The electronic structure of impurities and other point defects in semiconductors*

Sokrates T. Pantelides

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

A review is presented of the various theoretical methods that have thus far been developed for the study of states introduced by impurities and other point defects in semiconductors. The main body of the paper is prefaced with brief sections on the role of impurities and defects in semiconductors and on the general aspects of experimental techniques, as an appropriate setting for the theoretical discourse. Theoretical methods, including those of the effective-mass type, and a wide range of methods appropriate to deep levels are then presented. Applications of these methods are discussed critically. Finally, the relative merits of the various approaches are compared and the prospects for future work are assessed.

CONTENTS

LIST OF SYMBOLS

a_q	Phonon annihilation operator
$a_q^\dagger$	Phonon creation operator
$D^0(E)$	Density of states of perfect crystal
$D(E)$	Density of states of perturbed crystal
$\Delta D(E)$	Change in the density of states
det	Determinant
e	Charge of a proton

*Work supported in part by the Air Force Office of Scientific Research under contract No. F49620-77-C-0005.

E_A	Thermal activation energy
E_B	Binding energy
E_c	Conduction band edge
E_v	Valence band edge
E_{cv}	Minimum band gap
E_g	Average band gap
E_I	Ionization energy
E^0_{nk}	Energy bands in perfect crystal
E_ν	Energy levels in perturbed crystal
$\mathcal{E}$	Electric field
F	Total angular momentum
F_{nk}	Expansion coefficients for impurity wave function in Bloch representation
f_{nk}	Expansion coefficients for impurity wave function in pseudo-Bloch representation
F_{nj}	Expansion coefficients for impurity wave function in Wannier representations
G^0	Green's function for perfect crystal
G	Green's function for perturbed crystal
H^0	One-electron Hamiltonian for perfect crystal
H	One-electron Hamiltonian for perturbed crystal
H_{pol}	Polaron Hamiltonian
H_{ph}	Phonon Hamiltonian
H_{ep}	Electron-phonon interaction Hamiltonian
Im	Imaginary part of a complex number
I_μ	μth ionization potential for an atom
I	Spin 1 angular momentum matrix
J	Spin 3/2 angular momentum matrix
$\mathbf{k}$	Wave vector of Bloch functions
k_B	Boltzmann's constant
$\mathbf{K}_p$	Reciprocal lattice vectors
$K_{\mu\lambda}$	Extended Hückel theory constants
m_0	Mass of an electron in vacuum
m^*_e	Effective mass of an electron
m^*_h	Effective mass of a hole
n	Principal quantum number for hydrogenic states
n	Band index in perfect crystal
$n(T)$	Carrier concentration at temperature T
p	Momentum
P_{ik}	Momentum Cartesian tensor
Q	$1-G^0U$
$\mathbf{R}_j$	jth atomic site
Re	Real part of a complex number
R_M	Model-potential radius
$S_{\mu\lambda}$	Overlap matrix of basis functions
T	Absolute temperature
Tr	Trace
u^0_{nk}	Cell-periodic part of Bloch function
U	Impurity or defect potential
U_b	Bare impurity potential
U_s	Screening potential
U_p	Impurity or defect pseudopotential
U_{pb}	Bare impurity or defect pseudopotential
U_{ps}	Screening pseudopotential
U_H	Hydrogenic potential
U_{pc}	Point-charge potential
V^0	Potential of perfect crystal
V	Potential of perturbed crystal
V_M	Model-potential parameter
V_q	Electron-phonon coupling constant
$w_n(\mathbf{r}-\mathbf{R}_j)$	Wannier function of nth band and jth site
Z	Atomic number
z	Valence of chemical elements
$\gamma_1, \gamma_2, \gamma_3$	Valence-band effective-mass parameters
$\mathbf{\Gamma}$	Width of resonances
δ	Valence-band effective-mass parameter
$\delta(E)$	Phase shift
$\Delta(E)$	Determinant used in methods described in Sec. IX
μ	Valence-band effective-mass parameter
μ_{xc}	Local-density functional for exchange and correlation
$\sigma(E)$	Photoionization cross section
Φ	Trial function in acceptor calculations
$\phi^0_{n\mathbf{k}}$	Pseudo-Bloch functions of the nth band
ϕ_ν	Pseudo-wave-functions in the perturbed crystal
$\chi_{\alpha\mathbf{k}}$	Bloch sum of atomic oribtals
$\psi^0_{n\mathbf{k}}$	Bloch function of nth band
ψ_ν	Wave function in perturbed crystal
ψ^0_{ct}	Core wave function of tth core level in perfect crystal
$\psi_{ct'}$	Core wave function of t'th core level in perturbed crystal
ω_q	Phonon frequencies

LIST OF ACRONYMS

EHT	Extended Hückel Theory
EMA	Effective-Mass Approximation
EME	Effective-Mass Equation
EMT	Effective-Mass Theory
ENDOR	Electron-Nuclear Double Resonance
EPR	Electron Paramagnetic Resonance
ESR	Electron Spin Resonance
HEMT	Hydrogenic Effective-Mass Theory
LCAO	Linear Combination of Atomic Orbitals
LED	Light-Emitting Diode
MV EME	Multi-Valley EME
Xα-SW	Xα-Scattered-Wave Method

I. INTRODUCTION

Solids are usually classified in terms of their electrical properties as conductors, insulators, and semiconductors. Conductors and insulators have intrinsic properties which make them very useful in applications. Semiconductors, on the other hand, would have very little practical use if it were not for the wide range of properties which can be attained via the incorporation of impurities. The role of impurities in semiconductors was recognized soon after the advent of quantum mechanics (see, for example, Wilson, 1932), but progress was very slow for about a decade. Then the urgent needs during World War II for efficient devices ushered in the era of semiconductors. Building on experience gained during the long secretive years of the war, Bardeen, Brattain, and Shockley (1948, 1949) invented the transistor. Since then, semiconductor technology has mushroomed, and has given us high-speed computers, junction lasers, light-emitting diodes, and the myriad of appliances and devices which proudly bear the insignia "solid state." It is fair to say that little of all this would have been possible if it were not for the effects produced by impurities when they are judiciously introduced in semiconductors.

The subject of impurities is, therefore, a vast one and no single monograph can cover all their important aspects. There exist many books devoted entirely to impurities and other defects (see bibliography) as well as a number of review papers focusing on particular aspects of the problem (Kohn, 1957; Dean, 1968, 1973; Williams, 1968; Queisser, 1974; Bassani, Iadonisi, and Preziosi, 1974; Roitsin, 1974; Grimmeiss, 1977; Miller, Lang, and Kimerling, 1977).

The main purpose of this paper is to provide a comprehensive account of theoretical methods and techniques that have thus far been used to describe the electronic structure and properties of impurities and defects in semiconductors. Instead of plunging directly into theory, however, a number of background sections are included to set the subject in perspective. Thus we begin with a section on basic concepts and a classification of impurities and defects, designed to facilitate subsequent discussion. In the next section, we trace the historical development of the theoretical understanding of the energy levels introduced by impurities, and give elementary descriptions of these levels. Finally, before we embark on our main subject of theoretical methods, we supply a brief section on the role of impurities in devices and a section on experimental techniques that are used to measure the positions of the energy levels introduced by impurities and defects.

Five sections and one appendix follow on theoretical methods. We begin with a section that gives a rigorous description of the quantum-mechanical problem and presents general results. We then devote two sections to effective-mass theory and its extensions, one section to other general techniques which are perturbative in nature, and one section to nonperturbative methods. We do not, by any means, exhaust the subject, especially the applications of effective-mass theory. For example, we do not discuss excitons or bound excitons, impurity bands, the analysis of ENDOR data, nor do we describe the changes of impurity states caused by external stresses or fields. For the other methods, which generally apply to deep levels, we try to be as comprehensive as possible, since most of them are still at the stage of infancy. In the last section we provide a comparative critique of methods and assess the prospects for future work.

II. BASIC CONCEPTS AND CLASSIFICATION OF DEFECTS AND IMPURITIES

A perfect crystal consists of a three-dimensional array of atoms arranged on a periodic lattice. The introduction of imperfections in such a crystal disrupts this periodic structure and alters the properties of the material in significant ways. Because of the variety of imperfections that can be present in a crystal and the variety of ways in which they affect the properties of the host material, we begin by discussing ways of classifying imperfections. Terminology will thus be introduced which will facilitate subsequent discussion.

At first, we distinguish between lattice defects and foreign atoms, or impurities. Lattice defects can be *point defects*, which correspond to misplaced atoms, *line defects*, which correspond to misplaced lines of atoms and are known as dislocations, and *planar defects*, which correspond to misplaced planes of atoms and are known as stacking faults. This review will not address the properties of either line or planar defects.

There are basically two lattice point defects (Fig. 1): the vacancy, i.e., a vacant atomic site, and the self-interstitial, i.e., an extra atom occupying an interstitial site. In compound semiconductors, such as those of type AB, a third possibility exists, namely an *antisite defect*, which corresponds to an A atom occupying a B site or vice versa. These elementary point defects may form complexes among themselves and with impurities (see below).

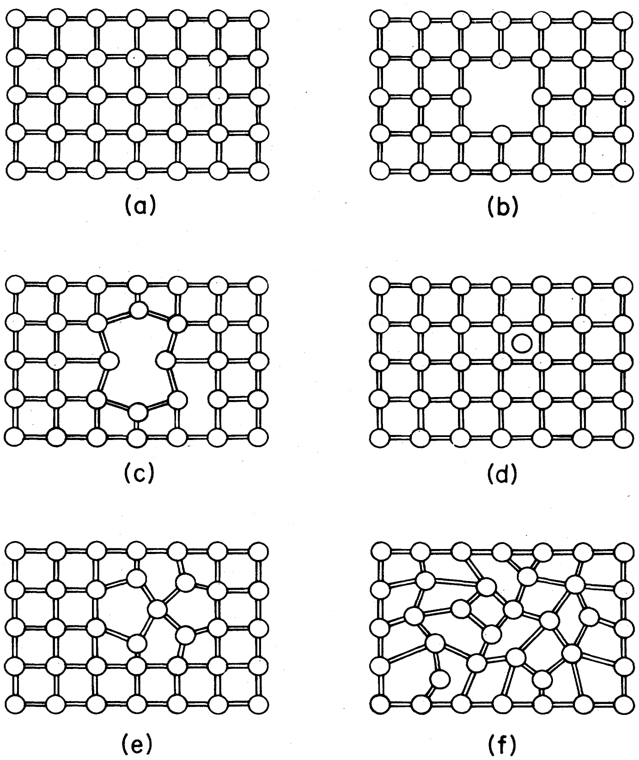

FIG. 1. Schematic illustration of various lattice defects: (a) perfect crystal, (b) ideal vacancy, (c) reconstructed vacancy, (d) self-interstitial, (e) simple interstitialcy, (f) extended interstitial.

Single impurities may be classified in a variety of ways. The simplest way is to classify them in terms of their physical location in the lattice. An impurity atom may replace one of the host atoms, in which case it is known as a substitutional impurity. Alternatively, it may occupy an interstitial site, in which case it may be either a simple interstitial, or what is known as an interstitialcy. The distinguishing criterion is whether the interstitial atom leaves the bonding of the host atoms undisturbed or if it disrupts the local bonding and forms a bridge between two host atoms (Fig. 1). A special case is a "split interstitial," in which a host atom is replaced by two atoms, symmetrically displaced with respect to the original site.

Finally, lattice point defects and impurities may form complexes. The simplest complexes are pairs, such as two impurity atoms at neighboring sites, a vacancy and an impurity atom at neighboring sites, and a divacancy. More extended complexes become more difficult to characterize, but some forms have been given special names. One example is that of an extended interstitial (Seeger and Chick, 1969; Van Vechten, 1977; see Fig. 1), in which case bonding is disrupted seriously in an extended region with the net result that the region contains an excess atom (an impurity atom may or may not be present). Other examples are the swirl defect in silicon (de Kock, 1973), which got its name from the way these defects are distributed on the surface of a wafer, and the dark-line defect in GaAs–GaAlAs heterostructure lasers (Petroff and Hartmann, 1973). The practice of giving special names to individual defects is more common in the case of defects in ionic crystals, which are known generically as color centers (Fowler, 1968).

Since most of our review will be on point defects, we now return to single impurities for further classification. Substitutional impurities may be classified according to their position in the periodic table of the elements relative to that of the host atom. Let Z be the atomic number and $\Delta Z = Z_{\text{impurity}} - Z_{\text{host}}$. Also let z be the chemical valence and $\Delta z = z_{\text{impurity}} - z_{\text{host}}$. Atoms from the same column of the periodic table as the host atom ($\Delta z = 0$) are known as isovalent impurities, because they have the same number of valence electrons as the host atoms. (Sometimes isovalent impurities are referred to as isoelectronic, which is an unfortunate practice. Isoelectronic is usually used to signify equal numbers of all electrons. For example, the crystals Ge and GaAs are isoelectronic since they have the same total number of electrons per unit cell. Ge and GaP are not isoelectronic even though they have the same number of valence electrons per unit cell.) From among nonisovalent impurities ($\Delta z \neq 0$), substitutional impurities from a column of the periodic table to the right of the column of the host atom ($\Delta z > 0$) are in general referred to as *donors*, because they have more valence electrons than the host atoms (the relative number of core electrons between host and impurity atom does not enter these considerations). On the other hand, substitutional impurities from columns of the periodic table to the left of the column of the host atom ($\Delta z < 0$) are in general known as *acceptors*, because they must accept electrons from host atoms in order to fulfill local bonding requirements. Donors and acceptors may be single, double, etc., depending on whether $|\Delta z|$ is one, two, etc., respectively (Fig. 2).

Position in the periodic table of the elements relative to the host provides an additional way of classifying impurities. Impurities from the same row of the periodic table as the host atom ($|\Delta Z| = 1, 2,$ or 3) are referred to as *isocoric* because the impurity core is isoelectronic with the core of the host atom.

Before we examine other methods of classifying impurities, it is worth noting that the terms "donor" and "acceptor" introduced above, though widely used, are rather ambiguous. For example, confusion may arise in compound semiconductors where the same impurity would be referred to as a donor or acceptor depending on which host atom it replaces. Furthermore, even in homopolar semiconductors, many atoms have been found to behave either as donors or acceptors depending on the circumstances, and have been referred to as amphoteric. A more precise and useful definition of the terms "donor" and "acceptor" will be given below. It will turn out that the definition given above is a special case of the more general and unambiguous one.

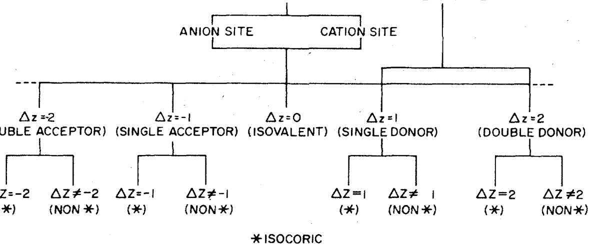

FIG. 2. Simple classification of impurities according to their position in the periodic table of the elements. Note that for interstitial impurities $\Delta Z = Z_{\text{impurity}}$ and $\Delta z = z_{\text{impurity}}$.

In order to arrive at the more general definition, we first need to turn to the electronic structure of perfect crystals and to the modifications it undergoes in the presence of impurities. We do not intend here to give an introduction to the band theory of solids. What is needed for the present purposes is the fact that the energy levels for electrons in a perfect crystal form a series of *bands*, separated by gaps. The various materials may be classified by the way these bands are occupied by electrons at 0 °K. First, all materials have the core bands (corresponding to the atomic core levels), which have negligible width, and are occupied by electrons. In metals, the core bands are followed by a set of contiguous bands which are not completely occupied. In insulators and semiconductors, on the other hand, the core bands are followed by the valence bands, which are occupied in their entirety at 0 °K; the valence bands are followed by a gap (the fundamental energy gap), which in turn is followed by the conduction bands which are empty at 0°K.

The wave functions of all states in a perfect crystal extend over the whole crystal and have the same probability amplitude in every unit cell. Periodicity does not allow the existence of *localized states*, namely states whose wave functions decay with distance outside a finite set of unit cells. When, however, an impurity or other defect is introduced, periodicity is broken and localized states are allowed. In most cases localized states appear in the fundamental gap. These are the states that have dramatic effects on the properties of semiconductors and are the main subject of this review.

An important consequence of the presence of these localized states in the otherwise forbidden energy gap is that the impurity[1] can exist in various *charge states* depending on whether the localized states are occupied or not. This fact leads to a precise and unambiguous definition for the terms "donor" and "acceptor" (Shockley, 1950): *Positively charged states of an impurity are defined as donor states, and negatively charged states are defined as acceptor states.* Neutral states bear no other distinctive name. Notice that this definition allows for the possibility that a given impurity can have only one or more donor states, in which case it can unambiguously be referred to as a donor impurity, as defined earlier in terms of Δz. Similarly, a given impurity can have

[1] From now on, the term "impurity" will be assumed to refer to both chemical impurities and lattice defects, such as vacancies and interstitials, unless an explicit distinction is made.

only one or more acceptor states, whereby it can unambiguously be referred to as an acceptor impurity. In fact, however, any impurity might in principle have both donor and acceptor states and many are observed to behave so. They are, therefore, referred to as amphoteric. Finally, note that according to the definition of donor and acceptor states given above, one can define donor and acceptor states for isovalent impurities as well, which, according to the definition of Δz (Fig. 2) would not be classified as either of the two.

We turn now to the concept of transitions from one charge state to another, caused by interaction with a perturbing field. Consideration of these transitions provides definitions for the energies that characterize donor and acceptor states. First we consider the ionization of a neutral impurity which is best described by a reaction of the form

$$X^0 \rightarrow X^+ + e^-, E_I, \tag{2.1}$$

for a donor, and

$$X^0 \rightarrow X^- + e^+, E_I, \tag{2.2}$$

for an acceptor (Van Vechten and Thurmond, 1976). Here X^0 is used to denote the neutral impurity, X^+ and X^- are used to denote charged states of the impurity, and E_I, the ionization energy, is the energy needed for the reaction to take place. The donor ionization reaction (2.1) corresponds to an emission of an electron e^- to the conduction bands. The acceptor ionization reaction corresponds to an emission of a hole e^+ to the valence bands, or, equivalently, the capture of an electron from the valence bands. The energy E_I in both cases defines what may be called a donor or acceptor *level*, respectively. These levels are conventionally marked on an energy level diagram as shown in Fig. 3. In the context of reactions (2.1) and (2.2) and Fig. 3, it should be mentioned that the forbidden band gap E_{cv} of a semiconductor is properly defined by the reaction (Van Vechten and Thurmond, 1976)

$$0 \rightarrow e^+ + e^-, E_{cv}. \tag{2.3}$$

The ratio of E_I and E_{cv} is often used to denote a level as *shallow* ($E_I \ll E_{cv}$) or *deep* ($E_I \lesssim E_{cv}$). These designations turn out to be useful in discussing the role of impurities in semiconductors since shallow and deep impurities affect the properties of the material in distinctly different ways (see Sec. IV).

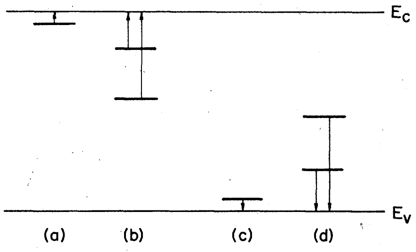

FIG. 3. Typical energy levels in the band gap of a semiconductor: (a) single donor, (b) double donor, (c) single acceptor, (d) double acceptor.

Ionization energies for double or triple donors and acceptors are defined in terms of reactions such as (2.1) and (2.2). For example, the reaction

$$X^+ \rightarrow X^{++} + e^-, E_I^{(2)} \tag{2.4}$$

defines the second ionization energy (second energy level) for a double donor (Fig. 3). Notice, however, that another reaction may be defined for X^+, namely

$$X^+ \rightarrow X^0 + e^+, E_{cv} - E_I. \tag{2.5}$$

This is not an ionization reaction. It is rather a deionization reaction since a charged impurity becomes neutral. In this particular instance, the deionization occurs by emitting a hole to the valence bands, a process that physically corresponds to capturing an electron from the valence bands. Deionization can, of course, also occur via the capture of an electron from the conduction bands in the reaction

$$X^+ + e^- \rightarrow X^0, -E_I. \tag{2.6}$$

The distinction between (2.5) and (2.6) is rather unique to impurities in semiconductors where there can be two sources of electrons, in contrast to the case of free atoms.

Many other reactions can be written down. For example, one can have:

$$X^+ + e^+ \rightarrow X^{++}, E_I^{(2)} - E_{cv} \tag{2.7}$$

which can be obtained by subtracting (2.3) from (2.4). Reactions (2.6) and (2.7) simply state that a given impurity state may act both as an electron and a hole trap.

At $T = 0$°K, the energy needed for or released by each of the above reactions is a change in the *enthalpy* H of the system. At finite temperatures, the reaction energy is a change in the Gibbs free energy F and is related to the change in enthalpy by

$$\Delta F = \Delta H + T \Delta S. \tag{2.8}$$

Finally, we conclude this section by stating the notation that will be used throughout this paper. For homopolar semiconductors, such as Si and Ge, we will employ the notation Si:P to denote a substitutional phosphorus impurity. A subscript i on the impurity atom will be used to denote an interstitial impurity. For example, Si:Li_i denotes a lithium interstitial in Si. For the vacancy, the symbol V will be used in the form Si:V. Since vanadium is not a common impurity in semiconductors, this notation should not cause confusion. In compound semiconductors, the notation for interstitial impurities remains the same. For substitutional impurities and vacancies, however, a subscript is added to the impurity atom to denote the site of substitution. For example, GaP:N_P stands for a nitrogen impurity at a phosphorus site. Finally, in all cases superscripts on the impurity atom symbol can be used to denote charge states. For example, ZnSe:V_{Se}^+ stands for the positively charged state of a Se vacancy in ZnSe.

III. ELEMENTARY THEORIES: HISTORICAL PERSPECTIVES

In this section we will discuss some elementary models for understanding the energy levels introduced by impurities in semiconductors and engage in a bit of nostalgia by tracing the historical development of the subject.

Some of the properties of semiconductors were known long before the advent of quantum mechanics. For example, rectification of alternating current at a contact between a metal and a semiconductor was first reported in 1874 (Braun). It was not, however, until the 1930's that semiconductor rectifiers became competitive with vacuum-tube diodes, when interest in higher and higher frequencies for radio waves was growing. This was the time when the application of quantum mechanics to the understanding of solids had just begun, and experiment and theory proceeded in parallel in a very fruitful way.

Understanding of the properties of solids in terms of quantum mechanics began with the work of Sommerfeld (1928) and of Bloch (1928). Bloch explained how the valence electrons in metals can behave as if they were free despite the immense attractive fields in the vicinity of the atomic nuclei. He showed that in a perfect lattice, conductivity would be infinite if it were not for the thermal motion of nuclei and for the presence of impurities, and calculated correctly the temperature dependence of metallic conductivity (Bloch, 1930). At about the same time, it was shown that the energy levels of electrons in a periodic potential break up into allowed and forbidden bands (Peierls, 1930; Morse, 1930; Brillouin, 1930; Kronig and Penney, 1931). Following these developments, Wilson (1931) proposed the now well-known energy-band picture of semiconductors and insulators (full valence bands followed by an energy gap, followed by empty conduction bands) and explained the basic difference between metallic and semiconducting conductivity as a function of temperature. At about the same time, experimental evidence became available that the observed conductivity of semiconductors was entirely due to the presence of impurities (Gudden, 1931). Immediately, Wilson (1932) proposed a qualitative explanation in terms of the energy levels introduced by impurity atoms in the otherwise forbidden energy gaps of semiconductors. Wilson's 1932 paper is thus the first quantum-mechanical theory of impurity states in semiconductors.

Wilson's model is an extreme tight-binding model, but presents a useful qualitative picture. The solid is assumed to be a collection of atoms on a periodic lattice and the energy bands are viewed as broadened atomic energy levels. An impurity atom, which has a different set of energy levels, could thus happen to have its highest occupied level lie within the energy gap between the full valence bands and the empty conduction bands, which are not affected by the presence of the impurity. The impurity electron can then be thermally excited into the empty bands where it will conduct.

Wilson's picture of a donor impurity is usually valid only in a qualitative way. The tight-binding picture can be a basis for quantitative calculations for deep levels, as we shall see in Secs. IX and X. Nevertheless, the tight-binding picture is not the proper framework for quantitative understanding of the shallow levels which are the ones that dominate conduction in semiconductors. As we shall soon see, the so-called effective-mass or hydrogenic model is more appropriate for this purpose.

During the 1930's, a limited amount of theoretical work was done to understand the behavior of impurities in semiconductors. Wannier in 1937 described the motion of an electron near the bottom of the conduction band and of a hole near the top of the valence bands in terms of *effective masses* and showed that during optical absorption the continuum of interband transitions would be preceded by discrete lines due to the formation of *excitons,* on account of the Coulomb interaction between the electron and the hole. The calculation of the exciton levels was shown to be isomorphic to the hydrogen atom, except that the effective rydberg is now reduced by the value of the effective mass and by the dielectric constant which screens the Coulomb force between electrons and holes. Exciton binding energies are thus a few meV, compared with 13.6 eV, the binding energy of the electron in the hydrogen atom.

Wannier did not discuss impurities, however. The first description of what came to be known as the effective-mass or hydrogenic theory for impurities is given by Mott and Gurney in their classic 1940 book *Electronic Process in Ionic Crystals.* Mott and Gurney's discussion of semiconductors is rather interesting in that it shows the state of the art at the time. The basic aspects of donor and acceptor impurities were described, but the terminology was not yet in use. The only type of donor impurities Mott and Gurney described, however, were interstitials (both host atoms and foreign atoms) and F-center-type defects (negative-ion vacancies in ionic crystals). The latter, though behaving as donors in principle, were not known to contribute to conductivity. Perhaps even more fascinating is the fact that though hole conductivity was understood, it was termed abnormal, and the only acceptor-type impurity discussed was the positive-ion vacancy. It should be noted that the only semiconductors that had been investigated extensively by then were oxides such as Cu_2O, ZnO, U_2O, etc., which are today known to be quite hard to work with, compared with silicon, germanium, etc. It was known at the time that most oxides could show either "normal" or "abnormal" conductivity, i.e., in today's terminology, could be made either n-type or p-type.

Much of the terminology and understanding of the behavior of semiconductors containing impurities was developed during the secretive years of World War II and allotment of credit becomes a very difficult task. It seems, however, that the first estimate of the binding energy of donors in a semiconductor, namely Si, in terms of the hydrogenic model, and the explanation of relevant data was done by Beth (1942).[2] By the end of the war, donor and acceptor states in semiconductors were well understood, and this understanding was a contributing factor to the invention of the transistor (Bardeen and Brattain, 1948, 1949). In the late 1940's and early 1950's, many investigators carried out rig-

[2] I found a copy of this war-time report some years ago at the Engineering Library of the University of Illinois. It bears the word "CONFIDENTIAL" in big red letters on every page. Donors are referred to as donators, which is the German word for the concept.

orous derivations of the effective-mass equations for realistic band structures and accurate numerical calculations. Some of the landmark papers were by Slater (1949), James (1949), Kittel and Mitchell (1954), Luttinger and Kohn (1955), Kohn and Luttinger (1955), and, finally, the classic review article of Kohn (1957), which is probably still one of the most widely referenced papers in semiconductor physics.

We shall now see what the elementary aspects of the effective-mass or hydrogenic model are. Rigorous derivations of these results will be discussed in subsequent sections. Let us start with the well-known band picture for the electronic energy levels in a pure and perfect crystalline semiconductor. We have the valence bands, which at 0°K are completely full, an energy gap, and the conduction bands, which are completely empty. Let us then introduce an extra electron in this crystal. At 0°K, this electron will occupy the lowest energy state available to it, which by definition is the bottom of the conduction band. In the language of band theory, the wave functions of band states are characterized by a propagation vector (wave vector) **k**. For simple bands, the energies of the states near the bottom of the bands are given by the simple formula

$$E(k)=E_c+\hbar^2k^2/2m_e^*, \tag{3.1}$$

where E_c is the energy at the band minimum (it may be taken to be the zero of the energy scale), and m_e^* is known as the effective mass. This result should be compared with the corresponding result for an electron in free space. In that case, the wave functions are also characterized by a wave vector **k** and the corresponding energies are given by

$$E(k)=\hbar^2k^2/2m_0, \tag{3.2}$$

where m_0 is the mass of a free electron. This explains the terminology for the constant m_e^* in Eq. (3.1). The effective mass embodies in an average way the effect of the crystal potential so that the dynamical behavior of an extra electron in the conduction bands under the influence of external forces is the same as that of a free electron with a mass m_e^*.

Let us now go back to the perfect and pure material and replace one of the host atoms by an atom from the column of the periodic table next to the column of the host atom. The impurity atom then has one extra valence electron. Let us assume for a moment that somehow we held onto that extra electron and did not let it enter the crystal. The crystal, then, has the same number of valence electrons as it had before, so that it still has completely full valence bands and completely empty conduction bands. The only significant change is that although all host atoms are on the average neutral, the impurity atom is by necessity positively charged. It therefore sets up a Coulomb field in addition to all the other crystal fields that existed before the introduction of the impurity. Since the positive charge is in a dielectric medium, the Coulomb potential is not given by $U(r)=e/r$ but by $U(r)=e/\epsilon r$, where ϵ is the dielectric constant of the material. Let us then introduce the extra electron in the crystal, so the whole sample is once more neutral. According to the above discussion, we can then view it as a free electron with mass m_e^*, which is now acted upon by the Coulombic field $e/\epsilon r$. The situation is therefore isomorphic to the hydrogen atom except that now the "proton" has a charge equal to e/ϵ and the electron has a mass m_e^*. Since the hydrogen atom has bound states below the ionization continuum whose energies are given by

$$E_n^{(H)}=-e^4m_0/2\hbar^2n^2, \tag{3.3}$$

where $n=1,2,3,\ldots\infty$, we conclude immediately that bound states for the electron are introduced below the conduction-band edge at energies given by

$$E_n=E_c-e^4m_e^*/2\hbar^2\epsilon^2n^2, \tag{3.4}$$

or

$$E_n=E_c-E_n^{(H)}(m_e^*/\epsilon^2), \tag{3.5}$$

where m_e^* is in units of m_0, and where again $n=1,2,3,\ldots\infty$. Since typical values of ϵ are of order 10, and values of m_e^* range from about $0.03m_0$ to about m_0, Eq. (3.5) shows that the ionization energy of the donor energy level (equal to E_1) ranges from 10^{-4} to 10^{-2} Ry or from about 50 to about 130 meV. Compared with band gaps of order 1eV or more, such hydrogenic levels are clearly shallow.

The qualitative picture for acceptors is analogous to that for donors. One first considers removing an electron from the otherwise perfect and pure crystal whereby a hole is introduced in the otherwise full valence bands. The dynamics of the hole near the top of the valence bands are again describable in terms of an expression similar to (3.1) but now with a minus sign, i.e.,

$$E(k)=E_v-\hbar^2k^2/2m_h^*, \tag{3.6}$$

where now E_v is the top of the valence bands, and m_h^* is the effective mass of the hole. When an impurity with fewer valence electrons than the host is introduced, a negatively charged center is created, setting up a screened Coulomb potential to which positively charged holes are attracted. Hydrogenic energy levels are therefore once more introduced in the band gap, this time above the top of the valence bands, given by

$$E_n=E_v+E_n^{(H)}(m_h^*/\epsilon^2), \tag{3.7}$$

where again m_h^* is in units of m_0. Numerical estimates for the binding energies of holes are similar to those described above for electrons bound to donor impurities.

The hydrogenic model, whose rigorous foundations will be discussed in Sec. VII, is a valid description of those impurities which introduce approximately Coulombic potentials in the crystal, such as nonisovalent chemical impurities. Its quantitative success varies from excellent to poor, depending on the complexity of the energy bands and the values of m^* and ϵ. It does best when the relevant band extremum is of the simple form (3.1) or (3.6) and for single donors and single acceptors. The hydrogenic model can be extended to describe double donors and double acceptors by using a screened Coulomb potential of two charges, whereby the energy levels are simply four times deeper. In general, the excited states come out reasonably well, but binding energies calculated from such formulas for double donors and acceptors are smaller than experimental values by a factor larger than two and as much as ten or more. Effective-mass theory, however, which is the rigorous version of the above qualitative theory, remains valid for many potentials which are not Coulombic in the immediate vicinity of the impurity, and gives good quantitative results for many shallow as well as some moderately deep levels. The theory will be discussed in detail in Sec. VIII. For most deep-level impurities and defects, however, effective-mass theory is inadequate and other techniques have been developed. These theories will be discussed in Secs. IX and X.

IV. THE ROLE OF IMPURITIES IN SEMICONDUCTORS

Before we go on to examine experimental and theoretical methods that are used to determine the electronic energy levels and other properties of impurities in semiconductors, we wish to discuss briefly the role that impurities play in the various applications of semiconductors.

The role a given impurity can play in a semiconductor depends strongly on the kind of localized energy levels it introduces in the otherwise forbidden band gap, on the concentration with which it can be incorporated in a sample, and on the nature of other impurities present in the sample. By far the most important role of shallow donors and acceptors is to control conductivity. At room temperature, almost all such impurities are ionized and contribute to the conductivity because their ionization energies are comparable to k_BT (k_B= Boltzmann's constant). They also contribute to resistivity as scattering centers, but this effect is secondary. Most shallow donors and acceptors can be incorporated in semiconductors in arbitrary concentrations up to about one part per thousand (i.e., about 10^{20} cm^{-3}). The range of conductivities that can be attained at room temperature is, therefore, enormous, namely about twelve orders of magnitude, from about 10^{-9} $(\text{ohm cm})^{-1}$ to about 10^3 $(\text{ohm cm})^{-1}$. This should be compared with metals, all of which have conductivities of order 10^6 $(\text{ohm cm})^{-1}$. In good insulators, conductivity can be as low as 10^{-22} $(\text{ohm cm})^{-1}$. Furthermore, conductivity in most semiconductors can be dominated by either electrons (n-type) or holes (p-type).

What really makes semiconductors useful, however, is the fact that concentrations of shallow donor and acceptor impurities can be made nonuniform in carefully chosen and controlled ways. By judicious choices of such inhomogeneities, the densities and currents of electrons and holes, both in the absence and in the presence of applied electrostatic potentials, can be exploited to produce a variety of effects that can be used in devices. The simplest example of an inhomogeneous semiconductor is a p-n junction, which consists of a region doped with acceptors (i.e., p-type) adjacent to a region doped with donors (i.e., n-type). This arrangement is such that it allows the flow of current only in one of the two directions perpendicular to the interface and the device acts as a rectifier of alternating current. When some other conditions are also met, a p-n junction may act as an emitter of radiation, either as a laser or as a light-emitting diode (LED). The transistor has three consecutive regions, p-n-p or n-p-n, and it acts as an amplifier of signals. Other devices, such as modulators, detectors, photocells, etc., consist of similar or more complicated structures formed with n- and p-type regions of various donor and acceptor concentrations. The physics of these devices is beyond the scope of this paper and will not be discussed.

Deep-level impurities, on the other hand, play an entirely different role. In general, they can be incorporated in a crystal in smaller concentrations, usually of order 10^{12}-10^{13} cm^{-3} in the case of Si, or as high as 10^{17} cm^{-3} in some compounds. They usually contribute negligibly to the concentration of current carriers. Instead, their function in most cases is to act as a catalyst for the recombination of electrons and holes. They accomplish this by providing a level somewhere in the middle of the band gap. Since for an electron and a hole to recombine (i.e., for an electron in the conduction bands to drop into an empty state in the valence bands), an amount of energy equal to the band gap must be dissipated, recombination is made more likely if that energy can be dissipated in smaller fractions. In fact the most efficient recombination centers are those deep-level impurities whose ionization energy is of order half the band gap and which at the same time have a series of hydrogenic excited states near one or the other of the band edges. The electron (or the hole) can then cascade through these excited states by losing energy to the lattice in small quantities at a time, i.e., by emitting phonons whose energies are equal to separations between excited states, thus making the recombination cross section larger.

In view of their function as recombination centers, the most important role of deep-level impurities is to control the lifetime of carriers. Clearly, then, if the device calls for long carrier lifetimes, deep impurities must be avoided. An example of this case is a photocell which is used for the conversion of solar energy into electrical energy. When sunlight strikes the cell, it generates electron-hole pairs by exciting electrons from the valence bands into the conduction bands. It is important that the lifetime of the carriers be long so that they can be drifted to their respective electrodes for collection without substantial loss. Another example is a junction laser where the presence of deep-level recombination centers would limit the efficiency. On the other hand, when the device calls for short carrier lifetimes, deep-level impurities must be judiciously incorporated in the device. An example of such a case is a photocell which is used as a fast switch. When light strikes the cell, it generates electron–hole pairs which produce conductivity (the process is called photoconduction). For a fast switch, conduction of current must last only for a very short period of time, so efficient recombination centers are necessary to destroy the carriers quickly.

The above discussion of shallow versus deep impurities referred to nonisovalent impurities. Isovalent impurities belong to a class of their own. They also have useful applications, in particular in the manufacture of LED's. In order to produce LED's of a given color, a material must be found whose band gap is equal to an energy corresponding to the desired wavelength of radiation. Moreover, the gap must be direct (conduction-band minimum and valence-band maximum at the same **k** vector) for efficient radiative recombination. The material must also be suitable for the fabrication of a *p-n* junction. If a material is found which satisfies all the required criteria but has an indirect gap, the situation can be rectified if an isovalent impurity introduces a level near one of the band edges. As we shall see in Sec. IX, such a level has a wave function which can be constructed as a linear combination of band wave functions from large regions of k space. In particular, if the localized wave function contains substantial contributions from the region of **k** space where the other band has its extremum, recombination between the localized level and the other band becomes very efficient. Light is thus obtained at a wavelength corresponding to an energy slightly less than the band gap. The isovalent impurity thus "converts" the band gap from indirect to direct and produces efficient radiative recombination. The most notable example of such a situation is nitrogen in $GaAs_xP_{1-x}$ alloys, which is widely used for red and green diodes. The same end of "converting" the band gap from indirect to direct is sometimes accomplished by introducing donor-acceptor pairs in a judicious way. An example of this is GaP doped with Zn and O which, depending on the Zn–O separation, produces light of different colors.

Finally, most lattice defects, such as those introduced when a sample is irradiated, cause nothing but trouble. Those of an extended nature are particularly bad in that they usually make devices inoperable. The best-known such defect is the dark-line defect which is the primary source of degradation of GaAs-$Ga_{1-x}Al_x$As double heterostructure lasers (Kishino *et al.*, 1976, and references therein).

V. EXPERIMENTAL METHODS

Before we embark on a detailed description of the theoretical methods that have thus far been used to study the electronic structure and properties of impurities and defects in semiconductors, we turn to a brief survey of experimental methods. The most fundamental properties that can be measured are, of course, the positions of the energy levels and the wave functions of the states introduced by the impurity or defect in the otherwise forbidden energy gap of the semiconductor. The positions of localized energy levels are usually obtained experimentally by detecting transitions of electrons from the localized level of interest to other localized levels or to one of the band continua, or, conversely from one of the band continua to the localized level. Transitions may correspond to *excitation*, namely when electrons go to states of higher energy by absorbing energy, or to *de-excitation*, when electrons go to states of lower energy by releasing energy. The form in which the energy is supplied or released provides a convenient means to classify experiments.

A. Thermal experiments

At $T = 0°K$, all electrons occupy the lowest one-electron energy levels available to them and no transitions are possible in the absence of external perturbations. At finite temperatures, however, the lattice vibrates and thus stores energy which we sense as heat. In quantum-mechanical language, lattice vibrations correspond to phonons, which may interact with electrons and be annihilated, imparting their energy and momentum to electrons. Alternatively, the vibrating lattice emits black-body radiation in the form of light quanta (photons) which electrons may absorb. In the case of donor states, electrons may make transitions from a localized level to the conduction bands (electron emission) while some others drop back into empty localized levels (electron capture). At a given temperature, a steady state occurs and one can measure the density of excess carriers $n(T)$. Usually $n(T)$ is measured indirectly by measuring either the *conductivity* $\sigma(T)$ or the *Hall coefficient* $R_H(T)$, both of which are simple functions of $n(T)$. In the simplest case, $n(T)$ obeys the activation formula

$$n(T) = A(T)\exp(-E_A/k_B T), \tag{5.1}$$

where E_A is referred to as the *thermal activation energy*, and corresponds to the ionization energy defined in Sec. II. In the case of acceptor states, electrons make transitions from the valence bands into the localized level (hole emission) and vice versa (hole capture) and similar results follow. In general, one has both donors and acceptors and a more complicated analysis is required.

Conductivity and Hall-effect measurements were the first methods used to study semiconductors and in fact led to the conclusions that impurities dominated their electrical behavior (Wilson, 1932, and references therein). They have been particularly useful for shallow donors and acceptors. For deep levels, however, one has to work with high-resistivity materials (low concentration of shallow dopants) so that effects due to the deep levels themselves can be detected. In general, however, one may not be able to reduce the shallow-dopant concentration to values that are substantially lower than the available concentrations of deep levels, which makes the experiments very hard. In recent years, *junction techniques* have been developed which eliminate the need of high-resistivity materials. Such a technique was first reported by Williams (1966), but the foundations of a large family of junction techniques were laid in independent work by Sah and his co-workers[3] (Sah, Forbes, Rosier, and Tasch, 1976; Sah, Chan, Fu, and Walker, 1972; for a review, see Sah, 1976, 1977a, 1977b). These techniques study deep-level impurities situated in the transition region of a *p-n* junction (Fig. 4) where the Fermi level lies somewhere in the middle of the gap and the shallow donor impurities are all ionized. By applying a reverse bias, the electrons are swept away, thus perfectly simulating a high-resistivity material. Electrons from the deep levels are then thermally excited to the conduction bands. One then has a choice of measuring the so-called dark current, or the change in the capacitance of the *p-n* junction caused by

[3] As an interesting historical note (Sah, 1976), the junction method evolved from a homework problem assigned by Sah to students in an undergraduate solid-state electronics course in the spring of 1965.

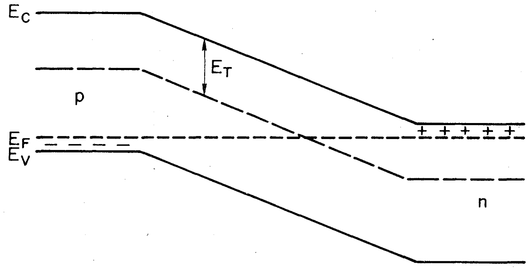

FIG. 4. Schematic representation of a *pn* junction with no external field. E_F is the Fermi level.

the depopulation of the deep levels. Once again, one ends up with an emission rate from which one can extract a thermal activation energy. In contrast to experiments done on bulk high-resistivity samples, measurements in *p-n* junctions are carried out in the presence of high electrostatic fields whose effect on the emission rates is not thoroughly understood (see recent review by Grimmeiss, 1977). Junction techniques, on the other hand, provide high versatility and there exist numerous variations, depending on how initial conditions are established, what one actually measures, etc. (Sah, *et al.*, 1970; Grimmeiss, 1977). Recently, junction techniques which allow one to detect energy-level positions as peaks in a continuous spectrum have been shown to be particularly efficient [such techniques have been referred to as deep level transient spectroscopy (DLTS). See Lang, 1974; Miller *et al.*, 1977; Grimmeiss, 1977.] We will not delve here into the subject of classifying the various junction techniques and comparing their technical advantages and disadvantages.

B. Optical experiments

Optical experiments differ from thermal experiments in that the energy supplied by the excitation or released by de-excitation is in the form of radiation. Transitions are now caused by an external photon field, instead of the phonon field. Optical experiments provide additional flexibility since photons, unlike phonons, can be supplied in a well-controlled manner, with frequencies and intensities of one's choice.

By analogy with thermal conductivity measurements, one can perform a photoconductivity experiment, in which carriers are excited from localized states into the band by absorbing light. By scanning through a continuum of light frequencies one measures the resulting conductivity, which should be zero for frequencies smaller than the separation between the localized level and the band edge. The earliest such measurements were by Burstein and co-workers (1951, 1953). The process of extracting an accurate threshold, however, which corresponds to the true ionization energy defined in Sec. II, is not quite straightforward because photoconductivity becomes nonzero in a somewhat gradual manner. Secondary processes, such as two-step photothermal ionization [a process in which electrons are optically excited to an excited state and then are thermally excited to the band edge (Lifshitz and Nad, 1965; Kogan and Sedunov, 1967)], may smear out the threshold in a hopeless manner.

