-current peak) of the resonant energy level as the number of duons or of a high-field domain between the quantum well and barrier decreases.

Let G be the maximum rate of e-h generation by Zener tunneling from the barrier valence band to the drain conduction band for a given bias. This constant rate G is a function only of the applied voltage, or more appropriately of the biasing field at the depletion layer of the drain region (refer to Fig. 15). It is therefore our measure of the applied bias at the drain contact for a given device structure. It is clear that the maximum rate of generation of e-h by Zener tunneling occurs in the absence of any hole charge trapped in the barrier, and hence in the absence of duons which we denote by P. The reason for this is that the highest field in the depletion layer for a given bias occurs in the absence of duons, as indicated in Fig. 16(a).

The effective generation rate for e-h by Zener tunneling will of course decrease with polarization since the region between the quantum well and the barrier becomes a growing high-field region at the expense of the voltage drop in the depletion region. We expect this rate of decrease to be proportional to the rate of production of duons. On the other hand, the generation rate of the duons autocatalytically depends on the concentration of existing duons. The generation of duons involves three interacting components, namely, (a) trapped holes in the barrier, (b) generated conduction electrons, and (c) existing duons to stimulate the net transfer of conduction electron from the emitter to the quantum well to be paired with the trapped hole in the barrier to form more duons. One can view this autocatalytic "pairing" of conduction electron in the quantum well and trapped hole charge in the barrier as a stimulated transformation of the e-h pair generated by Zener tunneling into a duon with the electron created by Zener tunneling recombining at the drain contact, leaving only the "polarization pair" (duon) as depicted in Fig. 16(b).

The duon generation rate with three interacting components can thus be expressed as

$$\left.\begin{array}{c} \textit{Stimulated duon}(P)\textit{generation rate} \\ (h_{barrier} + e_{drain} + P \Rightarrow 2P) \end{array}\right\} = \tilde{\Delta}(\mathcal{N}_B)^2 P \qquad (54)$$

where $\mathcal{N}_B$ is the number of "unpaired" holes which is equal to the number of "exiting" conduction electrons created by Zener tunneling. $\tilde{\Delta}$ is the rate parameter [per number of electrons and per number of holes produced by Zener tunneling as indicated in Fig. 16(b)].

We can now write the "effective" generation rate of unpaired trapped holes in the barrier as

$$\frac{\partial \mathcal{N}_B}{\partial t} = G - \tilde{\Delta}(\mathcal{N}_B)^2 P \qquad (55)$$

Note that the total concentration of trapped holes in the barrier, Q_B, at any time is given by

$$Q_B = \mathcal{N}_B + P \qquad (56)$$

where P is given as the duon concentration.

To obtain the rate equation for P, we need to formulate the process in Fig. 16(c) describing the decay of the high-field domain between the quantum well and barrier. This decay rate for P is expected to saturate to a constant rate for very large P. Let N be the total number of matching states for the holes in the barrier to transition to. Let N_P be the number of matching states in the valence band no longer available by virtue of holes already transitioning to these states. The production rate of N_P is proportional to the product of the available number of matching states and P. Let λ be this proportionality constant. And let γ the decay rate of N_P by virtue of recombination of holes and screening: electrons in the valence band. Then, we can write the rate equation for N_P as

$$\frac{\partial N_P}{\partial t} = \lambda(N - N_P)P - \gamma N_P \qquad (57)$$

where the first term is also the decay rate of P. The process described by the last term of Eq. (57), i.e., recombination of holes and screening electrons in the valence band, is the fastest process in the problem. N_P is therefore expected to relax much faster than P and $\mathcal{N}_B$. Thus by adiabatic elimination of fast variables, we can let $\frac{\partial N_P}{\partial t} \Rightarrow 0$. Then the decay rate of P is equal to γN_P. We obtain

$$N - N_P = \frac{N}{1 + \frac{\lambda P}{\gamma}} \qquad (58)$$

Upon substituting the expression of Eq. (58) in the first term of Eq. (57), we obtain the decay rate of P given by

$$\gamma N_P = \frac{\lambda NP}{(1+\frac{\lambda P}{\gamma})}$$

Likewise, the decay rate via tunneling of conduction electrons from the quantum well to the drain is limited, through self-consistency, by the limit set on the trapped hole decay process. Thus, we can express the decay rate of the duon concentration as

$$Duon\,(P)\,decay\,rate = \frac{\alpha P}{1+\beta P} \tag{59}$$

where $\frac{\alpha}{\beta} = \gamma N, \frac{1}{\beta}$ is proportional to the sum of available states in the valence band of the quantum-well region and the states participating in thermal recombination, otherwise, it represents the actual number of hole states in the barrier in the case of the loss of bound hole state. Equation (58) is similar to the Michaelis-Menten decay law in chemical kinetics (37). The parameter $\alpha = \lambda N$ is the decay rate constant and $\frac{\alpha}{\beta}$ is the value of the saturated decay rate of duons. Therefore, we can now write the rate equation for P as

$$\frac{\partial P}{\partial t} = \tilde{\Delta}(\mathcal{N}_B)^2 P - \frac{\alpha P}{1+\beta P} \tag{60}$$

As seen in Eq. (62) below, the physical situation corresponds to $\frac{\alpha}{\beta} > G$. This means that to sustain this two-dimensional dynamical nonlinear system with a stimulated intermediate process, the maximum e-h generation rate by Zener tunneling for a given bias must be less than the maximum possible (saturation) value of the decay rate of duons. Therefore, in the formation of the high-field domain, the maximum discharging rate is larger than the maximum build-up rate as conjectured before. Indeed, we can estimate that $\frac{1}{\tau_B} \approx G$ and $\frac{1}{\tau_L} \approx \frac{\alpha}{\beta}$. Therefore, $\frac{\alpha}{\beta} > G$ implies that $\tau_B > \tau_L$.

Equations (55) and (60) describe a two-parameter ($\tilde{\Delta}$, G) and two-dimensional ($\mathcal{N}_B$, P) dynamical system. The intermediate process depicted in Fig. 16(b) is a catalytic process whose form occurs in several other systems. Similar coupled equations were considered by Pimpale *et al.* (38) for analyzing the stimulated production of excitons in the presence of recombination centers in optically excited semiconductors. We give the details of our calculations in what follows.

Stationary Solution and Stability Analysis. The stationary solution to the coupled rate equations, Eqs. (55) and (60), is given by

$$G = \tilde{\Delta}(\mathcal{N}_B)^2 P = \frac{\alpha P}{1+\beta p} \tag{61}$$

It follows from Eq. (55) that the effective Zener tunneling has stopped. The total stationary trapped hole concentration, Q_B, is thus given by

$$Q_B = \mathcal{N}_B + P = \left(\frac{\alpha - \beta G}{\Delta^{\sim}}\right)^{1/2} + \frac{G}{\alpha - \beta G} \tag{62}$$

It is important to point out that the more accurate average value of the total trapped hole charge under a limit cycle oscillation is shown later to be a slowly decreasing function of bias. This is an important factor in comparing our theory with the existing experiment (35).

For a fixed bias, the total terminal current of the RTD is the sum of polarization current, due to the time varying polarization, and the RTD conduction electron resonant-tunneling current component. With Q_B and P constant for a given bias, Eq. (61) states that the production of duons is balanced by the decay of duons. This means that the corresponding polarization current is also a constant, equal to zero. Since Q_B and P are constants, the duon production rate is via the transfer of conduction electrons from the emitter to the quantum well and the duon decay rate is via transfer of conduction electrons from the quantum well to the drain. At steady state these two conduction-electron-mediated decay and generation processes are balanced resulting in net steady-state resonant-tunneling current across the double-barrier structure.

It should be emphasized that it is the effective e-h generation rate that describes the actual Zener tunneling process, to account for self-consistency of the potential in the presence of duons. This effective Zener tunneling rate, $G_{eff}(\mathcal{N}_B, P) = G - \tilde{\Delta}(\mathcal{N}_B)^2 P$, reduces to zero at steady state. In other words, the steady-state operation no longer involves interband processes, as schematically shown in Fig. 18. However, in dynamical situation, $\frac{\partial \mathcal{N}_B}{\partial t} = 0$, while the duon generation may still be nonzero, which may lead to an elliptical limit-cycle behavior.

For the following stability and nonlinear analyses, it is convenient to simplify the fundamental rate equations and write them in terms of dimensionless variables as

$$\frac{\partial}{\partial \tau}\Pi = \Delta \mathcal{Q}^2 \Pi - \frac{\Pi}{1+\Pi} \tag{63}$$

$$\frac{\partial}{\partial \tau}\mathcal{Q} = \mathcal{G} - \Delta \mathcal{Q}^2 \Pi \tag{64}$$

where

$$\Pi = \beta P$$

$$\mathcal{Q} = \beta \mathcal{N}_B$$

$$\Delta = \frac{\tilde{\Delta}}{\alpha}(\frac{1}{\beta})^2 \tag{65}$$

$$\mathcal{G} = \frac{G}{\alpha/\beta} \tag{66}$$

$$\tau = \alpha t$$

Since, in general, the duon formation also involves a higher-order process involving Zener tunneling from the valence band of the barrier to the conduction band in the drain coupled with tunneling of electron from the emitter to the quantum well to form a polarization pair, we expect the frequency of duon formation to be much less than the frequency for duon decay which may consist of several parallel single-event channels. Therefore, Δ must be considerably less than one. Similarly, the dimensionless pa-

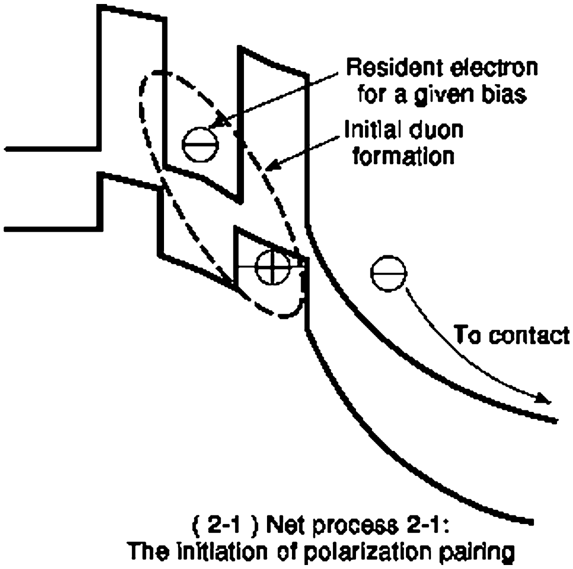
Resident electron for a given bias
Initial duon formation
To contact
(2-1) Net process 2-1:
The initiation of polarization pairing

Figure 17. More efficient generation of initial polarization pairs or duons is due to the initial resident excess electron in the quantum well for a given applied bias.

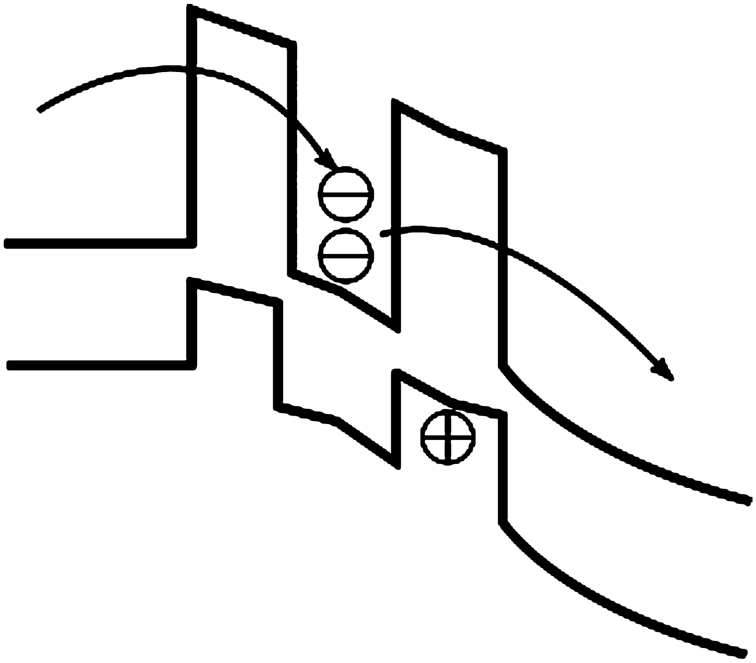

Figure 18. At steady state the two conduction-electron-mediated duon decay and generation processes are balanced, resulting in steady-state current across the double barrier structure.

rameter $\mathcal{G}$ in Eq. (66) is basically a ratio of the maximum e-h generation rate to the maximum duon decay rate. $\mathcal{G}$ is less than one since the e-h generation rate has only one channel, whereas the duon decay rate has several parallel channels. Clearly, in the presence of stimulated intermediate process, the maximum decay rate for duons must be larger than. the maximum e-h generation rate, otherwise the whole 2D dynamical system cannot be sustained or will become unbounded. It is estimated that $\mathcal{G}$ is much less than 0.5 for all the pertinent range of biasing condition before the current peak.

