Bridging the Gap between Crystal Theory and Semiconductor Physics

William R. Frensley

Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX 75080
(Dated: June 5, 2018)

The theory of perfect crystals, founded upon the Bloch theorem, gives an understanding of extended quantum states grouped into energy bands, and permits the derivation of the dynamics of electrons in those states. The semiconductor physics used to explain the operation of electronic devices treats the (imperfect) semiconductor crystal as a uniform effective medium in which positively and negatively charged quasiparticles mostly obey Newtonian dynamics, and in which the chemistry of impurity atoms is far different from that of those same atoms in free space. The connection between these two pictures can be made by made by invoking a mathematical transformation that takes the finite-temperature, impure device structure and algebraically subtracts from it a perfect crystal, leaving only the residual differences to be analyzed. This notion of the residual difference offers a conceptual basis for understanding many aspects of semiconductor physics, including the properties of impurity states and heterogeneous interfaces. The mesoscopic transformation that underlies the residual-difference picture provides the systematic way to define a concept that is essential to the understanding of semiconductor devices: a position-dependent band structure.

I. INTRODUCTION

The physics of semiconductor devices is taught in a number of contexts, most prominently in courses dedicated to this topic in Electrical Engineering curricula, but also in Modern Physics courses (particularly when an associated laboratory includes measurements of the Hall effect), and other courses in solid-state physics and materials science. Semiconductor devices are also the prototypical nanostructured systems of technological significance, and an understanding of how this field handles abstraction can benefit those who wish to pursue work in nanotechnology.

The traditional expositions of semiconductor physics make a rather awkward leap from the microscopic physics of perfect crystals to the meso- or macroscopic picture in which all subsequent discussion is grounded. In the microscopic picture we treat only perfect, unbounded crystals to which the Bloch theorem applies. We make the assumption that we have access to a sufficiently complete understanding of quasi-particle moves as if it has an “effective mass” $m^*$ that is typically of the order of $0.1m_0$, $m_0$ being the mass of the electron in free space. Impurity atoms can be embedded within this medium by technological processes, and once they are in place they generally act as spatially fixed charges of unit magnitude and either sign. The semiconductor is also a dielectric with relative permittivity $K$ which is typically a bit greater than 10. From this picture the development of the subject proceeds to the equilibrium distributions of electrons and holes, and their nonequilibrium interactions including electron-hole pair generation and recombination, and transport in response to applied electric and magnetic fields and concentration gradients. These last elements are the basis for describing the operation of diodes, transistors, and other semiconductor devices.

II. THE MESOSCOPIC TRANSFORMATION

There is an intermediate step in the logical development that is the key to bridging the gap described above. It is generally described as “effective-mass theory,” and the seminal works are those of Slater and of Luttinger and Kohn. In the present work we will consider these to be particular implementations of a broader category of mesoscopic transformations. Effective-mass theory is generally presented as an approach to treating a system consisting of a periodic potential $V_{\text{periodic}}$ and an “external” slowly-varying potential $V_{\text{ext}}$. It is generally assumed that one unique $V_{\text{periodic}}$ will adequately deter-
mine the energies and wavefunctions of all the states of interest, typically those of the valence bands and lower conduction bands. By factoring out a basis function that approximates the behavior of the microscopic wavefunction within the crystal unit cell, the rest of the wavefunction becomes a slowly-varying mesoscopic wavefunction \( \Psi_b \) (often called the envelope function when the Luttinger-Kohn formulation is used). The mesoscopic wavefunction is always associated with a particular band \( b \), hence the need to include a band index in the notation.

The mesoscopic wavefunction obeys the effective-mass Schrödinger equation which is of the general form:

\[
i\hbar \frac{\partial \Psi_b}{\partial t} = E_b(-i\nabla)\Psi_b + V_{\text{ext}}\Psi_b, \tag{3}\]

where \( E_b \) is here the functional form of the dispersion relation applied to the gradient operator. We usually expand this function about its extremum (assumed for the present to be located at \( k = 0 \)), yielding:

\[
E_b(k) \approx E_b(0) + \frac{\hbar^2 k^2}{2m^*},
\]

and we now see the origin of the effective mass. The effective-mass Schrödinger equation now takes a more familiar form:

\[
i\hbar \frac{\partial \Psi_b}{\partial t} = -\frac{\hbar^2}{2m^*} \nabla^2 \Psi_b + (E_b + V_{\text{ext}}) \Psi_b. \tag{4}\]

There is a distinct equation for each energy band of interest. Note that the periodic potential remains in the problem only through the dispersion relations \( E_b(k) \).