Accurate optical measurements are usually made for shallow impurities by directly measuring the *absorption coefficient* as a function of photon frequency, a process which is not possible in thermal experiments. Sharp peaks are then obtained at frequencies corresponding to energy separations between the ground state and excited hydrogenic states. A typical spectrum is shown in Fig. 5. Since the positions of the excited states are well known from effective-mass theory (see Sec. VII for more details), band edges can be located very accurately. The method has proved extremely accurate for donors and acceptors in Si and Ge (Burstein, Picus, Henvis, and Wallis, 1956; Aggarwal and Ramdas, 1965; Jones and Fisher, 1965; Fisher and Ramdas, 1965), including some relatively deep ones (Ho and Ramdas, 1972; Kleiner and Krag, 1970). Similar measurements in compound semiconductors become complicated by the presence of many phonon sidebands arising from stronger coupling to the lattice in partially ionic materials (Ahlburn and Ramdas, 1968, 1969; Onton, 1969; Onton and Taylor, 1970; Carter, Dean, Skolnik, and Stradling, 1977).

Detailed information in compound semiconductors has been obtained mostly from photoluminescence measurements, which are the opposite of photoabsorption. In such experiments, electrons drop into lower-energy states and the energy is released in the form of photons, which are detected. Direct luminescence arising from a band-to-localized-level transition is not a very efficient technique in indirect-gap materials (Haynes and Westphal, 1956). A number of intriguing developments in the 1960's, however, made luminescence a very powerful technique. First, the so-called pair spectra were observed (Hopfield, Thomas, and Gershenzon, 1963), which correspond to an electron bound to a donor dropping into the bound state of an acceptor. One could thus extract accurate *sums* of acceptor and donor binding energies. An important breakthrough came in 1967 (Dean,

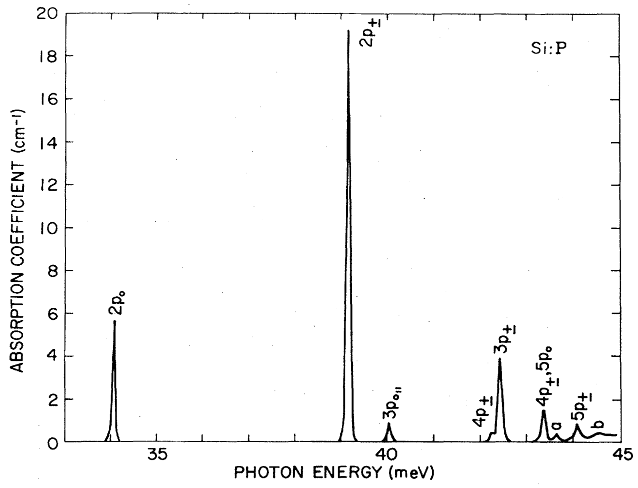

FIG. 5. The infrared-absorption spectrum of phosphorus-doped Si, as measured by Aggarwal and Ramdas (1956).

Cuthbert, Thomas, and Lynch) with the observation of two-electron partially-radiative transitions. In these experiments, an exciton bound to a donor recombines but the released energy is only partially released as a photon. The rest of the energy is picked up by a donor electron which goes into an excited state or the conduction band. The excited states reached by this type of experiment have even parity, unlike those reached by infrared absorption, which have odd parity. They can be used to obtain accurate donor binding energies. Accurate acceptor binding energies are then obtained from the pair spectra (Dean, Cuthbert, Thomas, and Lynch, 1967; Dean, 1968). Another example of luminescence is that reported by Dean and Henry (1968) in which they observed a transition from an excited state of the oxygen donor in GaP to the ground state of the *same* impurity atom.

More recently, with the advent of tunable dye lasers, a variation of luminescence spectroscopy has become very powerful. The method is called *luminescence excitation spectroscopy* (Street and Senske, 1976; Cohen and Sturge, 1977). It differs from conventional luminescence spectroscopy in the way the crystal is excited. Instead of exciting carriers by band gap irradiation, this technique uses a tunable dye laser whereby selected states can be excited at will and then allowed to decay radiatively. In this way, one can measure the luminescence intensity as a function of the excitation energy. Cohen and Sturge (1977) used this technique and were able to detect excited states of excitons bound to pairs of isovalent nitrogen impurities in GaP. Street and Senske (1976) showed that the method is unique in giving detailed information about binding energies and excited states of acceptors in a direct way. The method is basically the same as the luminescence of pair spectra described earlier. Instead of looking at all the pair lines in the luminescence spectrum, however, one selects a particular line and measures its intensity as a function of the excitation energy supplied by a dye laser. In this manner, the measured intensity has a peak whenever the excitation energy of the dye laser coincides with the energy separation between the donor ground state and one of the excited states of the acceptor. The complete acceptor spectrum is thus obtained without needing the band gap, the donor binding, or the donor–acceptor interaction energy. Furthermore, because transitions occur from one center to another, dipole selection rules do not operate and both s-like and p-like excited states can be detected. This feature is unique to this method. In contrast, infrared absorption reaches only p-like excited states, while "two-electron" luminescence (discussed above) and Raman spectroscopy (see below) reach only s-like excited states.

Raman spectroscopy makes use of inelastic light scattering, and thus corresponds to a combination of absorption and luminescence measurements. Light of a given energy is absorbed and light of a different energy is emitted. The difference is taken up or supplied by one or more excitations in the solid, depending on whether the energy of the emitted photon is smaller or larger than the energy of the absorbed photon. The *shift*, therefore, measures the energy of the excitation, which can be either a phonon or an electronic excitation or both. The technique can therefore be used to study the excited states and ionization energies of impurities. Unlike photoabsorption, however, Raman spectra reach excited states with the same parity as the ground state. The reason is that quantum mechanically Raman scattering is a two-step process. The first step is the absorption of the incoming photon with the system going into a virtual state of opposite parity. The second step is the emission of the outgoing photon with the system reaching a state of the same parity as the initial (ground) state. The technique was first successfully applied to electronic states in semiconductors by Henry, Hopfield, and Luther (1966), who studied the Zn and Mg acceptors in GaP, and has been used later by many others.

The techniques described above work well for shallow levels and on occasion for deep levels as well. For the latter, however, junction techniques are more widely used. As we mentioned earlier in the context of thermal experiments, there exist an immense variation of junction techniques depending on initial conditions and what one actually measures. The use of optical excitation or luminescence adds vast new flexibility. Apart from technical details, however, in all cases one ends up extracting optical emission or capture cross sections for either electrons or holes, which are proportional to the absorption coefficients that one might measure by direct light absorption in a bulk sample, except for the complication added by the high electric fields. In general, however, excited states are not observed, and the threshold for transitions to the bands is not easily determined.

C. What experiments really measure

The experimental techniques we discussed thus far detect transitions of electrons from one state to another, at some finite temperature. The transitions are induced either optically or thermally. Usually, one would like to be able to analyze such data and extract quantities that may also be obtainable from theoretical calculations. In particular, one would like to be able to construct an energy-level diagram at $T = 0°\mathrm{K}$.

The first complication arises from the fact that optical and thermal experiments do not measure the same transition. The difference was well understood quite early (deBoer and van Gell, 1935; Mott and Gurney, 1940): In an optical experiment, when an electron absorbs a photon and makes a transition to another state, the surrounding ions do not move in the process (Franck-Condon principle). The process is usually illustrated in terms of a configuration-coordinate diagram as in Fig. 6(a). Optical transitions are "vertical," leaving the lattice configuration unchanged. Once the transition occurs, the system is likely to relax to a new minimum-energy lattice configuration, before recombination occurs. As Fig. 6(a) illustrates, the energy of the emitted photon (CD) is then smaller than that of the absorbed photon (AB). Thermal experiments, on the other hand, do not measure either AB or CD. In such experiments, one detects the number of electrons in the upper level after equilibrium is reached at each temperature. The thermal activation energy corresponds to the energy dif-

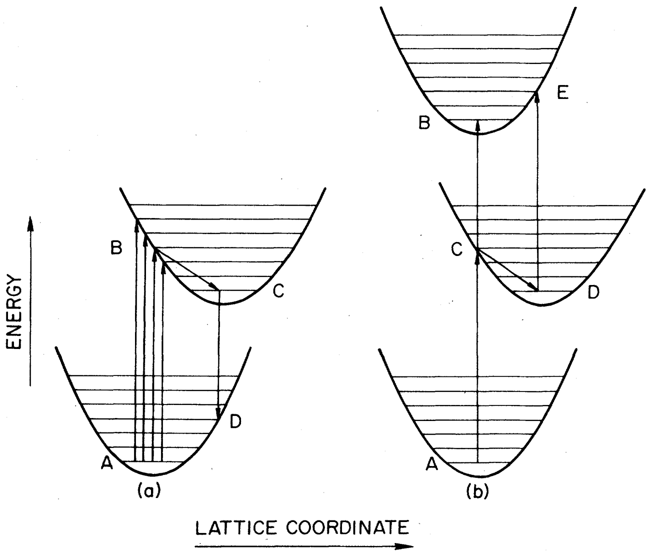

FIG. 6. Configuration-coordinate diagrams. (a) An optical transition corresponds to one of the arrows marked AB. The various arrows correspond to phonon sidebands. After the initial excitation, the electron may relax to the energy C so that the luminescence energy CD may be smaller than the absorption energy AB (Franck–Condon shift). (b) Two optical experiments would yield the two energies AC and DE which do not add up to the band gap AB.

ference AC, which is therefore always smaller than or equal to the optical threshold AB.

Another complication is due to the fact that it is possible during an optical experiment to create or destroy phonons (see, e.g., the various AB vertical lines in Fig. 6a). At low temperatures a discrete spectrum may appear for each electronic transition, known as a *phonon sideband*, whereas at high temperatures one simply has thermal broadening which can be discussed in classical terms. In all cases, the truly demanding task is to extract a threshold which corresponds to an electronic transition, for comparison with theoretical calculations.

In the case of shallow impurities in Si and Ge, low-temperature (liquid-helium) experiments have given clean, sharp absorption spectra which have been interpreted unambiguously. In the case of shallow impurities in the compound semiconductors, where electron-lattice interaction is stronger due to their partially ionic character, phonon sidebands complicate the spectra considerably, but the huge variety of experiments that have been possible have allowed the unambiguous extraction of energy levels. [For a compilation of unambiguously identified levels in Si, Ge, and some compounds, see the recent book by Watts, 1977.]

Deep levels are another matter, however. Most experiments are done in *p-n* junctions or Schottky barriers where high electric fields are present, and usually at temperatures between 100 and 400°K. Observed cross sections are therefore considerably broadened, and the extraction of energy levels in an unambiguous way is hindered. The procedure may also be complicated further by lattice relaxation being different for different charge states, which would cause the valence-to-trap and trap-to-conduction thresholds not to add up to the total band gap [Mott and Gurney, 1940; Kukimoto, Henry, and Merritt, 1973; Henry and Lang, 1977; see Fig. 6(b)]. A similar effect may be caused if one of the two thresholds does not correspond to transitions from or to the respective band extremum (White *et al.*, 1977).

The theory of the temperature dependence of energy levels and cross sections is rather primitive and not ripe for review. An elementary model has recently been proposed by Van Vechten and Thurmond (1976), in terms of which they analyzed available data and were led to new assignments for the observed centers in Au- and Co-doped Si (see below). On the other hand, theories of multiphonon processes have in general been available (Huang and Rhys, 1950; Lax, 1952; Gummel and Lax, 1955; Kubo and Toyozawa, 1955; Kovarskii, 1962; Kovarskii and Sinyavskii, 1962), but the use of such theories to analyze cross section line shapes is not a straightforward task. Valiant efforts have recently been made by Henry and Lang (1977) and by Samuelson and Monemar (1977) to include the effect of the lattice on bound-to-free and free-to-bound transitions and to interpret data on deep levels.

For deep levels, an even more serious problem is that of actually identifying the impurity or defect center one measures. Particular attention to this problem is paid by experimentalists in the analysis of data on radiation-induced defects (see, e.g., the review by Corbett, 1964), where it is the central issue. For chemical impurities, however, which are diffused in the sample in a controlled manner, the problem is often given only minimal consideration. Any measured levels in the gap are often associated with the particular impurity that was diffused in. No problems arise with impurities which can be incorporated at high concentrations and introduce shallow donor or acceptor levels. The risk is high, however, for impurities that can be incorporated only at intermediate concentrations of order 10^{12}–10^{14} cm^{-3}, as is generally the case with deep-level impurities. The risk has been highlighted by a series of experiments by Sah and coworkers (Yau and Sah, 1974, and references therein; Sah and Wang, 1975) who heat-treated a p^+n junction at 1200°C for several hours and then cooled it to room temperature and found that the process introduces two deep donor levels probably associated with the same center. The concentration of this center was determined to be 10^{13}–10^{14} cm^{-3}! Since impurities are often diffused in at high temperatures, the task of identifying measured levels with the impurity that was diffused in is highly demanding.

Carrying the line of thought one step further, even if enough experiments are carried out to ensure that a particular measured level is in fact associated with a certain chemical impurity, the task still remains to identify whether the center is a simple substitutional impurity, an interstitial, a vacancy-impurity complex, or a complex involving the ever-present shallow dopant of the *p-n* junction. A recent example of this problem is the analysis carried out by Fagelström and Grimmeiss (1977) on the published data on GaP:Cu. Cross sections published by several authors showed disturbingly different thresholds. A careful analysis, however, revealed that the different curves correspond to samples with a different shallow-donor dopant. Some experiments were repeated and it was confirmed that GaP:Cu reproducibly gives a cross section with a threshold that depends on the shallow-donor dopant. One might be tempted to conclude that the center is a Cu-shallow-donor complex, but, clearly, more experimental work is needed. For example, cross sections measured on uniaxially stressed samples or electron-spin-resonance (ESR) data may provide clues to the symmetry of the center.

Another case that raises similar questions is that of Si:Au. The data available in the literature are not in very good agreement with one another (see tables in the recent review by Grimmeiss, 1977). More recently, Lang (1977) performed similar measurements on a Au-doped Si sample but found dramatically different behavior. In particular he found a different threshold and much stronger temperature dependence than found by Engström and Grimmeiss (1975). In an effort to resolve the discrepancy, Lang and Grimmeiss exchanged samples and repeated measurements. They reproduced each other's data, establishing that the centers in the two samples are not the same. Interestingly, Van Vechten and Thurmond (1976) have already argued on theoretical grounds that the Au center in Si is not simple substitutional, as has been invariably assumed, but Au-vacancy complex.

The very recent developments described above demonstrate that a meticulous re-examination of what is known and accepted about deep-level centers in semiconductors is in order. New experiments will have to be done and more theories will have to be developed to help the process.

VI. THE QUANTUM-MECHANICAL PROBLEM: GENERAL PROPERTIES OF ELECTRONIC STATES

In this section we will give a general description of the electronic states in semiconductors containing a single isolated impurity. The remainder of the paper will be concerned with calculational techniques. The treatment in this section in the most part will apply to any defect of a localized nature; it will not apply to line or plane defects, as discussed in Sec. II. We begin by giving a summary of known results for the perfect crystal in order to establish terminology and set up the framework for the description of the imperfect crystal.

A. Electron states in perfect crystals

A semiconductor is in principle described by a "many-body" Hamiltonian $\mathcal{H}$, which describes the correlated motion of all the electrons and the nuclei in the sample. A convenient way to proceed, which is adequate for the purposes of this paper, is to make use of the Born-Oppenheimer approximation (see, e.g., Ziman, 1964, p. 169), whereby the nuclear motion can be separated from the electronic motion. One is then left with a "many-body" Hamiltonian for the electrons. From this Hamiltonian, one can deduce an effective one-electron Hamiltonian H^0 which describes in an approximate way the behavior of each electron in the presence of fields arising from both the nuclei and the other electrons. The most widely used form of the effective one-electron potential is that obtained by making the so-called local density approximation to exchange and correlation effects. This class of potentials includes both those used in the $X\alpha$ method (Slater, 1951, 1972) and those based on the use of the homogeneous electron gas to estimate exchange and correlation effects (Hohenberg and Kohn, 1964; Kohn and Sham, 1965; Hedin and Lundquist, 1971). The choice of effective one-electron potentials, while basic to all electronic-structure calculations, is neither specific to nor different in the study of impurity states; we shall therefore not discuss this issue further. Individual potentials, used in the applications of various methods for impurity states, will be discussed in the appropriate contexts.

The one-electron Hamiltonian can be written as

$$H^0 = T + V^0 , \tag{6.1}$$

where $T = -\hbar^2\nabla^2/2m_0$ is the kinetic energy (m_0 is the mass of a free electron), and V^0 is the one-electron potential, which has the full symmetry of the lattice. (Superscripts 0 will be used throughout to denote perfect-crystal quantities.) The corresponding eigenvalue problem is

$$H^0\psi^0_{n\mathbf{k}} = E^0_{n\mathbf{k}}\psi^0_{n\mathbf{k}} . \tag{6.2}$$

The solutions are characterized by the wave vector $\mathbf{k}$, which is a consequence of the periodicity of V^0. One could allow $\mathbf{k}$ to take any value in the three-dimensional $\mathbf{k}$ space, in which case the index n in (6.2) would not be necessary (this representation is known as the extended-zone representation). However, since $\mathbf{k}$ space is also periodic, it is possible to describe all states in terms of a $\mathbf{k}$ that lies within a primitive unit cell of $\mathbf{k}$ space, the Brillouin zone. The index n then becomes necessary to enumerate the various solutions at the same $\mathbf{k}$ (this representation is known as the reduced-zone representation).

The eigenvalues $E^0_{n\mathbf{k}}$ form the well-known bands of allowed energies, which are separated by energy gaps. The energies $E^0_{n\mathbf{k}}$ are usually plotted as a function of $\mathbf{k}$ along directions of high symmetry. For illustration purposes, the energy bands of Si are shown in Fig. 7. Near a band extremum at $\mathbf{k}_0$, $E^0_{n\mathbf{k}}$ can be expanded in terms of $\mathbf{k}-\mathbf{k}_0$ and the leading term is of order $(\mathbf{k}-\mathbf{k}_0)^2$. In the case of a single band, the expansion may be written in the form

$$E^0_{n\mathbf{k}} = E^0_{n\mathbf{k}_0} + (\hbar^2/2)(\mathbf{k}-\mathbf{k}_0)\overleftrightarrow{\mathrm{m}}^{-1}(\mathbf{k}-\mathbf{k}_0) , \tag{6.3}$$

where $\overleftrightarrow{\mathrm{m}}^{-1}$ is the tensor

$$\overleftrightarrow{\mathrm{m}}^{-1} = (1/\hbar^2)\,\nabla_{\mathbf{k}}\cdot(\nabla_{\mathbf{k}}E^0_{n\mathbf{k}})\,|_{\mathbf{k}=\mathbf{k}_0} . \tag{6.4}$$

In cubic materials at $\mathbf{k}_0 = 0$, symmetry requires that the tensor $\overleftrightarrow{\mathrm{m}}^{-1}$ be diagonal with $m^{-1}_{xx} = m^{-1}_{yy} = m^{-1}_{zz} = m^{*-1}$, whereby (6.3) becomes

$$E^0_{n\mathbf{k}} = E^0_{n0} + (\hbar^2/2m^*)k^2 , \tag{6.5}$$

and m^* is referred to as the effective mass. More complicated expressions obtain for extrema at $\mathbf{k}$ points with lower symmetry or at extrema of degenerate bands. These cases will be dealt with when needed, in Sec. VII.

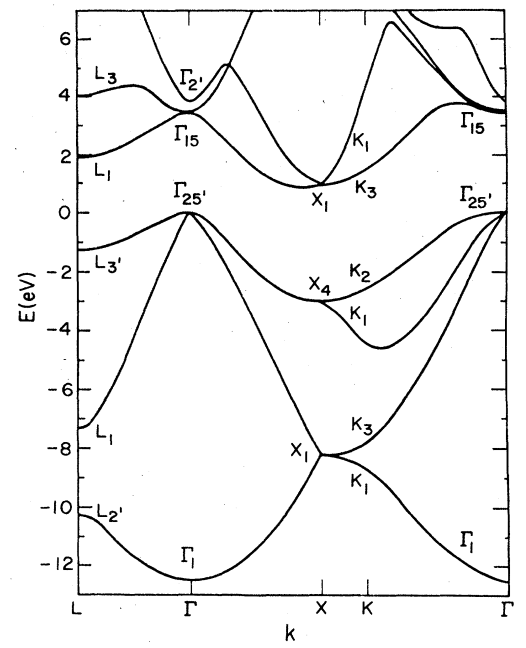

FIG. 7. The energy bands of silicon, calculated by using the empirical pseudopotential method and the potential form factors of Cohen and Bergstresser (1966).

The wave functions $\psi^0_{n\mathbf{k}}$ in (6.2) are known as Bloch functions, after Bloch (1928) who demonstrated that they can be written in the form

$$\psi^0_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u^0_{n\mathbf{k}}(\mathbf{r}) , \tag{6.6}$$

where $u^0_{n\mathbf{k}}(\mathbf{r})$ is a periodic function, i.e.,

$$u^0_{n\mathbf{k}}(\mathbf{r}+\mathbf{R}_j) = u^0_{n\mathbf{k}}(\mathbf{r}) , \tag{6.7}$$

where $\mathbf{R}_j$ is any lattice vector. The Bloch functions, therefore, extend over the whole crystal and have the same amplitude $|\psi^0_{n\mathbf{k}}(\mathbf{r})|^2$ in every unit cell.

In terms of the Bloch functions one may define another set of functions, the Wannier functions (Wannier, 1937)

$$w^0_n(\mathbf{r}-\mathbf{R}_j) = N^{-1/2} \sum_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{R}_j} \psi^0_{n\mathbf{k}}(\mathbf{r}) , \tag{6.8}$$

where N is the total number of lattice sites. By multiplying (6.6) by $e^{i\mathbf{k}\cdot\mathbf{R}_{j'}}$, summing over k and using the theorem (Reitz, 1955)

$$(1/N) \sum_k e^{i\mathbf{k}\cdot(\mathbf{R}_j-\mathbf{R}_{j'})} = \delta_{jj'} , \tag{6.9}$$

one immediately obtains

$$\psi^0_{n\mathbf{k}}(\mathbf{r}) = N^{-1/2} \sum_j e^{i\mathbf{k}\cdot\mathbf{R}_j} w^0_n(\mathbf{r}-\mathbf{R}_j) . \tag{6.10}$$

Both sets of functions are orthonormal, namely

$$\langle \psi^0_{n'\mathbf{k}'} | \psi^0_{n\mathbf{k}} \rangle = \delta_{n'n} \delta_{\mathbf{k}'\mathbf{k}} , \tag{6.11}$$

and

$$\langle w^0_{n'j'} | w^0_{nj} \rangle = \delta_{n'n} \delta_{j'j} . \tag{6.12}$$

In contrast to the Bloch functions, which extend over the whole crystal, the Wannier functions are localized about their corresponding site $\mathbf{R}_j$, except for some weak oscillations that persist at large distances from $\mathbf{R}_j$, which are necessary for the orthogonality expressed by (6.12).

Another quantity that is useful in the description of electronic states in crystals is the Green's function. It is defined formally, as an operator for energies E with an infinitesimal imaginary part by

$$G^0(E) = 1/(E-H^0) , \tag{6.13}$$

so that

$$G^0 \psi^0_{n\mathbf{k}} = [1/(E-E^0_{n\mathbf{k}})] \psi^0_{n\mathbf{k}} . \tag{6.14}$$

It can be expressed as a function of $\mathbf{r}$ and $\mathbf{r}'$

$$G^0(E; \mathbf{r}, \mathbf{r}') = \sum_{n\mathbf{k}} \frac{\psi^{0*}_{n\mathbf{k}}(\mathbf{r}) \psi^0_{n\mathbf{k}}(\mathbf{r}')}{E-E^0_{n\mathbf{k}}} . \tag{6.15}$$

Its matrix in the Bloch representation is diagonal, i.e.,

$$\langle \psi^0_{n'\mathbf{k}'} | G^0 | \psi^0_{n\mathbf{k}} \rangle = \delta_{n'n} \delta_{\mathbf{k}'\mathbf{k}} / (E-E^0_{n\mathbf{k}}) , \tag{6.16}$$

whereas in the Wannier representation it is

$$\langle w^0_{n'j'} | G^0 | w^0_{nj} \rangle = \delta_{n'n} N^{-1} \sum_{\mathbf{k}} \frac{e^{i\mathbf{k}\cdot(\mathbf{R}_{j'}-\mathbf{R}_j)}}{E-E^0_{n\mathbf{k}}} . \tag{6.17}$$

Any property of the system may be expressed in terms of the wave functions or in terms of the Green's function. Quite often it is a tradeoff between calculational simplicity and physical clarity. For our purposes, a useful property of the Green's function is that the imaginary part of its trace, which is independent of representation, may be identified with the density of states $D(E)$

$$D^0(E) = -\frac{1}{\pi} \operatorname{Im} \operatorname{Tr} G^0(E) . \tag{6.18}$$

This expression can be directly obtained from Eq. (6.16) by allowing E to have an infinitesimal imaginary part and using the Dirac result (Merzbacher, 1967, p. 490),

$$\lim_{\eta\to 0} \int \frac{1}{\omega \pm i\eta} = P \int \frac{1}{\omega} \mp \pi i \delta(\omega) , \tag{6.19}$$

and the conventional definition of $D^0(E)$,

$$D^0(E) = \sum_{n\mathbf{k}} \delta(E-E^0_{n\mathbf{k}}) . \tag{6.20}$$

In the Wannier representation, (6.18) becomes

$$D^0(E) = -\frac{1}{\pi} \sum_n \operatorname{Im} \langle w^0_{n0} | G^0 | w^0_{n0} \rangle \tag{6.21}$$

since $\langle w^0_{nj} | G^0 | w^0_{nj} \rangle$ is independent of j.

B. Electron states in imperfect crystals

As in the case of the perfect crystal, it is adequate for our purposes to describe the imperfect crystal in the one-electron approximation. (Many-body effects will be discussed in Sec. X.) The corresponding Hamiltonian H may be written as

$$H = T + V , \tag{6.22}$$

where T is again the kinetic energy as in (6.1), and V is the new one-electron potential. (As a matter of convention in notation, all quantities with superscript zero will refer to quantities in the perfect crystal. The corresponding quantity in the imperfect crystal will be denoted without a superscript. Quantities which have the same expression in both crystals, e.g., the kinetic energy T, will have no superscript.) V may be written as

$$V = V^0 + U\,, \tag{6.23}$$

where U represents the change or perturbation introduced by the impurity, and will therefore be referred to as the *impurity potential*. In view of (6.1) and (6.23), H may also be written as

$$H = H^0 + U\,. \tag{6.24}$$

The eigenvalue problem for the imperfect crystal is

$$H\psi_\nu = E_\nu \psi_\nu\,, \tag{6.25}$$

and our task here is to describe the nature of its solutions.

We observe immediately that two types of solutions may exist: (a) those with energy E_ν within the allowed energy bands of the perfect crystal; and (b) those with energy E_ν within the forbidden band gaps.

1. States within the band gaps

The first question is one of existence, and we distinguish two cases, namely impurity potentials which extend over an effectively finite range, and impurity potentials which decay as c/r, where c is a constant.

In the first case, states may or may not exist within a given gap, depending on the strength of the impurity potential and the nature of the energy bands. If they do exist, their wave functions decay exponentially far from the impurity or defect, and they are therefore referred to as bound states. In order to see that gap states must be bound, we note that the wave function ψ_ν far from the impurity must be expressible as a linear combination of all the eigenfunctions of H^0 at the energy E_ν. However, H^0 has no propagating solutions in the gaps; instead, the only solutions correspond to imaginary k's, so that they either decay or grow exponentially. The latter solutions are physically ruled out, thus requiring that ψ_ν for a gap state is localized.

The case of an impurity potential with a Coulombic c/r tail is more complicated in that its effect remains non-negligible all the way to infinity. An elegant proof may be given, however, (see Mott and Gurney, 1940) that if $c<0$ (potential attractive to electrons) an infinite number of bound states exists in the gap below the upper-band edge. Alternatively, if $c>0$ (potential attractive to holes), an infinite number of bound states exists in the gap above the lower-band edge. In either case, the states closest to the band edge are hydrogenic in nature and may be assigned an integral quantum number n. As $n\to\infty$, the orbit of the bound state tends to infinity so that bound and propagating states merge in a continuous way at the band edge.

Formally, the bound-state solutions of (6.25) may be written down by first rewriting (6.25) as

$$(E_\nu - H^0)\psi_\nu = U\psi_\nu\,, \tag{6.26}$$

and then using the definition (6.13). We get

$$\psi_\nu = G^0(E_\nu)\,U\psi_\nu\,. \tag{6.27}$$

Note that $G^0(E_\nu)$ is a well-defined real function for all real energies E_ν in the gaps and (6.27) is simply an integral equation for $\psi_\nu(r)$. By expressing the operators G^0 and U in any representation, such as the Wannier representation, (6.27) becomes a set of linear algebraic equations for the expansion coefficients of ψ_ν in that representation, whereby the bound-state energies are the zeroes of the determinant of $1-G^0U$. Since the determinant of an operator is independent of representation, any basis set can be used to represent the operators G^0 and U. We conclude that the criterion for the existence of bound states in the gap is that the determinant of $1-G^0U$ as a function of energy must go through zero. Most of the remainder of this paper will be devoted to methods of determining bound-state energies and wave functions for a variety of impurity potentials.

2. States within the energy bands

Within the region of the energy bands of the perfect crystal, a state with energy E_ν is degenerate with an energy E^0_{nk}, whereby the most general solution of (6.26) is no longer (6.27), but

$$\psi_\nu = \psi^0_{n\mathbf{k}} + G^0(E_\nu)U\psi_\nu\,, \tag{6.28}$$

where $\psi^0_{n\mathbf{k}}$ is a Bloch function corresponding to the energy $E^0_{n\mathbf{k}} = E_\nu$. Here $G^0(E_\nu)$ is understood to mean $\lim_{\eta\to 0} G^0(E_\nu + i\eta)$.

The states ψ_ν described by (6.28) are clearly scattering solutions, as they asymptotically approach one of the unperturbed solutions $\psi^0_{n\mathbf{k}}$. In fact, Eq. (6.28) is the solid-state analog of the Lippman–Schwinger equation of scattering theory. An elaborate theory exists for these states (Koster, 1954; Lifshitz and Kaganov, 1959; Callaway, 1964; Preziosi, 1971; Garcia-Moliner, 1971; see also Bassani, Iadonisi, and Preziosi, 1974) constructed along the lines of formal scattering theory for free electrons. Thus one defines scattering amplitudes, scattering cross sections, and phase shifts by complete analogy with free-electron results, except that the algebra and the details are considerably more complicated. We will not review that literature, but, instead, refer the interested reader to the original literature cited above. Many applications of scattering theory are actually done in the effective-mass approximation, when one can use all the results of free-electron scattering theory. The theory yields scattering cross sections for impurities and hence lifetimes of carriers caused by impurity scattering and mobilities (Conwell and Weisskopf, 1950; Brooks, 1951, 1955).

We turn to another aspect of the problem which is important to the overall purposes of this paper, namely the changes in the distribution of states, i.e., density of states, caused by the impurity potential. For this purpose we define the Green's function $G(E)$ for the imperfect crystal, by analogy to (6.13):

$$G(E) = 1/(E-H)\,. \tag{6.29}$$

Again, E is viewed as a complex variable, and for applications, its small imaginary part will be let go to zero in an appropriate way. The new density of states is given by

$$D(E) = -(1/\pi)\,\mathrm{Im}\,\mathrm{Tr}\,G(E)\,. \tag{6.30}$$

By using the fact that $dG^{-1}/dE = 1$, one can write

$$G(E) = -(d/dE)\ln G(E)\,. \tag{6.31}$$

The trace in Eq. (6.30) can then be evaluated in a representation which diagonalizes G (any representation would give the same result since Tr is an invariant) to get

$$D(E) = \frac{1}{\pi}\,\mathrm{Im}\frac{d}{dE}\ln\det G(E)\,, \tag{6.32}$$

where, again, the determinant $\det G(E)$, as an invariant, may be evaluated in any representation. An identical expression holds for $D^0(E)$ in terms of $G^0(E)$.

Now, by using (6.24) and (6.13) in (6.29) one gets Dyson's equation for G, namely

$$G = G^0 + G^0 UG\,, \tag{6.33}$$

which can be solved formally to give

$$G = (1 - G^0 U)^{-1} G^0\,. \tag{6.34}$$

Using this in (6.32), we get

$$D(E) = D^0(E) + \Delta D(E)\,,$$

where ΔD, the *change* in the density of states, is given by

$$\Delta D(E) = \frac{1}{\pi}\,\mathrm{Im}\frac{d}{dE}\ln\det(1 - G^0 U)\,. \tag{6.35}$$

We note that the operator $1 - G^0 U$ enters once more, and find it convenient to denote it by Q.

An important theorem, which is a consequence of the analytic properties of Q, is that

$$\int_{-\infty}^{\infty} \Delta D(E)dE = 0\,. \tag{6.36}$$

(See Garcia-Moliner, 1971.) It is known as Levinson's theorem and expresses the conservation of states. It means that for every state appearing in a band gap, a state must be missing in the bands. Therefore, if we have a total of N_b bound states in gaps, (6.36) becomes

$$\int_{\text{bands}} \Delta D(E)dE = -N_b\,, \tag{6.37}$$

where the integral is now only over the regions of the energy bands.

In many cases, the function $\Delta D(E)$ is negligible over most of the energy axis within the bands and is appreciable only over restricted regions of energy. When the integral over such a region is a negative integer, compensating for one or more of the bound states in the gaps, such a region of energy is called an *antiresonance*. Alternatively, when the integral is a positive integer, the region is called a resonance. Resonances, just like bound states, must be compensated by antiresonances, in order to satisfy (6.36). As far as charge is concerned, the wave functions ψ_v in a resonance region are such that they build up a charge equal to the integral of $\Delta D(E)$ in the vicinity of the impurity or defect. Similarly, the wave functions ψ_v in an antiresonance region correspond to a depletion of charge in the vicinity of the impurity or defect. $\Delta D(E)$, therefore, represents a change in the local density of states in the vicinity of the impurity or defect.

An interesting connection with free-electron scattering theory may be made by noting that if z is the complex number $a + ib$, then $\mathrm{Im}\ln z = \tan^{-1}(b/a)$. We therefore define the phase shift (Callaway, 1967) $\delta(E)$ by

$$\delta(E) = -\tan^{-1}(\mathrm{Im}\det Q/\mathrm{Re}\det Q)\,, \tag{6.38}$$

whereby (6.35) becomes

$$\Delta D(E) = \frac{1}{\pi}\,\frac{d\delta(E)}{dE}\,. \tag{6.39}$$

Therefore an extra state is introduced or removed in every energy interval in which $\delta(E)$ changes by π. The position of the resonance or antiresonance is usually defined to be the center of such an interval, whereby resonances and antiresonances are located at positions where $\delta(E) = m(\pi/2)$, with $m = 1, 3, 5\ldots$. According to this result, and using (6.38), resonances and antiresonances occur when $\mathrm{Re}\det Q = 0$. Note that this condition is identical with the one defining bound states in the gaps. The only difference is that in the gap $\mathrm{Im}\det Q = 0$, whereby $\Delta D(E)$ is a δ function at the energy of the bound state. For resonances and antiresonances, this is not so, and it is therefore interesting to determine the form of $\mathrm{Im}\det Q$ [i.e., the form of $\Delta D(E)$] in those regions of energy. In order to do that, we let E_0 be the value where $\mathrm{Re}\det Q = 0$. We then expand $\det Q(E)$ about E_0, letting E be a complex variable and seeking the value $E = E_r + i\Gamma$ where $\det Q(E) = 0$. This result is (Callaway, 1967)

$$E_r = E_0 - II'/(R'^2 + I'^2) \simeq E_0 - II'/R'^2\,, \tag{6.40}$$

$$\Gamma = 2IR'/(R'^2 + I'^2) \simeq 2I/R'\,, \tag{6.41}$$

where we have used $I = \mathrm{Im}\det Q$, $R = \mathrm{Re}\det Q$, and the primes denote derivatives with respect to energy. The approximate forms are for $R' \gg I'$. If we now expand $\delta(E)$ about E_0, we get for $\Delta D(E)$

$$\Delta D(E) \simeq \frac{\Gamma}{2\pi}\,\frac{1}{(E - E_0)^2 + \Gamma^2/4} \tag{6.42}$$

which is the well-known Lorentzian form. Note that it is the sign of Γ that determines whether we have a resonance or an antiresonance.

VII. HYDROGENIC EFFECTIVE-MASS THEORY

A. General results

Effective-mass theory (EMT) consists of a set of approximations which, in the case of bound states, allow the transformation of the eigenvalue problem (6.25) into an equivalent eigenvalue problem of the form

$$H_{\text{eff}} F_v(r) = E_v F_v(r)\,, \tag{7.1}$$

where H_{eff} is given by

$$H_{\text{eff}} = T_{\text{eff}} + U\,. \tag{7.2}$$

As we will see, the form of T_{eff} depends on the nature of the energy bands of the host crystal. By comparing (7.2) with (6.22) and (6.23), we see immediately that T_{eff} absorbs the effect of the periodic potential V^0.

The starting point for the derivation of (7.1) is the one-electron eigenvalue problem for the imperfect crystal, namely

$$H\psi_v = E_v\psi \,. \tag{7.3}$$

One can then proceed by expanding ψ_v in terms of either the Bloch functions $\psi^0_{n\mathbf{k}}$ of the host crystal, or the corresponding Wannier functions $w^0_n(\mathbf{r}-\mathbf{R}_j)$, or the set of functions $\psi^0_{n\mathbf{k}_0}(\mathbf{r})e^{i(\mathbf{k}-\mathbf{k}_0)\mathbf{r}}$, where $\mathbf{k}_0$ is some judiciously chosen point in the Brillouin zone. Detailed derivations for each such choice may be found in the original literature (Koster and Slater, 1954a; Luttinger and Kohn, 1955; Kittel and Mitchell, 1954; Kohn, 1957). We will follow here a simple derivation in order to identify the major approximations of the theory.

Let us then expand ψ_v in terms of the Bloch functions of the perfect crystal

$$\psi_v(\mathbf{r}) = \sum_{n\mathbf{k}} F_{n\mathbf{k}}\psi^0_{n\mathbf{k}}(\mathbf{r}) \,. \tag{7.4}$$

The expansion is exact as long as the sum is over all the bands, including the core bands. By substituting (7.4) in (7.3), making use of (6.2), multiplying on the left by $\psi^0_{n'\mathbf{k}'}$, integrating over all space using (6.11), and interchanging primed and unprimed symbols, one gets

$$E^0_{n\mathbf{k}}F_{n\mathbf{k}} + \sum_{n'\mathbf{k}'} \langle\psi^0_{n\mathbf{k}}|U|\psi^0_{n'\mathbf{k}'}\rangle F_{n'\mathbf{k}'} = E_v F_{n\mathbf{k}} \,. \tag{7.5}$$

This is an exact result. It may be viewed as a set of coupled linear algebraic equations for the coefficients $F_{n\mathbf{k}}$, or, by converting the sum over k in the Brillouin zone into an integral, as a set of coupled integral equations for $F_n(\mathbf{k})$.

Equation (7.5) may be developed further by using (6.6) for the Bloch functions, and expanding the product $u^0_{n\mathbf{k}}{}^*(\mathbf{r})u^0_{n'\mathbf{k}'}(\mathbf{r})$ in terms of plane waves. Since we are dealing with a periodic function, only the reciprocal lattice vectors $\mathbf{K}_p$ contribute to the expansion so that

$$u^0_{n\mathbf{k}}{}^*(\mathbf{r})\,u^0_{n'\mathbf{k}'}(\mathbf{r}) = \sum_p C^{nn'}_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p)e^{i\mathbf{K}_p\cdot\mathbf{r}} \,. \tag{7.6}$$

The matrix element in (7.5) then becomes

$$\langle\psi^0_{n\mathbf{k}}|U|\psi^0_{n'\mathbf{k}'}\rangle = \sum_p C^{nn'}_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p)U(\mathbf{k}-\mathbf{k}'-\mathbf{K}_p) \,, \tag{7.7}$$

where we have defined $U(\mathbf{q})$ to be the Fourier transform of $U(\mathbf{r})$

$$U(\mathbf{q}) = \int d^3r\, U(\mathbf{r})e^{-i\mathbf{q}\cdot\mathbf{r}} \,. \tag{7.8}$$

Note also that

$$C^{nn'}_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p) = \int d^3r\, u^*_{n\mathbf{k}}(\mathbf{r})u_{n'\mathbf{k}'}(\mathbf{r})e^{-i\mathbf{K}_p\cdot\mathbf{r}} \,. \tag{7.9}$$

Equation (7.5) can now be written as

$$E^0_{n\mathbf{k}}F_{n\mathbf{k}} + \sum_{n'\mathbf{k}'}\sum_p C^{nn'}_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p)U(\mathbf{k}-\mathbf{k}'-\mathbf{K}_p)\,F_{n'\mathbf{k}'} = EF_{n\mathbf{k}} \tag{7.10}$$

which is still exact.

In order to proceed further, approximations must be made and specific assumptions about the energy bands of the host crystal and the impurity potential must be made. We will consider for the time being only bound states in the fundamental energy gap of a semiconductor. To date, the only practically useful forms of the EMT are those that retain either only conduction bands or only valence bands in the expansion (7.4). More specifically, one usually retains only one band, unless more than one band is degenerate at the relevant band edge, in which case all the degenerate bands are retained. This approximation, of course, limits the range of applicability of the resulting equations to particular forms of the impurity potential. The theory was originally developed for the hydrogenic potential

$$U(r) = U_H(r) = -e^2/\epsilon r \tag{7.11}$$

for which the approximation can be justified in most materials. We therefore turn to the derivation of the effective-mass equations for several specific cases, having in mind the potential (7.11). We shall refer to this theory as the hydrogenic effective-mass theory (HEMT). Other potentials will be considered in Sec. VIII.

B. Simple bands

The simplest case to describe is that of a single band with a nondegenerate extremum at $\mathbf{k}=0$. Dropping the index n altogether, Eq. (7.10) becomes

$$E^0_{\mathbf{k}}F_{\mathbf{k}} + \sum_{\mathbf{k}}\sum_p C_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p)U(\mathbf{k}-\mathbf{k}'-\mathbf{K}_p)F_{\mathbf{k}'} = EF_{\mathbf{k}} \,. \tag{7.12}$$

At this stage, one anticipates solutions for which $F_{\mathbf{k}}$ is localized about $\mathbf{k}=0$ so that

$$|\mathbf{k}-\mathbf{k}'| \ll \mathbf{K}_p \,. \tag{7.13}$$

This assumption immediately leads to the following approximations:

(i) The $\mathbf{K}_p \neq 0$ terms in (7.12) are dropped, since

$$|U_H(\mathbf{k}-\mathbf{k}'-\mathbf{K}_p)| \ll |U_H(\mathbf{k}-\mathbf{k}')| \,, \tag{7.14}$$

a result which follows immediately from (7.13) and the fact that the Fourier transform of U_H is

$$U_H(q) = -4\pi e^2/\epsilon q^2 \,. \tag{7.15}$$

(ii) $C_{\mathbf{k}\mathbf{k}'}(0)$ may be approximated by

$$C_{\mathbf{k}\mathbf{k}'} \simeq C_{\mathbf{k}\mathbf{k}}(0) = 1 \,. \tag{7.16}$$

The second part of (7.16) is exact and follows from (7.9) and (6.11).

(iii) Sums over k and k′, which should be over the first Brillouin zone, may be converted into integrals over all of k space. We note in passing that when the sums over k and k′ are extended over all k space, the approximation (i) above no longer corresponds to dropping the $\mathbf{K}_p \neq 0$ terms in (7.12), but to retaining them, with all the constants $C_{\mathbf{k}\mathbf{k}'}(\mathbf{K})$ taken to be unity. The distinction should of course be immaterial if (7.14) is well satisfied. It would be relevant, however, if one were to seek corrections to approximation (i).

(iv) $E^0_{\mathbf{k}}$ is expanded about $\mathbf{k}=0$ to order k^2. In cubic materials, symmetry requires that a nondegenerate band at $\mathbf{k}=0$ be isotropic, whereby

$$E^0_{\mathbf{k}} \simeq E^0_0 + \hbar^2k^2/2m^* \,, \tag{7.17}$$

where m^* is the effective mass.

