PROGRESS IN SEMICONDUCTORS

VOLUME 5

General Editor
ALAN F. GIBSON, B.Sc., Ph.D.

American Editor
Prof. R. E. BURGESS
Vancouver, B.C.

European Editor
Dr. F. A. KRÖGER
Salford

LONDON
HEYWOOD & COMPANY LTD.
1960

PREFACE

If there is any noticeable bias in this, our fifth volume, it is towards the optical properties of semiconductors. In the last two or three years it has been realized that an enormous amount of detailed quantitative information can be obtained by high resolution spectroscopic examination of the absorption spectra of semiconductors, with and without an external magnetic field. The theory and application of these powerful techniques form the subject matter of the papers by Dr. McLean and by Dr. Lax and Mr. Zwerdling. I believe both papers will be very valuable contributions to the literature of our subject, collecting and collating as they do information otherwise rather scattered through the learned journals.

The paper by Dr. Moss on indium antimonide also shows an optical bias as the technological importance of this material rests to a large extent on its photoproperties. Indium antimonide has certainly arrived as a respectable semiconductor and a review of its properties is overdue. We are also very pleased to include a second 'material' paper, that by Professors Haering and Mrozowski on graphite, as it would appear that the properties of this interesting material are not as widely known as they should be. In the more general materials field Drs. Mooser and Pearson contribute a paper on the classification of semiconductors in terms of the chemical bond picture, a subject on which they are the acknowledged authorities.

Our two remaining papers are on the theory of the thermal conductivity of semiconductors and on the electrical properties of semiconductor surfaces. The paper by Dr. Appel on thermal conductivity gives a very complete account of this important aspect of transport phenomena. Reviews of the steady progress being made towards a full understanding of semiconductor surface physics have appeared from time to time but not in our series and nowhere (to the best of my knowledge) recently. I expect that the contribution by Dr. Watkins will be welcomed as both timely and authoritative.

A. F. Gibson

CONTENTS

THE ELECTRICAL PROPERTIES OF SEMICONDUCTOR SURFACES

T. B. WATKINS, M.Sc., D.I.C., Ph.D., A.M.I.E.E.

Mullard Research Laboratories, Salfords, Surrey, U.K.

MS. received December 1959

THE ELECTRICAL PROPERTIES OF SEMICONDUCTOR SURFACES

1. INTRODUCTION

In the bulk of a crystal the atoms are arranged in a regular periodic manner but at the surface the pattern comes to an abrupt end. It must be expected therefore that the energy band structure is modified. This problem was first examined in 1932 by Tamm[1] who considered an idealized semi-infinite crystal to be represented by a terminated series of potential wells and showed that it was possible to have energy levels whose wave functions were localized at the surface of the crystal. Further theoretical studies were made by a number of workers[2–5] and in 1939 Shockley[6] carried out a critical investigation and found the conditions under which surface states are occupied in a normal crystal.

The next major advance took place in 1947 when Bardeen,[7] while studying rectification by semiconductors, developed the very important idea of a double layer or space charge region at a free surface. Any charge trapped in the surface states must be neutralized by an equal and opposite charge within the crystal. He showed that with a relatively high density of surface states a space charge region could extend into the crystal to a depth of 10^{-6} to 10^{-4} cm.

Shortly after this, some evidence indicating the existence of surface states was found during an experiment carried out by Shockley and Pearson.[8] An attempt was made to produce a semiconductor amplifier by modulating the conductance of a slab of semiconductor by using it as one plate of a parallel plate capacitor. The other plate was a metal electrode which, in effect, applied a transverse field to the semiconductor and by this means it was hoped to change the number of free carriers. The change in conductance measured was very much smaller than that expected and this was explained on the basis of the added carriers being trapped in surface states.

Further surface studies on germanium crystals by means of wire probes led to the discovery of transistor action by Bardeen and Brattain in 1948.[9] This stimulated intensive surface studies and very rapid and spectacular advances were made so that by 1955 the basic picture of a germanium surface had emerged and this was described in a review article by Kingston.[10] This picture has been further refined and the work has resulted in a great increase in the fundamental knowledge of the properties of surfaces. The understanding of the behaviour of surfaces in the vicinity of P–N junctions has explained changes of diode and transistor parameters due to surface conditions.

2. THE SURFACE MODEL

Neutrality conditions are maintained in the bulk of a semiconductor as the charge of the impurity atoms is equal and opposite to that of the free carriers. To neutralize any charge in the surface states, therefore, the energy bands must bend at the

surface so that the carrier concentrations are changed to give the required space charge. This is illustrated in Figure 1 for positive and negative charges in the surface states on N- and P-type material. When the surface layer is of a conductivity type different to the bulk there is an inversion layer (Figure 1(a) and (d))

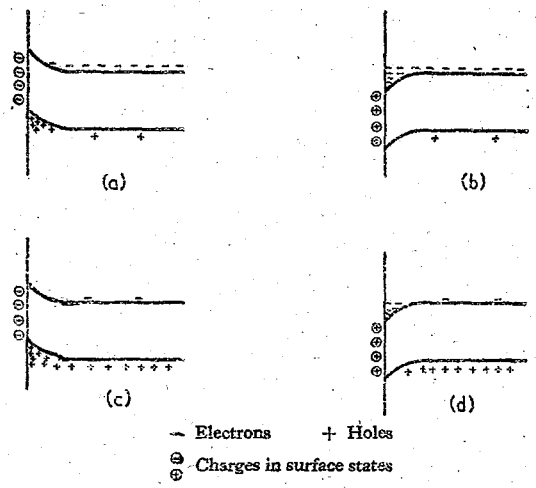


Figure 1. *Surface barriers on semiconductors: (a) Inversion layer on N-type material; (b) Accumulation layer on N-type material; (c) Accumulation layer on P-type material; (d) Inversion layer on P-type material*

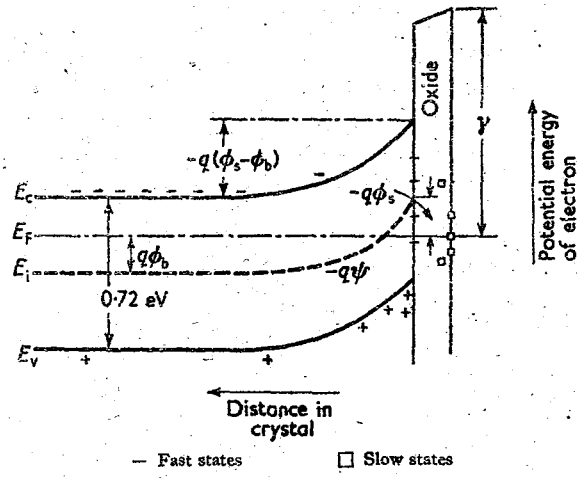


Figure 2. *Model of germanium surface*

and when there is an increase in the concentration of majority carriers at the surface there is an accumulation layer (Figure 1(b) and (c)). A layer intermediate between the two types, in which the total conductance of a specimen is decreased due to the presence of a surface barrier, is known as an exhaustion layer.

We now consider the detailed surface model for germanium as there is greater knowledge of its surface properties than any other material. The model for silicon is believed to be qualitatively similar. The energy diagram of a germanium surface is given in Figure 2.

The symbols shown are those commonly used in surface work and are defined as follows:

E_c = energy of bottom of conduction band
E_v = energy of top of valence band
$E_c - E_v$ = forbidden energy gap (0·72 eV for germanium)
E_F = Fermi level
E_i = intrinsic Fermi level
ψ = electrostatic potential ($-q\psi$, the potential energy of an electron is shown in the Figure). This is arbitrarily taken to be coincident with the intrinsic Fermi level.
$q\phi = E_F + q\psi$, where q is the electronic charge

so that the carrier concentrations of electrons and holes are respectively

$$n = n_i \exp(q\phi/kT)$$

$$p = n_i \exp(-q\phi/kT)$$

where n_i is the intrinsic concentration, k is Boltzmann's constant, and T is the absolute temperature.

The subscripts 'b' and 's' denote values in the bulk and at the surface respectively.

The surface barrier height or surface potential is $+q(\psi_s - \psi_b)$ or $q(\phi_s - \phi_b)$ and thus positive values correspond to the bands bending downwards. Later it will be convenient to use ϕ_s to describe the barrier height, as ϕ_b is fixed by the bulk resistivity. Units of kT will sometimes be used where $kT = 0{\cdot}026$ eV at room temperature. The work function γ is as shown and thus depends on the barrier height. On the germanium surface is an oxide layer which is about 10 Å thick immediately after a normal etching procedure. On standing in air this grows[11] to about 50 Å and the thickness may be greatly increased by various treatments. A potential barrier extends to about 10^{-4} cm into the bulk crystal.

We see that there are three places where charge can reside:

(1) In states which are associated with the oxide layer called the slow states;
(2) In states situated in the forbidden energy gap which are associated with the interface between the oxide and the germanium called the fast states;
(3) In the surface barrier or space charge region extending into the crystal.

Since there can be no net charge on the crystal the sum of these charges must be zero so that the charge in the states is neutralized by the charge in the surface barrier. Letting Q denote the charge per square centimetre of surface area this may be written

$$Q_{ss} + Q_{fs} + Q_{sc} = 0$$

where the subscripts are self-explanatory.

The slow states are so called because their communication with the free carriers is very limited and thus after any disturbance equilibrium with the interior is slowly established. The time constants involved range from milliseconds to minutes and with artificially thickened layers they may extend to many hours. It appears that the great majority of these states lie on the outside of the oxide layer and their numbers and energy levels are controlled by the ambient atmosphere. Their density is of the order of 10^{13} cm^{-2} to 10^{15} cm^{-2}.

The fast states appear to have time constants of the order of 10^{-8} sec and are far less numerous than the slow states, their density being of the order of 10^{11} cm^{-2}. Their energy levels are mainly discrete. Some states have large capture cross-sections for both electrons and holes so that recombination can occur, while it is possible that others may communicate with only one type of carrier.

The slow states are probably due to adsorbed atoms or molecules from the ambient atmosphere on the surface of the oxide layer and one would expect defects in the oxide layer to give states also. The structure of the fast states is not understood but may be due to misfits between the oxide and the germanium crystal lattices.

The type of surface we have considered so far is called a real surface, but recently experimental techniques have been developed to give surfaces with no apparent oxide layer. These are known as clean surfaces and may be produced by argon bombardment or by cleavage. High vacuum conditions are necessary to ensure freedom from contamination while measurements are being made. These clean surfaces are in effect perfect crystal planes and so correspond to the case considered by Tamm.[1] According to his theory the number of states should be of the same order as the number of surface atoms giving a density of about 10^{15} cm^{-2}. As there will be incomplete bonds at the surface the states should be acceptors. Present experimental evidence is not sufficient to give an accurate model for this type of surface but clean surfaces are certainly very highly P-type as would be expected.

3. BASIC THEORETICAL CONSIDERATIONS

As the surface barrier height changes a number of parameters associated with the surface are varied. For example the number of free carriers in the barrier is changed so that the space charge and the conductance of the surface layer are altered. Also the effectiveness of the fast states as recombination centres is changed due to variations in their occupation and in the carrier concentrations at the surface.

Calculations of the variation of surface parameters have been given in a number of papers. Kingston and Neustadter[12] have studied the changes of space charge with barrier height and bulk resistivity. These calculations have been extended to the case of degenerate carrier distributions by Stratton[13] and generalized by Seiwatz and Green.[14] The surface conductance has been investigated by Schrieffer[15,16] and the effect of the surface barrier on the carrier mobility considered. The change of surface recombination with barrier height has been studied theoretically and experimentally by Stevenson and Keyes.[17] Garrett and Brattain[18] have given a comprehensive exposition of the basic physical theory of surfaces. The transient behaviour of the fast states under non-equilibrium conditions has been treated theoretically by Low.[19]

In this section the main results of the basic calculations are presented.

3.1. The Space Charge

A generalized surface model is taken to include non-equilibrium conditions. When a semiconductor is illuminated, excess carriers are present in the bulk and in the surface barrier. The electron and hole concentrations are therefore described by $q\phi_n$ and $q\phi_p$, respectively, which are the differences between quasi-Fermi levels and the intrinsic Fermi level. This case is shown in Figure 3(a). Normally, the electron

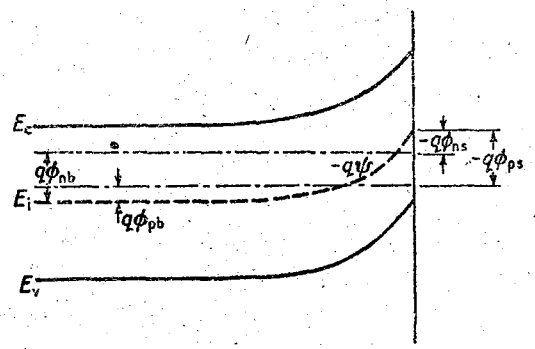


Figure 3(a). *Surface barrier with excess carriers*

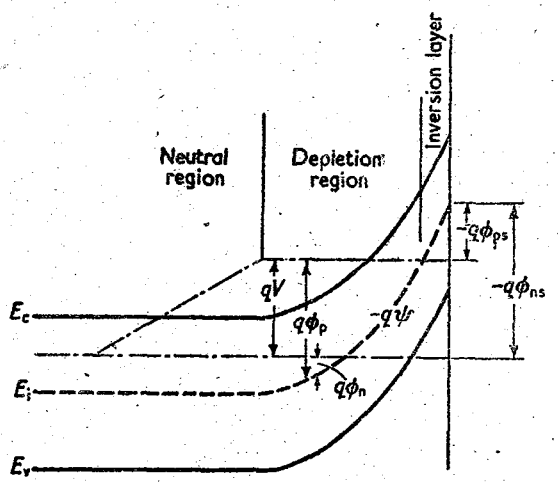


Figure 3(b). *Surface barrier with reverse bias*

and hole currents that are flowing to the surface to recombine are sufficiently small to permit the quasi-Fermi levels to be drawn straight across the barrier region without introducing an appreciable error. In Figure 3(b) the non-equilibrium case where a reverse bias is applied across the surface is shown. As we shall see later this may arise at surfaces in the vicinity of P–N junctions when in effect the surface layer acts as part of the junction. It may be divided into three regions: (1) an inversion layer at the surface, (2) a depletion region where there are ionized

impurities and very few carriers, and (3) a neutral region where the concentrations of the electrons and holes differ from their equilibrium values by equal amounts. The voltage between the quasi-Fermi levels is equal to the bias voltage V.

Consider first Figure 3(a). The charge in the surface barrier may be found by calculating the field at the interface by an integration of Poisson's equation (c.g.s. units)

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = -\frac{4\pi}{\kappa}\sigma$$

where x is the distance measured from the surface into the semiconductor, σ is the charge density at x, and κ is the dielectric constant. If N_D and N_A are the donor and acceptor concentrations

$$\begin{aligned}\sigma &= q[(N_\mathrm{D}-N_\mathrm{A})+(p-n)]\\ &= q[(p-p_\mathrm{b})-(n-n_\mathrm{b})]\end{aligned}$$

Since $N_\mathrm{D}-N_\mathrm{A}=n_0-p_0$ where n_0 and p_0 are the equilibrium concentrations and the excess carrier concentrations in the bulk $\Delta p = p_\mathrm{b}-p_0$ and $\Delta n = n_\mathrm{b}-n_0$ are equal.

In the bulk σ is, of course, zero but in the barrier region p and n are changed from their bulk values so that a space charge results.

Also

$$\begin{aligned} p &= n_\mathrm{i}\exp(-q\phi_\mathrm{p}/kT); \qquad & p_\mathrm{b} &= n_\mathrm{i}\exp(-q\phi_\mathrm{pb}/kT)\\ n &= n_\mathrm{i}\exp(q\phi_\mathrm{n}/kT); & n_\mathrm{b} &= n_\mathrm{i}\exp(q\phi_\mathrm{nb}/kT)\end{aligned}$$

Integrating Poisson's equation, we have

$$\left(\frac{\mathrm{d}\psi}{\mathrm{d}x}\right)^2 = \frac{-8\pi}{\kappa}\int_{\psi_\mathrm{b}}^{\psi}\sigma\,\mathrm{d}\psi = \frac{-8\pi}{\kappa}\int_{\phi_\mathrm{p,\,nb}}^{\phi_\mathrm{p,\,n}}\sigma\,\mathrm{d}\phi_\mathrm{p,\,n}$$

using $\phi_\mathrm{p} = \mathrm{const}_1+\psi$, and $\phi_\mathrm{n} = \mathrm{const}_2+\psi$. Therefore,

$$\begin{aligned}\left(\frac{\mathrm{d}\psi}{\mathrm{d}x}\right)^2 = \frac{8\pi}{\kappa}qn_\mathrm{i}\Big\{&\frac{kT}{q}[\exp(-q\phi_\mathrm{p}/kT)-\exp(-q\phi_\mathrm{pb}/kT)]-\\ &-\frac{kT}{q}[\exp(q\phi_\mathrm{n}/kT)-\exp(q\phi_\mathrm{nb}/kT)]+\\ &+[\exp(-q\phi_\mathrm{pb}/kT)-\exp(q\phi_\mathrm{nb}/kT)](\psi-\psi_\mathrm{b})\Big\} \qquad \ldots(1)\end{aligned}$$

This gives the field at x. The voltage profile in the space charge region may be found by integrating this equation. For the field we may write

$$E_x = -\frac{2kT}{q}\frac{1}{\mathscr{L}}\left[(p+n)-(p_\mathrm{b}+n_\mathrm{b})+\frac{q}{kT}(\psi-\psi_\mathrm{b})(p_0-n_0)\right]^{1/2} \qquad \ldots(2)$$

where $\mathscr{L}$ is a characteristic length called the Debye length and is equal to

$$\left(\frac{\kappa kT}{2\pi n_\mathrm{i}q^2}\right)^{1/2}$$

This length gives a measure of the width of the space charge region. The field at the surface is thus

$$E_s = -\frac{2kT}{q}\frac{1}{\mathscr{L}}\left[(p_s+n_s)-(p_b+p_n)+\frac{q}{kT}(\psi_s-\psi_b)(p_0-n_0)\right]^{1/2} \quad \ldots (3)$$

This gives the charge in the barrier region, since

$$Q_{sc} = -4\pi E_s$$

The case of Figure 3(b) follows through in a similar manner and the same expressions result. Here for the integration we take the space charge region to

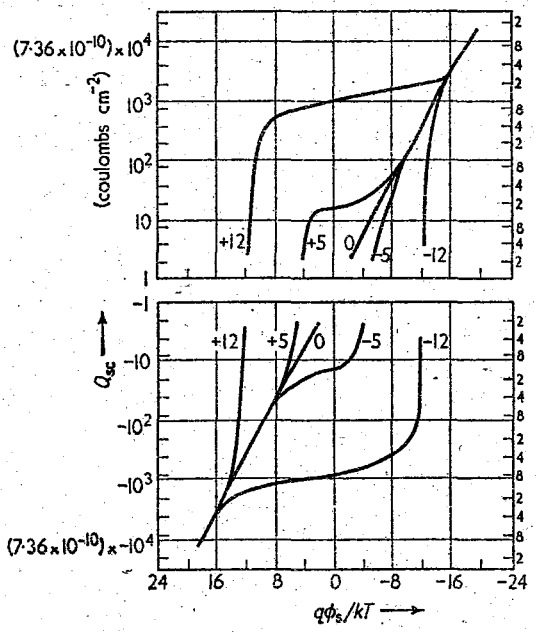


Figure 4. Q_{sc} *as a function of* ϕ_s *for various bulk resistivities (after Dousmanis). Figures on curves refer to values of* $q\phi_b/kT$

be the region over which both quasi-Fermi levels are horizontal; the region beyond this is neutral.

For the equilibrium case p_s and n_s are simply the equilibrium values.

Equation (3) thus applies to the three cases:

For the equilibrium case:

$$p_s = p_0 \exp\left[-q(\psi_s-\psi_b)/kT\right], \quad p_b = p_0$$
$$n_s = n_0 \exp\left[q(\psi_s-\psi_b)/kT\right], \quad n_b = n_0$$

For the case with excess carriers:

$$p_s = (p_0+\Delta p)\exp\left[-q(\psi_s-\psi_b)/kT\right], \quad p_b = p_0+\Delta p$$
$$n_s = (n_0+\Delta n)\exp\left[q(\psi_s-\psi_b)/kT\right], \quad n_b = n_0+\Delta n$$

For the case with the reverse bias:

$$p_s = p_0 \exp\left[-q(\psi_s - \psi_b + V)/kT\right], \quad p_b = 0$$
$$n_s = n_0 \exp\left[q(\psi_s - \psi_b)/kT\right], \quad n_b = n_0 - p_0$$

We see therefore that equation (3) gives the space charge in terms of the barrier height, the excess carrier concentrations, and the bulk resistivity. When the bands bend downwards analogous expressions result, but the opposite sign must be taken for the field.

The results are illustrated in Figure 4 for the equilibrium case where Q_{sc} is plotted as a function of ϕ_s for various bulk resistivities.

3.2. The Surface Conductance

It is evident that the increase or decrease of conductance due to the surface layer will be a function of barrier height and to evaluate it we introduce the concept of the surface excesses.[18] This may be defined as the change in the number of the free carriers per square centimetre of surface due to the presence of the surface barrier. For simplicity we consider the equilibrium case.

For holes, we may write for the excess

$$\Gamma_p = \int_0^\infty (p - p_b)\,dx$$
$$= -\int_{\phi b}^{\phi s} (p - p_b)\frac{dx}{d\phi}\cdot d\phi \qquad \ldots(4a)$$

and for electrons
$$\Gamma_n = -\int_{\phi b}^{\phi s} (n - n_b)\frac{dx}{d\phi}\cdot d\phi \qquad \ldots(4b)$$

The charge in the surface barrier may be written

$$Q_{sc} = q(\Gamma_p - \Gamma_n)$$

The change in conductance due to the surface barrier is given by

$$G_s = q(\mu_p \Gamma_p + \mu_n \Gamma_n) \qquad \ldots(5)$$
$$= q\mu_p(\Gamma_p + b\Gamma_n)$$

where
$$b = \frac{\mu_n}{\mu_p}$$

μ_n and μ_p are the electron and hole mobilities.

Equations (5), (4), and (1) give the surface conductance as a unique function of the barrier height. Curves showing the relationship are given in Figure 5 for different bulk resistivities.

The conductance is zero when

$$\Gamma_p = \Gamma_n = 0$$

and when

$$\Gamma_p = -b\Gamma_n$$

and between the barrier heights corresponding to these two conditions an exhaustion layer exists. Beyond these barrier heights there is an inversion or accumulation layer. The minimum value of conductance occurs when

$$\frac{\partial \Gamma_p}{\partial \phi_s} = -b\frac{\partial \Gamma_n}{\partial \phi_s}$$

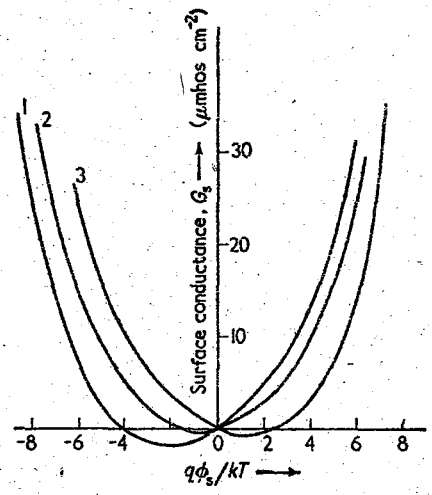


Figure 5. *Surface conductance as a function of ϕ_s for various bulk resistivities (after Schrieffer). Curve 1:* $q\phi_b = 0{\cdot}036$ *eV* (15 *Ωcm N-type); Curve 2:* $q\phi_b = 0$ (*intrinsic*); *Curve 3:* $q\phi_b = 0{\cdot}065$ *eV* (10 *Ωcm P-type*)

Using equations (4) the condition is $(p_s - p_0) = -b(n_s - n_0)$. That is,

$$\frac{\exp\left[-q(\phi_s-\phi_b)/kT\right]-1}{\exp\left[q(\phi_s-\phi_b)/kT\right]-1} = -b\left(\frac{n_0}{p_0}\right)$$

when

$$|\phi_s - \phi_b| \gg \frac{kT}{q}$$

this becomes

$$-(\phi_s-\phi_b) = \frac{kT}{q}\ln b\frac{n_0}{p_0}$$

Since

$$-\phi_b = \tfrac{1}{2}\frac{kT}{q}\ln\frac{p_0}{n_0}$$

it is seen that for $b = 1$, $\phi_s = -\phi_b$ at the minimum conductance point.

We have assumed that the mobilities of the carriers at the surface are the same as the bulk values. When there is a large positive or negative barrier a potential well exists for one or other type of carrier and under certain circumstances this can limit the mean free path. This has been studied by Schrieffer[15, 16] who has shown that the mobility at the surface may be considerably reduced. We consider the case where there is a large potential well for electrons and assume that when an electron collides with the surface it loses all its drift velocity. For simplicity the transverse field is taken as constant so that the potential profile is as shown in Figure 6. A rough estimate of the effective mobility μ' may be made by taking the mean free time τ_m' to be $2w/v$. w is the distance from the surface at which the potential is sufficient to reduce the average thermal velocity v to zero, i.e.

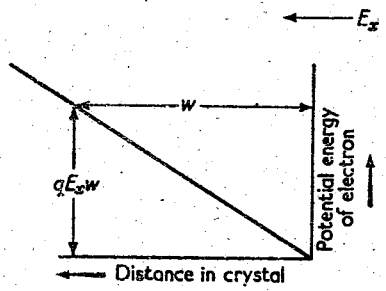


Figure 6. *Potential well with uniform field*

$qE_xw = kT$. Letting τ_m be the bulk mean free time the ratio of effective to bulk mobility may be written

$$\frac{\mu'}{\mu} = \frac{\tau_m'}{\tau_m} \simeq \frac{2w}{\tau_m v} \simeq \frac{kT}{\tau_m qE_x} \Big/ \sqrt{\frac{kT}{m^*}}$$

$$\simeq \frac{\sqrt{(kTm^*)}}{q\tau_m E_x} \equiv \sqrt{2}.\alpha$$

We see therefore that the reduced mobility is proportional to $1/E_x$.

In practice the case of deep wells, as in Figure 3(b), closely approaches this and since the field mainly arises from the constant ionized impurity concentration usual depletion layer theory shows that, since $V \propto E_x^2$, $\mu' \propto 1/\sqrt{V}$.

Schrieffer has considered the case of the linear potential gradient rigorously and, assuming as above completely diffuse scattering, he finds

$$\frac{\mu'}{\mu} = 1 - \exp(\alpha^2)(1 - \operatorname{erf}\alpha)$$

where α is as defined above.

For very large E_x this reduces to

$$\frac{\mu'}{\mu} \simeq 2\alpha\pi^{1/2}$$

which compares with the approximate expression found above.

He has also investigated the more general case where the field is derived as in sub-section 3.1 and where there are non-equilibrium conditions due to a reverse bias. The effective mobility is found to be a function of two parameters which

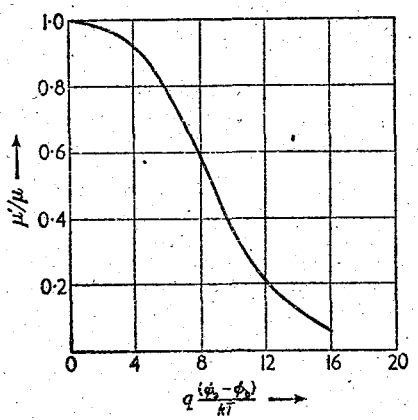


Figure 7. *Surface mobility as a function of barrier height (after Schrieffer).* $N_A = 10^{18}\ cm^{-3}$, $m^*/m = 0{\cdot}25$

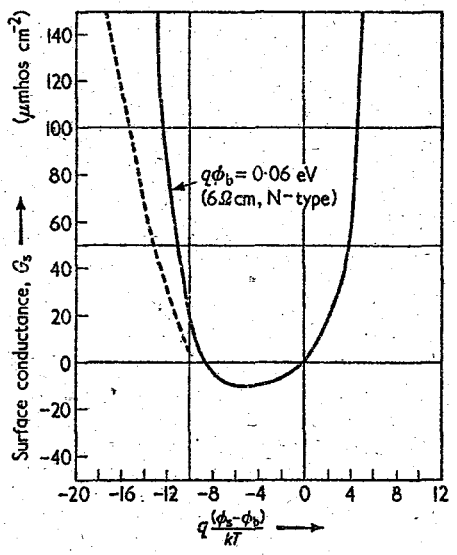


Figure 8. *Surface conductance as a function of barrier height showing Schrieffer correction*

depend on the total barrier height, which depends on V and the carrier concentrations at the surface and in the bulk.

To illustrate the results two Figures are given. Figure 7 shows the variation of mobility as a function of the barrier height $q(\phi_s - \phi_b)$ for the equilibrium case with

no bias applied. P-type germanium is assumed with a potential well for electrons with an effective mass of 0·25 of the free space mass. The resistivity is nearly intrinsic with $N_A = 10^{13}$ cm^{-3}. The curve is slightly modified with increased impurity concentration due to the change in shape of the barrier potential profile. The modification to a surface conductance curve is given in Figure 8. The corrected conductance due to the reduced mobility is shown by the dotted curve, only the inversion layer side giving sufficiently deep wells at low surface conductances to give a significant correction. It should be noted here that the surface conductance curves of Figure 5 are actually based on the corrected mobility.

As mentioned previously Schrieffer assumes in his work that the scattering is completely diffuse. If, however, there is some specular reflection of carriers from the surface, i.e. the carriers are reflected without any change of energy and longitudinal velocity then the reduction in mobility would be less. Total specular reflection would give no change from the bulk mobility.

3.3. The Fast States

The fast states, situated at the germanium–germanium oxide interface, may be specified by their energy level in the forbidden gap, their density, and their capture cross-sections for electrons and holes. Their occupation may be described by a Fermi–Dirac distribution function f_t so that if there are N_t states (or traps) per square centimetre at an energy level E_t the number of electrons in them is

$$n_t = \frac{N_t}{1+\frac{1}{2}\exp\left[(E_t - E_{Ft})/kT\right]} = N_t f_t \qquad \ldots (6)$$

E_{Ft} is the trap quasi-Fermi level describing the occupation of the states. Under equilibrium conditions this coincides with the Fermi level E_F.

The $\frac{1}{2}$ occurs because each state can be occupied by one electron with plus or minus spin but not with two electrons of opposite spin.

Assuming that the states communicate with the conduction band and if c_n is the probability per second that an electron is captured by a state then, in accordance with the Shockley–Read theory,[20] there will be an electron current to the states of $c_n n_s N_t(1-f_t)$ and an emitted current of $c_n n_1 N_t f_t$, where

$$n_1 = N_c \exp\left[-(E_c - E_t)/kT\right]$$

where N_c is the density of states in the conduction band. The net electron current to the states is thus

$$U_n = c_n n_s N_t(1-f_t) - c_n n_1 N_t f_t \qquad \ldots (7a)$$

Similarly if the states communicate with the valence band and c_p is the probability per second that a hole is captured by a state, then there will be a net hole current to the states

$$U_p = c_p p_s N_t f_t - c_p p_1 N_t(1-f_t) \qquad \ldots (7b)$$

$N_t(1-f_t)$ is the number of holes in the states, p_t, and p_1 is defined in a similar manner to n_1

$$p_1 = N_v \exp\left[(E_v - E_t)/kT\right]$$

where N_v is the density of states in the valence band.

Under equilibrium conditions the two electron currents and the two hole currents are equal so that $U_n = 0$ and $U_p = 0$.

We now consider the quasi-equilibrium case illustrated in Figure 3(a). If c_n and c_p are both not zero the excess carriers will recombine through the states and there will be equal electron and hole currents flowing to them, i.e. $U_n = U_p = U$. Using equations (7) to eliminate f_t and since $p_s n_s = p_b n_b$ and $p_1 n_1 = p_0 n_0$, we find

$$U = \frac{c_p c_n N_t (p_b n_b - p_0 n_0)}{c_n(n_s + n_1) + c_p(p_s + p_1)}$$

Defining the surface recombination velocity $s = U/\Delta p$ gives

$$s = \frac{c_p c_n N_t (p_0 + n_0)}{c_n(n_s + n_1) + c_p(p_s + p_1)}$$

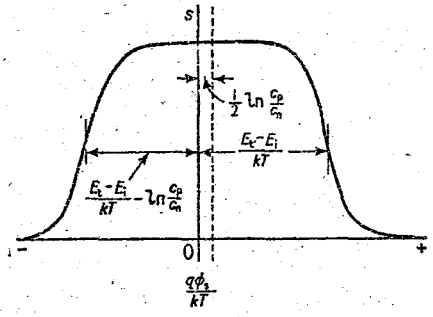


Figure 9. *Surface recombination velocity as a function of* ϕ_s

with a little manipulation this may also be written

$$s = \frac{N_t c_p (p_0 + n_0)}{2n_i \exp(q\phi_0/kT)\{\cosh[(E_t - E_i - q\phi_0)/kT] + \cosh[q(\phi_s - \phi_0)/kT]\}} \quad \cdots (8)$$

where $$\phi_0 = \frac{kT}{2q}\ln\frac{c_p}{c_n}$$

We see from the expression that the maximum value occurs at a barrier height such that $\phi_s = \phi_0$.

When $$\cosh\left(\frac{E_t - E_i - q\phi_0}{kT}\right) \gg 1$$

the half maximum values are given by

$$\frac{q\phi_s}{kT} = \frac{E_t - E_i}{kT}$$

and $$\frac{q\phi_s}{kT} = \frac{-(E_t - E_i)}{kT} + \ln\frac{c_p}{c_n}$$

A typical curve of the variation of s with ϕ_s is shown in Figure 9.

This derivation assumes that the Maxwell–Boltzmann distribution law is obeyed. With high fields due to large barrier heights this is not true and this case has been studied by Berz[21] who has given modified s versus ϕ_s curves.

Another non-equilibrium case must now be considered. We have seen that under equilibrium conditions currents flow between the space charge region and the fast states and the charge in each remains at a stationary value. We now suppose that the charge in the space charge region is suddenly changed by a small amount. It will be seen that this may be done by means of an external electrode. Some of the charge will be transferred to the fast states until they are in equilibrium (or more correctly quasi-equilibrium) with the space charge region. This will take place with a characteristic time constant τ_s.

Let the added charges in the space charge region and the fast states at time t after the initial change be δQ_{sc} and δQ_{fs}, then

$$\delta Q_{sc}+\delta Q_{fs} = \text{total added charge} = \text{constant}$$

The net rate of charge transfer is $q(U_p'-U_n')$ where the primes denote non-equilibrium values, so

$$q(U_p'-U_n') = -\frac{d\,\delta Q_{sc}}{dt} = \frac{d\,\delta Q_{fs}}{dt}$$

When the fast states and space charge region reach equilibrium, $U_p'=U_n'$, but before this they will have different values because of changes δf_t, δp_s, and δn_s in f_t, p_s, and n_s. Thus, from equations (7),

$$U_p'-U_n' = c_p N_t[\delta f_t(p_s+p_1)+f_t\,\delta p_s]+c_n N_t[\delta f_t(n_s+n_1)-(1-f_t)\,\delta n_s]$$

δQ_{fs} differs from its equilibrium value by $qN_t\delta f_t$ and

$$\frac{d\delta Q_{fs}}{dt} = -qN_t\frac{d\delta f_t}{dt}$$

The time constant is given by

$$\frac{1}{\tau_s} = \frac{U_p'-U_n'}{N_t\,\delta f_t}$$

$$= c_p\left(p_s+p_1+\frac{f_t\,\delta p_s}{\delta f_t}\right)+c_n\left(n_s+n_1-\frac{1-f_t}{\delta f_t}\cdot\delta n_s\right)$$

Assuming that the variation of charge in the space charge region is produced only by changes in barrier height, we have

$$\delta n_s = n_s\frac{q}{kT}\delta\phi_s \quad \text{and} \quad \delta p_s = -p_s\frac{q}{kT}\delta\phi_s$$

Therefore

$$\frac{1}{\tau_s} = c_p(p_s+p_1)\left[1-\frac{q}{kT}\frac{f_t p_s}{(\delta f_t/\delta\phi_s)(p_s+p_1)}\right]+$$

$$+c_n(n_s+n_1)\left[1-\frac{q}{kT}\frac{(1-f_t)n_s}{(\delta f_t/\delta\phi_s)(n_s+n_1)}\right]$$

Using $n_s/(n_s+n_1)=f_t$ and $p_s/(p_s+p_1)=1-f_t$, and since for changes in barrier height the change in charge in the space charge region and the fast states are equal and opposite

$$N_t\frac{\mathrm{d}\delta f_t}{\mathrm{d}\phi_s}=\frac{1}{q}\frac{\mathrm{d}\delta Q_{sc}}{\mathrm{d}\phi_s}=\frac{\mathrm{d}\delta(\Gamma_p-\Gamma_n)}{\mathrm{d}\phi_s}$$

so
$$\frac{1}{\tau_s}=c_p(p_s+p_1)+c_n(n_s+n_1)\left[1-\frac{q}{kT}\frac{N_t}{\mathrm{d}(\Gamma_p-\Gamma_n)/\mathrm{d}\phi_s}f_t(1-f_t)\right] \quad \ldots(9)$$

4. MEASUREMENTS

After the advent of the transistor the first significant quantitative experiments on the electrical properties of surfaces were perhaps those of Bardeen and Brattain in 1953.[11] An essential part of this work was the finding of a method to give reproducible changes in surface barrier height of the germanium specimens and a method to measure the change. The variations in barrier height were produced by changing the ambient atmosphere in contact with the specimen and contact potential measurements showed changes in the work function and hence in the barrier height. It was found that electronegative gases and vapours such as ozone, chlorine, hydrogen peroxide, oxygen, etc., gave P-type surfaces while electropositive ones such as wet nitrogen and alcohol gave N-type surfaces. Barrier heights between the extremes could be obtained by dry nitrogen and air and the total change of the surface barrier was 0·5 eV. The variation at ambients cycling the surface barrier from one extreme to the other is called the Bardeen–Brattain ambient cycle.