In the absence of any data, we can make a rough estimate for $\mathcal{G}$. Assume that there are $10^{11}\, cm^{-2}$ electrons available in the barrier capable of Zener tunneling to the drain, but only $10^{10}\, cm^{-2}$ were able to execute tunneling in $10^{-12}s$ (or $10^{9} cm^{-2}$ in $10^{-13}\, s$). Then, we have $\mathcal{G} = 10^{22}\, cm^{-2}\, s^{-1}$. We can also estimate the intravalence band tunneling time of the trapped holes to be around $10^{-13}\, s$ leading to $\alpha = 10^{13}\, s^{-1}$. If we take $\frac{1}{\beta} \simeq 10^{11}\, cm^{-2}$, then $\mathcal{G} = 10^{-2}$. Using these values, a reasonable value for Δ comes out to be about same order of magnitude as $\mathcal{G}$, which can easily lead to the inequality $(1 - \mathcal{G})^3/4 > \Delta$. We will. see that this last inequality has a very important role in our limit cycle analysis. In what follows we take $\mathcal{G} < 0.5$ to cover the physical range for $\mathcal{G}$.

In terms of these dimensionless variables, the stationary values of Q and Π are given by

$$\Pi^{\circ} = \frac{\mathcal{G}}{1-\mathcal{G}} \tag{67}$$

$$Q^{o} = (\frac{1-\mathcal{G}}{\Delta})^{1/2} \tag{68}$$

As mentioned before, for the physical situation in the relevant RTD, $0 < \mathcal{G} < 1.0$.

The question whether there is a nonstationary solution to our fundamental rate equations, involving interband processes as depicted in Fig. 16, can first be answered by examining the stability of the stationary point in (Π, Q) space. This is done by examining the neighborhood of the stationary point. Let us denote the coordinates of this neighborhood by

$$\begin{array}{rcl} \Pi & = & \Pi^{\circ} + p \\ Q & = & Q^{\circ} + q \end{array} \tag{69}$$

Substituting these in the coupled rate equations, Eqs. (63) and (64), and retaining only linear terms in p and q, we have

$$\frac{\partial}{\partial\tau}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \mathcal{G}(1-\mathcal{G}) & 2\mathcal{G}\Big[\dfrac{\Delta}{(1-\mathcal{G})}\Big]^{1/2} \\ -(1-\mathcal{G}) & -2\mathcal{G}\Big[\dfrac{\Delta}{(1-\mathcal{G})}\Big]^{1/2} \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix} \tag{70}$$

The solution for the trajectories in (Π, Q) space about the equilibrium point is given by

$$\begin{pmatrix} p \\ q \end{pmatrix} = A_1 e^{\lambda_1 \tau}\begin{pmatrix} V_1^p \\ V_1^q \end{pmatrix} + A_2 e^{\lambda_2 \tau}\begin{pmatrix} V_2^p \\ V_2^q \end{pmatrix} \tag{71}$$

where λ_1 and λ_2 are the eigenvalues of the matrix (M) defined by Eq. (70), and the corresponding eigenvectors are the column vectors V_1 and V_2, respectively, which can readily be determined from the knowledge of the eigenvalues. The eigenvalues are given by

$$\lambda_{1,2} = \frac{Tr(M)}{2} \pm \frac{1}{2}\sqrt{[Tr(M)]^2 - 4det(M)} \tag{72}$$

The character of the stationary point can thus be determined with the help of the invariants of the matrix (M), namely. $Tr(M)$, $det(M)$, and $D(M) = [Tr(M)]^2 - 4det(M)$. The stationary point cannot be a saddle point for physical reasons since $det(M) = 2\mathcal{G}(1-\mathcal{G})^{3/2}\sqrt{\Delta} > 0$. The physical processes depicted in Fig. 16 also suggest that the stationary point can only be any one of the following cases: stable focus, $Tr(M) < 0$, center, $Tr(M) = 0$, or unstable focus, $Tr(M) > 0$. This means that $D(M) < 0$ or $4\det(M) > [Tr(M)]^2$. Thus, there are two out of three chances that the oscillating processes depicted in Fig. 16 are maintained, depending on the value of the rate parameter Δ in relation to $\mathcal{G}$. Physically, we expect the limit cycle oscillation for uniqueness and stability.

For the unstable focus we have to demonstrate that a limit cycle exists. The region in parameter space where the structurally stable limit cycle solution is possible is the area under the bifurcation curve of Fig. 19, [locus of $Tr(M) = 0$]. Towards the end of this article, we will present experimental evidence that our physical model is correct by showing that the averaged trapped hole charge in the barrier is a slowly and linearly decreasing function of the bias when a limit cycle exists. This lead to results which give excellent qualitative agreement with the experiment (34,35,25,26). In practice, for a given material parameter Δ, we choose the operating bias $\mathcal{G}$ such that $Tr(M) > 0$. For a range of $\mathcal{G}$ where this is satisfied, we can optimize the operating point $\mathcal{G}$ to realize a THz source with optimum power and frequency.

The trace of (M) is given by $Tr(M) = \mathcal{G}\left\{(1-\mathcal{G}) - 2\Big[\dfrac{\Delta}{(1-\mathcal{G})}\Big]^{1/2}\right\}$. Thus, $Tr(M) > 0$ implies $(1-\mathcal{G})^3 > \Delta$. On the other hand, $D(M) < 0$ implies $(1-\mathcal{G})^3 < 4\Delta + 8\mathcal{G}^{-1}(1-\mathcal{G})^{5/2}\Delta^{1/2} + 4\sqrt{2}\mathcal{G}^{-1}(1-\mathcal{G})^{5/4}\Delta^{3/4}$. In the next section, we will employ a nonlinear perturbation technique using the method of multiple time scales with values of the rate parameter near the bifurcation curve, Fig. 19. As we shall show in the following nonlinear analysis, a unique limit cycle indeed occurs at $Tr(M) > 0$ with rate parameter expanded around the parameters for $Tr(M) = 0$. The amplitude and frequency of oscillation is expected to depend on the actual values of the two-parameters Δ and $\mathcal{G}$ in the region where $Tr(M) > 0$.

Nonlinear analysis and limit cycle solution. Retaining nonlinear terms for p and q measured from the stationary point, the rate equation, from Eqs. (10) and (11), in matrix form becomes

$$\frac{\partial}{\partial\tau}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \mathcal{G}(1-\mathcal{G}) & 2\mathcal{G}\Big[\dfrac{\Delta}{(1-\mathcal{G})}\Big]^{1/2} \\ -(1-\mathcal{G}) & -2\mathcal{G}\Big[\dfrac{\Delta}{(1-\mathcal{G})}\Big]^{1/2} \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix} + \begin{pmatrix} N_p \\ N_q \end{pmatrix} \tag{73}$$

where

$$\begin{pmatrix} N_p \\ N_q \end{pmatrix} = \begin{pmatrix} \left\{\begin{array}{l} (1-\mathcal{G})^3\,p^2 + 2[\Delta(1-\mathcal{G})^{1/2}\,pq + \dfrac{\Delta\mathcal{G}}{(1-\mathcal{G})}\,q^2 + \Delta\,pq^2] \\ +\displaystyle\sum_{n=3}^{\infty}(-1)^n(1-\mathcal{G})^{n+1}\,p^n \end{array}\right\} \\ -2[\Delta(1-\mathcal{G})]^{1/2}\,pq - \dfrac{\Delta\mathcal{G}}{(1-\mathcal{G})}q^2 - \Delta\,pq^2 \end{pmatrix} \tag{74}$$

The perturbation technique employed in what follows essentially transforms the above nonlinear equation into a hierarchy of solvable and simpler equations, obtained by equating coefficients of powers of the smallness parameter. Near $Tr(M) = 0$, we use as our smallness parameter the departure of $Tr(M)$ from zero, i.e., the departure of the equality $(1-\mathcal{G})^3/4 = \Delta_c$.

Let the smallness parameter be $\epsilon = \sqrt{\left\{\Delta - \dfrac{[(1-\mathcal{G})^3]}{4}\right\}\mathcal{D}}$ where $\mathcal{D}$ is to be determined from the expansion of Δ in powers of ϵ. Since $\mathcal{G}$ is assumed constant at fixed bias, i.e.. a function only of the external bias, ϵ is also a measure of the departure of Δ from Δ_c,. We make the following expansion:

$$\Delta = \sum_{j=0}^{\infty}\varepsilon^j\Delta_j, \quad \text{where} \quad \Delta_0 = \Delta_c \tag{75}$$

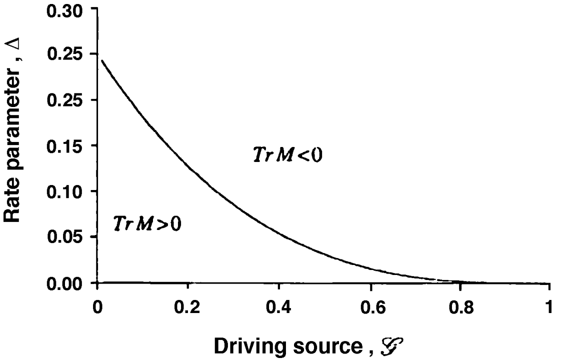

Figure 19. Plot of the bifurcation curve, $Tr(M) = 0$, which is the curve for $\Delta = \Delta_c$.

As it turns out, $D \simeq \Delta_2$ to zeroth order in ϵ in the analysis that follows [refer to Eqs. (85) and (125) below]. We also expand the matrix (M) in powers of ϵ through direct Taylor expansion in powers of $(\Delta - \Delta_c)$, as

$$(M) = (M_c) + \epsilon\Delta_1 \left(\frac{\partial M(\Delta)}{\partial \Delta}|_{\Delta=\Delta_c} \right) + \frac{1}{2}\epsilon^2 \left[\begin{array}{c} \Delta_2 \left(\dfrac{\partial M(\Delta)}{\partial \Delta}|_{\Delta=\Delta_c} \right) \\ +\Delta_1^2 \left(\dfrac{\partial^2 M(\Delta)}{\partial \Delta^2}|_{\Delta=\Delta_c} \right) \end{array} \right] + O(\epsilon^3) \quad (76)$$

Using $(1-\mathcal{G})^3 = 4\Delta_c$, we obtain the following expressions:

$$(M_c) = \mathcal{G}(1-\mathcal{G}) \begin{pmatrix} 1 & 1 \\ -\mathcal{G}^{-1} & -1 \end{pmatrix} \quad (77)$$

$$(M_1) = \left(\frac{\partial M(\Delta)}{\partial \Delta}|_{\Delta=\Delta_c} \right) = 2\mathcal{G}(1-\mathcal{G})^{-2} \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \quad (78)$$

$$(M_2) = \frac{1}{2} \left(\frac{\partial^2 M(\Delta)}{\partial \Delta^2}|_{\Delta=\Delta_c} \right) = -2\mathcal{G}(1-\mathcal{G})^{-5} \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \quad (79)$$

We let the solution depends on time τ in a combination $\tau_o = \tau$ and $\tau_1 = (\Delta - \Delta_c)\tau$. Thus, instead of determining the solution in terms of τ we seek the solution as a function of τ_o, τ, and ϵ. This method of doing the nonlinear perturbation analysis is well known and is often referred to as the method of multiple time scales (39). This has the virtue that it separates the dependence of the solution into the fast and slow time scales. For limit cycle behavior, for example, we expect that the amplitude of the oscillation is only a function of the slow time scale. The left side of the rate equation can now be written a

$$\frac{\partial}{\partial \tau} \begin{pmatrix} p(\tau_o, \tau, \epsilon) \\ q(\tau_o, \tau, \epsilon) \end{pmatrix} = \left[\frac{\partial}{\partial \tau_o} + (\Delta - \Delta_c) \frac{\partial}{\partial \tau_1} \right] \begin{pmatrix} p(\tau_o, \tau, \epsilon) \\ q(\tau_o, \tau, \epsilon) \end{pmatrix} \quad (80)$$