With the above approximations, (7.12) becomes

$$(\hbar^2k^2/2m^*)F(\mathbf{k})+\int d^3k\,U(\mathbf{k}-\mathbf{k}')F(\mathbf{k}')=E_BF(\mathbf{k})\,, \qquad (7.18)$$

where we have written $F(\mathbf{k})$ in place of $F_{\mathbf{k}}$ and $E_B=E-E_0^0$. This equation is immediately recognized as isomorphic to a Schrödinger equation in momentum space for a particle of mass m^* in the presence of the potential U. It can therefore be transformed to real space by defining

$$F(\mathbf{r})=\int d^3k\,F(\mathbf{k})\,e^{i\mathbf{k}\cdot\mathbf{r}}\,. \qquad (7.19)$$

The result is

$$[-\hbar^2\nabla^2/2m^*+U(\mathbf{r})]\,F(\mathbf{r})=E_BF(\mathbf{r})\,. \qquad (7.20)$$

For the potential $U_H(\mathbf{r})$, Eq. (7.20) is then isomorphic to the equation of the hydrogen atom and the solutions are therefore given by (3.8). The corresponding "wave functions" $F(\mathbf{r})$ are the hydrogenic wave functions. In particular, the "1s" ground state is given by

$$F(r)=\pi^{-1/2}a^{*-3/2}e^{-r/a^*}\,, \qquad (7.21)$$

where a^*, the effective Bohr radius is

$$a^*=\hbar^2\epsilon/m^*e^2=a_0\epsilon(m_0/m^*)\,, \qquad (7.22)$$

where a_0 is the Bohr radius of the hydrogen atom. We note that $F(k)$ for the ground state is given by

$$F(k)=8\pi^{1/2}a^{*-5/2}/[k^2+1/a^{*2}]^2 \qquad (7.23)$$

so that it extends appreciably in $\mathbf{k}$ space to approximately $\bar{k}\sim 1/a^*$. We also note that the complete wave function $\psi(\mathbf{r})$ is now given approximately by (Kohn, 1957)

$$\psi(\mathbf{r})\simeq F(\mathbf{r})u_0^0(\mathbf{r})=F(\mathbf{r})\psi_0^0(\mathbf{r})\,. \qquad (7.24)$$

The validity of the approximations can now be tested *ex post facto*. In particular, (7.14) is satisfied if

$$K_p^2\gg\bar{k}^2\,. \qquad (7.25)$$

Since K_p's are of order $2\pi/a$, (7.25) is satisfied if

$$(2\pi a^*)^2\gg a^2 \qquad (7.26)$$

and the error is of order $(a/2\pi a^*)^2$. A similar result is obtained for the other approximations since all of them are consistently good to order k^2. We conclude that the approximations are best for $a^*\gg a$, but (7.26) suggests that the approximations are also reasonable for values of a^* of the same order of magnitude as a. Finally, by using (7.22), (7.26) may be written as

$$2\pi\epsilon(m_0/m^*)\gg a/a_0\,. \qquad (7.27)$$

Since most lattice constants are of order $10a_0$ this result suggests that the EMT is valid for a hydrogenic potential if the ratio of the dielectric constant to the effective mass is larger than about two. Note, however, that the above statements refer to the validity of the EMT for a hydrogenic potential in a given crystal. They do not refer to the validity of the hydrogenic potential in describing real impurities in real crystals. This is a separate issue, which will be discussed later.

We are now in a position to discuss the approximations of retaining only one band. By second-order perturbation theory, we obtain for the correction to the one-band energy

$$\Delta E=\sum_{n'\mathbf{k}'}\frac{|\langle\psi|U|\psi^0_{n'\mathbf{k}'}\rangle|^2}{E^0_{n\mathbf{k}}-E^0_{n'\mathbf{k}'}}\,, \qquad (7.28)$$

whereby we see immediately that the sign of the correction depends only on whether the band is above or below the primary band (bands from above push the level down, bands from below push the level up) and the magnitude of the correction is inversely proportional to the band separation. In order to estimate (7.28), we use (7.4) and (7.7) and we immediately realize that we have two kinds of terms to consider. The first is for $\mathbf{k}'\sim\mathbf{k}$. For such terms

$$C^{nn'}_{\mathbf{kk}'}(0)\simeq C^{nn'}_{\mathbf{kk}}(0)=0\,, \qquad (7.29)$$

the second result following from (7.9) and (6.11). This is an important result since it shows that "direct" interband coupling is via the $\mathrm{K}_p\neq 0$ terms. By estimating

$$U_H(\mathrm{K})\sim(a/a^*)^2E_H\,, \qquad (7.30)$$

where E_H is the hydrogenic ground state, we get

$$\Delta E/E_H\sim(a/a^*)^4(E_H/E_g)\,, \qquad (7.31)$$

where E_g is an average interband separation. Thus if only $\mathbf{k}'\sim\mathbf{k}$ terms are important, the one-band approximation is well justified if E_H is a fraction of the band gap, no other bands are close by an energy of order E_H, and $a^*>a$. As we shall see later on, in many cases of interest, these requirements are well satisfied.

Terms with $\mathbf{k}'\neq\mathbf{k}$ in (7.28) may, however, spoil this. For such terms, (7.29) may not be satisfied. For particular $\mathbf{k}'$ values, $C^{nn'}_{\mathbf{kk}'}(0)$ may be considerably different from zero whereby such "indirect" interband coupling is via the $\mathrm{K}_p=0$ term of the potential which is not negligible. Such effects are not easy to estimate and must be addressed for individual materials.

C. Band with several equivalent extrema

The next case we consider is a band with several, say L, equivalent extrema along some crystallographic directions, away from $\mathbf{k}=0$. One can then start with (7.12), but $F_{\mathbf{k}}$ can no longer be assumed localized at a given extremum. Instead, one writes

$$F_{\mathbf{k}}=\sum_{i=1}^{L}\alpha_iF_{\mathbf{k}}^{(i)}\,, \qquad (7.32)$$

where $F_{\mathbf{k}}^{(i)}$ is assumed localized about the ith extremum and the constants are determined from symmetry considerations. Substituting (7.32) in (7.10) we get

$$E^0_{\mathbf{k}}\sum_i\alpha_iF_{\mathbf{k}}^{(i)}+\sum_{\mathbf{k}'}\sum_p C_{\mathbf{kk}'}(\mathrm{K}_p)\,U(\mathbf{k}-\mathbf{k}'-\mathrm{K}_p)\sum_i\alpha_iF_{\mathbf{k}}^{(i)}=E\sum_i\alpha_iF_{\mathbf{k}}^{(i)}\,. \qquad (7.33)$$

Note that now there will be so-called intravalley terms for $\mathbf{k}$ and $\mathbf{k}'$ near the same extremum (valley) and intervalley terms for $\mathbf{k}$ and $\mathbf{k}'$ near different extrema. In early applications of the theory, using $U_H(\mathbf{r})$, intervalley terms were neglected on the grounds that $F_{\mathbf{k}}^{(i)}$ at different valleys do not overlap substantially. The result then is a hydrogenic equation of the form (7.18) or (7.20), except that the kinetic energy is now anisotropic. The anisotropy arises from the fact that each extremum is at $\mathbf{k}\neq 0$. If the extrema are along a crystallographic direction, which may be defined to be the z axis, then $E_{\mathbf{k}}^{0}$ is expanded in the form

$$E_{\mathbf{k}}^{0}=E_{0}^{0}+(\hbar^{2}/2)\{(k_{x}^{2}+k_{y}^{2})/m_{t}^{*}+k_{z}^{2}/m_{l}^{*}\}, \tag{7.34}$$

where m_t^* and m_l^* are the transverse and longitudinal effective masses, respectively. The resultant one-valley EME is then

$$-\frac{\hbar^{2}}{2}\left\{\frac{1}{m_{t}^{*}}\left(\frac{\partial^{2}}{\partial x^{2}}+\frac{\partial^{2}}{\partial y^{2}}\right)+\frac{1}{m_{l}^{*}}\frac{\partial^{2}}{\partial z^{2}}\right\}F^{j}(\mathbf{r}) + U(\mathbf{r})F^{j}(\mathbf{r})=EF^{j}(\mathbf{r}). \tag{7.35}$$

Solutions of this equation can only be obtained variationally and we shall discuss them in the next subsection. Considerations concerning the validity of this approximation are similar to the simple band case except that $\bar{k}$ for each $F_{\mathbf{k}}^{i}$ must now be small compared with the intervalley separation, not the typical K's.

The total wave function is now given to first order in $\mathbf{k}$ by

$$\psi(\mathbf{r})=\sum_{j}\alpha_{j}F^{j}(\mathbf{r})\,\psi_{\mathbf{k}_{j}}^{0}(\mathbf{r}), \tag{7.36}$$

where

$$F^{j}(\mathbf{r})=\int d^{3}k\, e^{i(\mathbf{k}-\mathbf{k}_{j})\cdot\mathbf{r}}F^{j}(\mathbf{k}). \tag{7.37}$$

This many-valley form of $\psi(\mathbf{r})$ is required by symmetry (Kohn, 1957) even though intervalley coupling is neglected in the EME.

Many-valley effective-mass equations (MV EME) were first derived by Twose (reported by Fritzsche, 1962). The assumption that went into that derivation is that the same approximations used for the intravalley terms can also be used for the intervalley terms even though $\mathbf{k}-\mathbf{k}'$ is no longer small compared with $\mathbf{K}_p$. This assumption will be scrutinized in Sec. VIII. The MV EME is given by

$$\sum_{i}\alpha_{i}\,e^{i\mathbf{k}_{i}\cdot\mathbf{r}}\{T_{i}(-i\nabla)+U-E\}F^{i}(\mathbf{r})=0, \tag{7.38}$$

where $T_i(\mathbf{k})$ stands for the expression (7.34) with the z axis being in the direction of the ith extremum. Note that if intervalley terms are omitted, (7.38) reduces to (7.35). Note further that Eq. (7.38) contains only one unknown function, say $F^1(\mathbf{r})$, since the others can be related to it by symmetry operations.

D. Band with inequivalent extrema

The case of a band with more than one set of equivalent extrema so that members of one set are not equivalent with members of another set can be handled in a similar way. Let us assume two such sets for simplicity, whereby one can write, instead of (7.32)

$$F_{\mathbf{k}}=C_{1}\sum_{i=1}^{L_{1}}\alpha_{i}F_{1\mathbf{k}}^{(i)}+C_{2}\sum_{j=1}^{L_{2}}\beta_{j}F_{2\mathbf{k}}^{(j)}, \tag{7.39}$$

where now C_1 and C_2 are not determinable from symmetry. The result is two coupled equations of the form (7.38) (Bassani, Iadonisi, and Preziosi, 1969, 1974; Altarelli and Iadonisi, 1971).

E. Bands with a degenerate extremum

In this case, several bands are degenerate or nearly degenerate near the absolute extremum and all of them must be retained. The derivation of the effective-mass equations is more complicated and the reader is referred to the excellent original treatments by Kittel and Mitchell (1954) and Luttinger and Kohn (1955). The basic approximations are, however, the same, namely, that coupling with other bands is neglected and all $\mathbf{K}_p\neq 0$ terms are also dropped. The resultant effective-mass equations are of the form

$$\{D(-i\nabla)+U(\mathbf{r})I\}F(\mathbf{r})=E_{B}F(\mathbf{r}), \tag{7.40}$$

where $D(\mathbf{k})$ is a matrix containing terms up to k^2 whose size is equal to the number of bands retained in the "degenerate" set. I is the unit matrix of the same order and F is a column vector. When diagonalized by itself, $D(\mathbf{k})$ yields the energy bands at each $\mathbf{k}$ in the vicinity of $\mathbf{k}=0$ (Kane, 1956). The coefficients of the various terms in $D(\mathbf{k})$ are parameters similar to effective masses and can be expressed in terms of the second derivatives of the energy bands in particular directions. In this approximation, the impurity wave function is given by

$$\psi(\mathbf{r})=\sum_{m=1}^{M}F_{m}(\mathbf{r})\,u_{m0}^{0}(\mathbf{r}), \tag{7.41}$$

where $F_m(\mathbf{r})$ are the components of the vector $F(\mathbf{r})$, and $u_{m0}^{0}(\mathbf{r})$ are the Bloch functions at $\mathbf{k}=0$ of the M degenerate or nearly degenerate bands. The structure of (7.40) and (7.41) will be discussed at length later on in specific applications. Note, however, that (7.40) reduces to (7.20) when $M=1$ (only one band).

F. "Two-band" models

The term "two-band" effective-mass theory has often been used to refer to formalisms that retain both valence and conduction bands for a level in the fundamental gap (Keldysh, 1963; Glodeanu, 1969a). We distinguish two cases:

(a) If the valence-band maximum and the conduction-band minimum are at the same $\mathbf{k}$ point, we saw already that the impurity potential cannot couple the two sets of bands as long as only the $\mathbf{K}_p=0$ term is retained in the expansion of Eq. (7.7). A "two-band" effective-mass equation can, however, be obtained by constructing a $\mathbf{k}\cdot\mathbf{p}$ matrix for the two or more bands one may wish to include. The procedure would be identical with that used to obtain the acceptor effective-mass equation (Kittel and Mitchell, 1954; Luttinger and Kohn, 1955), which is in fact a three-band (without spin) or six-band (with spin) equation. An "eight-band" $\mathbf{k}\cdot\mathbf{p}$ matrix has been given by Kane (1957) for the top valence bands and lowest conduction band of InSb, which would yield an "eight-band" effective-mass equation of a form similar to (7.40). Keldysh's "two-band" equations correspond to a model 2×2 $\mathbf{k}\cdot\mathbf{p}$ matrix, which he used to study the analytical structure of "two-band" solutions.

(b) If the valence-band maximum is not at the same $\mathbf{k}$

point as the conduction band minimum, the impurity potential may couple these bands via the $K_p = 0$ term even if the two bands are not assumed coupled via a $\mathbf{k}\cdot\mathbf{p}$ matrix. No work along these lines has been pursued.

G. Applications to real materials–successes and failures

Soon after the derivation of the rigorous effective-mass equations for realistic band configurations, it became apparent that the method, in conjunction with a hydrogenic potential, provided an excellent framework for the quantitative description of the excited states of shallow donors and acceptors. It also became apparent that in most cases the method failed to produce satisfactory ground-state energies (i.e., binding energies) and was totally inadequate for the description of deep donors and acceptors. For this reason, work on excited states over the last twenty years has concentrated almost exclusively on obtaining more accurate solutions of the hydrogenic EME's and on classifying the excited states in more useful ways. On the other hand, work on understanding the failure of the HEMT to produce accurate binding energies, and on procedures to improve the situation, has been along many different directions. For this reason, we first concentrate on reviewing the literature on excited states. In doing this, we will at the same time review and evaluate the various techniques that have thus far been introduced in solving the EME's.

1. Excited states

a. Donors

Silicon and Germanium: Both these materials have a conduction band with several equivalent minima. Si has six minima along the (100) directions at about $0.85(2\pi/a)$. The effective masses corresponding to Eq. (7.34) are

$$m_l^* = 0.9163\, m_0\,,\qquad m_t^* = 0.1905\, m_0\,, \tag{7.42}$$

for Si (Hensel, Hasegawa and Nakayama, 1965), and

$$m_l^* = 1.588\, m_0\,,\qquad m_t^* = 0.08152\, m_0\,, \tag{7.43}$$

for Ge (Levinger and Frankl, 1961). The one-valley EME is then precisely of the form (7.35). The symmetry is no longer spherical, as in the hydrogen atom problem, but only cylindrical. Nevertheless, the solutions may be conveniently labeled by the hydrogenic notation (1s, 2s, etc.), except that now the notation indicates the hydrogenic solution into which the actual solution would reduce in the limit $m_t^*/m_l^* \to 1$. The only complication is that states with different azimuthal quantum number m are split by the nonspherical terms in the Hamiltonian. The m value is then indicated by a superscript on the hydrogenic notation. For example, $2p^0$, $2p^{\pm}$ stand for $2p$ with $m = 0$ and $m = \pm 1$, respectively.

Energy-level calculations were first done by Kohn and Luttinger (1955) who employed a trial function of the form

$$F^j(\mathbf{r}) = \exp[-a(x^2 + y^2) - bz^2]^{1/2} \tag{7.44}$$

for the minimum j along the z direction. More detailed calculations were done by Faulkner (1969) who expanded $F^j(\mathbf{r})$ in terms of spherical harmonics and Laguerre polynomials. Faulkner obtained a complete set of solutions for an arbitrary ratio $\gamma = m_t^*/m_l^*$ in units of the effective Rydberg defined by $e^4 m_t^*/2\hbar^2\epsilon$ and the effective Bohr radius defined by $\hbar^2\epsilon/m_t^* e^2$. Results for individual semiconductors are then obtained by specifying γ for the material and converting the effective Rydbergs into absolute units. For comparison with experiment, the results are best displayed as energy differences between the various states, because that is what is directly extracted from experiment. Table I shows that

TABLE I. Spacings of selected excited states of shallow donors in Si and Ge as calculated by Faulkner (1969) compared with experimental[a] values. All energies are in meV.

	States	Theory	P	As	Sb	Bi	Li	S^0	$S^+/4$
Si									
	$2p^{\pm}-2p^0$	5.11	5.06	5.12	5.06	4.94	5.13	5.2	5.15
	$3p^0-2p^{\pm}$	0.92	0.93	0.86	0.95	0.93	0.88	0.7	1.08
	$4p^0-2p^{\pm}$	3.07	3.11	2.6	...	2.61	...	...	...
	$3p^{\pm}-2p^{\pm}$	3.28	3.27	3.25	3.34	3.31	3.28	3.1	3.45
	$4p^0-2p^{\pm}$	4.07	...	...	...	...	...	...	...
	$4p^0-2p^{\pm}$	4.17	4.21	4.3	4.33	4.34	4.19	...	4.35
	$4p^{\pm}-2p^{\pm}$	4.21	4.21	4.3	4.33	4.35	4.19	4.35	...
	$5p^{\pm}-2p^{\pm}$	4.97	4.95	4.9	...	5.26	4.93	...	...
Ge									
	$2p^{\pm}-2p^0$	3.02	3.03	3.02	3.02	3.02	3.08	3.04	
	$2p^{\pm}-3p^0$	0.83	0.83	0.83	0.84	0.88	0.84	...	
	$3p^{\pm}-2p^{\pm}$	0.69	0.69	0.70	0.69	0.66	0.67	...	

[a] Experimental values from Aggarwal and Ramdas (1965) for Si:P, As, Sb, Bi; from Reuszer and Fisher (1964) for Ge:P, As, Sb, Bi; from Aggarwal, Fisher, Mourzine and Ramdas (1965) for Si:$\mathrm{Li_i}$ and Ge:$\mathrm{Li_i}$; and from Krag and Zeiger (1962) for Si:S.

TABLE II. Theoretical and experimental binding energies in meV of shallow donors in GaAs.

HEMT	5.72[a]
Si	5.81[b]
Ge	6.08[b]
S	6.10[b]
Se	5.89[b]

[a] Using $m_e^*=0.0665$ from Fetterman *et al*. (1971).
[b] Summers, Dingle, and Hill (1970).

for Si and Ge excellent agreement with experiment was obtained for all the p-like states detected in infrared absorption experiments.

III-V and II-VI Compounds: From among the III-V and II-VI compounds, GaAs, InP, InAs, InSb, CdTe, and CdSe have the conduction-band minimum at Γ. Effective masses are in general very small, of order $0.1m_0$, whereas dielectric constants are of order 10. Binding energies are therefore of order 10 meV and effective Bohr radii are of order $100a_0$. The hydrogenic model is then at its best. All approximations are satisfied with high accuracy. Systematic experimental studies are not available, however. The most reliable data for m_e^* and for donor binding energies are for GaAs (Table II).

GaP, one of the most thoroughly studied III-V compounds, and AlSb have conduction bands which are similar to that of Si. In GaP the minima were thought to be at the X points, but it has recently been established that they are actually at a very small distance from X along the (001) axes, similar to the situation in Si (Dean and Herbert, 1976). Effective masses have not been accurately determined by cyclotron resonance as in Si and Ge. Attempts have in fact been made to extract m_t^* and m_l^* from analysis of the infrared spectra of donors in terms of the HEMT (Onton, 1969; Carter, Dean, Skolnick, and Stradling, 1977).

Finally, GaSb is a very intriguing case. The absolute minimum of the conduction band is at Γ, but the minima at L are believed to be only about 80 meV higher (Vul *et al.*, 1970) with substantially higher effective masses. Similarly the X minima are only about 300 meV (Vul *et al.*, 1970). Such a configuration allows for the possibility that the level in the gap has strong contributions from more than one species of minima and would be described by a wave function of the form (7.38). Shallow donors in this material and their association with different minima have been discussed by Kosicki and Paul (1966), Kosicki *et al.* (1969) and by Vul *et al.* (1970). These associations were vividly demonstrated by applying pressure, which alters the relative energy positions of the various minima, and studying the resistivity at room temperature. No theoretical calculations are available, however, for this intriguing material. A theoretical study of the effect of pressure on inequivalent minima and effective-mass binding energies has been carried out for GaAs by Altarelli and Iadonisi (1972).

b. Acceptors

All diamond-type and zinc-blende-type semiconductors have similar valence bands in the vicinity of the band gap. The bands are degenerate, anisotropic, and spin-orbit split (Fig. 8). The maximum is fourfold degenerate (Γ_8 symmetry) and slightly below it is the split-off band, with its maximum of Γ_7 symmetry. The EME is therefore of the form (7.39) where $D(-i\nabla)$ is a 6×6 matrix. The form of this matrix was first derived by Dresselhaus, Kip, and Kittel (1955) who also determined the "effective-mass" constants that appear in it. The corresponding EME was derived simultaneously by Kittel and Mitchell (1954) and by Luttinger and Kohn (1955). These early expressions for the matrix D are rather complicated and may be found in the original literature. Solutions of the corresponding EME were rather cumbersome to obtain (Kohn and Schechter, 1955; Schechter, 1962; Mendelson and James, 1964; Suzuki, Okazaki, and Hasegawa, 1964; Mendelson and Schultz, 1969). The basic difficulty lay in constructing appropriate trial functions for states of each allowed symmetry, a procedure analogous to the construction of cubic harmonics, and then in evaluating the multitude of integrals. The resulting classification of excited states was also rather awkward since the analogies with the hydrogenic spectrum were limited.

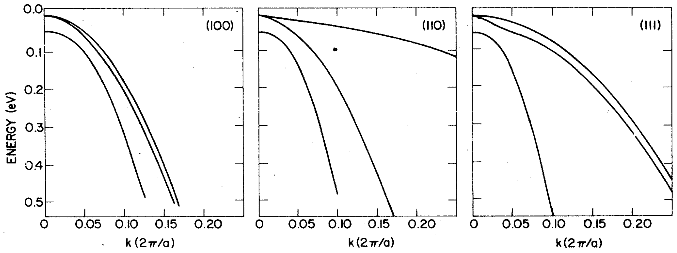

FIG. 8. The energy bands of Si near the top of the valence bands, plotted along three symmetry directions to demonstrate the large anisotropy. Note the spin-orbit splitting which is 0.044 eV at $k=0$. Such plots were first given by Kane (1956).

A more compact expression for the matrix D was obtained by Luttinger (1956), in terms of standard angular momentum matrices in the two limits of zero and infinite spin-orbit interactions. He found that symmetry allows only three independent parameters to enter the construction of the most general matrix. In the case of zero spin-orbit interaction, the matrix D is given by

$$D(p)=(\gamma_1+4\gamma_2)\frac{p^2}{2m_0}-\frac{3\gamma_2}{m_0}(p_x^2I_x^2+p_y^2I_y^2+p_z^2I_z^2)-\frac{6\gamma_3}{m_0}[\{p_xp_y\}\{I_xI_y\}+\{p_yp_z\}\{I_yI_z\}+\{p_zp_x\}\{I_zI_x\}], \tag{7.45}$$

where $\{ab\}=(ab+ba)/2$, γ_1, γ_2, and γ_3 are "inverse effective-mass constants," and I is the angular momentum operator corresponding to spin 1. In the case of infinite spin-orbit interaction, D is given by

$$D(p)=(\gamma_1+\frac{5}{2}\gamma_2)\frac{p^2}{2m_0}-\frac{\gamma_2}{m_0}(p_x^2J_x^2+p_y^2J_y^2+p_z^2J_z^2)-\frac{2\gamma_3}{m_0}[\{p_xp_y\}\{J_xJ_y\}+\{p_yp_z\}\{J_yJ_z\}+\{p_zp_x\}\{J_zJ_x\}], \tag{7.46}$$

where J is the angular momentum operator for spin 3/2. These forms of the matrix D, however, did not make the problem of solving the acceptor EME any easier.

More recently, Lipari and Baldereschi (1970; see also Baldereschi and Lipari, 1973), rewrote these forms in alternative ways, which turned out to also be more useful. They first observed that $D(p)$ may be written in a very compact form in terms of the following second-rank Cartesian tensors:

$$P_{ik}=3p_ip_k-\delta_{ik}p^2\,, \tag{7.47a}$$

$$I_{ik}=\frac{3}{2}(I_iI_k+I_kI_i)-\delta_{ik}I^2\,, \tag{7.47b}$$

and

$$J_{ik}=\frac{3}{2}(J_iJ_k+J_kJ_i)-\delta_{ik}J^2\,, \tag{7.47c}$$

where the indices $i,k=1,2,3$, mean x,y, and z, respectively. For example, in the limit of infinite spin-orbit coupling, $D(p)$ becomes

$$D(p)=\frac{\gamma_1}{2m_0}-\frac{1}{9m_0}[\gamma_3-(\gamma_3-\gamma_2)\delta_{ik}]P_{ik}J_{ik}\,, \tag{7.48}$$

where δ_{ik} is the Kronecker delta and repeated indices are summed over (Einstein convention). Form (7.47) is indeed more compact than (7.45), but would not be more useful if one had to go back to (7.45) for actual calculations. The substantial contribution of Lipari and Baldereschi was in noting that the tensors P_{ik}, I_{ik}, and J_{ik}, which are Cartesian tensors of rank two, may be reduced to spherical tensors. In general, the reduction would yield spherical tensors of rank 0, 1, and 2, but the fact that the traces of the tensors P_{ik}, I_{ik} and J_{ik} are zero eliminates the zero-rank spherical tensor. Similarly, the fact that the three tensors are symmetric (in the sense $T_{ik}=T_{ki}$) eliminates all spherical tensors of order 1. The net result is that each of the three Cartesian tensors T_{ik} (where T stands for P, I, or J) is reducible into spherical tensors of rank two, which are denoted by $T_q^{(2)}$, with $q=2,-1,0,1,2$. In view of (7.48), one must then obtain products between the $P_q^{(2)}$ and the $I_q^{(2)}$ or $J_q^{(2)}$. Such products are in general spherical tensors of rank 0, 1, 2, 3, and 4, but for the problem at hand only a tensor of rank 0 (i.e., the scalar product) and three different components of a tensor of rank 4 contribute. Having done all this, one may conveniently express $D(p)$ in units of the effective Rydberg $(e^4m_0/2\hbar^2\epsilon^2\gamma_1)$ as follows:

$$\begin{aligned}D(p)=&\frac{1}{\hbar^2}\,p^2-\frac{1}{9\hbar^2}\,\mu(P^{(2)}\cdot J^{(2)})\\&+\frac{1}{9\hbar^2}\,\delta\{[P^{(2)}\times J^{(2)}]_4^{(4)}+\frac{1}{5}\sqrt{70}\,[P^{(2)}\times J^{(2)}]_0^{(4)}\\&+[P^{(2)}\times J^{(2)}]_{-4}^{(4)}\}\,,\end{aligned} \tag{7.49}$$

where the definition of the "scalar" and "vector" products of the spherical tensors may be found in the paper by Baldereschi and Lipari (1973). A similar expression is obtained for the case of zero spin-orbit coupling. The new constants μ and δ are related to the γ's by

$$\mu=(6\gamma_3+4\gamma_2)/5\gamma_1\,, \tag{7.50a}$$

$$\delta=(\gamma_3-\gamma_2)/\gamma_1\,. \tag{7.50b}$$

The form (7.49) appears rather forbidding but contains a major simplification. Whereas in the Luttinger forms (7.45) and (7.46) only the p^2 term had full spherical symmetry, in the new form (7.49) an additional term [the second term] with full spherical symmetry is isolated. It turns out that in most semiconductors μ is substantially larger than δ so that the cubic term [last term in (7.49)] may be either ignored or included by perturbation theory, at least for excited states (Baldereschi and Lipari, 1973).

As a first step, neglecting the cubic term in (7.49) allows a systematic classification of all states in terms of their total angular momentum, in complete analogy to atomic spectroscopy [the second term in (7.49) plays the role of a spin-orbit interaction term]. For example the ground state would have total angular momentum $F=\frac{3}{2}$ (the spherical analog of Γ_8 cubic symmetry), where

$$\mathbf{F}=\mathbf{L}+\mathbf{J}\,. \tag{7.51}$$

Since $J=\frac{3}{2}$, L could take any integral value beginning with 0, except that parity conservation requires that only even L values contribute. The total wave function would thus be of the form

$$\Phi(S_{3/2})=\sum_L f_L(r)\left|L\,,J=\frac{3}{2}\,,F_z\right\rangle. \tag{7.52a}$$

Similarly one can construct the excited p-like states

$$\Phi(P_{1/2})=\sum_L f_L(r)\left|L,J=\frac{3}{2}\,,F=\frac{1}{2}\,,F_z\right\rangle, \tag{7.52b}$$

$$\Phi(P_{3/2})=\sum_L f_L(r)\left|L,\,J=\frac{3}{2}\,,F=\frac{3}{2},F_z\right\rangle, \tag{7.52c}$$

and

$$\Phi(P_{5/2})=\sum_L f_L(r)\left|L,\,J=\frac{3}{2}\,,F=\frac{5}{2},F_z\right\rangle, \tag{7.52d}$$

where now only odd L values contribute. In (7.52), the $f_L(r)$ are undetermined radial functions. Having written the solutions in the form (7.52), the matrix elements of the cubic part of Hamiltonian (7.49) are directly evaluated using the reduced matrix element (or double-bar-matrix-element) techniques (Edmonds, 1960). The net result is a set of coupled differential equations for the $f_L(r)$ which are then solved variationally. In their original application, Lipari and Baldereschi employed $L=0$ and 2 for s-like states and $L=1$ and 3 for p-like states, as was done in previous applications. Another nice feature of the new approach is the fact that within the infinite spin-orbit limit and the spherical approximation, one can use the effective Rydberg as the unit of energy and obtain complete solutions as a function of μ, which is analogous to the results of Faulkner (1969) for donor states. When the cubic term is included by perturbation theory (Baldereschi and Lipari, 1974), only states of angular momentum $\frac{5}{2}$ split into a Γ_8 and a Γ_7 state.

More recently, Baldereschi and Lipari (1976) (also Lipari and Baldereschi, 1978) included the cubic term of the Hamiltonian (7.49), the spin-orbit split-off band which introduces the extra term $\frac{2}{3}(\frac{1}{2}-\mathbf{J}\cdot\mathbf{S})\Delta$ in the Hamiltonian (where Δ is the spin-orbit splitting at Γ) and L

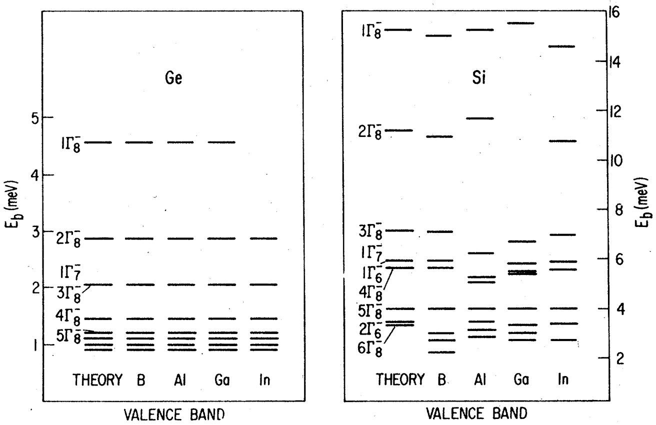

FIG. 9. The latest theoretical values for the excited states of acceptors in Si and Ge, calculated by Baldereschi and Lipari. (1978). The experimental values are from Skolnick *et al.* (1974) and from Haller and Hansen (1974).

values up to $L = 7$ for convergence. Results have been reported for Si and Ge and are shown in Fig. 9. The agreement with the experimental data is indeed very impressive and demonstrates once more how powerful the EMT can be for excited states, if the calculations are done accurately.

The situation in the compound semiconductors is somewhat more uncertain, largely because the effective-mass constants γ_1, γ_2, and γ_3 (or γ_1, μ, and δ) are not known accurately from experiments. Values commonly used are those estimated by Lawaetz (1964). Recently, it has proved advantageous to reverse the procedure and try to fit the data on excited states by varying γ_1, γ_2, and γ_3 and thus extracting the crystal parameters from the measurements (Street and Senske, 1976). This should be listed as one of the triumphs of the EMT, though caution must be exercised in view of possible slight variations among various acceptors and the possible nonuniqueness of numerical fits. A similar feat was accomplished earlier by Faulkner (1969), who deduced from the donor excited states that the appropriate dielectric constants for Si and Ge should be 11.4 and 15.36, respectively, instead of the usual 12 and 16. In fact, it turns out that Faulkner's values are the correct dielectric constants at the low temperatures at which the impurity spectra are measured (Cardona, Paul, and Brooks, 1959).

2. Ground states

As we noted already, the HEMT does not always do very well for ground-state energies. The most notable failure was Si: For donors, the HEMT binding energy obtained by Kittel and Mitchell (1954) and by Kohn and Luttinger (1955) was 29 meV. The more accurate calculation of Faulkner (1969) raised this to 31.2 meV. The experimental values, on the other hand, are 45.5, 53.7, and 42.5 meV for substitutional P, As, and Sb, respectively (Aggarwal and Ramdas, 1965). Numbers close to these were known back in the mid-50's, whereby it became immediately clear that the HEMT was not adequate. The situation became worse in 1964 when Aggarwal discovered that the ground state of the donors was actually split into three levels (a singlet, with lowest energy, a doublet, and a triplet), instead of being sixfold degenerate, as predicted by the one-valley HEMT. (Theory had predicted, e.g., Kohn, 1957, that terms beyond the one-valley EMT could cause a splitting, but no quantitative estimates were possible.) A similar situation existed for acceptors. Early calculations (Kohn and Schechter, 1955; Schechter, 1962) produced a binding energy of about 31 meV. Later this number was improved to 35.2 meV by Suzuki, Okazaki, and Haseqawa (1964) and to 37.1 by Mendelson and Schultz (1969). It turns out that the most accurate *hydrogenic* binding energy for acceptors in Si is about 44 meV (Baldereschi and Lipari, 1976; Bernholc and Pantelides, 1977), which is a far cry from most of the experimental values of about 45, 68, 71, and 151 meV for B, Al, Ga, and In, respectively. Finally, the HEMT fails even more blatantly for deep impurities. For example, if the hydrogenic potential of two charges was used to describe binding of one electron to the double S donor in Si, the binding energy would be about 4 × 31 = 124 meV, which compares disastrously with the experimental value of 613 meV (Kleiner and Krag, 1970).

The situation in Ge is not so bad. Experimental values range from 11 to 13 meV for both donors and acceptors and theoretical values could easily come close to them (Kohn and Schechter, 1955; Schechter, 1962; Mendelson and James, 1964; Suzuki, Okazaki, and Hasegawa, 1964; Faulkner, 1969; Baldereschi and Lipari, 1974). The ground state of donors, however, was also found to be split into a singlet and a triplet instead of being fourfold degenerate, as predicted by the HEMT.

The situation in the compound semiconductors varies from material to material. Compound semiconductors, however, have problems that are unique to them and we defer their discussion to the next section. Our main task now is to examine the procedures that have been introduced to improve and go beyond the HEMT in Si and Ge, which are the two materials that have been studied the most (Sec. VII.H below and Sec. VIII). Before we do that, however, we briefly discuss the hydrogenic model for resonant states and for impurity pairs.

3. Resonant states

According to the HEMT, bound states appear below or above the extremum of every band. In many cases one is only interested in those states that happen to lie within the fundamental band gap. Hydrogenic states do appear, however, above or below other band extrema and lie within the band continuum. They are therefore quasi-bound in the sense that they are degenerate with propagating Bloch states. They are called resonant states and have been observed in many materials, e.g., in acceptors in Si, associated with the spin-orbit split-off band (Zwerdling, Button, Lax, and Roth, 1960; Onton, Fisher, and Ramdas, 1967), and in the case of donors in GaP, associated with a higher conduction band at X (Onton, 1971). Extensive theoretical studies of the analytical structure of such states have been carried out by

Bassani and co-workers (see, e.g., Bassani *et al.*, 1974; Altarelli and Iadonisi, 1971).

4. Impurity pairs

The HEMT has also been generalized to study the binding of electrons and/or holes to pairs of impurities, separated by a distance R. Pioneering work was done by Williams (1960), and later by Shaffer and Williams (1964, 1970) and by Kaczmareck (1966). A comprehensive review of the subject has been given by Williams (1968) and by Dean (1973). The extension of the HEMT to pairs is straightforward and follows the Heitler-London treatment of the hydrogen molecule. In the case of an electron and a hole bound to a donor-acceptor pair, the exchange integral is missing because the two particles are distinguishable. The experimentally interesting quantity is the electron-hole annihilation energy which is given by (Williams, 1968).

$$E_a = E_{cv} - E_D - E_A + J + e^2/\epsilon R \,, \qquad (7.53)$$

where E_{cv} is the minimum valence-to-conduction band gap, E_D is the donor binding energy, E_A is the acceptor binding energy, and R is the pair separation distance. J is given by

$$J = \frac{e^2}{\epsilon} \int\int F_D(\mathbf{r})F_A(\mathbf{r}')\left(\frac{1}{|\mathbf{r}_A - \mathbf{r}|} + \frac{1}{|\mathbf{r}_D - \mathbf{r}'|} - \frac{1}{|\mathbf{r} - \mathbf{r}'|}\right) \times F_D(\mathbf{r})F_A(\mathbf{r}')d^3r d^3r' \,. \qquad (7.54)$$

Here $\mathbf{r}_A$ and $\mathbf{r}_D$ are the acceptor and donor positions, respectively; F_A and F_D are the acceptor and donor ground-state effective-mass envelope functions, respectively. The above expression is valid to first order in perturbation theory. Second-order, or Van der Waals interactions, have also been estimated (Hoogenstraaten, 1958), but in general they are rather small. Equation (7.53) has been extremely useful in analyzing the luminescence spectra obtained when electrons and holes annihilate, as we saw in Sec. V. Other interesting aspects of the impurity-pair problem may be found in the review papers by Williams (1968) and by Dean (1973).

H. Corrections to the hydrogenic effective-mass theory: Chemical shifts and central-cell corrections

In the years following the derivation of the EME's for donors and acceptors in Si and Ge (Kittel and Mitchell, 1954; Luttinger and Kohn, 1955), many attempts have been made to go beyond the hydrogenic effective-mass model and account for the large discrepancies between the hydrogenic binding energy E_H and the observed value E_0. This quantity, namely

$$\Delta E = E_0 - E_H \,, \qquad (7.55)$$

came to be known as the *chemical shift*, since the particular chemical nature of the impurity was thought to be responsible for its presence. Since the chemical details of the impurity would contribute to the total impurity potential only inside the so-called "central-cell" region, the outstanding problem was often referred to as the calculation of central-cell corrections to the EMT. In this vein, a number of authors (Reiss, 1956, Kaus, 1958; Müller, 1964, 1965; Breitenecker, Sexl, and Thirring, 1964) defined a cavity radius r_0 and replaced $U_H(r)$ inside the cavity by a function-reflecting reduced dielectric screening. Csavinszky (1963, 1965) also used the concept of a cavity radius and inside the cavity he included the difference between the potentials of the impurity and host atoms *in addition* to U_H, as a correction. Weinreich (1959), Shinohara (1961), and Morgan (1970) estimated a correction due to the strain field arising from the misfit of an impurity at a substitutional site. Appel (1964) and Sham (1966) included more subtle effects such as s shifts, mass-velocity relativistic corrections, and exchange-correlation corrections. Schechter (1969) evaluated corrections arising from the spatial variation of $U_H(r)$ and other forms of potentials. Jaros (1969, 1971) used an impurity potential calculated from model potentials (Animalu and Heine, 1965) and also introduced a position-dependent effective mass. Haug (1970) used $U_H(r)$ multiplied by an effective charge $Z_{\text{eff}} \neq 1$ obtained from Slater's (1930) rules for free-atom wave functions. Finally, Phillips (1970) focused on understanding the chemical shifts ΔE directly by making use of his dielectric theory of electronegativity plus other corrections, using a number of adjustable parameters.

The above-cited papers were able to account for the observed discrepancies in whole or in part. Stoneham (1975) has compiled a table of their results and has remarked that in several cases the underlying theory is disputable. We do not find it very useful to analyze the conceptual inadequacies present in some of the papers cited, except to note one important shortcoming: In the case of donors, all the above papers ignored intervalley coupling and worked entirely within the one-valley effective-mass theory, which results in an n-fold ($n = 6$ for Si, 4 for Ge) degenerate ground state. Experimentally, however, it was established in 1964 (Aggarwal, 1964; Aggarwal and Ramdas, 1965) that the ground state of donors in Si is *split* into a singlet of A_1 symmetry, a doublet of E symmetry, and a triplet of T_2 symmetry. The A_1 level is well separated from the other two and is the true ground state, whereas the E and T_2 levels have energies near the one-valley effective-mass value for the sixfold degenerate state. In the case of donors in Ge, the fourfold degenerate state splits into a singlet (A_1) and a triplet (T_2) (Reuszer and Fisher, 1964). These experimental results proved unequivocally that what had been referred to as "chemical shift" must arise almost entirely from intervalley mixing, which causes the split. Nevertheless, a number of papers, cited above, continued to appear till about 1970, seeking to explain the observed binding energies or chemical shifts without including intervalley coupling, but by invoking a variety of other mechanisms. The first theoretical demonstration of the importance of intervalley coupling was given by Morita and Nara (1966). These authors used the hydrogenic one-valley EME outside a sphere of radius r_0. Inside the sphere, they used screened differences of atomic potentials (see Sec. VIII), and integrated numerically, using the correct linear combinations of Bloch functions for the A_1, the T_2, and the E states, thus automatically including intervalley coupling. Their results were in good agreement with experiment for the shallow donors in Si (see Sec. VIII for further discussion of the Morita-Nara work).

The first effective-mass-type calculation that demonstrated the importance of intervalley mixing was by Baldereschi (1970), who noticed that the intervalley matrix element of $U_H(r)$ should not be divided by the static dielectric constant ϵ, which corresponds to $q=0$, but by the appropriate values $\epsilon(q), q$ being the intervalley separation. This reduction in screening increases the intervalley matrix elements. The splitting arises because the coefficients α_j appearing in the total wave function (7.36) are different for states of different symmetries (Kohn, 1957). (Only in a one-valley calculation do the values of the coefficients α_j become irrelevant and a sixfold degeneracy occurs.) Baldereschi (1970) estimated the splittings by perturbation theory and obtained numbers of the same order of magnitude as experiment [$\Delta E(T_2-E)=10.6$ meV, compared with 11.85, 9.94, and 21.15 for P, As, and Sb, respectively, and $\Delta E(E-T_2)$ $=1.1$ meV, compared with 1.35, 2.50, and 1.42 for the same impurities. For Ge, the estimated singlet-triplet splitting was 0.6 meV, compared with 2.83, 4.23, and 0.32 for the same impurities]. Baldereschi's calculation demonstrated that the bulk of the "chemical shift," as measured from the one-valley EME value, is in fact not chemical at all, but simply due to intervalley coupling, caused by species-independent reduction of screening. The variation of the observed binding energies with chemical species, however, was still outside that picture, and Baldereschi's conclusion was that the new chemical shifts, defined from the new value for the A_1 level, could be understood by techniques such as those introduced by Phillips (1970), once the free parameters are properly readjusted. At about the same time, Ning and Sah (1970, 1971a) independently demonstrated the importance of intervalley coupling by using a phenomenological two-parameter impurity potential.