Measurements of changes in contact potential with illumination of the surface and surface recombination velocity were also made and the results were interpreted in terms of a surface model. Soon after, this work was refined by Morrison,[22] Bardeen and Morrison,[23] and Stevenson and Keyes.[17] The changes in surface barrier height produced by the variation of ambient atmosphere were found to produce changes in surface conductance and surface recombination velocity. The changes in conductance together with the changes in contact potential indicated that the main alteration of the potential occurred in the surface barrier and not the oxide layer. The variations of the s.r.v. were explicable in terms of changes in barrier height, the numbers and energy levels of the states remaining constant. The application of a transverse electric field to the surface was also tried and this gave changes of surface conductance with which long time constants were associated. From all this emerged the basic ideas of the surface model. As the barrier height was controlled by the ambient atmosphere it was evident that states associated with the outside of the oxide layer were produced with a fairly large density. In contrast to this, the states controlling the s.r.v. were relatively unaffected by the ambient atmosphere conditions and so were probably situated at the germanium–germanium oxide interface. As they did not control the equilibrium barrier height they must be much less numerous than the ambient sensitive states. It seemed probable that the states on the outside of the oxide layer were associated with the long time constants.

The next progress was the more accurate determination of the properties, i.e. numbers, energy levels, and capture cross-sections for holes and electrons of the

interface or fast states. Two main methods were pursued. One was by the application of the field effect where an external electrode placed in close proximity to the surface applied a transverse field[24–30] while the other was based on inversion layer channels formed across the middle region of a P–N–P or N–P–N structure.[31,32] A number of workers have determined the state parameters of germanium surfaces by these methods and mainly good agreement has been found.

No such detailed description of the fast states of silicon surfaces is available at the present time. This seems to be mainly due to the difficulty in swinging the barrier height by reasonable amounts by the field effect. Fast-state determinations by channel measurements have been carried out by Statz et al.[31,32] and they have found somewhat similar state distributions to those in germanium but with a density about ten times as large.

Precise measurements by Many and Gerlick,[33] and Harnik and Margoninski,[34] on germanium have shown that the fast states are slightly ambient sensitive, particularly immediately after etching. This dependence on the degree of oxidation at the surface has thrown some light on their structure. It is evident that they cannot be Tamm states as their density is too low (10^{11} cm^{-2} as compared with about 10^{15} cm^{-2}). It may be postulated that the broken bonds of the surface atoms become completed by oxidation. Experiments by Wang and Wallis[35] on the changes in the fast states with oxidation and by Feuersanger (reported by Many[36]) using argon bombardment indicate that the fast states may be due to misfits between the oxide and germanium lattices.

Slow-state measurements have been carried out mainly by the field effect.[37–39] Experiments by Lasser, Wysocki, and Bernstein[40,41] have shown that the majority probably reside on the outside of the oxide layer. Water vapour and other condensible vapours appears to give a very large number of states and Statz et al.[42,43] have found motion of charge in the states in this case.

Concurrently with the work on real surfaces the much more difficult study of clean surfaces under high vacuum has been carried on. The technique of ion bombardment has been developed by Farnsworth and his co-workers[44,45] and cleaned surfaces have been investigated by several workers[46–51] who have measured work function, photoconductance, surface conductance, field effect, and Hall effect. Although not yet clear cut the results are consistent with the existence of a large number of acceptor levels of the Tamm type.

4.1. The Field Effect

Field effect measurements have proved to be the most powerful method of obtaining information about the surface states. A transverse field is applied to the surface of the semiconductor by means of a plate arranged as in a parallel plate capacitor so that a charge of opposite sign to that on the plate is induced in the surface. If δQ_p is the charge on this plate, then we have at any instant

$$\delta Q_p + \delta Q_{ss} + \delta Q_{fs} + \delta Q_{sc} = 0$$

Let us consider the sudden application of a field to the surface of a specimen. An additional charge will appear at the surface and will distribute itself in the space charge region and the fast and slow states. The conductance of the specimen will therefore change, any variation in the space charge region giving a difference in

the number of free carriers; any charge trapped in the states is immobile and so does not contribute to the conductance.

We now examine the process in more detail. To be definite a thin slab of N-type germanium 1 cm^2 with a strong inversion layer at the surface is assumed. The sequence of events on the sudden application of the field is illustrated in Figure 10. The plate is taken to be charged negatively with N electronic units (3 in the Figure). In Figure 10(a), no field is applied and the specimen is in equilibrium.

In (b), the field is suddenly applied and to neutralize the charge on the plate there will be a flow of minority carriers to the space charge region from the bulk. A flow of electrons out of the specimen supplies the negative charge on the plate. The barrier height therefore changes so that there is an excess charge of N holes

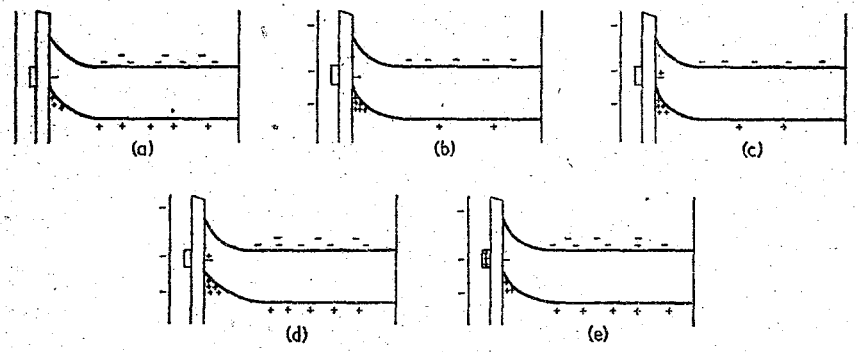


Figure 10. *Diagrammatic representation of sequence of events following application of field*

in the space charge region and there will be a change of $-N$ electrons and $-N$ holes in the bulk (for reasons of neutrality). The change in conductance is therefore

$$\delta G = qN(\mu_p' - \mu_n - \mu_p)$$

where the prime denotes mobility in the surface well. This will take place very quickly as it is an electromagnetic phenomenon.

In (c), the surface states are coming into equilibrium with the added charge in the barrier region. Because of their large capture cross-sections the fast states will be the first to reach equilibrium. The barrier height therefore adjusts itself so that there is a total charge of N holes in the fast states and the barrier. Let λ be the fraction of the N holes that occupy the fast states. The change of conductance is now

$$\delta G = qN[(1-\lambda)\mu_p' - \mu_n - \mu_p]$$

This will take place with the relaxation time of the fast states τ_s which was calculated in sub-section 3.3 and is about 10^{-8} sec.

In Figure 10(d), equilibrium in the bulk is restored by the generation of hole–electron pairs through recombination centres in the bulk and at the surface. We then have

$$\delta G = qN(1-\lambda)\mu_p'$$

This will take place in a time equivalent to the lifetime of the specimen which is normally of the order of 10^{-5} sec.

In (e), the additional holes in the fast states and space charge region are gradually transferred to the slow states. Because of their large number the slow states eventually capture all the added charge so that the specimen is effectively shielded from the field. The barrier height thus returns to its original value and the conductance is unchanged. This process is controlled by the relaxation time of the slow states which is usually a matter of minutes.

We see, therefore, that from this type of measurement a considerable amount of information can be obtained. From (b) it would appear to be possible to measure the corrected mobility in potential wells. From (c) the relaxation times of the fast states can be determined. (d) may be used to give a measure of the lifetime of the specimen. After this transient is complete the change of conductance depends on the number of carriers trapped in the fast states so that information on their densities and energy levels is possible. Using (e) the slow states may be investigated.

4.2. Field Effect Measurements on the Fast States

The majority of the work on the field effect has been on measurements of the fast states and this will be dealt with first. A determination of the parameters of

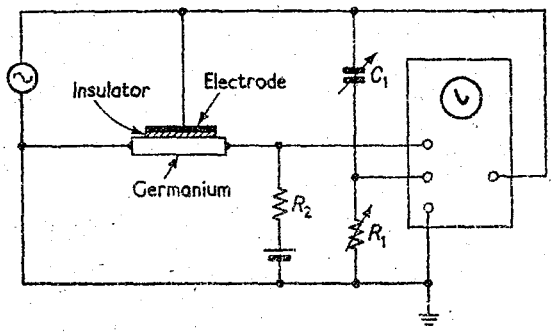


Figure 11. *Diagram of apparatus for measuring a.c. field effect (after Montgomery and Brown)*

the fast states (energy levels, densities, capture cross-sections) requires a knowledge of s and the change in charge in the fast states ΔQ_{fs} as a function of ϕ_s. Here the symbol Δ is used generally for total changes in Q_{fs} produced when the barrier height is varied by any means over a wide range. The symbol δ is reserved for small variations produced by the application of a field. In principle ϕ_s may be directly determined from the surface conductance because of their unique relationship. Measurement of s may be made by the injection of excess carriers while knowing the bulk lifetime, and ΔQ_{fs} may be found from the field effect by comparing the total added charge with that added to the space charge region as determined by the change in the surface conductance. For the latter it is necessary for the field to vary fast enough so that there is no charge transfer to the slow states. We may thus write $\delta Q_{fs} = -\delta Q_p - \delta Q_{sc}$.

The principle of the method of determining ϕ_s and ΔQ_{fs} in practice may be illustrated by describing the work of Montgomery and Brown.[27] In this a field is applied to a thin slice of germanium by an electrode separated from the surface

by a sheet of insulator. The voltage applied is of the order of hundreds of volts and the frequency above about 30 c/s. A typical circuit is shown in Figure 11. The field modulates the surface conductance so that a variation of voltage appears across R_2. There is also a charging current for the electrode which flows through the specimen and causes an additional voltage across R_2. This may be compensated for by R_1C_1 so that the voltage giving the Y deflection of the oscilloscope is that due to the change in the surface conductance. The X deflection of the oscilloscope is proportional to the voltage applied to the electrode. The display is thus a graph of surface conductance against field which is of the same general shape as the G_s versus ϕ_s curves since large barrier heights correspond to large fields (sub-section 3.1). Due to breakdown effects there is an upper limit to the value of the applied field so that only a portion of the curve may be produced. Use of the Bardeen–Brattain ambient cycle gives considerable changes of barrier height so

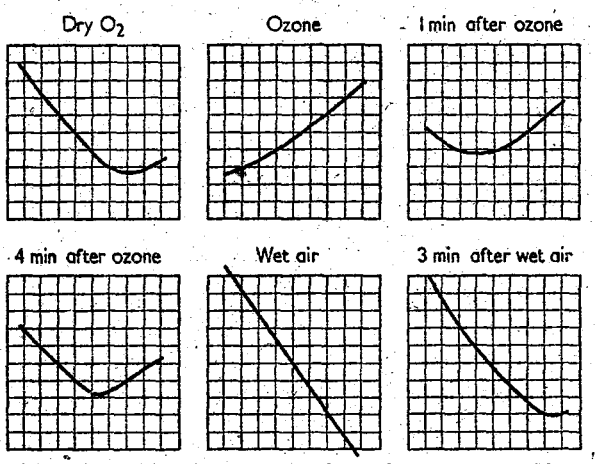


Figure 12. *Field effect patterns in various gaseous ambients (after Montgomery and Brown)*

that the mean operating point for the field effect measurements may be shifted. A portion of the curve may be obtained for each ambient and this is shown in Figure 12. The total curve may therefore be built up from the portions at each barrier height. They may be positioned vertically from the d.c. conductance changes occurring with the different ambients and their horizontal position may be determined by lining them up. A built-up curve for a 40 Ωcm P-type germanium specimen is shown in Figure 13. The horizontal scale is the total added charge Q_p which is proportional to the field as $E=4\pi Q_p$. The zero is usually taken arbitrarily as the mid-point of a curve obtained with an ambient giving an intermediate barrier height. The vertical scale is the conductance per square centimetre G_s, the minimum value being taken as zero. Also in the Figure is shown a theoretical curve of surface conductance versus charge. The charge here is that in the space charge region, and the curve may be calculated by the methods of subsection 3.2, the bulk resistivity having been previously determined. The Schrieffer correction is taken into account. In the Figure the curve is placed horizontally so that the charge zero corresponds to that when there is no potential barrier ($\phi_s = \phi_b$).

The value of the minimum conductance of the measured curve must be the same as that of the theoretical as it is independent of any trapped charge. The barrier height in electron volts is marked on each curve.

For any barrier height the charge trapped in the fast states may now be found. For any given conductance both curves have the same charge in the space charge region and the same barrier height. The change in trapped charge is therefore found by subtracting the charges of the two corresponding points on the curves for the same conductance. This gives a curve of ΔQ_{fs} against ϕ_s. The zero of charge is arbitrary but the variation enables the curve to be fitted to determine the density and energy levels of the fast states on the basis of equation (6) as $E_T - E_F$ varies as ϕ_s.

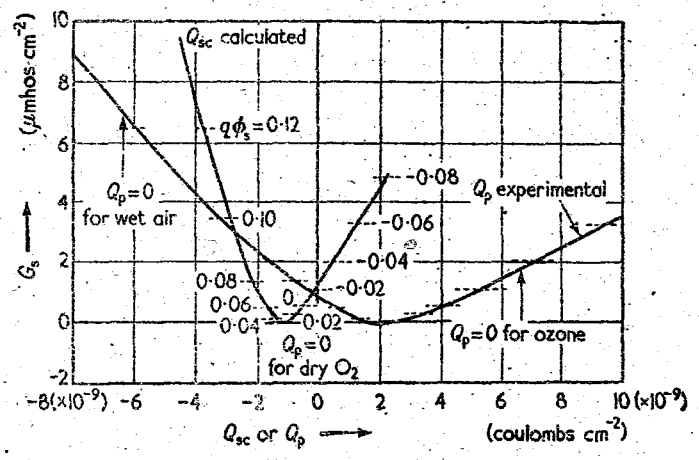


Figure 13. *Calculated and experimental surface conductance as a function of charge (after Montgomery and Brown)*

Accurate measurements of the fast state parameters have been made by Many et al.[33] Careful experimental techniques enabled very large barrier height swings to be obtained (of the order of 0·3 eV peak to peak) on CP4 etched surfaces by the use of field alone so that any changes in the fast states due to ambient atmosphere variations were avoided. As the conductance minimum point was included in the swing the barrier height could be determined. The value of s was measured by a bridge method in which excess carriers are injected by an end contact and the change of impedance of the filament under the action of a pulse is measured by a resistance–capacitance bridge. The filament resistance is also measured by the same bridge so that surface conductance changes could be evaluated to determine ϕ_s. The surface barrier was swung by an a.c. field of 50 c/s of variable amplitude. At the positive and negative peaks, pulses were applied to the filament and bridge measurements of lifetime and specimen resistance taken. From these point by point graphs of s and ΔQ_{fs} versus ϕ_s could be plotted and fitted by theoretical curves. The formula for s given by equation (8) shows that c_p/c_n may be determined by the shift of the maximum from $\phi_s = 0$ but, since the cosh function is even, only

$$\left|\frac{E_t - E_i - q\phi_0}{kT}\right|$$

and not its actual value, may be found. This difficulty may be overcome by repeating the measurements at a different temperature. Each temperature gives two possible values of $(E_t - E_i)/kT$ but the correct values should be in the ratio of T_1/T_2. Typical results are produced in Figures 14–16. The specimens were N-type of 21 Ωcm resistivity and the measurements were carried out under vacuum. The energy level of the recombination centres are determined from the s/s_{max} versus ϕ_s curves and may be used to help to fit the ΔQ_{fs} versus ϕ_s curves. We see from Figure 14 that the energy level of the recombination centres is such that, at 289° K (T_1), $E_t - E_i = 4{\cdot}1\ kT_1$, and the ratio of the capture cross-sections is given by

$$1{\cdot}1 = \frac{kT}{2q} \ln \frac{c_p}{c_n}$$

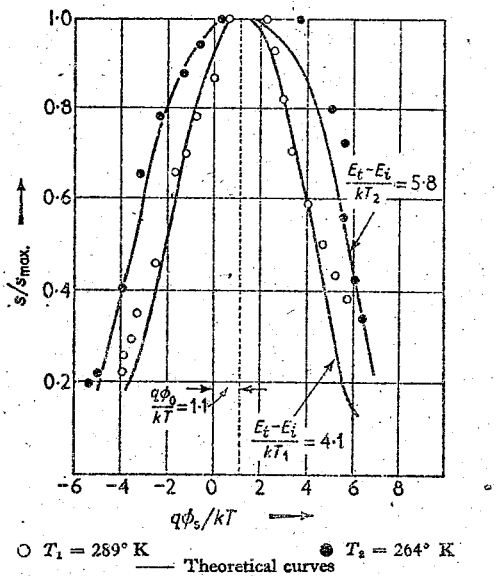
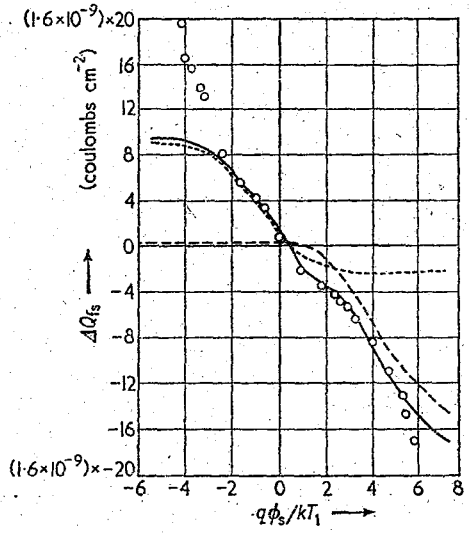


Figure 14. *Surface recombination velocity as a function of ϕ_s (after Many and Gerlich)*

These parameters are used to fit the ΔQ_{fs} versus ϕ_s curves of Figure 15. The densities of the states may be found by noting that the slope of the curve is $N_t/4$ when the Fermi level passes through the energy level of the states, i.e. when $q\phi_s = E_t - E_i$. As this is in the range of barrier swing an accurate estimate may be made to give $N_t = 1{\cdot}5 \times 10^{11}$. It is seen that a complete fit cannot be obtained by the recombination states only but another set of states which are not active recombination centres has to be introduced. These states have $E_t' - E_i = -1\ kT_1$ and $N_t' = 1{\cdot}2 \times 10^{11}$. The above values are consistent with those found at 264° K (T_2) as shown by Figure 16.

The donor or acceptor nature of the states may not be absolutely determined due to the arbitrary zero for the ΔQ_{fs} axis. At negative and extreme positive barrier heights we see that the two sets of states are not sufficient to give a good fit.



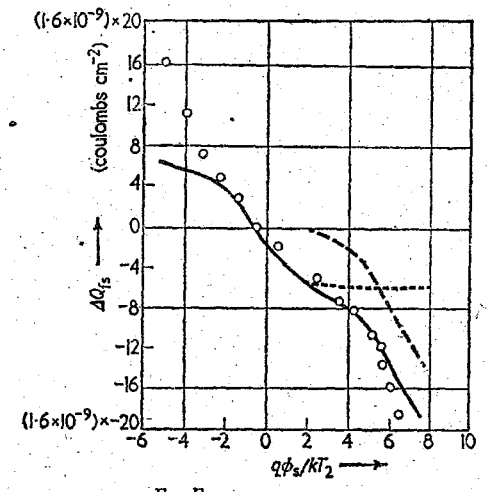
—— — $\frac{E_t - E_1}{kT_1} = 4{\cdot}1$; $N_t = 1{\cdot}5\times10^{11}$

---- $\frac{E_t' - E_1}{kT_1} = 1$; $N_t' = 1{\cdot}2\times10^{11}$

—— Composite curve

O Experimental data ($T_1 = 287°$ K)

Figure 15. ΔQ_{fs} *as a function of* ϕ_s *(after Many and Gerlich)*



—— — $\frac{E_t - E_1}{kT_2} = 5{\cdot}8$; $N_t = 1{\cdot}5\times10^{11}$

---- $\frac{E_t' - E_1}{kT_2} = 1$; $N_t' = 1{\cdot}2\times10^{11}$

—— Composite curve

O Experimental data ($T_2 = 264°$ K)

Figure 16. ΔQ_{fs} *as a function of* ϕ_s *(after Many and Gerlich)*

There appear to be two more sets of states situated well above and below the centre of the gap coming into action. These are estimated to be about $\pm 6\,kT$ from the centre.

A number of other workers[26–30] have used the above principles to determine the fast state parameters but have made the measurements by other methods. Wang and Wallis[30] have measured s by photoconductive response methods and have applied the field effect to dark and illuminated specimens at room temperature. The excess conductance of an illuminated specimen is directly proportional to the effective lifetime τ. For a thin specimen of thickness d we may write

$$\frac{1}{\tau} = \frac{1}{\tau_b} + \frac{2s}{d}$$

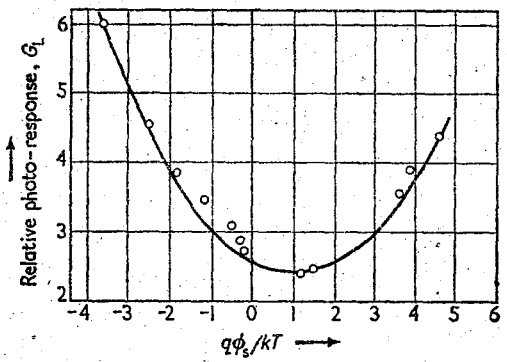


Figure 17. *Photoconductance G_L as a function of ϕ_s. The experimental points are fitted with theoretical curves for recombination centres with* $E_t - E_i = 1{\cdot}1 \pm 2{\cdot}8$ kT *and* $c_p/c_n = 9$ *(after Wang and Wallis)*

where τ_b is the bulk lifetime. The photo response is therefore a function of s and for very thin samples the excess conductance G_L will be roughly inversely proportional to s. The general shape of a G_L versus ϕ_s curve will therefore be that of an inverted s versus ϕ_s curve and the results may be fitted to give recombination centres parameters. Wang and Wallis changed the barrier height by the Bardeen–Brattain ambient cycle and measured the surface conductance G_s so that ϕ_s could be determined. Typical results are shown in Figure 17 for 35 Ωcm N-type germanium, the points giving the experimental results.

As the gaseous ambients were changed the field effect measurements were taken at a frequency of 32 c/s, the peak field being kept constant. This enabled the change of surface conductance δG_s due to the field to be measured and changes in barrier height $\delta\phi_s$ to be calculated. From the knowledge of the induced charge the change in charge in the fast states δQ_{fs} may be found and hence $\delta Q_{fs}/\delta\phi_s$. An integration gives the total variation of charge ΔQ_{fs}. These results are shown in Figure 18. The field effect on an illuminated specimen was also measured. In this case the changes in ϕ_s produced by the field cause a change in G_s and G_L. For sufficiently weak illumination δG_s for dark and illuminated specimens may be

taken to be the same. From the measured conductance change δG we can find $\delta G_L = \delta G - \delta G_s$. The results are given in Figure 19 where $\delta G_L/G_L$ is plotted

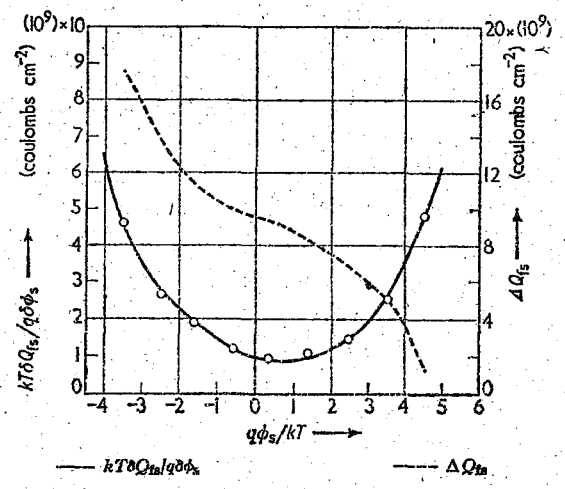


Figure 18. $\delta Q_{fs}/\delta\phi_s$ *and* ΔQ_{fs} *as a function of* ϕ_s. *The experimental points are fitted with theoretical curves of the recombination centres of Figures* 17 *and* 19 *with densities* 4×10^{10} cm^{-2} *and also with two outer states (after Wang and Wallis)*

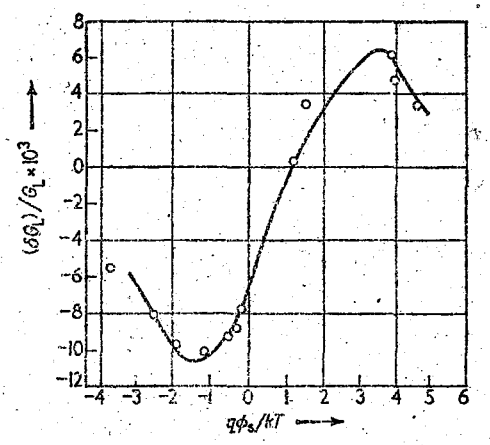


Figure 19. $\delta G_L/G_L$ *as a function of* ϕ_s. *The experimental points are fitted with theoretical curves for recombination centres with* $E_t - E_i = 1{\cdot}1 \pm 2{\cdot}8$ kT *and* $c_p/c_n = 9$ *(after Wang and Wallis)*

against ϕ_s. The solid lines in all the Figures are the fitted curves. The G_L and $\delta G_L/G_L$ versus ϕ_s curves are given by two sets of recombination centres with $q\phi_0/kT = 1{\cdot}1$ so that $c_p/c_n = 9$ and with $E_t - E_i = 1{\cdot}1 \pm 2{\cdot}8\ kT$. The two values

are taken because of the cosh function being even. The results in Figure 18 were fitted with these states with a density of 4×10^{10} cm^{-2} each and with another pair of states $E_t - E_i > 6\ kT$ and $< -5\ kT$. The $\delta G_L/G_L$ curve is of interest because its zero is controlled by c_p/c_n. This may be seen from the fact that changes in barrier height do not produce changes in s at the maximum of the s versus ϕ_s curve, so that at this point, which depends only on c_p/c_n, δG_L is zero. Also it may be shown that the slope of the $\delta G_L/G_L$ curve near its zero is largely determined by the energies of the recombination centres and their c_p/c_n and is very insensitive to their density. This property is useful for determining which parameters vary when changes are taking place in the recombination centres.

Some recent work on fast states has been concerned with changes of their parameters under varying conditions in attempts to throw light on their physical structure. Most of this work has been carried out by Many and his co-workers[34–36] and Wang and Wallis[53] using their respective methods described above. Agreement is found in that changes take place under changing ambient atmosphere conditions particularly on freshly etched surfaces; on aged surfaces the changes are small. There is some disagreement on which parameters change. Wang and Wallis have found that their results are mainly explicable in terms of a change of the density N_t of the recombination centres only, E_t and c_p/c_n remaining unchanged. This is deduced from the constancy of the $\delta G_L/G_L$ versus ϕ_s curves which are independent of N_t. The outer states appeared to remain constant. On the other hand Many et al. have found changes in N_t, c_p/c_n, and E_t of the recombination states. Changes in E_t with temperature have also been found and it may be seen from Figure 16 that a better fit is obtainable with a somewhat higher value of E_t.

The states produced by a number of etches and the effects of baking have also been studied by Wang and Wallis.[54] Four discrete sets of states are always found with two types of the centre states with different E_t and c_p values according to the etch. An excellent critical survey of the state of the above work on the changes in the fast states produced by ambient variations is given in reference 54.

Low temperature field effect measurements on fast states have been made by some workers. Montgomery and Brown[27] have applied a.c. fields down to 170° K on germanium. Their measurements were interpreted to yield the change in the charge in fast states against barrier height by the methods described at the beginning of this section. Their results are consistent with the fast states being unchanged by temperature. At the lower temperatures their δG_s versus added charge curves began to exhibit hysteresis effects which are not yet understood. This effect has been further studied by Banbury et al.[55]

Morrison[56] has made d.c. field effect measurements on germanium and silicon down to about 80° K and studied the resultant transients. On the application of the field there was an immediate change of surface conductance which decayed over a period of the order of tens of seconds to a smaller value. This decay is believed to be due to the charging of the fast states, their relaxation time having increased. With large fields strongly asymmetrical decays were found depending on the polarity of the applied field. This may be explained in terms of the charging time of the fast states being controlled by the barrier height. This mechanism had previously been considered by Morrison to explain the slow-state charging characteristics and is given in detail later (sub-section 4.4).

4.3. The High Frequency Field Effect

In the general description of the field effect we saw that further information could be obtained on the fast states and the surface mobility from the initial part of the transient. Experimentally it is more convenient to study this by means of an a.c. applied field of high frequency. Measurements of the field effect have been made up to a frequency of 5×10^7 c/s by Montgomery,[57] and an analysis of the field effect up to the high frequency range has been carried out by Garrett.[58] Here we will carry through a simple analysis illustrating the basic physical principles.

It was seen in sub-section 4.1 that non-equilibrium carrier concentrations in the bulk can occur in field effect experiments. These concentrations may be described by quasi-Fermi levels which may be extended across the barrier region so that the carrier concentrations there are modified. Following Garrett we will neglect this modification to the surface concentrations for simplicity, although Berz,[59] who extended the analysis in a more rigorous manner, has shown that this may not in every case be justified.

An N-type specimen 1 cm × 1 cm and thickness d, small compared to a diffusion length, is considered. The field of angular frequency ω is applied to one side of the specimen and any effects from the other surface are ignored.

Taking small fields we let the added charge be

$$\delta Q_s \exp(i\omega t)$$

($\delta Q_s = -\delta Q_p$). Then there will be a barrier height change

$$\delta\phi_s \exp(i\omega t)$$

a change in the space charge

$$\delta Q_{sc} = q\delta(\Gamma_p - \Gamma_n^{\mathbf{N}}) \exp(i\omega t)$$

a change in the charge in the fast states

$$\delta Q_{fs} = -qN_t \delta f_t \exp(i\omega t)$$

and a change in the minority carrier density in the bulk

$$\delta p \exp(i\omega t)$$

The above are vector quantities and represent small variations from the zero field or equilibrium values.

These changes in charges vary the conductance of the specimen and to find the variation we divide the specimen into two parts, the surface region under the electrode and the remainder. Then

$$\delta G = \delta G_s + \delta G_b$$

where δG_b refers to the latter region. The field effect mobility is defined as

$$\mu_{FE} = \frac{\delta G}{\delta Q_s}$$

The change in surface conductance

$$\delta G_s = q(\mu_p' \delta\Gamma_p + \mu_n' \delta\Gamma_n)$$

and the change in bulk conductance

$$\delta G_b = q(\mu_p \delta p . d + \mu_n \delta n . d)$$

The primes denote surface values. To find $\delta\Gamma_p$ and $\delta\Gamma_n$ it is necessary to calculate the change in barrier height $\delta\phi_s$ produced by δQ_a. We have

$$\delta\Gamma_p = \frac{\partial\Gamma_p}{\partial\phi_s} . \delta\phi_s \quad \text{and} \quad \delta\Gamma_n = \frac{\partial\Gamma_n}{\partial\phi_s} . \delta\phi_s$$

where $\partial\Gamma_p/\partial\phi_s$ and $\partial\Gamma_n/\partial\phi_s$ are known functions which may be calculated from the results of Section 3. $\delta\phi_s$ is controlled by the fact that $\delta Q_a = \delta Q_{sc} + \delta Q_{fs}$ and is frequency dependent as charge transfer takes place between the space charge region and the fast states with a time constant τ_s as shown in sub-section 3.3. The change in the hole concentration δp in the bulk is due to the need to supply holes to or receive holes from the surface as the number of holes in the space charge region and the fast states change. This is only significant when there is an inversion layer. Because of neutrality requirements there is a change in the electron concentration such that $\delta n = \delta p$, the variation in the number of electrons being supplied by the ohmic contacts. In addition to the frequency dependent requirements of the surface for holes, the generation–recombination processes in the specimen produce a frequency dependence of δp associated with the specimen lifetime τ. We will first deal with the surface region. $\delta\phi_s$ may be found by calculating δQ_{sc} from its continuity equation and using

$$\delta\phi_s = \frac{\partial\phi_s}{\partial Q_{sc}} . \delta Q_{sc} \qquad \ldots (10)$$

At frequencies which are low compared with $1/\tau_s$, δQ_{sc} and δQ_{fs} are in equilibrium. Using the suffix 0 to denote equilibrium values we may find the change in barrier height at this condition

$$\delta Q_a = \frac{\partial Q_{sc}}{\partial\phi_s} . \delta\phi_{s_0} + \frac{\partial Q_{fs}}{\partial\phi_s} . \delta\phi_{s_0}$$

Therefore

$$\delta\phi_{s_0} = \frac{\delta Q_a}{(\partial Q_{sc}/\partial\phi_s) + (\partial Q_{fs}/\partial\phi_s)} \qquad \ldots (11)$$

The continuity equation of δQ_{sc} is

$$\frac{d}{dt} . \delta Q_{sc} \exp(i\omega t) = \frac{d}{dt} \delta Q_a \exp(i\omega t) - \frac{(\delta Q_{sc} - \delta Q_{sc_0})}{\tau_s} \exp(i\omega t)$$

The first term on the right-hand side is the rate at which charge is entering the surface region and the second is the rate at which it is being captured by the fast states. δQ_{sc_0} is the value of δQ_{sc} when the barrier height has the value $\delta\phi_{s_0}$, so

$$\delta Q_{sc} = \frac{i\omega\tau_s \delta Q_a + \delta Q_{sc_0}}{1 + i\omega\tau_s}$$

Since $$\delta Q_{sc_0} = \frac{\partial Q_{sc}}{\partial \phi_s} . \delta \phi_{s_0}$$

and using equations (10) and (11), we obtain

$$\frac{\delta \phi_s}{\delta Q_a} = \frac{\delta \phi_{s_0}}{\delta Q_a}\left(1 + \frac{i\omega\tau_s}{1+i\omega\tau_s} . \frac{\partial Q_{fs}/\partial \phi_s}{\partial Q_{sc}/\partial \phi_s}\right) \qquad \ldots (12)$$

At high frequencies we see that

$$\delta \phi_s = \frac{\partial \phi_s}{\partial Q_{sc}} . \delta Q_a$$

showing that all the added charge is in the space charge region. The contribution μ_{FEs} to μ_{FE} from the surface region is then

$$\mu_{FEs} = q\left(\mu_p' \frac{\partial \Gamma_p}{\partial \phi_s} + \mu_n' \frac{\partial \Gamma_n}{\partial \phi_s}\right)\left(\frac{\partial Q_{sc}}{\partial \phi_s} + \frac{\partial Q_{fs}}{\partial \phi_s}\right)^{-1} \times$$

$$\times \left(1 + \frac{i\omega\tau_s}{1+i\omega\tau_s} . \frac{\partial Q_{fs}/\partial \phi_s}{\partial Q_{sc}/\partial \phi_s}\right) \qquad \ldots (13)$$

To find the contribution of the bulk it is necessary to calculate the change in the number of holes in the surface region as these have to come from the bulk. We may divide this into two parts. For the first part when $\omega \ll 1/\tau_s$ the fast states and the space charge region are in equilibrium and the frequency dependent effects are due to the lifetime τ. When $\omega \ll 1/\tau$ there is no contribution to the field effect mobility from the bulk. For the second part frequencies of the order of $1/\tau_s$ are considered when the charging of the fast states is taken into account.

