Extended-range tight-binding method for tunneling

J. N. Schulman
Hughes Research Laboratories, Malibu, California 90265

D. Z.-Y. Ting
Thomas J. Watson, Sr. Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125
(Received 29 March 1991)

The calculation of multiband tunneling in semiconductor multilayer structures, including resonant tunneling structures, has been hampered by computational problems such as limitations of numerical precision or excessive requirements of computer time. Recently, a few techniques have appeared which help eliminate these limitations. Here we present a related method and its derivation that solves the tight-binding equations for tunneling problems.

Modern interest in tunneling in multilayer semiconductor structures was stimulated by the work of Chang, Esaki, and Tsu in 1974 (Ref. 1) and clearly demonstrated by Sollner *et al.* in 1983.[2] This work, and most work for several years thereafter, was concentrated on conduction-band tunneling in GaAs/Ga-Al-As based systems. For this system, the simple one-band effective-mass model was adequate to describe the main features of the tunneling.[3]

Soon thereafter, interest grew in understanding other aspects of resonant tunneling. Mendez *et al.*[4] observed resonant tunneling due to holes in GaAs/Ga-Al-As double-barrier structures. Mendez, Calleja, and Wang[5] and Bonnefoi *et al.*[6] found indications of tunneling through X-point related conduction-band states, also in GaAs/Ga-Al-As structures. Subsequently, resonant tunneling of holes in Si/SiGe structures was detected by Rhee *et al.*[7] and Liu *et al.*[8] Recently, tunneling in the InAs/GaSb/AlSb group of tunnel structures has shown indications of providing higher current densities and peak-to-valley ratios than previous materials.[9,10] This system, which has a type-II band lineup, involves tunneling in which the carrier converts from electronlike to holelike depending on which layer it is in. In addition to these continuum types of tunneling cases, there is substantial interest in tunneling between confined states in coupled quantum wells, both due to conduction- and valence-band tunneling.

It was realized early on that the simple one-band model was inadequate to deal with these intrinsically multiband situations. An early attempt to deal with the multiband scattering problem was by Osbourn and Smith.[11] They solved the problem at a single interface in the context of the tight-binding model with complex-wave-vector bulk states, and calculated transmission and reflection coefficients at one interface. Schulman and Chang developed a more efficient technique for calculating the complex states and showed how to extend the tight-binding model to calculate tunneling through multilayer structures using tight-binding transfer matrices.[12] Calculations of multiband tunneling within the tight-binding, $\mathbf{k}\cdot\mathbf{p}$, and pseudopotential frameworks were carried out by a number of groups and explained several important features of hole tunneling and tunneling involving conduction X-point related states.

While all of the models had some success, they shared a common problem. The repeated multiplication of transfer matrices resulted in a loss of numerical precision, thus limiting the width of the structures that could be modeled. The limitation was most severe for the more complete models which included more bands. This was because the more complete models included larger values of the imaginary wave vectors of the complex bulk states. Since the eigenvalues of the transfer matrices were essentially the exponential of the wave vectors, the tranfer matrices in these models were more ill behaved.

There have been three solutions presented to date to these problems. One solution was that of Ko and Inkson.[13] Their technique involves rearranging the transfer matrix to explicitly separate the growing and decaying complex bulk states, at every transfer step. This was very effective and allowed transfers over regions large enough to model realistic structures. Since the model is based on extended states at every transfer step, in contrast to local orbital type models, it is most appropriate for $\mathbf{k}\cdot\mathbf{p}$ and pseudopotential frameworks. As it is, the model involves inversions and multiplications of matrices with dimensions equal to the size of the bulk basis set at every step. Although somewhat time-consuming, the method is probably the best way to deal with the problem within these two models. It can be adapted for local orbital models, but this would involve the transformation between the local orbital basis and the extended state basis at every step, an even more time-consuming process.

Very recently, two solutions have been presented appropriate for local orbital models. The first involves the creation of a large band matrix whose bandwidth is on the order of the number of basis orbitals per unit cell (times a small factor depending on the orientation of the layers and the number of neighbors included in the

model). This method was developed by Ting, Yu, and McGill[14] based on a suggestion by Frensley[15] to adapt the waveguide technique of Lent and Kirkner[16] to the multiband problem. The dimension of the matrix is equal to the number of basis orbitals times the total number of layers plus a few more to describe the boundary conditions on either side of the structure. The matrix represents the coefficients in a set of linear equations. Although the matrix is very large, the fact that it is a band matrix makes it necessary to store only the nonzero elements and there is software available that can solve it very efficiently.