Following these works, Pantelides and Sah (1972, 1974) and Pantelides (1973, 1974, 1975) pointed out a fundamental conceptual difficulty with the traditional approach of seeking corrections to the hydrogenic or modified-hydrogenic effective-mass binding energies. Stated simply, the difficulty is that the hydrogenic potential $U_H(r)$ does not in general represent a meaningful contribution to the total impurity potential $U(r)$ in the central cell region. Therefore corrections to U_H cannot always be identified and calculated unless one first calculates $U(r)$ [which asymptotically becomes equal to $U_H(r)$ at large r] and then subtracts $U_H(r)$ in order to obtain the "corrections." It is worth noting that in the original papers (Kittel and Mitchell, 1954; Kohn and Luttinger, 1955; Kohn, 1957) the hydrogenic potential $U_H(R)$ was not used because it was thought to be a good approximation to the actual impurity potential in all space, but because it was thought that the wave function would be so spread out that the central cell region would not contribute appreciably to the binding energy. In such a case, the details of the impurity potential in the central cell would simply not matter. Such thinking, however, anticipates binding energies which are the same for all shallow donors or acceptors in a given material. Experimental data that reveal the contrary should suggest that the details of the impurity potential do in fact matter, so that accurate impurity potentials would have to be calculated.

The one case for which $U_H(r)$ is in fact a meaningful approximation to $U(r)$, in the sense that it can be systematically improved, is the case of substitutional impurities from the same row in the periodic table of the elements as the host, which have been referred to as *isocoric* impurities (Pantelides and Sah, 1974) because they have the same number of core electrons as the host. The significance of this property will become more transparent in the discussion in Sec. VIII. For the time being, let us illustrate how one can arrive at $U_H(r)$ as an approximation to the impurity potential for isocoric impurities by using Si:P as an example. Since P differs from Si by only one extra nuclear proton and one extra electron (the donor electron), we can rigorously construct the impurity potential for Si:P by placing a proton (positive point charge) on a Si nucleus, thus converting it into a P nucleus, and calculating the resultant change in the crystal potential. Clearly, $U_H(r)$ results if the response of the crystal to the point charge is calculated in the most elementary, q-independent approximation of static screening. [Carrying out the screening in terms of $\epsilon(q)$, instead of simply the constant ϵ used in $U_H(r)$, is therefore an improvement toward the goal of constructing $U(r)$ for isocoric impurities. This point of view will be discussed further in Sec. VIII.] An identical picture holds for the isocoric acceptor Si:Al, except now the point charge is negative (to "neutralize" one of Si's protons in order to "convert" it into Al). In contrast, the concept of a point charge for nonisocoric impurities is no help. True, at values of r larger than a central-cell "radius," $(\mathrm{As}^+)-(\mathrm{Si}^0)$ looks like a positive charge which must be screened, etc., but inside the central cell the situation is much more complicated and $U(r)\rightarrow+19e/r$! (See Sec. VIII for further discussion of these potentials.)

We conclude that the concepts of chemical shifts and central-cell corrections are ill defined, except for isocoric impurities, and are therefore inappropriate as a tool to describe the systematics of classes of impurities (such as all the shallow donors or all the shallow acceptors in a given material, etc.). They have in fact held up progress by drawing undue attention. We further conclude that the first step in going beyond the HEMT must be to construct accurate impurity potentials, which reflect the chemical nature of both the host atom and the foreign atom. This step was not taken for a long time largely because it was thought that realistic impurity potentials would be too "violent" for the effective-mass approximations to be valid. In fact, in two instances for which realistic impurity potentials were constructed (Csavinszky, 1963, 1965; Morita and Nara, 1966) no attempt was made to use them directly in EME's. However, recent work by Pantelides and Sah (1972, 1974), demonstrated that realistic impurity potentials can in fact be used in EME's, as long as one deals appropriately with the orthogonality requirements imposed on the localized wave function by the core orbitals of the foreign atom. In that context, a natural distinction between isocoric and nonisocoric impurities is made and the equations are shown to be adequate for shallow as well as moderately deep levels. We refer to the procedure as generalized effective-mass theory (GEMT) and discuss it at length in the next section.

VIII. GENERALIZED EFFECTIVE-MASS THEORY

In contrast to the approaches discussed in the previous section whose motivation was to obtain corrections to the hydrogenic effective-mass theory and thus calculate "chemical shifts," in this section we discuss the alternative procedure whereby realistic impurity potentials are constructed and used directly with effective-mass equations to predict binding energies for individual impurities. We refer to this procedure as the generalized effective-mass theory (GEMT). The task ahead of us, therefore, consists of describing various forms of impurity potentials, checking the validity of the EMA for such potentials, and discussing the results obtained thus far.

A. Isocoric impurities: "True" potentials and point-charge models

By definition, the impurity potential $U(r)$ is given by

$$U = V - V^0 \tag{8.1}$$

[cf. Eq. (6.23)], where V and V^0 are the one-electron potentials for the imperfect and perfect crystal, respectively. The latter potentials depend on the eigenfunctions of their respective eigenvalue problems (6.25) and (6.2). As a good approximation, however, one may take the core solutions in the crystals to be $\mathbf{k}$-independent linear combinations of the atomic core wave functions ($\mathbf{k}$ independence implies flat, zero-width bands), whereby the one electron potential may be written as

$$V^0 = \sum_j v_c^0(\mathbf{r}-\mathbf{R}_j) + V_e^0 \,, \tag{8.2}$$

where $v_c^0(\mathbf{r})$ is similar to the potential of an atom stripped of all its valence electrons, and V_e^0 is the part of the potential arising from the valence electrons in the crystal. Similarly,

$$V = \sum_j v_{cj}(\mathbf{r}-\mathbf{R}_j) + V_e \,, \tag{8.3}$$

where now the core potential depends on the site j and is appropriately labeled. Using the general form of the local approximation to exchange and correlation, we have

$$v_c^0(\mathbf{r}) = -\frac{Z^0 e^2}{r} + \int d^3r' \, \frac{e^2\rho^0(r')}{|\mathbf{r}-\mathbf{r}'|} + \mu_{xc}[\rho^0(\mathbf{r})] \tag{8.4}$$

where Z^0 is the nuclear charge,

$$\rho^0(r) = \sum_t \psi_{ct}^{0*}(\mathbf{r})\psi_{ct}^0(\mathbf{r}) \tag{8.5}$$

with t ranging over the core wave functions of the atom, and μ_{xc} is the local-density functional for exchange and correlation (Kohn and Sham, 1965). Similar expressions hold for v_{cj} in terms of the corresponding quantities without superscript 0. If the approximation is made that $v_{cj}(\mathbf{r}-\mathbf{R}_j) = v_c^0(\mathbf{r}-\mathbf{R}_j)$ for j other than the impurity site, we immediately obtain

$$U(\mathbf{r}) = U_b(\mathbf{r}) + U_s(\mathbf{r}) \,, \tag{8.6}$$

where

$$U_b(\mathbf{r}) = v_c(\mathbf{r}) - v_c^0(\mathbf{r}) \tag{8.7a}$$

for substitutional impurities and

$$U_b(\mathbf{r}) = v_c(\mathbf{r}) \tag{8.7b}$$

for interstitial impurities. Note that the impurity is assumed to be at the origin ($j=0$) and the j label has been dropped on v_{cj}. In (8.6), $U_s(\mathbf{r})$ is given for both substitutional and interstitial impurities by

$$U_s(\mathbf{r}) = V_e(\mathbf{r}) - V_e^0(\mathbf{r}) \,. \tag{8.8}$$

In the case of substitutional impurities, the potential $U_b(r)$, which we refer to as the "bare" impurity potential, is the change that occurs from the replacement of a host ion with an impurity ion, where by ion we mean the nucleus plus the core electrons. In the case of interstitial impurities, U_b is just the potential of the impurity ion. U_s represents the effect of the redistribution that the valence electrons undergo in response to the introduction of U_b. If U_b is sufficiently weak, U_s may be calculated by linear response theory. The degree of weakness required for linear response theory cannot be stated in simple terms, but one can safely say that the theory breaks down when the bare perturbation potential introduces new bound states. When valid, the rigorous result of linear response theory is as follows (Adler, 1962; Wiser, 1963): First define the Fourier transform of $U_b(r)$ by

$$U_b(\mathbf{r}) = \sum_{\mathbf{k}} \sum_p U_b(\mathbf{k}+\mathbf{K}_p) e^{i(\mathbf{k}+\mathbf{K}_p)\cdot\mathbf{r}} \,, \tag{8.9}$$

where the sum over k is restricted to the first Brillouin zone. The total impurity potential $U(\mathbf{r})$ is then given by

$$U(\mathbf{r}) = \sum_{\mathbf{k}} \sum_p U(\mathbf{k}+\mathbf{K}_p) e^{i(\mathbf{k}+\mathbf{K}_p)\cdot\mathbf{r}} \,, \tag{8.10}$$

where

$$U(\mathbf{k}+\mathbf{K}_p) = \sum_{p'} \epsilon_{pp'}^{-1}(\mathbf{k}) U_b(\mathbf{k}+\mathbf{K}_{p'}) \,. \tag{8.11}$$

where $\epsilon_{pp'}^{-1}(\mathbf{k})$ is the inverse of the dielectric tensor $\epsilon_{pp'}(\mathbf{k})$.

Impurity potentials were first constructed in the spirit of the above derivation by Csavinszky (1963, 1965). He used the Thomas–Fermi statistical theory (Gombâs, 1956) to calculate atomic potentials, but the difference potential for substitutional impurities was viewed as a correction to $U_H(r)$ within a cavity, to be treated by perturbation theory, an assumption which cannot be properly justified. Later, Morita and Nara (1966) and Nara and Morita (1967) constructed impurity potentials for shallow donors in Si according to Eqs. (8.6) and (8.7) and included U_s in terms of linear response theory by dropping the off-diagonal, so-called Umklapp elements of $\epsilon_{pp'}(\mathbf{k})$, i.e., writing

$$\epsilon_{pp'}(\mathbf{k}) = \epsilon(\mathbf{k}+\mathbf{K}_p)\delta_{pp'} \,, \tag{8.12}$$

whereby (8.11) becomes

$$U(q) = U_b(q)/\epsilon(q) \,, \tag{8.13}$$

where q extends over all of reciprocal space. [Nara

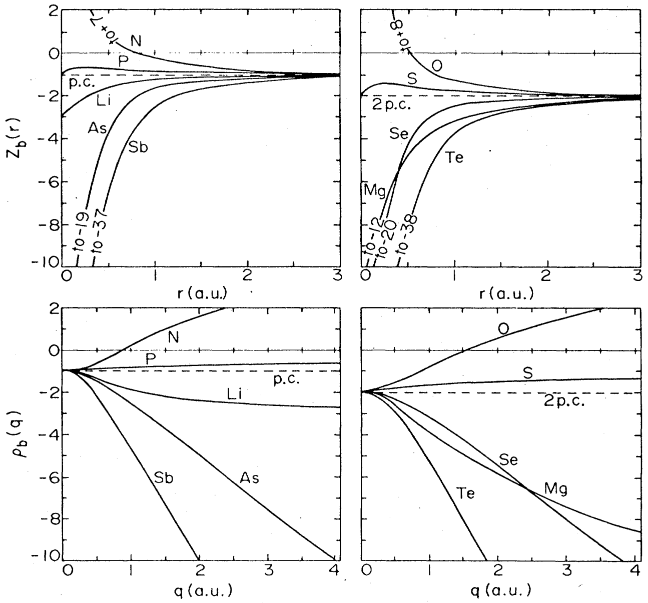

FIG. 10. Bare impurity potentials for donors in Si as calculated by Pantelides and Sah (1972, 1974). The quantities plotted, namely $Z_b(r)$ and $\rho_b(q)$, are defined by $U_b(r) = Z_b(r)/r$ and $U_b(q) = 4\pi e^2 \rho(q)/q^2$. Note that nonisocoric impurites have strong potentials for small r, which correspond to large q values.

(1965) estimated the effect of Umklapp terms to be of order 10%.] The use of (8.13), however, is properly justified only for Si:P, for which $U_b(r)$ does not bind additional core electrons (see further discussion of this point below). Morita and Nara did not use these potentials in effective-mass-type equations. As discussed in Sec. VII, they used them within a cavity of radius r_0, integrated them numerically, and matched effective-mass wave functions at the boundary.

More recently Pantelides and Sah (1972, 1974a) calculated bare impurity potentials for group-V donors and group-VI double donors, using Eqs. (8.6) and (8.7), for possible use in effective-mass equations. The results, shown in Fig. 10, demonstrated explicitly that with the possible exception of Si:P and Si:S all other potentials have strong high Fourier components and would definitely violate the analog of (7.14). The reason is actually very simple. Si:P and Si:S are the only substitutional donors for which impurity and host have the same number of core electrons [hence the term isocoric (Pantelides and Sah, 1974a)]. In all other cases, extra core electrons are present in the crystal and $U(r)$ is strong enough to account for their binding. $U(r)$ being strong in the core means that the $U(\mathbf{k}+\mathbf{K})$ are very large, and they definitely violate the main EMT criterion, Eq. (7.14). By the same token, U_s for these impurities cannot be calculated by linear response theory.

Figure 10 also reveals that the bare impurity potential U_b for isocoric impurities is very nearly equal to that of one or two point charges for single and double donors, respectively. It is further clear that the same thing would be true for all isocoric impurities. The origins of the potentials for isocoric impurities may be viewed as follows: As we saw at the end of Sec. VII, one can "create" a P impurity in Si by simply adding a proton (point charge) to a Si nucleus and supplying an extra electron (the donor electron). The total impurity potential $U(r)$ would then be the bare Coulomb potential, i.e., $-e^2/r$, plus whatever response it generates in the crystal. As we saw in Sec. VII, $U_H(r)$ represents the most elementary approximation of linear response theory in that it views the extra charge as embedded in a macroscopic dielectric medium. In order to understand the calculated potentials for Si:P and Si:S, we go back to the bare Coulomb potential $-e^2/r$ and seek the response in a microscopic picture. The electronic response may be divided into that of the core electrons and that of the valence electrons. Clearly, the core electrons are drawn closer to the nucleus and provide some "screening." Hence the net bare potentials for Si:P and Si:S are *weaker* than the bare point charge or two point charges respectively, just as Fig. 10 shows them to be. Note that for acceptors, where one or two negative charges are added to the nucleus (to effectively "remove" one or two protons), the core electrons would move out and thus again *weaken* the net negative charge(s). Finally, the response of the valence electrons can be included to a good approximation by linear response theory, either using the full dielectric tensor as in Eq. (8.11), or, more approximately, the "diagonal" function $\epsilon(q)$, as in Eqs. (8.12) and (8.13). Note that one could also include lattice relaxation, say in terms of a strain field, as a response of the nuclei to the introduction of the point charge (Morgan, 1970).

The above discussion reveals the convenience of defining an $\epsilon(q)$-screened Coulomb potential, which Pantelides and Sah (1972, 1974) referred to as the point-charge model, U_{pc}, in contradistinction from the hydrogenic model U_H. U_{pc} is defined by (8.13) with U_b being a bare Coulomb potential, so that

$$U_{pc}(q) = -4\pi e^2/\epsilon(q)q^2 . \tag{8.14}$$

This definition may be viewed as a generalization of the corrections to U_H introduced by Müller (1964, 1965), Breitenecker, Sexl, and Thirring (1964), and Baldereschi (1970). It is stressed, however, that $U_{pc}(r)$ is a meaningful potential only for *isocoric* impurities and represents an approximation that views the core wave functions of the isocoric impurity as identical with those of the host atom.

We turn now to the use of the above potentials in EME's. It might be argued that the use of $\epsilon(q)$ is outside the realm of validity of the EMT. This is in fact not necessarily true. If we go back to (7.12), we recall that one requires that

$$|U(\mathbf{k}-\mathbf{k}'+\mathbf{K}_p)| \ll |U(\mathbf{k}-\mathbf{k}')| , \tag{8.15}$$

whereby (7.12) becomes

$$E^0_{\mathbf{k}} F_{\mathbf{k}} + \sum_{\mathbf{k}'} C_{\mathbf{k}\mathbf{k}'}(0)\, U(\mathbf{k}-\mathbf{k}') F_{\mathbf{k}'} = E_\nu F_{\mathbf{k}} , \tag{8.16}$$

where k and k′ are allowed to extend over all k space. The next approximation is $C_{\mathbf{k}\mathbf{k}'}(0) \simeq C_{\mathbf{k}\mathbf{k}}(0) = 1$; but if we write $U(\mathbf{k}-\mathbf{k}')$ as

$$U(\mathbf{k}-\mathbf{k}') = U_b(\mathbf{k}-\mathbf{k}')/\epsilon(\mathbf{k}-\mathbf{k}'), \qquad (8.17)$$

we observe that, since it is $U(\mathbf{k}-\mathbf{k}')$ that appears in (8.16), and $\mathbf{k}$ and $\mathbf{k}'$ are kept intact, it makes a more internally consistent theory to use $\epsilon(\mathbf{k}-\mathbf{k}')$ in full, as in (8.17), instead of replacing it by $\epsilon(0)$, even if $\mathbf{k} \sim \mathbf{k}'$.

At the same time, we would also like to point out that it would be *outside* the realm of the EMT to go beyond the "diagonal" dielectric function $\epsilon(q)$ and attempt to use the full tensor $\epsilon_{pp}(\mathbf{k})$. This statement can be verified by using (8.11) in (7.12) to get

$$E^0_{\mathbf{k}} F_{\mathbf{k}} + \sum_{\mathbf{k}'} \sum_{p} C_{\mathbf{k}\mathbf{k}'}(\mathbf{K}_p) \sum_{p'} \epsilon^{-1}_{pp'}(\mathbf{k}-\mathbf{k}')\, U_b(\mathbf{k}-\mathbf{k}'+\mathbf{K}_{p'}) F_{\mathbf{k}'} = E_\nu F_{\mathbf{k}}. \qquad (8.18)$$

Clearly, if the $p \neq p'$ terms are viewed as important, nonzero $\mathbf{K}_p$ vectors would have to be retained, which is beyond the EMT.

We have so far shown that the use of $\epsilon(\mathbf{k}-\mathbf{k}')$ is not in principle outside the realm of the EMT. On the other hand it does make the requirement (8.15) somewhat less satisfied. *The choice is therefore between a less accurate impurity potential that satisfies subsequent approximations well, and a more accurate impurity potential which satisfies subsequent approximations not so well.* Numerical calculations thus far suggest that the second choice is the better one. Before we discuss numerical results, however, we turn to the choice of $U_b(\mathbf{k}-\mathbf{k}')$ for use in EME's. The equations show that as long as (8.15) is satisfied and the nonzero $\mathbf{K}_p$ can be dropped, it should not matter what the particular form of $U_b(\mathbf{k}-\mathbf{k}')$ is. In fact, Fig. 10 shows that the full impurity potentials for isocoric impurities satisfy (8.15) even *better* than the point-charge potential.

Numerical calculations for the point-charge model have thus far been carried out for donors in Si (Pantelides and Sah, 1972, 1974a), for acceptors in Si and Ge (Baldereschi and Lipari, 1976; Bernholc and Pantelides, 1977) and for acceptors in zinc-blende-type compound semiconductors (Bernholc and Pantelides, 1977; Lipari and Baldereschi, 1978). A variety of dielectric functions were used, as calculated by several authors using pseudopotential band structures (see discussion in Bernholc and Pantelides, 1977). For the full isocoric potentials calculations, have been carried out only for donors in Si (Pantelides and Sah, 1972, 1974a). All the results for Si and Ge are shown in Table III. Compound semiconductors will be discussed separately at the end of this section. Si:S is the only case where both one-electron and two-electron calculations have been carried out and the two ionization energies are denoted by S^0 and S^+, the superscript denoting the charge state before removal of the electron.

TABLE III. Binding energies in meV of isocoric impurities in Si and Ge using a variety of approximations. See text. The ranges given for the P.C. model correspond to different choices of $\epsilon(q)$.

	P.C. model	"True"	Pseudo	Model	Expt.
Si:Donors					
Si:P	48.8[a]	42.4[a]	44.3[a]	44.8[b]	45.5[c]
Si:S^+	1085.3[a]	659.3[a]	709.8[a]	874.0[b]	613.6[d]
Si:S^0	489.0[a]	297.1[a]	334.4[a]	406.5[b]	314.0[e]
Si:Acceptors					
Si:Al	70.5–130[f]	...	...	...	70.0[g]
	48.8–54.4[h]	...	...	48.7[h]	70.0[g]
Si:Mg	1409–2351[h]	...	...	...[i]	...[i]
Ge:Acceptors					
Ge:Ga	11.2[f]	...	...	...	11.3[j]
Ge:Ga	10.7[h]	...	...	10.6[h]	11.3[j]
Ge:Zn	99[f]	...	...	...	95[k]
	72.4–76.9[h]	...	...	54.3[h]	95[k]
Ge:Cu	1152–1442[h]	...	...	246[h]	530[k]

[a] Pantelides and Sah (1974a).
[b] Pantelides (1974).
[c] Aggarwal and Ramdas (1965), corrected by Faulkner (1969).
[d] Krag, Kleiner, Zeiger, and Fischler (1966).
[e] Rosier and Sah (1971); corrected by allowing 12 meV for excited states.
[f] Baldereschi and Lipari (1976); Lipari and Baldereschi (1978).
[g] Onton, Fisher, and Ramdas (1967), corrected by Baldereschi and Lipari (1976).
[h] Bernholc and Pantelides, 1977.
[i] No data are available for the isocoric Si:Mg as a substitutional acceptor. For comparison, the model value of Si:Be is 486 meV compared with an experimental value of 420 meV. The respective quantities for Si:Zn are 428 and 620 meV.
[j] Jones and Fisher, 1965; corrected by Baldereschi and Lipari (1976).
[k] Quoted by Milnes (1973).

The heading "true" indicates the potentials described above (Fig. 10). The other headings will be discussed later.

The results for Si:P and Si:S demonstrated for the first time that "true" impurity potentials can in fact be used successfully in EME's, but *only* for isocoric impurities. At the same time, they demonstrated that *deep* isocoric impurities can also be treated along the same lines. Having the periodic table of the elements in mind, one could say that the EMT with "true" potentials has been shown to do well for impurities in a given *row*, whereas the hydrogenic EMT is known to do well for impurities in a given *column*. Qualitatively, this result can be understood as follows. As we have seen, the impurity wave function for the ground state is given by a nodeless envelope function $F(r)$ times the Bloch functions at the extremum (extrema) of the band of interest. Take Si:donors, for example. The conduction-band Bloch functions have a $3s/3p/3d$-like character over every Si atom. When multiplied by a nodeless envelope, they retain that character, which is appropriate for isocoric impurities, both shallow and deep, but totally inappropriate for nonisocoric impurities, whether shallow or deep. They are inappropriate in the sense that they are not orthogonal to the new core orbitals, so that if the "true" potentials were to be used in EME's in a variational calculation, the ground state would tend to "collapse" into some core state (orthogonality catastrophe). (This limitation of "true" potentials was suspected by Csavinszky, 1963, but no remedy was offered. See subsection C below for further discussion.) Core states are, of course, not describable by the EMT. This argument shows that, with true potentials, a many-band expansion in the vicinity of the impurity would be inevitable for nonisocoric impurities, even the shallow ones, such as Si:As or the shallower Si:Li_i but a one-band solution *may* be adequate for the deep isocoric impurities. The question that remains to be addressed is how valid the one-band EME is for deep levels. We postpone this issue for the moment and discuss further the first two columns of Table III.

First we note that point-charge and "true" potential do about equally well for Si:P, though the "true" potential has a smaller binding energy, in agreement with Fig. 10. Similarly, the "true" potential for Si:S gives a smaller binding energy than the two-point-charge model, but in this case the difference is substantial, since the value 1085.3 is unacceptably large (the level is almost in the valence bands and the approximations definitely break down). The results of acceptors in Si point in the same direction. The point-charge model for the single isocoric acceptor is seen to do very well for particular choices of $\epsilon(q)$. Lipari and Baldereschi's (1978) values are more accurate because more L values (up to $L=6$) were included in the trial wave function, as compared with that of Bernholc and Pantelides. (The spread in the values for the acceptor point-charge model correspond to uncertainties in $\epsilon(q)$, which are actually present in all the theoretical values in Table III.) Once more, the two-point-charge model fails completely for the double acceptor. No "true" potential calculation is available, but the model-potential results of Bernholc and Pantelides (to be described later on; see Table III under "Model") show that a more realistic potential would again improve the situation. The results for Ge are very similar, but now the point-charge model seems to do very well for both single and double acceptors, while better potentials would be needed for triple acceptors.

The calculcations of Baldereschi and Lipari (1976) (also Lipari and Baldereschi, 1978) have produced another interesting result: A state of Γ_7^+ symmetry having strong s-like contribution from the split-off band and d-like contributions from the top valence bands was found at 24 meV below the ground state. The energy separation agrees very well with a Raman line observed in B-doped Si at 23.4 meV (Wright and Mooradian, 1968). The degeneracy and symmetry of the Γ_7^+ state are consistent with Raman selection rules. The puzzle, however, is that no such state is observed in Si:Al, which is the isocoric system for which a point-charge calculation applies. Si:B has a Γ_8^+ ground state at 45.6 meV, compared with Si:Al's 70.0 meV, and the corresponding Γ_7^+ states might be expected to lie at comparably different energies. A similar Raman line has been observed for some acceptors in GaAs and GaP (Manchon and Dean, 1970; Chase, Hayes, and Ryan, 1977), but no theoretical calculations have been reported.

Finally, we turn to examine the validity and accuracy of the calculations. We have already seen that the use of $\epsilon(q)$ with either a point-charge model or an isocoric "true" impurity potential is not necessarily outside the EMT and that, qualitatively, a one-band calculation for isocoric impurities contains all the essential features of the solution. The question is one of accuracy determined by the neglect of $U(\mathbf{q}+\mathbf{K})$ as compared with $U(\mathbf{q})$. Since the extent of q is roughly $1/\langle r\rangle$, where $\langle r\rangle$ is the mean radius of the envelope function $F(r)$, the approximation is justified for the point-charge model if

$$R=\frac{\epsilon(1/\langle r\rangle)}{\epsilon(2\pi/a)}\cdot\left(\frac{a}{2\pi\langle r\rangle}\right)^2\ll 1\,.$$

This ratio is of order unity when $\langle r\rangle$ is about 5 a.u., indicating large uncertainties, but it is gratifying to see that R is smaller for "true" potentials than for point-charge potentials, and that $\langle r\rangle$ values don't get much less than 10 a.u. due to the Pauli exclusion principle. This is illustrated in Fig. 11 where a parametric model potential was used and $\langle r\rangle$ is plotted against binding energy (from Bernholc and Pantelides, 1977). The above argument holds for acceptors and for the intravalley terms of donors. For the intervalley terms, the relevant $\bar{q}$ is not $1/\langle r\rangle$, but the intervalley separation. In that case, for some K's, the corresponding R is actually larger than 1. On the face of it, the approximations would be unacceptable. In fact, Shindo and Nara (1976), Resta (1977), and Altarelli, Hsu, and Sabatini (1977) have recently demonstrated that the $\mathbf{K}_p\neq 0$ terms are substantial and cannot be dropped, if the intervalley kinetic energy matrix elements are assumed negligible (these papers are discussed further in Sec. VIII, E below). On the other hand, Twose's many-valley equations, which were used in the numerical results discussed above, include *both* kinetic- and potential-energy intervalley matrix elements to order k^2, as is done for intravalley terms. An attempt to assess the validity of this approximation

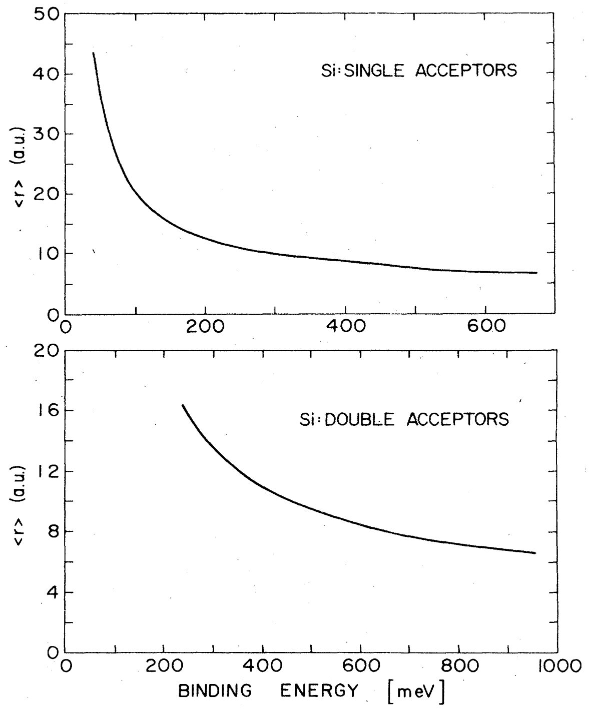

FIG. 11. The average radius of the impurity wave function as a function of the binding energy, as calculated by Bernholc and Pantelides (1977).

was made by Pantelides and Sah (1974). It was shown that higher-order terms of kinetic and potential energies are of opposite sign and tend to cancel. The more recent calculations cited above show that these cancellations involve large terms. The numerical results thus far—those shown in Table III and also results to be discussed in the next two subsections on nonisocoric impurities—show that the cancellations of higher-order terms are quite effective for Si:donors, perhaps due to a combination of the particular nature of the Si band structure and the other approximations used in the calculations (spherically averaged effective-mass and simple hydrogenic trial functions). Folland (1977) has calculated higher-order intervalley terms and found a strong cancellation to occur. Other cancellations are also known to occur, contributing to the adequacy of the EME's. For example, as discussed in Sec. VII contributions from the higher conduction bands cancel contributions from the valence bands. Numerical work by Jaros (1974) has shown that individual terms are in fact large, but the cancellations are quite effective. We are certainly dealing with regions where the validity of the EME's cannot be demonstrated rigorously, so that further probing of the relevant approximations is necessary.

In the end, our conclusions are that EME's seem to work well even when simple estimates suggest that the approximations are no longer valid. It is perhaps fortuitous and rather dangerous that terms that are left out are individually large, but tend to cancel each other. Such terms include contributions from other bands, which for midgap levels tend to cancel, since, as we have seen in Sec. VII, bands from above push down and bands from below tend to push up. As a result, the position of the energy level may be determined rather well, but the wave function is poor. For example, the need for many-band wave functions, even for shallow levels, was demonstrated by Ivey and Mieher (1975) in their analysis of ENDOR data. It is hoped that future research will shed more light on the cancellations and other approximations, such as the isotropic-mass approximation for donors, which result in the good agreement with experiment reflected in Table III and in the pseudopotential calculations for nonisocoric impurities which we describe next.

B. Nonisocoric impurities: Pseudopotentials

As we saw earlier in this section, the difficulty for nonisocoric impurities arises because of the presence of extra (or fewer) core levels. These difficulties were recognized early (see, e.g., Kohn, 1957), but no techniques were available to deal with them. Csavinszky (1963, 1965) in fact recognized that "true" potentials cannot be used in a variational calculation in EME's. Procedures for handling variational collapses to core states had already been developed [orthogonalized plane waves (OPW), Herring, 1940; pseudopotential theory, Phillips and Kleinman, 1959; Bassani and Celli, 1961; Cohen and Heine, 1961; Austin, Heine, and Sham, 1962], but it was not clear how to handle the orthogonalization part when a *difference* of two atomic potentials is involved. The first attempt to accomplish this task was by Morita and Nara (1966), whose approach is comparable to the OPW (Herring, 1940) method in that the "true" potential is retained but the variational function is properly orthogonalized. In the impurity problem, the "plane-wave" part is the smooth Bloch function $\phi_{\mathbf{k}}^0(\mathbf{r})$ from a pseudopotential band-structure calculation. Morita and Nara then make the ansatz definition

$$\tilde{\phi}_{\mathbf{k}}^0 = \phi_{\mathbf{k}}^0 + \sum_t \langle \psi_{ct}^0 | \phi_{\mathbf{k}}^0 \rangle \psi_{ct}^0 - \sum_{t'} \langle \psi_{ct'} | \phi_{\mathbf{k}}^0 \rangle \psi_{ct'} \,. \tag{8.19}$$

where ψ_{ct}^0 are host core wave functions and $\psi_{ct'}$ are impurity core wave functions, and expand the impurity wave function in terms of the $\tilde{\phi}_{\mathbf{k}}^0$. Equation (8.19) is motivated by the OPW expression

$$|OPW\rangle = |\mathbf{k}\rangle - \sum_t \langle \psi_{ct}^0 | \mathbf{k} \rangle |\mathbf{k}\rangle \,, \tag{8.20}$$

where $|\mathbf{k}\rangle$ is a plane wave. Morita and Nara's objective was to write down a wave function which by construction is orthogonal to the $\psi_{ct'}$. It can immediately be demonstrated, however, by evaluating $\langle \tilde{\phi}_{\mathbf{k}}^0 | \psi_{ct'} \rangle$, that $\tilde{\phi}_{\mathbf{k}}^0$ is *not* orthogonal to the $\psi_{ct'}$, not even approximately so. It is unclear why Morita and Nara's calculation did not collapse to core levels. An expression similar to (8.19), but differing in an important aspect, was first obtained by Pantelides and Sah (1974b) and will be derived and discussed shortly.

An alternative approach to handle orthogonality collapses is via the rigorous pseudopotential theory, as developed in papers cited above. The first formulation of the impurity problem in a pseudopotential scheme was given by Glodeanu (1965a). He mixed "true" and "pseudo" quantities by expanding pseudo-wave-functions in terms of true Bloch functions. A variety of cases were studied formally in a series of papers (1965, 1968, 1969a, 1969b), but no pseudopotentials were acutally calculated. In particular, Glodeanu's work after several reformulations did not show how substitutional impurities should be handled, i.e., which set of core states ought to appear where.

Another attempt to deal with the problem was made by Yi-Hsiang and Yu-Ping (1966). They introduced the rather unwieldy concept of orthogonalized Bloch functions, i.e.,

$$\tilde{\psi}_{\mathbf{k}} = \psi^0_{\mathbf{k}} - \sum_{ct'} \langle \psi_{ct'} | \psi^0_{\mathbf{k}} \rangle \psi_{ct'} . \tag{8.21}$$

These functions are in fact orthogonal to the $\psi_{ct'}$ and would thus help avert an orthogonality catastrophe, but they are not necessarily suitable for the expansion of the impurity wave function for an EMT-type theory. Note that the $\psi^0_{\mathbf{k}}$ have a number of nodes already, so that the resulting nodal structure of $\tilde{\psi}_{\mathbf{k}}$ could be detrimental. It was not demonstrated that effective-mass-type equations could be derived in an internally consistent development, and no calculations were pursued.

The impurity problem was formulated in terms of pseudopotential theory independently by Hermanson and Phillips (1967). In their identification of the perturbation potential in the pseudorepresentation, however, they separated out as the unperturbed Hamiltonian a quantity which was not the perfect-crystal pseudo-Hamiltonian and did not have periodicity. The impurity pseudo-potential ended up depending only on the imperfect-crystal core wave functions. They were aware of the problem, but suggested that the symmetry-breaking terms would be negligible. As we shall soon demonstrate, however, these "symmetry-breaking" terms must be included in the perturbation potential for an internally consistent theory and are in fact not negligible. Hermanson and Phillips applied their theory to determine central-cell corrections to the hydrogenic potential for rare-gas solids.

In a later formulation, Ning and Sah (1971) introduced a mixed representation (expanding pseudofunctions in terms of true Bloch functions), arriving at an impurity pseudopotential identical to that of Hermanson and Phillips (1967), i.e., containing only the impurity core wave functions. No calculations using the form of the pseudopotential were pursued.

The above difficulties were resolved by a formulation of the impurity problem in the pseudopotential representation given by Pantelides (1973) and by Pantelides and Sah (1974b). The starting point is the general theory of pseudopotentials, as given by Cohen and Heine (1961) and Austin, Heine, and Sham (1962), which applies to any system described by a Hamiltonian H whose eigenstates ψ_n, obtained from

$$H\psi_n = (T+V)\psi_n = E_n\psi_n , \tag{8.22}$$

may be divided into two groups, the "core" states ψ_c and the remaining states ψ_v. One then defines a pseudo-Hamiltonian H_p by

$$H_p = H + \sum_c |\psi_c\rangle\langle F_c| , \tag{8.23}$$

where the F_c are arbitrary functions. It is then shown (Austin, Heine, and Sham, 1962) that

$$H_p\phi_n = E_n\phi_n \tag{8.24}$$

and

$$\psi_n = \phi_n - \sum_c \langle\psi_c|\phi_n\rangle\psi_c . \tag{8.25}$$

We will discuss convenient choices of F_c later on.

The application to the impurity problem is then straightforward, as follows. Apply the above transformation to the perfect-crystal Hamiltonian (6.1). According to Eqs. (8.23)-(8.25), the resulting pseudo-eigenvalue problem is

$$H^0_p\phi^0_{n\mathbf{k}} = \left[T + V^0 + \sum_t |\psi^0_{ct}\rangle\langle F^0_{ct}|\right]\phi^0_{n\mathbf{k}} = E^0_{n\mathbf{k}}\phi^0_{n\mathbf{k}} , \tag{8.26}$$

where

$$\psi^0_{n\mathbf{k}} = \phi^0_{n\mathbf{k}} - \sum_t \langle\psi^0_{ct}|\phi^0_{n\mathbf{k}}\rangle\psi^0_{ct} . \tag{8.27}$$

Now apply the transformation separately to the imperfect-crystal Hamiltonian (6.22). Again according to Eqs. (8.23)-(8.25), the result is

$$H_p\phi_v = \left[T + V + \sum_{t'} |\psi_{ct'}\rangle\langle F_{ct'}|\right]\phi_v = E_v\phi_v , \tag{8.28}$$

where

$$\psi_v = \phi_v - \sum_{t'} \langle\psi_{ct'}|\phi_v\rangle\psi_{ct'} . \tag{8.29}$$

Note that by dealing with the two crystals separately no ambiguity is present and the two sets of core states ψ^0_{ct} (all the core orbitals in the perfect crystal) and $\psi_{ct'}$ (all the core orbitals in the imperfect crystal) cannot be confused. Now the task is to identify the terms in (8.28) which are to be separated out as a perturbation to the pseudo-Hamiltonian (8.26). The result is also unambiguous. We get for the impurity pseudopotential

$$U_p = H_p - H^0_p = \left[V + \sum_{t'} |\psi_{ct'}\rangle\langle F_{ct'}|\right] - \left[V^0 + \sum_t |\psi^0_{ct}\rangle\langle F^0_{ct}|\right] . \tag{8.30}$$

This is the most general form of the impurity pseudopotential. Notice that if $F^0_{ct} = F_{ct'} = 0$, everything reduces to the "true" quantities, including U_p, which becomes $V - V^0 = U$. The pseudopotential transformation, therefore, will not help unless a judicious choice of the functions F^0_{ct} and $F_{ct'}$ is made. Since we wish to derive effective-mass-type equations, U_p must not bind any core states, which turned out to be the effective criterion for the validity of the EMT. This requirement might be accomplished if neither of the two terms in square brackets in (8.30) binds core states. Austin, Heine, and Sham (1962) have demonstrated that the choice $\langle F_c| = -\langle\psi_c|V$ for the general system (8.23) accomplishes maximum cancellation and the smoothest pseudofunctions. This choice is therefore best for developing an EME from (8.28). We therefore have, from (8.30)

$$U_p = U + U_R \tag{8.31}$$

where

$$U_R = \sum_t |\psi^0_{ct}\rangle\langle\psi^0_{ct}| V^0 - \sum_t |\psi_{ct'}\rangle\langle\psi_{ct'}| V . \tag{8.32}$$

Now, for the development of a pseudo-EMT, the procedure is again straightforward and unambiguous. We expand

$$\phi_v = \sum_{n\mathbf{k}} f_{n\mathbf{k}} \phi^0_{n\mathbf{k}} \tag{8.33}$$

and proceed in a manner analogous to that used in Sec. VII, except now *everything* is "pseudo" (see Pantelides and Sah, 1974b). The result is the same EME's as before with U_p also appearing as before. Note that it is *not* correct that one can simply have an EME with a "true" potential, a pseudopotential, or something else that one may wish to construct. Recall that we found that when "true" qualities are used, the EME is valid *only* for isocoric impurities. Now, when pseudoquantities are used, the EME *may* be valid for nonisocoric impurities depending on the choice of the F_c. The choice (8.32) is the best candidate. It has the important property that it reduces to the "true" EME for *isocoric* impurities for which $U_R \simeq 0$ since $\psi^0_{ct} \simeq \psi_{ct'}$ (this statement has been verified in calculations; see below).

Before we develop U_R for calculations, let us first turn to the form of the wave functions arising from this theory, when a one-band approximation is made in (3.38). Note that the one-band approximation is for the smooth pseudofunction ϕ_v, not the true function ψ_v, which is given by (8.29). Substituting (8.35) in (8.29) and dropping the sum over bands n, we get

$$\psi_v = \sum_{\mathbf{k}} f_{\mathbf{k}} \left[\phi^0_{\mathbf{k}} - \sum_{t'} \langle\psi_{ct'}|\phi^0_{\mathbf{k}}\rangle \psi_{ct'} \right] . \tag{8.34}$$

Clearly, ψ_v is implicitly expanded in terms of the functions in square brackets which are orthogonal to the $\psi_{ct'}$, whereby an orthogonality catastrophe is averted. The $\phi^0_{\mathbf{k}}$ are the smooth pseudo-Bloch functions, however, not the oscillatory Bloch functions used in an *ad hoc* fashion by earlier formulations. The functions in square brackets are the ones that Morita and Nara (1966) would have to use in their formalism in order to accomplish the desired result. Notice that here (8.34) is not postulated but is derived from a natural and systematic application of pseudopotential theory.

An alternative and physically transparent form of (8.34) can be obtained by using (8.27) for $\phi^0_{n\mathbf{k}}$. We get

$$\psi_v = \sum_{\mathbf{k}} f_{\mathbf{k}} \left[\psi^0_{\mathbf{k}} + \sum_t \langle\psi^0_{ct}|\phi^0_{n\mathbf{k}}\rangle \psi^0_{ct} - \sum_{t'} \langle\psi_{ct'}|\phi^0_{n\mathbf{k}}\rangle \psi_{ct'} \right] . \tag{8.35}$$

This is the formula that accomplishes *reorthogonalization* of the $\psi_{\mathbf{k}}$. When compared with that postulated by Morita and Nara, Eq. (8.19), we note that the important difference is the first term in the square brackets of (8.40), which is a true Bloch function, not a pseudo-Bloch function. When (8.35) is compared with (8.31) and (8.32) we note that U_R is a result of the need for *re*orthogonalization of the $\psi^0_{\mathbf{k}}$, whereby both sets of core orbitals appear in a natural setting. For isocoric impurities, both (8.31) and (8.35) reduce to the "true" counterparts.

Another interesting observation is that the expansion (8.35) for ψ_v appears to implicitly be a many-band expansion. First the main band contributes via the $\psi^0_{\mathbf{k}}$. Then the host's core bands contribute via the ψ^0_{ct}. Finally, the $\psi_{ct'}$ may correspond to higher conduction bands of the host, as in the case of, say, Si : Sb, where the 4*s*, 4*p*, and 4*d* core states of Sb belong to the same subspace as some of the conduction bands of Si. In other words, the pseudopotential formalism implicitly makes use of whatever core states are needed to construct an appropriate "true" trial function, so that a one-band expansion for the smooth pseudofunction should be adequate. Along the same vein, in the pseudoequations, it is the "pseudoelectron," described by ϕ_v, which has a kinetic energy given by the effective-mass expression $\hbar^2k^2/2m^*$ and a potential energy $U + U_R$. For the "true electron," described by ψ_v, U_R is kinetic energy in disguise, so its net kinetic energy is actually $\hbar^2k^2/2m^* + U_R$. As we shall see, U_R is nonzero only in the impurity cell so that the true electron ends up having an effective-mass kinetic energy outside the impurity cell, but "something else" inside. For isocoric impurities, $U_R \simeq 0$ everywhere, and there is no distinction between pseudoelectron and true electron.

For calculational purposes (8.32) is developed further by writing V and V^0 as in Eqs. (8.2) and (8.3). The net result is

$$U_p = U_{pb} + U_{ps} , \tag{8.36}$$

where

$$U_{pb} = [v_c - v^0_c] + \sum_t |\psi^0_{ct}\rangle\langle\psi^0_{ct}| v^0_c - \sum_{t'} |\psi_{ct'}\rangle\langle\psi_{ct'}| v_c \tag{8.37a}$$

for substitutional impurities, and

$$U_{pb} = v_c - \sum_{t'} |\psi_{ct'}\rangle\langle\psi_{ct'}| v_c \tag{8.37b}$$

for interstitial impurities. Expression (8.37a) may be compared with the Hermanson-Phillips result which has $\psi_{ct'}$ in both terms in the second square brackets. The discrepancy, a term equal to

$$\sum_t |\psi^0_{ct}\rangle\langle\psi^0_{ct}| v^0_c - \sum_{t'} |\psi_{ct'}\rangle\langle\psi_{ct'}| v^0_c , \tag{8.38}$$

was "absorbed" in the unperturbed Hamiltonian. The term indeed breaks the symmetry and is not negligible for nonisocoric impurities, where the whole issue is the fact that the $\psi_{ct'}$ are drastically different from the ψ^0_{ct}.