When $\omega \ll 1/\tau_s$, $\delta \phi_s = \delta \phi_{s_0}$ so that there is a change in the number of holes in the space charge region

$$\delta \Gamma_p = \frac{\partial \Gamma_p}{\partial \phi_s} . \delta \phi_{s_0}$$

In the fast states there is change in the number of holes $= \delta Q_{fs_0}/q$. Not all of these come from the valence band as the states communicate with the conduction band also and the emission of an electron from a state to the conduction band also supplies holes to the states. We may then write for the fraction of the total number of holes in the states captured from the valence band: *the probability of hole capture from the valence band* divided by *the probability of hole capture from the valence band* plus *the probability of electron emission to the conduction band.*

These probabilities follow from the Shockley–Read theory which was treated in sub-section 3.3, and the expression for the relevant number of holes is

$$\frac{\delta Q_{fs_0}}{q} . \frac{c_p p_s f_t}{c_p p_s f_t + c_n n_1 f_t}$$

Using

$$U_n = U_p$$

$$f_t(1-f_t) = \frac{kT}{q}\frac{\partial f_t}{\partial \phi_s} \quad (E_t \text{ varies with } \phi_s)$$

$$\delta Q_{fs_0} = -qN_t\frac{\partial f_t}{\partial \phi_s}.\delta\phi_{s_0}$$

and

$$\frac{1}{f_t} = 1+\frac{n_1}{n_s} = 1+\frac{p_s}{p_1}$$

the fraction becomes

$$-\frac{N_t c_p p_s f_t}{c_n(n_s+n_1)+c_p(p_s+p_1)}q.\frac{\delta\phi_{s_0}}{kT} = -qN_t\frac{c_p p_s f_t}{\Omega}.\frac{\delta\phi_{s_0}}{kT}$$

where

$$\Omega = c_n(n_s+n_1)+c_p(p_s+p_1) \qquad \ldots (14)$$

The total number of holes required by the surface from the bulk is then

$$\delta P_s = \left(\frac{\partial \Gamma_p}{\partial \phi_s}-\frac{q}{kT}.\frac{N_t c_p p_s f_t}{\Omega}\right)\delta\phi_{s_0}$$

The continuity equation for holes in the bulk is

$$\frac{d}{dt}.\delta p.d\exp(i\omega t) = \frac{-d}{dt}.\delta P_s\exp(i\omega t)-\frac{\delta p.d\exp(i\omega t)}{\tau}$$

where

$$\frac{1}{\tau} = \frac{1}{\tau_b}+\frac{s}{d}$$

Therefore

$$\delta p.d = -\frac{i\omega\tau\delta P_s}{1+i\omega\tau}$$

Substituting for δP_s and $\delta\phi_{s_0}$, the contribution to the field effect mobility from the bulk μ_{FEb} is thus

$$\mu_{FEb} = -q(\mu_p+\mu_n)\frac{i\omega\tau}{1+i\omega\tau}\left(\frac{\partial\Gamma_p}{\partial\phi_s}-\frac{q}{kT}.\frac{N_t c_p p_s f_t}{\Omega}\right)\left(\frac{\partial Q_{sc}}{\partial\phi_s}+\frac{\partial Q_{fs}}{\partial\phi_s}\right)^{-1} \qquad \ldots (15)$$

For the case when ω is of the same order as $1/\tau_s$ the change in barrier height is no longer $\delta\phi_{s_0}$ and the charge in the fast states has not its equilibrium value. The change in the barrier height has been calculated previously and δQ_{fs} may be found simply from its continuity equation

$$\frac{d}{dt}.\delta Q_{fs}\exp(i\omega t) = -\frac{(\delta Q_{fs}-\delta Q_{fs_0})\exp(i\omega t)}{\tau_s}$$

which gives

$$\delta Q_{fs} = \frac{\delta Q_{fs_0}}{1+i\omega\tau_s}$$

Taking into account, as above, that not all of the change of the number of holes in the states come from the valence band, we have for number of holes required from the bulk

$$\frac{\partial \Gamma_p}{\partial \phi_s}.\delta\phi_s - \frac{c_p p_s f_t N_t}{\Omega}\frac{q}{kT}.\delta\phi_{s_0}\frac{1}{1+i\omega\tau_s}$$

We see that when $\omega \gg 1/\tau_s$, all the added holes are in the space charge region. The generalization of equation (15) for the contribution of the bulk to the field effect mobility is thus

$$\mu_{FEb} = -q(\mu_p+\mu_n)\frac{i\omega\tau}{1+i\omega\tau}\left(\frac{\partial Q_{sc}}{\partial\phi_s}+\frac{\partial Q_{fs}}{\partial\phi_s}\right)^{-1}\times$$
$$\times\left[\frac{\partial\Gamma_p}{\partial\phi_s}\left(1+\frac{i\omega\tau_s}{1+i\omega\tau_s}.\frac{\partial Q_{fs}/\partial\phi_s}{\partial Q_{sc}/\partial\phi_s}\right)-\frac{q}{kT}.\frac{N_t c_p p_s f_t}{\Omega}\frac{1}{1+i\omega\tau_s}\right] \quad \ldots(16)$$

and the total field effect mobility is the sum of equations (13) and (16).

Following Garrett these may be combined in a simple form by neglecting the Schrieffer correction, so that $\mu_p' = \mu_p$ and $\mu_n' = \mu_n$. δQ_{sc} and δQ_{fs} are expressed in terms of $\delta\Gamma_p$ and $\delta\Gamma_n$, and N_t and δf_t, respectively. We also take $\tau_s \ll \tau$ and after some manipulation

$$\mu_{FE} = -\mu_n + \frac{A_n}{1+i\omega\tau} + \frac{B_n}{1+i\omega\tau_s} \quad \ldots(17a)$$

where

$$A_n = (\mu_n+\mu_p).\frac{(\partial\Gamma_p/\partial\phi_s)-(q/kT)(N_t/\Omega).(c_p p_s f_t)}{[\partial(\Gamma_p-\Gamma_n)/\partial\phi_s]-N_t(\partial f_t/\partial\phi_s)}$$

$$B_n = \frac{(\mu_n+\mu_p)(q/kT)(N_t/\Omega).(c_p f_t p_s)-\mu_n N_t(\partial f_t/\partial\phi_s)}{[\partial(\Gamma_p-\Gamma_n)/\partial\phi_s]-N_t(\partial f_t/\partial\phi_s)}$$

By analogy the formula for P-type is

$$\mu_{FE} = \mu_p - \frac{A_p}{1+i\omega\tau} + \frac{B_p}{1+i\omega\tau_s} \quad \ldots(17b)$$

where

$$A_p = (\mu_n+\mu_p)\frac{-(\partial\Gamma_n/\partial\phi_s)-(q/kT)(N_t/\Omega)c_n(1-f_t)n_s}{[\partial(\Gamma_p-\Gamma_n)/\partial\phi_s]-N_t(\partial f_t/\partial\phi_s)}$$

$$B_p = \frac{(\mu_n+\mu_p)(q/kT)(N_t/\Omega)[c_n(1-f_t)n_s]-\mu_n N_t(\partial f_t/\partial\phi_s)}{[\partial(\Gamma_p-\Gamma_n)/\partial\phi_s]-N_t(\partial f_t/\partial\phi_s)}$$

A plot of these results for P-type germanium of 20 Ωcm resistivity is given in Figure 20 for various barrier heights. The fast-state parameters were chosen to give the best fit to Montgomery's experimental results. The energy level assumed was in the upper half of the energy gap 7 kT from the centre with a density 7×10^{11} cm^{-2} and the capture cross-sections give τ_s to be about 10^{-8} sec. The full lines are for inversion layers, the greatest variations in μ_{FE} occurring for the most

extreme values of barrier height. The shape for inversion layers may be explained by noting that at low frequencies the addition of a positive charge to the surface causes the barrier to flatten so that the change in conductance is negative. As the frequency increases to about the reciprocal of the lifetime the conductance increases because the injection of electrons into the bulk from the surface region gives excess carriers. At frequencies above that corresponding to $1/\tau_s$ the conductance tends to decrease because of the greater tendency of the barrier to flatten due to the fast states being inactive. With less extreme inversion layers and accumulation layers there is no increase in field effect mobility as there is no injection of minority carriers into the bulk. The effects of charging of the states are also negligible as their energy level is now well removed from the Fermi level. μ_{FE} now approaches a constant value equal to the bulk mobility of holes. The dotted line is for illustration only and applies to an accumulation layer with a set of states assumed in the lower half of the energy gap and shows that charging effects are now noticeable.

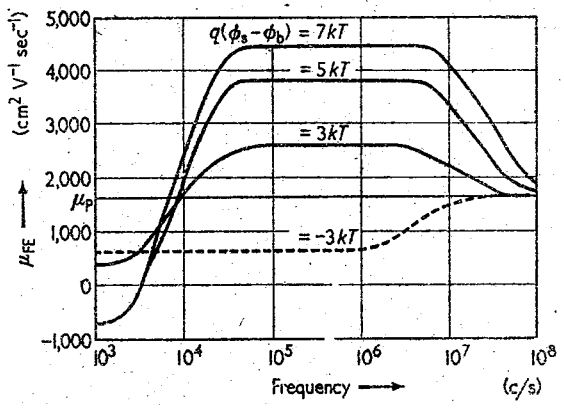


Figure 20. *Field effect mobility as a function of frequency (after Garrett)*

The theoretical curves with the chosen parameters gave a good fit to Montgomery's results. There is no indication of the existence of any other surface states, but a detailed comparison with the fast-state model found from other experiments is not possible as the barrier height was not accurately known.

4.4. Field Effect Measurements on the Slow States

In field effect experiments the slow states are distinguished from the fast states by being much more numerous and having larger time constants. The former property prevents their being filled by the field induced charge and there is usually little or no detectable change in barrier height under stationary conditions with normal values of the field. Accurate determination of their densities and energy levels as described in sub-section 4.2 for the fast states is not possible, but it may be estimated from their clamping effect on the barrier height that their density is greater than 10^{13} cm^{-2} and it seems likely that the barrier is controlled so that the Fermi level passes very near their mean energy level.

Investigations using the field effect have been mainly concerned with evaluating their time constants. Kingston *et al.*[39] have used a.c. fields of variable frequency and have measured the variation of conductance of a germanium specimen against frequency at low frequencies. Morrison[37,38] has also made an extensive study using a d.c. field and measuring the conductance transient. From the results of these and other workers a number of properties of the time constants of the slow states have been established. For the majority of surface treatments the slow decay is not exponential and thus cannot be described by a single time constant. For an etched surface there is an apparent range of time constants from milliseconds to minutes. There have, however, been some cases with low applied fields when an exponential decay has been observed. In most cases the time constants have been found to be very temperature dependent, increasing with decreasing temperatures. Work by Lasser, Wysocki, and Bernstein[40,41] has shown that when the thickness of the oxide layer has been increased by prolonged heating in oxygen there has been a marked increase in the relaxation time. This indicates that the majority of states are on the outside of the oxide layer. When the ambient atmosphere surrounding the specimen is varied changes in the decay time occur. Ambients containing water vapour usually shorten the decay time. On etched surfaces ozone gives shorter relaxation times than dry ambients.

To explain the apparent range of time constants, two models have been proposed. One, due to Kingston and McWhorter, assigns capture probabilities for carriers by the slow states and the theory given in sub-section 3.3 for the fast states will apply. The capture probabilities are much smaller and also when calculating their charging time as in the latter part of sub-section 3.3 a modification would have to be made to include the effects of carrier transfer from fast states in addition to transfer from the space charge region. The range of time constants is obtained by dividing the surface up into small areas with different capture probabilities. If it is assumed that the capture probabilities depend on the distance between the interface and the slow states then they may either be distributed throughout an oxide layer of constant thickness at different depths or alternatively they may be on the outside of an oxide layer of varying thickness.

This model may only be applied where the changes of charge in the slow states are relatively small. Morrison on the other hand has proposed a model which produces a non-exponential decay from a uniform slow-state distribution when the changes in the charge in the states and hence the barrier height variations are large. At equilibrium, before the field is applied, the electron currents between the conduction band and the slow states are equal and opposite and similarly with the hole currents. Assuming for simplicity that only electrons are involved, then let there be an increase in the number in the surface region δN_s when the field is suddenly applied and let there be a barrier height change $\delta\phi_s$, directly proportional to δN_s. The electron current from the conduction band to the slow states will be changed by a factor $\exp(q\delta\phi_s/kT)$ because of the additional potential barrier the electrons have to surmount to reach the slow states. The electron current from the slow states to the conduction band is not affected as the potential barrier it has to surmount is in the oxide layer and is unchanged. We may therefore write

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta N_s = A[1-\exp(\beta\delta N_s)] \qquad \ldots(18)$$

where $$\beta\delta N_s = \frac{q\delta\phi_s}{kT}$$

and β is inversely proportional to the capacitance between the slow states and the induced carriers. This type of equation will give a non-exponential decay for $\beta\delta N_s > 1$ and an exponential decay for $\beta\delta N_s < 1$. Asymmetrical effects are also predicted as for δN_s large and positive

$$\frac{d}{dt}\delta N_s \to -A\exp(\beta\delta N_s)$$

and for δN_s large and negative

$$\frac{d}{dt}\delta N_s \to A$$

These effects have been observed experimentally for large voltages.

Two possible methods of charge transfer through the oxide layers have been proposed. In one, suggested by Kingston and McWhorter, the mechanism is that of quantum mechanical tunnel penetration. This gives capture probabilities which depend on the distance between the slow state and the interface and would make the relaxation time very sensitive to the width of the oxide layer if the majority of the states are on the outside. On the other hand the relaxation time might be expected to be relatively temperature independent. The other mechanism, proposed by Morrison, is that of thermal excitation of the carrier over a potential barrier in the oxide layer. This should give relaxation times which are relatively insensitive to oxide layer width changes but which are very temperature dependent.

The present experimental evidence does not give a clear-cut indication for or against any of the relaxation time and charge transfer models and it seems that a combination of them is necessary to explain all the results.

4.5. The Channel Effect

An alternative method to field effect experiments for making measurements of surface states is that of the channel effect. This differs essentially from the field effect in that a field is applied to the surface internally by means of a P–N junction instead of applying the field externally, by means of an electrode. A typical arrangement is shown in Figure 21. A P–N–P or N–P–N transistor-like structure is used and the surface is specially treated so that an inversion layer is formed on the middle region. The circuit is arranged so that a reverse bias is applied between the middle region and the end regions which are joined by the inversion layer on the middle region. In this way a channel, the conductance of which may be measured, is formed between the end regions and the transverse field applied internally to this channel may be varied by changing the reverse bias. Figure 3(b) gives the energy diagram for the channel. For the case shown the conductance per square centimetre of the inversion layer is given by

$$G_s = q\mu_p' p'$$

where p' is the number of holes per square centimetre of the P-type skin. μ_p' may be found from Schrieffer's theory and is a function of the applied voltage V and ϕ_s. (ϕ_s is written for ϕ_{ps} for convenience.) p' depends on ϕ_s and the effective width of the channel which is governed by V and ϕ_s. Thus channel conductance measurements enable ϕ_s to be determined.

As with the field effect we consider what happens with a sudden change in field. The field at the surface arises from three sources, the holes in the channel, the ionized donors in the depletion layer, and the electrons in the edge of the depletion layer in the N-type region. The total charge which must be balanced by that in the surface states is thus determined by ϕ_s and V. As the reverse bias is increased the field at the surface increases and the charge in the surface states is varied. Neglecting any transient effects due to the fast-state charging time, we see that initially the change in charge will be in the fast states only and the quasi-Fermi level at the interface will rise to increase the number of electrons in the states in the lower half of the energy gap. It may be assumed that the quasi-Fermi level for the states and holes are coincident[32] so that the number of holes in the channel will

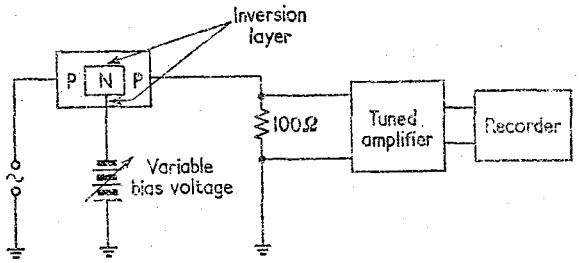


Figure 21. *Diagram of circuit used to measure conductance of inversion layers as a function of bias voltage (after Statz et al.)*

decrease as will the conductance. The slow states will gradually come into equilibrium as the charge is transferred to them from the fast states, and ϕ_s and the conductance will return with a slow decay to their original values because of the large number of slow states and thus ϕ_s is independent of the bias voltage.

Using this type of experiment we see that the change in charge of the fast states may be found as a function of ϕ_s so that their densities and energy levels may be determined. Channel measurements were first carried out by Brown[60] and later by Kingston.[61] They have been highly developed by Statz et al.,[31,32,42] and have the advantage over the field effect that they may be used easily on silicon as well as germanium. However, only the states in the lower half of the energy gap may be determined with a P–N–P structure, an N–P–N structure being necessary for determinations in the upper half. Fast states near the middle of the energy gap are difficult to determine due to the very low channel conductance values occurring when the quasi-Fermi level is near their energy levels.

A simplified outline of the methods used by Statz et al. in determining the fast-state parameters will now be given. It is necessary to find ϕ_s from the measured values of G_s and V and also the total charge must be calculated from V and ϕ_s. For biases greater than about 1 volt we assume that all the space charge comes

from the ionized donors and using Poisson's equation we find (referring to Figure 3(b))

$$\left(\frac{d\psi}{dx}\right)^2 = -\frac{8\pi}{\kappa} q \int_{\psi_b}^{\psi_s} N_D \, d\psi$$

$$= \frac{8\pi}{\kappa} q N_D (\phi_{nb} + V - \phi_{ps})$$

Therefore $$\frac{Q_{sc}}{4\pi} = -\mathcal{E}_s = \left[\frac{8\pi}{\kappa} q N_D (\phi_{nb} + V - \phi_{ps})\right]^{1/2}$$

p' may be found by assuming E_s to be constant over the effective part of the channel

$$p' = n_i \int_0^\infty \exp\left[\frac{-q(\phi_{ps} - E_s x)}{kT}\right] dx$$

$$= \frac{kT}{q} \frac{n_i}{E_s} \exp(-q\phi_{ps}/kT)$$

μ_p' may be found from Schrieffer's theory using the expression for deep wells and we see that both p' and μ_p' are inversely proportional to E_s so that, generally

$$G_s \propto \frac{1}{|V| + |\phi_s| + |\phi_b|}$$

For the easy determination of ϕ_s, calculated curves may be drawn of G_s against V for various values of ϕ_s, and Q_{sc} may be found from curves of Q_{sc} against V as a function of ϕ_s.

Some experiments on N-type inversion layers on P-type germanium are now described. Initially the constancy of ϕ_s under steady state conditions is checked by making a series of conductance measurements with a series of fixed values of V. From the theoretical curves the value of ϕ_s may be found and plotted against V. Some typical experimental results are shown in Figure 22. For constant ϕ_s a horizontal line should result. Only very slight variation is found.

The fast-states measurements must be carried out sufficiently quickly for the slow states to have no effect. The reverse bias is suddenly raised from a low value to a higher value and the change in conductance noted. If raised to a sufficiently high value a condition of cut-off is obtained when the quasi-Fermi level has moved to near the middle of the energy gap so that there is a very low channel conductance. The charge in the slow states remains constant and at cut-off the fast states near the bottom of the energy gap are fully occupied with electrons. For lower values of V less negative charge is required in the fast states to neutralize the charge in the space charge layer and the quasi-Fermi level moves towards the valence band as the states fill with holes. At any value of V, therefore, ϕ_s is known from the conductance measurement and the total charge in the space charge region is determined by V and ϕ_s. The change in charge in the fast states as a function of

ϕ_s may therefore be plotted and typical results are shown in Figure 23. The experimental points are fitted with a theoretical curve and in this case the parameters are $E_t - E_i = 5\ kT$ and $N_t = 1{\cdot}05 \times 10^{11}$.

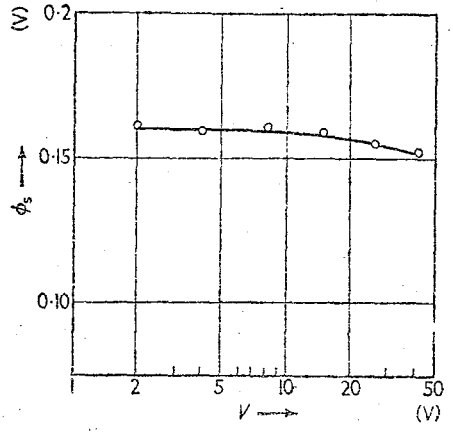


Figure 22. *Position of quasi-Fermi level at the surface as a function of the bias voltage under steady state conditions (after Statz et al.)*

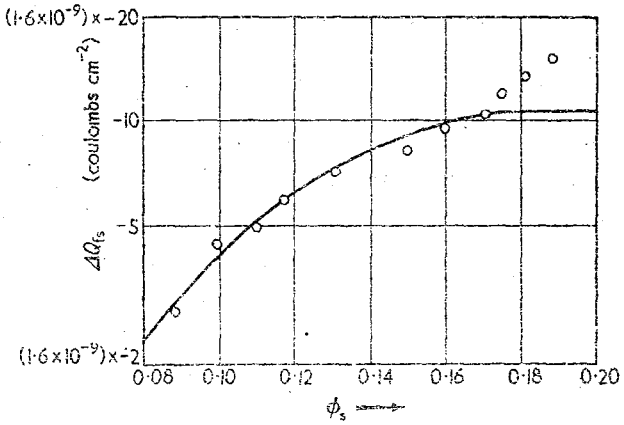


Figure 23. ΔQ_{fs} *as a function of the quasi-Fermi level at the surface (after Statz et al.)*

Statz et al. have made very comprehensive measurements of this nature on N–P–N and P–N–P structures in germanium and silicon. In silicon, the check on the constancy of ϕ_s with V under steady state conditions shows greater variation than in germanium. This may be explained by the thickness of the oxide layer on silicon. There may be a changing voltage across it due to the variation of charge in the slow states. This would give a different value of ϕ_s at the interface to that outside the oxide layer. The non-steady state measurements on silicon are made

easier by the longer relaxation times of the slow states which permit the use of a recorder instead of a point-by-point method as in germanium.

The results obtained may be summarized as follows. In germanium there are states with a density of about 10^{11} cm^{-2}, 5·5 kT below the centre of the energy gap, and states of density of 10^{10}–10^{11} cm^{-2}, 5 kT or 7 kT above the centre of the gap. These latter values depend on the state of oxidation of the surface. For silicon there are states of density 10^{12} cm^{-2}, 17 kT below the centre of the energy gap, and states of density 10^{11}–10^{12} cm^{-2}, 16 kT to 18·5 kT above the centre of the gap. The latter values again depend on the state of oxidation of the surface.

Channel effect measurements have also provided information on the physical structure of the fast states.[41] At high reverse voltages strong fields exist in the oxide layer and it has been shown that this can cause changes in the numbers and energy levels of the fast states. Models to explain these changes involving the motion of germanium and silicon ions in the oxide layer under the action of the field, have been given by Statz et al.[42]

Evidence for the motion of charge in the slow states has also been obtained from channel effect experiments. It was noticed that when steady state channel conductance measurements were made in the presence of vapours of certain liquids, water, acetone, droxane, etc., the conductance began to increase rather than decrease when the bias voltage was increased above a certain value. An explanation of this in terms of charge movement in the slow states has been investigated by Eriksen[62] and Statz and co-workers.[43] It was found that the application of an external transverse d.c. field, as in the field effect, to the channel of a P–N–P structure modulates the apparent channel conductance when the specimen is immersed in a polar liquid. Channel effect experiments on N–P–N structures in the presence of acetone vapour have shown complex transients in the conductance on the application of a bias voltage pulse. These have been quantitatively explained by considering the motion of charges in the absorbed vapour layer over the channel region. The charge moves between the middle and ends under the action of the voltage drop caused by the reverse current flowing in the channel. Mobilities not greater than 10^{-3} cm^2 V^{-1} sec^{-1} have been estimated.[43]

4.6. Measurements Associated with the Schrieffer Correction

The detailed interpretation of the majority of the experimental results in the previous section has involved a knowledge of the mobility in the surface barrier. Schrieffer's theory for modifying the mobility in deep wells from its bulk value has assumed there is completely diffuse scattering when a carrier collides with the surface. There is no good physical reason for this and experiments have been designed to test the validity of this assumption. Field effect experiments have been carried out by Millea and Hall.[63] The change in conductance of germanium and silicon filaments due to the known induced charge was measured. This change should be less than calculated using the bulk mobility value because of trapping in the surface states (fast states only in this case) and because of the Schrieffer correction. Graphs of conductance change against induced change were plotted and these were compared with theoretical curves neglecting trapping but taking the Schrieffer correction into account. Near the conductance minimum the experimental curve was below that of the theoretical because of trapping in the states. At larger barrier heights the experimental curve rose above the theoretical

curve. This shows that even though there may be a small amount of trapping the conductance change is greater than that calculated by Schrieffer with no trapping. It thus appears that not all the scattering is diffuse.

The Hall effect enables measurements to be made of carrier densities and mobilities. The theory and practice of this effect and also the magnetoresistance effect applied to surfaces have been developed by Petritz[64] and Zemel.[65] The presence of the bulk must be taken into account and Petritz has derived rigorous expressions for the Hall coefficient and magnetoresistance from the Boltzmann equations. Zemel has extended the Schrieffer correction calculations to include magnetic fields and has found that the Hall mobility differs by only about 15 per cent from the reduced mobility found by Schrieffer. Experiments by Zemel and Petritz[66] have shown that there is a variation of magnetoresistance and Hall effect with barrier height and the experimental data can be fitted to the theory. A three-carrier model with light and heavy holes is necessary and the existence of the diffuse scattering was confirmed. Further Hall effect measurements have been made by Missman and Handler[49] who have used the gradual oxidation of a clean surface to give a very large barrier height change. The existence of diffuse scattering was again confirmed but there were indications that some specular scattering was also occurring.

4.7. Other Measurements

In support of the measurements described above a great amount of work using other methods has been carried out. Here, to help to complete the picture, a small selection of recent work is given.

An experiment of Thomas and Rediker[67] using a reverse biased diode to measure the variation of s with ϕ_s has been extended and developed by Dousmanis.[68] A diagrammatic arrangement is shown in Figure 24. The diode is specially made and consists of a thin wafer of semiconductor with an alloy junction on one side and a free surface to which a transverse electric field may be applied on the other. In the reverse biased condition a substantial proportion of the saturation current arises from generation on the free surface and this is proportional to s. The application of an a.c. field will therefore modulate the field at the interface so that the barrier height and s are varied. The connections to the X and Y plates were arranged so that the variation of s with an applied field could be displayed on the oscilloscope. Dousmanis has used this method to determine the fast-state parameters. It is necessary to guess a value of c_p/c_n to fix the barrier height at which the maximum value of s occurs (see sub-section 3.3). A knowledge of the relationship between ϕ_s and the field at the surface E as in sub-section 3.1 then allows the oscilloscope traces to be transformed to s versus ϕ_s curves. Trapping in the fast states modifies the s versus E curves because then E arises from Q_{fs} as well as Q_{sc}. Effects from the fast states show up as kinks on the s versus E trace when the Fermi level is swept through the trap energy level and from this their number and energy level may be deduced and corrections to the ϕ_s versus E curve are made. The s versus ϕ_s curve is then compared with the theoretical one and if different another value of c_p/c_n is assumed. This method is independent of uncertainties due to the surface mobility correction, but determination of the barrier height depends on an accurate guess of c_p/c_n.

The surface photovoltage effect has been investigated by Johnson[69] as a method

of finding the barrier height without involving a knowledge of surface mobilities. When a semiconducting specimen is illuminated the quasi-Fermi levels in the bulk are governed by neutrality requirements. When the quasi-Fermi levels are extended across the barrier the space charge is varied and the barrier height must change so that the total surface charge is again zero. Normally the barrier tends to flatten and if the illumination is modulated at a frequency such that the slow states are inoperative the change in barrier height depends on the mean barrier height and the fast-state distribution. Johnson has measured the surface photovoltage in germanium using chopped light and a capacitance-coupled probe near the illuminated surface to measure the changes in contact potential. A large range of light intensities was employed and the results compared with theory. It was found that the changes in barrier height were as expected from changes in the charge in the space charge region and the effect of the fast states could be neglected. The method is thus not useful for the measurement of fast-state parameters but

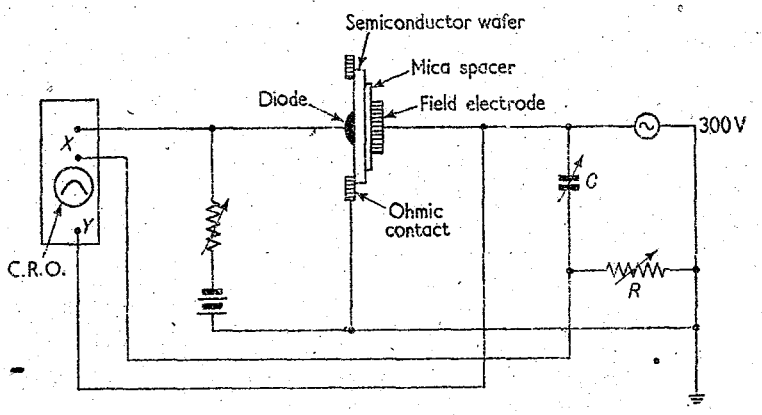


Figure 24. *Circuit for measurement of the effect of electric fields on surface recombination velocity (after Dousmanis)*

the barrier height may be determined to an estimated accuracy of about 0·026 eV providing the barrier has not too low a value.

A novel method of measuring fast-state parameters has been studied by Rupprecht.[70] A germanium P–N junction with a thin diffused N-type layer is used and connections are made so that the surface conductance of the N-type region may be measured. The thickness of the N-type region is less than a carrier diffusion length. The arrangement is shown in Figure 25. The reverse bias is suddenly charged by pulsing and produces a movement of the quasi-Fermi level at the surface of the N-type layer. The fast states are therefore no longer in equilibrium with the conduction and valence bands and they capture and release carriers with a characteristic time constant causing a variation of conductance of the N-type region. Measurement of this transient as a function of temperature enables the densities of states, their energy levels, and their capture cross-sections to be deduced. N- and P-type surfaces on the N-type region may be studied by varying the ambient atmosphere in the usual way. The state parameters found are in agreement with those of other workers.

As an alternative to varying the barrier height of germanium by the field effect or by ambient changes, use may be made of an electrolyte. This has been investigated by Brattain and Garrett[71] and it was found that the surface conductivity type was dependent on the polarity of the germanium relative to the solution. For example when the electrolyte is made negative a P-type surface results. This method has been extended and used for silicon by Harten.[72] The arrangement of Thomas and Rediker[67] with a silicon diode was used, the free surface being in contact with the electrolyte. Variation of the potential of the electrolyte produced changes in the reverse current corresponding to the usual shape of the s versus ϕ_s curve.

Much less work has been done on surface state measurements on silicon than germanium, due principally to the difficulty of obtaining large changes in barrier height. Buck and McKim[73] have used various etches to give different barrier heights on silicon and have modulated these heights by changing ambients.

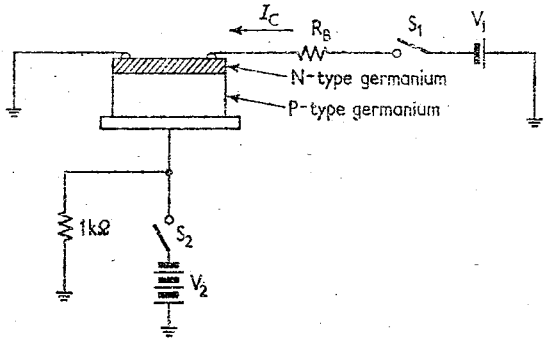


Figure 25. *Diagram of circuit for measurement of pulsed channel effect (after Rupprecht)*

Surface conductance and recombination measurements were carried out and complete curves obtained by piecing together the results for different etches. This enabled the fast-state parameters to be estimated. The estimations however are made uncertain by the possibility of the parameter values being a function of the etch.

Relatively little work has been carried out on the surface properties of other semiconductors. Some experiments have been made on lead sulphide which is of interest as a photoconductor.[74] Field effect measurements have shown that a surface model generally similar to that of germanium and silicon with fast and slow states may be postulated.

4.8. Measurements on Clean Surfaces

The study of clean surfaces is of great interest from the fundamental viewpoint because of the theoretically ideal conditions that are obtainable and because of the possibility of building-up controlled oxide layers. The cleaning of germanium surfaces was first investigated by Farnsworth and his co-workers[44,45] who developed a method based on argon bombardment and annealing under high vacuum. The crystal is subjected to prolonged outgassing at about 700° C under high vacuum

conditions, then bombarded with positive argon ions, and the resulting surface damage is annealed at temperatures around 500° C. The bombardment–annealing treatments may be cycled several times. An examination of the surfaces by low voltage electron diffraction then gives patterns that are consistent with an atomically clean condition. This technique has been used by a number of groups to prepare clean germanium surfaces for electrical measurements. The work function, photoconductance, surface conductance, field effect, and Hall effect[46–52] have been investigated and a qualitative picture of the electrical properties has emerged. The surface is very strongly P-type, almost independent of the bulk resistivity, and the estimated number of acceptor-type states is between 1×10^{13} and 4×10^{14} cm^{-2}. This is to be compared with the density of predicted Tamm states which is about 10^{15}. The admission of oxygen at low pressure changes the electrical properties. The surface becomes more P-type and may even become degenerate; then as the pressure of oxygen increases it becomes less P-type than its clean value. This behaviour is demonstrated by measurements of surface conductance and Hall effect.

It is found that the s.r.v. has a very high value after bombardment and annealing decreases it. To obtain the lowest values, prolonged annealing at high temperatures is necessary but the minimum values are higher than those of etched surfaces. The changes found by different groups on the exposure to oxygen at room temperature after annealing are conflicting, but it appears that provided the anneal is long enough and at a high temperature there is little change. However, for short anneals at lower temperature there is a substantial increase in s with oxygen pressure.[52]

Wang and Wallis[50] have made a study of ion bombarded surfaces with varying degrees of anneal using N- and P-type germanium. Their measurement techniques were similar to those described earlier in sub-section 4.2. It was found that the surface was N-type immediately after the ion bombardment and became increasingly P-type as the annealing treatment progressed. Their results are in general agreement with those of other workers but they have interpreted the details in terms of two sets of recombination centres, one set near the centre of the energy gap whose number appears to decrease with annealing and the other set near the valence band which affects s mainly through changes of barrier height.

As an alternative to the above method of obtaining clean surfaces, Barnes and Banbury[75] have cleaved germanium under high vacuum and made electrical measurements on the exposed surface. D.C. and a.c. field effect and photoconductance measurements have been made, but due to the possibility of the existence of stray fields between the electrode and the uncleaved surfaces no clear cut picture of the properties of a cleaved surface has yet emerged.

5. APPLICATIONS

The importance of surface properties on diode and transistor parameters has long been realized and it has been shown by the direct application of the field effect to devices that the behaviour of surfaces near P–N junctions is explicable in terms of the free surface model.[76,77]

Surface properties are involved in devices in a number of ways. In the point-contact transistor the behaviour of a metal wire pressed against a surface was a

basic consideration although this was complicated by complex conditions in the neighbouring bulk material. Devices using more controllable metal–semiconductor contacts have been studied and the results have been interpreted in terms of surface properties.[78–81] In the usual type of alloy junction transistor the current amplification factor is largely controlled by the properties of the surface around the emitter and the reverse current at the collector may be greatly influenced by surface conditions nearby. The breakdown voltage, too, depends on the surface barrier height near the collector.[82] It has also been found that the level of $1/f$ noise in semiconductor filaments and devices is greatly dependent on the surface treatment so that it must be considered as a basic surface property.[83–85]

With the growth of knowledge of surfaces by analytical methods, interest is now turning to the tailoring of surfaces for devices. The surface may be specified to have desirable electrical properties and these should remain stable with changes of ambient temperature and atmosphere.

5.1. Metal–Semiconductor Contacts

Although, in general, point-contacts have proved too complex to be tractable to theory some success has been achieved in understanding large-area plated and evaporated contacts.

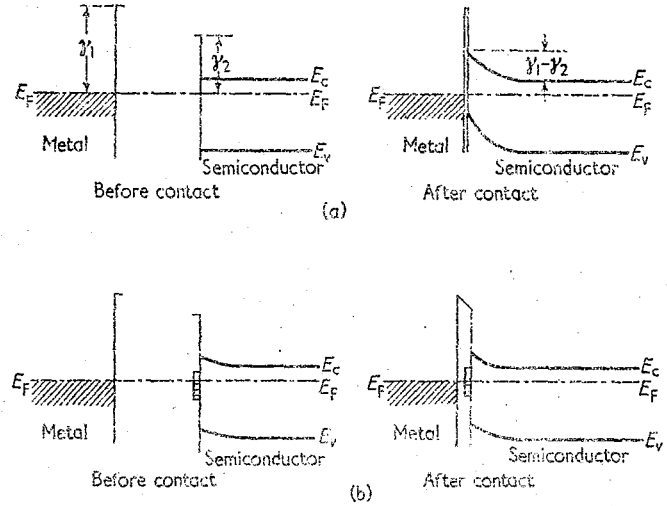


Figure 26. *Diagram of metal–semiconductor contact with and without surface states*

When a metal and a semiconductor with no surface states are brought into contact, the barrier height is determined by the difference in work functions. When apart, there is a contact potential difference equal to the difference in the work functions. When together, the charge in the space charge region is neutralized by an equal and opposite charge on the surface of the metal. This is illustrated in Figure 26(a). To explain cases where experimental results did not show dependence on the difference in work functions of various metals Bardeen[7] invoked

surface states. In this way small changes in barrier height can give large changes in surface charge so that the contact potential difference is largely dropped across the small spacing between the metal and semiconductor. This is shown in Figure 26(b). The barrier height is thus relatively unchanged by bringing the metal into contact with it. Experimental and theoretical work on germanium has been carried out using plated, evaporated, and painted contacts by Gunn[79, 80] and Bocciarelli[81] and good agreement obtained. The results show that the rectification properties are chiefly a function of the semiconductor surface and it appears that an inversion layer is necessary for good rectification.

Not in all cases has the barrier height been independent of the contact potential and Harrick[86] has investigated this on germanium using the infra-red absorption by free carriers at the surface to measure barrier height. Changes in barrier height as it was approached by a metal were found.

5.2. The Effect of the Surface on Reverse Currents

It is found experimentally that the reverse current of a P–N junction is highly dependent on the surface conditions in the neighbourhood of the junction.

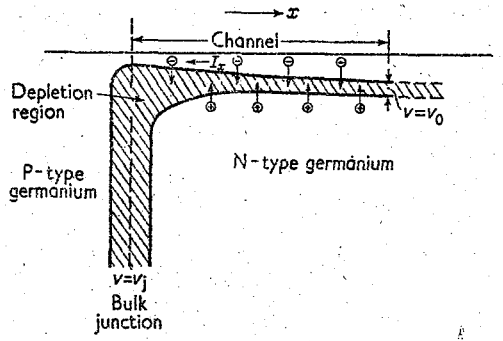


Figure 27. *Diagram of channel on N-type germanium*

According to the simple theory of Shockley[87] the reverse current is due to the generation of electron–hole pairs within a diffusion length of the junction, but in practice a number of complicating factors have to be taken into account.

When a reverse bias is suddenly applied the reverse current does not immediately assume its theoretical value but a number of transients occur which may be associated with changing surface conditions. The shorter transients of the order of seconds appear to be due to the slow states near the junction coming into equilibrium[88] and the larger transients with durations of minutes are caused by the formation of channels.[89, 90] In addition the current is substantially increased by the presence of an ambient atmosphere of water vapour or vapours of organic liquids such as acetone, dioxane, methyl alcohol.[90, 91]

Considerable work has been done on the channel effect and its basic principles are now fairly well understood. A diagrammatic representation of a channel on a single P–N junction is given in Figure 27. A P-type inversion layer is formed on the N-type region so that the junction is effectively extended over the surface of

the N-type region. Electron–hole pairs are generated at the surface of the inversion layer, the electron passing immediately to the N-type region and the hole flowing along the channel to the P-type region. At the beginning of the channel the voltage drop between the inversion layer surface and the N-type region is the same as that across the junction in the bulk. The energy level diagram is similar to that of the case of Figure 3(b) examined in sub-section 3.1. Due to the voltage drop caused by the longitudinal current, the voltage between the channel and the bulk decreases and the channel will extend until the reverse bias is less than about kT/q volts when the full saturation current will no longer be taken.