The analysis is greatly simplified if we take the leading order of the last term in Eq. (73) as second order in the smallness parameter. Thus, for the solution we adopt the following expansion:

$$\begin{pmatrix} p \\ q \end{pmatrix} = \sum_{j=0}^{\infty} e^{j+1} \begin{pmatrix} p_j \\ q_j \end{pmatrix} \quad (81)$$

Therefore, any finite solution that will found in this analysis will invariably indicate that the limit cycle occurs for values of the parameter away from the critical point, $Tr(M) = 0$, i.e., for nonzero smallness parameter. This holds for example in our numerical simulation for the limit cycle of conventional RTD operating at the NDR region (33). With Eq. (81), the non-linear term in Eq. (73) acquires the following expansion in terms of the smallness parameter:

$$\begin{pmatrix} N_p \\ N_q \end{pmatrix} = \epsilon^2 \begin{pmatrix} N_2^p \\ N_2^q \end{pmatrix} + \epsilon^3 \begin{pmatrix} N_3^p \\ N_3^q \end{pmatrix} + O(\epsilon^4) \quad (82)$$

where

$$\begin{pmatrix} N_2^p \\ N_2^q \end{pmatrix} = (1-\mathcal{G})^2 \begin{pmatrix} \left[p_o q_o + \left(\dfrac{\mathcal{G}}{4} \right) q_o^2 + (1-\mathcal{G})\, p_o^2 \right] \\ -\left[p_o q_o + \left(\dfrac{\mathcal{G}}{4} \right) q_o^2 \right] \end{pmatrix} \quad (83)$$

$$\begin{pmatrix} N_3^p \\ N_3^q \end{pmatrix} = (1-\mathcal{G})^2 \begin{pmatrix} \left[\begin{array}{l} p_o q_1 + p_1 q_o + 2(1-\mathcal{G})\, p_o\, p_1 + \left(\dfrac{\mathcal{G}}{2} \right) q_1 q_o + \dfrac{(1-\mathcal{G})}{4}\, p_o q_o^2 \\ + \left(\dfrac{\Delta_1 \mathcal{G}}{(1-\mathcal{G})^3} \right) q_o^2 + 2 \left\{ \dfrac{\Delta_2}{(1-\mathcal{G})^3} \right\}^{1/2} p_o q_o - (1-\mathcal{G})^2\, p_o^3 \end{array} \right] \\ -\left[\begin{array}{l} p_o q_1 + p_1 q_o + \left(\dfrac{\mathcal{G}}{2} \right) q_1 q_o + \dfrac{(1-\mathcal{G})}{4}\, p_o q_o^2 \\ + \left(\dfrac{\Delta_1 \mathcal{G}}{(1-\mathcal{G})^3} \right) q_o^2 + 2 \left\{ \dfrac{\Delta_2}{(1-\mathcal{G})^3} \right\}^{1/2} p_o q_o \end{array} \right] \end{pmatrix} \quad (84)$$

We did not show nonlinear terms with fractional powers of ϵ in Eq. (82) associated with Δ_1 in Eqs. (74) since the left-hand side of the rate equation does not contain fractional powers of ϵ. To eliminate the occurrence of these fractional powers of ϵ, we have to make $\Delta_1 = 0$ in the expansion of Δ, Eq. (75), and also in Eqs. (76) and (84).

Upon substituting all the expanded quantities in the non-linear rate equation, Eq. (73), we obtain a hierarchy of simpler equations. Those arising from the first up to the third powers of ϵ are given below,

$$\mathcal{L}_o \begin{pmatrix} p_o \\ q_o \end{pmatrix} = 0 \quad (85)$$

$$\mathcal{L}_o \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} N_2^p(p_o, q_o) \\ N_2^q(p_o, q_o) \end{pmatrix} \quad (86)$$

$$\mathcal{L}_o \begin{pmatrix} p_2 \\ q_2 \end{pmatrix} + \Delta_2 \mathcal{L}_1 \begin{pmatrix} p_o \\ q_o \end{pmatrix} = \begin{pmatrix} N_3^p(p_o, q_o;\ p_1, q_1) \\ N_3^q(p_o, q_o;\ p_1, q_1) \end{pmatrix} \quad (87)$$

where

$$\mathcal{L}_o = \left(\frac{\partial}{\partial \tau_o} - (M_c) \right) \tag{88}$$

$$\mathcal{L}_1 = \left(\frac{\partial}{\partial \tau_1} - (M_1) \right) \tag{89}$$

The first equation in the hierarchy turns out to be a simple eigenvalue problem, analogous to our linear-stability analysis before. The only difference is that the present eigenvalue problem has to be solved with values of the parameter at the critical point, where $Tr(M) = 0$, using the matrix (M_c).

Using the matrix expression of Eq. (77) for (M_c), the eigenvalues for Eq. (85) are

$$\gamma_{1,2} = \pm i \sqrt{\mathcal{G}} (1 - \mathcal{G})^{3/2} = \pm i\omega \tag{90}$$

and the corresponding eigenvectors are

$$Y_o^{1,2} = \begin{pmatrix} 1 \\ -1 \pm \dfrac{i\omega}{\mathcal{G}(1-\mathcal{G})} \end{pmatrix} \tag{91}$$

The solution can be written in the form

$$\begin{aligned} \begin{pmatrix} p_o \\ q_o \end{pmatrix} &= \Theta(\tau_1)\Psi(\tau_o) + c.c. \\ &= \Theta(\tau_1) \left\{ \exp(i\omega\tau_o) \begin{pmatrix} 1 \\ -1 \pm \dfrac{i\omega}{\mathcal{G}(1-\mathcal{G})} \end{pmatrix} \right\} + c.c. \end{aligned} \tag{92}$$

where the separation between the slow and fast time scales is explicitly written. Note that Eq. (87) determines the nature of the dependence of the solution on the slow time scale τ_1 by virtue of the presence of the operator $\mathcal{L}_1$.

Next, we obtain the solution for $\begin{pmatrix} p_1 \\ q_1 \end{pmatrix}$ by solving Eq. (86). The right-hand side is now known since it is only a function of $\begin{pmatrix} p_o \\ q_o \end{pmatrix}$. The solution can be obtained separately for terms involving fast and slow time scales, where the fast time scale occurs only in the exponential terms. The right-hand side of Eq. (86) can be written as

$$\begin{aligned} \begin{pmatrix} N_2^p(p_o, q_o) \\ N_2^q(p_o, q_o) \end{pmatrix} &= \begin{pmatrix} A^p \\ A^q \end{pmatrix} |\Theta(\tau_1)|^2 \\ &\quad + \left[\begin{pmatrix} B^p \\ B^q \end{pmatrix} \Theta(\tau_1)^2 \exp(2i\omega\tau_o) + c.c. \right] \end{aligned} \tag{93}$$

where

$$\begin{pmatrix} A^p \\ A^q \end{pmatrix} = (1-\mathcal{G})^2 \begin{pmatrix} \dfrac{1}{2} - 2\mathcal{G} \\ \dfrac{3}{2} \end{pmatrix} \tag{94}$$

$$\begin{pmatrix} B^p \\ B^q \end{pmatrix} = (1-\mathcal{G})^2 \begin{pmatrix} -\left(\dfrac{1+2\mathcal{G}}{4} \right) + \dfrac{i\omega(2-\mathcal{G})}{2\mathcal{G}(1-\mathcal{G})} \\ \left(\dfrac{5-2\mathcal{G}}{4} \right) - \dfrac{i\omega(2-\mathcal{G})}{2\mathcal{G}(1-\mathcal{G})} \end{pmatrix} \tag{95}$$

If we write the solution for $\begin{pmatrix} p_1 \\ q_1 \end{pmatrix}$ as

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} \alpha^p \\ \alpha^q \end{pmatrix} |\Theta(\tau_1)|^2 + \left[\begin{pmatrix} \beta^p \\ \beta^q \end{pmatrix} \Theta(\tau_1)^2 \exp(2i\omega\tau_o) + c.c. \right] \tag{96}$$

then we have to solve the following equations for the coefficients:

$$\begin{pmatrix} \alpha^p \\ \alpha^q \end{pmatrix} = -(M_c)^{-1} \begin{pmatrix} A^p \\ A^q \end{pmatrix} \tag{97}$$

$$\begin{pmatrix} \beta^p \\ \beta^q \end{pmatrix} = \begin{pmatrix} 2i\omega - \mathcal{G}(1-\mathcal{G}) & -\mathcal{G}(1-\mathcal{G}) \\ (1-\mathcal{G}) & 2i\omega + \mathcal{G}(1-\mathcal{G}) \end{pmatrix}^{-1} \begin{pmatrix} B^p \\ B^q \end{pmatrix} \tag{98}$$

Substituting the expressions given by Eqs. (94) and (95) in Eqs. (97) and (98), we obtained the following expressions:

$$\begin{pmatrix} \alpha^p \\ \alpha^q \end{pmatrix} = \frac{(1-\mathcal{G})}{\mathcal{G}} \begin{pmatrix} 2\mathcal{G} \\ -\dfrac{1}{2} \end{pmatrix} \tag{99}$$

$$\begin{aligned} \begin{pmatrix} \beta^p \\ \beta^q \end{pmatrix} &= \frac{1}{6\mathcal{G}(1-\mathcal{G})} \\ &\quad \cdot \begin{pmatrix} 4(1-\mathcal{G})^3 + i\omega(1+2\mathcal{G}) \\ (\dfrac{15}{2}\mathcal{G} - \dfrac{9}{2} - 3\mathcal{G}^2)(1-\mathcal{G}) - i\omega(8 - 3\mathcal{G} - 2\mathcal{G}^{-1}) \end{pmatrix} \end{aligned} \tag{100}$$

and thus $\begin{pmatrix} p_1 \\ q_1 \end{pmatrix}$ is determined by Eq. (96).

At this stage of the calculation, we can expect that all of $\begin{pmatrix} p_j \\ q_j \end{pmatrix}$ contains, as factors, various powers of $\Theta(\tau_1)$ as well as powers of its absolute value, and their combinations. Therefore, in order to find out if a limit cycle exists, it is important to examine the τ_1 dependence of $\Theta(\tau_1)$ and thereby determine if a well-defined finite limit exists for $\Theta(\tau_1)$ as $t \Rightarrow \infty$. Moreover, before we can calculate $\begin{pmatrix} p_2 \\ q_2 \end{pmatrix}$, we need to know the τ_1 derivative of $\Theta(\tau_1)$ in Eq. (87). This information can be obtained by imposing the "solvability" condition. This condition makes use of the property of the solution to the adjoint of $\mathcal{L}_o$, denoted as $\mathcal{L}_o^\dagger$. Let $\mathcal{L}_o^\dagger \mathcal{R} = 0$, then $\left\langle \mathcal{R}, \mathcal{L}_o \begin{pmatrix} p_2 \\ q_2 \end{pmatrix} \right\rangle = \left\langle \mathcal{L}_o^\dagger \mathcal{R}, \begin{pmatrix} p_2 \\ q_2 \end{pmatrix} \right\rangle = 0$, where the scalar product is defined by $\langle \nu, \mu \rangle \equiv \lim_{T \to \infty} \dfrac{1}{T} \displaystyle\int_0^T \nu^* \cdot \mu d\tau$. Therefore, from Eq. (87) we must have

$$\left\langle \mathcal{R}, -\Delta_2 \mathcal{L}_1 \begin{pmatrix} p_o \\ q_o \end{pmatrix} + \begin{pmatrix} N_3^p(p_o, q_o;\ p_1, q_1) \\ N_3^q(p_o, q_o;\ p_1, q_1) \end{pmatrix} \right\rangle = 0 \tag{101}$$

We refer the readers to Morse and Feshback (40), in showing that the cigensolutions of $\mathcal{L}_o$ and $\mathcal{L}_o^\dagger$ form biorthogonal set of eigenvectors, where $\mathcal{L}_o^\dagger$ here is given by

$$\mathcal{L}_o^\dagger = \left(-\frac{\partial}{\partial \tau_o} - \mathcal{G}(1-\mathcal{G}) \begin{pmatrix} 1 & -\mathcal{G}^{-1} \\ 1 & -1 \end{pmatrix} \right) \tag{102}$$