Independently a second model was developed by Boykin, van der Wagt, and Harris.[17] It also involves the creation of a large matrix, but the matrix size is reduced by a factor equal to the number of layers that can be transferred over before numerical-precision problems result. This reduction is done by using the tight-binding transfer matrices to relate the coefficients of the orbitals within these subregions. Since, depending on the model, this can be a significant factor, their matrix can be much smaller than that of Ting, Yu, and McGill. For example, for a nearest-neighbor zinc-blende s^*sp^3 tight-binding model the factor can be about 30.

There are two tradeoffs for this improvement. First, the matrix is not a band matrix and therefore cannot make use of the special software which greatly speeds up the solution of the band matrix problem. Second, while there is some dependence on the computer and the software used, the time gained by decreasing the size of the matrix is approximately compensated by the time required to multiply the transfer matrices together. This is discussed further below.

Here we present a third model which is closely related to the second of the above two models. It also uses transfer matrices to reduce the overall size of the matrix. However, the matrix is a band matrix, as in the first method. The derivation presented below is somewhat more intuitive than that presented in Ref. 17, and brings out the band nature of the matrix in a simple manner. However, the result is almost the same, and a simple reordering of the basis in Ref. 17 would have produced the same result as given below.

The present method follows as an extension of the original tight-binding tunneling method and it is necessary to briefly review it here.[12] In that paper the structure was divided into three regions as shown in Fig. 1. The carrier is incoming from the right, region III. The region of varying composition and potential is region II. The transmission occurs in region I. The coefficients of the bulk complex-wave-vector states that make up the total wave function are called f^{I} and f^{III} in those regions. The f's are organized so that the top half of each includes the coefficients of the rightward propagating real states and of the complex states that exponentially decay to the right. The bottom halves are for the leftward propagating real states and complex states that decay to the left. This can be written $f^{\mathrm{I,III}} = (f_+^{\mathrm{I,III}}, f_-^{\mathrm{I,III}})$. f^{I} and f^{III} are related by the total transfer matrix

$$f^{\mathrm{III}} = M f^{\mathrm{I}}. \tag{1}$$

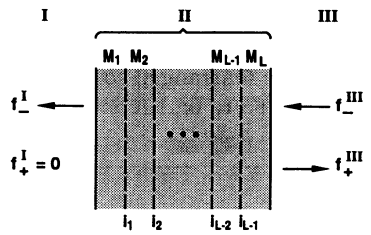

FIG. 1. Diagram of multilayer regions. The carrier is incoming and reflected in region III. Region II is the varying composition and/or potential region. The transmitted carrier is in region I. Region II is divided into L subregions traversed by the M^L matrices.

The incoming real state is a given and the exponentially growing states on the right must have zero coefficients, so f_-^{III} is known. Similarly f_+^{I} represents the states exponentially growing to the left and it must be zero. Equation (1) must then be solved for f_+^{III} and f_-^{I}. This was done by writing

$$M = \begin{pmatrix} M_{++} & M_{+-} \\ M_{-+} & M_{--} \end{pmatrix} \tag{2}$$

with the result

$$\begin{aligned} f_-^{\mathrm{I}} &= (M_{--})^{-1}(f_-^{\mathrm{III}} - M_{-+} f_+^{\mathrm{I}}), \\ f_+^{\mathrm{III}} &= M_{+-} f_-^{\mathrm{I}} + M_{++} f_+^{\mathrm{I}}. \end{aligned} \tag{3}$$

The loss of numerical precision occurs in the formation of the matrix M, which can be written

$$M = (S^{\mathrm{III}})^{-1} U S^{\mathrm{I}}. \tag{4}$$

U is the multiplication of the transfer matrices over all of the N layers in region II plus one, to get over both bordering interfaces:

$$U = T_{N+1} T_N, \ldots, T_2 T_1. \tag{5}$$

S^{I} and S^{III} are the matrices which give the local orbital coefficients of the complex bulk states. They are matrices formed by putting the local orbital coefficients of the complex bulk states into columns.

In this method M is split into a product of L matrices, each one of which is a product of a chain of the T matrices whose number is small enough so that precision is not lost, as in Ref. 17. We can write $M = M^L M^{L-1}, \ldots, M^2 M^1$, where for convenience we let $(S^{\mathrm{III}})^{-1}$ be considered as part of M^L and S^{I} be included in M^1. Let $C(i)$ be the column vector of the coefficients of the local orbitals on layer i. Then, by transferring from left to right, we have a chain of equations relating the coefficients $f^{\mathrm{I,III}}$ and $C(i)$:

$$\begin{aligned} C(i_1) &= M^1 f^{\mathrm{I}}, \\ C(i_2) &= M^2 C(i_1), \\ &\vdots \\ C(i_{L-1}) &= M^{L-1} C(i_{L-2}), \\ f^{\mathrm{III}} &= M^L C(i_{L-1}), \end{aligned} \tag{6}$$