Before we examine the conceptual consequences of this absorption of the periodic potential we need to look more clearly at what is actually contained in that potential. First, we will assume that the ideal crystal bandstructure computation used the pseudopotential approach. The pseudopotential is an effective potential that is much weaker than the realistic potential within an atom, which is constructed to allow no bound core states, but to produce accurate wavefunction energies for states with energies well above those of the tightly-bound core electron states. The reason for insisting on the use of pseudopotentials (apart from the fact that all modern electronic-structure computations use this approach) is that this allows us to assert that the ion core (pseudo)potentials for different elements from the same column of the periodic table differ by only small amounts, even though their nuclear charge and core electron configurations differ a great deal. In addition to the core pseudopotentials, \( V_{\text{periodic}} \) will include contributions from the Hartree potential, which must be self-consistent with the valence-electron charge distribution, and the appropriate exchange-correlation potential.

Now we can articulate a more general interpretation of the mesoscopic transformation than merely cancelling out the periodic potential. When we apply this transformation, we will not only cancel out the periodic potential, we will cancel, or subtract, every quantity with which the potential is self-consistent. The mesoscopic transformation algebraically subtracts an ideal crystal from the non-ideal system that is of interest. This subtraction includes the ion-core potentials, the valence-electron charge distribution and the occupation probabilities for each electron state in the system. And, as was the original intent of the approach, the strong periodic crystal potential is subtracted out, leaving only the residual Hartree potential due to the charges that are present in the residual differences that define the system of interest. There are two other consequences of this transformation: The residual charge carriers (electrons and holes) obey the effective-mass Schrödinger equation of the appropriate energy band, and electrostatic interactions within the residual system must be computed using the dielectric constant of the semiconductor material. The last requirement follows from the fact that we may be able to abstract away the valence charge distribution, but we cannot abstract away its self-consistent response to applied electric fields. With this in mind, we can specify that \( V_{\text{ext}} \) should be the Hartree potential derived from the residual system by solving Poisson’s equation in the dielectric medium for the residual charge distribution, subject to boundary conditions derived from state of any electrical contacts to the system. (The particle densities in the residual system are sufficiently dilute that exchange interactions are generally negligible.)

The workings of the mesoscopic transformation may be visually illustrated using three Figures. Fig. 1 schematically illustrates a perfect silicon crystal. Fig. 2 shows a Si crystal in which impurity atoms have been introduced. Finally, Fig. 3 shows the enormous simplification that is evident in the residual system after the algebraic subtraction of the ideal crystal.
FIG. 2. Schematic of a crystal with impurity atoms introduced as is common in device fabrication processes.

FIG. 3. Schematic of the simplified system resulting from the mesoscopic transformation which algebraically subtracted Fig. 1 from Fig. 2. The open circle where an electron would have been in the perfect crystal represents the hole quasiparticle.

III. OCCUPATION FACTORS AND QUASIPARTICLES

Expressing the effect of the mesoscopic transformation as a subtraction operation clarifies a number of points. Most importantly it explains why quantities that are manifestly non-negative like particle spatial densities and quantum state occupation probabilities can now take negative values. In particular there can be negative occupation probabilities for states in the valence band, otherwise known as a positive density of holes. The dynamic properties of a hole are indistinguishable from those of a positively-charged particle with positive mass, particularly in measurements of the Hall effect, and therefore we can regard it as a quasiparticle with those properties. The point that is essential when one attempts to explain how the absence of a negatively-charged electron can move like a positively-charged particle is that the state from which the electron is missing has a negative effective mass. So the short explanation is that the absence of a negatively-charged, negative effective-mass particle is equivalent to the presence of a positively-charged, positive mass particle.

But it is often useful to expound upon this point by making a more detailed examination of the dynamical observables, including energy, momentum and charge current. This explanation can be a bit tricky, and it is very easy to draw erroneous conclusions that appear to contradict the positive quasiparticle picture. Thus it is worthwhile to document the details of this analysis. We begin by defining a reference state of the crystal in Fig. 4 that is in particular required for the evaluation of the kinetic energies of the quasiparticles. In this state there is one extra electron in the conduction band and one missing electron (hole) in the valence band, both assumed to be at \( k = 0 \) in this initial condition. In what follows, one must bear in mind that the particle velocity is the group velocity (1) and that the charge current is in reality carried by electrons and is defined by:

\[
  j = -q v.
\]  

We quickly see that the conduction-band electron in Fig. 4 has zero momentum, velocity, and current. For the valence band, however, it is still the electrons that contribute the momentum and current. Thus, we need to sum these quantities over all of the occupied valence-band states. These sums are greatly simplified by the Kramers degeneracy of \( E_b(k) \), which guarantees that the dynamical contributions of a state \( |k, b\rangle \) will be cancelled out by the contributions of state \( |-k, b\rangle \), provided that both are occupied. This cancellation is represented in Fig. 4 by opposing arrows. Taking this cancellation into account, we readily see that the valence-band system of Fig. 4 has zero momentum and current. The unoccupied hole state also has a zero velocity due to its position at an extremum of the band structure.