The simplified theory of sub-section 4.5 may be applied. The channel resistance is given by

$$R = B(|V|+|\phi_s|+|\phi_b|)$$

where B is a constant and ϕ_s, clamped by the slow states, is taken as constant along the channel.

Letting J_c = generated hole current per unit length of channel, the longitudinal channel current at x is

$$I_x = J_c(l-x)$$

where l is the length of the channel. Then

$$\frac{dV}{dx} = -RI_x$$

Integrating gives for the channel length

$$l = \left[\frac{2}{J_c B}\ln\frac{|V_j|+|\phi_s|+|\phi_b|}{|V_0|+|\phi_s|+|\phi_b|}\right]^{1/2}$$

where V_j is the voltage across the bulk junction and V_0 is the voltage at the end of the channel ($\simeq kT/q$). The reverse current is thus

$$I = J_c l = \left[\frac{2J_c}{B}\ln\frac{|V_j|+|\phi_s|+|\phi_b|}{|V_0|+|\phi_s|+|\phi_b|}\right]^{1/2}$$

A more rigorous analysis by Eriksen et al.[91] taking into account the presence of charge carriers as well as the ionized impurities results in a variation of current with voltage which is more nearly logarithmic.

A number of experiments have given results which are consistent with theory. The length of the channel may be measured by the photo response produced when the surface is scanned by a light spot.[92] While on this channel a photocurrent component is produced in the reverse current and at either end the photocurrent falls towards zero when the light spot is more than a diffusion length from the end. By this means it has been shown that the reverse current varies linearly with channel length, and a logarithmic behaviour with voltage has been found. The growth of channels on the sudden application of a voltage takes place with time constants of the order of minutes and would seem to be associated with the slow states. On the other hand some very rapid decays of the length of the channel have been observed when the bias is reduced and no good explanation for this has so far been given.[93]

The additional leakage current which is observed in the presence of water vapour and other condensible vapours varies linearly with voltage. This current usually becomes dominant at relative humidities greater than about 50 per cent and voltages greater than about 10 V. Initially the current was thought to be due to ionic conduction in the absorbed layers, but careful experiments have shown that there is no evolution of gas or mass transfer from one side of the junction to the other.[91] It was then thought to be of an electronic nature and studies by Eriksen[62,91] and Statz et al.[42] on germanium and silicon N–P–N structures and diodes in the presence of water vapour and acetone indicated mobilities in acetone of about 50 $cm^2 V^{-1} sec^{-1}$. The more recent work by Statz et al.[43] described at the end of sub-section 4.5, however, has cast doubts on this simple picture. The experiments on the transients in N–P–N structures in acetone vapour have indicated mobilities not greater than 10^{-3} $cm^2 V^{-1} sec^{-1}$. Very little conduction in the absorbed layers is therefore thought possible. Effects due to absorbed vapours may therefore be due to the effect of a redistribution of charge in the layers over the channel rather than a straightforward conduction process.

When the voltage across a P–N junction is raised to a large value, breakdown effects occur giving rise to a larger reverse current.[82] The breakdown voltage may be smaller than that for the bulk material and is a function of the surface conditions. This effect has been investigated by Garrett and Brattain.[94] Using a light-spot technique on an N^+–P alloyed junction they demonstrated that under breakdown conditions multiplication of the photocurrent took place at the surface. This showed that surface breakdown, like usual cases of bulk breakdown, is an avalanche process. They also found that the breakdown occurred in localized regions of the surface which is again similar to the behaviour of bulk breakdown. The highest breakdown voltages occurred when channels were present on the P-type region although the values were lower than that of the true bulk breakdown value. The lowest values occurred when the gaseous ambient tended to give an accumulation layer on the P-type region. This is understandable by considering that the surface charge is then negative which will give a very high field from the donors in the N^+ region to the surface around the perimeter of the junction. Statz et al.[43] in their experiments on channels in N–P–N structures have investigated the occurrence of breakdown voltages lower than the bulk values. An explanation is given in terms of the movement of charges in the slow states away from the ends of the channel which will result in high fields between the end N-type regions and the surface at the ends of the P-type region. As an alternative the impact ionization of the fast states is considered as this would occur at a relatively low voltage.

5.3. 1/*f* Noise

1/*f* noise occurs very commonly when a current flows in semiconducting substances. It is characterized by being nearly inversely proportional to the frequency and directly proportional to the square of the current. It extends over a remarkable frequency range and has only a small temperature dependence. No completely satisfactory explanation of the mechanism has been proposed in any instance of its occurrence.

It is observed in germanium and silicon single crystals and devices and is found to be very sensitive to surface treatment and ambient atmosphere.[83–85] This

establishes it as a surface property and, with the growth of detailed knowledge of the surface, theories have been expounded to explain it. The most comprehensive theory is that of McWhorter.[95] During experiments with Kingston he found that the time constants of the slow states had a range of values and a distribution suitable for processes which could give a $1/f$ spectrum over a large frequency range. For example a process which decays exponentially with a time constant λ can give a spectrum of the form $\lambda/(1+\omega^2\lambda^2)$. If the number of sources with time constant λ is proportional to $1/\lambda$ when $\lambda_1 > \lambda > \lambda_2$, a $1/f$ spectrum results when $\omega\lambda_2 \ll 1 \ll \omega\lambda_1$ holds. McWhorter has calculated the statistical fluctuations of the charge in the slow states due to the trapping of carriers. This causes variations of the surface barrier height and hence fluctuations in conductance. Using the measured values of the time constants and the calculated values of the fluctuations he showed that noise with a $1/f$ spectrum and the correct order of magnitude was possible by this mechanism. It was also postulated that the communication between the slow states and the bulk was by tunnel penetration which would give a temperature-independent noise mechanism. Morrison[96] has shown also that a $1/f$ spectrum may arise from the slow states by the use of equation (18) which effectively gives a range of time constants.

A major prediction of these theories is that the noise is caused by fluctuations of the surface barrier height. On a germanium filament, therefore, the noise level should have a minimum value at the surface conductance minimum because here small changes in barrier height produce no effect on the conductance. In devices where excess carriers are present there should be a noise minimum at the top of the s versus ϕ_s curve as barrier height fluctuations produce no variations in the number of excess carriers. Experiments have been designed to check this prediction.[97, 98] The barrier height may be varied by slowly changing the ambient atmosphere or by applying a field and using the long decay time of the slow states. Up to the present the noise minimum has not been observed so that the association of noise with fluctuations of barrier height may not be valid. Present experimental results show that as the surface barrier changes from that giving an accumulation layer to that giving an inversion layer there is a marked increase in noise.

5.4. Stable Surfaces

The surface of a device should have certain properties. In a transistor the surface of the base should have a low number of recombination centres so that the current amplification factor is large and the reverse current small. Extremes of barrier height should be avoided as inversion layers give channels and accumulation layers lower the breakdown voltage. It is desirable to have large values of $E_t - E_i$ to minimize changes of s with barrier height (Figure 9). These properties should remain stable with time and two approaches may be made to achieve this. In one the ambient atmosphere may be controlled so that changes in barrier height due to alterations in the slow states are avoided. In the other the surface may be shielded from the effects of changing gaseous ambient by means of a protective layer. Although the first approach seems straightforward by the use of encapsulation techniques, it is difficult to control the water vapour content inside the envelope because of the absorption of water vapour molecules on the internal walls. The other approach has been pursued by studying the effect of thick oxide layers on the surface properties of germanium and silicon.

In the experiments of Lasser, Wysocki, and Bernstein,[40, 41] oxide layers of about 0·25 micron thickness have been formed on germanium filaments by prolonged heating in oxygen. D.C. field effect measurements gave immeasurably long decay times, and changes in ambient, with the exception of water vapour, produced no changes in barrier height. The effect of water vapour was marked. High relative humidities and exposure periods of one hour gave quite short decay times on even the thickest oxide layers. The surface, however, could be restored by dessication showing no permanent damage had occurred. The oxidation consistently produced N-type surfaces on both N- and P-type crystals. A combination of two factors is suggested to explain the protective action. The communication between the bulk and the slow states is greatly reduced and therefore the time for the slow states to reach equilibrium is increased. Also because of the thick oxide layer the potential drop across the layer produced by charge variations due to changes in the slow states is greatly increased so that large potential changes of the outer surface of the oxide layer relative to the Fermi level may occur with comparative small changes in the internal surface barrier height.

More recently a very intensive study of thermally grown oxides on silicon has been reported by Atalla et al.[99] Both filaments and diodes were investigated, the oxide layer thickness being up to the order of microns. Very stable surfaces were produced, the d.c. field effect giving an infinite decay time and there were no changes with different ambient atmospheres, including that of water vapour of 40 per cent relative humidity. Fast-state measurements were made on the filaments and gave a characteristic structure which depended on pre-oxidation treatment and certain impurities in the crystal. Both N- and P-type surfaces could be produced. The diodes retained stable characteristics when left for periods of months in room air and it was found also that the $1/f$ noise was considerably reduced by the oxidation treatment. It therefore seems possible that in the not too distant future the surface may become a designable part of devices.

6. CONCLUSIONS

In recent years the basic surface model of germanium and other semiconductors has been well established but there are many aspects which are still obscure.

The parameters of the fast states are fairly well defined for etched germanium surfaces and silicon may be qualitatively similar. Less is known about the fast states of other semiconductors. In germanium there are four sets of states with discrete energy levels, two near the centre of the gap about 3 kT above and 1 kT below the centre. The two outer sets are about $\pm$ 6 kT from the centre. The density of each set is of the order of 10^{11} cm^{-2} and only the inner sets seem to be active in recombination and have larger capture cross-sections for holes than electrons. In silicon, two outer sets 18·5 kT above and 17 kT below the centre of the gap have been found with densities about 10^{12} cm^{-2}. Any states near the centre have not yet been definitely determined. The physical structure and the dependence of the parameters of the fast states on the state of oxidation of the surface is not yet clear. Their behaviour at low temperature is not completely understood and it would appear that more work is necessary on the high frequency field effect. Regarding the slow states the chief problems seem to be the mechanism of their communication with the bulk and whether there are any states within the oxide layer.

The results obtained from experiments on clean surfaces are consistent with the states being of the Tamm type although this has not yet been quantitatively established.

In the interpretation of surface measurements the corrected mobility in surface wells is often involved and this has not yet been definitely experimentally ascertained.

On the more applied aspects of surfaces the outstanding problems are the causes of $1/f$ noise and the tailoring of surfaces by the control of an oxide or other layer.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the very great help he has received from Dr. F. Berz, with whom he has had many valuable discussions. Dr. K. Hoselitz has read the manuscript and suggested a number of improvements.

REFERENCES

1. I. Tamm. *Phys. Z. Sowjet.* **1**, 733 (1932)
2. R. H. Fowler. *Proc. roy. Soc.* **A141**, 56 (1933)
3. S. Rijanow. *Z. Phys.* **89**, 806 (1934)
4. A. W. Maue. *Z. Phys.* **94**, 717 (1935)
5. E. T. Goodwin. *Proc. Camb. phil. Soc.* **35**, 205, 221, 232 (1938)
6. W. Shockley. *Phys. Rev.* **56**, 317 (1939)
7. J. Bardeen. *Phys. Rev.* **71**, 717 (1947)
8. W. Shockley and G. L. Pearson. *Phys. Rev.* **74**, 232 (1948)
9. J. Bardeen and W. H. Brattain. *Phys. Rev.* **74**, 230 (1948)
10. R. H. Kingston. *J. appl. Phys.* **27**, 101 (1956)
11. W. H. Brattain and J. Bardeen. *Bell Syst. Tech. J.* **32**, 1 (1953)
12. R. H. Kingston and S. F. Neustadter. *J. appl. Phys.* **26**, 718 (1955)
13. R. Stratton. *Proc. phys. Soc. Lond.* **B68**, 746 (1955)
14. R. Seiwatz and M. Green. *J. appl. Phys.* **29**, 1039 (1958)
15. J. R. Schrieffer. *Phys. Rev.* **97**, 641 (1955)
16. J. R. Schrieffer. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
17. D. T. Stevenson and R. J. Keyes. *Physica* **20**, 1041 (1954)
18. C. G. B. Garrett and W. H. Brattain. *Phys. Rev.* **99**, 376 (1955)
19. G. G. E. Low. *Proc. phys. Soc. Lond.* **B69**, 1331 (1956)
20. W. Shockley and W. T. Read. *Phys. Rev.* **87**, 835 (1952)
21. F. Berz. *Proc. phys. Soc. Lond.* **71**, 275 (1958)
22. S. R. Morrison. *J. phys. Chem.* **57**, 860 (1953)
23. J. Bardeen and S. R. Morrison. *Physica* **20**, 873 (1954)
24. A. Many, E. Harnik, and Y. Margoninski. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
25. A. Many and D. Gerlick. *Phys. Rev.* **107**, 404 (1957)
26. J. Bardeen, R. E. Coovert, S. R. Morrison, J. R. Schrieffer, and R. Sun. *Phys. Rev.* **104**, 47 (1956)
27. H. C. Montgomery and W. L. Brown. *Phys. Rev.* **103**, 865 (1956)
28. W. H. Brattain and C. G. B. Garrett. *Bell Syst. Tech. J.* **35**, 1019 (1956)
29. W. L. Brown, W. H. Brattain, C. G. B. Garrett, and H. C. Montgomery. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
30. S. Wang and G. Wallis. *Phys. Rev.* **105**, 1459 (1957)
31. H. Statz, G. A. De Mars, L. Davis, Jr., and A. Adams, Jr. *Phys. Rev.* **101**, 1272 (1956)
32. H. Statz, G. A. De Mars, L. Davis, Jr., and A. Adams, Jr. *Phys. Rev.* **106**, 455 (1957)
33. A. Many and D. Gerlich. *Phys. Rev.* **107**, 404 (1957)
34. E. Harnik and Y. Margoninski. To be published
35. S. Wang and G. Wallis. *Phys. Rev.* **107**, 947 (1957)
36. A. Many. *J. Phys. Chem. Solids* **8**, 87 (1958)

37. S. R. MORRISON. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
38. S. R. MORRISON. *Phys. Rev.* **102**, 1297 (1956)
39. R. H. KINGSTON and A. L. MCWHORTER. *Phys. Rev.* **103**, 534 (1956)
40. M. LASSER, C. WYSOCKI, and B. BERNSTEIN. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
41. M. LASSER, C. WYSOCKI, and B. BERNSTEIN. *Phys. Rev.* **105**, 491 (1957)
42. H. STATZ, G. A. DE MARS, L. DAVIS, JR., and A. ADAMS, JR. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
43. H. STATZ and G. A. DE MARS. *Phys. Rev.* **111**, 169 (1958)
44. H. E. FARNSWORTH, R. E. SCHLIER, T. H. GEORGE, and R. M. BURGER. *J. appl. Phys.* **26**, 252 (1955)
45. R. E. SCHLIER and H. E. FARNSWORTH. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
46. J. A. DILLON, JR. and H. E. FARNSWORTH. *J. appl. Phys.* **28**, 174 (1957)
47. J. T. LAW and C. G. B. GARRETT. *J. appl. Phys.* **27**, 656 (1956)
48. P. HANDLER. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
49. R. MISSMAN and P. HANDLER. *J. Phys. Chem. Solids* **8**, 109 (1958)
50. G. WALLIS and S. WANG. *Trans. Electrochem. Soc.* **106**, 231 (1959)
51. J. A. DILLON and H. E. FARNSWORTH. *J. appl. Phys.* **29**, 1195 (1958)
52. H. H. MADDEN and H. E. FARNSWORTH. *Phys. Rev.* **112**, 793 (1958)
53. S. WANG and G. WALLIS. *Phys. Rev.* **107**, 947 (1957)
54. G. WALLIS and S. WANG. *Trans. electrochem. Soc.* **106**, 231 (1959)
55. P. C. BANBURY, CAJKOVSKI, and J. D. NIXON. *Bull. Amer. phys. Soc. Ser.* II, **3**, 376 (1958)
56. S. R. MORRISON. *Phys. Rev.* **114**, 437 (1959)
57. H. C. MONTGOMERY. *Phys. Rev.* **106**, 441 (1957)
58. C. G. B. GARRETT. *Phys. Rev.* **107**, 478 (1957)
59. F. BERZ. *J. Electron. Control* **6**, 97 (1959), and work to be published.
60. W. L. BROWN. *Phys. Rev.* **91**, 518 (1953)
61. R. H. KINGSTON. *Phys. Rev.* **98**, 1766 (1955)
62. W. T. ERIKSEN. *J. appl. Phys.* **29**, 730 (1958)
63. M. F. MILLEA and T. C. HALL. *Phys. Rev. Letters* **1**, 276 (1958)
64. R. L. PETRITZ. *Phys. Rev.* **110**, 1254 (1958)
65. J. N. ZEMEL. *Phys. Rev.* **112**, 762 (1958)
66. J. N. ZEMEL and R. L. PETRITZ. *Phys. Rev.* **110**, 1263 (1958)
67. J. E. THOMAS and R. H. REDIKER. *Phys. Rev.* **101**, 984 (1956)
68. G. C. DOUSMANIS. *Phys. Rev.* **112**, 369 (1958)
69. E. O. JOHNSON. *Phys. Rev.* **111**, 153 (1958)
70. G. RUPPRECHT. *Phys. Rev.* **111**, 75 (1958)
71. W. H. BRATTAIN and C. G. B. GARRETT. *Bell Syst. Tech. J.* **34**, 129 (1955)
72. H. U. HARTEN. Private communication
73. T. M. BUCK and E. S. MCKIM. *Trans. electrochem. Soc.* **105**, 1 (1958)
74. R. L. PETRITZ, F. L. LUMMIS, H. E. SORROWS, and J. F. WOODS. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
75. G. A. BARNES and P. C. BANBURY. *J. Phys. Chem. Solids* **8**, 111 (1958)
76. J. R. A. BEALE, D. THOMAS, and T. B. WATKINS. *Proc. Phys. Soc. Lond.* **72**, 910 (1958)
77. J. H. FORSTER and H. S. VELORIC. *J. Appl. Phys.* **30**, 906 (1959)
78. W. E. BRADLEY et al. *Proc. Inst. Radio Engrs, N.Y.* **41**, 1702 (1953)
79. J. B. GUNN. *Proc. Phys. Soc. Lond.* **B67**, 409 (1954)
80. J. B. GUNN. *Proc. Phys. Soc. Lond.* **B67**, 575 (1954)
81. C. V. BOCCIARELLI. *Physica* **20**, 1020 (1954)
82. H. BERNSTEIN and R. H. KINGSTON. *Phys. Rev.* **98**, 1566(A) (1955)
83. H. C. MONTGOMERY. *Bell Syst. Tech. J.* **31**, 950 (1952)
84. T. G. MAPLE, L. BESS, and H. A. GEBBIE. *J. appl. Phys.* **26**, 490 (1955)
85. A. SLOCUM and J. N. SHIVE. *J. appl. Phys.* **25**, 406 (1954)
86. N. J. HARRICK. *J. Phys. Chem. Solids* **8**, 106 (1958)
87. W. SHOCKLEY. *Bell Syst. Tech. J.* **28**, 435 (1949)
88. T. B. WATKINS. *Proc. Phys. Soc. Lond.* **B69**, 1353 (1956)
89. A. L. MCWHORTER and R. H. KINGSTON. *Proc. Inst. Radio Engrs, N.Y.* **42**, 1376 (1954)
90. J. T. LAW. *Proc. Inst. Radio Engrs, N.Y.* **42**, 1367 (1954)
91. W. T. ERIKSEN, H. STATZ, and G. A. DE MARS. *J. appl. Phys.* **28**, 133 (1957)
92. H. CHRISTENSEN. *Phys. Rev.* **96**, 827 (1954)
93. A. F. PLUMMER. *Proc. Phys. Soc. Lond.* **B69**, 539 (1956)
94. C. G. B. GARRETT and W. H. BRATTAIN. *J. appl. Phys.* **27**, 299 (1956)

95. A. L. McWHORTER. *Semiconductor Surface Physics* (University of Pennsylvania Press, Philadelphia, 1957)
96. S. R. MORRISON. *Phys. Rev.* 99, 1404 (1955)
97. T. B. WATKINS. *Proc. Phys. Soc. Lond.* 73, 59 (1959)
98. A. U. MACRAE. Private communication
99. M. M. ATALLA, E. TANNENBAUM, and E. J. SCHEIBNER. *Bell Syst. Tech. J.* 38, 749 (1959)

THE ABSORPTION EDGE SPECTRUM OF SEMICONDUCTORS

T. P. McLEAN, B.Sc., Ph.D.

Royal Radar Establishment, Malvern

MS. received July 1959

THE ABSORPTION EDGE SPECTRUM OF SEMICONDUCTORS

1. INTRODUCTION

A feature of the optical absorption spectrum common to all semiconductors is the rapid increase in absorption which occurs over a small energy range when the absorbed radiation has an energy roughly equal to the energy gap of the semiconductor. This is called the absorption edge of the material, and in germanium and silicon, for example, it covers a fraction of an electron volt within which the absorption coefficient increases by three to four orders of magnitude. This large increase in absorption is, of course, due to the onset of absorption in which electrons are raised from the valence band across the forbidden energy gap to the conduction band, so that a study of the position of this edge and any internal structure it may have will yield information about the energy gap and the properties of those electron states just above it at the conduction band edge and below it at the valence band edge. This type of information is extremely useful as it is just these states close to the conduction and valence band edges which determine the electrical properties of the semiconductor. Consequently, measurement and analysis of the absorption edge spectrum of a semiconductor is an important feature in the study of its properties.

This spectrum has been studied in many semiconductors, a typical but far from comprehensive list including germanium,[1–3] silicon,[3,4] indium antimonide and gallium antimonide,[5] the lead salts lead sulphide, lead selenide, and lead telluride,[6] silicon carbide,[7] and diamond.[8] The most extensive and profitable work up to the present has been carried out on germanium and silicon[1–4] and we intend in this article to concentrate most of our attention on the absorption edge of these two substances, although most of the theory and much of the general discussion to be given will apply to any semiconductor. Much work has been carried out recently, mostly in germanium, in the study of the effects of a uniform magnetic field on the position and structure of the absorption edge. This work, however, will be discussed in an accompanying article by B. Lax and we shall not pursue it in any detail here.

Having defined the scope of the article, something of what might be called the historical background to the subject is appropriate at this point. If one considers, as we shall in sub-section 2.1, the electronic transitions induced by the interaction between the electrons in a semiconductor and optical or infra-red radiation, then one finds that in these transitions the momentum, or more accurately the wave vector, of the electrons does not change. This has no significant effect on the position of the absorption edge of a semiconductor with a simple band structure, i.e. with conduction and valence band extrema at the centre of the first Brillouin zone, since transitions between the valence and conduction band edges are possible without any change of the electron wave vector. However, as is well known now and was indicated in 1954 by the calculations of Herman and Callaway[9,10] and by

preliminary cyclotron resonance experiments,[11,12] the conduction and valence band extrema do not lie at the same points in **k**-space in germanium. It was therefore puzzling that the measured absorption edge in germanium[13] should begin at an energy $\simeq 0{\cdot}7$ eV, i.e. a wavelength $\simeq 1{\cdot}8$ microns, as this energy corresponds to the energy gap between these extrema deduced from conductivity and Hall constant measurements[14] and so is associated with transitions in which the electron wave vector must change. This was resolved by Hall, Bardeen, and Blatt[15] who pointed out that transitions could be made by the electrons in which their wave vector is changed by taking into account the interaction between the electrons and the lattice vibrations of the crystal. Transitions of this type, involving the simultaneous interaction of the electrons with the electromagnetic radiation and lattice vibrations, they called indirect as opposed to the direct transitions involving only the interaction between the electrons and radiation. Transitions of this indirect type could clearly explain the rise of the absorption edge in germanium at an energy corresponding to that of the energy gap. The first detailed comparison between this theory of Hall, Bardeen, and Blatt and the experimentally determined absorption around the absorption edge was carried out for germanium by Macfarlane and Roberts[16] who also realized the importance of the finite energy of the lattice vibrations involved in the indirect transitions, a factor which had been previously neglected. They were able to explain the available experimental data by the theory obtaining values for the energy gap in germanium as a function of temperature and also some information about the lattice vibrational energies in germanium. They also carried out a similar study of silicon[17] and found that its absorption edge was due to indirect transitions. Similar studies have been carried out in several semiconductors some of which are given in references 5–8.

Recently, however, the study of the absorption edge has become a much more powerful tool in the investigation of the properties of semiconductors due to the development of means whereby the measurement of the absorption can be made in much greater detail than was previously possible. This is the result of the development of techniques for the production of diffraction gratings suitable for use in the infra-red region of the spectrum. The use of such a diffraction grating in place of the prism previously used increases the resolving power of the optical system used in measuring the absorption as a function of wavelength around 1 micron by about an order of magnitude. The measurement of the absorption edge in germanium[1,2] and silicon[4] with this higher resolving power revealed a wealth of detailed structure in the edge, previously unresolved, and which, when analysed, gave a great deal of information about these substances. Having developed the theory of direct and indirect transitions in Section 2 we shall devote the remainder of this article to a discussion of this analysis and the information obtained from it.

2. THEORY

2.1. Direct Transitions

We consider first of all the absorption of electromagnetic radiation by a semiconducting crystal due solely to the interaction between the radiation and the electrons in the crystal, and we shall assume that the crystal is perfect so that we

can neglect the effects of electrons trapped in impurities or defects. The interaction between the radiation and electrons is described by the Hamiltonian

$$H_{\text{rad}} = -\frac{ie\hbar}{mc}\mathbf{A}.\text{grad} \qquad \ldots(1)$$

e being the electronic charge, m the free electron mass, c the velocity of light, $\hbar$ Planck's constant divided by 2π, and $\mathbf{A}$ the vector potential characterizing the radiation. The electronic states in the crystal are characterized by a band suffix n and a wave vector $\mathbf{k}$ lying inside the first Brillouin zone such that an electron whose wave function is the Bloch function $\psi_{n\mathbf{k}}(\mathbf{r}) = u_{n\mathbf{k}}(\mathbf{r})\exp(i\mathbf{k}.\mathbf{r})$, $u_{n\mathbf{k}}(\mathbf{r})$ having the lattice periodicity, has energy $\epsilon_n(\mathbf{k})$. The only electronic transitions (accompanied by the absorption of a quantum of radiation) which the interaction (1) can induce transfer an electron from a state with wave function $\psi_{i\mathbf{k}_1}(\mathbf{r})$ to one with wave function $\psi_{j\mathbf{k}_2}(\mathbf{r})$. The probability that such a transition will take place is governed by the square of the matrix element

$$\int d^3r\,\psi^*_{j\mathbf{k}_2}(\mathbf{r})\,H_{\text{rad}}\psi_{i\mathbf{k}_1}(\mathbf{r})$$
$$= -\frac{ie\hbar}{mc}\int d^3r\,u^*_{j\mathbf{k}_2}(\mathbf{r})\exp(-i\mathbf{k}_2.\mathbf{r})\,\mathbf{A}.\text{grad}\,u_{i\mathbf{k}_1}(\mathbf{r})\exp(i\mathbf{k}_1.\mathbf{r}) \qquad \ldots(2)$$

For radiation of wave vector $\mathbf{q}$ and angular frequency ω

$$\mathbf{A} = A_0\boldsymbol{\alpha}\exp(i\mathbf{q}.\mathbf{r}), \qquad |A_0|^2 = \frac{2\pi\hbar\omega N}{q^2} \qquad \ldots(3)$$

where $\boldsymbol{\alpha}$ is a unit vector in the direction of polarization of the radiation and N is the number of quanta of radiation of energy $\hbar\omega$ per unit volume. Using equation (3), equation (2) can be reduced to

$$-A_0\frac{ie\hbar(2\pi)^3}{mc\,\Omega}\int_c d^3r\,u^*_{j\mathbf{k}_2}(\mathbf{r})\,\boldsymbol{\alpha}.\text{grad}\,u_{i\mathbf{k}_1}(\mathbf{r})\,\delta(\mathbf{k}_2-\mathbf{k}_1-\mathbf{q}) \qquad \ldots(4)$$

where Ω is the volume of a unit cell of the crystal and the subscript c on the integration denotes that the integration is to be taken through this volume. The radiation of interest to us lies in the infra-red region, so that its wavelength is much larger than the typical lattice spacing of crystals which is a few Ångström units. Hence the electronic wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$, being of the order of the reciprocal lattice spacing, are much larger than the wave vector $\mathbf{q}$ of the radiation of interest. We can therefore neglect $\mathbf{q}$ in the δ-function appearing in equation (4). This amounts to saying that the radiation gives a negligible momentum to the electron. Defining

$$\mathbf{p}_{ji}(\mathbf{k}_1) = -i\hbar\frac{(2\pi)^3}{\Omega}\int_c d^3r\,u^*_{j\mathbf{k}_1}(\mathbf{r})\,\text{grad}\,u_{i\mathbf{k}_1}(\mathbf{r}) \qquad \ldots(5)$$

we find that

$$\int d^3r\, \psi^*_{\mathrm{j}\mathbf{k}_2}(\mathbf{r})\, H_{\mathrm{rad}}\, \psi_{\mathrm{i}\mathbf{k}_1}(\mathbf{r}) = A_0 \frac{e}{mc} \boldsymbol{\alpha}.\mathbf{p}_{\mathrm{ji}}(\mathbf{k}_1)\, \delta(\mathbf{k}_2 - \mathbf{k}_1) \qquad \ldots (6)$$

the electron wave vector being unchanged by the transition.

It follows from equation (6) that the number of transitions occurring per unit time per unit volume in which the absorption of a quantum of radiation of energy $\hbar\omega$ causes an electron to be raised from band i to band j is given by

$$\frac{2\pi}{\hbar} |A_0|^2 \left(\frac{e}{mc}\right)^2 \frac{2}{(2\pi)^3} \int d^3k_1 |\boldsymbol{\alpha}.\mathbf{p}_{\mathrm{ji}}(\mathbf{k}_1)|^2\, \delta[\epsilon_{\mathrm{j}}(\mathbf{k}_1) - \epsilon_{\mathrm{i}}(\mathbf{k}_1) - \hbar\omega]$$

the integration being over all available electron states consistent with the conservation of energy taking into account the two spin states of the electron. This is of course just the number of quanta of radiation of energy $\hbar\omega$ absorbed per unit time. We have made the usual assumption here that the bands i and j are essentially full and empty respectively so that we are not restricted by the exclusion principle. This will normally be valid at temperatures T such that kT is less than the energy gap of the semiconductor. If we consider the radiation in the form of a beam falling normally on to the face of specimen of semiconductor, the absorption coefficient K of the semiconductor is given by the number of quanta absorbed per unit time per unit volume divided by the flux density cN/n of quanta in the material of refractive index n. Thus the absorption coefficient due to electron transitions from band i to band j is

$$K_{\mathrm{ji}} = \frac{4\pi^2 e^2}{nm^2 c\omega} \frac{2}{(2\pi)^3} \int d^3k\, |\boldsymbol{\alpha}.\mathbf{p}_{\mathrm{ji}}(\mathbf{k})|^2\, \delta[\epsilon_{\mathrm{j}}(\mathbf{k}) - \epsilon_{\mathrm{i}}(\mathbf{k}) - \hbar\omega] \qquad \ldots (7)$$

where we have used the definition of $|A_0|^2$ given in equation (3).

For a semiconductor in its ground state with completely filled valence bands and empty conduction bands, the lowest energy transitions of this type are those in which electrons are raised from the uppermost valence band v, usually referred to as the valence band, to the lowest conduction band c, usually called the conduction band. The energy gap $\epsilon_{\mathrm{c}}(\mathbf{k}) - \epsilon_{\mathrm{v}}(\mathbf{k})$ between the valence and conduction bands will have a minimum value ϵ_0 at some set of wave vectors $\mathbf{k}_0$ which will either be zero or a set of points in $\mathbf{k}$-space having the symmetry of the crystal lattice. ϵ_0 is the direct energy gap of the material and is the threshold energy for direct electronic inter-band transitions in the crystal. Let us assume for simplicity that ϵ_0 occurs at $\mathbf{k}_0 = 0$ and at this point the valence and conduction bands have extrema characterized by isotropic effective masses m_{v} and m_{c} respectively, so that

$$\epsilon_{\mathrm{c}}(\mathbf{k}) = \epsilon_{\mathrm{c}}(0) + \frac{\hbar^2 k^2}{2m_{\mathrm{c}}} \qquad \ldots (8a)$$

$$\epsilon_{\mathrm{v}}(\mathbf{k}) = \epsilon_{\mathrm{v}}(0) - \frac{\hbar^2 k^2}{2m_{\mathrm{v}}} \qquad \ldots (8b)$$

$$\epsilon_0 = \epsilon_{\mathrm{c}}(0) - \epsilon_{\mathrm{v}}(0) \qquad \ldots (8c)$$

for $\mathbf{k}$ close to zero. Then, from equation (7), we find the absorption coefficient for direct transitions in the region close enough to $\hbar\omega = \epsilon_0$ for transitions only between the valence and conduction bands to be contributing

$$K = \frac{2e^2}{nm^2c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}|\boldsymbol{\alpha}.\mathbf{p}_{cv}(\mathbf{k})|^2(\hbar\omega-\epsilon_0)^{1/2} \qquad \ldots(9)$$

where

$$|\mathbf{k}| = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}(\hbar\omega-\epsilon_0)^{1/2} \qquad \ldots(10)$$

and

$$m^* = \frac{m_c m_v}{m_c+m_v} \qquad \ldots(11)$$

For energies so close to the threshold that we can expand $\boldsymbol{\alpha}.\mathbf{p}_{cv}(\mathbf{k})$ as

$$\boldsymbol{\alpha}.\mathbf{p}_{cv}(\mathbf{k}) = \boldsymbol{\alpha}.\mathbf{p}_{cv}(0)+\mathbf{k}\left[\frac{\partial}{\partial\mathbf{k}}\boldsymbol{\alpha}.\mathbf{p}_{cv}(\mathbf{k})\right]_{\mathbf{k}=0}+\ldots \qquad \ldots(12)$$

and retain only the first non-zero term, the complete energy dependence of K can be found. Two separate cases must be considered: (1) $\mathbf{p}_{cv}(0) \neq 0$ when the transitions are said to be allowed; and (2) $\mathbf{p}_{cv}(0) = 0$ when the transitions are said to be forbidden. We find from equation (9) that the absorption coefficient K_a for allowed transitions is

$$K_a = \frac{2e^2}{nm^2c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}|\boldsymbol{\alpha}.\mathbf{p}_{cv}(0)|^2(\hbar\omega-\epsilon_0)^{1/2} \qquad \ldots(13)$$

and for forbidden transitions is

$$K_f = \frac{2e^2}{3nm^2c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{5/2}\left|\frac{\partial}{\partial\mathbf{k}}\boldsymbol{\alpha}.\mathbf{p}_{cv}(\mathbf{k})\right|^2_{\mathbf{k}=0}(\hbar\omega-\epsilon_0)^{3/2} \qquad \ldots(14)$$

Thus the shape of the absorption curve as a function of energy is different according as the transitions are allowed or forbidden, this being determined purely by the symmetry of the electronic wave functions at the points $\mathbf{k}_0 = 0$ in the valence and conduction bands (cf. equation (5)).

It would seem that by carrying out an appropriate experiment one could determine from equations (13) and (14) the direct energy gap ϵ_0 of a semiconductor and distinguish between allowed and forbidden transitions. However, this is not the case, as equations (13) and (14) do not contain the whole story and we must now go back and examine some of the assumptions made in deriving these expressions.

In the derivation of equations (13) and (14) we have assumed that the electron raised to the conduction band was free to move throughout the crystal unaffected by all the other electrons, equivalent to a positively charged hole, remaining in the valence band. However, as we shall see later, the experimental measurements of the absorption in germanium and silicon show features which can be explained only if one takes into account the mutual attraction of the electron and hole produced in the absorption process. This attraction allows bound states of the electron and hole—excitons—to be formed in the crystal with energy less than that of the free pair of particles. In these states the relative motion of the electron and hole is localized and the pair of particles move through the crystal together.

Absorption can take place into these exciton states at energies $\hbar\omega < \epsilon_0$ and, moreover, the continuous absorption beginning at ϵ_0 is modified in shape from that given by equations (13) and (14) by this attraction. Frenkel[18] advanced the concept of an exciton in 1931 and recently much work has been carried out by Gross[19] and his co-workers and by Nikitine[20] and his co-workers on the study of structure attributable to excitons in the optical absorption spectra of several ionic semiconductors. Wannier[21] originally formulated a theory of the exciton suitable for application to semiconducting materials where electrons and holes are characterized by their band properties. His treatment applied only to materials with simple band structures but it was extended by Dresselhaus[22] to cover crystals with degenerate band edges like germanium and silicon. Elliott[23] has given an extensive discussion of the absorption mechanisms in semiconductors taking into account exciton effects. We shall not go through the details of his arguments but shall present them briefly and in such a way that the physical ideas are brought out.