where the i_l designate the layers at the boundaries between the regions transferred over by the M^i. These equations must now be rearranged so that the unknowns, f_+^{III}, f_-^{I}, and the $C(i)$'s are on the left of the equal sign, and the knowns, f_-^{III} and f_+^{I}, are on the right. This is easily done in a similar fashion as above for M itself by splitting M^1 into four submatrices. The result can be written in matrix form as

$$\begin{pmatrix} -M^1_{+-} & 1 & 0 & & & & & \\ -M^1_{--} & 0 & 1 & & & & & \\ & -M^2_{++} & -M^2_{+-} & 1 & 0 & & & \\ & -M^2_{-+} & -M^2_{--} & 0 & 1 & & & \\ & & & -M^3_{++} & -M^3_{+-} & 1 & 0 & \\ & & & -M^3_{-+} & -M^3_{--} & 0 & 1 & \\ & & & & & \ddots & & \\ & & & & & -M^L_{++} & -M^L_{+-} & 1 \\ & & & & & -M^L_{-+} & -M^L_{--} & 0 \end{pmatrix} \begin{pmatrix} f_-^{\mathrm{I}} \\ C_+(i_1) \\ C_-(i_1) \\ C_+(i_2) \\ C_-(i_2) \\ \vdots \\ C_+(i_{L-1}) \\ C_-(i_{L-1}) \\ f_+^{\mathrm{III}} \end{pmatrix} = \begin{pmatrix} M^1_{++} f_+^{\mathrm{I}} \\ M^1_{-+} f_+^{\mathrm{I}} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ -f_-^{\mathrm{III}} \end{pmatrix}. \tag{7}$$

Note that the M^i for i greater than 1 are also split into four parts, but for clarity of presentation only. The matrix equation has the familiar "$Ax = b$" form, where A is a known matrix, b is a known vector, and x is to be found. As discussed above, it can be solved very efficiently using conventional software, such as from LINPACK, using the band matrix versions. The main difference with the matrix of Ref. 17 is the placement of f_-^{I} and f_+^{III} at opposite ends of the column vector, instead of together at one end.

Figure 2 shows the results of a calculation for the transmission of carriers entering as heavy holes versus energy through an AlAs/GaAs/AlAs double-barrier structure with 6/12/6 monolayers, respectively (17 Å/34 Å/17 Å). The energy zero is the GaAs valence-band edge. There is a 0.05-V drop across the structure. The solid line was found using the method of Ref. 12 and the effects of the numerical instability can be seen. For even larger structures, the calculation deteriorates rapidly. In contrast, the results of our method, the dashed line, are smooth.

For the case shown in Fig. 2, region II was split up into only two subregions, i.e., $M = M^2 M^1$ was used. This

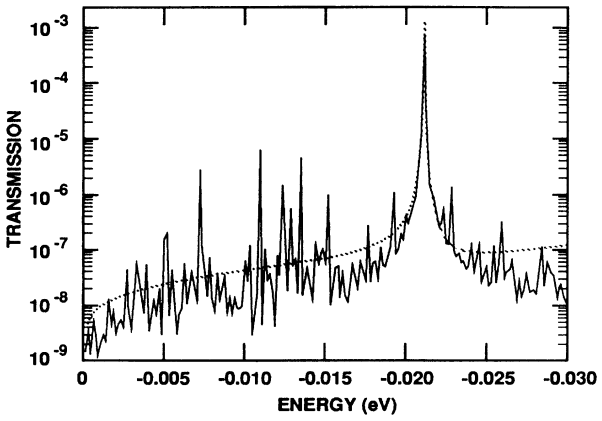

FIG. 2. Hole transmission vs energy for a double-barrier AlAs/GaAs/AlAs double-barrier tunnel structure. Solid line, method of Ref. 12; dashed line, our method.

was found to be usually sufficient for the region of the double-barrier structures alone. For more complicated problems, e.g., when the space-charge regions are also included,[14,17] much longer ranges are needed. However, for this simple case, Eqs. (6) and (7) simplify in that there is no need to introduce extra $C(i)$'s. Equation (6) becomes

$$C(i_1) = M^1 f^{\mathrm{I}}, \qquad f^{\mathrm{III}} = M^2 C(i_1). \tag{8}$$

Eliminating $C(i_1)$ gives

$$(M^2)^{-1} f^{\mathrm{III}} = M^1 f^{\mathrm{I}}. \tag{9}$$

This is then rearranged as above so that the unknowns are on the left and the knowns on the right, for solution by $Ax = b$ linear-equation software.