We now assume that an electric field \( \mathcal{E} \) in the posi-
The mesoscopic transformation also helps us explain the behavior of transition-metal impurities in semiconductors, even though their states are not at all those of the “effective atoms” described above. To understand the behavior of the transition metals, we must recognize that the ideal crystal which we subtract from the non-ideal system contains only valence-band states that are derived from s and p orbitals. A filled d shell will be treated as part of the ionic core, but empty or partially-filled d shells whose energies lie in proximity to the states of the valence s and p states will not be subtracted out of the problem, and thus they remain as a part of the residual system. The d orbital energies will of course be perturbed by the chemical environment of the crystal, but if they occur within the energy-band gap, they can appear at essentially any energy, and they will form the localized states known as “deep levels.” Because they are fundamentally the orbitals of the free atom, they will have a

\[ E_{\text{impurity}} = E_b - \frac{m_e^*}{K_s^4} R, \]
spatial extent of the order of 0.1 nm, unlike the much more extended hydrogenic states. Their distribution in momentum space will therefore span nearly the entire Brillouin zone of allowed momenta, and this points to the significance of such states: They can act as intermediate states in electron-hole recombination process, effectively breaking the momentum-conservation restriction. This is their main technological significance, increasing the rate of recombination. Often this is a parasitic process that we desire to minimize, but in some cases it is desirable and can be enhanced by the deliberate introduction of transition-metal impurities.

V. APPLICATION TO SEMICONDUCTOR HETEROSTRUCTURES

Until the mid-1970s the picture provided by the mesoscopic transformation was primarily of use as an intermediate step in the development of the purely classical picture of semiconductors as hosts for Newtonain quasi-particles. But then, the development of technologies for making high-quality semiconductor heterojunctions between chemically different materials led to the demonstration of size-quantization effects in such structures, and the quantum well became a common technological component. If the material composition changes abruptly the energy bandgap will also change abruptly, and there will necessarily be abrupt shifts in the band-edge energies. Because those energies are major components of the scalar potentials in equations (1), the resulting Schrödinger problems can very closely approximate the piecewise-constant potentials that had long appeared in textbooks on quantum mechanics. The mesoscopic picture was then and remains today the most convenient way to evaluate the quantum states of such structures. But, the presence of more than one type of ideal crystal poses complications for the formulation of the mesoscopic transformation.

To preserve simplicity in the mesoscopic model, we prefer a transformation which subtracts out the appropriate semiconductor composition in each region of the heterostructure. But to achieve this we need to more closely examine the technical details of the widely-used mesoscopic transformations.

The mesoscopic transformation always consists of a unitary transformation onto a basis that represents the mesoscopic wavefunction, and thus the exact formulation is spatially discrete like a tight-binding model; the continuum effective-mass Schrödinger differential equation is an approximation to the exact discrete difference equation.

The Luttinger-Kohn approach defines the mesoscopic wavefunction as a slowly-varying envelope function, which yields the microscopic wavefunction when multiplied by a zone-center Bloch function. The Fourier spectrum of the envelope function is limited to the first Brillouin zone, and any function with this property can be exactly specified by sampling its value at each lattice point. Thus the Luttinger-Kohn Hamiltonian can be formulated, in principle, as a discrete matrix of the same form as the Wannier-Slater Hamiltonian. (The discrete formulation makes a fleeting appearance in that paper, but the authors treated it as an approximation, not recognizing that the localized functions they defined are mutually orthogonal.)

Because these transformations can be expressed in terms of localized basis functions, it would appear to be straightforward to construct a mesoscopic transformation for a heterostructure by using a localized basis that is derived from the correct material at each lattice point. But noted that this does not generally produce a unitary transformation because the basis states from different materials are not necessarily orthogonal. The common practice in this field has been to neglect such considerations and assume that a suitable transformation can be defined in which the bulk effective-mass Schrödinger equation is valid right up to a heterojunction. On the whole such an approximation is well supported by agreement between experimental observations and theoretical calculations; there is no strong evidence that the electronic structure is significantly perturbed by the heterojunction beyond the atomic layers adjacent to the junction.