Raising an electron from a state with wave vector $\mathbf{k}_1$ in the valence band to a conduction band state with wave vector $\mathbf{k}_2$ is equivalent to producing an electron–hole pair, the electron having wave vector $\mathbf{k}_2$ and the hole $-\mathbf{k}_1$. In assuming that the electron and hole can move throughout the crystal independently of each other, we take the wave function describing this pair state to be essentially of the form

$$\psi = \exp(-i\mathbf{k}_1.\mathbf{r}_h)\exp(i\mathbf{k}_2.\mathbf{r}_e) \qquad \ldots (15)$$

where $\mathbf{r}_e$ and $\mathbf{r}_h$ are the electron and hole positions respectively. The complete pair wave function consists of ψ multiplied by appropriate functions of the type $u_{c\mathbf{k}_2}(\mathbf{r}_e)$ which describe the effect of the crystal lattice on the electron and hole. However, these functions do not concern our arguments here. Defining

$$\mathbf{r} = \mathbf{r}_e - \mathbf{r}_h \qquad \ldots (16)$$

the relative position of the pair and

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{r}_e + \mathbf{r}_h) \qquad \ldots (17)$$

the mean electron–hole position, we can write equation (15) as

$$\psi = \exp(i\mathbf{K}.\mathbf{R})\exp(i\mathbf{k}.\mathbf{r}) \qquad \ldots (18)$$

where

$$\mathbf{k} = \tfrac{1}{2}(\mathbf{k}_1 + \mathbf{k}_2) \qquad \ldots (19)$$

and

$$\mathbf{K} = (-\mathbf{k}_1) + \mathbf{k}_2 \qquad \ldots (20)$$

is the total wave vector of the pair. The first exponential factor in equation (18) now describes the motion of the electron–hole pair through the lattice and the second describes the relative motion of the particles, its form depending on the assumption that they do not have any mutual interaction. More generally, we can write in place of equation (18)

$$\psi = \exp(i\mathbf{K}.\mathbf{R})\,\phi_{s\mathbf{K}}(\mathbf{r}) \qquad \ldots (21)$$

where $\phi_{s\mathbf{K}}(\mathbf{r})$ is shown by Dresselhaus[22] to be a solution of the equation

$$[\epsilon_c(-i\nabla + \tfrac{1}{2}\mathbf{K} - \mathbf{k}_c) - \epsilon_v(i\nabla + \tfrac{1}{2}\mathbf{K} - \mathbf{k}_v) + V(\mathbf{r})]\phi_{s\mathbf{K}}(\mathbf{r}) = E_s(\mathbf{K})\phi_{s\mathbf{K}}(\mathbf{r}) \qquad \ldots (22)$$

and describes the relative motion of the electron and hole when they interact through a potential $V(\mathbf{r})$. $\mathbf{k}_c$ and $\mathbf{k}_v$ are the conduction and valence band extrema

in the region of which we assume $\mathbf{k}_2$ and $\mathbf{k}_1$ to lie and by $\epsilon_c(-i\nabla+\frac{1}{2}\mathbf{K}-\mathbf{k}_c)$ and $\epsilon_v(i\nabla+\frac{1}{2}\mathbf{K}-\mathbf{k}_v)$ is to be understood the expansion of these functions about $\mathbf{k}_c$ and $\mathbf{k}_v$ up to the quadratic terms. $E_s(\mathbf{K})$ is the total energy of the electron–hole pair and s is some set of quantum numbers describing the pair state.

The expression (6) along with the definition (20) of the total wave vector $\mathbf{K}$ of the pair shows that in direct transitions only pair states with $\mathbf{K}=0$ can be formed. For the case we are considering, $\mathbf{k}_c=\mathbf{k}_v=\mathbf{k}_0$, which we have taken to be zero, and the functions $\epsilon_c(\mathbf{k})$ and $\epsilon_v(\mathbf{k})$ are given in equation (8). Equation (22) then reduces to

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2+V(\mathbf{r})\right]\phi_{s_0}(\mathbf{r})=[E_s(0)-\epsilon_0]\phi_{s_0}(\mathbf{r}) \qquad \ldots(23)$$

where the reduced effective mass m^* is defined in equation (11). For values of $\mathbf{K}$ close to zero, Dresselhaus shows that

$$E_s(\mathbf{K})=E_s(0)+\frac{\hbar^2K^2}{2(m_c+m_v)} \qquad \ldots(24)$$

For $V(\mathbf{r})=0$, equation (23) of course gives $\phi_{s_0}(\mathbf{r})\equiv\phi_{\mathbf{k}_0}(\mathbf{r})=\exp(i\mathbf{k}.\mathbf{r})$, in agreement with equations (15) and (21). However, $V(\mathbf{r})$ will normally have the form of an attractive coulomb type of potential modified due to the crystal environment of the electron and hole. We shall find that, for semiconductors like germanium and silicon, the electron and hole, notwithstanding their mutual attraction, have a separation far in excess of the lattice spacing in these crystals. It is therefore appropriate to take

$$V(\mathbf{r})=-\frac{e^2}{\kappa r} \qquad \ldots(25)$$

the coulomb attraction being modified by the static dielectric constant κ of the material concerned.[24] Equation (23) is then just the familiar hydrogen atom equation with a reduced mass m^* and effective charge κ^{-1}. It has two types of solutions: one corresponding to bound states of the pair—the exciton states—and the other to unbound states. These solutions are well known and can be found in any standard text on quantum mechanics.[25] For the bound exciton states the energy depends on only one of the set s of quantum numbers characterizing the state, viz., n, the principal quantum number which can take any positive integral value. For small values of $\mathbf{K}$, by equation (24),

$$E_n(\mathbf{K})=\epsilon_0+\frac{\hbar^2K^2}{2(m_c+m_v)}-\frac{E_{ex}(0)}{n^2} \qquad \ldots(26)$$

where

$$E_{ex}(0)=\frac{m^*e^4}{2\hbar^2\kappa^2} \qquad \ldots(27)$$

is the binding energy of an exciton of zero wave vector in its ground state—$n=1$. The energies given by equation (26) lie in what is normally considered, in the one-electron approximation, to be a forbidden range of energies. Excitations into these levels become possible, however, when correlations between the motions of electrons in the crystal are taken into account as we have described. For the

unbound state solutions of equation (23), the energy depends on the magnitude of the wave vector $\mathbf{k}$ of the relative motion of the two particles; for small $\mathbf{K}$ it is

$$E_{\mathbf{k}}(\mathbf{K}) = \epsilon_0 + \frac{\hbar^2 K^2}{2(m_c + m_v)} + \frac{\hbar^2 k^2}{2m^*} \qquad \ldots (28)$$

and is in fact the same as the energy of the pair when the interaction is neglected. The inclusion of the coulomb attraction between the electron and hole, therefore, does not change the number of states per unit energy range available to the pair of particles in unbound states. It does of course alter considerably the wave function of the relative motion in these states.

The solutions of equation (22) will, in general, be of two types, one corresponding to unbound pair states and the other to bound exciton states. For the bound exciton states the total energy of the pair will depend on some quantum numbers, like n in equation (26), labelling the internal state of the exciton and the total wave vector of the exciton. We can therefore talk of exciton bands, each band corresponding to some internal state of the exciton and the shape of the band giving the dependence of the exciton energy on its wave vector. It should be pointed out, however, that care should be taken in discussing exciton and electronic energy bands in conjunction with one another as the exciton bands are the dispersion curves of a pair of particles moving through the lattice while the electronic bands are associated with the motion of single particles.

Having discussed the effects of an electron–hole interaction on the possible states of the pair, we must now see how this modifies our arguments in deriving equations (13) and (14). In deriving these expressions equal weight was given to all the possible states of the electron–hole pair formed in the absorption process regardless of the relative motion of the particles in these states. This was correct but its significance was missed due to the simplicity of the result. In actual fact, when the pair is formed, the electron and hole must be at the same position in the crystal and so they will be preferentially formed into states in which they are most likely to be together. Thus a final state, in which the wave vector of relative motion is $\mathbf{k}$, will be weighted by the probability $|\phi_{\mathbf{k}_0}(0)|^2$ that the electron and hole can be found at the same point in space in that state. $\mathbf{K} = 0$ in the function ϕ since electron–hole pairs can be formed only with a zero total wave vector in direct transitions. When the electron–hole interaction is neglected, $\phi_{\mathbf{k}_0}(\mathbf{r})$ describes a plane wave so that all final states have equal weight. However, when the interaction is taken into account, the unbound solutions of equation (23) give

$$|\phi_{\mathbf{k}_0}(0)|^2 = \frac{\pi\alpha \exp \pi\alpha}{\sinh \pi\alpha} \qquad \ldots (29)$$

where

$$\alpha = \sqrt{\left(\frac{2m^* E_{\mathrm{ex}}(0)}{\hbar^2 k^2}\right)} \qquad \ldots (30)$$

The detailed analysis of Elliott[23] leads to the expected result in the case of allowed transitions—we must weight the integrand in equation (7) by a factor $|\phi_{\mathbf{k}_0}(0)|^2$, obtaining for allowed transitions at energies $\hbar\omega$ close to ϵ_0

$$K_2 = \frac{2\pi e^2}{nm^2 c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} |\boldsymbol{\alpha}.\mathbf{p}_{cv}(0)|^2 E_{\mathrm{ex}}^{1/2}(0)\frac{\exp x}{\sinh x} \qquad \ldots (31)$$

where $$z = \pi \sqrt{\left(\frac{E_{ex}(0)}{\hbar\omega - \epsilon_0}\right)} \qquad \ldots (32)$$

When $\hbar\omega - \epsilon_0 \gg \pi^2 E_{ex}(0)$, the energy put into the relative motion of the electron and hole is much greater than the binding which the coulomb attraction can supply, so that the effects of the interaction are negligible and so, as one would expect, equation (31) reduces to equation (13). When $\hbar\omega - \epsilon_0 \to 0$, the absorption coefficient tends to a non-zero value

$$K_a \to \frac{4\pi e^2}{nm^2 c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} |\boldsymbol{\alpha}.\mathbf{p}_{cv}(0)|^2 E_{ex}^{1/2}(0) \qquad \ldots (33)$$

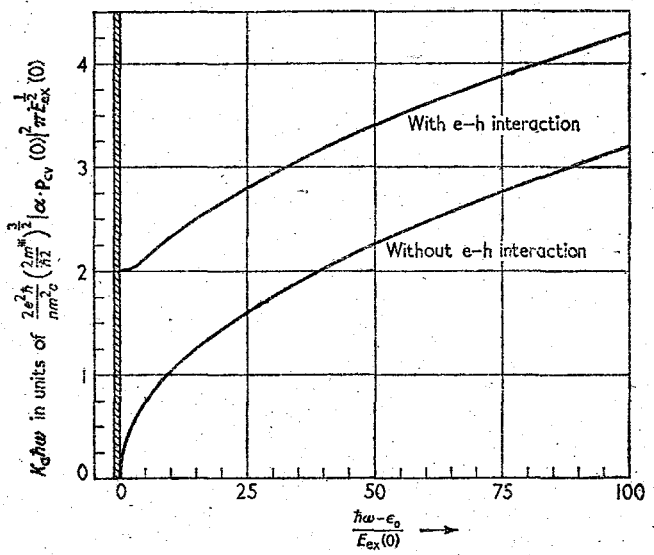


Figure 1. *A comparison of the form of the absorption coefficient due to allowed direct transitions between simple bands given by the theory with and without the inclusion of electron–hole interaction. The cross-hatched portion shows the extent in energy of the region of exciton absorption lines*

which increases as the strength of mutual attraction increases. Figure 1 shows the energy dependence of the two absorption coefficients given by equations (13) and (31) and we note from this Figure that even at energies greater than ϵ_0 by about one hundred times the exciton binding energy, the absorption given by equation (31) is still about 35 per cent greater than that given by equation (13).

Over and above this continuous absorption due to allowed transitions, absorption can also be produced by transitions into bound exciton states. Although each exciton state has a band of energies associated with it, absorption by direct transitions can excite only one discrete state in each band, viz., the state with zero wave vector, as we have already seen. We shall therefore have a series of discrete absorption lines at energies $E_{ex}(0)/n^2$, $n = 1, 2, 3 \ldots$, below the continuous absorption threshold, i.e. below the energy ϵ_0 normally regarded as the threshold for

direct transitions. Absorption into state n will be proportional to $|\phi_{n_0}(0)|^2$ which is non-zero only for s-states of the exciton and then has the value

$$\left(\frac{m^* e^2}{\hbar^2 \kappa}\right)^3 \frac{1}{\pi n^3}$$

The absorption into the exciton states will therefore fall in magnitude as n^{-3} and will eventually merge into the continuous absorption given by equation (31). Hopfield[26] has recently pointed out difficulties in using the normal type of theory to describe absorption into discrete exciton states and has treated the problem in a more rigorous fashion. However, this does not alter the results but merely the

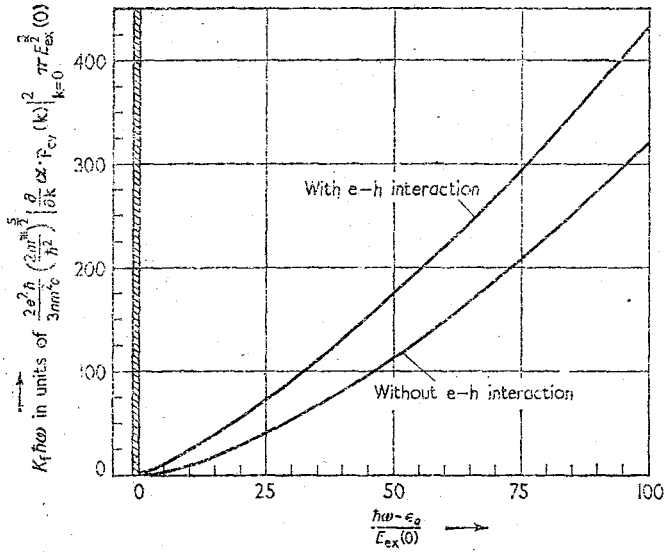


Figure 2. *A comparison of the form of the absorption coefficient due to forbidden direct transitions between simple bands given by the theory with and without the inclusion of electron–hole interaction. The cross-hatched portion shows the extent in energy of the region of exciton absorption lines*

interpretation of the absorption process so that one regards the energy absorbed in the exciton transitions as residing not in the exciton states but in energy modes of the crystal to which the excitons, formed by absorption, are coupled.

For absorption due to forbidden transitions, Elliott's analysis shows that the absorption into a particular final state characterized by $\phi_{s_0}(\mathbf{r})$ is proportional to $|\partial\phi_{s_0}/\partial r|^2_{\mathbf{r}=0}$. For the unbound states, this leads to an absorption coefficient

$$K_f = \frac{2\pi e^2}{3nm^2c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{5/2}\left|\frac{\partial}{\partial \mathbf{k}}\boldsymbol{\alpha}\cdot\mathbf{p}_{cv}(\mathbf{k})\right|^2_{\mathbf{k}=0} E_{ex}^{3/2}(0)\left(1+\frac{\pi^2}{z^2}\right)\frac{\exp z}{\sinh z} \quad \ldots (34)$$

which tends to the expression (14) for $\hbar\omega-\epsilon_0 \gg \pi^2 E_{ex}(0)$. For $\hbar\omega \to \epsilon_0$,

$$K_f \to \frac{4\pi e^2}{3nm^2c\omega}\left(\frac{2m^*}{\hbar^2}\right)^{5/2}\left|\frac{\partial}{\partial \mathbf{k}}\boldsymbol{\alpha}\cdot\mathbf{p}_{cv}(\mathbf{k})\right|^2_{\mathbf{k}=0} E_{ex}^{3/2}(0) \quad \ldots (35)$$

The relative energy dependence of equations (14) and (34) is shown in Figure 2. For the bound exciton states $|\partial\phi_{n_0}/\partial r|^2_{r=0}$ is non-zero only for p-states and then has the value

$$\left(\frac{m^* e^2}{\hbar^2 \kappa}\right)^5 \frac{n^2-1}{3\pi n^5}$$

so that a series of lines corresponding to $n = 2, 3, 4$ is observed at energies $E_{ex}(0)/n^2$ below ϵ_0, this series merging at $\hbar\omega = \epsilon_0$ with the continuous absorption given by equation (34).

2.2. Indirect Transitions

In many semiconductors and in particular in germanium and silicon, the direct energy gap ϵ_0 is not the minimum energy gap ϵ_g between the valence and conduction bands since the edges of these bands do not occur at the same points in $\mathbf{k}$-space. ϵ_0 is the threshold energy for direct optical transitions and one would not normally expect absorbing transitions to take place at energies less than this value due to the fact that, in transitions involving only the interaction between the electrons and radiation, we have seen that the electron wave vector does not change and conduction and valence band states separated by less than ϵ_0 must correspond to different wave vectors. However, as well as interacting with the incident radiation, the electrons in the semiconductor will interact with the various types of defects present in the crystal, e.g. lattice vibrations, impurity atoms, dislocations, etc. These defects can be regarded as sources of momentum or wave vector for the electron system providing, however, little or no energy for them. The simultaneous interaction of the electrons with the radiation and crystal defects provides a mechanism by which an electron can make a transition in which its wave vector changes, most of the energy required by the transition coming from the radiation and all the momentum or wave vector from the defect. Such a transition is called indirect and ϵ_g can be regarded as the indirect energy gap. It is close to but is not identical with the threshold energy for indirect transitions, as we shall see presently.

Cheeseman[27] was the first to consider the effects on the absorption edge of the simultaneous interaction of the electrons in an insulator with radiation and lattice vibrations and, in fact, he pointed out that this could produce absorption at energies lower than the direct energy gap for an appropriate band structure. Hall, Bardeen, and Blatt[15] used these same ideas and have given a discussion of the effects of indirect transitions on the absorption edge of a semiconductor with a band structure like germanium. In presenting the theory of these transitions we shall basically follow their approach and shall begin by neglecting exciton effects. We shall show how these effects alter the absorption at a later stage.

We shall denote by $|\mathbf{k}_s, s; \mathbf{k}_t, t)$ an electronic state in which an electron with wave vector $\mathbf{k}_s$ is present in a normally empty conduction band s and an electron of wave vector $\mathbf{k}_t$ is missing from a normally full valence band t, i.e. a hole of wave vector $-\mathbf{k}_t$ is present in band t. The ground state of the electrons, with empty conduction and full valence bands, we shall denote by $|0)$, and the Hamiltonian describing the interaction of the electrons with a crystal defect by H_d. We shall assume with Hall, Bardeen, and Blatt that H_d can be treated as a small perturbation of the electron system so that we shall be interested in the lowest

order perturbation theory matrix element which describes transitions, due to the interactions H_{rad} and H_{d}, from the ground state $|0)$ to some state $|\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{t}}, t)$ where $\mathbf{k}_{\text{t}}$ is not necessarily equal to $\mathbf{k}_{\text{s}}$. This occurs in second order and using equation (6) can be reduced to

$$(\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{t}}, t|M|0) = A_0 \frac{e}{mc} \sum_{l=0}^{1} \left[\sum_{m} \frac{(\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{t}}, t|H_{\text{dl}}|\mathbf{k}_{\text{t}}, m; \mathbf{k}_{\text{t}}, t)}{\epsilon_{\text{s}}(\mathbf{k}_{\text{s}}) - \epsilon_{\text{m}}(\mathbf{k}_{\text{t}}) + (-1)^l \epsilon_{\text{d}}} \boldsymbol{\alpha} \cdot \mathbf{p}_{\text{mt}}(\mathbf{k}_{\text{t}}) - \right.$$

$$- \sum_{n} \frac{(\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{t}}, n|H_{\text{dl}}|0)}{\epsilon_{\text{s}}(\mathbf{k}_{\text{s}}) - \epsilon_{\text{n}}(\mathbf{k}_{\text{t}}) + (-1)^l \epsilon_{\text{d}}} \boldsymbol{\alpha} \cdot \mathbf{p}_{\text{nt}}(\mathbf{k}_{\text{t}}) +$$

$$+ \sum_{p} \frac{(\mathbf{k}_{\text{s}}, p; \mathbf{k}_{\text{t}}, t|H_{\text{dl}}|0)}{\epsilon_{\text{t}}(\mathbf{k}_{\text{t}}) - \epsilon_{\text{p}}(\mathbf{k}_{\text{s}}) - (-1)^l \epsilon_{\text{d}}} \boldsymbol{\alpha} \cdot \mathbf{p}_{\text{sp}}(\mathbf{k}_{\text{s}}) -$$

$$\left. - \sum_{q} \frac{(\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{t}}, t|H_{\text{dl}}|\mathbf{k}_{\text{s}}, s; \mathbf{k}_{\text{s}}, q)}{\epsilon_{\text{t}}(\mathbf{k}_{\text{t}}) - \epsilon_{\text{q}}(\mathbf{k}_{\text{s}}) - (-1)^l \epsilon_{\text{d}}} \boldsymbol{\alpha} \cdot \mathbf{p}_{\text{sq}}(\mathbf{k}_{\text{s}}) \right] \qquad \ldots (36)$$

The summations over m and p are over normally empty conduction bands and those over n and q are over normally full valence bands. ϵ_{d} is the energy exchanged between the electron system and the lattice defect, $l = 0$ or 1 corresponding respectively to interactions in which energy is lost or gained by the electrons. The only indirect transitions whose effects on the absorption edge have been observed are those associated with the lattice vibrations of the crystal and so we shall discuss equation (36) with this particular type of defect in mind.

Any particular mode of vibration of a crystal lattice can be considered as being composed of several basic types of lattice vibration. Each basic type of vibration is characterized by a wave vector $\mathbf{K}$ lying in the first Brillouin zone of the reciprocal lattice and a label b denoting to which particular branch of the vibration spectrum it belongs. The energy of a vibration of wave vector $\mathbf{K}$ in branch b is $k\theta_{\text{b}}(\mathbf{K})$, where k is Boltzmann's constant and we choose to measure energy in terms of a temperature $\theta_{\text{b}}(\mathbf{K})$. The dispersion curves of $\theta_{\text{b}}(\mathbf{K})$ as a function of $\mathbf{K}$ for the various branches b constitute the lattice vibrational spectrum of the crystal. The number of branches is equal to three times the number of atoms in a unit cell of the crystal so that, for semiconductors of the diamond-type structure like germanium and silicon, there are six branches in the spectrum. There are always three branches for which $\theta_{\text{b}}(\mathbf{K}) \to 0$ as $\mathbf{K} \to 0$ and they are called the acoustical branches or modes of vibration. The remaining branches have finite energy at zero wave vector and are called optical branches. Moreover, in certain symmetry directions of the crystal, the three acoustical branches consist of two in which the vibration is transverse to the direction of propagation $\mathbf{K}$ and one in which it is parallel to $\mathbf{K}$; these are known as the transverse acoustical (TA) and longitudinal acoustical (LA) branches. The three remaining optical branches in germanium and silicon have the same property, the branches being known as the transverse optical (TO) and longitudinal optical (LO) branches. When the lattice vibrational field is quantized, then the various basic modes of lattice vibration correspond to quanta of lattice vibrational energy and are called phonons. The excitation and de-excitation of lattice vibrations can then be thought of in terms of creation and

destruction of phonons. When the crystal is in thermal equilibrium at temperature T, the number of phonons excited with wave vector $\mathbf{K}$ and belonging to branch b of the vibrational spectrum is

$$n_b(\mathbf{K}) = \left[\exp\frac{\theta_b(\mathbf{K})}{T} - 1\right]^{-1} \qquad \ldots (37)$$

The Hamiltonian H_{lat} describing the interaction between an electron and the lattice vibrations can be written[28] as

$$H_{lat} = \sum_b \int d^3K \xi_b(\mathbf{K}).\mathbf{O}_b(\mathbf{K})$$

where

$$\int d^3r\, \psi^*_{j\mathbf{k}}(\mathbf{r})\,\mathbf{O}_b(\mathbf{K})\psi_{i\mathbf{k}'}(\mathbf{r}) = \sqrt{\left[\frac{\hbar^2}{2DVk\theta_b(\mathbf{K})}\right]}[j,\mathbf{k}|\mathbf{C}_b(\mathbf{K})|i,\mathbf{k}']\sqrt{[n_b(\mathbf{K})+\tfrac{1}{2}\pm\tfrac{1}{2}]}\,\delta(\mathbf{k}-\mathbf{K}-\mathbf{k}') \ldots (38)$$

D and V being the density and volume of the crystal. $\xi_b(\mathbf{K})$ is a unit vector in the direction of the lattice displacement associated with a phonon from branch b propagating with wave vector $\mathbf{K}$. With the matrix element of the operator $\mathbf{O}_b(\mathbf{K})$ written in the form of equation (38), the operator $\mathbf{C}_b(\mathbf{K})$ has the property that its matrix element between two closely neighbouring states of the same band with wave vectors $\mathbf{k}$ and $\mathbf{k}'$ is proportional to $(\mathbf{k}-\mathbf{k}')$, the constant of proportionality being the appropriate deformation potential for the band[29] at $\mathbf{k}$. The upper sign in equation (38) corresponds to the case of a phonon of wave vector $-\mathbf{K}$ being created ($l=0$), the lower to a phonon of wave vector $\mathbf{K}$ being destroyed ($l=1$). Using equation (38), equation (36) becomes in the case of lattice scattering

$$(\mathbf{k}_s,s;\mathbf{k}_t,t|M|0) = A_0\frac{e}{mc}\sum_{l=0}^{1}\sum_b \sqrt{\left[\frac{\hbar^2}{2DVk\theta_b(\mathbf{k}_s-\mathbf{k}_t)}\right]} X^{lb}_{st}(\mathbf{k}_s,\mathbf{k}_t)\times \times\sqrt{\left[n_b(\mathbf{k}_s-\mathbf{k}_t)+\frac{1+(-1)^l}{2}\right]} \qquad \ldots (39)$$

where

$$X^{lb}_{st}(\mathbf{k}_s,\mathbf{k}_t) = \sum_m \frac{\xi_b(\mathbf{k}_s-\mathbf{k}_t).[s,\mathbf{k}_s|\mathbf{C}_b(\mathbf{k}_s-\mathbf{k}_t)|m,\mathbf{k}_t]\,\boldsymbol{\alpha}.\mathbf{p}_{mt}(\mathbf{k}_t)}{\epsilon_s(\mathbf{k}_s)-\epsilon_m(\mathbf{k}_t)+(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)} - \sum_n \frac{\xi_b(\mathbf{k}_s-\mathbf{k}_t).[s,\mathbf{k}_s|\mathbf{C}_b(\mathbf{k}_s-\mathbf{k}_t)|n,\mathbf{k}_t]\,\boldsymbol{\alpha}.\mathbf{p}_{nt}(\mathbf{k}_t)}{\epsilon_s(\mathbf{k}_s)-\epsilon_n(\mathbf{k}_t)+(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)} + \sum_p \frac{\boldsymbol{\alpha}.\mathbf{p}_{sp}(\mathbf{k}_s)\xi_b(\mathbf{k}_s-\mathbf{k}_t)[p,\mathbf{k}_s|\mathbf{C}_b(\mathbf{k}_s-\mathbf{k}_t)|t,\mathbf{k}_t]}{\epsilon_t(\mathbf{k}_t)-\epsilon_p(\mathbf{k}_s)-(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)} - \sum_q \frac{\boldsymbol{\alpha}.\mathbf{p}_{sq}(\mathbf{k}_s)\xi_b(\mathbf{k}_s-\mathbf{k}_t).[q,\mathbf{k}_s|\mathbf{C}_b(\mathbf{k}_s-\mathbf{k}_t)|t,\mathbf{k}_t]}{\epsilon_t(\mathbf{k}_t)-\epsilon_q(\mathbf{k}_s)-(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)} \qquad \ldots (40)$$

It follows from equation (39) that the number of quanta of radiation of energy $\hbar\omega$ absorbed per unit time per unit volume as a result of indirect transitions from band t to band s is

$$\frac{2\pi}{\hbar}|A_0|^2\left(\frac{e}{mc}\right)^2\frac{\hbar^2}{2D}\frac{2}{(2\pi)^6}\sum_{l=0}^{1}\sum_{b}\int d^3k_s\,d^3k_t\,\frac{|X_{st}^{lb}(\mathbf{k}_s,\mathbf{k}_t)|^2}{k\theta_b(\mathbf{k}_s-\mathbf{k}_t)}\times$$
$$\times\left(n_b(\mathbf{k}_s-\mathbf{k}_t)+\frac{1+(-1)^l}{2}\right)\delta[\epsilon_s(\mathbf{k}_s)-\epsilon_t(\mathbf{k}_t)+(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)-\hbar\omega] \quad \ldots(41)$$

taking into account the two spin states of the electron. Dividing this by the flux density of quanta yields the absorption coefficient K_{st} due to these transitions

$$K_{st}=\sum_{l=0}^{1}\sum_{b}K_{st}^{lb} \quad \ldots(42)$$

where

$$K_{st}^{lb}=\frac{4\pi^2e^2}{nm^2c\omega}\frac{\hbar^2}{D(2\pi)^6}\int d^3k_s\,d^3k_t\,\frac{|X_{st}^{lb}(\mathbf{k}_s,\mathbf{k}_t)|^2}{k\theta_b(\mathbf{k}_s-\mathbf{k}_t)}\left(n_b(\mathbf{k}_s-\mathbf{k}_t)+\frac{1+(-1)^l}{2}\right)\times$$
$$\times\delta[\epsilon_s(\mathbf{k}_s)-\epsilon_t(\mathbf{k}_t)+(-1)^l k\theta_b(\mathbf{k}_s-\mathbf{k}_t)-\hbar\omega] \quad \ldots(43)$$

and is the absorption coefficient due to indirect transitions from band t to band s in which a phonon belonging to branch b of the vibrational spectrum is either emitted into the lattice ($l=0$) or absorbed from the lattice ($l=1$).

For a semiconductor, whose valence band edge is at $\mathbf{k}_v$ and conduction band edge at $\mathbf{k}_c$, the lowest energy transitions of this type which can take place are those in which an electron is raised from a valence band state close to $\mathbf{k}_v$ to a conduction band state close to $\mathbf{k}_c$. We shall be interested in transitions of this type particularly in germanium and silicon and so we shall carry through our calculations using a model which simulates the band structures of germanium and silicon fairly well, viz., we shall assume that $\mathbf{k}_v=0$ and that we have a single isotropic band at this point whose energy is given by equation (8*b*); we shall assume that $\mathbf{k}_c$ lies at a set N_c of equivalent points on symmetry directions in $\mathbf{k}$-space and that close to one of these points

$$\epsilon_c(\mathbf{k})=\epsilon_c(\mathbf{k}_c)+\frac{\hbar^2}{2m_t}(k_1-k_{c_1})^2+\frac{\hbar^2}{2m_t}(k_2-k_{c_2})^2+\frac{\hbar^2}{2m_l}(k_3-k_{c_3})^2 \quad \ldots(44)$$

where the axes (1, 2, 3) have the 3-axis along the symmetry direction and m_l and m_t are the longitudinal and transverse masses characterizing the spheroidal energy surface. The principal drawback of this model is of course that it neglects the degenerate character and not completely isotropic shape of the valence band edge. However, the model has the advantage that it allows us to carry our analysis further and gives all the main features of the absorption spectrum correctly.

For absorption close to the threshold, equation (43) therefore becomes

$$K_{cv}^{lb}=\frac{4\pi^2e^2N_c}{nm^2c\omega}\frac{\hbar^2}{D(2\pi)^6}\int d^3k_1\,d^3k_2\,\frac{|X_{cv}^{lb}(\mathbf{k}_1,\mathbf{k}_2)|^2}{k\theta_b(\mathbf{k}_1-\mathbf{k}_2)}\left(n_b(\mathbf{k}_1-\mathbf{k}_2)+\frac{1+(-1)^l}{2}\right)\times$$
$$\times\delta[\epsilon_c(\mathbf{k}_1)-\epsilon_v(\mathbf{k}_2)+(-1)^l k\theta_b(\mathbf{k}_1-\mathbf{k}_2)-\hbar\omega] \quad \ldots(45)$$

where $\hbar\omega \simeq \epsilon_g$. We can expand $X_{cv}^{lb}(\mathbf{k}_1, \mathbf{k}_2)$ in equation (45) about $\mathbf{k}_1 = \mathbf{k}_c$ and $\mathbf{k}_2 = 0$ and retain only the first non-zero term. As in the case of direct transitions, the transitions are called allowed or forbidden according as the first non-zero term is $X_{cv}^{lb}(\mathbf{k}_c, 0)$ or not. The integrations in equation (45) are then easily carried out with the help of equations (8*b*) and (44) and noting that the variation of θ_b with wave vector is much slower than the variation of ϵ_c or ϵ_v. This latter fact is easily seen if we consider a simple calculation. In germanium, $\mathbf{k}_c$ lies in a $\langle 111 \rangle$ direction and, in this direction, the most rapidly varying branch of the vibrational spectrum is the LA branch which rises from zero at $\mathbf{k} = 0$ to 320° K at the zone edge. If we assume that this variation is quadratic in $\mathbf{k}$ then it is characterized by an 'effective mass' which is of the order of a hundred times larger than the effective masses of electrons at the valence or conduction band edges. We can therefore neglect the variation of $\theta_b(\mathbf{k}_1 - \mathbf{k}_2)$ in the δ-function in equation (45) and replace it by $\theta_b(\mathbf{k}_c)$, a similar calculation to that described showing that this is valid for silicon as well as germanium. The integral in equation (45), therefore, reduces in the case of allowed transitions to a sum over all possible final states of the electron–hole pair formed in the transition, consistent with the conservation of energy. This integration is easily carried out and leads to the following expression for the absorption coefficient close to threshold

$$K_s = \frac{N_c}{32\pi D}\frac{e^2\hbar^2}{nm^2c\omega}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2}\left(\frac{2m_v}{\hbar^2}\right)^{3/2}\sum_{l=0}^{1}{\sum_b}'\frac{|X_{cv}^{lb}(\mathbf{k}_c, 0)|^2}{k\theta_b(\mathbf{k}_c)}\times$$
$$\times\left(n_b(\mathbf{k}_c)+\frac{1+(-1)^l}{2}\right)[\hbar\omega-\epsilon_g-(-1)^l k\theta_b(\mathbf{k}_c)]^2 \quad \ldots(46)$$

where
$$m_c^* = (m_l m_t^2)^{1/3} \quad \ldots(47)$$

and the prime on the summation over b restricts it to those branches for which the transitions are allowed. The main features of this expression for the absorption coefficient are best illustrated if we use equation (37) and write it in the form

$$K_s = \frac{A}{\hbar\omega}{\sum_b}'\frac{1}{k\theta_b(\mathbf{k}_c)}\left\{\frac{|X_{cv}^{ob}(\mathbf{k}_c, 0)|^2}{1-\exp[-\theta_b(\mathbf{k}_c)/T]}[\hbar\omega-\epsilon_g-k\theta_b(\mathbf{k}_c)]^2+\right.$$
$$\left.+\frac{|X_{cv}^{lb}(\mathbf{k}_c, 0)|^2}{\exp[\theta_b(\mathbf{k}_c)/T]-1}[\hbar\omega-\epsilon_g+k\theta_b(\mathbf{k}_c)]^2\right\} \quad \ldots(48)$$

where
$$A = \frac{N_c}{4\pi D}\frac{e^2(m_c^* m_v)^{3/2}}{nm^2c\hbar^3}$$

For forbidden transitions, the expression for the absorption coefficient has the same structure as equation (48). The Xs are replaced by their first derivatives, the sum is over branches for which the transitions are forbidden, and each term depends on the cube and not the square of the energy. The general level of the absorption is of course lower than in the case of allowed transitions.

It follows, therefore, that for incident radiation energies around ϵ_g, the absorption spectrum consists in general of several pairs of components, there being as many pairs as branches in the vibrational spectrum. The two components in each pair have the same energy dependence, each component rising from some threshold

as the square or the cube of the energy above this threshold according as the transitions associated with the pair are allowed or forbidden. The two thresholds associated with any given pair are equally spaced above and below ϵ_g by the energy $k\theta_b(\mathbf{k}_c)$ required by the phonon emitted into the lattice or given up by the phonon absorbed from the lattice. Moreover, the absorption spectrum is quite strongly temperature dependent in contrast to the temperature independent absorption coefficient corresponding to direct absorbing transitions which we obtained in our analysis in the previous section. (This remark neglects any movement of the entire absorption curve with temperature which is not due to the absorption mechanism but to the temperature dependence of the energy gaps ϵ_0 and ϵ_g.) As the temperature is decreased, the component of each pair associated with the transitions involving the absorption of a phonon gradually disappears as the number of phonons present in the lattice decreases, and at very low temperatures only components associated with the emission of phonons are present. At any temperature T, the relative magnitude of the two components of any pair at the same energy above their thresholds is given by

$$\frac{\text{component with phonon emission}}{\text{component with phonon absorption}} = \rho_b \exp\left[\frac{\theta_b(\mathbf{k}_c)}{T}\right] \qquad \ldots (49)$$

where

$$\rho_b = \frac{|Z_{cv}^{0b}(\mathbf{k}_c, 0)|^2}{|Z_{cv}^{1b}(\mathbf{k}_c, 0)|^2}$$

Z_{cv}^{0b} being equal to X_{cv}^{0b} for allowed transitions and to its first derivative for forbidden transitions, and similarly for Z_{cv}^{1b}. We see from equation (40) that the only difference in Z_{cv}^{0b} and Z_{cv}^{1b} is in the phonon energy contribution to the various energy denominators. In most cases, the phonon energy forms a negligible part of these denominators so that Z_{cv}^{0b} and Z_{cv}^{1b} are effectively equal and their ratio in equation (49) is unity. However, an exceptional case arises in germanium which we shall discuss in detail later.

In both germanium and silicon, $\mathbf{k}_c$ lies on a symmetry direction of the crystal so that we can talk of transverse and longitudinal modes of vibration. Moreover, the transverse acoustical modes are degenerate as are the transverse optical modes, so that one should expect to see only four pairs of components at most in the absorption spectrum of these materials.

These are the features to be expected of the absorption due to indirect transitions when exciton effects are neglected. We must now consider how these features are modified by the inclusion of exciton effects. We shall find that the number, temperature dependence, and relative magnitudes of the components are not altered but that their detailed shapes and the interpretation of their threshold energies is considerably modified.