Slight modifications of the method make it suitable for bound states or periodic systems, e.g., superlattices. For a bound state there is no incoming wave and the right-hand side of Eq. (7) is zero and the matrix equations are overdetermined. The bound-state energy is identified as the energy for which the overdetermined equations are mutually consistent. This can be done by finding the energies for which the determinant is zero, or other equivalent techniques.

For periodic systems the f^{I} and f^{III} are eliminated. There is an additional condition to be satisfied, the Bloch condition, $M^L C(i_{L-1}) = C(i_L) = e^{ikD} C(1)$, where k is the wave vector of the Bloch state and D is the width of the repeated region. This condition appears in the upper right corner of the matrix equation for $C(1), C(i_1), C(i_2), \ldots, C(i_{L-1})$. The equations can be solved for given k in a manner similar to the bound-state problem.

In conclusion, we have presented a derivation of a numerically well-behaved solution for the tight-binding model of tunneling. It is similar to the model of Boykin, van der Wagt, and Harris, Jr. with this derivation clearly revealing the band nature of the matrix. The size of

the matrix involved in the present method and that of Boykin, van der Wagt, and Harris, Jr. is significantly smaller than that employed by Ting, Yu, and McGill as stated above. However, a closer analysis reveals that the time required will not differ significantly. The solution of Ting, Yu, and McGill for the band matrix linear equation problem for the (100) nearest-neighbor model requires about $4m^3N$ multiplications, where m is the number of orbitals and N is the total number of individual layers (each layer consisting of a cation-anion pair). The matrix of the present method is smaller and thus requires about $4m^3L$ multiplications for its solution. However, the L M^i matrices are constructed by multiplying together the T matrices [see Eq. (5)], and this requires about $4m^3(N-L)$ multiplications, making the net effort about the same. Actually, the forms of the matrices are somewhat different in the two models, making precise comparisons difficult. Depending on this, other details of the model, the orientation of the interfaces, and the efficiency of the implementations of the computer codes, the factor "4" will vary, but not significantly. Thus the total time taken is similar for the two methods.

[1] L.L. Chang, L. Esaki, and R. Tsu, Appl. Phys. Lett. **24**, 593 (1974).
[2] T.C.L.G. Sollner, W.D. Goodhue, P.E. Tannenwald, C.D. Parker, and D.D. Peck, Appl. Phys. Lett. **43**, 588 (1983).
[3] R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).
[4] E.E. Mendez, W.I. Wang, B. Ricco, and L. Esaki, Appl. Phys. Lett. **47**, 415 (1985).
[5] E.E. Mendez, E. Calleja, and W.I. Wang, Phys. Rev. B **34**, 6026 (1986).
[6] A.R. Bonnefoi, T.C. McGill, R.D. Burnham, and G.B. Anderson, Appl. Phys. Lett. **50**, 344 (1987).
[7] S.S. Rhee, J.S. Park, R.P.G. Karunasiri, Q. Ye, and K.L. Wang, Appl. Phys. Lett. **53**, 204 (1988).
[8] H.C. Liu, D. Landheer, M. Buchanan, and D.C. Houghton, Appl. Phys. Lett. **52**, 1809 (1988).
[9] J.R. Söderström, D.H. Chow, and T.C. McGill, Appl. Phys. Lett. **55**, 1094 (1989).
[10] L.F. Luo, R. Beresford, and W.I. Wang, Appl. Phys. Lett. **55**, 2023 (1989).
[11] G.C. Osbourn and D.L. Smith, Phys. Rev. B **19**, 2124 (1979).
[12] J.N. Schulman and Y.-C. Chang, Phys. Rev. B **27**, 2346 (1983).
[13] D.Y.K. Ko and J.C. Inkson, Phys. Rev. B **38**, 9945 (1988).
[14] D.Z.-Y. Ting, E.T. Yu, and T.C. McGill, Appl. Phys. Lett. **58**, 292 (1991).
[15] W.R. Frensley (private communication).
[16] C.S. Lent and D.J. Kirkner, J. Appl. Phys. **67**, 6353 (1990).
[17] T.B. Boykin, J.P.A. van der Wagt, and J.S. Harris, Jr., Phys. Rev. B **43**, 4777 (1991).

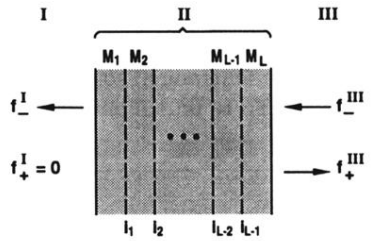

FIG. 1. Diagram of multilayer regions. The carrier is incoming and reflected in region III. Region II is the varying composition and/or potential region. The transmitted carrier is in region I. Region II is divided into L subregions traversed by the M^L matrices.