When we include the simplest models for both heterostructure and band structure the effective-mass Schrödinger equation becomes (in one dimension):

\[ i\hbar \frac{\partial \Psi_b(z)}{\partial t} = \]
\[ -\frac{\hbar^2}{2} \frac{\partial^2}{\partial z^2} m^*(z) \frac{\partial}{\partial z} \Psi_b + [E_b(z) + V_{ext}(z)] \Psi_b(z). \] (8)

We must write the kinetic-energy term in a Sturm-Liouville form to preserve the Hermiticity of the Hamiltonian. This is an adequate approximation when we can neglect energy-band nonparabolicity and the coupling of unlike bands across the heterojunction. It leads to the continuity of \((1/m^*) (\partial \Psi_b / \partial z)\) as the interface matching condition for the wavefunction.

A Hermitian mesoscopic Hamiltonian will determine the form of both interface matching conditions and of the current-density expression. If these are derived from the same Hamiltonian they will necessarily be self-consistent. However, for more complicated Hamiltonians, one must actually derive the current density by deriving the continuity equation. Generalizations of the Green identity required for such a derivation exist for all orders of derivative, for multi-component wavefunctions, and for spatially discrete formulations. There is a widespread misconception that the current density can be derived from the Heisenberg equation for \(dz/dt\). While this does
give the correct answer for the motion of the centroid of a wavepacket, it does not give a current density which obeys the expected continuity equation in discrete formulations.

VI. THE POSITION-DEPENDENT BAND EDGE

Finally, let us point out the role of the mesoscopic transformation in defining the notion that has underlain all of the preceding discussion, and indeed all of semiconductor device physics: the concept of a position-dependent band-edge. In principle such a concept runs into difficulties with the uncertainty principle, since energy bands are only well-defined in momentum space. In practice, we will see that the only useful definition of such a quantity is as that potential which appears in an appropriate effective-mass Schrödinger equation. The local band edge is not a directly observable quantity. Consider the problem of trying to infer the energy-band profile from a microscopic electronic-structure computation, such as are done with density-functional methods. Such a computation produces a set of wavefunctions and their energies, and with this information one can construct a local density of states, such as that illustrated in Fig. 6 (though this particular example was generated using a mesoscopic model). One can discern that there is an energy barrier near the center that extends into both conduction and valence band, and that an overall bias voltage has been applied, due to the offset of the energy gap.

When the mesoscopic potentials are drawn on the same plot as shown in Fig. 7, the situation becomes a great deal more obvious. One can readily understand how the how

VII. SUMMARY

The concept of the mesoscopic transformation is simple enough to be grasped and used by those who are just beginning the study of semiconductor physics, even though the rigorous demonstration of its validity is a far more advanced task. A distinguished precedent for the employment of such a result can be found in Richard Feynman’s use of the Heaviside radiation formula to teach the fundamental concepts of electromagnetic radiation and of radiating systems before developing Maxwell’s equations. This allowed him to explain not only simple properties like interference and polarization, but also such advanced phenomena as bremsstrahlung and synchrotron radiation. Similarly, the mesoscopic transformation makes it easy to grasp the basic properties of electrons, holes and dopants, but also offers insight into more advanced topics like deep-level impurities and size quantization in heterostructures. Neither the Heaviside formula nor the mesoscopic transformation is necessarily the optimum approach if highly accurate quantitative predictions are the goal, but each of them helps one develop an ability to identify the most likely behavior of a physical system without recourse to detailed computation. And such an ability is the operational definition of physical intuition.
1. H. Kroemer, Prog. IEEE 63, 988 (1975).
2. H. Kroemer, Am. J. Phys. 54, 177 (1986).
3. J. C. Slater, Phys. Rev. 76, 1592 (1949).
4. J. M. Luttinger and W. Kohn, Phys. Rev. 97, 869 (1955).
5. H. Kroemer, Phys. Scr. 1996, 10 (1996).
6. R. Dingle, W. Wiegmann, and C. H. Henry, Phys. Rev. Lett. 33, 827 (1974).
7. M. G. Burt, Semiconductor Science and Technology 3, 739 (1988).
8. D. J. BenDaniel and C. B. Duke, Phys. Rev. 152, 683 (1966).
9. W. R. Frensley, “Current density operator in systems with non-parabolic, position-dependent energy bands,” (1995), arXiv:1509.05929 [cond-mat.mes-hall].
10. R. P. Feynman, R. B. Leighton, and M. Sands, The Feynman Lectures on Physics, Vol. I (Addison-Wesley Publishing Co., Inc., Reading, Mass.-London, 1963) Chap. 28.