Numerical calculations have been carried out thus far for single and double donors in Si, both substitutional and interstitial, using the same multivalley EME's discussed earlier (Pantelides and Sah, 1974b). Figure 12 illustrates the cancellations occuring in the pseudopotential representation. The results for binding energies of nonisocoric impurities are given in Table IV. The range of energies over which good numerical results are obtained is quite large, from the shallowest single interstitial Li to the intermediate double interstitial Mg and the deeper double substitutional group-VI elements.

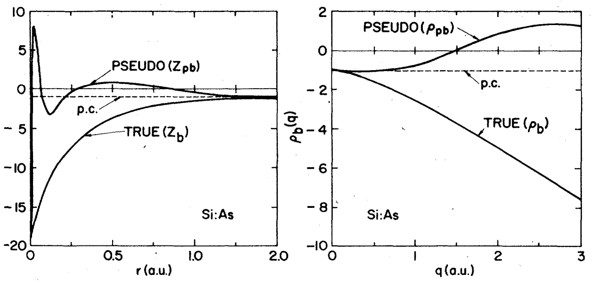

FIG. 12. The "true" and "pseudo" impurity potential for Si:As demonstrating the cancellation in both real and reciprocal space.

The failure for Si : Sb and Si : Te is probably due to their large size and an additional correction has been investigated by Pantelides and Sah (1974b) due to the contraction of their core orbitals in the Si lattice. A rather interesting case is the prediction that Si : N is deep and can bind a second electron, which is indirectly confirmed by the fact that GaP : O_P, for which the impurity pseudopotential is almost identical with that of Si : N, is known to behave in the same way. The predicted binding energy of Si : N compared with those of Si : P, As, Sb, scales nicely with that of GaP : O_P compared with those of GaP : S, Se, Te (Pantelides, 1974b).

The pseudopotential is also found to produce small corrections to the "true" potential binding energies of isocoric donors. (See Table III) They are somewhat less reliable because they are differences of two large numbers. Their physical significance is rather interesting in that they reflect a tiny change in kinetic energy arising from the fact that the Bloch function must modify its nodal structure slightly in response to the tightening of the core orbitals.

TABLE IV. Binding energies in meV for nonisocoric donors in Si using rigorous pseudopotential and the Appapillai–Heine model potentials. Theoretical values are from Pantelides and Sah (1974b) and Pantelides (1974a, 1975).

	Impurity	Pseudo	Model	Experiment
A. Interstitial				
	Li	33.8	33.1	31.0 [a]
	Na	34.6	32.4	···
	Be^+	385.6	239.6	···
	Be^0	146.5	103.6	···
	Mg^+	259.0	190.2	256.5 [b]
	Mg^0	98.0	81.3	107.5 [b]
B. Substitutional				
	N^0	335.9	27.0	···
	N^-	52.5	no binding	(45.0) [c]
	As	53.1	40.7	53.7 [d]
	Sb	31.7	35.3	42.7 [d]
	Se^+	921.3	559.0	···
	Se^0	358.4	265.3	···
	Te^+	246.0	256.2	···
	Te^0	71.9	111.7	···

[a] Aggarwal, Fisher, Mourzine, and Ramdas (1965).
[b] Ho and Ramdas, 1972.
[c] Zorin, Parlor an Tetel'baum, 1968; tentative assignment by Pantelides and Sah (1974b).
[d] Aggarwal and Ramdas, 1965.

Pseudopotential calculations for *shallow* donors in Si have also been reported by Schechter (1973, 1975). In the first paper, linear screening was used to screen the "true" potentials. The impropriety of this procedure was pointed out by Pantelides and Sah (1974b), and Schechter repeated his calculations in the second paper. The results changed considerably and are similar to those shown in Table 4. Schechter also used his wave functions to analyze ENDOR data.

C. Model impurity potentials

A model potential is generally defined as any potential that gives the same eigenvalue spectrum as the true potential, over a certain energy range. Useful model potentials are defined so as to have no core states, their lowest-energy state coinciding with the first valence state of the atom, molecule, or solid. Such model potentials are often referred to as empirical pseudopotentials because they are usually parametric forms, with the parameters fit to measured quantities. One disadvantage of model potentials is that they often have to be energy-dependent. This requirement arises from the fact that it is not possible to construct a model potential that reproduces the entire energy eigenvalue spectrum. Thus a given model potential is valid over a certain energy range.

Model potentials actually preceded pseudopotentials (see discussion by Heine and Abarenkov, 1964 and Abarenkov and Heine, 1965), but they became popular shortly after the advent of pseudopotentials, since the theory of the latter provided more fundamental justification for their use. A particular form, introduced by Heine and Abarenkov (1964), proved to be quite useful and extensive tables of atomic model potentials are available (Animalu, 1965; Appapillai and Heine, 1972). The Heine-Abarenkov model potential for an ion of charge z is defined by

$$\begin{aligned} V_M(r) &= \sum_l A_l P_l, \quad r < R_M \\ &= -z/r, \quad r > R_M \end{aligned} \tag{8.39}$$

where R_M is a radius of order 1 Å, P_l are angular momentum projection operators, and A_l are energy constants. Thus, for each l value, $V_M(r)$ is a square well with a Coulombic tail (Fig. 13).

Heine–Abarenkov model potentials were first used for the impurity problem by Jaros (1969) for shallow donors in Si. He used the one-valley EME, however, and good agreement with experiment was obtained by introducing a position-dependent effective mass which starts out with the free-electron mass at $r = 0$ and asymptotically becomes m^*. It can be seen from Eq. (6.22), however, that the free-electron mass cannot be used with the perturbation potential. Instead, if used in the central cell, one must use the total potential in that region (Pantelides, 1974a).

Model potentials somewhat akin to the Heine–Abaren-

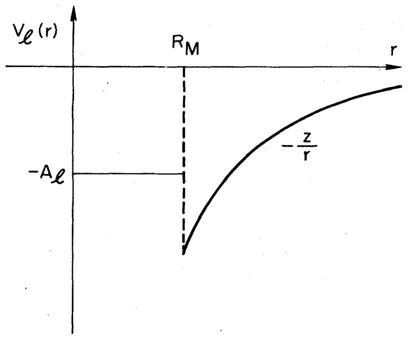

FIG. 13. The Abarenkov–Heine-model potential for atoms and ions. See text for the symbols.

kov form were subsequently used by Ning and Sah (1973) for donors in Si. These authors introduced a single parametrized model potential for the impurity potentials and fitted its two parameters to the A_1 and T_2 states of each impurity. The study confirmed the applicability of the many-valley EME for Si and showed that a plausible potential could be found to reproduce the deep Si : S levels (approximately double the model potential obtained for Si : P). Predictions were thus made for the double Si : Se and Si : Te donors by doubling the model potentials of Si : As and Si : Sb, respectively. The wave functions obtained for Si : P, As, and Sb were used to analyze ENDOR experimental data and to calculate optical absorption cross sections.

Heine–Abarenkov model potentials for shallow and deep donors in Si were employed by Pantelides (1974a) for the purpose of comparing with the results obtained previously in terms of "true" impurity potentials and rigorous pseudopotentials. The results, using model potentials from Appapillai and Heine (1972), are also shown in Table IV. The most outstanding disagreement is for Si : N, for which the model potential gives much weaker binding than the first-principles pseudopotentials.

More recently, model potentials were employed by Bernholc and Pantelides (1977) in a study of acceptors in Si and Ge. Detailed agreement with experiment was not very good, perhaps due to insufficiently flexible trial functions (only $l = 0$ and $l = 2$ spherical harmonics were used, whereas the work of Baldereschi and Lipari, 1976, on point-charge potentials showed that higher l values are important). One important result of that work is that for double acceptors in Si and for triple acceptors in Ge where the point-charge potential was found to give extremely deep levels, model potentials result in levels in the range of the observed values.

Model potentials have the advantage that they are simpler to work with and may be the ones that will eventually give a successful effective-mass theoretic description for most impurities, except for first-row elements. (Transition-metal impurities and lattice defects are, however, outside the EMT.) Model potentials must be constructed very accurately and then the EME's must be solved accurately as well. When this is accomplished, this author believes that, with the possible exception of the heavier elements, single and double acceptors in Si and single, double, and perhaps triple donors and acceptors in Ge will be described adequately by the EMT. The III-V and II-VI compounds are another story, however, and we turn to those for some rather disappointing conclusions.

D. Compound semiconductors

In the case of compound semiconductors, two new factors severely complicate the problem. The first complication arises because the compound semiconductors consist of alternately postively and negatively charged ions, which respond to the introduction of foreign charges differently from the neutral atoms of the homopolar semiconductors, such as Si and Ge. In order to understand the complication we first consider an electron in the conduction bands of a perfect polar crystal. The electron constantly interacts with the lattice, which even at $T = 0\,^\circ\mathrm{K}$ has the "zero-point motion." In polar crystals, this interaction is qualitatively different and much stronger than in homopolar materials because moving charged ions set up long-range Coulomb fields. The true stationary states in the crystal are the quasiparticles known as *polarons*, which consist of an electron and a cloud of virtual longitudinal optical (LO) phonons. The electrons and the phonons are constantly exchanging energy and momentum among themselves, while maintaining a constant energy and constant momentum for the polaron. The problem of binding carriers to impurity potentials may thus be viewed as a problem of binding polarons. Unlike ordinary electrons in homopolar semiconductors, however, polarons are dynamic entities in that, speaking in pictorial terms, they may "shed" part or all of the phonon cloud when they get bound. The quantum-mechanical description of this process is very complicated and can be carried out only in certain limiting cases (Schultz, 1962; Platzman, 1962; Bajaj, 1972; Larsen, 1972). Since virtually no applications have been carried out for real impurities in real semiconductors we will only give a brief account of the subject.

The Hamiltonian for a polaron is

$$H_{\mathrm{pol}} = H^0 + H_{ph} + H_{ep} \tag{8.40}$$

where H^0 is the electronic Hamiltonian of the perfect crystal; H_{ph} is the lattice Hamiltonian which is most conveniently written in second-quantized form as

$$H_{ph} = \sum_q \hbar\omega_q a_q^\dagger a_q\,, \tag{8.41}$$

where ω_q are the phonon frequencies, and $a_q^\dagger$ and a_q are the phonon creation and annihilation operators, respectively (see Kittel, 1963). Finally, H_{ep} represents the electron–phonon interactions, and is given by

$$H_{ep} = \sum_q \left[V_q a_q e^{i\mathbf{q}\cdot\mathbf{r}} + V_q^* a_q^\dagger e^{-i\mathbf{q}\cdot\mathbf{r}}\right], \tag{8.42}$$

where V_q represents the strength of electron–phonon coupling. V_q was first evaluated by Fröich (1937, 1962) by requiring that the electron–phonon coupling reduce the interaction between two well separated electrons from $e^2/\epsilon_\infty r$ to $e^2/\epsilon_0 r$. Here ϵ_∞ is the *high-frequency* or optical dielectric constant of the solid, which corresponds to screening by electrons only (it is measured with fields whose frequency is much larger than phonon frequencies, at which phonons cannot respond); ϵ_0 is the *static* dielectric constant, which includes full screening by both electrons and lattice (it is measured with static, zero-frequency fields, at which phonons do respond). The electronic Hamiltonian H^0 i s treated in an effective-mass approximation and the result is

$$V_q = -(i\hbar\omega_q/q)(\hbar/2m^*\omega_q)^{1/4}(4\pi\alpha/\Omega)^{1/2}. \tag{8.43}$$

The dimensionless coupling constant α, introduced by Frölich, is defined by

$$\alpha = \frac{e^2}{\hbar}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0}\right)\left(\frac{m^*}{2\hbar\omega_q}\right)^{1/2}. \tag{8.44}$$

In the above equations, m^* is the effective mass for the carrier (the theory has not been extended to treat other than simple, nondegenerate parabolic band extrema) and Ω is the volume of the crystal. Solutions of $H_{\rm pol}$ may be obtained by treating H_{ep} as a perturbation. In the case of dispersionless phonons, one gets a constant shift of the carrier energy of $-\alpha\hbar\omega$ and a change of m^* into $m^*/[1-\alpha/6]$.

We now turn to the problem of binding an electron to an impurity in an ionic or polar solid. The only case that has been studied is that of a point-charge impurity. In the absence of electron–phonon interactions, the impurity potential is assumed to be hydrogenic, screened by ϵ_∞. Including electron-phonon interactions *and* point-charge-phonon interactions, the total Hamiltonian is

$$H = H^0 + H_{\rm ph} - (e^2/\epsilon_\infty r) + H', \tag{8.45}$$

where H' is now

$$H' = \sum_q [V_q a_q(e^{i\mathbf{q}\cdot\mathbf{r}} - 1) + V_q^* a_q^\dagger(e^{-i\mathbf{q}\cdot\mathbf{r}} - 1)]. \tag{8.46}$$

V_q is assumed the same as in the perfect crystal.

Clearly, a number of limiting cases emerge (Schultz, 1962), depending on the values of α and the phonon frequencies ω. We wish here to concentrate on a given material, for which α and ω are fixed. We note immediately that, for small r, H' becomes negligible, because $e^{i\mathbf{q}\cdot\mathbf{r}} \approx e^{-i\mathbf{q}\cdot\mathbf{r}} \approx 1$, so that the lattice can be neglected entirely. Physically, what happens is that the electron is very close to the point charge, in effect neutralizing it, so that the lattice does not respond at all. Accordingly, for very tightly bound states, coupling to the lattice can be neglected. It has often been said that $-e^2/\epsilon_\infty r$ can then be used (as opposed to $-e^2/\epsilon_0 r$), but, in actuality, one must do much better. For a point charge, one must include the q dependence (or, equivalently, r dependence) of ϵ_∞. For real impurities, a good potential is necessary. More importantly, for H' to be negligible, the radius of the bound state must be of the order of the interatomic spacing, in which case effective-mass theory breaks down.

The other limit, naturally, is that of very large orbits for the bound states. In that case, H' cannot be neglected. The usual procedure is to first make the canonical transformation

$$\hat{H} = S^{-1}HS, \tag{8.47}$$

where

$$S = \sum_q \exp[-(a_q^\dagger + a_q)V_q/\hbar\omega]. \tag{8.48}$$

The result is

$$\hat{H} = H^0 + H_{\rm ph} - (e^2/\epsilon_0 r) + H'', \tag{8.49}$$

where

$$H'' = \sum_q [V_q a_q e^{i\mathbf{q}\cdot\mathbf{r}} + V_q^* a_q^\dagger e^{-i\mathbf{q}\cdot\mathbf{r}}]. \tag{8.50}$$

The effect of this transformation is to eliminate the point-charge–lattice interaction in H' and replace $-e^2/\epsilon_\infty r$ by $-e^2/\epsilon_0 r$. After the canonical transformation, it becomes appropriate to first leave out the impurity potential and solve the free polaron problem. Subsequently, the perturbation $-e^2/\epsilon_0 r$ is introduced and the polaron is bound to it in a hydrogenic orbit. Clearly, however, this limit is valid when orbits are very large so that a free polaron does not change its character appreciably when it gets bound. This is the limit that is usually used in describing shallow donors and acceptors in polar semiconductors. All lattice effects are simply lumped into ϵ_0 and the experimental values of m^*, which is the polaron mass. The approximation works well when m^* is of order $0.1m_0$ and ϵ_0 is large (>10) so that binding energies are of order 10 meV and orbits are >100 a.u. Examples are donors in GaAs, InSb, etc.

An improved solution of Eq. (8.49) can be obtained by treating H'' as the perturbation and treating the rest of the Hamiltonian first. The unperturbed solution is again hydrogenic, but m^* is the bare effective electron mass (as opposed to the polaron effective mass). Inclusion of H'' by perturbation theory involves rather tedious algebra and additional approximations. In the simplest of the calculations, valid for $\alpha \ll 1$, first done by Platzman (1962) and then improved by Sak (1971), one gets for the binding energy

$$E_B = (1 + \frac{\alpha}{6} + \frac{1}{24}\alpha E_B^0/\hbar\omega)E_B^0, \tag{8.51}$$

where E_B^0 is the hydrogenic binding energy with bare effective mass. This expression is an expansion in powers of E_B^0/ω and is, therefore, valid for $E_B^0 \ll \omega$. We see that the first correction, $\alpha/6$, simply amounts to replacing the bare effective mass by the polaron effective mass. The next correction, $\alpha E_B^0/24\hbar\omega$, is a true improvement over the simpler approximation discussed earlier, in which all lattice effects were lumped in ϵ_0 and m^*. Similar expansions have been obtained for the degenerate $2s$ and $2p$ hydrogenic levels (Sak, 1971) and the result is a splitting analogous to the Lamb shift caused by the photon field in the free hydrogen atom.

Alternative approximations in evaluating H'' by perturbation theory have been made by Stoneham (1970) and by Bajaj (1970, 1971). Engineer and Tzoar (1972) evaluated the correction numerically, without needing to impose $E_B^0 \ll \omega$. All these calculations, however, are for a simple parabolic band and for a potential of the form $-e^2/\epsilon_\infty r$. In fact as the binding energy begins to approach $\hbar\omega$, central-cell effects begin to become important and a more accurate impurity potential is needed. Recently, Bernholc and Pantelides (1977) calculated binding energies for a point-charge acceptor impurity in compound semiconductors using the full 6×6 $\mathbf{k}\cdot\mathbf{p}$ matrix and $\epsilon_\infty(q)$, which amounts to neglecting lattice screening. Their use of the effective-mass parameters of Lawaetz (1971), which are bare with respect to phonons, makes the calculation internally consistent. However, the resulting orbits are fairly large (making the EMT valid), so that the effect of electron–phonon interaction cannot be neglected. Bernholc and Pantelides (1977) also carried out the calculations using $\epsilon_0(q)$, obtained by scaling $\epsilon_\infty(q)$ so that $\epsilon_0(0)=\epsilon_0$. The calculation corresponds to a potential that has the proper asymptotic forms but only one of many possible interpolated forms at intermediate r values. The proper calculation would have to treat lattice rearrangement in a self-consistent way, as recently done for an exciton in a polar semiconductor with simple parabolic bands by Pollmann and Büttner (1975). The conclusion, therefore, is that so far there have not been any calculations which properly take into account electron-phonon interactions for impurities in real semiconductors, except for the trivial cases where a hydrogenic calculation with ϵ_0 and the polaron mass m^* is adequate.

Even if polaron effects could be included accurately, however, impurities in polar semiconductors have another major difficulty associated with them (Pantelides, 1975a; Bernholc and Pantelides, 1977). This difficulty is best illustrated by looking at the two isocoric single acceptors in GaP, namely GaP : $\mathrm{Zn_{Ga}}$ and GaP : $\mathrm{Si_P}$. The point-charge model should describe both of them well. However, the two acceptors have substantially different binding energies, namely 64 meV (Dean, Faulkner, Kimura, and Ilegems, 1971) and 204 meV (Dean, Frosch, and Henry, 1968), respectively. The difference between cation single acceptors and anion single acceptors are even more pronounced in II-VI compounds, such as CdS. These differences have been attributed by Phillips (1973) to electronegativity differences. More recently, Pantelides (1975) and Bernholc and Pantelides (1977) pointed out two factors that contribute to these differences in the context of a quantum-mechanical calculation. One is the site dependence of screening which is neglected when $\epsilon(q)$ is used. Physically, the site dependence arises because the electrons are distributed nonuniformly, the majority being around anions. Point charges, positive or negative, would thus be screened more effectively on *anions*. (Bernholc and Pantelides, 1977, made an inappropriate distinction between the screening of positive and negative charges.) Site-dependent screening, therefore, would tend to push point-charge binding energies in the wrong direction. The second factor, which must then overcome the site-dependent screening and yield the observed anion–cation difference, is the need for evaluating the matrix ele- $\langle\psi|U|\psi\rangle$ by going beyond the effective-mass approximation. Recall that ψ is expanded in terms of Bloch functions, and the latter are expanded in terms of plane waves of the reciprocal lattice vectors $\mathbf{K}$. [Cf. Eqs. (7.4)–(7.7)] By retaining only $\mathbf{K}=0$, site dependence is lost. If more $\mathbf{K}$ vectors are retained, then the two sites are going to be differentiated. In particular, for acceptors, Bloch functions at the top of the valence bands have larger amplitudes around anions, whereby $\langle\psi|U|\psi\rangle$ and the resulting binding energy is larger for anion-site acceptors than for cation-site acceptors. Clearly, such calculations are outside the EMT, as developed thus far.

E. Recent developments

In recent papers, Shindo and Nara (1976) and Altarelli, Hsu, and Sabatini (1977) offer an alternative form of many-valley effective-mass equations. They suggest that (a) intravalley terms should remain as in the traditional EME's discussed earlier; (b) intervalley kinetic energy terms should be neglected as small; and (c) intervalley potential-energy terms should be evaluated by approximating

$$\langle\psi^0_{\mathbf{k}}|U|\psi^0_{\mathbf{k}'}\rangle \approx \langle u_{\mathbf{k}_i}e^{i\mathbf{k}\cdot\mathbf{r}}|U|u_{\mathbf{k}_j}e^{i\mathbf{k}'\cdot\mathbf{r}}\rangle \tag{8.52}$$

in contrast to the usual

$$\langle\psi^0_{\mathbf{k}}|U|\psi^{0\prime}_{\mathbf{k}} \approx \langle e^{i\mathbf{k}\cdot\mathbf{r}}|U|e^{i\mathbf{k}'\cdot\mathbf{r}}\rangle\,. \tag{8.53}$$

In the language used by Altarelli *et al.* (1977), (8.52) corresponds to including the Umklapp ($\mathbf{K}_p\neq 0$) terms in the expansion of $u^*_{\mathbf{k}_i}(\mathbf{r})u_{k_j}(\mathbf{r})$, instead of just the $\mathbf{K}_p=0$ term.

At first glance, the use of (8.52) instead of (8.53) for intervalley terms appears to be a definite improvement. A closer examination, however, reveals that it may not necessarily be so. First note that if Umklapp terms are included in the intervalley terms, they must *also* be included in the intravalley terms as well, in order to maintain internal consistency. This requirement, however, can lead to trouble: Intravalley terms (Luttinger and Kohn, 1955; Kittel and Mitchell, 1954) are systematically calculated to order k^2 within $\mathbf{k}\cdot\mathbf{p}$ theory, which corresponds to dropping Umklapp terms in the potential matrix elements. Both the original papers cited above pointed out that one cannot carry some terms to higher order than others, whereby inclusion of Umklapp terms in the potential matrix elements may necessitate going beyond the k^2 approximation in the kinetic energy. Furthermore, in order to keep all terms of the same order, interband matrix elements must also be included.

Numerical results obtained with the point-charge model by Baldereschi (1970), Pantelides and Sah (1974a), and by Altarelli *et al.* (1977) are shown in Table V. Though substantially different approximations were used, the results are comparable. It turns out that the choice of dielectric function alone can introduce large uncertainties (see, e.g., Bernholc and Pantelides, 1977; also Lipari and Baldereschi, 1978).

TABLE V. Binding energies of the split ground states for donors in Si and Ge from three different calculations using the point-charge model and from experiment for the respective isocoric donors.

	Silicon			Germanium	
	A_1	T_2	E	A_1	T_2
Baldereschi (1970)	40.5	29.9	28.8	10.1	9.5
Pantelides and Sah (1974a)	48.8	31.1	30.4	⋯	⋯
Altarelli *et al.* (1977)	47.5	31.4	30.6	12.5	9.7
Experiment (Si:P;Ge:As)	45.5[a]	33.9[a]	32.6[a]	14.0[b]	9.8[b]

[a] Aggarwal and Ramdas (1965).

[b] Reuszer and Fisher (1964).

The role of the Umklapp terms, therefore, still remains to be investigated further. First one must check whether interband matrix elements are in fact negligible. If they are not, many-band formulations would become necessary. If they are, one should next check the need for including k^4 and higher-order terms in the kinetic energy matrix elements. If all goes well, one may then proceed to include Umklapp terms. Contrary to what Shindo and Nara (1976) and Altarelli *et al.* (1977) have done, however, the present author believes that Umklapp terms should be included in *both* intervalley and intravalley terms. One can then quickly deduce that the necessary modification of Twose's MV EME's is to first drop the intervalley kinetic energy and replace

$$U(\mathbf{r}) \to U(\mathbf{r}) u^*_{\mathbf{k}j}(\mathbf{r}) u_{\mathbf{k}j}(\mathbf{r}) \tag{8.54}$$

for *both* $i=j$ (intravalley) and $i \neq j$ (intervalley). For acceptors, the modification would be more troublesome. Whereas in the conventional 6×6 EMT matrix Hamiltonian the impurity potential appears only on the diagonal, inclusion of Umklapp terms would introduce the potential in the off-diagonal elements as well. The new Hamiltonian (Pantelides, 1978) would be

$$D_{ij}(-i\nabla) + U(\mathbf{r}) M_{ij}(\mathbf{r}) , \tag{8.55}$$

where $D_{ij}(\mathbf{k})$ was defined in Sec. VII, and $M_{ij}(\mathbf{r})$ is

$$M_{ij}(\mathbf{r}) = u^*_{i0}(\mathbf{r}) u_{j0}(\mathbf{r}) . \tag{8.56}$$

Here $u_{i0}(\mathbf{r})$ and $u_{j0}(\mathbf{r})$ are the six Bloch functions at the top of the valence bands at $\mathbf{k}=0$ which contain spinors explicitly.

Test calculations must be carried out before the validity and usefulness of these generalized equations become clear. In the meantime we may remark that Umklapp terms have another unpleasant consequence. Recall that in their absence one simply had to choose an impurity pseudopotential, without needing the corresponding Bloch functions. Now, however, Bloch functions are needed, and they must correspond to the same choice for *internal consistency*. For example, if Heine-Abarenkov-model potentials are used for the impurity- and host-atom potentials, the Bloch functions should come from a band structure that made use of the same host-atom model potential. This requirement is very stringent. In particular, it makes the point-charge model somewhat meaningless, because different sets of pseudo-Bloch functions (corresponding to the many possible choices of crystal pseudopotentials) may result in different binding energies. Finally, on the positive side, Umklapp terms are the very ones that were identified by Bernholc and Pantelides (1977) (see also discussion in Sec. VIII.D above) as being responsible for the strong site dependence of acceptor binding energies of acceptors in compound semiconductors. Their proper inclusion in EME's would, therefore, have substantial import.

IX. GENERAL PERTURBATIVE METHODS

In this section, we discuss methods which, like effective-mass theory, build on, or start from, a knowledge of the electronic structure of the perfect host solid. In all these methods, the total crystal potential is written as in Eq. (6.23), namely

$$V = V^0 + U , \tag{9.1}$$

whereby the eigenvalue problem to be solved is given by

$$(H^0 + U)\psi_\nu = E_\nu \psi_\nu . \tag{9.2}$$

Identical equations may be written down, whether the above quantities are viewed as "true," pseudo, or model, in the sense discussed at length in the previous section. The term perturbative in the section title is not meant to imply that the methods are based on a power series in the perturbation U, but rather that they are designed to calculate directly the *changes* produced by the perturbation.

In this section, we describe the mathematical foundations of various perturbative methods and discuss applications to real systems that have been reported thus far deferring a critical comparison of all the methods to Sec. XI, after we have introduced and discussed the nonperturbative methods of Sec. X.

A. Secular-matrix methods

These methods are most conveniently described in the conventional Hamiltonian/wave-function representation in which one seeks to determine the eigenenergies E_ν and eigenfunctions ψ_ν by expanding ψ_ν in terms of a complete set of functions φ_λ in the form

$$\psi_\nu = \sum_\lambda F_\lambda \varphi_\lambda . \tag{9.3}$$

Upon substitution in (9.2), multiplication on the left by φ^*_λ, and integration, one gets the following set of coupled linear algebraic equations

$$\sum_\lambda [\langle \varphi_{\lambda'} | H^0 | \varphi_\lambda \rangle + \langle \varphi_{\lambda'} | U | \varphi_\lambda \rangle - E_\nu \langle \varphi_{\lambda'} | \varphi_\lambda \rangle] F_\lambda = 0 \tag{9.4}$$

or

$$\sum_\lambda [H^0_{\lambda'\lambda} + U_{\lambda'\lambda} - E_\nu S_{\lambda'\lambda}] F_\lambda = 0 \tag{9.5}$$

in obvious notation. The energies E_ν are then the solutions of the generalized eigenvalue problem (9.5). If the basis set is orthonormal so that $S_{\lambda'\lambda} = \delta_{\lambda'\lambda}$, the energies E_ν are simply the eigenvalues of the secular matrix $H^0_{\lambda'\lambda} + U_{\lambda'\lambda}$. The size of this matrix is equal to the number of functions φ_λ one needs to include in the expansion (9.3).

1. The Bloch representation

The most natural choice for the φ_λ is the set of Bloch functions $\psi^0_{n\mathbf{k}}$, which are eigenfunctions of H^0 and are orthonormal. Thus

$$\psi_\nu(\mathbf{r}) = \sum_{n\mathbf{k}} F_{n\mathbf{k}} \psi^0_{n\mathbf{k}}(\mathbf{r}) \tag{9.6}$$

and the set of equations (9.4) becomes

$$E^0_{n\mathbf{k}} F_{n\mathbf{k}} + \sum_{n'\mathbf{k}'} \langle \psi^0_{n\mathbf{k}} | U | \psi^0_{n'\mathbf{k}'} \rangle F_{n'\mathbf{k}'} = E_\nu F_{n\mathbf{k}} . \tag{9.7}$$

The secular matrix to be diagonalized is

$$E^0_{n\mathrm{k}}\delta_{nn'}\delta_{\mathrm{kk}'}+\langle\psi^0_{n\mathrm{k}}|U|\psi^0_{n'\mathrm{k}'}\rangle\,. \tag{9.8}$$

The size of this matrix is equal to the number of k points in the Brillouin zone times the number of bands one includes in the expansion (9.6). Note that a single diagonalization would yield a quasicontinuum of band states plus any bound states in the gaps. In a practical implementation of this method, the necessarily finite sampling of the Brillouin zone makes the method least appropriate for states for which $F_{n\mathrm{k}}$ varies rapidly with k, such as diffuse, weakly bound, effective-mass-like impurity levels.

2. The Wannier representation

Another possible choice for the φ_λ are the Wannier functions $w^0_n(\mathbf{r}-\mathbf{R}_j)$ of the perfect crystal. We write

$$\psi_\nu(\mathbf{r})=\sum_{nj}F_{nj}w^0_n(\mathbf{r}-\mathbf{R}_j)\,. \tag{9.9}$$

The coefficients are written as F_{nj} in order to explicitly indicate their relationship to the $F_{n\mathrm{k}}$ of (9.6): since the $\psi^0_{n\mathrm{k}}$ and the w^0_{nj} are related by (6.9), we immediately obtain

$$F_{n\mathrm{k}}=N^{-1/2}\sum_j F_{nj}e^{-i\mathrm{k}\cdot\mathrm{R}_j} \tag{9.10}$$

and

$$F_{nj}=N^{-1/2}\sum_{\mathrm{k}}F_{n\mathrm{k}}e^{i\mathrm{k}\cdot\mathrm{R}_j}\,. \tag{9.11}$$

The set of equations resulting from (9.9) is

$$\sum_{n'j'}[E_n(\mathbf{R}_j-\mathbf{R}_{j'})\delta_{nn'}+\langle w^0_{nj}|U|w_{n'j'}\rangle]F_{n'j'}=EF_{nj}\,, \tag{9.12}$$

where

$$E_n(\mathbf{R}_j-\mathbf{R}_{j'})=N^{-1}\sum_{\mathrm{k}}E^0_{n\mathrm{k}}e^{i\mathrm{k}\cdot(\mathrm{R}_j-\mathrm{R}_j')}\,. \tag{9.13}$$

The size of the secular matrix is equal to the number of sites times the number of bands one includes in the expansion (9.9). The number of sites is in turn dictated by the range of the bound-state wave function.

3. Other localized-function representations

One clearly could use any set of localized functions $\phi_\alpha(\mathbf{r}-\mathbf{R}_j)$, in particular the same functions that one might employ in an LCAO-type energy-band calculation for the host crystal. Thus

$$\psi_\nu(\mathbf{r})=\sum_\alpha F_{\alpha j}\phi_\alpha(\mathbf{r}-\mathbf{R}_j)\,, \tag{9.14}$$

and the corresponding set of equations is

$$\sum_{\alpha'j'}[\langle\phi_{\alpha j}|H^0|\phi_{\alpha'j'}\rangle+\langle\phi_{\alpha'j}|U|\phi_{\alpha'j'}\rangle-E_\nu\langle\phi_{\alpha j}|\phi_{\alpha'j'}\rangle]F_{\alpha'j'}=0\,, \tag{9.15}$$

where presumably one knows the matrix elements $\langle\phi_{\alpha j}|H^0|\phi_{\alpha'j'}\rangle$ from the band-structure calculation. Again, the size of the secular determinant is equal to the number of functions that must be included in (9.14) for an accurate expansion of $\psi_\nu(\mathbf{r})$.

B. Determinantal methods

These methods follow along the same lines as the secular-matrix methods described above, but are especially designed for cases in which the range of the perturbing potential is significantly shorter than the range of the bound-state wave function. In order to exploit this distinction one then needs a set of basis functions φ_λ which are appropriately localized.

1. The Koster-Slater method in the Wannier representation

The starting point is the set of equations (9.7) in the Bloch representation which are first rewritten as

$$F_{n\mathrm{k}}+\sum_{n'\mathrm{k}'}\frac{\langle\psi^0_{n\mathrm{k}}|U|\psi^0_{n'\mathrm{k}'}\rangle}{E^0_{n\mathrm{k}}-E_\nu}F_{n'\mathrm{k}'}=0\,. \tag{9.16}$$

The unit operator in the Wannier representation

$$1=\sum_{nj}|w^0_{nj}\rangle\langle w^0_{nj}| \tag{9.17}$$

(representing closure or completeness for the set of functions) is then inserted on both sides of U. Using the result

$$\langle\psi^0_{n\mathrm{k}}|w^0_{nj}\rangle=N^{-1/2}e^{-i\mathrm{k}\cdot\mathrm{R}_j}\delta_{nn'}\,, \tag{9.18}$$

multiplying (9.16) by $e^{i\mathrm{k}\cdot\mathrm{R}_j}$, summing over k and using (9.11), one gets (Koster and Slater, 1954)

$$F_{nj}+N^{-1}\sum_{n'j'}\sum_{j''}\sum_{\mathrm{k}}\frac{e^{i\mathrm{k}\cdot(\mathrm{R}_j-\mathrm{R}_j'')}}{E^0_{n'\mathrm{k}}-E_\nu}\langle w^0_{nj''}|U|w^0_{n'j'}\rangle F_{n'j'}=0\,, \tag{9.19}$$

where we immediately recognize the Green's function in the Wannier representation, Eq. (6.17), so that (9.19) may be written as

$$F_{nj}-\sum_{n'j'}\sum_{j''}G^0_{nnjj''}(E_\nu)U_{nn'j''j'}F_{n'j'}=0 \tag{9.20}$$

in obvious notation. Since the secular matrix in this form depends nonlinearly on E_ν, the solutions are not found by diagonalization, but by seeking the zeros of the determinant $\Delta(E)$ defined by

$$\Delta(E)=\det\|\delta_{nn'}\delta_{jj'}-\sum_{j''}G^0_{nnjj''}(E)U_{nn'j''j'}\|\,. \tag{9.21}$$

The most important feature of this transformation is the size of the matrix. Notice that in the set of equations (9.12) the size of the secular matrix was determined by the number of bands and the number of sites over which the *wave function* ψ_ν extends, as required by the expansion (9.9). Now, in (9.21), the size of the matrix is determined by the number of bands and the number of sites over which the *perturbation potential* U extends, which may be substantially smaller. The simplest model, which Koster and Slater (1954) first solved, is the famous one-band/one-site model for which all $U_{nn'j''j}$ but one are taken to be zero. The matrix in (9.21) is then 1×1 and the problem is trivial. The resultant wave function, however, extends over many lattice sites, so that a much larger effort would be required with (9.12). We will have occasion to discuss this simple model in the context of isoelectronic traps later on in this section. Other applications will be discussed in subsection D below.

2. The Koster-Slater method in other localized representations

Koster and Slater (1954a) suggested that instead of Wannier functions, which may require considerable effort to construct, one could employ any arbitrary complete orthonormal set of localized orbitals $\phi_\alpha(\mathbf{r}-\mathbf{R}_j)$. It would then be natural to make use of the closure relation

$$1=\sum_{\alpha j}|\phi_{\alpha j}\rangle\langle\phi_{\alpha j}|, \tag{9.22}$$

in (9.16), and follow the same steps. Instead, Koster and Slater (1954a) chose to define the Bloch sums

$$\chi_{\alpha\mathbf{k}}(\mathbf{r})=N^{-1/2}\sum_j e^{i\mathbf{k}\cdot\mathbf{R}_j}\,\phi_\alpha(\mathbf{r}-\mathbf{R}_j) \tag{9.23}$$

and expand $\psi_\nu(r)$ in terms of these

$$\psi_\nu(\mathbf{r})=\sum_{\alpha\mathbf{k}}F_{\alpha\mathbf{k}}\chi_{\alpha\mathbf{k}}(\mathbf{r}). \tag{9.24}$$

The corresponding set of secular equations is then (assuming the $\phi_{\alpha j}$ are orthonormal)

$$\sum_{\alpha'\mathbf{k}'}[H^0{}_{\alpha\alpha'}(\mathbf{k})\delta_{\mathbf{k}\mathbf{k}'}+\langle\chi_{\alpha\mathbf{k}}|U|\chi_{\alpha'\mathbf{k}'}\rangle]F_{\alpha'\mathbf{k}'}=E_\nu F_{\alpha\mathbf{k}}, \tag{9.25}$$

where we have used

$$\langle\chi_{\alpha\mathbf{k}}|H^0|\chi_{\alpha'\mathbf{k}'}\rangle=H_{\alpha\alpha'}(\mathbf{k})\delta_{\mathbf{k}\mathbf{k}'}. \tag{9.26}$$

Equation (9.25) can then be transformed to a form similar to (9.16) by defining $A_{\beta\beta'}(\mathbf{k},E_\nu)$ to be the inverse of $H_{\alpha\alpha'}(\mathbf{k})-E_\nu\delta_{\alpha\alpha'}$. The result is

$$F_{\alpha\mathbf{k}}+\sum_\beta\sum_{\alpha'\mathbf{k}'}A_{\alpha\beta}(\mathbf{k},E_\nu)\langle\chi_{\beta\mathbf{k}'}|U|\chi_{\alpha'\mathbf{k}'}\rangle F_{\alpha'\mathbf{k}'}=0. \tag{9.27}$$

It is at this point that Koster and Slater (1954) use (9.22), define $F_{\alpha j}$ by analogy with the F_{nj}, and obtain an equation which is analogous to Eq. (9.19), namely

$$F_{\alpha j}+N^{-1}\sum_{\alpha'j'}\sum_{\beta l}\sum_{\mathbf{k}}e^{i\mathbf{k}\cdot(\mathbf{R}_j-\mathbf{R}_l)}A_{\alpha\beta}(\mathbf{k},E_\nu)\times\langle\phi_{\beta l}|U|\phi_{\alpha'j'}\rangle F_{\alpha'j'}=0. \tag{9.28}$$

The corresponding determinant is

$$\Delta(E)=\det\|\delta_{\alpha\alpha'}\delta_{jj'}-\sum_{\beta l}\sum_k e^{i\mathbf{k}\cdot(\mathbf{R}_j-\mathbf{R}_l)}A_{\alpha\beta}(\mathbf{k},E)U_{\beta l\,\alpha'j'}\|. \tag{9.29}$$

Clearly, however, another form is possible (Bernholc and Pantelides, 1978). As noted earlier, (9.22) may be used directly in (9.16) in place of (9.17), so that (9.28) is replaced by

$$F_{\alpha j}+N^{-1}\sum_{\alpha'j'}\sum_{\beta l}\sum_{n\mathbf{k}}\frac{\langle\phi_{\alpha j}|\psi^0_{n\mathbf{k}}\rangle\langle\psi^0_{n\mathbf{k}}|\phi_{\beta l}\rangle}{E^0_{n\mathbf{k}}-E_\nu}\times\langle\phi_{\beta l}|U|\phi_{\alpha j'}\rangle F_{\alpha'j'}=0, \tag{9.30}$$

and the corresponding determinant is

$$\Delta(E)=\det\|\delta_{\alpha\alpha'}\delta_{jj'}-\sum_{\beta l}\sum_{n\mathbf{k}}\frac{\langle\phi_{\alpha j}|\psi^0_{n\mathbf{k}}\rangle\langle\psi^0_{n\mathbf{k}}|\phi_{\beta l}\rangle}{E-E^0_{n\mathbf{k}}}U_{\beta l\,\alpha'j'}\|. \tag{9.31}$$

The forms (9.29) and (9.31) are equivalent. Note that (9.29) needs the inverse of $H_{\alpha\alpha'}(\mathbf{k})-E$, whereas (9.31) needs to first diagonalize $H_{\alpha\alpha'}(\mathbf{k})$ and obtain the eigenvalues $E_{n\mathbf{k}}$ and eigenvectors $\psi^0_{n\mathbf{k}}$.

3. The Bassani-Iadonisi-Preziosi-Jaros (BIPJ) method

This method is similar in spirit to the Koster–Slater-type methods discussed above. One starts with (9.16), but first splits $U(\mathbf{r})$ into a product of two functions

$$U(\mathbf{r})=U_1(\mathbf{r})U_2(\mathbf{r}). \tag{9.32}$$

The original suggestion of Bassani *et al.* (1969) was $U_1=U_2=U^{1/2}$. The method works, however, for an arbitrary product, giving $U(\mathbf{r})$ as in (9.32) (Jaros and Brand, 1976). One then introduces a complete orthonormal set of functions $g_m(\mathbf{r})$ at the impurity site so that

$$1=\sum_m|g_m\rangle\langle g_m|. \tag{9.33}$$

By inserting (9.33) between U_1 and U_2 and proceeding with steps analogous to the ones employed before, (9.16) becomes

$$a_m+\sum_{m'}\sum_{n\mathbf{k}}\frac{\langle g_m|U_1|\psi^0_{n\mathbf{k}}\rangle\langle\psi^0_{n\mathbf{k}}|U_2|g_{m'}\rangle}{E^0_{n\mathbf{k}}-E_\nu}a_{m'}=0, \tag{9.34}$$

and the corresponding determinant is

$$\Delta(E)=\det\|\delta_{mm'}-\sum_{n\mathbf{k}}\frac{\langle g_m|U_1|\psi^0_{n\mathbf{k}}\rangle\langle\psi_{n\mathbf{k}}|U_2|g_{m'}\rangle}{E-E^0_{n\mathbf{k}}}\|, \tag{9.35}$$

4. Green's-function derivation of the Koster-Slater and BIPJ equations

The form of (9.19), which may be written in terms of the Green's function G^0 as in (9.20), suggests that the same results may be obtained directly in terms of a Green's-function formalism. The basic equations were given in Sec. VI, where we saw that the wave function ψ_ν for a localized state in the gap obeys Eq. (6.27), namely

$$Q(E)\psi_\nu=0, \tag{9.36}$$

where Q is the operator $1-G^0(E)U$. If ψ_ν is expanded as in (9.3) in terms of any complete set of functions φ_λ, (9.36) becomes a set of algebraic equations,

$$\sum_{\lambda'}Q_{\lambda\lambda'}(E)F_\lambda=0. \tag{9.37}$$

These equations have a solution if the determinant

$$\Delta(E)=\det\|Q_{\lambda\lambda'}(E)\| \tag{9.38}$$

becomes equal to zero. Notice that the determinant of an operator is invariant, i.e., it has the same value no matter what basis functions φ_λ are used. One can therefore write

$$\Delta(E)=\det\|1-G^0(E)U\|=0. \tag{9.39}$$

In the Wannier representation $\Delta(E)$ becomes precisely (9.21). In the $\phi_{\alpha j}$ representation, $\Delta(E)$ becomes (Bernholc and Pantelides, 1978)

$$\Delta(E)=\det\|\delta_{\alpha\alpha'}\delta_{jj'}-\sum_{\beta l}G^0_{\alpha\beta jl}(E)U_{\beta l\,\alpha'j'}\|, \tag{9.40}$$

where

$$G^0_{\alpha\beta jl}(E) = \sum_{nk} \frac{\langle \phi_{\alpha j} | \psi^0_{nk} \rangle \langle \psi^0_{nk} | \phi_{\beta l} \rangle}{E - E^0_{nk}} . \tag{9.41a}$$

Note that (9.40) is identical to (9.31). It is also identical to 9.29 with $G^0_{\alpha\beta jl}$ given by the alternative form

$$G^0_{\alpha\beta jl}(E) = \sum_{nk} e^{ik\cdot(R_j - R_l)} A_{\alpha\beta}(k, E) . \tag{9.41b}$$

By formulating the problem directly in a Green's-function formalism, one does obtain some advantages, however. (Callaway, 1964; Callaway and Hughes, 1967; Callaway, 1971). In particular, one now has a formalism which can deal with states in the band continuum, as we saw in Sec. VI.