For example, the eigensolutions to $\mathcal{L}_o^\dagger \mathcal{R} = 0$ with eigenvalues $\mu_1 = i\omega$ and $\mu_1 = -i\omega$ are given by $\mathcal{R}_1 = \exp(-i\omega\tau_o) \begin{pmatrix} \mathcal{R}_1^p \\ \mathcal{R}_1^q \end{pmatrix}$ and $\mathcal{R}_2 = \mathcal{R}_1^*$, respectively, where

$$\begin{pmatrix} \mathcal{R}_1^p \\ \mathcal{R}_1^q \end{pmatrix} = \begin{pmatrix} 1 \\ \mathcal{G}(1 - \dfrac{i\omega}{\mathcal{G}(1-\mathcal{G})}) \end{pmatrix} \tag{103}$$

The eigensolution $\mathcal{R}_1 = \exp(-i\omega\tau_o) \begin{pmatrix} \mathcal{R}_1^p \\ \mathcal{R}_1^q \end{pmatrix}$ is orthogonal to the eigenvector of $\mathcal{L}_o$ for the same eigenvalue, i.e., $\left\langle \mathcal{R}_1 | \exp(i\omega\tau_o) Y_o^1 \right\rangle = 0$, since $\mu_1^* = -i\omega = \gamma_2$ of $\mathcal{L}_o$, Eq. (90). Thus, we are lead to the following relation:

$$\langle \mathcal{R}_2 | \Psi(\tau_o) \rangle = 2(1-\mathcal{G}) + \frac{2i\omega}{(1-\mathcal{G})} \tag{104}$$

where only the Y_o^1 [in Eqs. (91) and (92)] contributes by virtue of the biorthogonality, i.e., the complex conjugate part of $\Psi(\tau_o)$ also does not contribute in Eq. (104). Thus, with $\mathcal{R}$ chosen to be equal to $\mathcal{R}_2$, the scalar product in Eq. (101) can be evaluated, and defines the differential equation for $\Theta(\tau_1)$. We obtain

$$\frac{\partial\Theta(\tau_1)}{\partial\tau_1} = \frac{\langle\mathcal{R}, (M_1)\Psi(\tau_o)\rangle}{\langle\mathcal{R}, \Psi(\tau_o)\rangle}\Theta(\tau_1) + \frac{\Delta_2^{-1}}{\langle\mathcal{R}, \Psi(\tau_o)\rangle} \cdot \left\langle \mathcal{R}, \begin{pmatrix} N_3^p(p_o, q_o;\ p_1, q_1) \\ N_3^q(p_o, q_o;\ p_1, q_1) \end{pmatrix} \right\rangle \tag{105}$$

Note that in Eq. (105) nonzero τ_o integration comes only from p_oq_1, p_1q_o, p_op_1, q_1q_o, and $p_oq_o^2$ terms in $\begin{pmatrix} N_3^p(p_o, q_o;\ p_1, q_1) \\ N_3^q(p_o, q_o;\ p_1, q_1) \end{pmatrix}$, Eq. (84), while the rest including the complex conjugate part do not contribute as $T \Rightarrow \infty$ by virtue of the appearance of delta function of the frequency sum after the τ_o integration. Therefore, by taking into account only the contributing terms, we can write Eq. (105) in a simpler form as

$$\frac{\partial\Theta(\tau_1)}{\partial\tau_1} = \eta\Theta(\tau_1) + \frac{\sigma}{\Delta_2}\Theta(\tau_1)|\Theta(\tau_1)|^2 \tag{106}$$

where

$$\eta = \frac{\langle\mathcal{R}, (M_1)\Psi(\tau_o)\rangle}{\langle\mathcal{R}, \Psi(\tau_o)\rangle} \tag{107}$$

and

$$\sigma\Theta(\tau_1)|\Theta(\tau_1)|^2 = \frac{(1-\mathcal{G})^2}{\langle\mathcal{R}, \Psi(\tau_o)\rangle}\left\langle \mathcal{R}, \begin{pmatrix} Q_3(p_o, q_o;\ p_1, q_1) - (1-\mathcal{G})^3\, p_o^3 \\ -Q_3(p_o, q_o;\ p_1, q_1) \end{pmatrix} \right\rangle \tag{108}$$

with

$$Q_3(p_o, q_o;\ p_1, q_1) = p_oq_1 + p_1q_o + 2(1-\mathcal{G})\, p_o\, p_1 + \left(\frac{\mathcal{G}}{2}\right) q_1q_o + \frac{(1-\mathcal{G})}{4}\, p_oq_o^2 \tag{109}$$

We thus obtain

$$\eta = -\mathcal{G}(1-\mathcal{G})^{-2} + i\omega(1-\mathcal{G}) - 3 \tag{110}$$

and

$$\begin{aligned}\sigma\Theta(\tau_1)|\Theta(\tau_1)|^2 \;=\; & \lim_{T\Rightarrow\infty}\frac{1}{T}\int_0^T \exp(-i\omega\tau_o)d\tau_o(1-\mathcal{G}) \\ & \times\{[(1-\mathcal{G})^2 + i\omega]Q_3 - (1-\mathcal{G})^4\, p_o^3\} \\ & \times\left(2(1-\mathcal{G}) + \frac{2i\omega}{(1-\mathcal{G})}\right)^{-1}\end{aligned} \tag{111}$$

Upon performing the τ_o integration in Eq. (111), we obtain

$$\begin{aligned}\sigma\Theta(\tau_1)|\Theta(\tau_1)|^2 \;=\; & \frac{1}{4}\{[(1-\mathcal{G})^2 + i\omega]\hat{Q}_3 + 2(1-\mathcal{G})\hat{p}_o\hat{p}_1 - (1-\mathcal{G})^3\hat{p}_o^3\} \\ & \times(2(1-\mathcal{G}) - \frac{2i\omega}{(1-\mathcal{G})})\end{aligned} \tag{112}$$

where

$$\begin{aligned}\hat{Q}_3 \;=\; & A_pC_q + A_p^*B_q + A_qC_p + A_q^*B_p + \frac{\mathcal{G}}{2}(A_qC_q + A_q^*B_q) \\ & - \frac{(1-\mathcal{G})}{4}(2A_p|A_q|^2 + A_p^*A_q^2)\end{aligned} \tag{113}$$

$$\begin{aligned}\hat{p}_o\hat{p}_1 &= A_pC_p + A_p^*B_p \\ \hat{p}_o^3 &= 3A_p|A_p|^2\end{aligned} \tag{114}$$

$$\begin{aligned}A_p &= \Theta(\tau_1), \\ A_q &= \Theta(\tau_1)\left[-1 + \frac{i\omega}{\mathcal{G}(1-\mathcal{G})}\right]\end{aligned} \tag{115}$$

$$\begin{aligned}C_p &= 2(1-\mathcal{G})|\Theta(\tau_1)|^2 \\ C_q &= -\left[\frac{(1-\mathcal{G})}{2\mathcal{G}}\right]|\Theta(\tau_1)|^2\end{aligned} \tag{116}$$

$$\begin{aligned}B_p &= \frac{[4(1-\mathcal{G})^3 + i\omega(1+2\mathcal{G})]\Theta(\tau_1)^2}{6\mathcal{G}(1-\mathcal{G})} \\ B_q &= \frac{\left[\left(\frac{15}{2}\mathcal{G}\text{-}\ \frac{9}{2} - 3\mathcal{G}^2\right)(1-\mathcal{G}) - i\omega(8 - 3\mathcal{G} - 2\mathcal{G}^{-1})\right]\Theta(\tau_1)^2}{6\mathcal{G}(1-\mathcal{G})}\end{aligned} \tag{117}$$

Carrying out the operation in Eq. (112), using Eqs. (113)–(115), we obtain

$$\begin{aligned}\sigma \;=\; & -\frac{(1-\mathcal{G})^3}{16\mathcal{G}}(4 + 19\mathcal{G} - 8\mathcal{G}^2) \\ & - \frac{i\omega(1-\mathcal{G})}{48\mathcal{G}^2}(8 - 31\mathcal{G} + 17\mathcal{G}^2 + 24\mathcal{G}^3)\end{aligned} \tag{118}$$

We note that for $0.0 < \mathcal{G} < 1.0$, $\text{Re}\eta < 0.0$, and $\text{Re}\sigma < 0.0$. As a check we also note that both η and σ in Eq. (110) and Eq. (118), respectively, goes to zero nonlinearly as $\mathcal{G} \Rightarrow 1.0$.

We solve for the absolute value and phase of $\Theta(\tau_1)$ by writing this in polar form and equating the real and imaginary parts on both sides of Eq. (106). With $\Theta(\tau_1) = |\Theta(\tau_1)|\exp i\phi(\tau_1)$, we obtain exactly solvable equations,

$$\frac{\partial}{\partial\tau_1}|\Theta(\tau_1)| = \text{Re}\eta|\Theta(\tau_1)| + \frac{\text{Re}\sigma}{\Delta_2}|\Theta(\tau_1)|^3 \tag{119}$$

$$\frac{\partial}{\partial\tau_1}\phi(\tau_1) = \text{Im}\eta + \frac{\text{Im}\sigma}{\Delta_2}|\Theta(\tau_1)|^2 \tag{120}$$

A Solution of Eq. (119) in which the $|\Theta(0)|$ can be arbitrarily independent of the limiting value $|\Theta(\infty)|$, which is the possible limit cycle value, is of the form

$$|\Theta(\tau_1)| = \frac{|\Theta(0)||\Theta(\infty)|\exp(\text{Re}\eta\tau_1)}{[|\Theta(\infty)|^2 + \{\exp[2\text{Re}\eta\tau_1 - 1]|\Theta(0)|^2\}]^{1/2}} \tag{121}$$

where $|\Theta(\infty)| = \left[-\frac{(\text{Re}\eta\Delta_2)}{\text{Re}\sigma}\right]^{1/2}$ which is a real value if $\Delta_2 < 0$, since $\text{Re}\eta < 0$ and $\text{Re}\sigma < 0$.

Indeed, in real time $|\Theta(\infty)| = \lim_{t\to\infty}|\Theta(\tau_1)|$ only if $\text{Re}\eta\tau_1 \Rightarrow \infty$ as $t \Rightarrow \infty$ hence only if $(\Delta - \Delta_c) < 0$ or $\Delta < \Delta_c$, in Eq. (121), i.e. $\Delta_2 < 0$. This is consistent with the criteria for unstable focus in the linear analysis. Otherwise, $\lim_{t\to\infty}|\Theta(\tau_1)| = 0$ if $(\Delta - \Delta_c) > 0$ or $\Delta > \Delta_c$. Thus, a well-defined limiting value of $|\Theta(\tau_1)|$ as $t \Rightarrow \infty$ exists only for $\Delta < \Delta_c$. This is the limit cycle value of $|\Theta(\tau_1)|$. Substituting the now known functional form of $|\Theta(\tau_1)|$ in Eq. (121), we can also integrate Eq. (120). The result is

$$\begin{aligned}\phi(\tau_1) \;=\; & Const + \text{Im}\eta\tau_1 \\ & + \frac{\text{Im}\sigma}{\Delta_2}\,\frac{|\Theta(\infty)|^2}{2\text{Re}\eta}\left\{ln\,\frac{[|\Theta(\infty)|^2 + \exp[2\text{Re}\eta\tau_1 - 1]|\Theta(0)|^2]}{[|\Theta(\infty)|^2 - |\Theta(0)|^2]}\right\}\end{aligned} \tag{122}$$

The limiting values of $\phi(\tau_1)$ are

$$\lim_{t\Rightarrow\infty}\phi(\tau_1)=\begin{cases} Const+\mathrm{Im}\eta\tau_1,\ \text{if } \Delta>\Delta_c \\ Const+\mathrm{Im}\eta\tau_1+\dfrac{\mathrm{Im}\sigma}{\Delta_2}\left|\Theta(\infty)\right|^2\tau_1 \\ +ln\left\{\dfrac{|\Theta(0)|^2}{[|\Theta(\infty)|^2-|\Theta(0)|^2]}\right\},\ \text{if } \Delta<\Delta_c \end{cases} \tag{123}$$