Elliott's work shows that for allowed indirect transitions, as in the case of allowed direct transitions, each possible electron–hole pair state must be weighted by the probability of finding the two particles together in this state. Thus equation (45) reduces, for energies $\hbar\omega \simeq \epsilon_g$, to a sum over all possible pair states consistent with the conservation of energy, the states being weighted by $|\phi_{s,\mathbf{k}_c-\mathbf{k}_v}(0)|^2$ where $\phi_{s,\mathbf{k}_c-\mathbf{k}_v}(x)$ is a solution of equation (22) with $k_y = 0$ and the functions $\epsilon_v(\mathbf{k})$ and $\epsilon_c(\mathbf{k})$ are given by equations (8b) and (44). This equation has the form of equation (23) but is complicated by the anisotropic conduction band mass. Let us assume

for the present, with Elliott, that the conduction band is isotropic around its edge and take the effective mass of electrons near the edge to be m_c. $\phi_{s,\mathbf{k}_1-\mathbf{k}_2}(\mathbf{r})$ then satisfies an equation identical in form to equation (23), whose solutions we have already discussed in the previous section, but with ϵ_0 replaced by ϵ_g and $E_s(0)$ by $E_s(\mathbf{k}_c)$. For absorption producing unbound electron–hole pairs due to allowed transitions, the energy dependence of a component of the absorption coefficient is therefore given by

$$\frac{1}{\hbar\omega}\int d^3k d^3K \frac{\pi\alpha \exp\pi\alpha}{\sinh\pi\alpha}\delta\left[\epsilon_g+\frac{\hbar^2K^2}{2M}+\frac{\hbar^2k^2}{2m^*}+(-1)^l k\theta_b(\mathbf{k}_c)-\hbar\omega\right] \quad \ldots (50)$$

for $\hbar\omega \simeq \epsilon_g$. In this expression, α is given by equation (30) with $E_{ex}(0)$ replaced by $E_{ex}(\mathbf{k}_c)$ and we have used

$$M = m_c+m_v \quad \ldots (51)$$

as the total mass of the pair. Equation (50) reduces to an expression which apart from some constants is

$$\frac{E_{ex}^2(\mathbf{k}_c)}{\hbar\omega}\int_0^\Delta \frac{\sqrt{(\Delta-z)}}{1-\exp\left(-\frac{2}{\sqrt{z}}\right)}dz \quad \ldots (52)$$

where

$$\Delta = \frac{1}{\pi^2 E_{ex}(\mathbf{k}_c)}[\hbar\omega-\epsilon_g-(-1)^l k\theta_b(\mathbf{k}_c)] \quad \ldots (53)$$

For $\hbar\omega \simeq \epsilon_g$, equation (52) yields

$$\frac{2E_{ex}^{1/2}(\mathbf{k}_c)}{3\pi^3\hbar\omega}[\hbar\omega-\epsilon_g-(-1)^l k\theta_b(\mathbf{k}_c)]^{3/2} \quad \ldots (54)$$

so that the component rises from its threshold as the three-halves power of the energy above this threshold. Although equations (50) and (46) are not valid physically when $\hbar\omega \gg \epsilon_g$, we know that analytically they should become identical if we assume the same band parameters, etc., since in this region the mutual interaction of the electron and hole becomes negligible. In this region equation (52) certainly has the correct energy dependence; it reduces to

$$\frac{1}{16\pi^3\hbar\omega}[\hbar\omega-\epsilon_g-(-1)^l k\theta_b(\mathbf{k}_c)]^2 \quad \ldots (55)$$

We can therefore take

$$K_a = \frac{\pi^2 N_c}{2D}\frac{e^2\hbar^2}{nm^2c\omega}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2}\left(\frac{2m_v}{\hbar^2}\right)^{3/2}E_{ex}^2(\mathbf{k}_c)\sum_{l=0}^{1}\sum_{b}{}'\frac{|X_{cv}^{lb}(\mathbf{k}_c,0)|^2}{k\theta_b(\mathbf{k}_c)}\times$$

$$\times\left[n_b(\mathbf{k}_c)+\frac{1+(-1)^l}{2}\right]\int_0^\Delta \frac{\sqrt{(\Delta-z)}}{1-\exp\left(-\frac{2}{\sqrt{z}}\right)}dz \quad \ldots (56)$$

which becomes identical to equation (46) in the high energy limit. For energies $\hbar\omega \simeq \epsilon_g$, equation (56) gives

$$K_a = \frac{N_c}{3\pi D}\frac{e^2\hbar^2}{nm^2c\omega}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2}\left(\frac{2m_v}{\hbar^2}\right)^{3/2} E_{ex}^{1/2}(\mathbf{k}_c)\sum_{l=0}^{1}\sum_{b}{}' \frac{|X_{cv}^{lb}(\mathbf{k}_c,0)|^2}{k\theta_b(\mathbf{k}_c)}\times$$
$$\times\left[n_b(\mathbf{k}_c)+\frac{1+(-1)^l}{2}\right][\hbar\omega-\epsilon_g-(-1)^l k\theta_b(\mathbf{k}_c)]^{3/2} \qquad \ldots(57)$$

which is the absorption coefficient near threshold for allowed indirect transitions when exciton effects are taken into account. By forcing equation (56) to be identical to equation (46) in the high energy limit, we have ensured that the anisotropy of the conduction band edge has been taken correctly into account in summing over the final states. Equation (57) is, of course, still based on $|\phi_{s,\mathbf{k}_1-\mathbf{k}_2}(0)|^2$ appropriate to isotropic bands. This, however, should not produce any significant error in equation (57).

Absorption can also take place into bound exciton states and in this case the absorption coefficient has the form

$$\frac{1}{\hbar\omega}\sum_{n=1}^{\infty}|\phi_{n\mathbf{k}_c}(0)|^2\int d^3K\,\delta\left[\epsilon_g+\frac{\hbar^2K^2}{2M}-\frac{E_{ex}(\mathbf{k}_c)}{n^2}+(-1)^l k\theta_b(\mathbf{k}_c)-\hbar\omega\right] \qquad \ldots(58)$$

which leads to an expression of the form

$$\frac{1}{\hbar\omega}\sum_{n=1}^{\infty}\frac{1}{n^3}\left[\hbar\omega-\epsilon_g+\frac{E_{ex}(\mathbf{k}_c)}{n^2}-(-1)^l k\theta_b(\mathbf{k}_c)\right]^{1/2} \qquad \ldots(59)$$

The absorption into exciton states due to indirect transitions thus gives a continuous absorption in contrast to the absorption lines produced by direct transitions. This is due essentially to the fact that, by absorbing or emitting phonons of the appropriate wave vector, transitions can be made to any point of an exciton band so that the summation over all possible phonon wave vectors implicit in equation (58) leads to a continuous absorption.

For forbidden transitions, the energy dependences given by equations (57) and (59) are modified. For transitions to unbound electron–hole states the absorption will rise as the $\frac{5}{2}$ power of the energy and, for transitions to exciton states, it will rise as the $\frac{3}{2}$ power.

We can summarize this discussion of exciton effects on the indirect absorption in the following way. For incident radiation energies $\hbar\omega \sim \epsilon_g$, the absorption coefficient is given by

$$K = \frac{B}{\hbar\omega}\sum_{b}\left\{\frac{|Z_{cv}^{0b}(\mathbf{k}_c,0)|^2}{1-\exp[-\theta_b(\mathbf{k}_c)/T]}F_b[\hbar\omega-\epsilon_g+E_{ex}(\mathbf{k}_c)-k\theta_b(\mathbf{k}_c)]+\right.$$
$$\left.+\frac{|Z_{cv}^{1b}(\mathbf{k}_c,0)|^2}{\exp[\theta_b(\mathbf{k}_c)/T]-1}F_b[\hbar\omega-\epsilon_g+E_{ex}(\mathbf{k}_c)+k\theta_b(\mathbf{k}_c)]\right\} \qquad \ldots(60)$$

where B is a constant. The functional dependence of the functions $F_b(E)$ on the energy E depends on whether the transitions associated with the phonons from

branch b are allowed or forbidden. In each case $F_b(E)$ consists of a series of contributions each one increasing with energy. These contributions begin at a series of energies, the lowest one corresponding to the onset of transitions into the lowest available exciton state and the highest to the onset of transitions producing unbound electron–hole pairs in the lattice, this being identical to the threshold of the component given by the simple formula (48). For allowed transitions, each contribution except the highest rises as the square root of the energy above its threshold and the highest rises as the $\frac{3}{2}$ power. For forbidden transitions, the corresponding power laws are $\frac{3}{2}$ and $\frac{5}{2}$. Also the contributions due to transitions into the various exciton bands are weighted relative to one another by $1/n^3$ for allowed transitions and $(n^2-1)/n^5$ for forbidden transitions, $n=1$ corresponding to the exciton ground state, so that the main exciton contribution comes from the lowest available state, $n=1$ in the allowed case and $n=2$ in the forbidden. Thus, for an allowed transition, $F_b(E)$ will rise from $E=0$ as $E^{1/2}$, and at an energy several times the exciton binding energy will be rising almost as the $\frac{3}{2}$ power of the energy. For a forbidden transition the initial rise will be as $E^{3/2}$ developing into a $\frac{5}{2}$ power law.

So far we have been discussing indirect transitions between band extrema lying at different points of $\mathbf{k}$-space so that the phonons must supply a significant amount of momentum to the electrons making the transition. However, phonons of all possible wave vectors are available so that in fact indirect transitions can be made between any two band states regardless of their separation in $\mathbf{k}$-space. Dumke[30] has pointed out that, for a substance with conduction and valence band extrema at the same point in $\mathbf{k}$-space giving rise to absorption by direct transitions, as we have discussed earlier, indirect transitions can also take place between these extrema. In these transitions, the electrons must emit or absorb phonons of essentially zero wave vector and the absorption will consist of components equally spaced in pairs above and below the *direct* transition threshold by the energies of phonons of zero wave vector. Since absorption by indirect transitions is much weaker than that by direct transitions, indirect absorption above the direct threshold will not be resolved, only those components of the indirect absorption which begin at energies below this threshold being seen. These components are due to transitions in which optical phonons with zero wave vector are absorbed, these being the only phonons with finite energy at zero wave vector. This absorption will therefore appear on the direct absorption edge as a tail which vanishes at low temperatures.

3. EXPERIMENTAL METHODS

The measurement of the absorption coefficient of a semiconductor in the region of the absorption edge can normally be carried out by measuring the transmission of samples of known thickness for radiation of a series of different wavelengths. The apparatus consists essentially of some source of radiation and a spectrometer to produce a nearly monochromatic beam, some device for accommodating a specimen whose absorption is to be measured, and a detector for measuring the intensity of the radiation beam. As an example of such an apparatus, a diagram of that used by the R.R.E. group in making the measurements on germanium and silicon[1–4] is shown in Figure 3. If I_T and I_0 are the intensities of radiation falling

on the detector with the specimen respectively in and out of the beam, then the absorption coefficient is given by the formula

$$t \equiv \frac{I_T}{I_0} = \frac{(1-R)^2 \exp(-Kd)}{1-R^2\exp(-2Kd)} \qquad \ldots (61)$$

where d is the thickness of the specimen, R its reflectivity, and t the transmission ratio. This expression takes into account multiple reflections inside the specimen but is an approximation to the correct formula given by Stratton[31] which includes interference effects among the multiply reflected beams. The approximation implicit in equation (61) is normally valid; it assumes that the specimen is many wavelengths thick and that the radiation beam is never truly monochromatic. The only situation in which equation (61) may not be valid is when very thin specimens are used and when in fact interference fringes appear. However, such thin specimens are only used in measurements of very high absorption coefficients and then

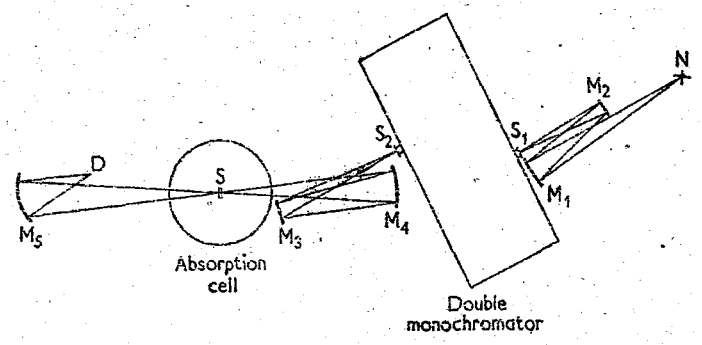


Figure 3. *A diagram of the experimental apparatus used in making the measurements on germanium and silicon discussed in this article.* D *is the source,* N *the detector,* S *the specimen,* S_1 *and* S_2 *the exit and entrance slits of the monochromator, and* M_1–M_5 *mirrors*

the denominator in equation (61) is effectively unity, the beam only passing through the specimen once so that interference effects are absent.

In the apparatus shown in Figure 3 the source used for the measurements on germanium and silicon was a tungsten-strip filament lamp and the detector a lead sulphide cell. To exploit fully the temperature dependence of the absorption edge, the specimens were mounted in an absorption cell[32] so that they could be held at any desired temperature normally lower than room temperature. The spectrometer was one devised by Roberts[33] to give a higher resolving power than a normal Leiss double monochromator by replacing the second prism of such a monochromator by a diffraction grating which could be rotated independently of the fore-prism. Used in conjunction with a lead sulphide cell, such a monochromator can give a maximum resolving power of the order of 10^4 in the wavelength region around the germanium and silicon absorption edges which is an order of magnitude better than the resolution obtainable using a normal monochromator with two flint glass prisms. In the measurements of absorption one cannot operate at this maximum resolving power; this is due essentially to the lowering of the

transmitted beam intensity. However, in the measurements on germanium, the resolving power never fell below 2×10^3 and was higher in many regions. That such a resolution is required to observe structure in the absorption due to exciton formation is easily seen from a rough calculation of the exciton binding energy to be expected in germanium. Using some mean values for the effective masses at the conduction and valence band edges, we find from equation (27) the exciton binding energy to be of the order of 10^{-3} eV. To detect structure lying within an energy range of this magnitude around 1 eV therefore requires a resolving power of the order used.

To obtain values of the absorption coefficient ranging over several orders of magnitude, one must use a series of specimens cut from high-purity single crystals and of varying thicknesses, the thickness of the specimen required to measure the absorption around some value K being such that $Kd \simeq 1$. The lower limit of the absorption which can be measured with any particular specimen is set by one's ability to measure the difference between the value of t taken at some wavelength and the almost constant value

$$t_0 = \frac{1-R}{1+R} \qquad \ldots (62)$$

to which t levels off at longer wavelengths, and which is used to fix the value of R for that specimen. As t approaches t_0, the absorption coefficient becomes proportional to the difference $(t_0 - t)$ which can be found only with an accuracy much less than that with which t and t_0 are measured in the region of $t \simeq t_0$. This limitation is not usually severe as the absorption in this region can be obtained from measurements on a thicker specimen. However, there is normally some limitation on the thickness of specimen which can be accommodated and hence a lower limit to the level of absorption which can be measured by any apparatus. Moss and Hawkins[34] have recently obtained a measure of the absorption to much lower levels than would normally be possible with a specimen by essentially using the specimen as a detector and measuring the photoconductive current induced in it as a result of the absorption. An upper limit to the absorption measurable with any specimen is set by the fact that it must transmit a detectable amount of radiation.

To prepare specimens for the measurement of absorption up to levels of the order of a few hundred cm^{-1} presents no great difficulty. However, for higher absorption levels the specimens required must only be of the order of some microns thick. Measurements made on evaporated films of germanium and silicon of this thickness indicated to Dash and Newman[35] that the absorption obtained in this way was not characteristic of the bulk single-crystal material. As a result of this, they developed a technique for producing thin specimens from a single crystal by a difficult and painstaking method of grinding and polishing. They were able in this way to study the absorption in germanium and silicon up to levels of 10^5 cm^{-1} by preparing specimens as thin as 0·5 micron. The thinnest specimens (less than 2 microns thick) prepared by this technique of Dash and Newman are permanently stuck down on to some backing material. This backing material is usually glass. Recently Roberts and Quarrington[36] have prepared thin specimens both backed and free from any backing. Measurements on these specimens[2,37] have shown that as the temperature is lowered the difference in the expansion coefficient of a

backed specimen and its backing can produce considerable strain in a thin specimen. This has two effects. Since the energy gap is altered by stressing the material, the position of the absorption edge will change. This amounts to as much as 0·007 eV in germanium. Secondly, the application of a stress can alter the band edge if this is normally degenerate and so produce additional structure in the edge.[37,38] Care must therefore be taken to mount any thin specimen in a strain-free manner so that values of absorption characteristic of the normal material can be obtained. In this connection, care must also be taken in comparing the results of different experiments to ensure that the specimen conditions are similar.

4. THE ABSORPTION EDGE OF GERMANIUM

We shall now discuss the information which can be gained from a detailed study of the absorption edge of germanium, this discussion being based on the work

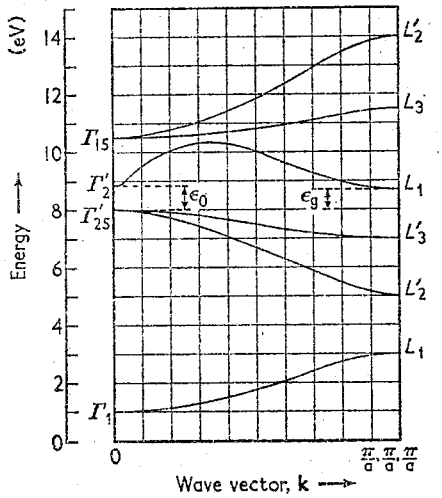


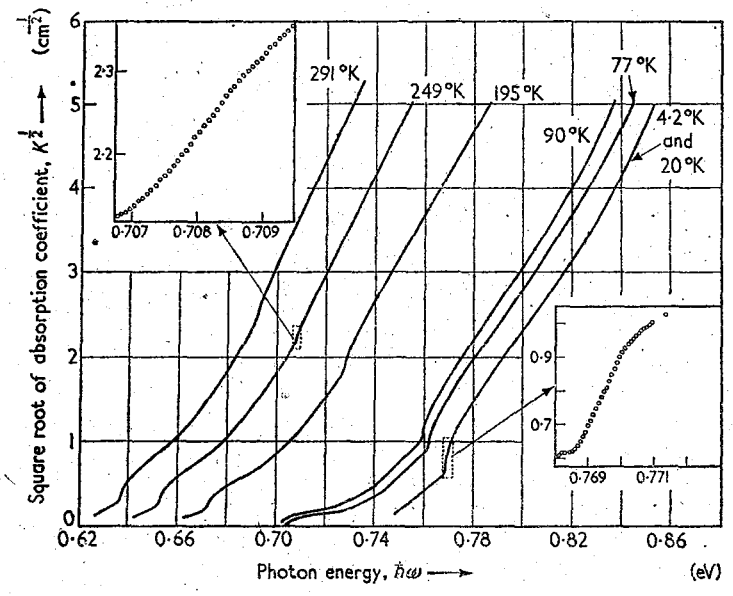
Figure 4. *A diagram of the band structure of germanium in a* ⟨111⟩ *direction as it is currently accepted. The irreducible representations according to which the electron wave functions at various points in* **k***-space transform are shown. The notation used here and elsewhere in this article for the representations of a diamond-type lattice is that of Herring* (J. Franklin Inst. 233, 525 (1942)). *Except for the energy gaps* ϵ_0 *and* ϵ_g, *the energy scale should be considered as only approximate*

described in references 1–3. The main features of the energy band structure of germanium will play an important part in this discussion and so, for convenience, we give in Figure 4 a diagram of the energy bands along a ⟨111⟩ direction in which the conduction band minima lie. We see from this Figure that the conduction band has a secondary minimum at $\mathbf{k} = 0$ and, as a consequence, the direct transition threshold occurs at this point. We shall break the discussion of the absorption

edge into two parts. Firstly, we shall examine the region of low level absorption ($K < 10$ cm^{-1}) where the absorption is due to indirect transitions producing electron–hole pairs, the electron being associated with states around the principal minima in the conduction band and the hole with states around the valence band maximum. We shall then consider the region of higher absorption ($K \simeq 10^3$ cm^{-1}) which we shall find is due to direct transitions between states close to the valence and conduction band edges at $\mathbf{k} = 0$.

4.1. The Indirect Transition Region

The experimentally determined absorption curves close to the beginning of the absorption edge are shown for a series of different temperatures between 4·2° K



(By courtesy of the Editor, *Physical Review*)

Figure 5. *The absorption edge at low levels in germanium at various temperatures. The inserts indicate the accuracy with which the experimental points define the curves*

and room temperature in Figure 5 where, for convenience, $K^{1/2}$ is plotted as a function of the absorbed radiation energy. Bearing in mind the features which we showed in sub-section 2.2 are to be expected of the absorption due to indirect transitions, an inspection of these curves shows that they possess many of these features. At the lowest temperatures, for example, the absorption appears to consist of two components, one beginning at an energy of about 0·75 eV and a second beginning at about 0·77 eV. As the temperature rises, two other components appear beginning at lower energies than the two which persist down to the lowest temperatures and, of these four components, the outer pair appear to have the same general shape, as also do the inner pair. At 77° K, the four components appear to begin at roughly energies of 0·705, 0·725, 0·745, and 0·760 eV. In

terms of indirect transitions, one would interpret the two low temperature components as being due to transitions in which phonons of two different energies are emitted into the lattice and the two components, which appear as the temperature rises, as being due to transitions in which phonons of these same energies are absorbed from the lattice. At still higher temperatures, further absorption begins at energies lower than that at which these four components appear.

To obtain the maximum amount of information from the absorption curves, one must follow up these indications of the composition of the curves, which can be obtained merely by inspection, by a detailed study of all the curves, examining and comparing the shapes, magnitudes, and positions of their several components. This can be done by methods of curve extrapolation and correlation and, although basically simple, the analysis of a set of curves like that of Figure 5 can be a long and painstaking job in which one learns much by experience. The details of the methods used can be found in the original descriptions of this work[1,4] and we shall not discuss them here but merely present the results. In presenting and discussing these results, we shall work in terms, not of the absorption coefficient K, but in terms of $\alpha \equiv K\hbar\omega$, so that the energy dependent factor $\hbar\omega$ always appearing in the denominator of expressions for K will not trouble us.

The absorption curve at any temperature is found to consist, in general, of four pairs of components. We shall denote the two components of the i'th pair ($i = 1,2,3,4$) by α_{ia} and α_{ie}, the subscripts a and e denoting the components associated with transitions in which a phonon is respectively absorbed or emitted. The threshold energies of α_{ia} and α_{ie} will be denoted by ϵ_{ia} and ϵ_{ie}, respectively, and their mean value by $\bar{\epsilon}_i$. It follows from the discussions following equations (48) and (60) that $(\epsilon_{ie}-\epsilon_{ia})$ is twice the energy $k\theta_i$ of the phonons absorbed or emitted in the transitions associated with the components α_{ia} and α_{ie}, and that $\bar{\epsilon}_i$ is the energy gap less the binding energy of the excitons formed in the transition. We shall associate larger phonon energies with larger values of i so that $\theta_i < \theta_{i+1}$ and consequently $\epsilon_{ia} > \epsilon_{i+1,a}$ and $\epsilon_{ie} < \epsilon_{i+1,e}$. Table 1 gives the values of the threshold energies and the corresponding phonon temperatures found from the analysis of the curves given in Figure 5, and some measurements at 363° K and 416° K not shown there. We see from this that most of the features of the curves shown in Figure 5 are due solely to transitions involving the two lower energy phonons with temperatures 90° K and 320° K. The two components comprising the curves at the lowest temperatures are due to transitions in which these two types of phonon are emitted and the other two components, appearing at lower energies as the temperature is increased, are due to transitions involving the absorption of these phonons. The close agreement between $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ at each temperature and the constancy of θ_1 and θ_2 with temperature are convincing evidence of the correctness of the analysis and its interpretation. The additional low level absorption appearing at the higher temperatures rather like a tail to the α_{2a} component is in fact composed of two components α_{3a} and α_{4a} associated with the absorption of phonons of temperatures 350° K and 420° K. The analysis of the absorption due to these two components is difficult due to the low absorption levels. One might hope to overcome this by raising the temperature and increasing the number of phonons available to assist in the transitions. However, as the temperature is raised more and more free carriers are generated and one then runs into difficulties due to the large amount of free carrier absorption present which is not easy to subtract out

Table 1. The Values of the Threshold Energies ϵ_{1a}, ϵ_{2a}, ϵ_{2a}, ϵ_{4a}, ϵ_{1c}, and ϵ_{2c} Found from the Analysis of the Absorption Curves and the Corresponding Values of the Energy Gaps $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$ at Various Temperatures T for Germanium.† Also Shown are the Four Phonon Temperatures θ_1, θ_2, θ_3, and θ_4 and the Half-width σ of the Gaussian Broadening Suffered by the Exciton Absorption

T (°K)	ϵ_{4a} (eV)	ϵ_{2a} (eV)	ϵ_{2a} (eV)	ϵ_{1a} (eV)	ϵ_{1c} (eV)	ϵ_{2c} (eV)	θ_1 (°K) = $\frac{1}{2k}(\epsilon_{1c}-\epsilon_{1a})$	θ_2 (°K) = $\frac{1}{2k}(\epsilon_{2c}-\epsilon_{2a})$	θ_3 (°K) = $\frac{1}{k}(\epsilon-\epsilon_{2a})$	θ_4 (°K) = $\frac{1}{k}(\epsilon-\epsilon_{1a})$	$\tilde{\epsilon}_1$ (eV) = $\frac{1}{2}(\epsilon_{1c}+\epsilon_{1a})$	$\tilde{\epsilon}_2$ (eV) = $\frac{1}{2}(\epsilon_{2c}+\epsilon_{2a})$	$\sigma\times10^4$(eV)
4·2					0·7485	0·7686					0·7407‡	0·7410§	2·7
20					0·7480	0·7679					0·7402‡	0·7403§	2·7
77			0·7063	0·7264	0·7419	0·7614	90	320			0·7342	0·7339	3·0
90			0·7035	0·7238	0·7393	0·7587	90	320			0·7315	0·7311	3·0
195	0·6633	0·6694	0·6718	0·6919	0·7073	0·7269	89	320	349	420	0·6996	0·6994	4·0
249	0·6435	0·6492	0·6523	0·6718	0·6868	0·7072	87	319	356	422	0·6793	0·6797	8·0
291	0·6273	0·6332	0·6367	0·6570	0·6724	0·6913	88	317	357	426	0·6647	0·6640	9·9
363	0·5980	0·604	0·6073						359	426		0·6349¶	16
416			0·5877									0·6153¶	20

† All these energy values and those given in Table 2 have been corrected by the factor 0·99955 as explained in reference 4.
‡ Calculated from $\tilde{\epsilon}_1 = \epsilon_{1c} - k\theta_1$ with $\theta_1 = 90°$K.
§ Calculated from $\tilde{\epsilon}_2 = \epsilon_{2c} - k\theta_2$ with $\theta_2 = 320°$K.
¶ Calculated from $\epsilon_3 = \epsilon_{2a} + k\theta_2$ with $\theta_2 = 320°$K.

with any great accuracy. It may seem surprising that no values of ϵ_{3e} and ϵ_{4e} are given in Table 1. However, a rough calculation shows that, from the magnitudes of α_{3a} and α_{4a}, α_{3e} and α_{4e} would not be expected to show any detectable effect when added to α_{1e} and α_{2e}, and this is borne out by a detailed examination of the absorption curves.

In discussing the shapes and magnitudes of the components, little can be said about the shapes of the components α_{3a} and α_{4a} due to the low absorption levels at which they appear. The two components α_{2a} and α_{2e} are found to have the same shape at each temperature where they both appear, as one would expect, but this shape changes slightly with temperature. This change takes the form of a general smoothing out of the shape of a component as the temperature rises and is found to be due to a broadening of Gaussian type whose magnitude can be characterized by a half-width σ which increases with temperature. We shall have more to say about this broadening later. At present, we merely note that the value of σ at each temperature is given in Table 1. If ϵ_2 is the threshold energy of the component α_2 associated with either phonon absorption or emission, then the basic shape of α_2 is made up of three parts and is given in form by

$$\alpha_2 = a[(\hbar\omega-\epsilon_2)^{1/2}+b(\hbar\omega-\epsilon_2-0\cdot0010)^{1/2}+c(\hbar\omega-\epsilon_2-0\cdot0027)^{3/2}] \quad \ldots (63)$$

where b and c are constants, a is a function of temperature, and we are measuring energies in electron volts. The shape of α_{2a} and α_{2e} at any temperature can be obtained from equation (63) by applying to it a Gaussian broadening whose magnitude is determined by the value of σ in Table 1. The energies of $0\cdot0010$ eV and $0\cdot0027$ eV appearing in equation (63) are of course determined from the experimental absorption curves and have estimated errors of $\pm0\cdot0001$ eV and $\pm0\cdot0004$ eV respectively. The values of ϵ_{2a} and ϵ_{2e} given in Table 1 are those required in equation (63) so that, after broadening, α_{2a} and α_{2e} derived from equation (63) fit the components obtained from the absorption curves. ϵ_{2a} and ϵ_{2e}, therefore, do not correspond to the energies at which the broadened components actually begin but to the energies at which the associated basic shape begins. At any temperature, the ratio of α_{2e} to α_{2a} should be given by expression (49) and since we know the phonon temperature, θ_2, we can find the value of ρ_2 associated with transitions involving phonons of this energy. It turns out that $\rho_2 \simeq 2$, this value being independent of temperature as one would expect. This does not agree with the results quoted in reference 1 where ρ_2 is implicitly given as unity. This is due to an unfortunate error which was made in the analysis at that time when the components α_{3a} and α_{4a} were neglected due to the lack of good enough experimental data. This had the effect of increasing the magnitude of α_{2a} and so decreasing ρ_2 to what appeared to be unity, so that the significance of this factor was missed.

The same remarks regarding the shapes of the components α_{1a} and α_{1e} hold as do for the components α_{2a} and α_{2e}. At any temperature they have the same shape, but this changes slightly with temperature due to broadening. The basic shape α_1, associated with α_{1a} and α_{1e}, is found to be

$$\alpha_1 = a'(\hbar\omega-\epsilon_1)^{3/2} \quad \ldots (64)$$

ϵ_1 being the threshold energy and a' an energy independent but temperature dependent quantity. The ratio of α_{1e} to α_{1a} and the value of $\theta_1 = 90°$ K

is found to lead to a value from equation (49) of $\rho_1 = 1 \cdot 01 \pm 0 \cdot 005$ at all temperatures.

Having presented all the data which has been obtained from a detailed analysis of the absorption curves in Figure 5, we shall now proceed in the following section to discuss what information can be deduced from it about various properties of germanium.

4.2. Interpretation of the Indirect Absorption Curves—Band Structure, Excitons, and Phonons

We shall first of all discuss the interpretation of the shapes of the various components, bearing in mind our remarks about these shapes at the end of sub-section 2.2. From these remarks, the interpretation of the form of the components α_{2a} and α_{2e} given in equation (63) is clear. The initial rise as the square root of the energy is due to allowed transitions into the ground state of the exciton. The contribution from the second term is due to transitions into probably the first excited state of the exciton lying $0 \cdot 0010 \pm 0 \cdot 0001$ eV above the ground state. Absorption into higher excited states of the exciton is negligible, the third term arising from allowed transitions producing unbound electron–hole pairs in the crystal. The threshold for these transitions is $0 \cdot 0027 \pm 0 \cdot 0004$ eV above the threshold for transitions into the exciton ground state, so that this energy difference is equal to the binding energy of an exciton in germanium with a wave vector equal to that of the conduction band minima, i.e.

$$E_{ex}(\mathbf{k}_c) = 0 \cdot 0027 \pm 0 \cdot 0004 \text{ eV} \qquad \ldots (65)$$

By increasing the values of $\bar{\epsilon}_2$ given in Table 1 by this value of $E_{ex}(\mathbf{k}_c)$ we obtain a highly accurate set of values for the indirect energy gap ϵ_g as a function of temperature. We shall have more to say about this energy gap in sub-section 4.3, when it will be compared with the direct energy gap. It is shown as a function of temperature there in Figure 8.

The production of excitons in the absorption process is also evident from the photoconductive spectrum. From a knowledge of the photoconductive response of a specimen as a function of the absorbed radiation energy and the corresponding absorption coefficient, the photoconductive current J per unit quantum of radiation absorbed can be found. At low absorption levels when the radiation intensity is practically uniform throughout the specimen, any variation of J with energy must be attributed to a variation in quantum efficiency. Now, if each quantum of radiation absorbed produces an unbound electron–hole pair, the quantum efficiency should be constant with energy in this region. The fact that the measured photoconductivity[1] leads to a value of J which falls rapidly just at those energies at which the absorption rises abruptly, indicates that in these regions the absorption cannot be completely due to processes producing free current carriers. The production of excitons by the absorption process can clearly account for this.

The simple rise as the three-halves power of the energy from threshold of the components α_{1a} and α_{1e} suggests that no exciton producing transitions contributes to these components at all. However, it would follow from this that the values of $\bar{\epsilon}_1$ in Table 1 should be larger than the values of $\bar{\epsilon}_2$ by the binding energy $0 \cdot 0027$ eV of the excitons but this is not the case, the values of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ disagreeing at most by $\pm 0 \cdot 0007$ eV. The virtual equality of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ forces us to conclude that exciton

absorption is present in α_{1a} and α_{1e}. It might, therefore, seem that the absorption in α_{1a} and α_{1e} is due to forbidden transitions and that the separate contributions from absorption into different exciton states and unbound pair states has not been resolved by the analysis.

Having determined accurate values of the energy gap ϵ_g as a function of temperature, these can be used in the expression giving the intrinsic carrier concentration n_i as a function of temperature; this is

$$n_i = 4{\cdot}82\times 10^{15}\, T^{3/2}(\bar{m}/m)^{3/2} N_c^{1/2} \exp(-\epsilon_g/2kT) \qquad \ldots (66)$$

where $\bar{m}$ is the combined density-of-states effective mass of the valence and conduction band edges. For simple bands with effective masses m_c and m_v, $\bar{m} = (m_c m_v)^{1/3}$ but for germanium the expression is more complicated due to the spheroidal and degenerate band edges. Lax and Mavroides[39] have developed an expression for $\bar{m}$ in germanium and using this in conjunction with the valence and conduction band parameters determined from cyclotron resonance experiments[40,41] carried out at 4·2° K, one finds that $\bar{m} \simeq 0{\cdot}275\, m$ at this temperature. The intrinsic carrier concentration has been found by Morin and Maita[14] to be well represented over the temperature range from 250° to 500° K by the expression

$$n_i = 1{\cdot}76\times 10^{16}\, T^{3/2} \exp(-0{\cdot}785/kT) \qquad \ldots (67)$$

Comparing expressions (66) and (67) in the temperature range where both are valid, it turns out that $N_c(\bar{m}/m)^3$ must increase with temperature above 250° K. If the conduction band minima in germanium are assumed to lie inside the first Brillouin zone, then, since they lie in ⟨111⟩ directions, $N_c = 8$. With this value of N_c, one finds that the effective mass $\bar{m}$ must originally decrease as the temperature rises from zero and pass through a minimum value before rising to a value of $\simeq 0{\cdot}26\, m$ at room temperature, which is lower than its value at 4·2° K. On the other hand, if the conduction band minima are assumed to be at the zone edge, then $N_c = 4$ and $\bar{m}$ is found to increase with temperature from 4·2° K, having increased by about 15 per cent by the time room temperature is reached. This monotonic increase of $\bar{m}$ with temperature is the more acceptable temperature dependence and we take it to indicate that $N_c = 4$ in agreement with other work on germanium.[42–44] It is interesting to note in connection with this discussion that, if we consider the exciton binding energy used in calculating ϵ_g to be a variable, then with $N_c = 8$ a binding energy about four times greater than our value must be assumed to produce an effective mass $\bar{m}$ increasing monotonically with temperature. This would be in violent disagreement with the calculated binding energy which we shall discuss later. Furthermore, with $N_c = 4$, a non-zero value of the exciton binding energy must be assumed to produce a suitable temperature dependence for $\bar{m}$ which can be taken as further evidence for the correctness of our interpretation of the absorption components.

It follows from this that the phonons assisting in the indirect transitions have wave vectors at a ⟨111⟩ direction zone edge and we can identify the temperatures 90° K, 320° K, 350° K, and 420° K with the energies of the TA, LA, LO, and TO phonons respectively, at this point. These energies are in excellent agreement with the values found by Brockhouse and Iyengar[45] from neutron scattering experiments for the phonon energies corresponding to this wave vector. The determination of the temperatures 90° K and 320° K of the TA and LA modes[1]

gave the first indication that the lattice vibrational spectrum of germanium could not be explained in terms of the theory developed by Helen Smith[46] for the vibrations of a diamond-type lattice, assuming that each atom had an appreciable interaction only with closely neighbouring atoms. This has been pointed out even more convincingly by Herman[47] who has shown that the interactions between each atom and up to at least its fifth nearest neighbours must be included before the phonon dispersion curves determined by Brockhouse and Iyengar can be fitted. Lax[48] has suggested that the long-range interactions which are apparently present are due to the importance of quadrupole–quadrupole interactions in a diamond-type lattice and calculations on the basis of this suggestion are currently being made.

We are now in a position to discuss the selection rules governing the indirect transitions in germanium, i.e. the determination of which transitions are allowed and which are forbidden, and the intermediate states principally used in these transitions. The transitions were defined to be allowed or forbidden according as $X_{cv}^{lb}(\mathbf{k}_c, 0)$, defined by equation (40), is non-zero or zero and this can be determined by group theory from a knowledge of the transformation properties of the initial and final state electron wave functions and the operators $\mathbf{O}_b$ and $\mathbf{p}$ describing the interaction of the electrons with the phonons and electromagnetic radiation. For phonons with wave vector lying at a $\langle 111 \rangle$ direction zone edge, the operator $\mathbf{O}_b$ transforms as the irreducible representations L_3', L_2', L_1, and L_3 of the group L for phonons of type TA, LA, LO, and TO, respectively, and $\mathbf{p}$ transforms as a polar vector. The irreducible representations for which the electron wave functions form bases are shown in Figure 4 for $\mathbf{k} = 0$ and lying at a $\langle 111 \rangle$ direction zone edge for the valence and conduction bands and several neighbouring bands which can provide suitable intermediate states. The initial electron state—the valence band edge—transforms as Γ_{25}' and the final state—the conduction band edge—as L_1. We can neglect the spin of the electron in this discussion as neither of the operators $\mathbf{O}_b$ or $\mathbf{p}$ operates on spin.