The BIPJ form of the secular determinant, Eq. (9.35), can be derived from the general Green's-function formalism as well, as shown by Bernholc and Pantelides (1978). One starts with the general results that bound states are given by the zeros of $\det\| 1 - G^0(E)U \|$. Writing U as in (9.32) and multiplying on the left by U_2, we see that bound states also correspond to zeros of $\det \| 1 - U_2 G^0 U_1 \|$. When this last determinant is expressed in the representation defined by the complete set of functions $|g_m\rangle$ and G^0 is expressed as in (6.15), the expression (9.35) follows.

C. General remarks

The methods described above are strictly computational in nature, except when applied to simple models (see, e.g., Koster and Slater, 1954b; Bassani *et al.*, 1969). In general, the problem is twofold: First, given a band structure and an impurity potential one must choose one of the representations discussed above, set up the appropriate matrix, and determine the eigenenergies, either by diagonalizing a secular matrix or by seeking the zeros of the corresponding determinant. For accuracy, one must demonstrate that *convergence* has been achieved in the number of basis functions used, the number of bands, the number of k points in the Brillouin zone or whatever else may apply. The second aspect of the problem is the construction of impurity potentials. Ideally *self-consistency* should be achieved, which means that V should be constructed from the charge density arising from all the ψ_v, U should be determined from (9.1), and the cycle should be repeated until the resultant ψ_v's are the same as those used to generate V. In case self-consistency is not attempted, the impurity potential may be constructed in a fashion similar to that described in Sec. VIII, but *internal consistency* needs to be maintained. This means that one may work in a "true", pseudo, or model representation of the Hamiltonian, but both impurity potential and Bloch functions must be in the same representation. In other words, if one uses the "true" impurity potential, one must use the "true" Bloch functions; if one uses a pseudopotential for the impurity potential, one must use the pseudo-Bloch functions that correspond to that particular choice of pseudopotential. Otherwise, the equations derived above are not valid. This requirement appears to be self-evident, but as we shall see, it has often been violated in applications to real systems without exploring its consequences.

D. Applications

The methods described above have been used in quantitative calculations for a variety of systems. We will proceed by examining applications of one method at a time, more or less in chronological order, except that we collect all applications to isovalent impurities in a separate subsection. The purpose here is to review the merits of each specific application. As stated earlier, we defer a review of the relative merits of methods to Sec. XI.

1. Determinantal method in the Wannier representation

Calculations using the determinantal method in the Wannier representation were reported for the neutral unrelaxed vacancy and the neutral unrelaxed divacancy in Si by Callaway and Hughes (1967a, b) and by Callaway (1971). Calculations for the self-interstitial in Si were reported later by Singhal (1971, 1972). Applications of the same method for isovalent impurities have been carried out by Faulkner (1968) and by Baldereschi and Hopfield (1972). The latter will be discussed under a separate heading later on.

Callaway and Hughes (1967) and Singhal (1971, 1972) used the same basic approximations. The band structure was calculated in the empirical pseudopotential scheme (see, e.g., Cohen and Heine, 1970) using the empirical form factors of Brust (1964), but only a limited number of plane waves. The defect potential was constructed as follows: First, an empirical pseudopotential for a Si atom was constructed by interpolating the form factors $V^0(\mathrm{K})$ that enter the band structure and obtaining $v^{\smile}(\mathbf{k})$ in all of $\mathbf{k}$ space. The vacancy potential was then taken to be the negative of an atomic Si pseudopotential, the divacancy potential was taken to be the negative of two such atomic pseudopotentials located at neighboring sites, and the self-interstitial potential was taken to be an atomic pseudopotential located at the chosen interstitial site. Thus constructed, the defect pseudopotential is of course not self-consistent (i.e., the redistribution of the electronic charge caused by the introduction of the defect is neglected) and ignores lattice relaxation (i.e., the movement of the nuclei in the vicinity of the defect). On the other hand, the defect potential *is internally consistent* in the sense described above.

Given the defect potential and the band structure, the determinant (9.21) had to be set up and its zeros had to be found. This process is extremely laborious. A tedious symmetry analysis had to be carried out to make the task feasible. The construction of matrix elements of the defect potential was particularly time consuming, since it involved double sums over the Brillouin zones. These difficulties took their toll by limiting the computation to include only a small number of bands and sites.

For the vacancy in Si, Callaway and Hughes (1967a) found no bound state in the gap, even with the maximum number of bands and sites they could include. For this reason, they introduced a scaling factor λ for the defect potential and presented results for various values of λ. They also presented results for a variety of combinations of bands and sites. We wish to review these results by addressing two distinct ques-

TABLE VI. Results of convergence of the Callaway–Hughes calculation for the vacancy in Si with $\lambda = 1.20$. Energies are measured from the top of the valence bands in eV. NBS stands for No Bound State.

Number of bands	2-site bound state	3-site bound state
3	0.86	— —
4	0.43	0.36
5	NBS	NBS
5	0.06	NBS
5	0.32	0.26
6	0.12	0.07

tions. The first is one of convergence: For a given potential, i.e., for a given value of λ, how well does the calculation converge? We choose $\lambda = 1.20$ as a typical value, for which many combinations of bands and sites were used. Collecting information from various tables given by Callaway and Hughes (1976), we supply the results in Table VI. The Table shows that the results are not convergent. Clearly, adding the sixth band in both cases makes a dramatic change. Similarly, adding the extra site produces a change of 40% in the bound-state energy. Similar results were obtained for other λ values.

The second question we wish to address is the overall achievement of the work by Callaway-Hughes-Singhal. As far as the determination of bound states associated with the defects, no conclusive answer was obtained. Callaway (1971) extracted a value of the formation energy from a study of the phase shifts in the valence bands, which yields approximate changes of total energies, but the accuracy of the results is also uncertain due to the lack of convergence. Similarly, Singhal calculated the total energy for the interstitial at various locations and concluded that the bond-centered site is likely to be preferred, in agreement with Watkins' suggestion (see Watkins, Messmer, Weigel, Peak, and Corbett, 1971), but the conclusion was not firm. Nonetheless, the calculations by Callaway and co-workers represent the first major step in the application of Koster–Slater methods to real systems. The troublesome and costly aspect of these calculations was the construction of the Wannier functions. In the years following the original work no further applications have been reported, indicating lack of promise in the method. As we saw, however, in Sec. IX. B. 2, the Koster-Slater method does not require the use of Wannier functions. The use of LCAO basis sets appears to be a more promising technique (see below).

2. Determinantal method in LCAO representations

The last idea, namely the use of some other localized basis set instead of Wannier functions, was first exploited by Lannoo and Lenglart (1969), who used a set of s and p atomic orbitals on each atom, performed an LCAO energy-band calculation by treating the Hamiltonian matrix elements (only first-neighbor interactions) as parameters to be fit to known energy bands, and then looked at the determinant (9.40), taking $\phi_\alpha(\mathbf{r} - \mathbf{R}_j)$ to be the same set of s and p orbitals. The ideal vacancy is then approximated by removing an atom and leaving everything else unchanged. In the determinantal method this can be accomplished by either removing all the interactions of the central atom with its neighbors of various orders, or, equivalently, setting the diagonal Hamiltonian matrix element of the central atom at a very large energy, so that all the other atoms do not "see" it. (See also Bernholc and Pantelides, 1978). Using symmetry, the net result is that states of T_2 symmetry are the zeros of $G^0_{pp00}(E)$ and states of A_1 symmetry are the zeros of $G^0_{ss00}(E)$. The results of the calculation for diamond are qualitatively interesting but quantitatively unreliable, due to the crudeness of the band structure employed. The calculation yields two bound states in the gap which are nearly degenerate in energy: one of A_1 symmetry and one of T_2 symmetry with a total of four electrons available for them. As we will see in the next section, more realistic calculations obtain the A_1 state within the valence bands as a resonance and the T_2 in the gap with only two electrons in it. The wave function obtained by Lannoo and Lenglart is approximately 70% localized on the first shell of neighbors, which compares well with the localization extracted by Watkins (1972) from EPR data of the vacancy in Si.

More recently, calculations similar to those of Lannoo and Lenglart have been performed for the vacancy in Si, Ge, and GaAs by Bernholc and Pantelides (1978). The band-structure parametrization was that of Pandey and Phillips (1976), which reproduces empirical-pseudopotential valence bands very accurately and does reasonably well for the lowest conduction band, but not so satisfactorily for higher conduction bands. This parametrization (first- and second-neighbor interactions) was found to be very good for surface states. The qualitative results for the single vacancy are again as found by others using different methods (see next section), namely a T_2 level in the gap and resonances and antiresonances within the bands. Quantitatively, the results can only be compared with other model calculations and we postpone this until other calculations are described. One particular qualitative result, however, is especially interesting (Bernholc and Pantelides, 1978). As we saw above, the T_2 level in the gap is determined by the zeros of $G^0_{pp00}(E)$. With slight manipulation, this function may be written as

$$G^0_{pp00}(E) = \int dE' \, D_p(E')/(E - E') , \qquad (9.42)$$

where D_p is the partial density of states of p symmetry at the central atom. The integral is over the entire energy axis. For levels in the gap, it is clear from (9.42) that the valence bands contribute to $G^0_{pp00}(E)$ with a positive sign, while the conduction bands contribute with a negative sign. The zero of G^0_{pp00} is therefore obtained where the two contributions cancel exactly. This observation implies that both valence and conduction bands play an equally important role in determining the position of the bound state. This conclusion, of course, pertains only to this particular model. A many-site calculation may in fact result in a weaker role for the conduction bands, as found by Callaway and Hughes (1967).

3. BIPJ method

Another series of applications using the methods described earlier in this section has been carried out by Jaros and co-workers. Initially, Jaros (1973), Jaros and Ross (1973), and Ross and Jaros (1973, 1974) attempted to set up and diagonalize the secular matrix in the Bloch representation, i.e., Eq. (9.7). Symmetry was used to great benefit, but the number of k points that could be included was still too small to ensure convergent results. The authors carried out calculations in a number of systems, the earliest one being acceptors in Si (Jaros and Ross,1973), and later on GaP:O_P, GaP:N_P, and ZnTe:O_P. Soon, however, the technique was abandoned for its inability to produce convergent results and the Bassani–Iadonisi–Preziosi (1969) method was adopted, with the functions g_m taken to be products of associated Laguerre polynomials and spherical harmonics. Applications have been carried out for GaP:O_P (Jaros, 1975), the vacancy in GaP, GaAs, and InSb, and the divacancy and vacancy-oxygen complex in GaAs (Jaros and Brand, 1976).

In all their calculations, Jaros and co-workers employed empirical-pseudopotential band structures similar to the one used by Callaway and Hughes (1967a, b). The more recent calculations, which make use of the Bassani-Iadonisi-Preziosi method, are convergent with respect to the number of k points in the Brillouin zone and the number of bands. On the other hand, no tests have been carried out for convergence of the angular part of the intermediate functions g_m.

We now turn to examine the choices of impurity potentials and review the results. The method of constructing the impurity or defect potential has not been discussed in depth in the published papers. In particular, the question of internal consistency, as described above, has not been addressed, and from what we can infer, the impurity/defect pseudopotentials used in the calculations thus far appear to violate this requirement. We are unable to assess the consequences of this choice of potentials. More specifically, the oxygen potential, which has been used in a series of papers, including the most recent work, was obtained in 1973 by extrapolating unpublished tables of model potentials by Animalu (1956) using a value for the effective Fermi level determined by Jones and Lettington in unpublished work by fitting experimental data on GaN. Jaros and Ross (1973) were aware of the uncertainties involved and stated their "hope that it is a good estimate." This situation is rather discomforting, especially since oxygen, like other first-row elements lacking *p* core states, is known to be a troublesome case for pseudopotentials (see, e.g., Cohen and Heine, 1970). In a recent study of oxygen, Chelikowsky and Schlüter (1977) were able to construct an accurate pseudopotential, but nonlocality was found to be essential. This latter potential may provide a better basis for calculations of oxygen impurities in the future.

Finally we turn to the actual results and accomplishments of Jaros and co-workers. One series of calculations was on GaP:O_P. The latest calculation on this system gave a binding energy for the extra electron at the oxygen site of 1 eV, in very good agreement with the experimental value of 0.9 eV. This result may be viewed as indication that the oxygen pseudopotential used in the calculation is adequately accurate. On the other hand, it may be argued that the agreement with experiment may be masking other issues that may be important for such a pathological system as GaP:O_P. (The system is pathological because the other group VI elements, i.e., S, Se and Te are shallow donors in GaP with binding energies of order 100 meV. Oxygen, on the other hand, has a binding energy nine times larger and is capable of binding a second electron; see below.) One question is whether the eigenvalue may be identified as the ionization energy in the sense of Koopmans' theorem. For example, GaP:O_P is very likely to introduce a bound state below the bottom of the valence band corresponding to the O 2*s* state) just like Si:S has been found to do (see Sec. X.C.2). When the bound electron is removed to the conduction bands, the rearrangement of charge may cause this level to move and produce a correction to the ionization energy. This particular question of electronic relaxation is perhaps more relevant for the two-electron state of GaP:O_P for which a dramatic lattice relaxation has been suggested to accompany binding (Henry and Lang, 1977). Jaros has included lattice relaxation by moving the nearest neighbors a certain distance and calculating the additional perturbation potential. He was thus able to reproduce the experimentally observed optical cross section with a threshold at 1.4 eV. These results were not presented as conclusive, however. In particular, the threshold is actually fitted, in accordance with a particular interpretation of the data (Henry and Lang, 1977), and therefore it cannot rule out the alternative interpretation of the data by Grimmeiss, Ledebo, Ovren, and Morgan (1974) and by Morgan (1975). Furthermore, as experiments with higher resolution become available, the accepted picture may change dramatically. For example, recent data by Samuelson and Monemar (1977) suggest an ionization threshold for the second electron of about 1 eV. Moreover, Morgan (1978) is now proposing an alternative interpretation of old data on this system. No attempt will be made to resolve the "GaP:O_P controversy" here. It appears that as more data are becoming available the situation will become clearer. Theoretical calculations, such as those of Jaros, definitely contribute positively to the process by establishing quantitative guidelines. It is hoped that more accurate impurity pseudopotentials will be employed in the future so that firmer quantitative predictions can help the interpretation of the data more directly.

The more recent work of Jaros (Jaros, 1975; Jaros and Brand, 1976) has been on the vacancy, divacancy, and vacancy-oxygen complex in a number of III-V compounds. These pioneering calculations are the first quantitative work on such systems. Jaros and Brand (1976) gave a very cautious assessment of the assumptions they made, pointing out the main limitations, such as lack of self-consistency in both the electron distribution and the lattice relaxation. They presented calculations of the bound states by scaling the defect potential in the manner of Callaway and Hughes (1967) and demonstrated that for a small number of bands (nonconvergent solutions) the bound-state energies are very sensitive to the scaling factor λ, as also found by Callaway and Hughes. For fully convergent solutions, however, the dependence on scaling was found to be quite small, suggesting that self-consistency may not be very important. Jaros and Brand were therefore able to conclude that the Ga vacancy in GaAs introduces a triplet T_2 level in the gap near the top of the valence bands. The As vacancy, on the other hand, introduces a triplet T_2 level which is resonant with conduction states somewhat above the conduction band edge. (The uncertainty in this level is larger because the formalism is valid only in the regions outside the bands.) The V_{As} level is, however, much simpler to interpret: Since in the neutral state of the vacancy the level would contain one electron, the center would be expected to behave as a donor. The V_{Ga} T_2 level, however, would contain three electrons in the neutral state, so that electron–electron interactions are likely to change its position in the gap. Jaros and Brand suggest that the calculated level may correspond to a V_{Ga}^{++} state, but the assignment would be inconsistent with a neutral potential, unless one argues that the Coulomb tail would not affect the calculation substantially. Jaros and Brand then calculated the energy levels of the V_{Ga}^{+} system treating the two bound electrons in the Hartree approximation. For the one-electron V_{As}^{++} system, they also performed the calculation by including a trigonal Jahn–Teller distortion, which splits the triplet into a doublet that goes up in energy and a singlet (singly occupied) that goes down in energy toward the valence bands. Similar results were obtained for anion and cation vacancies, respectively, in other III–V compounds. The tentative final conclusion was that cation vacancies are likely to be single acceptors and the anion vacancies are likely to be single donors, in agreement with intuitive predictions (Bube, 1960). Similar calculations were performed for the $V_{\text{Ga}}-\text{O}$ complex where it is found that the extra electron supplied by O fills up the hole in the valence bands created by the Ga vacancy and the complex in its neutral state does not behave like an acceptor. Calculations on the divacancy ($V_{\text{Ga}}-V_{\text{As}}$) led only to very uncertain and rather speculative results.

Overall, Jaros and co-workers have made a very valuable contribution to the field of deep-level impurities and defects by demonstrating that calculations of this magnitude can be practical. They have persevered in their intent to carry the calculations to convergence and have contributed valuable guidelines for the interpretation of experiments.

E. Applications to isovalent impurities

Isovalent impurities (also called isoelectronic impurities) were first identified independently by Aten, Haanstra, and deVries (1965) and by Thomas, Hopfield, and Frosch (1965). At first, isovalent impurities were observed to bind excitons (electron-hole pairs), but Thomas, Hopfield, and Lynch (1966) proposed that binding may be viewed as a two-step process: The neutral impurity first binds one of the two particles (either the electron or the hole) and becomes charged. The resultant Coulomb field then binds the second particle in an effective-mass fashion. The electronegativity of the impurity atom was thought to be the factor determining whether an electron or a hole or no particle at all could be initially trapped by the neutral impurity. This picture was in agreement with observations: GaP:N$_{\text{P}}$ was observed to bind an electron with about 10 meV, whereas GaP:Bi$_{\text{P}}$ was observed to bind a hole with about 38 meV (Dean, Cuthbert, and Lynch, 1969). There is a vast literature on experimental work on isoelectronic traps, especially on nitrogen and pairs of nitrogens in a variety of phosphide and arsenide alloys. Reviewing this literature is beyond the scope of this article. We will instead focus only on bona fide attempts to develop a theoretical description of binding by isoelectronic traps.

The first theoretical model proposed to describe binding by isovalent impurities was the Koster–Slater one-band/one-site model mentioned earlier (Faulkner, 1968). By taking the potential matrix elements to be of the form

$$\langle w_{nj}^{0}|U|w_{n'j'}^{0}\rangle = U_0\delta_{nn'}\delta_{n0}\delta_{jj'}\delta_{j0} \tag{9.43}$$

Eq. (9.21) becomes

$$1+U_0\sum_{\mathbf{k}}\frac{1}{E_{\mathbf{k}}^{0}-E}=0\,, \tag{9.44}$$

which can be solved numerically for a given band structure and a given U_0 to yield the bound-state energy. The dependence of binding on U_0 becomes clearer if one uses the identity

$$\frac{1}{\epsilon-E}=\frac{1}{\epsilon}+\frac{1}{\epsilon}\frac{E}{\epsilon-E}\,, \tag{9.45}$$

whereby (9.44) becomes

$$1+\frac{U_0}{\langle E\rangle}=-U_0\sum_{\mathbf{k}}\frac{E}{E_{\mathbf{k}}^{0}}[E_{\mathbf{k}}^{0}-E]^{-1}\,, \tag{9.46}$$

where

$$\frac{1}{\langle E\rangle}=\sum_{\mathbf{k}}\frac{1}{E_{\mathbf{k}}^{0}}\,. \tag{9.47}$$

Equation (9.49) reveals that for a potential that is attractive to electrons ($U_0<0$), a bound state *below* the band edge exists if

$$|U_0|>\langle E\rangle\,. \tag{9.48}$$

An identical condition holds for potentials that are attractive to holes ($U_0>0$) for bound states *above* the band edge.

Faulkner proceeded to evaluate $\langle E\rangle$ using a simple effective-mass expansion for the conduction of GaP and then treated U_0 as a parameter. The model was found to be inadequate for the description of binding energies: if U_0 was fit to give the binding energy for one electron bound to a single nitrogen, the resulting binding energy of an electron bound to a pair of nitrogens did not agree with experiment. Furthermore, the model was incapable of producing any excited states. The usefulness of the model was in providing a qualitative understanding of the absorption processes that go on near the interband absorption threshold of GaP when N is present (see Faulkner, 1968).

Faulkner (1968) proceeded to carry out a quantitative calculation of bound states by using the Koster–Slater determinantal method and going beyond the one-band/one-site approximation. The calculation was similar in spirit to that of Callaway and Hughes (1967) for the vacancy in Si described earlier. Faulkner performed his calculations by retaining only two bands. No convergence studies were carried out. The impurity pseudopotential used by Faulkner was the difference between atomic pseudopotentials similar to those described in Sec. VIII, Eq. (8.37). Since the energy-band structure for the host crystal was obtained from an empirical-pseudopotential calculation, Faulkner's impurity potential does not satisfy the requirement of internal consistency described earlier. The numerical calculations resulted in a bound state about 1 eV below the conduction band edge, two orders of magnitude larger than the observed value of 0.01 eV. A scaling factor λ was then introduced, in the manner of Callaway and Hughes (1967) and the bound state recalculated for various values of λ. The binding energy was found to be extremely sensitive to the value of λ ($\lambda = 0.501$ gave 0.008 eV, the observed value, whereas $\lambda = 0.504$ gave 0.013 eV). Using $\lambda = 0.501$, which reproduces the experimental value for the single nitrogen, calculations for a pair of nitrogens were carried out. The results did not agree in detail with experiment, but the range of energies and the average spacings are of the correct order of magnitude.

More recently, the problem of binding by isovalent impurities was investigated by Baldereschi and Hopfield (1972) (see also Baldereschi, 1973). These authors assumed that the Koster–Slater one-band/one-site model can provide reliable information on binding of holes if generalized to a three-band/one-site model for the top three valence bands of tetrahedral semiconductors. Instead of Wannier functions for each of the three bands, Baldereschi and Hopfield defined three linear combinations which transform like x, y, and z. This transformation enabled them to obtain a criterion for binding which is formally identical with (9.49). Now, however, U_0 is the matrix element of U with any of three "symmetrized" Wannier functions, and $\langle F\rangle$ is the threefold degenerate eigenvalue of the matrix $M_{\alpha\alpha'}$ given by

$$M_{\alpha\alpha'} = -\sum_{\mathrm{k}} H^{-1}_{\alpha\alpha'}(\mathrm{k})\,, \qquad (9.49)$$

where $H_{\alpha\alpha'}(\mathrm{k})$ is the Hamiltonian matrix in the new "symmetrized" Wannier representation. (These equations assume the zero of energy at the top of the valence bands.)

Baldereschi and Hopfield focused their attention on constructing impurity potentials U for isovalent impurities, calculating the matrix element U_0 and comparing it with $\langle E\rangle$ to determine whether a bound state exists or not. The impurity potentials were calculated in the Heine–Abarenkov model-potential representation as described in Sec. VIII. These potentials were then screened by the Penn (1962) form of the dielectric function $\epsilon(q)$. Local screening effects, which are beyond linear response theory, were included by scaling the Fermi momentum, Fermi energy, and plasma frequency that enter the Penn formula according to the local density in the impurity cell. An additional term was included in the impurity potential arising from lattice relaxation around the impurity. This latter contribution was calculated in terms of a simple model making use of bond lengths observed in compounds of the impurity atom (which corresponds to "maximum" relaxation) and subtracting corresponding crystal potentials. The actual lattice relaxation was estimated by a simple spring model and the potential for this particular configuration was obtained by interpolating between the zero-relaxation and maximum-relaxation potentials. Finally, a spin-orbit correction was added to the impurity potential, as suggested by Allen (1971).

The calculations showed that screening reduces the potential matrix element U_0 substantially, thus confirming a speculation by Faulkner (1968) on the origins of the very large binding energy he obtained with an unscreened pseudopotential. (Baldereschi and Hopfield did not study GaP:$\mathrm{N_P}$, however; they limited their study to states near the valence bands.) On the other hand, lattice relaxation was found to increase U_0 and hence binding. In the end, Baldereschi and Hopfield were able to predict which isovalent impurities can bind a hole and which cannot. An attempt to calculate actual binding energies was not successful, however. The resultant binding energies were more than an order of magnitude larger than the observed values. Such a result suggests that the prediction that a given impurity *does* bind a hole is very safe despite the approximations, the most serious of which is the use of a one-site perturbation in the Koster–Slater equations. Perhaps the most important result of the work is the quantitative demonstration of the fact that screening, local screening, and lattice relaxation play a very important role in the determination of binding by isovalent impurities.

The one-band/one-site Koster–Slater model has been used extensively to analyze and interpret data on isovalent N in $\mathrm{GaAs}_x\mathrm{P}_{1-x}$ alloys by Holonyak and co-workers (e.g., Scifres *et al.*, 1971, 1972). Newer data on these systems have been recently interpreted independently by Kleinman (1977) and by Hsu, Dow, Wolford, and Streetman (1977). In both cases, the one-band/one-site model was found inadequate and longer-range potentials were assumed, with the potential matrix elements fitted to experimental data. The model of Wolford *et al.* corresponds to a one-band/two-shell Koster-Slater model, whereas Kleinman attributed his long-range potential to strain fields extending uniformly over a range of 25 Å.

Finally, studies of isovalent impurities have been carried out by Jaros and co-workers, using the techniques which we discussed earlier in this section (Jaros and Ross, 1973; Ross and Jaros, 1974; Jaros and Ross, 1974). These studies were carried out in the original Bloch representation which was not brought to convergence (Jaros, 1974). The quantitative results are therefore unreliable. The conclusion that the binding energies are very sensitive to the strength of the potential agrees with the work of Faulkner (1968) and Baldereschi and Hopfield (1972). Recalling Jaros's later conclusion (Jaros and Brandt, 1976) that only when full convergence with respect to the number of bands is achieved does the sensitivity go away, we may conclude that no numerically convergent calculations for isovalent impurities have thus far been reported.

X. NONPERTURBATIVE METHODS (CLUSTER METHODS)

In this section we will discuss methods which do not separate V into V^0+U as in Eq. (9.1) but instead work directly with V. In other words, we will be discussing methods which attempt to solve directly the eigenvalue problem

$$H\psi_\nu = [-\hbar^2\nabla^2/2m_0 + V]\psi_\nu = E_\nu\psi_\nu . \qquad (10.1)$$

All these methods have one common feature: they treat the crystal as a finite *cluster* of atoms. They differ in whether they use small clusters or very large clusters. In the case of small clusters, they are additionally distinguished by the type of boundary conditions which are imposed at the surfaces. Cluster calculations have been carried out using a variety of one-electron Hamiltonians and the merits of the techniques seem to depend strongly on the choice of Hamiltonian and the method of solution of (10.1), once the cluster has been defined. We have therefore organized the discussion in this section according to the type of Hamiltonian that was employed in the various calculations. We will first discuss the "defect molecule" model which is a rather unique case. We will then discuss the various methods that have been used to deal with (10.1) when the Hamiltonian is defined in terms of its matrix elements, which in turn are defined in terms of empirical parameters, and, finally, we turn to a discussion of methods in which the potential V is constructed from first principles and (10.1) is solved self-consistently. As we did in Sec. IX, we defer a critical comparison of the various methods to Sec. XI.

A. Defect-molecule model

This approach is interesting for historical perspective rather than for its usefulness today. The method was first introduced for the vacancy in diamond by Coulson and Kearsley (1957), and a similar calculation for the vacancy and self-interstitial in diamond was reported later by Yamaguchi (1962, 1963). The model was extended later by Coulson and Larkins (1969, 1971) for the divacancy and by Friedel, Lannoo, and Leman, who used it to study Jahn-Teller relaxations around the vacancy in diamond.

The central assumption of the model is that the wave function of the vacancy electrons may be constructed from the four "dangling" sp^3 hybrid orbitals on the four neighboring carbon atoms, and that this wave function does not interact with the band states in the crystal. A related assumption is that the potential which the vacancy electrons see is just what arises from the four nearest-neighbor carbon atoms. One then constructs this potential from atomic potentials and, using appropriate linear combinations of atomic wave functions, one carries out a "many-body" configuration-interaction calculation for the vacancy electrons. Observed absorption can then be attributed to transitions between these localized states. The main limitation of the model, in addition to the rather drastic approximations, is that it cannot position the localized levels with respect to the crystal band edges.

From a conceptual point of view, it is perhaps important to distinguish the defect-molecule model from conventional calculations on clusters of four atoms (cluster calculations will be discussed at length later on in this section). Note that though the potential of a total of four atoms is taken into account, the wave function of the bound states is expanded only in terms of the four dangling hybrids, *not all sixteen hybrids* that exist on the four atoms. In other words only the bound states for the vacancy electrons are sought, while all other states are viewed to be unperturbed. In this way, the question of what to do with the surface bonds does not arise. The price one pays is that the bound states cannot be related to the energy band edges.

In subsequent work on the vacancy in diamond, some of which was mentioned in Sec. IX and some of which will be reviewed below (see especially Watkins and Messmer, 1974) it became clear that the defect-molecule assumption of a wave function which is completely localized on the nearest neighbors is not justified, and also causes the many-electron multiplet splitting to be too large. The results of this model will therefore not be discussed in further detail. We have mentioned it here because for more than ten years it served as the only guide to understanding a vast amount of data on vacancies and other radiation-induced defects.

B. Semiempirical LCAO methods

In semiempirical methods one does not calculate the potential V explicitly. Instead, a basis set for ψ_ν is chosen and the matrix elements of H are determined empirically, i.e., either evaluated using a well tested prescription or directly fit to experiment. For the defect problem, the most suitable basis set is a set of atomic-like orbitals. Let us therefore define a general problem for an arbitrary collection of atoms described by a Hamiltonian h. We seek solutions of the eigenvalue problem

$$h\psi_\nu = \epsilon_\nu\psi_\nu . \qquad (10.2)$$

We then introduce a set of atomic orbitals ϕ_μ, where μ runs over all atomic orbitals on each atom and over all atoms, and expand

$$\psi_\nu = \sum_\mu C_{\mu\nu}\,\phi_\mu . \qquad (10.3)$$

The eigenvalues ϵ_ν are then solutions of the secular equations

$$\sum_\mu (h_{\mu\lambda} - \epsilon_\nu S_{\mu\lambda})C_{\mu\nu} = 0 , \qquad (10.4)$$

where

$$h_{\mu\lambda} = \langle\phi_\mu|h|\phi_\lambda\rangle , \qquad (10.5)$$

and

$$S_{\mu\lambda} = \langle\phi_\mu|\phi_\lambda\rangle . \qquad (10.6)$$

The set of equations (10.4) is just a generalized eigenvalue problem and can be solved by standard techniques once $h_{\mu\lambda}$ and $S_{\mu\lambda}$ are known. Note that the size of the matrix is equal to the number of orbitals ϕ_μ so that for a direct solution of (10.4) the number of orbitals is limited by present computer capacity and accuracy to about 300. If only s and p (p_x, p_y, p_z) orbitals are imposed, the limit of cluster size is about 75 atoms, unless symmetry is taken advantage of.

1. Extended Hückel theory (EHT)

Extended Hückel theory (EHT) is a method that has its roots in organic chemistry (see, e.g., Hoffmann, 1963). It can be defined for an arbitrary collection of atoms by the equations given above and the following set of rules:

(i) Restrict the set of atomic orbitals ϕ_μ to the valence orbitals of each atom and choose them to be Slater-type atomic orbitals, i.e., the appropriate spherical harmonic multiplied by a radial function of the form

$$f(r) = Nr^{n-1}e^{-\zeta r}. \tag{10.7}$$

Here n is the principal quantum number of the orbital, ζ is an exponent, chosen, for example, according to Slater's rules (Slater, 1930), and N is a normalization constant. Having chosen the ϕ_μ, the overlap matrix $S_{\mu\lambda}$ is then evaluated directly.

(ii) Evaluate $h_{\mu\lambda}$ in terms of the prescription

$$h_{\mu\lambda} = -\tfrac{1}{2}K_{\mu\lambda}(I_\mu + I_\lambda)S_{\mu\lambda}, \tag{10.8}$$

where

$$K_{\mu\lambda} = \begin{cases} K \text{ for } \mu \neq \lambda \\ 1 \text{ for } \mu = \lambda \end{cases}. \tag{10.9}$$

Here K is a constant, $1 < K < 2$, usually taken to be 1.75 in applications to organic molecules (Hoffmann, 1963). I_μ is the μth orbital ionization potential from experimental data (see, e.g., Pople and Segal, 1965).

A separate assumption, but one that is usually made along with the assumptions stated above, is that the sum of eigenvalues

$$E_{\text{tot}} = \sum_\nu n_\nu \epsilon_\nu, \tag{10.10}$$

where n_ν is the degeneracy of the νth eigenvalue, is a good measure of the total energy of the system. Note that the sum in (10.10) is only over the occupied states.

a. The validitv of the EHT

The central assumption of the EHT is that the Hamiltonian matrix elements $h_{\mu\lambda}$ may be approximated by (10.8). The approximation was originally suggested by Mulliken (1949) as plausible, and a number of workers subsequently attempted to derive conditions under which the matrix elements of the one-electron Hartree-Fock operator can in fact be approximated by an expression of the form (10.8) (Boer, Newton, and Lipscomb, 1964; Newton, Boer, and Lipscomb, 1966; Blyholder and Coulson, 1968; Gilbert, 1970). All these workers showed that under no circumstances can the rigorous Hartree-Fock matrix elements be reduced to a form that is proportional to the overlap matrix elements. (The most serious problem is the kinetic energy.) The only "justification" of EHT is that it *works* more often than it does not, and detailed studies have shown that its success is due to fortuitous cancellations of errors (see, e.g., Boer *et al.*, 1964).

Expanded Huckel Theory (EHT), as a prescription for $h_{\mu\lambda}$, has another rather unsettling property, pointed out by Coulson (1972): The prescription is not invariant with respect to the choice of origin for the energy unless $K = 1$, a choice which has not been found to be useful. The usual choice of K is 1.5 or 1.75.

Finally, the prescription for the total energy, Eq. (10.10), can in fact be viewed as an approximation, but most studies (see, e.g., references quoted above) show that the other terms in the correct expression for the total energy are not necessarily negligible. Including the additional terms, however, would be entirely outside the spirit of EHT, and they have always been left out.

The above discussion suggests that applying the EHT to solids would be a dangerous path. Nevertheless, the method has provided very valuable information for an assortment of molecular problems in the absence of practical schemes for accurate calculations. From this point of view, the use of the EHT for studies of defects in solids is totally justified, as long as one proceeds with caution and carries out intermediate tests. As we shall see, the EHT has in fact provided very useful information for defects in covalent solids.

b. The limitations of EHT

Turning our attention to solids, it is worthwhile to identify at the outset the quantities which can and cannot be calculated within the EHT. For the purposes of this discussion let us assume that the eigenvalue problem (10.4) can be solved for either a perfect solid or a solid containing a defect. Clearly one then gets the one-electron energies ϵ_ν and the corresponding wave functions ψ_ν. By filling these levels according to the Pauli exclusion principle, one can then determine the Fermi level and the total energy as defined by (10.10). The calculation of the total energy can then be repeated for various atomic configurations and one can obtain elastic constants and lattice relaxation around a defect, trace the path for the diffusion of atoms or vacancies, and determine preferred positions for interstitial impurities and self-interstitials. On the negative side, EHT can study only a neutral center as it contains no prescription for altering the $h_{\mu\lambda}$ when electrons or holes are added or removed from a center. For this same reason, EHT cannot provide excitation energies between localized states or between a localized state and a band state, except when the excitation corresponds to removing an electron from a singly occupied localized state to an empty state or putting one electron in an otherwise empty bound state. We will see an example of this later on.

Another limitation of the EHT is that it is not a well defined prescription in the case of compounds. The atoms in such crystals are not neutral and the effective charge on the ions is an ill defined quantity making the choice of the I_α in (10.4) somewhat ambiguous, especially because of the Madelung energies that alter the I_α from their free-ion values. Nonetheless, one could supplant additional prescriptions to take care of these difficulties. Work on defects has, however, thus far been mainly on diamond and Si and we will therefore not address the question of ionic solids.

c. The EHT for solid-state systems

The EHT prescriptions, as taken from molecular theory and described above, can be used directly to describe perfect periodic crystals. For this purpose, one makes use of Bloch's theorem according to which the coefficients $C_{\mu\nu}$ in (10.3) may be written as

$$C_{\mu\nu}=C_{\alpha\nu}e^{i\mathbf{k}\cdot\mathbf{R}_j}, \tag{10.11}$$

where μ, which in (10.3) is a composite index running over atoms and orbitals, has been broken into an index j, labeling lattice sites, and an index α, labeling orbitals in the primitive unit cell. The index ν is now a composite band index and wave vector $\mathbf{k}$. The size of the resultant secular matrix at each wave vector $\mathbf{k}$ becomes equal to the number of orbitals in the unit cell. For diamond-type crystals inclusion of only the s and p valence orbitals on the two atoms per unit cell yields 8×8 secular matrix at each $\mathbf{k}$ point, which in turn yields 8 bands.

A calculation of the energy bands of diamond using an $s-p$ basis set and the EHT prescription for matrix elements has been performed by Messmer (1971), who found that the valence bands are reproduced very well, as compared with more sophisticated calculations, but the conduction bands are less satisfactory. A dramatic improvement of the conduction bands was accomplished, however, by a slight modification of K in (10.9) and one of the Slater exponents. The modified EHT parameters were also found to give elastic constants which agree very well with experimental values (Watkins and Messmer, 1973). Somewhat similar results were obtained by Lee and McGill (1973) for Si, who actually fit the EHT parameters to available energy bands: Good valence bands were obtained, but the conduction bands were unsatisfactory. One might think that the unsatisfactory conduction bands for Si may be due to the absence of d orbitals in the basis set, but the inadequacy can actually be traced to the EHT prescription (10.8), since recent work by Chadi (1977) has shown that excellent valence *and* conduction bands can be obtained for Si and Ge with an $s-p$ basis set, similar to that used by EHT, but using the empirical pseudopotential of Cohen and Bergstresser (1966) in terms of which to calculate the Hamiltonian matrix elements.

The above results for perfect crystals show that the EHT (or modified EHT) is likely to produce meaningful results in imperfect crystals as well, at least in diamond. The solution of the EHT secular equations for crystals containing defects is, however, no easy task. Periodicity is broken and Bloch's theorem does not hold any more, so that one is faced with a very large secular matrix. In all the applications of the EHT to the problem, therefore, *additional* approximations have been made, which, to a certain extent, mask the absolute success or failure of the EHT.

Three different approximations have thus far been used to deal with the crystal containing a defect. The first two simulate the real crystal with a small cluster of atoms, i.e., 30 to 70 atoms. In the one case the surface atoms are left free, while in the other the so-called dangling bonds are saturated with hydrogens. The third approximation is similar to the other two in that one again starts with a small cluster of atoms, but the cluster is judiciously chosen so that periodic boundary conditions can be imposed. The cluster may thus be viewed as a "molecular unit cell" (Messmer and Watkins, 1972) of an infinite crystal with a periodic array of defects. The single defect is thus replaced with an array or "superlattice" of defects in the hope that the interdefect distance can be made large enough to eliminate defect-defect interactions.

The first application of the EHT to defects in solids was by Walter and Birman (1967). Extensive applications were later (1970–1973) carried out by Messmer and Watkins and co-workers (see detailed references below) and by Larkins (1971) in terms of free and H-saturated clusters. By 1973, Watkins and Messmer (1973) concluded that the superlattice approach is by far superior, but only limited studies have been carried out using this method. The superiority of the superlattice method was demonstrated independently and at about the same time by Lee and McGill (1973), who carried out an EHT calculation for the divacancy in Si.

The superiority of the superlattice approach (or molecular-unit-cell-approach) over free or H-saturated clusters can be seen very quickly by comparing the corresponding results for the perfect crystal: Free or H-saturated clusters give a band gap which converges extremely slowly with cluster size, whereas a superlattice calculation immediately reproduces the infinite-crystal band gap obtained by a band-structure calculation (Fig. 14). The latter follows from the fact that the superlattice calculation is in fact a band-structure calculation, albeit with a larger-than-normal unit cell. On the other hand, it should also be noted that once an impurity or defect is introduced, a superlattice calculation does not necessarily reproduce the perfect-crystal band gap.

Another disadvantage of free clusters is that they yield a substantial number of surface states, some of which may lie within the forbidden gap or near the band edges, thus making the identification of true bulk states a difficult task. The surface states also present a problem

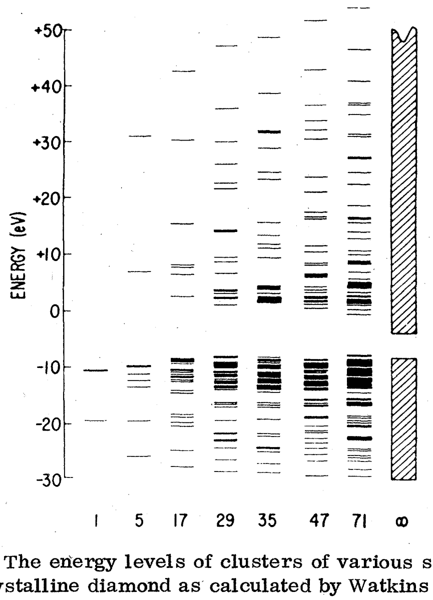

FIG. 14. The energy levels of clusters of various sizes simulating crystalline diamond as calculated by Watkins and Messmer (1973). The cluster with 71 atoms corresponds to including six shells of neighbors of the central atom. The result for the "infinite" cluster comes from a band-structure calculation using identical parameters.

in the calculation of the total energy because for a neutral cluster they should be half-filled. One way to get around this difficulty is to saturate the surface bonds with hydrogens (Larkins, 1971) or simply fill up all the states within the valence-band region, including the unidentifiable "surface" states, whereby one ends up with a heavily negatively-charged cluster (Messmer and Watkins, 1970; 1973). The hydrogen saturation may sound better, but it should be noted that the "surface" states are not removed, but only shifted in energy.

Another demonstration of the superiority of the superlattice method over free and H-saturated clusters can be made by comparing the results one expects from a calculation for a defect at the center of a cluster. Let us first take the case of a set of EHT parameters which, when used in a small free cluster, yield "surface" states in the middle of the band gap. These states ought to be localized near the "surfaces" and decay within the "bulk," though the smallness of the cluster probably makes the distinction somewhat ambiguous. Now, suppose a defect at the center of the cluster introduces a bound state also within the band gap. Clearly, for small clusters, the defect state will interact with surface states rather strongly, Note that the separation between the "centers" of these two types of "localized" states is R, the average "radius" of the cluster. If we now use the same EHT parameters and carry out a superlattice calculation with the same cluster size, there will be no surface states and the gap will be "clean." When a defect is introduced with a bound state in the gap, there will no longer be interactions with surfaces. Instead, there will be interactions between defects on the superlattice. These interactions are, however, *weaker,* because neighboring defects are now separated by $2R$. Finally, whereas defect-surface interactions in free clusters introduce *shifts* in the localized energy levels, which are hard to deal with, defect-defect interactions in superlattices introduce *dispersion.* The dispersion may be fitted to a tight-binding expression from which an approximate position for the level may be extracted (see, e.g., Louie *et al.*, 1976, and the discussion of that work in C1 below).

d. Applications to defects in covalent solids

We turn now to a discussion of particular applications of the EHT to defects. We first examine the ability of the EHT to predict the positions of energy levels associated with defects. The task is complicated by the fact that no reliable experimental values are available for any of the systems for which calculations have been done. Furthermore, no calculations have been reported which have either converged with cluster size (in the case of free clusters), or have reduced the dispersion to acceptable limits, say less than 0.1 eV (in the case of superlattice calculations). In fact, the results available thus far on the vacancy in diamond can be used as an illustration of the difficulties involved in extracting useful information about energy levels from small-cluster calculations. Let us start with free clusters. Messmer and Watkins (1973) noticed that the position of the level in the gap may be quoted either relative to the topmost valence level of the same cluster, or relative to the top of the valence bands of the band-structure calculation. Clearly, in the limit of a very large cluster the two numbers ought to approach the same unique value. Figure 15 shows the two values, as calculated by Messmer and Watkins (1973) for clusters of various sizes. Also shown in the figure is the range of dispersion-obtained by Watkins and Messmer (1973) with the superlattice method using their original EHT parameters. When the improved EHT parameters of Messmer (1971) are used (Messmer and Watkins, 1972), the top of the defect "band" is lowered to about 0.47 eV, a rather small change.