Taking the overall constant of integration equal to zero, we end up with the expression for the limit cycle value of $|\Theta(\tau_1)|$ given as

$$\lim_{t\Rightarrow\infty}\Theta(\tau_1)=\left[\frac{-\mathrm{Re}\eta\Delta_2}{\mathrm{Re}\sigma}\right]^{1/2}\exp i[\mathrm{Im}\eta(\Delta-\Delta_c)+\frac{\mathrm{Im}\sigma}{\Delta_2}|\Theta(\infty)|^2(\Delta-\Delta_c)]\tau \tag{124}$$

With the limit cycle value of $\Theta(\tau_1)$ known, $\begin{pmatrix} p_o \\ q_o\end{pmatrix}$ and $\begin{pmatrix} p_1 \\ q_1\end{pmatrix}$ at the limit cycle are completely determined. Thus, to second order in the smallness parameter, and by virtue of Eq. (69) and Eq. (75), we have the limit cycle solution given as

$$\begin{aligned}\begin{pmatrix}\Pi \\ \mathcal{Q}\end{pmatrix} &= \begin{pmatrix}\Pi^o \\ \mathcal{Q}^o\end{pmatrix}+\left|\tfrac{(\Delta-\Delta_c)}{\Delta_2}\right|^{1/2}\begin{pmatrix}p_o \\ q_o\end{pmatrix}+\left|\tfrac{(\Delta-\Delta_c)}{\Delta_2}\right|\begin{pmatrix}p_1 \\ q_1\end{pmatrix} \\ &\quad +O\left(\left|\tfrac{(\Delta-\Delta_c)}{\Delta_2}\right|^{3/2}\right)\end{aligned} \tag{125}$$

where from Eq. (92), we have

$$\begin{aligned}\begin{pmatrix}p_o \\ q_o\end{pmatrix} &= \tfrac{|\Theta(\infty)|}{\sqrt{\mathcal{G}}}\begin{pmatrix}2cos\Omega\tau \\ -2\sqrt{\mathcal{G}}cos\Omega\tau-\sqrt{(1-\mathcal{G})2sin\Omega\tau}\end{pmatrix} \\ &= \tfrac{2|\Theta(\infty)|}{\sqrt{\mathcal{G}}}\begin{pmatrix}cos\Omega\tau \\ -sin(\Omega\tau+\Phi)\end{pmatrix}\end{aligned} \tag{126}$$

where

$$tan\Phi=\sqrt{\frac{\mathcal{G}}{(1-\mathcal{G})}}>0 \tag{127}$$

$$\Omega=\sqrt{\mathcal{G}}(1-\mathcal{G})^{3/2}+[\mathrm{Im}\eta(\Delta-\Delta_c)+\frac{\mathrm{Im}\sigma}{\Delta_2}|\Theta(\infty)|^2(\Delta-\Delta_c)] \tag{128}$$

From Eqs. (96), (99), and (100), we also have

$$\begin{pmatrix}p_1 \\ q_1\end{pmatrix}=|\Theta(\infty)|^2\left\{\begin{matrix}\frac{(1-\mathcal{G})}{\mathcal{G}}\begin{pmatrix}2\mathcal{G} \\ -1/2\end{pmatrix} \\ +\frac{2}{6\mathcal{G}(1-\mathcal{G})}\begin{pmatrix}4(1-\mathcal{G})^3cos2\Omega\tau-\omega(1+2\mathcal{G})sin2\Omega\tau \\ \left[\begin{matrix}\left(\frac{15}{2}\mathcal{G}\text{-}\frac{9}{2}-3\mathcal{G}^2\right)(1-\mathcal{G})cos2\Omega\tau \\ +\omega(8-3\mathcal{G}-2\mathcal{G}^{-1})\end{matrix}\right]sin2\Omega\tau\end{pmatrix}\end{matrix}\right\} \tag{129}$$

We note that Eq. (129) also contains a time-independent term, indicating higher-order shifts of the center of the limit cycle from the stationary point ($\mathcal{Q}^o$, Π^o). Therefore, the average value of $\begin{pmatrix}\Pi \\ \mathcal{Q}\end{pmatrix}$ is given by

$$\begin{aligned}\begin{pmatrix}\Pi \\ \mathcal{Q}\end{pmatrix}_{average} &= \begin{pmatrix}\Pi^o \\ \mathcal{Q}^o\end{pmatrix}+\left|\tfrac{(\Delta-\Delta_c)}{\Delta_2}\right||\Theta(\infty)|^2\tfrac{(1-\mathcal{G})}{\mathcal{G}}\begin{pmatrix}2\mathcal{G} \\ -1/2\end{pmatrix} \\ &\quad +O(\epsilon^3)\end{aligned} \tag{130}$$

where the leading higher-order corrections comes from the time-independent terms.

Variation of the Average Value and Amplitude with Bias. We see that the limit cycle occurs within the range of values of the parameter Δ where the criterion for unstable focus $(1-\mathcal{G})^3>4\Delta$ (i.e., $\mathrm{Tr}(M)>0$) also holds, analogous to our calculation of the limit cycle of AlGaAs/GaAs/ AlGaAs double-barrier heterostructure operating in the NDR region (33). From Eqs. (62), (125), and (126), the leading average value of the total trapped hole charge in the barrier is given by

$$\begin{aligned}\beta Q_B &= \Pi^o+\mathcal{Q}^o \\ &= \frac{\mathcal{G}}{(1-\mathcal{G})}+\left(\frac{1-\mathcal{G}}{\Delta}\right)^{1/2}\end{aligned} \tag{131}$$

This is a slowly varying function of $\mathcal{G}$ since it is a sum of an increasing function and a decreasing function. For values of Δ under the critical point, i.e., limit cycle device operation, the rate off change with respect to $\mathcal{G}$ is decreasing. Indeed, we have

$$\frac{d(\beta Q_B)}{d\mathcal{G}}=\frac{1}{(1-\mathcal{G})^2}-\left(\frac{1}{4\Delta(1-\mathcal{G})}\right)^{1/2}\leq 0 \tag{132}$$

where the equality is obtained at the critical point $\Delta=\Delta_c$. Since $\mathcal{G}$ is our measure of the voltage applied at the drain contact, we conclude from Eq. (132), and by taking into account the higher-order correction terms which increase with $\mathcal{G}$ that the average total hole charge trapped in the barrier is slowly decreasing with the applied bias to the leading orders, in the range of $\mathcal{G}$ of physical interest. Moreover, it is linearly decreasing in the presence of oscillation, i.e., $\Delta<\Delta_c$. Note that for nonoscillatory case, $\Delta>\Delta_c$, the total trapped hole charge is a nonlinearly increasing function of $\mathcal{G}$. Later, we will show that the oscillatory limit cycle operation is supported by the experiment.

We now show that the amplitude of oscillation increases with applied bias in the range of physical interest. From Eqs. (125) and (126), this amplitude is given by the following expression:

$$\begin{aligned}&\left|\frac{(\Delta-\Delta_c)}{\Delta_2}\right|^{1/2}\frac{2|\Theta(\infty)|}{\sqrt{\mathcal{G}}} \\ &=\left|\frac{(\Delta-\Delta_c)}{\Delta_2}\right|^{1/2}\frac{2}{\sqrt{\mathcal{G}}}\left[-\frac{(\mathrm{Re}\eta\Delta_2)}{\mathrm{Re}\sigma}\right]^{1/2}\end{aligned} \tag{133}$$

where η is given by Eq. (110) and σ is determined from Eq. (118). From Eqs. (110) and (118), we have the final expression for the amplitude given by

$$\begin{aligned}&\left|\frac{\Delta-\Delta_c}{\Delta_2}\right|^{1/2}\frac{2|\Theta(\infty)|}{\sqrt{G}} \\ &= 4\left|\frac{(\Delta-\Delta_c)}{\Delta_2}\right|^{1/2}\left[\frac{4G|\Delta_2|}{(1-G)^5|\{4+19G-8G^2\}|}\right]^{1/2}\end{aligned} \tag{134}$$

The expression in the curly bracket in the denominator of Eq. (134) goes to zero as $\mathcal{G}$ increases and approaches 1, while the numerator increases with $\mathcal{G}$. We are thus left with an expression which is an increasing function of $\mathcal{G}$ which is proportional to the applied bias.

Equation (134) explicitly shows that the amplitude of oscillation increases with bias. Denoting the leading time-dependent part of $Q_B(t)$ as $\delta Q_B(t)$, the total trapped hole charge in the barrier oscillates with amplitude that increases with bias and is given by

$$\delta Q_B(t) = 4\left|\frac{(\Delta - \Delta_c)}{\Delta_2}\right|^{1/2}\left[\frac{4G|\Delta_2|}{(1-G)^5|\{4+19G-8G^2\}|}\right]^{1/2} \times[\cos\Omega_T - \sin(\Omega_T + \Phi)] \qquad (135)$$

This increase in the amplitude of oscillation is expected since the maximum electric field in the depletion region, and hence the maximum e-h generation rate, by Zener tunneling (10), increases with bias.

In summary, we have verified from the analysis of our physical model that the average trapped hole charge in the barrier is approximately a slanted step function as a function of the applied biasing field in the depletion layer of the drain region determined by $\mathcal{G}$. Furthermore, we have also shown that the amplitude of the oscillation of trapped hole charge in the barrier is an increasing function of the biasing field in the range of physical interest, while maintaining a slowly decreasing average value.

Average Barrier and Quantum-Well Charges

Since common experimental techniques are not capable of proving this very high-frequency oscillation, we calculate the time-averaged quantities leading to a current-voltage I-V characteristic which can be compared with experimentally measured I-V. We will show that our theoretical results are in excellent qualitative agreement with the existing experiment. The time-averaged hole charge in the barrier is denoted as *Q(AlGaSb)*. This average value of the hole charge oscillation has been demonstrated previously to be a slowly varying function of the applied biasing field, after an abrupt increase at $(k_z^D)^2 \approx (k_F^D)^2$ in Fig. 15. We are thus guided to approximate a 'slanted' step-function behavior, $Q_h\Theta[(k_z^D)^2 - (k_F^D)^2]$ for the time-averaged hole charge of the barrier, where $Q_h \approx eQ_B$ [refer to Eq. (62), e is the unit charge which is positive definite since only holes are trapped. We shall see that indeed this leads to a "slanted parallelepiped" hysteresis in the I-V characteristics, in full qualitative agreement with an existing experiment.

The slanted step-function behavior of the time-averaged hole charge is a property of our physical model and strongly suggests the limit cycle operation of the experimental device. From the nature of Zener transition, higher bias and higher electric field would mean "faster passage in an avoided crossing region," (10) and hence larger Zener transition probability. Thus, if duon generation and accompanying oscillation is absent, *Q(AlGaSb)* is expected to monotonically increase with biasing field. Therefore, the most probable way for the hole charge to be slowly decreasing with bias as indicated by Eq. (132), is for it to be oscillating between "charged" and "discharged" state and hence to become strongly limited. It is the amplitude of this oscillation which can monotonically increase with bias, as we have demonstrated in the analysis of our physical model, and discussed in more detail later, while maintaining the average value to be slowly varying with bias. On the other hand, the self-consistency of the potential alone in Fig. 15, which must also be satisfied, demands that the polarization and hence *Q(AlGaSb)* increases monotonically with biasing field or $(k_z^D)^2$. We shall see that the simultaneous solution to these two requirements, plus the continuity condition, leads to a slanted parallelepiped hysteresis of the trapped hole charge in the barrier.

In order to describe the convex energy band-edge (EBE) profile in the quantum well we need a minimum of two values of fields, and this is also true for describing the concave EBE profile in the barrier. Therefore, we need four different field parameters in the theory. However, the requirement of a faster voltage drop, by virtue of the presence of hole charge, occurring in the barrier region allows us to use only three field parameters while still maintaining the concave EBE profile in the barrier region. The inflection point or the transition region from the convex to concave EBE profiles is located at the barrier wall of the conduction-band quantum well; for our purpose we assumed this region to have a measure zero as far as the integration of the fields is concerned to obtain the total voltage drop across the device. For nonzero average value of $\mathcal{N}_B$, which is the number of "unpaired" trapped holes, we also expect a nonzero average polarization between the barrier and spacer layer of the drain region to be affecting the potential profile, as indicated by a simple 'kink' in the spacer region of Fig. 15.