As was originally pointed out by Elliott[23], transitions involving the absorption or emission of optical phonons are forbidden. This is a consequence of the fact that both the groups Γ and L contain the inversion operator so that the wave functions for states at the conduction and valence band edges have well-defined parity as also do the operators $\mathbf{O}_b$ and $\mathbf{p}$. The initial and final electronic states both have even parity whereas the electric dipole radiation absorbed in the transition and characterized by the operator $\mathbf{p}$ has odd parity. Thus the terms in equation (40) contributing to X_{cv}^{lb} will all be zero if the phonons involved in the transition have even parity and are therefore optical. This explains why the components in the absorption curves associated with the two optical phonons are weak relative to the other components. Unfortunately it has not so far proved possible to obtain the shapes of these components from experimental measurements so that this could be more accurately verified.

Transitions involving acoustical phonons are allowed since intermediate states can be found which give a non-zero contribution to X_{cv}^{lb}. It is found that transitions taking place through the intermediate valence band state with symmetry Γ_1 and the intermediate conduction band states with symmetry L_1 and L_3 are forbidden for all types of phonons. However, transitions through the intermediate states with symmetry Γ_{15}, L_2', and L_3' are allowed for both TA and LA phonons but

those through the intermediate state with symmetry Γ_2', i.e. the conduction band at $\mathbf{k} = 0$, are allowed only for LA phonons. We can now understand the values of $\rho_1 = 1 \cdot 01$ and $\rho_2 \simeq 2$ associated with the transitions involving the TA and LA phonons respectively. It was mentioned in the discussion of the theory of indirect transitions that the phonon energies will normally form only a small part of the energy denominators appearing in equation (40) so that ρ will have a value of about unity. However, this is not the case for transitions proceeding via the conduction band state in germanium at $\mathbf{k} = 0$. This is an intermediate state of the type m in equation (40) and the energy difference between it and the final state at the conduction band edge is only $\simeq 0 \cdot 15$ eV. This is only about five times greater than the LA phonon energy which corresponds to a temperature of 320° K. Not only that, but it is considerably smaller than any of the other energy denominators appearing in the expressions for X_{cv}^{lb}, so that we might expect transitions through this state to predominate when these transitions involve the LA phonons. If this were the case, we find that $\rho_2 \simeq 2 \cdot 2$, whereas transitions through any other of the possible intermediate states would lead to values of $\rho_2 < 1 \cdot 2$. The value of $\rho_2 \simeq 2$ deduced from the absorption curves therefore confirms that transitions involving the LA phonons proceed principally through the conduction band at $\mathbf{k} = 0$. The situation with regard to the components α_{1a} and α_{1e} due to transitions involving the TA phonon is not so satisfactory. $\rho_1 = 1 \cdot 01$ indicates that the conduction band state at $\mathbf{k} = 0$ is not predominating as an intermediate state so that it would seem that the transitions are proceeding through the other intermediate states and are therefore allowed. This does not agree, as was pointed out with our earlier supposition that these transitions are forbidden, deduced from the energy dependence and thresholds of the components. This difficulty remains to be cleared up.

Note added in proof

It has recently been pointed out by Kane that the symmetries given by Elliott,[23] and quoted in this article, for transverse phonons with wave vectors lying at the zone edge in a (111) direction are incorrect. These should be interchanged, the operator O_b transforming according to the irreducible representation L_3 for TA modes and L_3' for TO modes. Thus, for transverse lattice vibrations with this wave vector, the two atoms in the unit cell vibrate in phase in the optical modes and out of phase in the acoustical modes in contrast to the relative motion in optical and acoustical modes with zero wave vector.

This observation clarifies the interpretation of the indirect absorption region in germanium to a great extent. In particular, the previously unexplained shape of the components α_{1a} and α_{1e} associated with TA phonons can now be understood since transitions involving these phonons are forbidden. Presumably the shape of the component α_4 associated with TO phonons, if it could be determined, would correspond to allowed transitions. Some preliminary calculations have been carried out by Elliott and the author principally to investigate what appears at first to be the anomalously large magnitude of the components associated with the forbidden TA phonon transitions. It is found that the relative magnitudes of the components associated with the different types of phonon are all consistent with the appropriate non-zero matrix elements of the electron–phonon interaction

being of the same order of magnitude but with the matrix elements associated with transverse phonons being a few times larger than those associated with the longitudinal phonons. This greater effectiveness of the transverse modes at scattering is also seen in the absorption in silicon where, although group theory gives no forbidden transitions, the absorption components associated with longitudinal phonons are not observed and so must be about an order of magnitude smaller than those associated with the transverse phonons.

The magnitude of the components of the absorption due to indirect transitions can be used to obtain values of the quantities X_{cv}^{lb} or their derivatives according as the transitions associated with the various components are allowed or forbidden. These in themselves are not of any great value as the X_{cv}^{lb} defined in equation (40) consist in general of many terms arising from transitions through different intermediate states. However, as we have just shown, the transitions in germanium involving LA phonons proceed virtually exclusively through the conduction band state at $\mathbf{k} = 0$ so that we have

$$X_{cv}^{l,LA}(\mathbf{k}_c, 0) = \frac{\xi_{LA}(\mathbf{k}_c) \cdot [c, \mathbf{k}_c | \mathbf{C}_{LA}(\mathbf{k}_c)| c, 0] \alpha \cdot \mathbf{p}_{cv}(0)}{\epsilon_c(\mathbf{k}_c) - \epsilon_c(0) + (-1)^l k\theta_{LA}(\mathbf{k}_c)} \quad \ldots (68)$$

In the following section, where we discuss the direct transitions in germanium, it is shown how a measure of $|\mathbf{p}_{cv}(0)|^2$ can be obtained from the magnitude of the absorption due to direct transitions in which unbound electron–hole pairs are produced. This takes into account, in a simple manner, the degenerate nature of the valence band edge and gives a value of $|\mathbf{p}_{cv}(0)|^2$ in excellent agreement with the value of this quantity deduced from cyclotron resonance measurements by Dresselhaus et al.[40] Using this value of $|\mathbf{p}_{cv}(0)|^2$, the effective masses at the valence and conduction band edges and the binding energy of an exciton with wave vector $\mathbf{k}_c$ given in equation (65), we can reduce the expression for the absorption coefficient[49] K_{LA} due to indirect transitions involving the emission of LA phonons at low temperatures to

$$\hbar\omega K_{LA} = 1{\cdot}54 \times 10^{-14} [c, \mathbf{k}_c | \mathbf{C}_{LA}(\mathbf{k}_c)| c, 0]^2 [\hbar\omega - \epsilon_g - k\theta_{LA}(\mathbf{k}_c)]^{3/2} \ldots (69)$$

where energies are expressed in electron volts and lengths in centimetres. From an analysis of the absorption curve at $4{\cdot}2°$ K we find that

$$\hbar\omega K_{LA} \simeq 260 [\hbar\omega - \epsilon_g - k\theta_{LA}(\mathbf{k}_c)]^{3/2} \quad \ldots (70)$$

which probably is correct to ± 10 per cent. Thus

$$[c, \mathbf{k}_c | \mathbf{C}_{LA}(\mathbf{k}_c)| c, 0]^2 \simeq 1{\cdot}7 \times 10^{16} \text{ eV}^2 \text{ cm}^{-2} \quad \ldots (71)$$

In discussing the theory of indirect transitions we remarked that, for $\mathbf{k} \sim \mathbf{k}'$, $(c, \mathbf{k}|\mathbf{C}(\mathbf{k}-\mathbf{k}')|c, \mathbf{k}')$ is equal to $(\mathbf{k}-\mathbf{k}')$ times the deformation potential of band c at $\mathbf{k}$. Let us extend this relationship and put

$$[c, \mathbf{k}_c | \mathbf{C}_{LA}(\mathbf{k}_c)| c, 0] = (c, \mathbf{k}_c | \epsilon_{LA}| c, 0) \mathbf{k}_c \quad \ldots (72)$$

where $(c, \mathbf{k}_c|e_{LA}|c, 0)$ is some sort of generalized deformation potential for the conduction band; then we find that

$$(c, \mathbf{k}_c|e_{LA}|c, 0) \simeq 1{\cdot}3 \text{ eV} \qquad \ldots (73)$$

We can compare this with the value of the deformation potential which characterizes the scattering of electrons with wave vectors close to $\mathbf{k}_c$ by long-wave LA phonons, along the direction of $\mathbf{k}_c$. In the notation of Herring and Vogt,[50] this deformation potential is just $(\Xi_d + \Xi_u)$ which, according to their results, has a value of about 13 eV. This, of course, should not be the same as the quantity in equation (73) but on the other hand should not be vastly different from it. The relative magnitudes of the two quantities are entirely reasonable.

The broadening suffered by the basic shapes given in equations (63) and (64) can be studied in most detail in the components α_{2s} and α_{2c} due to the sharp initial rise as the square root of the energy. The values of σ given in Table 1 were obtained from a study of the broadening of this initial rise and so apply to the exciton-producing transitions. The slow rise of the absorption due to transitions into the unbound pair states is not sensitive to small variations in σ. From the values of σ we can deduce relaxation times $\tau = \hbar/\sigma$ for the excitons formed in these transitions. τ is found to have a constant value around $1{\cdot}5 \times 10^{-11}$ sec at low temperatures which begins to fall roughly as T^{-2} around 100° K reaching a value of $2{\cdot}1 \times 10^{-12}$ sec at 416° K.

Before discussing the excitons formed in the indirect transitions it is convenient to examine the absorption at higher levels. We shall find that this is due to direct transitions and that excitons are formed in these as well as in the indirect transitions. These we shall call direct excitons in distinction to the indirect excitons formed by indirect transitions. We shall discuss both types of excitons in sub-section 4.4.

4.3. The Direct Transition Region

In Figure 6 we show the complete absorption edge in germanium at 20° K made up from measurements on a series of specimens of several thicknesses down to 5 microns. It is similar to but contains more detail than the absorption curves measured earlier at 77° K and room temperature by Dash and Newman.[35] The absorption due to the indirect transitions we have just been discussing, lies in the region up to $\sim 10\text{-cm}^{-1}$ and is due to transitions between states close to the valence and conduction band edges. As the absorbed radiation energy increases, this absorption rises fairly slowly and is due to indirect transitions between valence and conduction band states further and further away from the band edges. It is seen from the band structure shown in Figure 4 that as the energy increases a point will be reached at which enough energy is available from the radiation to produce transitions across the direct energy gap at $\mathbf{k} = 0$. The absorption will then increase rapidly to the higher levels produced by direct transitions. This rise is seen to occur in Figure 6 at an energy of about 0·885 eV from ~ 100 cm^{-1} up to $\sim 3{,}000$ cm^{-1}. The absorption rises to a peak, falls slightly, and then continues to rise more slowly again. This peak is a broadened line due to direct transitions into an exciton state with $\mathbf{K} = 0$ and the further absorption at higher energies arises from transitions forming unbound electron–hole pairs.

The absorption due to these direct transitions has been studied in more detail[2] at a series of temperatures. The transitions producing this absorption between

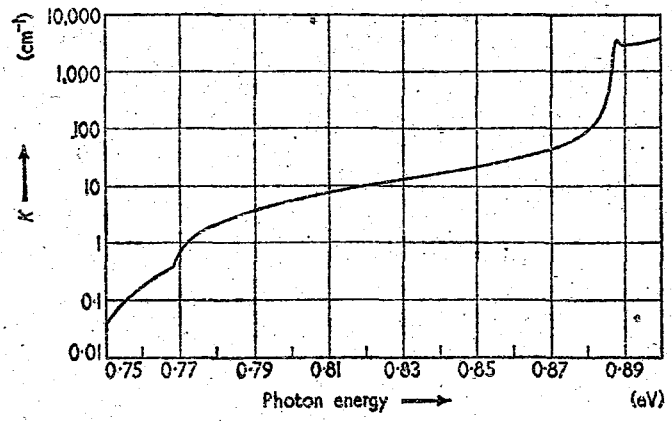
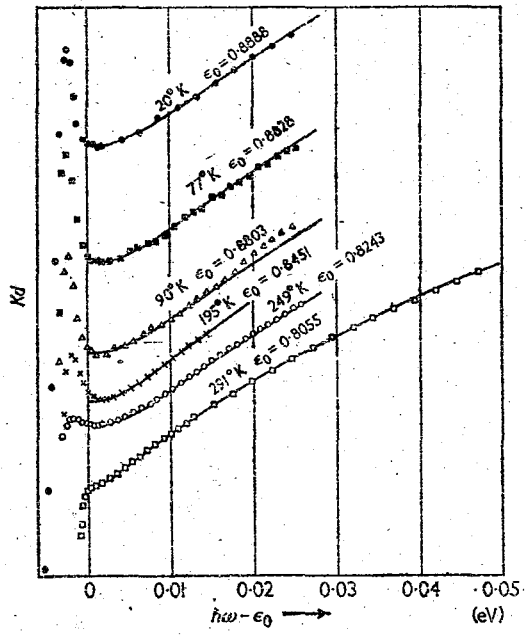


Figure 6. *The absorption edge in germanium at 20° K*



(By courtesy of the Physical Society)

Figure 7. *The variation of absorption* Kd *with absorbed radiation energy at various temperatures in germanium in the direct transition region. The points are the experimentally observed values and the full curves have been obtained from the theoretical expression* (31), *using the values of* $E_{ex}(0)$ *given in the text. The data at the various temperatures have been aligned in energy using the value of* ϵ_0 *associated with each curve and each set has been given an arbitrary vertical displacement. The general level of absorption in each curve is around* 3,300 cm^{-1}. *At room temperature the data have been plotted using a horizontal scale reduced by a factor of* 2 *from that shown.*

states of symmetry Γ'_{25} and Γ'_2 are allowed and so the absorption into the unbound pair states should have an energy dependence of the type given by equation (31). By fitting this expression to the experimental points one can deduce the value of the direct energy gap ϵ_0 at each temperature and also the binding energy $E_{ex}(0)$ of the excitons produced. $E_{ex}(0)$ must be such that it scales correctly in energy the shape given by equation (31) and is also equal to the energy difference between ϵ_0 and the position of the exciton absorption peak. Figure 7 shows the success with which this fitting can be carried out at different temperatures. The values of $E_{ex}(0)$ found were 0·0012 eV at 20°, 77°, and 90° K, 0·0010 eV at 195° K, 0·0011 eV at 249° K, and 0·0009 eV at room temperature where the results of the fitting

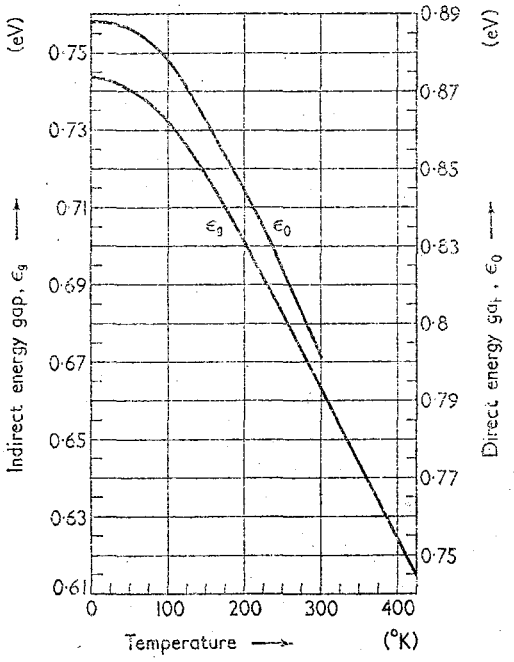


Figure 8. *The direct and indirect energy gaps as a function of temperature in germanium*

are less reliable due to the virtual disappearance of the exciton peak as a result of considerable broadening. The excellence of the agreement achieved between the experimental points and expression (31) is better than perhaps one should hope since expression (31) really applies only to simple band edges and not to the type of degenerate valence band edge encountered in germanium.

Figure 8 shows the variation of the direct energy gap ϵ_0, found from this analysis, as a function of temperature and on the same Figure the variation of the indirect energy gap ϵ_g with temperature. Both energy gaps have a similar temperature dependence but the difference $\epsilon_0 - \epsilon_g$ decreases by about 0·007 eV as the temperature increases from zero up to room temperature. This change in $\epsilon_0 - \epsilon_g$ is consistent with the increase with temperature of the mass associated with the conduction and valence band edges found earlier from the indirect transitions and

the decrease in the conduction band mass at $\mathbf{k}=0$ as the temperature increases found by Zwerdling, Lax, and Roth.[51] The agreement between our value of ϵ_0 at room temperature and that obtained from measurements of the oscillatory magneto-absorption[51] is excellent. However, as the temperature decreases the agreement breaks down, the values obtained in the two experiments differing by as much as 0·007 eV at low temperatures. The disagreement is explained by the specimens used in the two experiments being under different conditions. In the magneto-absorption measurements, the specimens were glued to a glass backing so that, as has been shown,[37] on lowering the temperature they were subjected to a considerable strain accompanied by a corresponding change in the energy gap.[52] The specimen used in the measurements shown in Figure 7 was free from any backing and mounted in as nearly as possible to a strain-free manner. We therefore take the values of the energy gap deduced from these measurements as characteristic of normal germanium.

We have so far been concerned only with fitting the shape given by expression (31) to the experimental points. However, by comparing the absolute magnitude of the absorption with an expression of the type (31) one can find[49] the magnitude of the momentum matrix element $\mathbf{p}_{cv}(0)$ between the bands at $\mathbf{k}=0$, which is closely related to the oscillator strength of the transitions causing the absorption. We say an expression of the type (31) since in this expression simple valence and conduction bands were assumed; in discussing the magnitude of the absorption in germanium we must modify expression (31) to account for the degenerate nature of the valence band edge. We do this in the following way. With spin–orbit coupling taken into account, the valence band edge is four-fold degenerate, the four states having Bloch functions $\frac{3}{2}u_j$, $j=\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}$, transforming as the four components of angular momentum. The conduction band edge is two-fold degenerate, the two Bloch functions $u_c^{\pm}$ transforming as the components of angular momentum $\frac{1}{2}$. Defining

$$p_{cv} = -i\hbar\frac{(2\pi)^3}{\Omega}\int_c d^3r\, u_c^{+*}(\mathbf{r})\frac{\partial}{\partial z}u_{1/2}(\mathbf{r}) \qquad \ldots(74)$$

we find that, neglecting the electron–hole interaction (cf. equation (13)),

$$K = 4{\cdot}44\times10^{-3}\frac{2e^2}{nm^2c\omega}\left(\frac{2m}{\hbar^2}\right)^{3/2}p_{cv}^2(\hbar\omega-\epsilon_0)^{1/2} \qquad \ldots(75)$$

The numerical factor in this expression comes from the effective masses of the conduction band[53] at $\mathbf{k}=0$ and the light and heavy hole valence bands[40, 41] whose fluting we have neglected. For simple bands our discussion in sub-section 2.1 shows that, to take into account the electron–hole interaction in the unbound states, we merely replace $(\hbar\omega-\epsilon_0)^{1/2}$ by $\pi E_{ex}^{1/2}(0)\exp z/\sinh z$. We shall assume we can make the same replacement in this case which cannot be a bad assumption considering the agreement which we obtained between expression (31) and the experimental points. We then find that at $\hbar\omega=\epsilon_0$, the absorption coefficient has the value

$$4{\cdot}44\times10^{-3}\frac{2e^2}{nm^2c\epsilon_0}\frac{\hbar}{}\left(\frac{2m}{\hbar^2}\right)^{3/2}p_{cv}^2\,2\pi E_{ex}^{1/2}(0) \qquad \ldots(76)$$

The most suitable quantity to work with is the oscillator strength F defined by Dresselhaus, Kip, and Kittel[40] which in terms of p_{cv} is given by

$$F = -\frac{3}{2}\frac{\hbar^2}{m^2}\frac{p_{cv}^2}{\epsilon_0} \qquad \ldots(77)$$

Using this in equation (76) along with the value 0·0011 eV for the exciton binding energy, the absorption coefficient at $\hbar\omega = \epsilon_0$ becomes

$$-116\frac{2mF}{\hbar^2}\text{cm}^{-1} \qquad \ldots(78)$$

The absorption level at $\hbar\omega = \epsilon_0$ at 20° K is 3,250 cm^{-1}, which is correct to ± 5 per cent. This value is lower than those given in references 2 and 3 but should be regarded as the better value. The earlier measurements were taken on a specimen which was in fact wedge shaped so that the actual specimen thickness through which the radiation was passing was difficult to assess. This uncertainty in thickness does not influence the shape of the absorption curve but does influence the absorption level. Indirect transitions clearly contribute a negligible amount to the absorption at $\hbar\omega = \epsilon_0$, so that we can compare this value of 3,250 cm^{-1} directly with equation (78). This leads to

$$F = -28{\cdot}0\frac{\hbar^2}{2m} \qquad \ldots(79)$$

which should be compared with the value obtained from cyclotron resonance[40] of $-28{\cdot}6\ \hbar^2/2m$. In making this comparison, one must consider the effects of the exchange potential in the Hamiltonian describing the electrons which we have so far neglected. This has been discussed recently by both Kane[54] and Phillips.[55] Formally, all the momentum matrix elements appearing in our expressions for the absorption coefficients should be replaced by the same matrix elements of the commutator

$$\frac{im}{\hbar}(H_{el}, \mathbf{r})$$

where H_{el} is the one-electron Hamiltonian. This operator has, of course, the same transformation properties as $\mathbf{p}$. Furthermore, an additional term is introduced into the expressions for the effective mass which results in the F obtained from cyclotron resonance measurements being not necessarily the same as that obtained from optical absorption, the difference being a measure of the strength of the exchange potential. However, as Kane has pointed out,[54] the excellent agreement between the measured value of the conduction band effective mass at $\mathbf{k} = 0$ and that calculated from the value of F obtained from measurements of the valence band effective mass[40] indicates that the effects of exchange are negligible. This is further supported by the value we obtain for F from the absorption measurements being essentially equal to that obtained from the measurements of effective mass.

4.4. The Direct and Indirect Excitons in Germanium

We have found evidence in the absorption edge that excitons are produced by both direct and indirect transitions. The excitons formed in the direct transitions

have zero wave vector and have a binding energy of $0{\cdot}0011 \pm 0{\cdot}0001$ eV. The indirect excitons have a wave vector lying at a $\langle 111 \rangle$ direction zone edge and have a binding energy of $0{\cdot}0027 \pm 0{\cdot}0004$ eV with another level lying $0{\cdot}0010 \pm 0{\cdot}0001$ eV above the ground state. Measurements of these same exciton energy levels have also been made by the Lincoln Laboratory group,[56,57] using the techniques of magneto-absorption discussed by Lax.[58] They obtain a binding energy for the direct exciton larger by about a factor of 2 than the value found by our analysis. For the indirect exciton, they also find two levels split by $0{\cdot}0010$ eV with the binding energy in the lowest level $0{\cdot}0033 \pm 0{\cdot}0004$ eV. The agreement in the indirect exciton levels found from the two experiments is quite satisfactory, the level splitting being identical in the two cases and binding energy agreeing to within the errors. The situation with regard to the direct exciton is far less satisfactory, however, and is at present unexplained. However, the binding energy of $0{\cdot}0020$–$0{\cdot}0025$ eV found from the magneto-absorption measurements would seem, on the basis of a simple calculation, to be abnormally large. This follows from the fact that, for any two particles interacting via an attractive potential, an upper limit to the binding attainable is obtained by taking the reduced mass of the pair equal to that of the lighter particle. In the case of the direct exciton, the electron has the smaller effective mass—$m_c = 0{\cdot}037m$—and this gives an upper limit to the binding energy of $0{\cdot}0020$ eV. One would expect the true binding energy to be considerably lower than this, as the light hole has a mass not much greater than that of the electron. This is borne out by a more exact calculation which, we shall see, gives a binding energy of $0{\cdot}0014$ eV for the exciton.

The two indirect exciton levels in which transitions are observed should not be thought of as two different hydrogenic type levels but rather as the splitting of one level of this type. This splitting has been discussed by Zwerdling et al.[59] and by McLean and Loudon.[60] In simple terms it is due to the degenerate nature of the valence band edge and to the fact that the conduction band edge consists of four minima in $\mathbf{k}$-space. The degenerate valence band will lead to degenerate exciton energy levels. This degeneracy is retained in the case of the direct exciton for which the electron is associated with the single conduction band minimum at $\mathbf{k} = 0$. For the indirect exciton, the electron can be associated with any one of four minima and this has the effect of lifting some of the degeneracy, splitting the exciton level into two separate levels. Indirect transitions involving acoustical phonons are allowed into both of these levels[60] and it is such transitions which are seen in the absorption.

The problem of calculating the energies of the various exciton levels is very similar to the calculation of the energy levels of shallow impurity levels. The calculation of these impurity levels, carried out in the effective mass approximation, is fully discussed by Kohn.[61] These calculations have been very successful especially when applied to excited states of the impurities of p-type symmetry. They tend to break down when applied to the s-type ground state for two reasons—the electron orbit is smaller than in the excited states and the wave function is non-zero at the impurity atom, whereas in the excited p-states the wave function vanishes at the impurity. For both reasons the electron in the ground state penetrates much closer to the impurity atom when in the ground state and so spends an appreciable time in the region where the assumed coulomb interaction between the electron and impurity breaks down. For simple hydrogen-type levels, the orbit

radius varies inversely as the binding energy so that one would expect the mean electron–hole separation in an exciton in germanium to be about ten times the orbit size in a corresponding impurity state. For this reason and also the fact that, in the exciton case, there is no distortion of the lattice by an impurity atom causing the coulomb interaction to be modified, one might expect that the effective mass approximation should provide a good estimate of the exciton energy levels. Calculations of the lowest direct and indirect exciton energy levels have been carried out[60] in the effective mass approximation taking into account all the complexities of the valence and conduction band edges and using a variational procedure developed by Schechter[62] for the calculations on shallow acceptor levels in germanium and silicon. These calculations give a value of 0·0014 eV for the binding energy of the direct exciton and the two lowest levels of the indirect exciton binding energies of 0·0035 eV and 0·0029 eV. In each case, the mean electron–hole separation is equal to many lattice spacings so justifying our comments in connection with equation (25). The value for the binding energy of the direct exciton is in fair agreement with the value obtained from our analysis. The situation with the indirect exciton is less satisfactory. From the two experimental values, the lowest level appears to have a binding energy of about 0·0030 eV compared with the calculated value of 0·0035 eV; both experiments give a splitting of the two levels of 0·0010 eV compared with the calculated splitting of 0·0006 eV. A disturbing feature of all these results is that the calculated energies are greater in all cases than the measured values. Any improvement in the calculations by way of slightly better trial functions or a modification of the potential for close distances of approach will not decrease the calculated energies and so will tend to break down what agreement there is at present.

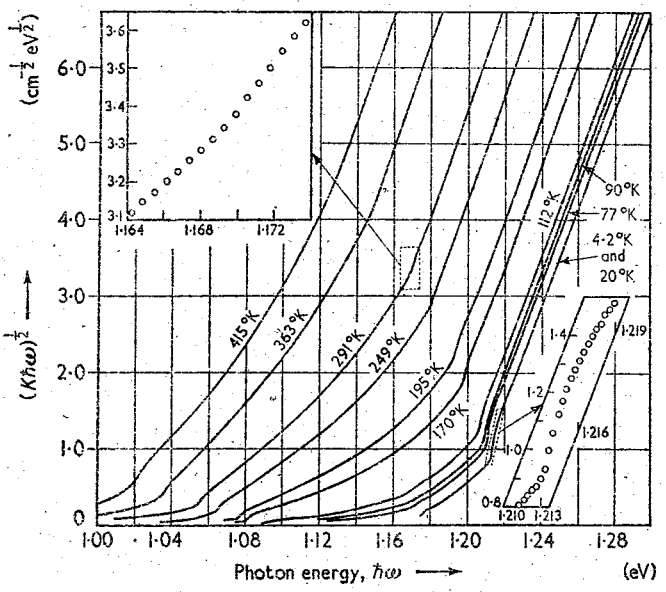
5. THE ABSORPTION EDGE OF SILICON

5.1. The Indirect Transition Region

Figure 9 shows the experimentally determined absorption curves of silicon[3,4] close to the beginning of the absorption edge for a series of temperatures from 4·2° to 415° K. In this case we have chosen to exhibit $\alpha \equiv K\hbar\omega$ as a function of the absorbed radiation energy. An inspection of these curves shows that they have a structure very similar to the corresponding curves for germanium shown in Figure 5. The analysis of this structure has been described previously[4] and we shall not repeat it here but merely state the results. These show that the absorption in this region is due to indirect phonon-assisted transitions, confirming the beliefs which the similarity with the germanium curves in the indirect transition region arouses. In giving these results, we shall use the same notation as we used for germanium.

Apart from the low-lying absorption tails which appear at the higher temperatures, the curves are found to consist in general of four components corresponding to transitions involving the absorption and emission of phonons of two different energies. At the lowest temperatures, only the two components involving the emission of these phonons, of course, contribute. These are clearly seen from the Figure to rise from thresholds at about 1·175 eV and 1·210 eV. The threshold energies ϵ_{1a}, ϵ_{1e}, ϵ_{2a}, and ϵ_{2e} of these four components are given in Table 2 at the various temperatures where measurements were made along with the temperatures

θ_1 and θ_2 of the two types of phonons assisting in these transitions, deduced from the difference in the thresholds corresponding to phonon emission and absorption processes. The close agreement between $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ at each temperature and the constancy of θ_1 and θ_2 as the temperature varies are again convincing evidence of the validity of the analysis of the curves and its interpretation. Furthermore, and consistent with our interpretation, the shapes of the components α_{1a} and α_{1e} are found to be the same at any one temperature; this shape changes slightly with temperature because of a gradual smoothing out of structure as the temperature rises. However, as in germanium, the shape at any temperature can be derived by applying a broadening of Gaussian type and of an appropriate magnitude to a basic



(By courtesy of the Editor, *Physical Review*)

Figure 9. *The absorption edge at low levels in silicon at various temperatures. The inserts indicate the accuracy with which the experimental points define the curves*

shape. The basic shape α_1 associated with the components α_{1a} and α_{1e} and rising from a threshold ϵ_1 is

$$\alpha_1 = a_1[(\hbar\omega-\epsilon_1)^{1/2}+\beta(\hbar\omega-\epsilon_1-0\cdot0055)] \qquad \ldots(80)$$

where $\beta(E)$ is a function rising initially from zero as $E^{1/2}$ and eventually taking up a more or less $\frac{3}{2}$-power dependence on the energy, and we are measuring energy in electron volts. The magnitude of the broadening to be applied to this basic shape to obtain the shape of the components at any temperature is again characterized by a half-width σ which is tabulated in Table 2 at the various temperatures at which measurements were made. The shapes of the components α_{2a} and α_{2e} have the same properties and can be derived from the basic shape

$$\alpha_2 = a_2[(\hbar\omega-\epsilon_2)^{1/2}+0\cdot41\,\beta(\hbar\omega-\epsilon_2-0\cdot0055)] \qquad \ldots(81)$$

Table 2. The Values of the Threshold Energies ϵ_{1a}, ϵ_{2a}, ϵ_{1e}, and ϵ_{2e} Found from the Analysis of the Absorption Curves and the Corresponding Values of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ at Various Temperatures for Silicon. Also Shown are the Phonon Temperatures θ_1 and θ_2 and the Half-width σ of the Gaussian Broadening Suffered by the Exciton Absorption

T (°K)	ϵ_{2a} (eV)	ϵ_{1a} (eV)	ϵ_{1e} (eV)	ϵ_{2e} (eV)	θ_1 (°K) = $\frac{1}{2k}(\epsilon_{1e}-\epsilon_{1a})$	θ_2 (°K) = $\frac{1}{2k}(\epsilon_{2e}-\epsilon_{2a})$	$\bar{\epsilon}_1$ (eV) = $\frac{1}{2}(\epsilon_{1e}+\epsilon_{1a})$	$\bar{\epsilon}_2$ (eV) = $\frac{1}{2}(\epsilon_{2e}+\epsilon_{2a})$	$\sigma \times 10^3$ (eV)
4·2			1·1735	1·2130			1·1552†	1·1553‡	1·2
20			1·1735	1·2130			1·1552†	1·1553‡	1·2
77			1·1710	1·2105			1·1527†	1·1527‡	1·3
90		1·1330	1·1700	1·2095	212		1·1515	1·1517‡	1·5
112	1·0912	1·1307	1·1672	1·2067	212	670	1·1490	1·1490	1·5
170	1·0822	1·1219	1·1583	1·1980	211	672	1·1401	1·1401	1·7
195	1·0772	1·1167	1·1532	1·1927	212	670	1·1350	1·1350	1·7
249	1·0655	1·1050	1·1415	1·1810	212	670	1·1233	1·1232	2·7
291	1·0552	1·0947	1·1312	1·1707	212	670	1·1130	1·1130	3·3
363	1·036			1·151		667		1·0935	4·8
415	1·021			1·137		672		1·079	6·5

† Calculated from $\bar{\epsilon}_1 = \epsilon_{1e} - k\theta_1$ with $\theta_1 = 212°$ K.
‡ Calculated from $\bar{\epsilon}_2 = \epsilon_{2e} - k\theta_2$ with $\theta_2 = 670°$ K.

the magnitude of the broadening required at any temperature being the same as for the components α_{1a} and α_{1e}. The magnitudes of the various components at each temperature taken along with the phonon temperatures of 212° and 670° K in expression (49) yield values for ρ_1 and ρ_2 of essentially unity at each temperature.

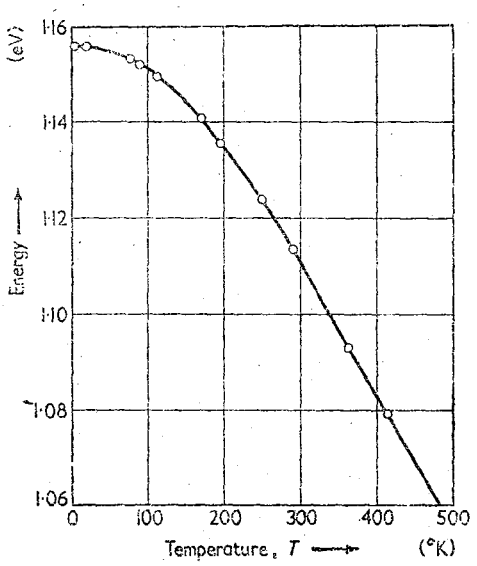
The absorption in the tails appearing at the higher temperatures is difficult to analyse due to the low absorption levels with the consequent errors introduced into the experimental points and the increased amount of smoothing out of detailed structure which occurs at these temperatures. The tails appear to consist of two components rising from thresholds with something like a quadratic energy dependence. The two thresholds lie at energies lower than $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ by equivalent temperatures of 1,050° and 1,420° K. These components were originally[3,4] thought to be due to transitions involving the absorption of phonons of these two temperatures but subsequent determination of the phonon dispersion curves in silicon[63] has shown this interpretation to be incorrect. We shall have more to say about this later.

5.2. Interpretation of the Indirect Absorption Curves—Band Structure, Excitons, and Phonons

The components which account for most of the structure in the absorption curves are due to transitions involving the absorption and emission of phonons with temperatures 212° and 670° K (cf. Table 2) and have essentially the same shapes, given in equations (80) and (81). From our discussions of the theory of indirect transitions, we interpret the two contributions to these components rising from thresholds 0·0055 eV apart as the square root of the energy as being due to allowed indirect transitions into the two lowest lying exciton levels which are separated by 0·0055 eV. The eventual $\frac{3}{2}$ power dependence on energy of the function β in equations (80) and (81) is due to the predominance at higher energies of transitions producing unbound electron–hole pairs. It has proved difficult to analyse the function β to obtain an accurate value for the threshold energy for these transitions. The best result to be obtained is that the threshold lies somewhere around 0·01 eV above the lowest exciton threshold. If we assume that the splitting of 0·0055 eV is between the two lowest hydrogenic levels we find a value of 0·0073 eV for the binding energy. However, this is probably not a good estimate of the binding energy since, as in germanium, the two lowest indirect exciton levels in silicon are formed by a splitting of a degenerate level by the symmetry of the conduction band edge.[60] Assuming a splitting proportionate to that in germanium leads to a binding energy of 0·015 eV which seems larger than that which can be reconciled with the function β. The uncertainty in the accurate value of the exciton binding energy produces a corresponding uncertainty in the value of the indirect energy gap ϵ_g deduced from $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$. Figure 10 shows a plot of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ as a function of temperature. This graph can be regarded as giving the temperature dependence of the indirect energy gap ϵ_g but with all values too low by an energy of the order of 0·01 eV. Calculations of the lowest indirect exciton energy levels in silicon have been carried out using the effective mass approximation[60] and taking into account the complexities of the conduction and valence band edges. They give a binding energy somewhere in the region 0·013–0·014 eV depending on the magnitude of the contribution to the exciton wave function of hole states in third valence band split off from the other two at $\mathbf{k} = 0$ by spin–orbit

coupling. The splitting between the ground state of the exciton and the next level is found to lie somewhere in the range 0·0006 to 0·001 eV, much smaller than the observed splitting. The agreement between the calculated binding energy and that estimated from the absorption curves is fair, but the calculated splitting between the first two levels is much smaller than that which is apparently observed. One wonders if the second level seen in the absorption curves does not correspond to some higher level of the exciton and not to the one whose binding energy was calculated.