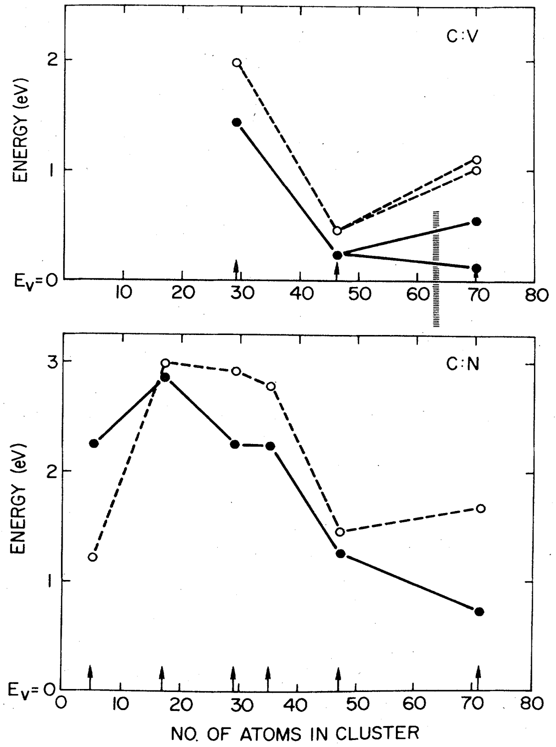

FIG. 15. The convergence of (a) the vacancy level and (b) the substitutional nitrogen level in diamond, from results reported by Messmer and Watkins (1973). The solid circles correspond to the levels being measured from the top of the valence bands of a band-structure calculation, whereas the open circles correspond to the levels being measured from the highest valence level of the respective cluster. Two different 70-atom clusters were used for the vacancy. The vertical line in the vacancy figure indicates the dispersion obtained from a 64-atom repeated-cluster calculation.

The case of nitrogen in diamond is even more revealing. Messmer and Watkins (1973) did a similar study of the bound state as a function of cluster size and obtained similar results, but the situation is actually more disturbing. The 35-atom undistorted cluster yields a bound state at about 1.5 eV below the cluster's conduction band. If one were to measure it from the conduction-band edge of the infinite crystal, the level would not be bound! (See Fig. 14). Now, the same cluster, after a Jahn–Teller trigonal distortion is introduced, yields a bound state which is 2.2 eV above the cluster's valence-band edge. Since one could follow the level being lowered in energy as the distortion is increased, one might want to continue measuring it from the conduction band edge. The final result would be 7.3 eV below the cluster's conduction-band edge, which, if measured from the infinite-crystal conduction-band edge, would be quoted as deep into the valence bands. Superlattice calculations have also been performed for C:N. The dispersion was found to be negligible (Watkins and Messmer, 1973), but the resulting energy level was not reported.

All in all, the "disappointing conclusion," in the words of Watkins and Messmer (1973), is that even larger clusters are needed, preferably with periodic boundary conditions, for a meaningful determination of bound-state energies. Again in the words of Watkins and Messmer (1973), this conclusion "should serve as a dramatic warning that results on a small cluster should be interpreted cautiously, at least for a strongly covalent material such as the elemental semiconductors." A similar conclusion was reached by Larkins (1971) who observed that a 35-atom cluster for the vacancy in Si produces a level in the gap, in apparent agreement with experiment, but a 41-atom cluster does not! Perhaps an even more disappointing conclusion is that even a superlattice (or repeated cluster) approach does not necessarily remove this difficulty. As we observed earlier, when the cluster contains an impurity or defect, a superlattice calculation does not necessarily reproduce the perfect-crystal band gap and the resulting bound state is broadened into a band.

Despite the inability of the cluster-EHT calculations to produce accurate energies for the localized states, the results have been extremely valuable from other points of view. For one, they demonstrated, perhaps in a painful way, that the wave functions of vacancies and other deep defects are not as localized as originally expected (e.g., Coulson and Kearsley, 1957; Yamaguchi, 1962). Furthermore, the overall picture of the wave function obtained by Messmer and Watkins (1972) from superlattice calculations is probably substantially correct, namely that a good part of it (perhaps 50%) is localized on the nearest neighbors, whereas the rest decays very slowly with a more-or-less constant amplitude. The correctness of this picture is substantiated by the fact that virtually identical wave functions were obtained at Γ and X of the small Brillouin zone of the superlattice (Watkins and Messmer, 1973). A similar picture for the wave function was also obtained by independent calculations, discussed in the previous section (Lannoo and Lenglart, 1969; Bernholc and Pantelides, 1978).

The calculations of Messmer and Watkins (1973) revealed another intriguing result about the wave function of the vacancy, which appears quite convincing. It was found that the molecular-orbital coefficients $C_{\mu\lambda}$ for the p functions on a given near-neighbor atom are not equal. In particular, for the wave function of the T_2 state that transforms like z, the coefficient of p_z is found to be larger than those of p_x and p_y. This result means that the "dangling hybrids" that would be expected to point toward the position of the missing atom are not pure sp^3 hybrids, as one might expect for an unreconstructed ideal vacancy. This "tilt" of the wave function was determined to be larger than +5.4° for all cluster sizes and was believed to be real. No experimental results are available for diamond, but EPR measurements for the vacancy in Si had already revealed such a tilt of +7.2° (Watkins, 1963), in support of the theoretical results. The calculations also demonstrated that the origin of the tilt need not be due to Jahn–Teller distortions. Additional calculations (Messmer and Watkins, 1973) demonstrated that the tilt is in fact caused by interactions with nearby valence-band states and concluded that the tilt is a measure of the proximity of the localized level from the valence bands. Experiment again supports this conclusion, since other vacancy-related defects with levels farther from the band edge display negligible tilt (see references in Messmer and Watkins, 1973).

The wave function for nitrogen in diamond (Messmer and Watkins, 1973) turned out to be very good and to vary little with cluster size. In this case, there exist detailed ENDOR data which actually measure the molecular-orbital coefficients $C_{\mu\lambda}$ of (10.4) (Watkins and Corbett, 1961). The agreement between theory and experiment is truly remarkable (Table VII). Similarly good agreement was obtained by comparing electric quadruple interaction parameters, also extracted from EPR and ENDOR data.

Finally, we turn to applications of the EHT which make use of the total-energy expression (10.10) to predict lattice relaxation, distortions and related quantities. Almost all such work has been on 35-atom free clusters for various defects in diamond. An indication of the validity of the procedure is provided by the fact that elastic constants, when calculated the same way, agreed well with experimental values (Messmer and Watkins, 1973). The approach was first used to study lattice distortion for substitutional nitrogen. As we mentioned already, it was determined that when a trigonal Jahn–Teller distortion is imposed, the bound state in the gap is lowered rather dramatically. This particular distortion is in agreement with EPR data. As we saw already, the wave function of the distorted configuration agrees well with ENDOR data.

Similar work was done for the vacancy in diamond by Messmer and Watkins (1973) and by Yip (1974) for the vacancy in Si and Ge. Symmetric lattice relaxation was determined to correspond to 13% outward relaxation of the nearest neighbors, but the Jahn–Teller distortion could not be unambiguously determined as tetragonal or trigonal. Estimates of the migration energy of the va-

TABLE VII. Wave-function coefficients $|C_\mu|^2$ determined by EHT cluster calculations by Messmer and Watkins (1973) for the nitrogen and nearby carbon atoms, and comparison with experiment.

	Orbital	35 atoms	47 atoms	71 atoms	Expt.
N(0, 0, 0)	$2s$	0.004	0.004	0.004	0.060
	$2p$	0.174	0.363	0.224	0.23
C(1, 1, 1)	$2s$	0.029	0.022	0.017	0.066
	$2p$	0.834	0.765	0.738	0.73
$C(\bar{1}, \bar{1}, 1)$	$2s$	<0.001	<0.001	<0.001	···
	$2p$	0.003	0.010	0.003	···

cancy were also obtained by comparing the total energy of the vacancy at (000), the vacancy at (111), and an intermediate configuration consisting of vacancies at both (000) and (111) and an atom halfway between the two sites, i.e., at $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$. No experimental values were available for comparison, but comparison with other theoretical estimates indicated a probable overestimate of the migration energy.

An interesting study along the same lines was carried out by Weigel, Peak, Corbett, Watkins, and Messmer (1973) for the self-interstitial in diamond. The goal was to determine the stable configuration. The conclusion was that neither the tetrahedral nor the hexagonal interstitial sites were the preferred locations, contrary to previous assumptions (Yamaguchi, 1963; Benneman, 1965; Singhal, 1971). The preferred configuration was instead found to be an *interstitialcy*. Two such configurations were found to be likely, i.e., a split-⟨100⟩ interstitial (i.e., two atoms sharing a substitutional site with the pair being in the 100 direction), or a bond-centered interstitial. The study also determined a migration path, which turned out to be quite insensitive to the choice of EHT parameters (the standard ones or the improved ones, obtained by Messmer, 1971, to fit the band structure better). The study was extended to describe charged interstitials, but the results are much less reliable for reasons discussed earlier. The predicted differences in preferred interstitial locations between the various charge states, did, however, suggest that low-temperature athermal migration of the interstitial would be possible by simply capturing carriers, a mechanism suggested earlier by Bourgoin (Bourgoin and Corbett, 1972).

e. A modified EHT for charged centers?

We just saw that Weigel *et al.* (1973) attempted to study charged interstitials in diamond by simply adding extra electrons to partially occupied states. They recognized the crudeness of the approximation in view of the fact that the new self-consistent potential of the charged center should produce substantially different energy levels. A similar approximation was also used by Yip (1974) in his study of charged vacancies in Si and Ge. Some time earlier, however, Shimizu and Minami (1971) used what may be referred to as modified EHT to study the positively charged state of Si:S. These authors added a new term to the off-diagonal $h_{\mu\lambda}$ of the form $-S_{\mu\lambda}e^2/\epsilon(R)R$, where R is defined by $\mathbf{R}^{-1}=(\mathbf{R}_j^{-1}+\mathbf{R}_{j'}^{-1})/2$. Here $\mathbf{R}_j,\mathbf{R}_{j'}$ are the atomic positions for ϕ_μ and ϕ_λ, respectively; $\epsilon(R)$ is a position-dependent dielectric function like the one that appears in the definition of the point-charge model in the EMT (Sec. VIII). Shimizu and Minami made use of symmetry and, by focusing only on states of A_1 symmetry, they were able to treat a cluster of 274 Si atoms. The result was an A_1 level 0.57 eV below the edge of the cluster conduction band. The agreement with experiment (0.61 eV) is very good. On the other hand, the cluster band gap is about 3 eV, which is closer to the direct gap at Γ instead of the minimum indirect gap, whereby the determination of the energy level is ambiguous. The new prescription for charged states is, however, acceptable in the overall EHT framework. It has not been tested adequately thus far.

f. A new "Renormalization-group" technique

Chui, Weigel, and Corbett (1977) have recently reported calculations using a new technique to solve the EHT equations for even larger clusters. The idea is to start with a small cluster, solve the EHT secular equations directly, then double the cluster and expand the new wave functions in terms of only selected *eigenfunctions* of the original cluster. The "special" eigenfunctions may be chosen, for example, from the vicinity of the bandgap. The procedure may be repeated several times to treat larger and larger clusters with decreasing accuracy. The technique obtains its name from the renormalization-group theory developed by Wilson (1975) for calculations of cooperative phenomena. No results have been published in the literature yet.

2. Other semiempirical methods: Large-cluster techniques

In recent years it has been popular to parametrize the energy bands of tetrahedral semiconductors in a rather direct way in terms of an LCAO calculation. Only s and p orbitals are used and only first or first and second nearest-neighbor interactions are included (Pandey and Phillips, 1974, 1976; Chadi and Cohen, 1975; Pantelides and Harrison, 1975). The same secular equations as in the EHT band-structure calculation must be solved, but the matrix elements $h_{\mu\lambda}$ are strictly treated as parameters. The overlap matrix $S_{\mu\lambda}$ is usually treated as diagonal (orthogonal basis set). The resulting parameters depend somewhat on whose energy bands one fits and which points in the Brillouin zone are fitted.

Once the Hamiltonian matrix elements are known, one can then use them to treat vacancies and self-interstitials by making an approximation very similar to that made by the EHT, i.e., that the Hamiltonian matrix elements in the perturbed crystal are the same as those of the perfect crystal. The only difference is that either an atom is missing (vacancy), or an extra atom is present (self-interstitial). We have already seen use of such an approximation by Lannoo and Lenglart (1969) and Bernholc and Pantelides (1978), who studied vacancies using a Green's-function perturbative approach (Sec. IX). The nonperturbative approaches we wish to review here are essentially cluster calculations and, except for one case, they are also techniques that can hand handle very large clusters (more than 2,000 atoms). The one exception is the recent work of Lowther (1977), who used an LCAO parametrization and studied vacancies in terms of 17-, 35-, and 47-atom clusters. Lowther, however, seems to have been totally unaware of the vast literature on small clusters using the EHT, and none of the problems we explored earlier in this section were addressed. His 35-atom clusters appear to reproduce the infinite-crystal band gaps rather well, with no surface states localized in the gaps. This last fact raises serious questions about the validity of these calculations, since Pandey and Phillips (1974, 1976) have shown that an accurate tight-binding calculation in fact produces surface states in the gap, in agreement with self-consistent surface-state calculations (Appelbaum and Hamann, 1974). Lowther also extracts several "vacancy levels" within the gap and within the valence bands but does not state how the identification was accomplished.

We turn now to the two techniques that have been used thus far to deal with very large clusters. Both of them are based on a Green's- function formalism, but, unlike the Koster-Slater approach, they calculate the new Green's function directly. They make use of Eq. (6.18) for the density of states, which in an LCAO basis set of functions ϕ_μ becomes

$$D(E) = -\frac{1}{\pi} \operatorname{Im} \sum_\mu G_{\mu\mu}(E) \tag{10.12}$$

where

$$G_{\mu\mu}(E) = \langle \phi_\mu | G | \phi_\mu \rangle . \tag{10.13}$$

Equation (10.12) enables one to define a local density of states $D_\mu(E)$ by

$$D_\mu(E) = -\frac{1}{\pi} \operatorname{Im} G_{\mu\mu}(E) \tag{10.14}$$

which can be immediately shown to be also given by [see, e.g., the way Eq. (6.20) was derived]

$$D_\mu(E) = \sum_\nu |\langle \phi_\mu | \psi_\nu \rangle|^2 \delta(E - E_\nu) , \tag{10.15}$$

where ψ_ν are the eigenfunctions and E_ν's are the eigenenergies of the system. Note that these equations are true for an arbitrary system and $D_\mu(E)$ is a weighted or projected density of states. Clearly, if one is interested in finding the localized states in the vicinity of a defect, it would be adequate to determine $D_\mu(E)$ with ϕ_μ being a basis function on either the impurity atom itself or a nearby atom (in the case of the vacancy). Within the band gap of a semiconductor containing an impurity, $D_\mu(E)$ would consist of one or more δ functions, but in an approximate calculation, the δ functions would in fact turn out to be broadened δ functions.

A number of general techniques have been developed to calculate $D_\mu(E)$ for an arbitrary collection of atoms by directly calculating $G_{\mu\mu}(E)$ instead of calculating the eigenvalues E_ν (see, e.g., an interesting overview in Haydock, Heine, and Kelly, 1972). One such method is the recursion or continued fraction method which ends up expressing $G_{\mu\mu}(E)$ for a particular μ as

$$G_{\mu\mu}(E) = \frac{1}{E - a_1 - b_1 g_1(E)} \tag{10.16}$$

with

$$g_n(E) = \frac{1}{E - a_{n+1} - b_{n+1} g_{n+1}(E)} \tag{10.17}$$

The coefficients a_n and b_n are given by somewhat complicated recursion formulas. Clearly (10.16) with (10.17) is an expansion of $G_{\mu\mu}(E)$ as a continued fraction. The details of the method may be found in the papers of Haydock, Heine, and Kelly (1972, 1975) and are not of particular interest to our discussion. This and similar techniques were designed to treat spatially disordered systems such as amorphous semiconductors; they do not exploit in any way the translational symmetry of the unperturbed crystal.

The continued-fraction method has been used by Kauffer, Pecheur, and Gerl (1976, 1977) to study the vacancy and self-interstitial in Si. They first use a "perfect" cluster of 2545 Si atoms and a parametrized s-p Hamiltonian. The cluster is found to reproduce the bulk density of states of Si well (it takes that many atoms, not 35 or 71, to get convergence with cluster size!). The central atom is then removed and the local density of states at a nearest neighbor is calculated. A broadened δ function is detected in the gap for states of T_2 symmetry and another such function is detected within the valence bands for states of A_1 symmetry. A similar calculation was done for the self-interstitial.

Another method that has been introduced to calculate $G_{\mu\mu}(E)$ for large clusters is the effective-Green's-function method of Joannopoulos and Mele (1976). In this method, one defines a *broadened* Green's function by

$$\hat{G}(E) = (2\pi\sigma^2)^{-1/2} \int_{-\infty}^{\infty} G(E') \exp[-(E-E')^2/2\sigma^2] \, dE' . \tag{10.18}$$

It is then shown that

$$\hat{D}_\mu(E) = -(1/\pi) \operatorname{Im} \hat{G}_{\mu\mu}(E) \tag{10.19}$$

represents a broadened density of states. $\hat{G}_{\mu\nu}$ is found to satisfy the equation

$$\sigma^2 (d/dE) \operatorname{Im} \hat{G}_{\mu\nu}(E) = \sum_\lambda H_{\mu\lambda} \operatorname{Im} \hat{G}_{\lambda\nu}(E) - E \operatorname{Im} \hat{G}_{\mu\nu}(E) \tag{10.20}$$

which reveals that $\hat{D}_\mu$ can be obtained by integrating (10.20) row by row. The advantage is that only a few rows of the otherwise large matrices must be specified. Joannopoulos and Mele used this technique to calculate the states around a vacancy in Ge using a parametrized tight-binding Hamiltonian and simply modeling the vacancy by the removal of an atom. Their Hamiltonian included only nearest-neighbor interactions so that the result was a degenerate T_2 and A_1 level in the gap, as previously found by Lannoo and Langlert (1969).

C. Self-consistent methods

In this subsection we describe methods which attempt to construct the potential V of the perturbed crystal and solve the eigenvalue problem (10.1) numerically in a self-consistent way.

1. The self-consistent pseudopotential method

In the self-consistent pseudopotential scheme (Appelbaum and Hamann, 1974; Cohen, Schlüter, Chelikowsky, and Louie, 1975), the pseudopotentials of the ionic cores are taken to be fixed model potentials, determined to fit free-atom properties, while the valence charge density is determined self-consistently. The pseudo-Bloch functions are usually expanded in plane waves. The method has been very successful in describing bulk band structures and surface states (Schlüter, Chelikowsky, Louie, and Cohen, 1975a, 1975b).

The method has been used to study the vacancy in Si by Louie, Schlüter, Chelikowsky, and Cohen (1976). They employed the "supercell" method which we have already described in conjunction with the EHT. In this method, the calculation is performed on a periodic array of vacancies in an otherwise perfect Si crystal. The problem then becomes one of standard band theory. The only complication is the large unit cell that must be employed. As a result, Louie *et al.* were able to perform the calculation only at high symmetry points where the size of the secular matrix could be reduced substantially.

The results were similar to those of Messmer and Watkins (1972). Dispersion of 1.2 eV was obtained for the vacancy level in the band gap. By using a tight-binding expression for the dispersion, the "average" position of the vacancy level in the gap was determined to be at about 0.5 from the conduction band edge. A strong dispersionless resonance at -8.2 eV from the top of the valence bands was also found. An interesting result of the self-consistent calculation is that the perturbation potential, as determined self-consistently, is rather similar to the total pseudopotential of a Si atom, which had been previously used by Callaway and Hughes (1967). Contrary to the Callaway-Hughes results, however, Louie *et al.* found a bound state deep in the gap, perhaps due to the fact that convergence had been attained.

Louie *et al.* (1976) also carried out self-consistent calculations for two different models of reconstruction (or Jahn-Teller distortion) of the single neutral vacancy in Si. As we saw previously, reconstruction comes about because the level in the gap (which in the limit of no dispersion would be a T_2 triplet) is only partially occupied, i.e., it contains only two electrons out of a possible six. Both models chosen for the study corresponded to a uniaxial distortion along the cubic [100] axis in agreement with observations (Watkins, 1965). The first model (Rec I) corresponded to a net relaxation of the nearby atoms *toward* the vacancy; the second model (Rec II) corresponded to a net relaxation of the nearby atoms *away* from the vacancy. The actual amount of displacement was estimated from previous theoretical work by Swalin (1961). The results of the calculations are interesting though not quantitatively conclusive. In both models the resonance in the valence bands does not move, but the triplet in the gap splits into a singlet and a doublet, as one would expect; the singlet goes down in energy and is fully occupied, whereas the doublet goes up and is completely empty. Thus both models provide Jahn-Teller stabilization and one cannot determine which type of reconstruction would be actually preferred. One might choose Rec I for which the singlet is at lower energies, but the large dispersions obtained in both cases would not warrant the conclusion. In fact, Louie *et al.* favor Rec II, because the atoms in this case move toward the *bulk*, an effect known to occur at free surfaces (Appelbaum and Hamann, 1974; Phillips, 1974).

The qualitative results of the reconstructed vacancies are rather interesting and instructive. In Rec I, the dispersion of the bound states is increased. This increase is traced to the fact that the movement of the nearby atom toward the vacant site weakens the back bonds, so that some charge is transferred to the second-neighbor bonds, thus making the bound-state wave function more spread out. On the contrary, in Rec II, the back bonds are strengthened and dispersion is decreased.

The above theoretical results may be compared with experiments which determine the charge state of the vacancy as a function of the position of the Fermi level in the gap. The energy level calculated by Louie *et al.* (1976), therefore, corresponds to the lowest possible position of the Fermi level for a crystal containing neutral vacancies. Unfortunately, experiments monitoring the Fermi level and the charge state of the vacancy in Si are not very precise. According to Watkins (1965), the neutral-vacancy Fermi level is about 50 meV above the valence band, whereas according to Naber, Mallon and Leadon (1973) the same level is at 440 meV above the valence band edge. More recently, Kimerling (1977) tentatively identified the level at 110 meV. It appears, therefore, that theory and experiment agree that the level is somewhere in the lower half of the gap, but neither can pin it down more precisely.

A more important observation is that the calculated energy separation between the T_2 level in the gap and the top of the valence band does not correspond to any optical excitation energy. The reason for this is that the T_2 level contains two electrons so that if one electron is removed optically (or another electron added) the position of the level would move substantially. A similar observation applies to the semiempirical results discussed earlier. The situation is analogous to the ionization of a He atom or the ionization of a Si:S double donor. For the purpose of calculating excitation energies one must calculate both the initial- and final-state charge configuration and subtract total energies. In lieu of total energies, one might monitor the changes produced by the movement of both the bound states in the gap and the resonance and antiresonance states in the valence bands. Such calculations have not been performed thus far.

2. The $X\alpha$-scattered-wave method

This method is a particular technique of solving the general eigenvalue problem (10.2) for a small collection of atoms. The potential is constructed self-consistently for the entire system, including the atomic cores (i.e., no pseudopotential approximation), using Slater's $X\alpha$ exchange. However, one important approximation is introduced: Spheres are drawn centered about each atom: the potential is then spherically averaged in each sphere and volume averaged to a constant in the interstitial region. The potential is also spherically averaged in a large sphere surrounding the atomic spheres. The eigenvalue problem is then solved by scattering methods (Slater and Johnson, 1972; Johnson and Smith, 1972). Excitation energies are calculated by the transition-state method of Slater (1972), i.e., the self-consistent solution is carried out with half of an electron in the initial state and half of an electron in the final state (or half of an electron missing for an ionization energy).

The method has had considerable success in treating the one-electron spectra of molecules and mixed success in determining internuclear separations by minimizing total energies. It has also been applied to a wide variety of solids by performing calculations on small clusters. These calculations have been used to interpret optical absorption spectra, x-ray emission spectra, photoemission spectra, etc., noting that these excitations are local in nature. The surfaces of the clusters are usually saturated either with hydrogens or by placing a charged sphere around the cluster to simulate the Madelung potential arising from the rest of the crystal in an ionic solid. The method has met with success in most cases but the results have not always been corroborated with studies of the effect of cluster size. We saw earlier in this section that the results are very sensitive to the cluster size and it is hard to see why a self-consistent calculation would eliminate the problem of surface states and surface-defect interactions. In fact one might expect self-consistency to make things worse since the self-consistent potential is appropriate to the cluster *per se*, viewed as a molecule, whereas in a semiempirical calculation the potential is not adjusted at the surfaces (see discussion by Messmer and Watkins, 1973). An example illustrating the potential pitfalls in simulating solids by small clusters is provided by the $X\alpha$ scattered-wave ($X\alpha$-SW) calculation by Tossel, Vaughn, and Johnson (1971), who were able to interpret all the observed spectra of SiO_2 in terms of an SiO_4^{4-} cluster. Similar results were obtained by an approximate Hartree-Fock calculation by Yip and Fowler (1974) for the same cluster, but slightly larger clusters were found to produce substantially different results.

The application of the $X\alpha$-SW method to deep defect levels in covalent solids was anticipated by Johnson, Norman, and Connolly (1972) who took note of the results of Messmer and Watkins (1973) and concluded that large clusters would in fact be necessary, perhaps with periodic boundary conditions, in which case the method reduces to the KKR energy-band-structure formalism (Korringa, 1946; Kohn and Rostoker, 1954). The first actual application of the method to defects in a tetrahedral semiconductor was by Cartling, Roos, and Wahlgren (1974) who used the technique to study impurities in Si. In that work, and subsequent work by Cartling (1975), a cluster of only five Si atoms saturated with hydrogen atoms was employed. The cluster appears to be very small for convergent results, but Cartling makes a good case that the simulation is reasonable, especially since a transition-state calculation of the band gap yields 0.95 eV, in good agreement with the observed value of 1.1 eV for the indirect gap.

Cartling did calculations on Si:S, which is a deep donor, Si:Zn, which is a deep acceptor, and Si:Fe, which is a transition-metal impurity. The choices are very good since these systems may be viewed as typical and the least pathological. Perhaps the most interesting result of the calculation is that Si:S is found to introduce an additional A_1 bound state *below the bottom of the valence band*, a conclusion which is very likely to survive the tests of higher accuracy and larger clusters. The bound state is a result of the very strong perturbation of the sulfur atom and has not been observed experimentally. If is also present in the results of Shimizu and Minami (1971) who used a modified EHT (see discussion earlier in this section) to describe the S^+ center. As expected, the bound state is even deeper in the S^+ center (~23 eV below the valence-band *top* in Shimizu and Minami's calculation) than in the S^0 center (~16 eV below the valence-band top in Cartling's calculation).

Cartling's results for the bound states in the fundamental gap are in reasonable agreement with experiment, but the uncertainty is somewhat large. Consider, for example, Si:S. In the perfect cluster, the bottom of the conduction band is an A_1 state. In the cluster containing sulfur, however, A_1 is lowered in energy to become the bound state and the bottom of the conduction band is a T_2. Cartling defines the impurity electron ionization energy to be the $A_1 \to T_2$ transition-state excitation energy, but the definition cannot be totally justified, since treating T_2 as the lowest conduction-band state yields a transition-state band gap of 1.23 eV, i.e., 25% larger than the perfect-cluster band gap. One would conclude that the conduction band edge is ill defined in the cluster containing a donor impurity, or alternatively, interpret the results as inherently uncertain by 25%. Clearly, these results call for calculations on larger clusters, which may remove this uncertainty.

Another particularly interesting application of the $X\alpha$-SW method has been reported by Watkins and Messmer (1974) who used the technique to study the vacancy in diamond. The objective of the work was to study the many-electron multiplet structure as a function of cluster size. The study demonstrated that as the cluster size is increased, the bound-state wave function becomes more delocalized (in agreement with previous work by the same authors) and as a result the multiplet structure is reduced dramatically. This conclusion settled the controversy concerning the importance of many-electron effects which were found to be large in the defect-molecule model, but were left out in cluster calculations using the EHT. It also established the inadequacy of the defect-molecule model to produce a reliable description of the electronic structure of defects.

A more recent application of the $X\alpha$-SW method to impurities in Si has been reported by Hemstreet (1977). Hemstreet used clusters similar to those used by Cartling, but he did not carry out transition-state calculations in order to determine excitation energies. As in Cartling's work, no tests were carried out to check the convergence of the results with cluster size. Hemstreet studied the elements of the first transition series (Cr through Zn) as substitutional impurities in Si, and found that there is a variety of trends as one goes through the series. Comparison with experiment, apparently good, was not conclusive, largely due to the uncertainties in the interpretation of data. The calculations predicted for each impurity whether it would be a donor or an acceptor, and therefore provided useful guidelines for reinterpretation of experiments.

3. The cushioned-cluster LCAO method

Recently, self-consistent calculations on clusters have also been reported by Menzel, Mednick, Lin, and Franklin (1975), Chaney and Lin (1976), and by Chaney (1976). The distinguishing feature of these calculations is that these authors use the Hamiltonian of an infinite crystal but expand the wave functions in terms of atomiclike orbitals on a small cluster of atoms. In order to avoid a variational collapse into the core orbitals of the surrounding atoms, core orbitals are placed on a "cushion" consisting of a few additional shells of atoms. Menzel *et al*. (1975) argue that by using the Hamiltonian of an infinite crystal "no physical surface is present" and surface states are eliminated. This assertion is, however, implausible, because surface states are introduced by the truncation of the Hamiltonian *matrix*, which in turn can be caused by the truncation of either the basis set or the Hamiltonian operator itself. For example, in the semiempirical cluster calculations we discussed earlier, one implicitly assumes an infinite-crystal Hamiltonian, since the matrix elements of the last few shells of atoms are taken to be identical with those in the bulk, but surface states exist nevertheless. Alternatively, the use of a truncated Hamiltonian, as in Cartling's self-consistent calculations, makes the Hamiltonian matrix elements near the surface different from those in the bulk, which simply shifts the surface states to different energies. It is not clear to the present author why the calculations of Lin and coworkers do not give surface states. It is more likely that surface states do in fact exist, but their energies are within the energy bands. Because of the smallness of the clusters, the surface states may not be easily identifiable as such. This question may be rather academic, however, if the bound states of a defect at the center of such clusters converge uniformly with cluster size. Applications have so far been reported for the F center in LiF (Chaney and Lin, 1976; Chaney, 1976). Convergence with cluster size was very good. A detailed discussion of the results is beyond the scope of the present paper. Applications to defects in semiconductors have not been reported.

XI. COMPARATIVE CRITIQUE OF THEORETICAL METHODS-CONCLUSIONS

We are now in a position to attempt a comparative critique of theoretical methods and to assess future prospects. For *shallow* levels, particularly for excited states, effective-mass theory, described in Sec. VII, is undisputably the only method that has provided detailed and accurate quantitative results. For ground-state energies, generalizations of effective-mass theory with realistic impurity potentials, described in Sec. VIII, seem to provide good results for a variety of shallow and moderately deep levels. The new developments involving Umklapp terms, discussed in Sec. VII.E, are likely to provide impetus for new work, which may lead to ways to obtain binding energies in the compound semiconductors, including their site dependence. However, sooner or later, effective-mass-type theories break down, and will never be able to handle accurately those levels whose wave functions must be built from large regions of k space and from many bands. We may list, for example, levels associated with vacancies, self-interstitials, transition-metal impurities, and complexes such as divacancies, vacancy-impurity pairs, etc. Attention must then inevitably focus on other methods, which are more powerful and, almost by necessity, more involved.

In order to compare the relative merits of the various methods we discussed in Secs. IX and X, which are especially designed to describe *deep* levels, we pose the following question: *Given a one-electron Hamiltonian for a crystal containing an isolated impurity or defect, which is the best technique to use in order to obtain accurate energy levels and additional information (such as transition energies, wave functions, etc.) needed to make contact with experiment?*

In Secs. IX and X we discussed the various methods in an order that was convenient for reviewing the available literature. Now we wish to look back and compare their merits in a critical way. The first choice that one must make is the basis set in which to expand the bound-state wave functions. There are two fundamentally distinct options. One can use (a) a set of propagating states; or (b) a set of localized states.

a. Propagating basis functions One can expand the bound-state wave function in terms of Bloch functions and end up with a secular matrix (see Sec. IV.A) whose size is determined by the number of bands and the number of k points used in the expansion. This method is a brute-force method and one runs out of computer capacity before convergence with respect to k points and bands can be reached. It was attempted by Jaros and co-workers and was abandoned soon thereafter in favor of determinantal techniques that employ localized basis functions.

Plane waves are another set of propagating states that can be used to expand the bound-state wave functions. At first sight, in the absence of periodicity, one might think that the number of plane waves needed (in a *continuous* k space) would be prohibitive. In fact, the self-consistent pseudopotential superlattice calculation for Si:V by Louie *et al*. (see Sec. X), is equivalent to a plane-wave expansion with a particular, but arbitrary, grid of k points. This grid is entirely determined by the choice of superlattice and the resultant dispersion in the bound state means that a slightly displaced grid of k points produces a different bound-state energy. There is in fact a one-to-one correspondence between the plane-wave expansion and the Bloch-function expansion, and it appears unlikely that either can be brought to acceptable convergence (elimination of dispersion) in the foreseeable future.

Another way of solving the same problem in terms of free-particle states is the $X\alpha$-SW method which employs scattering theory. The method is limited to small clusters and studies of convergence with respect to cluster size have not been extensive. An inherent limitation of the method, common to all small-cluster methods, is its inability to unambiguously relate the results to conventional band structures. One of the strengths of the method is that electrostatic self-consistency permits the description of the charge transfer associated with the charged impurities. On the other hand, its use of the muffin-tin approximation raises questions about its applicability to covalent systems. Though the method has produced useful information, it does not appear to have the promise of Green's-function techniques employing a basis of localized functions (see below).

b. Localized basis functions There is no doubt that, for deep-level impurities and defects, localized functions are a more natural choice for expanding bound-state wave functions. There exist, of course, several sets of localized functions one can choose from, such as Wannier functions, true atomic orbitals, atomiclike orbitals, Slater-type orbitals, Gaussians, etc. Assuming a choice has been made, the question that remains is one of technique, and two options are available: (1) The first option is to construct and diagonalize the secular matrix H_{ij}. (In this case, it would only be a minor detail whether one wants to exploit the fact that $H_{ij} = H_{ij}^0 + U_{ij}$). The important consideration is that the size of the matrix is governed by the *range of the bound-state wave function.* (2) The second option is to break H up into $H^0 + U$ and convert the secular matrix into a determinantal equation in the Koster-Slater manner (Sec. IX.B), or equivalently, set up the Lippmann-Schwinger Green's function equation in the chosen localized representation. In this formulation, the size of the determinant is governed by the *range of the perturbation potential*. For *neutral* potentials, which are usually localized over a few atoms, the determinantal or Green's-function formulation is undoubtedly the most promising technique.

A rather dramatic illustration of this promise is provided by a comparison of various LCAO-parametrized calculations of the vacancy in a covalent semiconductor. The work of Messmer and Watkins demonstrated that clusters of 70 some atoms are not adequate (secular matrix ~300 × 300). The work of Kauffer *et al*. demonstrated that a cluster of over 2400 atoms (corresponding to a matrix ~10 000 × 10 000) is needed for convergent results when the method of Haydock *et al*. (1972, 1975) is used. A similar result was obtained independently by Joannopoulos and Mele (1976). In the determinantal method, however, the vacancy problem reduces to a 1 × 1 matrix (!). This simple problem was in fact solved in 1969 by Lenglart and Lannoo, but it seems that its significance was not widely appreciated. Very recently, Bernholc and Pantelides (1978) reproduced the results of Kauffer *et al*. with the Koster–Slater 1 × 1 matrix. The methods of Kauffer *et al*., of Joannopoulos and Mele, and of Bernholc and Pantelides in fact accomplish the same thing, namely the calculation of a single density-of-states curve (read Green's function). The difference is that in the methods of Kauffer *et al*. and of Joannopoulos and Mele one must calculate a Green's function for the crystal containing the vacancy, and therefore ends up using the rather cumbersome and approximate methods introduced originally for amorphous materials, whereas in the Koster–Slater-type method used by Bernholc and Pantelides, one must calculate a Green's function of the *perfect* crystal, which can be calculated by summing over the Brillouin zone, a simple and highly accurate procedure. In addition, the method used by Bernholc and Pantelides yields directly the change in the density of states (including all resonances and antiresonances) introduced by the defect, whereas in the other two methods the changes must be obtained by subtracting two rather similar quantities, which increases the inherent uncertainty. The superiority of the method can be traced to the fact that it exploits both the short range of the defect potential and the translational symmetry of the host crystal, whereas the other two methods exploit neither.

What about the use of the determinantal Koster–Slater Green's function method in more realistic calculations, using some sort of first-principles potential, instead of an LCAO parametrization? Such Green's function techniques have already been used successfully for chemisorption studies (Lang and Williams, 1976). Callaway and co-workers pioneered the application of the method to defects in semiconductors but their use of Wannier functions rendered the method cumbersome. A more promising development is the use of LCAO-type basis sets, drawing on the vast experience present in the literature for LCAO-type basis sets for band structure calculations. The latter procedure is the subject of a forthcoming paper by Bernholc, Pantelides, and Lipari (1978). It builds on the Green's function machinery developed by Callaway and others. The calculation of bound-state energies, resonances, antiresonances, phase shifts, formation energies, etc., becomes feasible.

What about the Bassani–Iadonisi–Preziosi method, modified slightly and used extensively by Jaros and co-workers (BIPJ method)? Both the BIPJ method and the standard Koster–Slater Green's function method require an integration over the Brillouin zone of the host material. In the BIPJ method this integration must be carried out anew for each impurity potential U, whereas in the Koster–Slater method the zone integration depends only on the host and therefore has to be done only once for each host material. Since the zone integration is the most time consuming part of such calculations, this distinction is important.

The superiority of the determinantal Koster–Slater Green's-function method discussed above is limited to neutral potentials, which are nonzero only in a finite volume surrounding the point defect. When the impurity or defect potential has a Coulombic tail, the size of the determinant will be determined by the range of the wave function, as in the other methods. Jaros, however, has found that the Coulombic tail can be truncated without serious implications, making the determinantal method competitive.

The overall conclusion is that the most promising technique for deep-level defects and impurities in semiconductors is the Green's-function determinantal method. Its foundations were laid in 1954 by Koster and Slater. General and powerful formal results appeared in the 1960's. Pioneering calculations were performed by Callaway and co-workers. Let us look to the future for new applications which will set us on our way to better understanding the deep states introduced by defects and impurities in semiconductors.

ACKNOWLEDGMENTS

The author has benefited over the years from conversations with numerous colleagues whose interests overlap the subjects discussed in this paper. In particular, thanks are due to J. Bernholc, H. G. Grimmeiss, W. A. Harrison, L. Hedin, D. V. Lang, T. N. Morgan, J. L. Pollmann, T. D. Schultz, F. Stern, J. A. VanVechten, U. Von Barth, and A. R. Williams for discussions directly relevant to the material presented in this paper.

The author is indebted to H. G. Grimmeiss and D. V. Lang for their permission to mention their recent data on Si:Au.

APPENDIX: PHOTOIONIZATION CROSS SECTIONS

In this appendix, we give a brief review of theoretical work on photoionization cross sections. As we saw in Sec. V, the latter are measured by optically exciting electrons from localized levels into the conduction bands (electron emission) or from the valence bands into localized levels (hole emission). Either process is described by Fermi's golden rule (Schiff, 1955). We have

$$\sigma(E)=\left(\frac{\mathcal{E}_{\text{eff}}}{\mathcal{E}_0}\right)^2\frac{4\pi^2\hbar^2\alpha}{3m_0{}^2n}\,\frac{1}{E}\sum_f|\langle\psi_f|p|\psi_i\rangle|^2 \times\delta[E-(E_f-E_i)]\,,$$

where α is the fine-structure constant ($e^2/\hbar c\sim 1/137$), n is the refractive index of the material, $\mathcal{E}_0$ is the applied field and $\mathcal{E}_{\text{eff}}$ is the effective field at the impurity site, E is the photon energy, E_i and E_f are the initial- and final-state energies, respectively, and ψ_i and ψ_f are the the initial- and final-state wave functions, respectively. For the calculation of cross sections one therefore needs to know initial- and final-state energies and wave functions, as well as $\mathcal{E}_{\text{eff}}$. The latter is very hard to calculate, and the ratio $\mathcal{E}_{\text{eff}}/\mathcal{E}_0$ has generally been treated as an adjustable parameter for absolute values of σ. [It clearly does not affect the shape of $\sigma(E)$.]

The simplest calculations have been carried out in the effective-mass approximation (EMA). From the discussion of Secs. III and VI, the choices are obvious. For donors, the initial state ψ_i is a hydrogenic wave function: the final-state energies are given by the k•p expansion to order k^2, which, in the simplest case, is $\hbar^2k^2/2m^*$. The only unresolved question is the final-state wave function. One possibility is to take continuum solutions of the free hydrogen atom (Whittaker functions). The problem is then entirely isomorphic to the photoinionization of the free hydrogen atom, but the resulting characteristic cross section, which peaks at threshold, does not agree very well with experiment (Burstein, Picus, Henvis, and Wallis, 1965) (Fig. A 1). Alternatively, one can take ψ_f to be Bloch functions and calculate the momentum matrix element in the k · p approximation (Kane, 1956; Kohn, 1957). This calculation gives a cross section which peaks at 1.43 E_H above threshold, where E_H is the hydrogenic binding energy. The experimental curves for the shallow acceptors in Si (Burnstein *et al.*, 1956), however, show that the peak ranges from less than $1.43E_I$ to about $2E_I$, where E_I is the *observed* ionization energy. These shifts of the peak may also be termed chemical shifts, in analogy with the shifts of binding energies from the hydrogenic value.

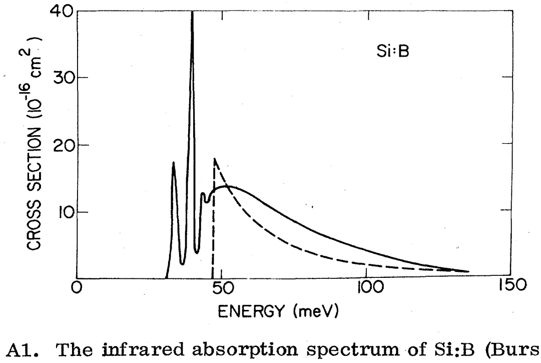

FIG. A1. The infrared absorption spectrum of Si:B (Burstein *et al.*, 1956) and the hydrogenic model showing the peak at threshold.

A number of attempts have been made to go beyond the simple hydrogenic models. Lucovsky (1965) proposed that the impurity potential for deep levels may be better simulated by a δ function instead of a Coulombic potential, as had been done in the theory of the photodissociation of deuterons. The bound-state wave function is then $e^{-\alpha r}/r$, compared with the hydrogenic wave function which is $e^{-\alpha r}$. In the case of a simple parabolic band, α is then given by $(m^*E_I)^{1/2}$ in atomic units. This expression is identical with that for α in a hydrogenic model, in which case E_I is just E_H. The resulting cross section is rather similar to the hydrogenic one, but now the peak is at $2E_I$, in very good agreement with the experimental data for the relatively deep Si-In (Lucovsky, 1965).