We only need to introduce a third field parameter, E_B, as shown in Fig. 15, instead of additional two field quantities to describe the concave EBE profile in the barrier. We can estimate the other field parameter in the second half of the barrier as proportional to E_R, as the figure suggests with proportionality factor (which may depend on the voltage) $\chi(V)$, and still maintain the physical requirement of concave EBE profile in this region. This takes into account the nonzero average $\mathcal{N}_B$, and the average presence of electrons in the spacer layer displaced from the barrier by Zener transition. Notice that the maximum concave curvature must be located at the middle of the barrier where the maximum density of highly confined holes occurs.

From Fig. 15, we can approximate the trapped hole charge in the second barrier by the expression: $\chi E_R - E_B = Q(AlGaSb)$, which follows from the Poisson equation. We estimate the proportionality factor χ is close to unity and positive. From the requirement of faster voltage drop in the barrier region in Fig. 15, we must have E_B more negative than χE_R, thus we obtain $Q(AlGaSb) > 0$ consistent with the trapped hole charge in the second barrier. Figure 2 shows schematically the conduction and valence band-edge profiles and serves to define the various quantities used in the present calculations.

Hysteresis of trapped hole charge in the barrier. The self-consistent calculation of the hole charge consists of two steps. The first step is to account for the work done on a positive charge e. This is related to the electric fields as follows: $Work = -e\int E \cdot ds$. With positive applied bias, V, and referring to Fig. 15, this translates to the following expression, with absolute value symbol explicitly indicated here:

$$eV = e|E_L|\left(b + \frac{w}{2}\right) + e|E_R|\frac{w}{2} + E_g + \frac{\hbar^2 (k_z^D)^2}{2m^*} - E_{hh} \qquad (136)$$

where w and b are the width of the quantum well and barrier, respectively. We now use the Poisson equation to eliminate $|E_R|$ in Eq. (136) in terms of *Q(AlGaSb)* and $|E_B|$. This means that $|E_R|$ of Eq. (136) will be absorbed in *Q(AlGaSb)*, which is determined self-consistently by the voltage drop in the left half of the barrier or $|E_B|$, as well as by $|E_L|$ which in turn determines Q_{w}. Since all fields on the average have negative sign for positive applied voltage, we may also write Poisson equation as

$$\begin{aligned} |E_R| - |E_L| &= \frac{|Q_w|}{\varepsilon} \\ |E_B| - \chi|E_R| &= \frac{Q(AlGaSb)}{\varepsilon} \end{aligned} \tag{137}$$

Therefore, we obtain the following expression for the trapped hole charge:

$$Q(AlGaSb) = \frac{2\varepsilon\chi}{ew}\left\{(E_g - E_{hh}) - \left[eV - e|E_L|\left(b + \frac{w}{2}\right)\right.\right. \\ \left.\left. -e|E_B|\frac{w}{2\chi}\right] + \frac{\hbar^2(k_z^D)^2}{2m^*}\right\} \tag{138}$$

Notice that the trapped hole charge is unambiguously an increasing function of $(k_z^D)^2$ in Eq. (138); the remaining terms are dependent on the applied voltage V and collectively serves as a parameter for a family of linear curves as a function of $(k_z^D)^2$.

The second step is to invoke the quantum transport phenomena of Zener tunneling. Following the preceeding discussions, this is expressed by

$$Q(AlGaSb) = Q_h[(k_z^D)^2]\Theta[(k_z^D)^2 - (k_F^D)^2] \tag{139}$$

where $Q_h[(k_z^D)^2]$ is a slightly decreasing function of average $(k_z^D)^2$ as dictated by our analytical result given by Eq. (132), which is plotted in Fig. 20(a). Here, we made the natural assumption that the average $(k_z^D)^2$ increases with the biasing field or $\mathcal{G}$. This assumption is consistent with a decreasing average trapped hole charge as a function of $\mathcal{G}$. The simultaneous solutions of Eqs. (138) and (139), incorporating the continuity condition, is shown graphically in Fig. 6(b), which yield the hysteresis of the trapped hole charge in the AlGaSb barrier. Note that the forward bias discontinuity is smaller than the reverse bias discontinuity in the hysteresis loop. This is the salient features of the present model which yield excellent qualitative agreement with experiment. The discussion given here concerning *Q(AlGaSb)* is more detailed than that in Refs. (25) and (26).

Hysteresis of the quantum-well charge. We give here the expression for the conduction-band quantum-well charge which determines the source-to-drain terminal current of the device. This is obtained by describing the whole length of the device by three independent fields, namely, E_{L}, E'_R, and E'_B. Note that the field χE_R used before is only valid in the right-half barrier region, by virtue of nonzero average number of unpaired trapped holes, $\mathcal{N}_B$, which may lead on the average to a polarization between the right-barrier edge and spacer-layer of the drain, as indicated in Fig. 15. The field E_{L} is as defined before, whereas E'_R is defined by the relation: $E'_R(b/2 + w/2) = E_R(w/2) + E_B(b/2)$, and E'_B is the constant-field approximation for the rest of the region of dimension $[(b/2) + c]$. As a consequence, we arrive at the following relation: $|E'_R| - |E_L| = \dfrac{\tilde{\alpha}|Q_w|}{\varepsilon}$ and $|E'_R| - |E'_B| = \dfrac{\tilde{\beta}|Q(AIGaSb)|}{\varepsilon}$, where $\tilde{\alpha}$ and $\tilde{\beta}$ are the proportionality constants. Thus, we may write for our present purpose

$$eV = e|E_L|\left(b + \frac{w}{2}\right) + e|E'_R|\left(\frac{b}{2} + \frac{w}{2}\right) \\ + e|E'_B|\left(\frac{b}{2} + c\right) \tag{140}$$

We then express $|E'_R|$ and $|E'_B|$ in terms of $|Q_w|$, ***Q(AlGaSb)***, and $|E_L|$. We also use the relation: $e|E_L| = [e(V - \phi_c)]/(2b + w)$, where $e\phi_c = eV - E_w + \dfrac{\hbar^2(k_z^s)^2}{2m^*}$ to obtain the final result

$$|Q_w| \\ = \frac{2\varepsilon}{\tilde{\alpha}[(b + c) + w/2]}\left\{\frac{\hbar^2(k_z^s)^2}{2m^*} - E_w + \frac{V}{2\xi} + \frac{\tilde{\beta}\,Q(AIGaSb)}{2\xi\left[\varepsilon/\left(\frac{b}{2} + c\right)\right]}\right\} \tag{141}$$

where we have $\xi = [2b + w + c]/(2b + w)$. Equation (141) is the self-consistency requirement for $|Q_w|$. The quantum transport requirement for Q_{w} was given by Buot and Rajagopal (25, 26) as

$$|Q_w((k_z^s))| \\ = \frac{em^*kT}{\pi\hbar^2}\,ln\left[1 + \exp\frac{1}{kT}\left(E_F - \frac{\hbar^2(k_z^s)^2}{2m^*}\right)\right]\left(\frac{\tau_d}{\tau_e}\right)\Theta((k_z)^2) \tag{142}$$

where $\dfrac{1}{\tau_d} = \dfrac{1}{\tau_e} + \dfrac{1}{\tau_c}, \dfrac{1}{\tau_c}$, is the effective rate of decay of Q_{w} into unoccupied collector states, $\dfrac{1}{\tau_e}$ is equal to the rate of supply of electrons from the emitter to the quantum well, and this can be assumed to be equal to the rate of the reverse process only for the near-equilibrium situation. Under steady state at significant bias, it is more appropriate to approximate $\tau_d = \tau_c$. The simultaneous solution of Eqs. 141 and 142, making use of the graphical solution of *Q(AlGaSb)*, is also graphically obtained as shown in Fig. 21.

Discussions

Average value of the oscillating current and I-V hysteresis. The RTD current can he approximated by Q_w/τ_c. One can immediately see that the resulting I-V characteristics have all the salient features of the experimental results (34, 35), exhibiting the slanted parallelepiped hysteresis inn the I-V characteristic before the current peak, Fig. 7(b). Because of the result obtained in Eq. (132), plotted in Fig. 20(a), concerning the decrease of the trapped hole charge with biasing field, the I-V hysteresis is more accurately shown in this article to have a smaller higher-current offset at forward-bias compared to the lower-current offset at reverse-bias. Note that the sensitivity of the time-averaged

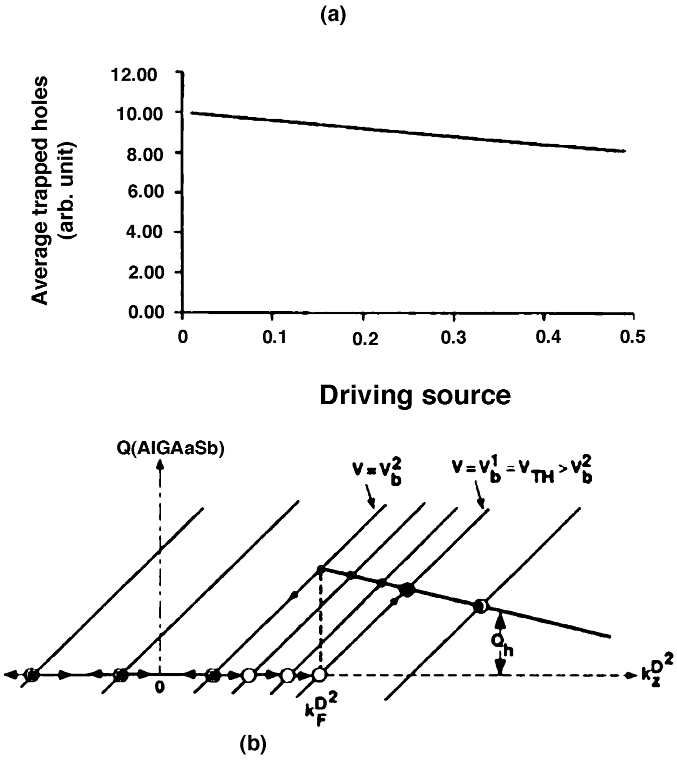

Figure 20. (a) Average trapped holes in the barrier, βQ_B of Eq. (131), as a function of $\mathcal{G}$ (proportional to applied bias). Note that the negative slope increases with decrease in Δ while the values of the total trapped holes increase. For the value of $\Delta = 0.01$, the $Tr(M) > 0$ region covers the range for $\mathcal{G} < 0.5$. (b) Graphical solution of Eqs. (138) and (139). Equation (138) for *Q(AlGaSb)* vs. $(k_z^D)^2$ is approximated by positive sloping lines, and Eq. (139) as a step function with value $Q_h[(k_z^D)^2]$ for $(k_z^D)^2 > (k_F^D)^2$. By applying the continuity condition, open circles and solid circles are solutions for the increasing and decreasing voltage sweep, respectively. A slanted parallelepiped hysteresis of trapped hole charge is clearly shown. [From Ref. 32 with permission].

hole charge, Q_{B}, to the applied bias is dictated by our physical model and limit cycle analysis. Specifically, for nonoscillatory case, the trapped hole charge increases nonlinearly with applied biasing field or $\mathcal{G}$, and is eliminated by the experiment since this would lead to a larger current offset at forward bias than for reverse bias contrary to the experimental I-V characteristics. Indeed, the analytical I-V result under limit cycle operation is in complete qualitative agreement with this one important salient feature of the hysteresis of the experimental I-V characteristics (35), which is reproduced in Fig. 21(c). Thus, we have experimental evidence indicating the correctness of our approach and the promising potential of this nanodevice as a novel all solid-state THz source.

Amplitude of the fundamental oscillation. Similarly, from Eq. (134), the amplitude $\mathcal{A}_o$ of the fundamental oscillation of trapped holes is given by

$$\mathcal{A}_o = \frac{1}{\beta}4\left|\frac{(\Delta - \Delta_c)}{\Delta_2}\right|^{1/2}\left[\frac{4\mathcal{G}|\Delta_2|}{(1-\mathcal{G})^5|\{4 + 19\mathcal{G} - 8\mathcal{G}^2\}|}\right]^{1/2} \qquad (143)$$

Note that the magnitude of $\frac{1}{\beta}$ is generally large, indicating that the amplitude of oscillation of the trapped holes can be quite significant. This could provide for a good charge control of the quantum-well resonant energy level yielding useful power at THz frequencies. Using our estimate of $\frac{1}{\beta} \approx 10^{11} cm^{-2}$, then $\mathcal{A}_o \approx (10^8 - 10^9) cm^{-2}$ or $\mathcal{A}_o \approx (10^{-11} - 10^{-10}) C \quad cm^{-2}$. Using a value of the capacitance C of about $10^{-10} F$, we see that corresponding voltage modulation can be on the order of 1.0 V.