A somewhat more elaborate calculation was carried out by Bebb and Chapman (1967) using the quantum-defect model, which is well known from the theory of atomic spectra. In this model, the bound-state wave function is taken to be the solution of the hydrogenic problem at the *observed* ionization energy. Such a wave function is not an eigenstate and therefore diverges at the origin, but without causing any harm. For final-state wave functions Bebb and Chapman used hydrogenic continuum solutions scaled according to the quantum-defect prescription. The quantum-defect parameter ν, which is usually obtained by fitting the excited-states spectrum to an expression of the form $1/(\nu+n)^2$, could not be unambiguously determined, and was thus treated as an adjustable parameter for each of the shallow acceptors in Si. Good agreement with experiment was obtained. A more detailed study of the quantum-defect model was subsequently carried out by Bebb (1969), but no other applications have been reported, perhaps due to the complicated nature of the calculations.

The fact that many deep-level cross sections did not fit very well either the hydrogenic or δ-function pattern led to a variety of alternatives to the above simple models. Grimmeiss and Ledebo (1975) assumed that the bound electron is described by the real mass m_0, instead of the band mass, in determining the wave-function exponent α for the Lucovsky model, and found that they could fit their data on GaAs:O rather well. The use of m_0 for the bound state, however alters the conceptual basis of the model in a dramatic way. While the original Lucovsky model simulated the perturbation potential by a δ function and used effective-mass theory, the Grimmeiss-Ledebo assumption corresponds to simulating the potential of an impurity atom by a δ function and neglecting the result of the crystal entirely. Thus, whereas the Lucovsky models stretches the limits of the EMT with a δ-function potential perturbation, the Grimmeiss–Ledebo model is an extreme-tight-binding model.

Grimmeiss and Ledebo (1975) could not obtain satisfactory fits to data on deep acceptors in Si and attributed the difficulty to the inadequacy of the parabolic-band approximation for the final states. They kept the δ-function-potential wave function and attempted to extract an average isotropic but nonparabolic "effective band" that reproduces experimental data on Si:Au. The "effective band" was moderately successful for other deep acceptors. The recent controversy on the experimental data on Si:Au (see Sec. V), and the possibility that "Si:Au" is actually a complex, as well as the unresolved temperature effects, raise a number of questions about the fit. From a theoretical point of view, the concept of an "effective band" is somewhat unsettling and has limited usefulness.

A rather intriguing calculation has recently been reported by Rynne, Cox, McGuire, and Blakemore (1976). These authors assumed a parabolic band of the form $\hbar^2k^2/2m^*$ for the final-state energies and plane waves for the final-state wave functions, and were able to invert the formula for $\sigma(E)$ and obtain $\psi_i(r)$ explicitly as an integral of $\sigma(E)$. The method is limited by the requirement that the band edge be parabolic and that the cross section must be known over a sufficient range of energies for the integral to be meaningful. Rynne *et al.* (1976) reported results for some acceptors in Si. Smooth, monotonically decreasing wave functions were extracted for Si:In, and for acceptors in Ge, including the relatively deep Ge:Hg. Si:B, which as the shallowest acceptor in Si might be thought to be the most hydrogenic of all, was found to behave strangely: its wave function was found to have a second maximum. An explanation for this peculiar effect has been given by the present author (Pantelides, 1976), who pointed out that the weakest of the assumptions made by Rynne *et al.* is that the valence bands of Si may be simulated by a parabolic band of the form $\hbar^2k^2/2m^*$. A test of this hypothesis was carried out by computing the density of states near the valence band top in terms of the full 6×6 $\mathbf{k}\cdot\mathbf{p}$ matrix discussed in Secs. VII and VIII (Pantelides and Bernholc, 1977). The resulting density of states was then compared with the corresponding density of states of a parabolic band, which has a simple $(m^*)^{3/2}E^{1/2}$ dependence. As Fig. A2 demonstrates, the binding energies of acceptors in Ge correspond to regions which are well simulated by a parabolic band. This is not true for acceptors in Si, where nonparabolic effects are strong, largely due to the presence of the split-off band at only 44 meV below the top. The parabolic approximation is seen to be better justified for the deeper Si:In for which one could in fact set the spin-orbit splitting equal to zero, but fails completely for Si:B.

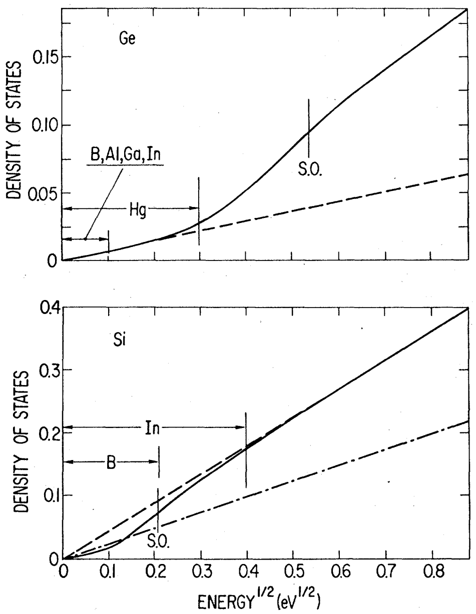

FIG. A2. The density of states near the top of the valence bands for (a) Ge (b) Si, from Pantelides (1976). See discussion in text.

More recently, Pantelides and Bernholc (1977) undertook a systematic study of cross sections in Si. The basic idea was to recognize that from among the three quantities that enter the calculation of σ, namely ψ_f, ψ_i, and ϵ_f, the final-state energies ϵ_f are the only ones that can be calculated very accurately with available band-theoretic techniques. The final-state energies were therefore obtained from diagonalizing the full 6×6 $\mathbf{k}\cdot\mathbf{p}$ matrix discussed earlier, instead of simulating them with a single parabola. The final-state wave functions ψ_f should in principle be products of some form of a Whittaker function times Bloch functions, but were approximated by pure Bloch functions. Similarly, the bound-state wave functions were approximated by linear combinations of Bloch functions, which made the calculation of transition matrix elements possible in terms of $\mathbf{k}\cdot\mathbf{p}$ theory. The Bloch-function coefficients were determined in part by symmetry, but their radial part in $\mathbf{k}$ space was taken to be a simple analytical expression, whose Fourier transform could be viewed as an envelope function with an adjustable radius. Good fits were obtained for a number of shallow and deep acceptors. For the first time, transitions to the split-off band were included and found to be important.

It is apparent from the above discussion that no definitive calculation of cross sections exists to date. The basic problem is that no accurate wave functions are available to make a direct prediction. One usually is forced to extract a wave function from the data. Such wave functions are usually of a simple form and only convey some average information, such as an average radius. As more accurate bound-state calculations are performed one may expect more accurate cross-section calculations will be feasible. As we saw in Sec. IX, Jaros has calculated some cross sections for deep levels, using his numerical wave functions, but temperature broadening still remains a problem which will have to be investigated further in the future. The effect of electron-lattice interactions on the photoionization cross sections has recently been studied at length for the particular case of GaP:O by Monemar and Samuelson (1976) and by Henry and Lang (1977).

Note added in proof: Since the completion of this review paper, fully self-consistent solutions for the single vacancy in Si have been reported independently by Bernholc, Pantelides, and Lipari and by Baraff and Schlüter (Bull. American Physical Society, March 1978). Papers by both groups have been submitted to Phys. Rev. Lett, and will also be presented at the 14th International Conference on the Physics of Semiconductors in Edinburgh, Scotland, Sept. 1978. Both calculations make use of the Koster-Slater Green's function formalism and an LCAO basis set. Another (non-self-consistent) study of Si:V using a reformulation of the BIPJ method (so that it corresponds to a very large cluster) has been reported by U. Lindefelt (J. Phys. C, in press).

REFERENCES

Abarenkov, I. V. and V. Heine, 1965, Philos. Mag. **12**, 529.

Adler, S. L., 1962, Phys. Rev. **126**, 413.

Aggarwal, R. L., 1964, Solid State Commun. **2**, 1963.

Aggarwal, R. L., P. Fisher, V. Mourzine, and A. K. Ramdas, 1965, Phys. Rev. **138**, A882.

Aggarwal, R. L., and A. K. Ramdas, 1965, Phys. Rev. **140**, A1246.

Ahlburn, B. T., and A. K. Ramdas, 1968, Phys. Rev. **167**, 717.

Ahlburn, B. T., and A. K. Ramdas, 1969, Phys. Rev. **187**, 932.

Allen, J. W., 1971, J. Phys. C **4**, 1937.

Altarelli, M., and G. Iadonisi, 1971, Nuovo Cimento B **5**, 21.

Altarelli, M., W. Y. Hsu, and R. A. Sabatini, 1977, J. Phys. C **10**, L605.

Animalu, A. O. E., 1965, Cavendish Laboratory Technical Reports 1-3, unpublished.

Animalu, A. O. E., and V. Heine, 1965, Philos. Mag. **12**, 1249.

Appapillai, M., and V. Heine, 1972, Tech. Rep. No. 5, Solid State Theory Group, Cavendish Laboratory, Cambridge, England, unpublished.

Appelbaum, J. A., and D. R. Hamann, 1974, Phys. Rev. Lett. **32**, 225.

Aten, A. C., J. H. Haanstra, and H. deVries, 1965, Philips Res. Rep. **20**, 395.

Austin, B. J., V. Heine, and L. J. Sham, 1962, Phys. Rev. **127**, 276.

Bajaj, K. K., 1970, Solid State Commun. **8**, 1423.

Bajaj, K. K., 1972, in *Polarons in Ionic Crystals and Polar Semiconductors*, edited by J. T. Devreese (North-Holland, Amsterdam), p. 193.

Baldereschi, A., 1970, Phys. Rev. B **1**, 4673.

Baldereschi, A., 1973, J. Lumin. **7**, 79.

Baldereschi, A., and J. J. Hopfield, 1972, Phys. Rev. Lett. **28**, 171.

Baldereschi, A., and N. O. Lipari, 1973, Phys. Rev. B **8**, 2697.

Baldereschi, A., and N. O. Lipari, 1974, Phys. Rev. B **9**, 1525.

Baldereschi, A., and N. O. Lipari, 1976, *Proceedings of the 13th International Conference on the Physics of Semiconductors*, edited by F. G. Fumi (Tipografia Marves, Rome), p. 595.

Bardeen, J., and W. H. Brattain, 1948, Phys. Rev. **74**, 230.

Bardeen, J., and W. H. Brattain, 1949, Phys. Rev. **75**, 1208.

Bassani, F., and V. Celli, 1961, J. Phys. Chem. Solids **20**, 64.

Bassani, F., and G. Iadonisi, and B. Preziosi, 1969, Phys. Rev. **186**, 735.

Bassani, F., G. Iadonisi, and B. Preziosi, 1974, Rep. Prog. Phys. **37**, 1099.

Bassani, F., and G. Pastori Parravicini, 1975, *Electronic States and Optical Transitions in Solids* (Pergamon, Oxford).

Bebb, H. B., 1969, Phys. Rev. **185**, 1116.

Bebb, H. B., and R. A. Chapman, 1967, J. Phys. Chem. Solids **28**, 2087.

Bernholc, J., and S. T. Pantelides, 1977, Phys. Rev. B **15**, 4935.

Bernholc, J., and S. T. Pantelides, 1978, Phys. Rev. B **18**, Aug. 15, 1978.

Bernholc, J., S. T. Pantelides, and N. O. Lipari, 1978, to be published.

Bethe, H. A., 1942, "Theory of the Boundary Layer of Crystal Rectifiers," MIT Radiation Laboratory Report 43-12.

Bloch, F., 1928, Z. Phys. **52**, 555.

Bloch, F., 1930, Z. Phys. **59**, 208.

Blyholder, G., and C. A. Coulson, 1968, Theor. Chim. Acta **10**, 316.

Boer, F. P., M. D. Newton, and W. N. Lipscomb, 1964, Proc. Natl. Acad. Sci. USA **52**. 890.

Bourgoin, J., and J. W. Corbett, 1972, Phys. Lett. A **38**, 135.

Braun, F., 1874, Ann. der Physik und Chemie **153**, 556.

Breitenecker, M., R. Sexl, and W. Thirring, 1964, Z. Phys. **182**, 123.

Brillouin, L., 1930, J. Phys. (Paris) **9**, 337.

Brooks, H., 1951, Phys. Rev. **83**, 879.

Brooks, H., 1955, in *Advances in Electronics and Electron Physics*, edited by L. Marton (Academic, New York), Vol. 7, p. 85.

Brust, D., 1964, Phys. Rev. **134**, A1337.

Bube, R. H., 1960, *Photoconductivity of Solids* (Wiley, New York).

Burstein, E., E. E. Bell, and J. W. Davisson, 1953, J. Phys. Chem. **57**, 849.

Burstein, E., J. J. Oberly, and J. W. Davisson, 1953, Phys. Rev. **89**, 331.

Burstein, E., J. J. Oberly, J. W. Davisson, and B. W. Henvis, 1951, Phys. Rev. **82**, 764.

Burstein, E., G. S. Picus, B. Henvis, and R. Wallis, 1956, J. Phys. Chem. Solids **1**, 65.

Callaway, J., 1964, J. Math. Phys. **5**, 783.

Callaway, J., 1967, Phys. Rev. **154**, 515.

Callaway, J., 1971, Phys. Rev. B **3**, 2556 (1971).

Callaway, J., and A. J. Hughes, 1967a, Phys. Rev. **156**, 860.

Callaway, J., and A. J. Hughes, 1967b, Phys. Rev. **164**, 1043.

Cardona, M., W. Paul, and H. Brooks, 1959, J. Phys. Chem. Solids **8**, 204.

Carter, A. C., P. J. Dean, M. S. Skolnick, and R. A. Stradling, 1977, J. Phys. C **10**, 5111.

Cartling, B., 1975, J. Phys. C **8**, 3183.

Cartling, B. G., B. Roos, and U. Wahlgren, 1974, Chem. Phys. Lett. **32**, 1244.

Chadi, D. J., 1977, Phys. Rev. B **16**, 790.

Chadi, D. J., and M. L. Cohen, 1975, Phys. Status Solidi B **68**, 405.

Chaney, R. C., 1976, Phys. Rev. B **14**, 4578.

Chaney, R. C., and C. C. Lin, 1976, Phys. Rev. B **13**, 843.

Chase, L. L., W. Hayes, and J. F. Ryan, 1977, J. Phys. C **10**, 2957.

Chelikowsky, J. R., and M. Schlüter, 1977, Phys. Rev. B **15**, 4020.

Chui, S. T., C. Weigel, and J. W. Corbett, 1977, Bull. Am. Phys. Soc. **22**, 267.

Clark, D. T., 1968, Tetrahedron **24**, 2663.

Cohen, E., and M. D. Sturge, 1977, Phys. Rev. B **15**, 1039.

Cohen, M. H., and V. Heine, 1961, Phys. Rev. **122**, 1821.

Cohen, M. L., and T. K. Bergstresser, 1966, Phys. Rev. **141**, 789.

Cohen, M. L., M. Schlüter, J. R. Chelikowsky, and S. G. Louie, 1975, Phys. Rev. B **12**, 5575.

Conwell, E., and V. F. Weisskopf, 1950, Phys. Rev. **77**, 388.

Corbett, J. W., 1964, Solid State Phys., Suppl. **7**.

Coulson, C. A., 1972, in *Radiation Damage and Defects in Semiconductors*, Conference Series No. 16, edited by J. E. Whitehouse (Institute of Physics, London), p. 249.

Coulson, C. A., and Mary J. Kearsley, 1957, Proc. R. Soc.

A **241**, 433.

Coulson, C. A., and F. P. Larkins, 1969, J. Phys. Chem. Solids **30**, 1963.

Coulson, C. A., and F. P. Larkins, 1971, J. Phys. Chem. Solids **32**, 2245.

Crawford, J. H., Jr., and L. M. Slifkin, 1975, editors, *Point Defects in Solids*, Volume 2 (Plenum, New York).

Csavinszky, P., 1963, J. Phys. Chem. Solids **24**, 1003.

Csavinszky, P., 1965, J. Phys. Soc. Jpn. **20**, 2027.

Dean, P. J., 1968, Trans. Met. Soc. of AIME, **242**, 1384.

Dean, P. J., 1973, Prog. Solid State Chem. **8**, 1.

Dean, P. J., J. D. Cuthbert, and R. T. Lynch, 1969, Phys. Rev. **179**, 754.

Dean, P. J., J. D. Cuthbert, G. D. Thomas, and R. T. Lynch, 1967, Phys. Rev. Lett. **18**, 122.

Dean, P. J., R. A. Faulkner, S. Kimura, and M. Ilegems, 1971, Phys. Rev. B **4**, 1926.

Dean, P. J., C. J. Frosch, and C. H. Henry, 1968, J. Appl. Phys. **39**, 5631.

Dean, P. J., and C. H. Henry, 1968, Phys. Rev. **176**, 928.

Dean, P. J., and D. C. Herbert, 1976, J. Lumin. **14**, 55.

deBoer, J. H., and W. C. vanGeel, 1935, Physica **2**, 286.

deKock, A. J. R., 1973, Philips Res. Rep. Suppl. **1**, 1.

Devreese, J. T., 1972, editor, *Polarons in Ionic Crystals and Polar Semiconductors* (North-Holland, Amsterdam).

Djafari-Rouhani, M., M. Lannoo, and P. Lenglart, 1970, J. Phys. **31**, 597.

Dresselhaus, G., A. F. Kip, and C. Kittel, 1955, Phys. Rev. **98**, 368.

Edmonds, A. R., 1960, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, N.J.).

Engineer, M., and N. Tzoar, 1972, in *Polarons in Ionic Crystals and Polar Semiconductors*, edited by J. T. Devreese (North-Holland, Amsterdam), p. 747.

Engström, O., and H. G. Grimmeiss, 1975, J. Appl. Phys. **46**, 831.

Fagelström, P. O., and H. G. Grimmeiss, 1977, to be published.

Faulkner, R. A., 1968, Phys. Rev. **175**, 991.

Fetterman, H. R., D. M. Larsen, G. E. Stillman, P. E. Tannewald, and J. Waldman, 1971, Phys. Rev. Lett. **26**, 975.

Fisher, P., and A. K. Ramdas, 1965, Phys. Lett. **16**, 26.

Flynn, C. P., 1972, *Point Defects and Diffusion* (Clarendon, Oxford).

Fowler, W. B., 1968, editor, *Physics of Color Centers* (Academic, New York).

Friedel, J., M. Lannoo, and G. Leman, 1967, Phys. Rev. **164**, 1056.

Fritzsche, H., 1962, Phys. Rev. **125**, 1560.

Frölich, H., 1937, Proc. R. Soc. A **160**, 230.

Frölich, H., 1962, in *Polarons and Excitons*, edited by G. G. Kuper and G. D. Whitfield (Plenum, New York), p. 1.

Garcia-Moliner, F., 1971, in *Theory of Imperfect Crystalline Solids*, Trieste Lectures, 1970 (International Atomic Energy Agency, Vienna), p. 1.

Gilbert, T. L., 1970, in *Sigma Molecular Orbital Theory*, edited by O. Sinanoglu and K. Wiberg (Yale University, New Haven, Conn.), p. 249.

Glodeanu, A., 1965a, Rev. Roum. Phys. **10**, 433.

Glodeanu, A., 1965b, Rev. Roum. Phys. **10**, 741.

Glodeanu, A., 1968, Phys. Lett. **A28**, 404.

Glodeanu, A., 1969a, Rev. Roum. Phys. **14**, 139.

Glodeanu, A., 1969b, Phys. Status Solidi **35**, 481.

Gombás, P., 1956, in *Encyclopedia of Physics*, edited by S. Flügge (Springer, Berlin), Vol. XXXVI, p. 109.

Grimmeiss, G. H., 1977, Ann. Rev. Mater. Sci. **7**, 341.

Grimmeiss, H. G., and L.-Å. Ledebo, 1975, J. Phys. C **8**, 2615.

Gudden, B., 1931, Ber. Phys. Soz. Erlangen **62**, 289.

Gummel, H., and M. Lax, 1955, Phys. Rev. **97**, 1469.

Gunnarson, O., J. Harris, and R. O. Jones, 1977, Phys. Rev. B **15**, 3027.

Haller, E. E., and W. L. Hansen, 1974, Solid State Commun. **15**, 687.

Harrison, W. A., 1966, *Pseudopotentials in the Theory of Metals* (Benjamin, New York).

Haug, A., 1970, Z. Naturforsch. **25a**, 143.

Haydock, R., V. Heine, and M. J. Kelly, 1972, J. Phys. C **5**, 2845.

Haydock, R., V. Heine, and M. J. Kelly, 1975, J. Phys. C **8**, 2591.

Haynes, J. R., and W. C. Westphal, 1956, Phys. Rev. **101**, 1676.

Hedin, L., and B. I. Lundqvist, 1971, J. Phys. C **4**, 2064.

Hemstreet, L. A., 1977, Phys. Rev. B **15**, 834.

Henry, C. H., J. J. Hopfield, and L. C. Luther, 1966, Phys. Rev. Lett. **17**, 1178.

Henry, C. H., and D. V. Lang, 1977, Phys. Rev. B **15**, 989.

Hensel, J. C., H. Hasegawa, and M. Nakayama, 1965, Phys. Rev. **138**, A225.

Hermanson, J., and J. C. Phillips, 1966, Phys. Rev. **150**, 652.

Hermanson, J., and J. C. Phillips, 1940, Phys. Rev. **57**, 1169.

Herring, C., 1954, in *Photoconductivity Conference*, Atlantic City, 1954, edited by R. C. Breckenridge, B. R. Russel, and E. E. Hahn (Wiley, New York), p. 81.

Ho, L. T., and A. K. Ramdas, 1972, Phys. Rev. B **5**, 462.

Hoffmann, R., 1963, J. Chem. Phys. **39**, 1397.

Hohenberg, P., and W. Kohn, 1964, Phys. Rev. **136**, 864.

Hopfield, J. J., G. D. Thomas, and M. Gershenzon, 1963, Phys. Rev. Lett. **10**, 162.

Hsu, W. Y., J. D. Dow, D. J. Wolford, and B. G. Streetman, 1977, Phys. Rev. B **16**, 1597.

Huang, K., and A. Rhys, 1950, Proc. R. Soc. A **204**, 406.

Ivey, J. L., and R. L. Mieher, 1975a, Phys. Rev. B **11**, 822.

Ivey, J. L., and R. L. Mieher, 1975b, Phys. Rev. B **11**, 849.

James, H. M., 1949, Phys. Rev. **76**, 1602.

Jaros, M., 1969, Phys. Status Solidi **36**, 181.

Jaros, M., 1971, J. Phys. C **4**, 1162.

Jaros, M., 1975, J. Phys. C **8**, 2455.

Jaros, M., and S. Brand, 1976, Phys. Rev. B **14**, 4494.

Jaros, M., and S. F. Ross, 1973, J. Phys. **6**, 1753.

Jaros, M., and S. F. Ross, 1974, Solid State Commun. **14**, 631.

Joannopoulos, J. D., and E. J. Mele, 1976, Solid State Commun. **20**, 729.

Johnson, K. H., J. G. Norman, and J. W. D. Connolly, 1973, in *Computational Methods for Large Molecules and Localized States in Solids*, edited by F. Herman, A. D. McLean, and R. K. Nesbet (Plenum, New York), p. 161.

Johnson, K. H., and F. C. Smith, 1972, Phys. Rev. B **5**, 831.

Jones, R. L., and P. Fisher, 1965, J. Phys. Chem. Solids **26**, 1125.

Kaczmarek, E., 1966, Acta Phys. Pol. **30**, 267.

Kane, E. O., 1956, J. Phys. Chem. Solids **1**, 82.

Kane, E. O., 1957, J. Phys. Chem. Solids **1**, 249.

Kauffer, E., P. Pecheur, and M. Gerl, 1976, J. Phys. C **9**, 2319.

Kauffer, E., P. Pecheur, and M. Gerl, 1977, Phys. Rev. B **15**, 4107.

Kaus, P., 1958, Phys. Rev. **109**, 1944.

Keldysh, L. V., 1963, Zh. Eksp. Teor. Fiz. 45, 364 [Sov. Phys.-JETP **18**, 253].

Kimerling, L. C., 1977, in *Radiation Effects in Semiconductors, 1976*, Conference Series No. 31, edited by N. B. Urli and J. W. Corbett (Institute of Physics, London), p. 221.

Kishino, S., N. Shinone, H. Nakashima, and R. Ito, 1976, Appl. Phys. Lett. **29**, 488.

Kittel, C., 1963, *Quantum Theory of Solids* (Wiley, New York).

Kittel, C., 1976, *Introduction to Solid State Physics*, 5th edition (Wiley, New York).

Kittel, C., and A. H. Mitchell, 1954, Phys. Rev. **96**, 1488.

Kleiman, G., 1977, Phys. Rev. B **15**, 802.

Kleiner, W. H., and W. E. Kragg, 1970, Phys. Rev. Lett. **25**, 1490.

Kleinman, L., and J. C. Phillips, 1959, Phys. Rev. **116**, 287.

Kogan, S. M., and B. I. Segunov, 1967, Soviet Phys.-Solid State **8**, 1898.

Kohn, W., 1957, Solid State Phys. **5**, 257.

Kohn, W., and J. M. Luttinger, 1955a, Phys. Rev. **97**, 883.

Kohn, W., and J. M. Luttinger, 1955b, Phys. Rev. **97**, 1721.

Kohn, W., and J. M. Luttinger, 1955c, Phys. Rev. **98**, 915.

Kohn, W., and N. Rostocker, 1954, Phys. Rev. **94**, 1111.

Kohn, W., and D. Schechter, 1955, Phys. Rev. **99**, 1903.

Kohn, W., and L. J. Sham, 1965, Phys. Rev. **140**, A1133.

Korringa, J., 1947, Physica **13**, 392.

Kosicki, B. B., and W. Paul, 1966, Phys. Rev. Lett. **17**, 246.

Kosicki, B. B., W. Paul, A. J. Strauss, and G. W. Iseler, 1966, Phys. Rev. Lett. **17**, 1175.

Koster, G. F., 1954, Phys. Rev. **95**, 1436.

Koster, G. F., and J. C. Slater, 1954a, Phys. Rev. **95**, 1167.

Koster, G. F., and J. C. Slater, 1954b, Phys. Rev. **96**, 1208.

Kovarskii, V. A., 1962, Fiz. Tverd. Tela **4**, 1636 [Sov. Phys.-Solid State **4**, 1200].

Kovarskii, V. A., and E. P. Sinyavskii, 1962, Fiz. Tverd. Tela **4**, 3202 [Sov. Phys.-Solid State **4**, 2345].

Kovarskii, V. A., and E. P. Sinyavskii, 1964, Fiz. Tverd. Tela **6**, 636 [Sov. Phys.-Solid State **6**, 498].

Krag, W. E., and H. J. Zeiger, 1962, Phys. Rev. Lett. **8**, 485.

Krag, W. E., W. H. Kleiner, H. J. Zieger, and S. Fischler, 1966, J. Phys. Soc. Jpn. Suppl. **21**, 230.

Kronig, R., and W. G. Penney, 1931, Proc. R. Soc. A **130**, 499.

Kubo, R., and I. Toyozawa, 1955, Prog. Theor. Phys. **13**, 160.

Kukimoto, H., C. H. Henry, and F. R. Merritt, 1973, Phys. Rev. B **7**, 2486.

Kuper, C. G., and G. D. Whitefield, 1962, editors, *Polarons and Excitons* (Plenum, New York).

Lang, D. V., 1974, J. Appl. Phys. **45**, 3023.

Lang, D. V., 1977, unpublished.

Lang, N. D., and A. R. Williams, 1976, Phys. Rev. Lett. **37**, 212.

Lannoo, M., and P. Lenglart, 1969, J. Phys. Chem. Solids **30**, 2409.

Larkins, F. P., 1971, J. Phys. C **4**, 3065.

Larkins, F. P., 1971, J. Phys. C **4**, 3077.

Larsen, D. M., 1972, in *Polarons in Ionic Crystals and Polar Semiconductors*, edited by J. T. Devreese (North-Holland, Amsterdam), p. 237.

Lawaetz, P., 1971, Phys. Rev. B **4**, 3460.

Lax, M., 1952, J. Chem. Phys. **20**, 1752.

Lax, M., 1960, Phys. Rev. **119**, 1502.

Lee, T. F., and T. C. McGrill, 1973, J. Phys. C **6**, 3438.

Levinger, B. W., and D. R. Frankl, 1961, J. Phys. Chem. Solids **20**, 291.

Lifshitz, J. M., and M. I. Kaganov, 1959, Usp. Fiz. Nauk **69**, 419 [1960, Sov. Phys.-Usp. **2**, 831].

Lifshitz, J. M., and F. Ya Nad, 1965, Sov. Phys.-Dokl. **10**, 532.

Lipari, N. O., and A. Baldereschi, 1970, Phys. Rev. Lett. **25**, 1660.

Lipari, N. O., and A. Baldereschi, 1978, Solid State Commun. **25**, 665.

Louie, S. G., M. Schlüter, J. R. Chelikowsky, and M. L. Cohen, 1976, Phys. Rev. B **13**, 1654.

Lowther, J. E., 1977, Phys. Rev. B **15**, 3928.

Lucovsky, G., 1965, Solid State Commun. **3**, 299.

Luttinger, J. M., 1956, Phys. Rev. **102**, 1030.

Luttinger, J. M., and W. Kohn, 1955, Phys. Rev. **97**, 969.

Manchon, D. D., and P. J. Dean, 1970, in *The Proceedings of the 10th International Conference on the Physics of Semiconductors* (U.S. AEC, Oak Ridge), p. 760.

Mendelson, K. S., and H. M. James, 1964, J. Phys. Chem. Solids **25**, 729.

Mendelson, K. S., and D. R. Schultz, 1969, Phys. Status Solidi **31**, 59.

Menzel, W. P., K. Mednick, C. C. Lin, and C. F. Dorman, J. Chem. Phys. **63**, 4708.

Merzbacher, E., 1967, *Quantum Mechanics* (Wiley, New York).

Messmer, R. P., 1971, Chem. Phys. Lett. **11**, 589.

Messmer, R. P., and G. D. Watkins, 1970, Phys. Rev. Lett. **25**, 656.

Messmer, R. P., and G. D. Watkins, 1972, in *Radiation Damage and Defects in Semiconductors*, Conference Series No. 16, edited by J. E. Whitehouse (Institute of Physics, London), p. 255.

Messmer, R. P., and G. D. Watkins, 1973, Phys. Rev. B **7**, 2568.

Miller, G. L., D. V. Lang, and L. C. Kimerling, 1977, Annu. Rev. Mater. Sci. **7**, 377.

Milnes, A. G., 1973, Deep Impurities in Semiconductors (Wiley, New York).

Monemar, B., and L. Samuelson, 1976, J. Lumin. **12/13**, 507.

Morgan, T. N., 1970, *Proceedings of the Tenth International Conference on the Physics of Semiconductors*, edited by S. P. Keller, J. C. Hensel, and F. Stern (U.S. AEC, Oak Ridge), p. 266.

Morgan, T. N., 1975, J. of Electronic Materials **4**, 1029.

Morgan, T. N., 1978, Phys. Rev. Lett. **40**, 190.

Morita, A., and H. Nara, 1966, J. Phys. Soc. Jpn. Suppl. **21**, 234.

Morse, P. M., 1930, Phys. Rev. **35**, 1310.

Moruzzi, V. L., A. R. Williams, and J. F. Janak, 1977, Phys. Rev. B **15**, 2854.

Mott, N. F., and R. W. Gurney, 1940, *Electronic Processes in Ionic Crystals* (Clarendon, Oxford).

Müller, A. M. K., 1964, Solid State Commun. **2**, 205.

Müller, A. M. K., 1965, Z. Naturforsch. **20a**, 1476.

Mulliken, R. S., 1949a, J. Chim. Phys. **46**, 497.

Mulliken, R. S., 1949b, J. Chim. Phys. **46**, 675.

Naber, J. A., C. E. Mallon, and R. E. Leadon, 1972, in *Radiation Damage and Defects in Semiconductors*, Conference Series No. 16, edited by J. E. Whitehouse (Institute of Physics, London), p. 26.

Nara, H., 1965, J. Phys. Soc. Jpn. **20**, 778.

Nara, H., and A. Morita, 1967, J. Phys. Soc. Jpn. **23**, 831.

Newton, M. D., F. P. Boer, and W. N. Lopscomb, 1966, J. Am. Chem. Soc. **88**, 2353.

Ning, T. H., and C. T. Sah, 1971a, Phys. Rev. B **4**, 3468.

Ning, T. H., 1971b, Phys. Rev. B **4**, 3482.

Onton, A., 1971, Phys. Rev. B **4**, 4449.

Onton, A., 1969, Phys. Rev. **186**, 786.

Onton, A., P. Fisher, and A. K. Ramdas, 1967, Phys. Rev. **163**, 686.

Onton, A., and R. C. Taylor, 1970, Phys. Rev. B **1**, 2587.

Pandey, K. C., and J. C. Phillips, 1974a, Solid State Commun. **14**, 439.

Pandey, K. C., and J. C. Phillips, 1974b, Phys. Rev. Lett. **32**, 1433.

Pandey, K. C., and J. C. Phillips, 1976, Phys. Rev. B **13**, 750.

Pantelides, S. T., 1973, doctoral thesis, Solid State Electronics Lab. T. R. No. 24, University of Illinois (unpublished).

Pantelides, S. T., 1974a, in *Proceedings of the 12th International Conference on the Physics of Semiconductors*, edited by M. H. Pilkuhn (Teubnes, Stuttgart), p. 396.

Pantelides, S. T., 1974b, Solid State Commun. **14**, 1255.

Pantelides, S. T., 1975, in *Festkörperprobleme*, edited by H. J. Queisser (Pergamon/Vieweg, Braunschweig), Vol. XV, p. 149.

Pantelides, S. T., 1976, unpublished.

Pantelides, S. T., 1978, to be published.

Pantelides, S. T., and J. Bernholc, 1977, in *Radiation Effects in Semiconductors 1976*, Conference Series No. 31, edited by N. B. Urli and J. W. Corbett (Institute of Physics, London), p. 465.

Pantelides, S. T., and W. A. Harrison, 1975, Phys. Rev. B **11**, 3006.

Pantelides, S. T., and C. T. Sah, 1972, Solid State Commun. **11**, 1713.

Pantelides, S. T., and C. T. Sah, 1974a, Phys. Rev. B **10**, 621.

Pantelides, S. T., and C. T. Sah, 1974b, Phys. Rev. B **10**, 638.

Peierls, R., 1930, Ann. Phys. **4**, 121.

Penn, D. R., 1962, Phys. Rev. **128**, 2093.

Petroff, P. M., and R. L. Hartmann, 1973, Appl. Phys. Lett. **23**, 469.

Phillips, J. C., 1970, Phys. Rev. B **1**, 1540.

Phillips, J. C., 1973, *Bonds and Bands in Semiconductors* (Academic, New York).

Phillips, J. C., 1974, Surf. Sci. **44**, 290.

Phillips, J. C., and L. Kleinman, 1959, Phys. Rev. **116**, 287.

Platzman, P. M., 1962, Phys. Rev. **125**, 1961.

Pollmann, J., and H. Büttner, 1975, Solid State Commun. **17**, 1171.

Pollmann, J., and S. T. Pantelides, 1978, Phys. Rev. B, to be published.

Pople, J. A., and G. A. Segal, 1965, J. Chem. Phys. **43**, S136.

Pople, J. A., and D. L. Beveridge, 1970, *Approximate Molecular-Orbital Theory* (McGraw-Hill, New York).

Preziosi, B., 1971, Nuovo Cimento **B6**, 131.

Queisser, H. J., 1971, in *Festkörperprobleme*, edited by O. Madelung (Pergamon/Vieweg, Braunschweig), Vol. XI, p. 45.

Rebane, K. K., 1970, *Impurity Spectra of Solids* (Plenum, 1970).

Reiss, H., 1956, J. Chem. Phys. **25**, 681.

Reitz, J. R., 1955, Solid State Phys. **1**, 1.

Resta, R., 1977, J. Phys. C **10**, L179.

Reuszer, J. H., and P. Fisher, 1964, Phys. Rev. **135**, A1125.

Roitsin, A. B., 1974, Fiz. Tekh. Poluprovodn. **8**, 3 (Sov. Phys.-Semicond. **8**, 1).

Ross, S. F., and M. Jaros, 1973, Solid State Commun. **13**, 1751.

Ross, S. F., and M. Jaros, 1974, J. Phys. C **7**, L235.

Rynne, E. F., J. R. Cox, J. B. McGuire, and J. S. Blakemore, 1976, Phys. Rev. Lett. **36**, 155.

Sah, C. T., 1976, Solid-State Electron. **19**, 975.

Sah, C. T., 1977a, IEEE Trans. Electron Devices **ED-24**, p.

Sah, C. T., 1977b, *Proceedings of the Third International Symposium on Silicon Materials, Science, and Technology*, edited by H. R. Huff and E. Sirtl (Electrochemical Society, Princeotn), Vol. 77, p. 868.

Sah, C. T., W. W. Chan, H. S. Fu, and J. W. Walker, 1972, Appl. Phys. Lett. **20**, 193.

Sah, C. T., L. Forbes, L. L. Rosier, and A. F. Tasch, 1970, Solid-State Electron. **13**, 759.

Sah, C. T., and C. T. Wang, 1975, J. Appl. Phys. **46**, 1767.

Sak, J., 1971, Phys. Rev. B **3**, 3356.

Samuelson, L., and B. Monemar, 1977, to be published.

Schlüter, M., J. R. Chelikowsky, S. G. Louie, and M. L. Cohen, 1975a, Phys. Rev. Lett. **34**, 1385.

Schlüter, M., J. R. Chelikowsky, S. G. Louie, and M. L. Cohen, 1975b, Phys. Rev. B **12**, 4200.

Schultz, T. D., 1962, in *Polarons and Excitons*, edited by G. G. Kuper and G. D. Whitfield (Plenum, New York), p. 71.

Schechter, D., 1962, J. Phys. Chem. Solids **23**, 237.

Schechter, D., 1969, J. Phys. Soc. Jpn. **26**, 8.

Seeger, A., and K. P. Chick, 1969, Phys. Status Solidi **29**, 455.

Shaffer, J. C., and F. Williams, 1964, *Proceedings of the International Conference on the Physics of Semiconductors* (Dunod, Paris), p. 811.

Shaffer, J., and F. Williams, 1970, Phys. Stat. Sol. **38**, 657.

Shimizu, T., and K. Minami, 1971, Phys. Status Solidi B **48**, K181.

Shindo, K., and H. Nara, 1976, J. Phys. Soc. Jpn. **40**, 1640.

Shinohara, S., 1961, Nuovo Cimento **22**, 18.

Shive, J. N., 1959, *Semiconductor Devices* (Van Nostrand, New York).

Shockley, W., 1949, Bell Syst. Tech. J. **28**, 453.

Shockley, W., 1950, *Electrons and Holes in Semiconductors* (Van Nostrand, New York).

Skolnick, M. S., L. Eaves, R. A. Stradling, J. C. Portal, and S. Askenazy, 1974, Solid State Commun. **15**, 1403.

Slater, J. C., 1930, Phys. Rev. **36**, 57.

Slater, J. C., 1949, Phys. Rev. **76**, 1592.

Slater, J. C., 1951, Phys. Rev. **81**, 385.

Slater, J. C., 1972, in *Advances in Quantum Chemistry*, edited by P.-O. Löwdin (Academic, New York), Vol. 6.

Slater, J. C., and K. H. Johnson, 1972, Phys. Rev. B **5**, 844.

Smith, R. A., 1959, *Semiconductors* (Cambridge University, Cambridge, England).

Sommerfeld, A. J. W., 1928, Z. Phys. **47**, 1.

Stoneham, 1970, J. Phys. C **3**, L131.

Stoneham, A. M., 1975, *Theory of Defects in Solids* (Clarendon, Oxford).

Street, R. A., and W. Senske, 1976, Phys. Rev. Lett. **37**, 1292.

Summers, C. J., R. Dingle, and D. E. Hill, 1970, Phys. Rev. B **1**, 1603.

Suzuki, K., M. Okazaki, and H. Hasegawa, 1964, J. Phys. Soc. Jpn. **19**, 930.

Swalin, R. A., 1961, J. Phys. Chem. Solids **18**, 290.

Thomas, D. G., J. J. Hopfield, and C. J. Frosch, 1965, Phys. Rev. Lett. **857**.

Thomas, D. G., J. J. Hopfield, and R. T. Lynch, 1966, Phys. Rev. Lett. **17**, 312.

Torrey, H. C., and C. A. Whitmer, 1948, *Crystal Rectifiers* (McGraw-Hill, New York).

Tossell, J. A., D. J. Vaughn, and K. H. Johnson, 1973, Chem. Phys. Lett. **20**, 329.

Van Vechten, J. A., 1977, *Radiation Effects in Semiconductors*, edited by N. B. Urli and J. W. Corbett, Conference Series No. 31 (Institute of Physics, London), p. 441.

Van Vechten, J. A., and C. T. Thurmond, 1976, Phys. Rev. B **14**, 3539.

Vul, A. Ya., G. L. Bir, and Ya. V. Shmartsev, 1970, Fiz-Tekh. Poluprovodn. **4**, 2331 [Sov. Phys.-Semicond. **4**, 2005].

Walter, W., and J. L. Birman, in *II-VI Semiconducting Compounds*, edited by D. G. Thomas (Benjamin, New York), p. 89.

Wannier, G., 1937, Phys. Rev. **52**, 191.

Watkins, G. D., 1963, J. Phys. Soc. Jpn. **18**, Suppl. II, 22.

Watkins, G. D., 1965, in *Proceedings of the Seventh International Conference on the Physics of Semiconductors, Paris, 1964*, edited by M. Hulin (Academic, New York), Vol. 3, p. 97.

Watkins, G. D., and J. W. Corbett, 1961, Phys. Rev. **121**, 1001.

Watkins, G. D., and R. P. Messmer, 1973, in *Computational Methods for Large Molecules and Localized States in Solids*, edited by F. Herman, A. D. McLean, and R. K. Nesbet (Plenum, New York), p. 133.

Watkins, G. D., and R. P. Messmer, 1974, Phys. Rev. Lett. **32**, 1244.

Watkins, G. D., R. P. Messmer, C. Weigel, D. Peak, and J. W. Corbett, 1971, Phys. Rev. Lett. **27**, 1573.

Watts, R. K., 1977, *Point Defects in Crystals* (Wiley, New York).

Weigel, C., D. Peak, J. W. Corbett, G. D. Watkins, and R. P. Messmer, 1973, Phys. Rev. B **8**, 2906.

Weinreich, G., 1959, J. Phys. Chem. Solids **8**, 216.

White, A. M., P. Porteous, W. F. Sherman, and A. A. Stadtmuller, 1977, J. Phys. C **10**, L473.

Whitehouse, J. E., 1972, editor, *Proceedings of the International Conference on Radiation Damage and Defects in Semiconductors*, Reading, Berkshire, England (Institute of Physics and Physical Society, London).

Williams, A. R., and J. van W. Morgan, 1974, J. Phys. C **7**, 37.

Williams, F., 1960, J. Phys. Chem. Solids **12**, 265.

Williams, F., 1968, Phys. Status Solidi **25**, 493.

Williams, R., 1966, J. Appl. Phys. **37**, 3411.

Wilson, A. H., 1931, Proc. R. Soc. A **133**, 458.

Wilson, A. H., 1932, Proc. R. Soc. A **134**, 277.

Wilson, K., 1975, Rev. Mod. Phys. **47**, 773.

Wiser, N., 1963, Phys. Rev. **129**, 62.

Wright, G. B., and A. Mooradian, 1968, in the *Proceedings of the Ninth International Conference on the Physics of Semiconductors*, Moscow, 1968 (Nauka, Leningrad), Vol. 2, p. 1067.
Yau, L. D., and C. T. Sah, 1974, Solid-State Electron. **17**, 193.
Yamaguchi, T., 1962, J. Phys. Soc. Jpn. **17**, 1359.
Yamaguchi, T., 1963, J. Phys. Soc. Jpn. **18**, 368.
Yi-Hsiang, C., and H. Yu-Ping, 1966, Chin. J. Phys. **22**, 24.
Yip, K. L., 1974, Phys. Status Solidi B **66**, 619.
Yip, K. L., and W. B. Fowler, 1974, Phys. Rev. B **10**, 1400.
Ziman, J. M., 1964, *Principles of the Theory of Solids* (Cambridge University, Cambridge, England).
Zunger, A., and A. J. Freeman, 1977a, Phys. Rev. B **15**, 4716.
Zunger, A., and A. J. Freeman, 1977b, Phys. Rev. B **15**, 5049.
Zunger, A., and A. J. Freeman, 1977c, Phys. Rev. B, to be published.
Zwerdling, S., B. Loux, K. J. Button, and L. M. Roth, 1960, Phys. Rev. Lett. **4**, 173.