The oscillatory behavior allows the oscillation amplitude to grow with the applied bias, in response to the increasing maximum electric field in the depletion region of the drain, while maintaining a slowly decreasing time-average value. Figure 22 shows the variation of amplitude as a function of $\mathcal{G}$.

Frequency of the fundamental oscillation. The fundamental frequency of oscillation is given by $w_o = \alpha\Omega$, where the α comes from the conversion of τ to real time. From Eq. (128), we have

$$w_o = \alpha[\sqrt{\mathcal{G}}(1-\mathcal{G})^{3/2} + [Im\,\eta(\Delta - \Delta_c) + \frac{Im\,\sigma}{\Delta_2}|\Theta(\infty)|^2(\Delta - \Delta_c)]] \qquad (144)$$

where $\alpha = \lambda N$. Using estimates for the valence band density of states and tunneling rates, the first term of Eq. (144) is estimated to be in the THz range of frequency. Using our conservative estimate for $\alpha \approx 10^{13} s^{-1}$ and $\mathcal{G} \approx 10^{-2}$, we obtain $w_o = 10^{12}$ for the leading term of Eq. (144). Figure 23 shows the variation of frequency as a function of $\mathcal{G}$. Comparing with Fig. 22, we see that there exists an optimum $\mathcal{G}$

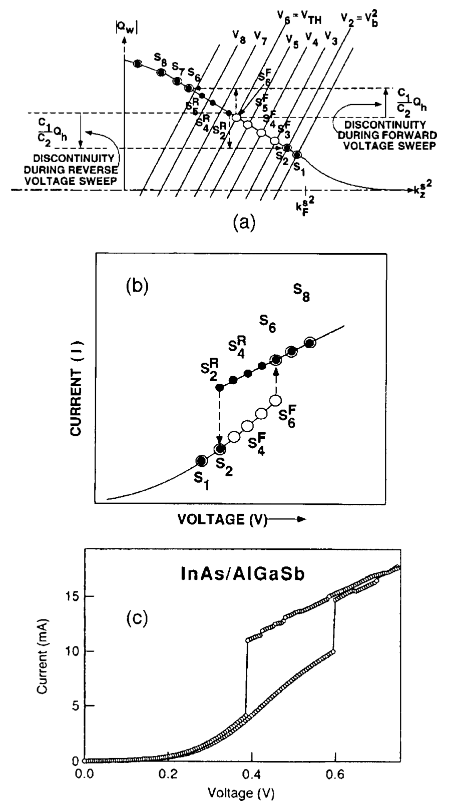

Figure 21. (a) Graphical solution of Eqs. (141) and (142). Equation (141) is approximated by parallel sloping lines, and Eq. (142) has a value zero for $(k_z^s)^2 \leq 0$. The solutions for *Q(AlGaSb)* are obtained from Fig. 20 and used here to create an offset in the sloping lines of Eq. (141), leading to solutions at higher values of Qw as indicated by the dotted arrows. For the increasing voltage sweep, the solutions for Q_{W} are given by the intersection points S_1, S_2, S_3, S_4, S_5, S_6 (low), S_6 (high), S_7, and S_8. For the reverse voltage sweep, the corresponding solution points are S_8, S_7, S_6, S_5^R, S_4^R, S_3^R, S_2^R [S_2(high)], S_2 [S_2(low)], and back to S_1. Note that the discontinuity at forward bias, where $Q_h \equiv Q_h(V_{TH})$ is less than that of reverse bias where $Q_h \equiv Q_h(V_b^2)$ in accordance with Fig. 20(a). (b) Solution for the *I-V* characteristic showing slanted parallelepiped hysteresis occurring before the RTD current peak, in agreement with the experiment. (c) Experimentally measured *I-V* characteristics of *AlGaSb/InAs/AlGaSb* double-barrier structure (from Ref. (35), Fig. 2(a), reprinted by permission).

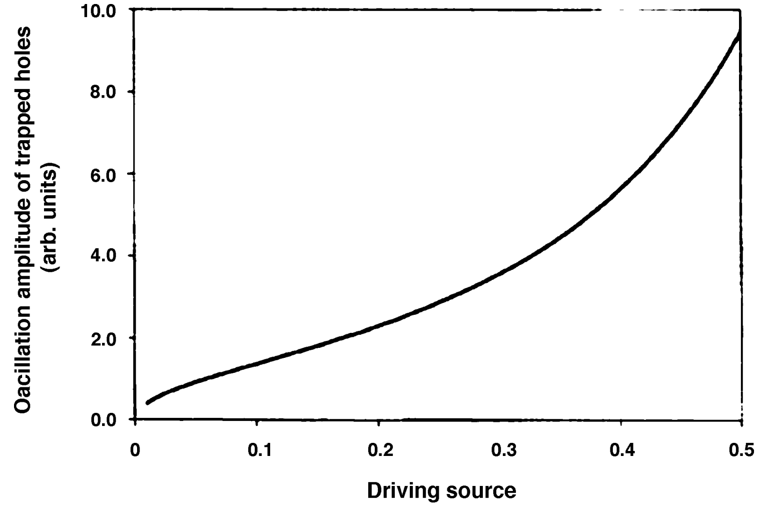

Figure 22. Plot of the amplitude of oscillation, square root of the expression within the bracket of Eq. (135), as a function of driving source $\mathcal{G}$ (applied bias)

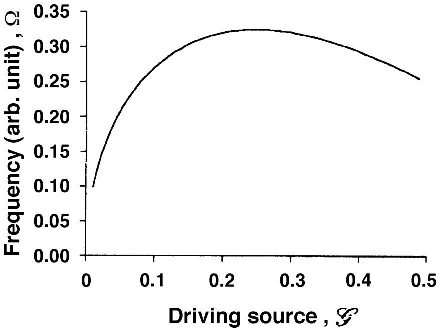

Figure 23. Plot of the frequency of oscillation, which is the leading term of Eq. (128), as a function of the driving source $\mathcal{G}$.

for a given rate parameter Δ, where the frequency and amplitude are optimized for maximum power output at THz frequency.

Figures 22 and 23 also point out a very attractive and unique feature of the device, namely, for some range of biasing conditions (physical range of $\mathcal{G}$), both the frequency and power delivered (amplitude of oscillation) increase with $\mathcal{G}$. Thus, one can optimize the device for the highest frequency at maximum power of operation.

Equivalent circuit model and anticipated THz power output. To appreciate the compact nature of the solid-state THz source, it is helpful to represent the proposed THz source in terms of its equivalent circuit, using the combination of familiar circuit elements. This is represented in Fig. 24, by a circuit consisting of one bipolar transistor, in combination with a capacitor and resistors to form a relaxation type oscillator. The capacitor is used to represent the generation of charge polarization pair or duon. The transistor represents the RTD with staggered band-edge alignment with the capability to trap holes which induces the conduction of current across the RTD. The associated physical process responsible for the high-frequency modulation of the current across the device is depicted in Fig. 16.

In Fig. 24, the capacitor charges (by virtue of polarization pair formation when Zener tunneling occurs) until the transistor conducts via the induced current which goes to a maximum value. At this point, the capacitor discharges rapidly into the transistor decreasing the "base" voltage. When the discharge of C has lowered the emitter-to-base voltage, the transistor is cut off (i.e., the induced current is cut off). The cycle is then repeated after a charge time determined by R_e, and C. Narrow pulses are thus available across the load R_L, coupled through some form of lossless resonator matching circuit. This load may be a circuit impedance used to drive an ultrafast timing circuitry in a computer, or an equivalent radiation resistance of a dipole (patch) antenna. The voltage across C is like a sawtooth-like pulse by virtue of the rapid discharge after a slower buildup of charge.

Note that in our proposed device the charging of the capacitance is through the Zener tunneling instability which is acting in parallel throughout the whole device area, rendering the charging time to be independent of the device area. This is expected to yield a very effective and simple power combining scheme. We can estimate the anticipated THz power output of our proposed THz source by observing that the polarization voltage is roughly in phase with the fluctuating current across the device. We estimated the current amplitude to be around $1.0 \times 10^5\ A/cm^2$, and the polarization voltage amplitude of about 0.3 V. These yield an anticipated THz power output of $1.5 \times 10^4\ W/cm^2$ or $15mW$ of power for a $10\mu m \times 10\mu m$ device area. To realize this often requires that some form of resonant guiding structure/antenna will be integrated with the proposed device to minimize losses (41). Electrically pumped THz sources (e.g., high performance GUNN diodes) operating tip to mW power levels have not been demonstrated (41) (also note that for the optically excited Bloch oscillator, the experimental emitted power is only on the order of nano- to microwatts (42)).

Advantages over conventional RTD circuit-based THz sources. The interband or Zener tunnelling high-frequency (ZTHF) source possess definite advantages when compared to the conventional intraband resonant-tunneling diode (CIRTD) oscillator source. It should be noted that CIRTDs have so far only been utilized as one of the components in a traditional two-terminal oscillator source circuit. Specifically, one biases the CIRTD in its NDR region. Here, any noise fluctuations are amplified when the CIRTD device and its inherent charge storage capacitance resonates in an unstable manner with some external charge storage element (e.g., inductance of the contact lead). The instability conditions necessary for operation of the CIRTD-based oscillator are intrinsically tied to the fact that the current and voltage oscillations are made completely out of phase by the NDR action. For example, at any constant bias within the NDR region a small decrease in voltage (i.e., due to noise fluctuations) leads to an increase in the current which acts to charge the device capacitance. During the next cycle the external element (inductor) is in turn charged by the discharging device capacitance. Over many successive cycles the gain, associated with the NDR, continues to increase the amplitude of the oscillation until a limit cycle is reached at the edges of the NDR.

The point is that CIRTD acts as an unstable gain mechanism and oscillations are produced by resonating with an external element. Hence, the energy associated with the oscillation must pass through a physical contact which will always possess some loss. Even more important, the CIRTD) is inherently unstable over a broad bandwidth. Specifically, the NDR of the CIRTD will encourage oscillation in the biasing circuitry down to zero frequency. There-

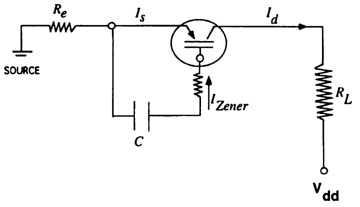

Figure 24. Equivalent-circuit model of an ITHS as a relaxation oscillator circuit with a common source.

fore, one must design the circuit coupled to the CIRTD-based oscillator so that it is low-frequency stable to prevent energy losses to lower-frequency modes. This requires the designer to reduce the CIRTD capacitance. This can only really be accomplished by reducing the area of the diode. This reduction in CIRTD device area severely limits the available output power of the oscillator at very high frequencies.

In contrast, the ZTHF device acts as a stand-alone high-frequency "generator" of nanometric dimensions in a "monolithic" construction. The fundamental physical principles upon which the ZTHF functions eliminates the previously described limitations associated with the CIRTD-based oscillator. Specifically, the ZTHF operates by charging and discharging the localized well regions separated by mesoscopic distances inside the semiconductor device. The process of charging the conduction-band quantum well in sub-picosecond time scales allows one to trigger a discharging of the valence-band quantum well. This process is equivalent to a mesoscopic scale feedback. Here, the real advantages are that this feedback is accomplished internally (therefore avoiding the loss and charging effects associated with external contacts) and that it can be done outside the NDR region, an operating region which in the CIRTD-based oscillator tends to amplify noise and subharmonics.

The only other known single-element high-frequency source device which exhibits this type of behavior (i.e., intrinsic oscillatory mechanism) are the IMPATT and GUNN diodes. However, the IMPATT and GUNN oscillators function by propagating a charge dipole domain along a transit channel. Hence, the IMPATT and GUNN oscillation frequency is inversely proportional to device length which can only be feasibly reduced to a minimum limit, much larger compared to the dimensional features of ZTHF source, to achieve subterahertz (i.e., approximately 100 GHz) performance. In contrast to IMPATT and GUNN devices, the ZTHF device operates on a fixed but oscillatory charge dipole domain which acts as a "gate signal" for modulating the conduction-band current of the CIRTD operation, without bringing the CIRTD operation into the NDR region.

Compared to other solid-state THz sources driven by optical lasers, the proposed device has a clear advantage in terms of simplicity, compactness, and monolithic integration capabilities with power combiners, matching/guided wave structures, and antennas. The proposed device also promises to yield much larger THz power output.

Hence, the ZTHF source is a potentially very important novel device because it can function on picosecond time scales, it is not a broad-bandwidth oscillator, and unlike CIRTD is not plagued by low-frequency stability constraints. We have performed preliminary numerical calculations and have found that, at reasonable dc biasing, significant levels of hole trapping results. One can modulate the conduction-band well energy-level by several meV by interband tunneling currents. This is important because it means that the conduction band current can be altered significantly ($\approx 50\%$ modulation) by the valence band current on THz frequency time scales.

Summary. In summary, the dynamical behavior of a coupled system of duon and unpaired trapped hole charge in the RTD with staggered band-gap alignment has been shown to give the fundamental physical explanation of the experimental I-V characteristic of $AlGaSb/lnAs/AlGaSb$ double-barrier structure. The stimulated production of duons and Zener tunneling of electrons leads to an autonomous control of the position of the energy level of the quantum well. The self-oscillatory character of the trapped hole charge provides the physical control mechanism behind a novel interband-tunnel high-frequency-source RTD device. To calculate the oscillating current, one simply replaced the average trapped charge *Q(AlGaSb)* in Eq. (141) with the oscillating trapped charge expression determined by Eqs. 131 and 135. This oscillation is useful for various high-bandwidth applications, well beyond the range of applications of the traditional IMPATT and Gunn effect devices. In practice, for a given material parameter Δ, we choose the operating bias, directly related to $\mathcal{G}$ such that $Tr(M) > 0$. For a range of $\mathcal{G}$ where this is satisfied, we can optimize the operating point $\mathcal{G}$ to realize a THz source with optimum power and frequency. Since in general the duon formation is a higher-order process involving Zener tunneling coupled with resonant tunneling to form a polarization pair, we expect Δ to be considerably less than one, as discussed before. Thus, in a realistic device $Tr(M) > 0$ should easily be satisfied.

The hysteresis of the trapped-hole charge in the$AlGaSb$ barrier plays a crucial role in the hysteresis of the I-V curve to occur before the current peak (25, 26). This is in sharp contrast to the I-V hysteresis and bistability commonly observed in RTD with conventional band-edge alignment where the I-V hysteresis occurs after the current peak, at the NDR region. The latter is caused by a combination of factors, namely, the nonlinear series resistance, quantum-well charge storage, and quantum inductance (33, 26).

Since the high-field domain associated with duon formation acts to modulate the resonant energy level in the conduction-band quantum well operating near and before the 'conventional' resonant-current peak, the transconductance of this "self-gated" structure can be quite large, yielding a novel high-frequency source with considerable power. Basically, the high-field domain acts as a self-gate of the built-in-transistor oscillator. These oscillations are expected to occur in the THz range. Application in communication and sensors are possible niches, and perhaps another, with the incorporation of defect engineering, as a triggering element in semiconductor lasers. Advances in power combining techniques using microfabrication give more grounds for the need of a serious research and development efforts on the RTD as potential solution for all solid-state and compact THz sources.

The theoretical technique used here have also been applied to the analysis of the self-oscillating behavior of conventional RTDs (27), also in full agrement with large-scale time-dependent simulation and experiments. Our present results qualitatively agree with the salient features of the experimental measurements of the I-V characteristic of a *AlGaSb / lnAs / AlGaSb* staggered band-gap double-barrier structure. However, exact numerical results call for large-scale time-dependent numerical simulation of multiband quantum transport equations which account for interband tunneling (31).

I would like to thank Dr. Dwight Woolard for some helpful discussions on the advantages over conventional RTD circuit-based THz sources.

OTHER RESEARCH ON LZS EFFECT

For those interested in more advanced work on LZS phenomena, there are several theoretical and experimental reports dealing with LZS effect in the presence of dissipation. For a good overview of the theoretical work, readers are referred to Ao and Rammer (13), and the monograph of Leggett et al. (30)). Experimental and theoretical work on superlattice heterostructures (2) and mesoscopic metal rings in also being pursued (31). There exist more mathematical treatments of Stark–Wannier energy levels and Zener tunneling; interested readers are referred to Refs. 32, 33.

ACKNOWLEDGMENT

The author is grateful to Dr. A. K. Rajagopal, for his critical reading of the manuscript and helpful comments.

BIBLIOGRAPHY

1. G. H. Wannier, Dynamics of band electrons in electric and magnetic fields, *Rev. Mod. Phys.*, **34**: 645–655, 1960.
2. C. Hamaguchi *et al*., Wanner-Stark localization in superlattices, *Jpn. J. Appl. Phys.*, **34**: 4519–4521, 1995.
3. S. M. Sze, *Physics of Semiconductor Devices*, New York: Wiley, 1981.
4. G. H. Wannier, Probability of violation of Ehrenfest principle in fast passage, *Physics*, **1**: 251–253, 1965.
5. D. R. Fredkin and G. H. Wannier, Theory of tunneling in semiconductor junction, *Phys. Rev.*, **128**: 2054–2061, 1962.
6. L. V. Keldysh, Behavior of non-metallic crystals in strong electric fields, *Sov. Phys. JETP*, **6**: 763–770, 1958.
7. R. A. Logan, Phonon-assisted semiconductor tunneling, in E. Burstein and S. Lundqvist (eds.), *Tunneling Phenomena in Solids*, New York: Plenum Press, 1969.
8. S. Mil'shtein, D. Karas, and C. Lee, Secondary electron imaging of metal-semiconductor field-effect transistor operation, *J. Vac. Sci. Technol.*, **B14**: 437–439, 1996.
9. C. Zener, Non-adiabatic crossing of energy levels, *Proc. Roy. Soc. London*, **A137**: 696–702, 1932.
10. C. Zener, A theory of dielectric breakdown of solid dielectrics, *Proc. Roy. Soc. London*, **A145**: 523–529, 1934.
11. F. A. Buot, P. L. Li, and J. O. Strom-Olsen, The influence of scattering on magnetic breakdown, *J. Low Temp. Phys.*, **22**: 535–556, 1976.
12. L. Esaki, New phenomenon in narrow germanium p-n junction, *Phys. Rev.*, **109**: 603–604, 1958.
13. P. Ao and J. Rammer, Quantum dynamics of a two-state system in a dissipative environment, *Phys. Rev.*, **B43**: 5397–5418, 1991.
14. E. O. Kane and E. I. Blount, Interband tunneling, in E. Burstein and S. Lundqvist (eds.), *Tunneling Phenomena in Solids*, New York: Plenum Press, 1969.
15. M. Tinkham, *Group Theory and Quantum Mechanics*, New York: McGraw-Hill, 1964.
16. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, New York: Academic Press, 1965.
17. J. R. Rubbmark *et al*., Dynamical effects at avoided crossings: A study of Landau-Zener effect using Rydberg atoms, *Phys. Rev.*, **A23**: 3107–3117, 1981.
18. K. L. Jensen and F. A. Buot, Numerical simulation of intrinsic bistability and high-frequency current oscillations in resonant tunneling structures, *Phys. Rev. Lett.*, **66**: 1078–1081, 1991.
19. B. A. Biegel and J. D. Plummer, Comparison of self-consistency iteration options for the Wigner function method of quantum device simulation, *Phys. Rev.*, **B54**: 8070–8082, 1996.
20. E. T. Yu. J. O. McCaldin, and T. C. McGill, in *Solid-State Physics, Advances in Research and Applications* (Academic, San Diego, 1992).
21. D. Z. Y. Ting, E. T. Yu, and T. C. McGill, Multiband treatment of quantum transport in interband tunnel devices, *Phys. Rev.* **B45**: 3583–3592,1992.
22. T. C. L. G. Sollner *et al*., Resonant tunneling through quantum wells at frequencies up to 2.5 THz, *Appl. Phys. Lett.*, **43**: 588–590,1983.
23. F. A. Buot and A. K. Rajagopal, High-frequency behavior of quantum-based devices: Equivalent circuit, nonperturbative response, and phase-space analyses, *Phys. Rev.*, **48**: 17217–17232, 1993.
24. F. A. Buot and K. L. Jensen, Intrinsic high-frequency oscillations and equivalent circuit model in the negative differential resistance region of resonant tunneling devices, *Int. J. Comp. Math. Elec. Electron. Eng. COMPEL*, **10**: 241–253, 1991.
25. F. A. Buot and A. K. Rajagopal, Hysteresis of trapped charge in AlGaSb barrier as a mechanism for the current bistability in AlGaSb/InAs/AlGaSb double-barrier structures,*Appl. Appl. Lett.*, **64**: 2994–2996, 1994.
26. F. A. Buot and A. K. Rajagopal, Theory of novel nonlinear quantum transport effects in resonant tunneling structures, *Mater. Sci. Eng.*, **B35**: 303–317, 1995.
27. F. A. Buot, P. Zhao, H. L. Cui, D. Woolard, K. L. Jensen, and C. M. Krowne "Emitter Quantization and Double Hysteresis in Resonant Tunneling Structures: A Nonlinear Model of Charge Oscillation and Current Bistability", *Phys. Rev.* **B61**, 5644–5665 (2000) .

28. F. A. Buot, General theory of quantum distribution function transport equations: superfluid systems and ultrafast dynamics of optically excited semiconductors, *La Rivista del Nuovo Cimento* **20**: No. 9, 1–75, 1997.
29. F. A. Buot, Foundation of Computational Nanoelectronics, in *Handbook of Theoretical and Computational Nanotechnology*, edited by M. Rieth, and W. Schommers, American Scientific Publishers, 2006, Vol.**1**
30. F. A. Buot, Generalized semiconductor Bloch equations, *J. Comp. Theor. Nanoscience* **1**: 144–168, 2004.
31. F. A. Buot, On the theory of novel solid-state teraherz souces: Renormalization and Bloch equations, *J. Comp. Theor. Nanoscience* **3**: 712–726, 2006.
32. F. A. Buot and C. M. Krowne, Double-barrier THz source based on electrical excitation of electrons and holes, *J. Appl. Phys.* **86**: 5215–5231, 1999. See also *J. Appl. Phys.* **87**: 3169, 2000.
33. D. L. Woolard *et al*., On the different roles of hysteresis and intrinsic oscillations in resonant tunneling structures, *J. Appl. Phys.*, **79**: 1515–1525, 1996.
34. F. A. Buot, *J. Phys. D: Appl. Phys.* **30**: 3016, 1997; *VLSI Design* **8**: 237, 1998.
35. D. H. Chow and J. N. Schulman, *AppL Phys. Lett.* **64**: 76, 1994.
36. H. W. Yoon and L. N. Pfeiffer. *Bull. Am. Phys. Soc.* **41**: 239, 1996.
37. See for example, S. Paddeu, V. Erokhin, and C. Nicolini, *Thin Solid Films*, **284–285**: 854, 1996.
38. A. Pimpale *et al*., Limit cycle in a bound exciton recombination model in non-equilibrium semiconductors, *J. Phys. Chem. Solids,* **42**: 873–881, 1981.
39. A. H. Nayfeh, *Introduction to Perturbation Techniques*, New York: Wiley, 1981.
40. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953, pp. 884–886.
41. N. J. Cronin, *Philos. Trans. R. Soc. London, Ser.* **A 354**: 2425, 1996.
42. K. Leo, *Semicond. Sci. Technol.* **13**, 249 (1998) .
43. C. Kidner, I. Mehdi, J. R. East, and G. I. Hadad, *IEEE Trans. Microwave Theory Tech.* **38**: 864, 1990.
44. A. J. Leggett *et al*., Dynamics of dissipative two-state system, *Rev. Mod. Phys.*, **59**: 1–85, 1987.
45. G. Blatter and D. A. Browne, Localization in small metal rings: Current saturation without dissipation, Physica Scripta, **T25**: 353–356, 1989.
46. G. Nencio, Dynamics of band electrons in electric and magnetic fields: rigorous justification of effective Hamiltonians, *Rev. Mod. Phys.*, **63**: 91–127, 1991.
47. V. Grecchi and A. Sacchetti, Crossings and anticrossing of resonances: The Wanner-Stark ladders, *Ann. Phys.*, **241**: 258–284, 1995.

Felix A. Buot
Naval Research Laboratory,
Washington, DC