A calculation, similar to that described for germanium, can be carried out using the temperature dependence of ϵ_g and the intrinsic carrier concentration values found by Morin and Maita,[64] Herlet,[65] and Putley and Mitchell[66] for temperatures



(By courtesy of the Editor, *Physical Review*)

Figure 10. *The indirect energy gap less the indirect exciton binding energy* ($\sim 0{\cdot}01$ *eV*) *as a function of temperature in silicon*

above 250° K. Again one finds that $N_c(\bar{m}/m)^3$ must increase with temperature when a binding energy $\sim 0{\cdot}01$ eV is taken for the exciton. With $N_c = 6$, which is now a well established value for silicon, the conduction band minima lying inside the Brillouin zone in ⟨100⟩ directions, one finds that $\bar{m}$ increases over its value[39–41] of $\simeq 0{\cdot}45\ m$ at 4·2° K by about 30–40 per cent as the temperature rises to 400° K. This is a substantial increase in effective mass and its effects should be significant in other properties of silicon, e.g. in the mobility as a function of temperature.

Again, as in germanium, the broadening suffered by the absorption components can be interpreted in terms of a relaxation time $\tau = \hbar/\sigma$ for the excitons, since the values of σ quoted in Table 2 were obtained from a study of the smoothing out of that part of the absorption rising as the square root of the energy and corresponding to transitions into exciton states. The temperature dependence of the relaxation time is very similar to that in germanium, being constant up to about 100° K and

then falling with increasing temperature roughly as T^{-2}. However, the actual relaxation times in silicon are about a quarter of the corresponding times in germanium.

The discussion of the phonon spectrum in silicon which was originally given[3,4] has proved to be incorrect in the light of the neutron scattering measurements of Brockhouse.[63] The phonons assisting in the indirect transitions in silicon must have wave vectors lying on a $\langle 100 \rangle$ direction $0 \cdot 85$ of the distance out to the edge of the zone, this being the position of the conduction band minima[67] and the valence band edge being at $\mathbf{k} = 0$. It is clear from Brockhouse's measurements that the two types of phonons, of energies 212° and 670° K, assisting in these transitions are TA and TO respectively, the two energies being in excellent agreement with his values at the appropriate points in $\mathbf{k}$-space. With this assignment, one is forced to interpret the two components forming the tails at the higher temperatures in terms of indirect transitions involving the absorption of two phonons. We have not discussed indirect transitions of this type so far but obviously they can take place providing the total wave vector of the two phonons absorbed equals the difference in wave vector of the valence and conduction band edges. Similar transitions involving the emission of two phonons or the absorption of one phonon and emission of another can take place providing the phonons satisfy the necessary requirements on conservation of wave vector. These different types of two-phonon processes will have different thresholds and temperature dependences. The threshold energies of the two components, of which the tails appear to consist, lie at energies equivalent to temperatures of 1,050° and 1,420° K, below $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$. They are therefore consistent with transitions involving the absorption respectively of two phonons coming from the TA branch somewhere in the outer half of the zone and from an optical branch around the centre of the zone where the phonon energy is 730° K and with two optical phonons, one from around the zone centre and the other from the TO branch somewhere in the outer half of the zone. Evidence for similar two-phonon processes has been found by Clark[8] in the absorption edge of diamond.

Brockhouse's results show that his earlier suppositions[68] are correct about the similarity of the vibrational spectra of germanium and silicon, viz., one can be obtained from the other by a simple scaling factor. Long range forces are therefore required to explain the vibrational spectrum in silicon just as in germanium.

The most surprising feature of the silicon absorption edge is that there appears to be no contributions to the absorption from transitions involving either the LA or LO phonons which at the appropriate wave vector have temperatures of 500° and 610° K respectively.[63] The absence of contributions from these phonons is not consistent with the selection rules governing these transitions. To see this, we note that for the phonons required in the transitions the operator $\mathbf{O}_b(\mathbf{K})$ in equation (38) transforms as the irreducible representations Δ_5, Δ_1, Δ_2', and Δ_5 of the group Δ for phonons of type TA, LA, LO, and TO respectively. Figure 11 shows the band structure of silicon in a $\langle 100 \rangle$ direction and the irreducible representations for which the electron wave functions form bases at the various parts of the zone for the valence and conduction bands and several neighbouring bands which can provide suitable intermediate states for the indirect transitions. The initial and final electron states transform as Γ_{25}' and Δ_1 respectively. Using these transformation properties of the various quantities appearing in the matrix elements

contributing to the X_{cv}^{lb}, one finds that there are no intermediate states through which transitions involving only transverse phonons are allowed. The most important intermediate state, on the grounds of the magnitude of the associated energy denominator, is the conduction band at $\mathbf{k} = 0$; through it transitions involving TA, LA, and TO phonons are allowed. Transitions through other states are allowed for either one of the longitudinal phonons or for both of the transverse phonons and either one or both of the longitudinal ones. There is no selection rule forbidding transitions involving any one particular type of phonon regardless of intermediate state as there is for the optical phonons in germanium. This is

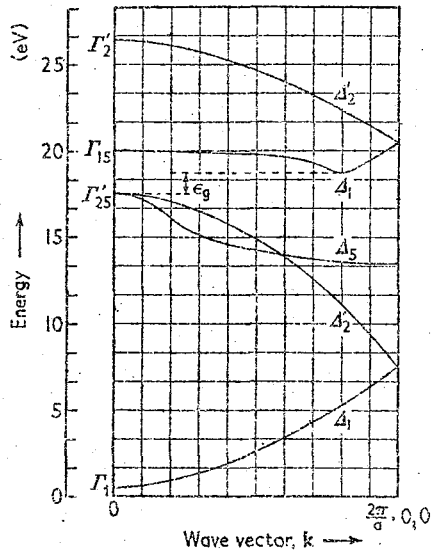


Figure 11. *A diagram of the band structure in a* ⟨100⟩ *direction in silicon as it is currently accepted. The irreducible representations according to which the electron wave functions at various points in* **k***-space transform are shown. Except for the indirect gap* ϵ_g *the energy scale should be considered as only approximate*

essentially due to the lack of inversion symmetry in the electron wave functions and phonon operators associated with a point on a ⟨100⟩ direction within the Brillouin zone. It is, therefore, difficult to understand the lack of any evidence in the absorption curves of transitions involving the longitudinal phonons. Even assuming that the conduction band minima are at the zone edge—a point with the symmetry of the group X—does not resolve the difficulty.

Spin–orbit coupling in silicon is fairly weak. The splitting of the degenerate valence band edge caused by this coupling[69] is only 0·05 eV so that one might expect to see evidence in the absorption edge of transitions involving the excitation of electrons from the lower valence band as well as from the two bands which form the normal valence band edge. The absorption curves were examined with this

very much in mind but no evidence for these transitions from the lower band was found.

5.3. High Absorption Levels

Measurements of the absorption edge in silicon were carried to levels around 5×10^4 cm^{-1} by Dash and Newman.[35] Their measurements extended over a radiation energy range from around 1 eV to 3·2 eV, i.e. from the infra-red through the visible to the edge of the ultra-violet and were made at room temperature and at 77° K. More recently these measurements have been repeated[70] using higher optical resolution than was available to Dash and Newman. The rise of the absorption edge to the higher absorption levels is considerably different in silicon than in germanium, a fact which is clearly shown in Figure 12 where the edges in

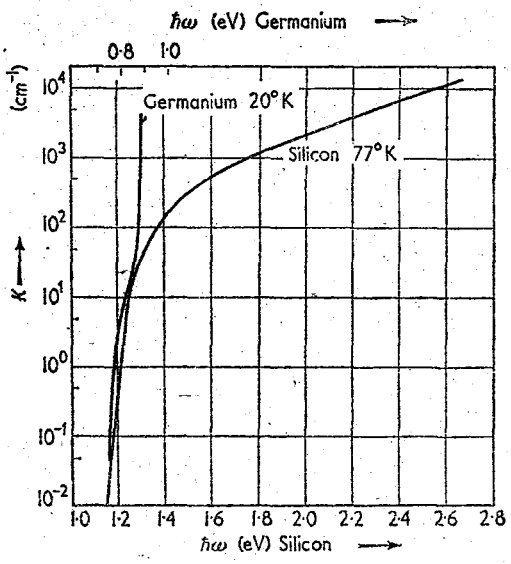


Figure 12. *A comparison of the absorption edges in germanium and silicon. The curve for germanium is based on early measurements and lies slightly low at the highest levels*

germanium at 20° K and silicon at 77° K[70] are shown so that the absorption in the indirect transition regions are superimposed as well as possible. We see that the germanium curve rises abruptly to the direct exciton peak, whereas the silicon curve continues to rise quite smoothly with energy. Empirically, the rise in the silicon curve follows quite accurately a quadratic dependence on the energy above 1·193 eV from levels of a few cm^{-1} up to about 2,000 cm^{-1} when it begins to rise slowly above this dependence. There is certainly no evidence of an abrupt rise in absorption comparable to the rise in the germanium curve at the onset of direct transitions. Dash and Newman's measurements show a small sharp rise in the absorption at 77° K at an absorbed radiation energy of about 2·5 eV and they suggest that this is an indication of the beginning of the direct transition region. However, the more recent measurements using higher resolution show no sign of this rise, as is seen in Figure 12 where the increase in absorption is quite smooth

up to energies of 2·7 eV. These measurements have in fact been extended to 3·0 eV and still show no sign of the onset of direct transitions. The direct energy gap is usually assumed to be at $\mathbf{k} = 0$ in silicon and to have a value of 2·5 eV based on the small rise at this energy in the absorption curve of Dash and Newman at 77° K. The complete absence of this rise in the measurements of Roberts and Quarrington casts some doubt on this value for the direct gap, and in fact from their measurements it would seem that $\epsilon_0 > 3 \cdot 0$ eV. Furthermore, the conduction band energy appears to decrease as $\mathbf{k}$ increases from zero out along a ⟨100⟩ direction, corresponding to a negative effective mass. The magnitude of this mass is not known but the fact that it is negative raises the possibility of the direct energy gap being at some non-zero value of $\mathbf{k}$. The direct gap lying in a region of negative electron effective mass raises interesting possibilities in the study of the direct exciton. A more detailed investigation of this high level absorption region of the silicon edge is required.

6. SUMMARY AND DISCUSSION

We have seen from our discussion of the absorption edges of germanium and silicon how a considerable amount of information about the properties of a semiconductor can be deduced from the shape, structure, and magnitude of the absorption in its absorption edge. This information concerns not only the electronic states close to the conduction and valence band edges and the interaction between electrons and holes in these states due to correlation effects in the motion of the many electrons in the solid, but also the lattice vibrational spectrum of the semiconductor. In our discussion we have not hesitated to point out the difficulties and uncertainties which still remain to be resolved. This is as it should be, but on the other hand it should not be allowed to conceal the considerable success which can be achieved in the analysis and interpretation of the absorption edge.

Although we have chosen to discuss the analysis only of those absorption curves measured in germanium and silicon with high resolution from which energy gap values of high accuracy can be obtained and also values of the energies of the individual phonons assisting in the absorption processes, progress can and has been made by examining the edges in other semiconductors measured in most cases with considerably lower resolution than that used in measuring the curves we have discussed. Clark[5] has examined the edge in diamond. He can account for the structure he observes in terms of indirect transitions in which both excitons with binding energy $\simeq 0 \cdot 1$ eV and unbound electron–hole pairs are produced. Braunstein, Moore, and Herman[71] have carried out an extensive study of the germanium–silicon alloy system measuring the absorption edge in a large number of alloys containing different relative amounts of germanium and silicon. Their measurements and analyses are similar to those in the original work of Macfarlane and Roberts[16, 17] on germanium and silicon where the resolution used was not great enough to show up all the structure in the edge. The indirect transition region appears to consist of only two components in general, each having a quadratic energy dependence; they are associated with transitions in which an 'average' phonon is either emitted or absorbed. A value for the energy gap accurate enough for many purposes can be obtained from this, and also a measure of the magnitude of the vibrational energies to be associated with the lattice. Measurements

and analyses of this type have now been carried out on several semiconductors, some of which were mentioned earlier.[5–7] However, there is much to be gained by studying the absorption edges in these materials with a resolution comparable to that used in germanium and silicon.

Finally, we should mention the work of Haynes, Lax, and Flood[72] in studying the spectral distribution of the radiation emitted from germanium and silicon due to the recombination of electron–hole pairs. This is really the inverse experiment to that of measuring the absorption coefficient and it consequently yields the same type of information. The recombination radiation spectrum consists in general of several pairs of components, each pair corresponding to indirect recombinations involving the absorption and emission of a phonon with appropriate wave vector. Each component in the emission spectrum can be associated with a corresponding one in the absorption and its shape can be obtained from the absorption component by multiplying by the usual factor $\exp(-\hbar\omega/kT)$, appropriate to the energy and temperature range we are considering. The emission components consequently do not continue to rise with increasing energy but rise to a peak and fall again as the energy of the photon emitted increases. The emission spectrum, therefore, has the form of a series of peaks whose relative magnitude changes with temperature. At the lowest temperatures, only peaks associated with transitions involving the emission of phonons are of course observed, but in contrast to the absorption curves these peaks occur at energies lower than the energy gap since in this case a certain amount of the energy available from the recombination is given up to the phonon. The results of the analysis of the emission spectra of germanium and silicon[63,72] are in very good agreement with the corresponding results obtained from the absorption curves.

ACKNOWLEDGEMENTS

The author would like to express his thanks to Dr. G. G. Macfarlane and Mr. V. Roberts for helpful criticisms of the original manuscript.

This paper is published by permission of the Controller, Her Majesty's Stationery Office.

REFERENCES

1. G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts. *Phys. Rev.* **108**, 1377 (1957)
2. G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts. *Proc. phys. Soc. Lond.* **71**, 863 (1958)
3. G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts. *J. Phys. Chem. Solids* **8**, 388 (1959)
4. G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts. *Phys. Rev.* **111**, 1245 (1958)
5. V. Roberts and J. E. Quarrington. *J. Electron.* **1**, 152 (1955)
6. W. W. Scanlon. *J. Phys. Chem. Solids* **8**, 423 (1959)
7. W. J. Choyke and L. Patrick. *Phys. Rev.* **105**, 1721 (1957)
8. C. D. Clark. *J. Phys. Chem. Solids* **8**, 481 (1959)
9. F. Herman. *Phys. Rev.* **88**, 1210 (1952)
10. F. Herman and J. Callaway. *Phys. Rev.* **89**, 518 (1953)
11. G. Dresselhaus, A. F. Kip, and C. Kittel. *Phys. Rev.* **92**, 827 (1953)
12. R. N. Dexter, H. J. Zeiger, and B. Lax. *Phys. Rev.* **95**, 557 (1954)
13. H. Y. Fan, M. L. Shepherd, and W. Spitzer. *Proceedings of the Atlantic City Conference on Photoconductivity* (Wiley, New York, 1954)
14. F. J. Morin and J. P. Maita. *Phys. Rev.* **94**, 1525 (1954)
15. L. H. Hall, J. Bardeen, and F. J. Blatt. *Phys. Rev.* **95**, 559 (1954); *Proceedings of the Atlantic City Conference on Photoconductivity* (Wiley, New York, 1954)

16. G. G. MACFARLANE and V. ROBERTS. *Phys. Rev.* **97**, 1714 (1955)
17. G. G. MACFARLANE and V. ROBERTS. *Phys. Rev.* **98**, 1865 (1955)
18. J. FRENKEL. *Phys. Rev.* **37**, 17, 1276 (1931)
19. E. F. GROSS. *Suppl. Nuovo Cimento* 3, 672 (1956)
20. S. NIKITINE. *Phil. Mag.* Vol. 4, No. 37, 1 (1959)
21. G. H. WANNIER. *Phys. Rev.* 52, 191 (1937)
22. G. DRESSELHAUS. *J. Phys. Chem. Solids* **1**, 14 (1956)
23. R. J. ELLIOTT. *Phys. Rev.* **108**, 1384 (1957)
24. H. HAKEN. *Halbleiterprobleme* 4. To be published (Ed. W. Schottky, Braunschweig, Vieweg); W. KOHN. *J. Phys. Chem. Solids* **8**, 45 (1959); L. M. ROTH and G. W. PRATT. *J. Phys. Chem. Solids* **8**, 47 (1959)
25. L. I. SCHIFF. *Quantum Mechanics* (McGraw-Hill, New York, 1949)
26. J. J. HOPFIELD. *Phys. Rev.* **112**, 1555 (1958)
27. I. C. CHEESEMAN. *Proc. phys. Soc. Lond.* **A65**, 25 (1952)
28. W. A. HARRISON. *Phys. Rev.* **104**, 1281 (1956)
29. R. ROSENBERG and M. LAX. *Phys. Rev.* **112**, 843 (1958)
30. W. P. DUMKE. *Phys. Rev.* **108**, 1419 (1957)
31. J. A. STRATTON. *Electromagnetic Theory*, p. 511 (McGraw-Hill, New York, 1941)
32. V. ROBERTS. *J. sci. Instrum.* **32**, 294 (1955)
33. V. ROBERTS. *J. sci. Instrum.* **29**, 134 (1952)
34. T. S. MOSS and T. D. H. HAWKINS. *Phys. Rev. Letters* **1**, 129 (1958)
35. W. C. DASH and R. NEWMAN. *Phys. Rev.* **99**, 1151 (1955)
36. V. ROBERTS and J. E. QUARRINGTON. Private communication
37. G. G. MACFARLANE, T. P. MCLEAN, J. E. QUARRINGTON, and V. ROBERTS. *Phys. Rev. Letters* **2**, 252 (1959)
38. W. H. KLEINER and L. M. ROTH. *Phys. Rev. Letters* **2**, 334 (1959)
39. B. LAX and J. G. MAVROIDES. *Phys. Rev.* **100**, 1650 (1955)
40. G. DRESSELHAUS, A. F. KIP, and C. KITTEL. *Phys. Rev.* **98**, 368 (1955)
41. R. N. DEXTER, H. J. ZEIGER, and B. LAX. *Phys. Rev.* **104**, 637 (1956)
42. C. ENZ. *Helv. phys. acta* **28**, 158 (1955)
43. E. M. CONWELL. *Phys. Rev.* **99**, 1195 (1955)
44. J. H. CRAWFORD, H. C. SCHWEINLER, and D. K. STEVENS. *Phys. Rev.* **99**, 1330 (1955); *Phys. Rev.* **100**, 1084 (1955)
45. B. N. BROCKHOUSE and P. K. IYENGAR. *Phys. Rev.* **108**, 894 (1957); *Phys. Rev.* **111**, 747 (1958)
46. H. M. J. SMITH. *Phil. Trans.* **241**, 105 (1948)
47. F. HERMAN. *J. Phys. Chem. Solids* **8**, 405 (1959)
48. M. LAX. *Phys. Rev. Letters* **1**, 131 (1958)
49. G. G. MACFARLANE, T. P. MCLEAN, J. E. QUARRINGTON, and V. ROBERTS. To be published
50. C. HERRING and E. VOGT. *Phys. Rev.* **101**, 944 (1956)
51. S. ZWERDLING, B. LAX, and L. M. ROTH. *Phys. Rev.* **108**, 1402 (1957)
52. W. PAUL and D. M. WARSCHAUER. *J. Phys. Chem. Solids* **5**, 89 (1958)
53. L. M. ROTH, B. LAX, and S. ZWERDLING. *Phys. Rev.* **114**, 90 (1959)
54. E. O. KANE. *J. Phys. Chem. Solids* **8**, 38 (1959)
55. J. C. PHILLIPS. *J. Phys. Chem. Solids* **7**, 52 (1958)
56. S. ZWERDLING, L. M. ROTH, and B. LAX. *J. Phys. Chem. Solids* **8**, 397 (1959)
57. K. J. BUTTON, L. M. ROTH, W. H. KLEINER, S. ZWERDLING, and B. LAX. *Phys. Rev. Letters* **2**, 161 (1959)
58. B. LAX and S. ZWERDLING. Accompanying article
59. S. ZWERDLING, B. LAX, L. M. ROTH, and K. J. BUTTON. *Phys. Rev.* **114**, 80 (1959)
60. T. P. MCLEAN and R. LOUDON. *J. Phys. Chem. Solids*. To be published
61. W. KOHN. *Solid State Physics* **5**, 257 (1957)
62. D. SCHECHTER. *Thesis*, Carnegie Institute of Technology (1958); *See also* W. KOHN and D. SCHECHTER. *Phys. Rev.* **99**, 1903 (1955)
63. B. N. BROCKHOUSE. *Phys. Rev. Letters* **2**, 256 (1959)
64. F. J. MORIN and J. P. MAITA. *Phys. Rev.* **96**, 28 (1954)
65. A. HERLET. *Z. angew. Phys.* **9**, 155 (1957)
66. E. H. PUTLEY and W. H. MITCHELL. *Proc. phys. Soc. Lond.* **A72**, 193 (1958)
67. G. FEHER. *J. Phys. Chem. Solids* **8**, 486 (1959)
68. B. N. BROCKHOUSE. *J. Phys. Chem. Solids* **8**, 400 (1959)
69. L. HULDT and T. STAFLIN. *Phys. Rev. Letters* **1**, 313 (1958)
70. V. ROBERTS and J. E. QUARRINGTON. Private communication
71. R. BRAUNSTEIN, A. R. MOORE, and F. HERMAN. *Phys. Rev.* **109**, 695 (1958)
72. J. R. HAYNES, M. LAX, and W. F. FLOOD. *J. Phys. Chem. Solids* **8**, 392 (1959)

THE CHEMICAL BOND IN SEMICONDUCTORS

E. MOOSER, Dr. Sc. Nat., Dipl. Phys. E.T.H.

and

W. B. PEARSON, D.F.C., M.A., D.Phil.

Division of Pure Physics, National Research Council, Ottawa, Canada

MS. received October 1959

THE CHEMICAL BOND IN SEMICONDUCTORS

1. INTRODUCTION

Cohen and Heine[1] in the introduction of a paper on the electronic band structure of monovalent metals make the following observations: 'Concerning the theoretical side, a band structure calculation is lengthy and subject to technical difficulties and considerable errors. Moreover, it has often focused attention on a single metal when what is needed is an understanding of systematic variations from metal to metal'. If in this quotation we replace the word 'metal' by 'semiconductor' we have admirably expressed the reasons for studying the chemical bonding in semiconductors since it is by this means that we can gain information on the band structure of semiconductors without actually carrying out band calculations. In addition to avoiding involved and cumbersome computations, a discussion of the chemical bonding permits a qualitative appraisal of many semiconducting materials simultaneously and thus allows the establishment of much needed rules governing systematic trends in the properties of semiconductors.

The chemical approach to semiconductivity is not new. As early as 1925, at a time when no band calculations were available, Friederich[2] discussed electronic conduction in solid inorganic compounds and concluded that high electrical conductivity occurs in compounds in which not all of the valence electrons are involved in the bonding. Friederich also noted the absence of an appreciable electronic conduction in molecular crystals. Nineteen years later, when Meyer[3] reviewed problems connected with electronic conduction in solid compounds, Friederich's basic question, namely, what are the necessary and sufficient conditions for the occurrence of electronic conduction in compounds, was still not rigorously answered. However, by this time a wealth of experimental data was available and Meyer, systematizing these data, was able to draw the following conclusions:

(1) The value of the electrical conductivity of a solid compound is related to the electron configurations of the component atoms.
(2) Semiconductors and insulators are normal valence compounds. Compounds which do not adhere to the valence rules and compounds with compositions deviating from stoichiometry are good (metallic) conductors.

Meyer's findings could very well have served as a guide in a search for new semiconducting materials. However, it so happened that by the end of World War II the attention of scientists was drawn to the two Group IV elements silicon and germanium, whose technology is relatively simple. Extensive programmes for their experimental and theoretical investigation were undertaken and the pursuit of these programmes proved very fruitful indeed, being crowned in 1948 by the discovery of the transistor.

The boom of solid state devices which followed the discovery of the transistor created an urgent need for new semiconducting materials suitable for a variety of different applications and it was only natural that at this stage the chemical

approach to semiconductivity was remembered. Thus, because of their chemical and structural similarity to the Group IVB elements (silicon, germanium, α-tin), Welker[4] in 1952 predicted and discovered the semiconducting properties of the $A^{III}X^{V}$ compounds. His speculations on the properties of these compounds indicate the potential power of the chemical approach. According to Welker the bonding in the elements silicon, germanium, and grey tin is purely covalent, but in the corresponding isoelectronic compounds (aluminium phosphide, gallium arsenide, indium antimonide) it has an ionic as well as a covalent component. Resonance between the covalent and ionic valence structures leads to an increase of the bond strength and Welker, therefore, concluded that the band gaps of the compounds must be larger than those of the corresponding elements. Moreover since, owing to the higher bond strength, the amplitudes of thermal vibration of the atoms in the compounds are smaller than in the elements, Welker predicted higher charge carrier mobilities for the compounds. Both these predictions were subsequently verified by experiment.

With his work Welker led the way for the development of a new chemical approach to semiconductivity. Indeed, his rules point to a programme for such an approach:

(1) To predict semiconductivity of a material from its chemical composition and (possibly) from its structure. More generally even, to predict whether a system of two or more chemical elements forms any semiconducting phases.
(2) To give an indication of the values of the most important parameters determining the properties of a semiconductor (e.g. energy gap, charge carrier mobility, etc.).

Present-day electron theories of solids are unable to cope successfully with such a programme. In fact, band calculations have frequently misled experimentalists to conclude that the periodicity of the crystal structure is a physical necessity for the occurrence of semiconductivity. Obviously this is not true. The semiconducting properties of amphorous and liquid selenium have long been known, but it was Joffé[5] who first pointed out that the short range order (i.e. the bonding) rather than the long range order (i.e. the structure) is responsible for semiconductivity. The importance of the bonding in semiconductors was stressed also by Krebs and Schottky[6] and a synthesis of their work and Pauling's valence bond theory of metals[7] as well as a survey of all known semiconductors led Mooser and Pearson[8, 9] to postulate a specific semiconducting bond. The concept of the semiconducting bond affords a chemical criterion equivalent to Wilson's[10] physical criterion of filled bands and thus permits a clear distinction to be made between metals and semiconductors. The semiconducting bond and its application to the prediction of semiconductivity will be discussed in Section 2 below. One of the immediate consequences of the bond treatment of semiconductors was the discovery of a classification of the structures of a large family of semiconductors. Indeed a crystal chemistry of semiconductors has since been developed, and this will form the topic of Section 3. Section 4 is devoted to the surface structure of semiconductors and, finally, in Section 5, empirical rules are discussed relating the energy gaps and charge carrier mobilities of semiconductors to their chemical composition.

2. THE SEMICONDUCTING BOND AND ITS APPLICATION TO THE PREDICTION OF SEMICONDUCTIVITY

In his discovery of the semiconducting properties of the $A^{III}X^{V}$ compounds, Welker[4] was guided by their chemical and structural similarity to the Group IVB elements. Subsequently, similarities of this type led to the prediction of a number of other semiconducting compounds. Thus Goodman and Douglas[11] in 1954 predicted and found semiconducting properties in the chalcopyrites of composition $A^{I}B^{III}X_2^{VI}$ which crystallize in a tetragonally distorted superlattice of the zinc blende structure (see also Glazov, Mirgalovskaya, and Petrakova[12]). On the basis of the work of Welker and of Goodman and Douglas, Folberth and Pfister[13] later predicted the existence, the structure, and the semiconducting properties of a series of compounds of composition $A^{II}B^{IV}X_2^{V}$ (see also Austin, Goodman, and Pengelly;[14] Goodman[15]). Independently of Welker, and at about the same time, Russian scientists also noted the chemical and structural similarities between the $A^{III}X^{V}$ compounds and the Group IVB elements. Their work has been summarized by Nasledov.[16] If, finally, we mention the compound indium telluride, which has a defect zinc blende type structure and whose semiconducting properties were found by Appel,[17] we have a fairly complete list of the materials whose semiconducting properties were predicted or suspected from their chemical and structural similarities to the Group IVB elements.

2.1. The Valence Bond Description

A number of qualitative criteria which are very useful as a guide in the search for new semiconducting materials has been discussed by Goodman.[18] However, in order to develop more general means for predicting semiconductivity, it was desirable to find a criterion which applies to all semiconductors and gives sufficient and necessary conditions for the occurrence of semiconductivity. Such conditions were given by Mooser and Pearson[8,9] in their discussion of the bonding in semiconductors.

Earlier, Pauling[7] had shown that the bonds in metals are essentially covalent. To account for the metallic conduction and to explain the high co-ordination numbers found in metals, Pauling introduced the concept of uninhibited resonance and showed that this resonance is possible only if some of the orbitals—the metallic orbitals—in the valence shells of the atoms remain empty. The metallic state, therefore, is characterized by covalent bonds which lead to partial filling of the valence orbitals of the atoms. Since the bonds in semiconductors are also predominantly covalent† (cf. Krebs and Schottky,[6] Krebs,[19] Tsidil'kovskii,[20] Garlick, Hough, and Fatehalley[21]), it was to be expected that the correspondence between

† By *predominantly covalent* bonds we understand electron pair bonds that lead to but *small* effective charges on the component atoms of a solid. This definition conflicts with the definition of purely covalent bonding which has recently been applied in estimating the percentage of ionic character of the bonds in compounds (cf. Jeffrey, Parry, and Mozzi;[22] Wolff and Broder[23]). According to this latter definition the bonds in cadmium telluride, for example, have 100 per cent covalent character if two electrons are transferred from each tellurium atom to a neighbouring cadmium atom, so that $Cd^{2-}Te^{2+}$ describes the covalent valence structure of this compound. Since $Cd^{2+}Te^{2-}$ on the other hand describes the ionic valence structure, we see that *uncharged* constituents would correspond to *50 per cent ionic character* of the bonds. In our opinion this is a rather unrealistic picture and we therefore do not adopt it here.

the partially filled bands and the partially filled atomic orbitals, as implied by Pauling's treatment of metals, should be paralleled by an analogous correspondence in semiconductors. A survey of the known semiconductors showed that this is indeed the case and it was, therefore, possible to define a specific semiconducting bond as follows: The bonds in semiconductors are predominantly covalent and they form a network which runs continuously throughout the whole crystal structure. By sharing electrons in forming the bonds, all atoms in elemental semiconductors acquire filled valence subshells. In semiconducting compounds only the most electronegative atoms (i.e. the 'anions') need acquire filled subshells. If empty orbitals occur on the electropositive atoms (the 'cations') then any bonds that might form between these atoms alone must not run continuously through the crystal.

Thus, the difference between the two types of electronic conductors—metals and semiconductors—can essentially be attributed to a difference in the degree of filling of the valence shells of the component atoms. The difference between semiconductors and insulators is less clear cut. Roughly speaking, one can distinguish between two classes of insulators, namely ionic crystals and molecular crystals. Representatives of both these classes may, under certain conditions, show electronic conduction. However, under thermal excitation this conduction is negligibly small so that the above definition of the semiconducting bond, which excludes solids with predominantly ionic bonds and molecular solids, is permissible.

The existence of a correspondence between the filling of atomic orbitals and the filling of bands, which is implied in our definition of the semiconducting bond, can readily be demonstrated by a simple molecular orbital treatment (see sub-section 2.2) which also emphasizes the importance of the *short range order* in semiconductors. Provided that the bonding in a solid is directional, and this seems to be the case in all semiconductors, the short range order allows the deduction of the likely valence states of the component atoms, and hence the possible valence structures of the solid can be determined (see also sub-section 3.2). *If all these valence structures lead to filled subshells on the anions then according to the concept of the semiconducting bond the solid is a semiconductor.* In estimating the relative importance of the different valence structures help can be obtained from the interatomic distances (relative bond strength) and the electronegativity difference (relative importance of ionic and covalent structures). Sometimes, in compounds containing transition elements and rare earths, a third factor, the magnetic properties, can be used. Moreover, it should be remembered that resonance can only occur among valence structures which have the same number of unpaired electrons (cf. Pauling[24]).

The general character of the conditions for semiconductivity as set up in the semiconductor bond concept is readily recognizable from the fact that all other criteria used for the prediction of semiconductivity can be derived from it. Thus, the criterion of chemical and structural similarity which we discussed at the beginning of this section obviously derives from the similarity of the bonds in these materials. Meyer's statement (see Introduction) that all semiconductors are normal valence compounds is closely related to the filling of subshells in the anions. This statement has been expressed in different forms. Goodman[18] shows that an ionic formula can be assigned to each semiconductor. Mooser and Pearson[8, 9] give

a formula governing the number of valence electrons in semiconductors in which the valence s and p orbitals on the anions are filled (see also Busch[25]). Making allowance for the possible formation of cation–cation bonds and in this way generalizing the original formula we can write:

$$\frac{n_e}{n_a}+b_a-b_c=8 \qquad \ldots(1)$$

where n_e is the number per formula unit of valence electrons not counting any unshared electrons on the cations, n_a is the number of anions per formula unit, b_a is the average number of anion–anion bonds formed by each anion, and b_c is the average number of cation–cation bonds formed by each cation. This formula has recently been discussed in some detail by Suchet.[73]

As illustrations let us consider the compounds silver antimony sulphide ($AgSbS_2$),[26] cadmium antimonide (CdSb),[27] and gallium telluride (GaTe).[28] In the first of these compounds, the antimony atoms act as cations having two unshared electrons. In the second, each anion forms a bond with a neighbouring anion, and, in the third, cation–cation bonds occur. The corresponding n_e, n_a, b_a, and b_c

Table 1

	n_e	n_a	b_a	b_c	$(n_e/n_a)+b_a-b_c$
$AgSbS_2$	$1+3+2\times6$	2	0	0	8
CdSb	$2+5$	1	1	0	8
GaTe	$3+6$	1	0	1	8

values are listed in Table 1 and we see that the three compounds do indeed adhere to equation (1).

It should be pointed out here that the majority of known semiconductors are characterized by filled s and p orbitals. They all contain Group IVB to VIIB anions—and this is another criterion used for predicting semiconductivity—because only these anions can fill their octets by electron sharing. The original definition[6,9] of the semiconducting bond, therefore, called for filled s and p orbitals. The present less-restrictive version accounts for the existence of some semiconducting hydrides in which the hydrogen atoms complete their 1s orbitals and it also allows for anomalies such as are observed in boron and some of its compounds. It seems that in these latter materials the filling of hybrid subshells containing only part of the valence p orbitals (e.g. sp^2 hybrids) is sufficient to give rise to semiconductivity. Moreover, the semiconducting properties of caesium auride (CsAu) reported by Spicer and Sommer[29] would suggest that in gold compounds the filling of the 6s orbitals on the gold atoms can also produce semiconductivity (see also Goodman[18]). A list of (partly hypothetical) semiconducting compounds in which only the s shells of the anions are completed has recently been compiled by Hulliger.[74]

It is of interest to note that the ionic formula criterion (Goodman[18]) does not set up sufficient conditions for the occurrence of semiconductivity. The reason for

this is that ionic formulae which can, for example, be attributed to metallic compounds such as indium bismuthide (InBi) and copper magnesium antimonide (CuMgSb)[30] need not correspond to an actual valence structure of these compounds. *Unless the valence structures described by the ionic formulae are warranted by the short range order, semiconducting properties do not result.* The importance of the short range order in assigning ionic formulae to compounds is clearly brought out in compounds such as iron disulphide and cadmium antimonide, which at first sight seem not to adhere to any valence rules. Inspection of the short range order shows that in these compounds ionic formulae of the form $Fe^{2+}(S\text{–}S)^{2-}$ and $2(Cd^{2+})(Sb\text{–}Sb)^{4-}$ are adequate (Goodman[18]). We have already seen that in equation (1) this situation is taken care of by the terms b_a and b_c which clearly are determined by the short range order.

2.2. A Simple Molecular Orbital Description

The postulated correspondence between the filling of bands and the filling of atomic orbitals is possible only if there is a close correspondence between the bands in a semiconductor and the energy levels of the component atoms. A simple molecular orbital treatment shows how this correspondence is brought about by the particular short range order in semiconductors and thus affords a theoretical justification for the semiconductor bond concept.

Before discussing solids let us briefly review the results obtained from a molecular orbital treatment of diatomic molecules (see, for example, Coulson[31]). Let ψ_A and ψ_B be two orbitals at the atoms A and B respectively and let E_A and E_B be the corresponding energy levels. We form the linear combination

$$\psi = \psi_A + \lambda\psi_B$$

in which the parameter λ is chosen so as to minimize the integral

$$\frac{\int \psi^* \mathscr{H} \psi \, d\tau}{\int \psi^* \psi \, d\tau}$$

where $\mathscr{H}$ is the Hamiltonian of the molecule and $d\tau$ is the volume element. This linear combination represents a good approximation for one of the orbitals which are found in the molecule AB if ψ_A and ψ_B fulfil the following requirements:

(1) They must show a considerable overlap;
(2) Their symmetry with respect to the molecular axis must be the same; and
(3) Their energies E_A and E_B must be nearly equal.

If in particular the atoms A and B are identical and if, moreover, the atomic orbitals ψ_A and ψ_B are of the same type (i.e. if they are both either 1*s*, or 2*s*, or 2*p*, or ... so that $E_A = E_B$) then by symmetry $|\lambda| = 1$, and on forming the molecule AB we find that the two orbitals ψ_A and ψ_B now give rise to two molecular orbitals $\psi_A + \psi_B$ and $\psi_A - \psi_B$. The energies E_+ and E_- associated with the orbitals $\psi_A \pm \psi_B$ differ by an amount which is essentially given by

$$\mathscr{H}_{AB} = \int \psi_A^* \mathscr{H} \psi_B \, d\tau$$

In all cases of interest to us $\mathscr{H}_{AB} < 0$ and hence $E_+ < E_-$. If, therefore, ψ_A and ψ_B were originally occupied by one electron each, then the molecular orbital $\psi_A + \psi_B$ which can accommodate two electrons will be completely filled.

It can be shown that the occupation by two electrons of the orbital $\psi_A + \psi_B$ results in an attractive force which holds the atoms A and B together. This electron configuration is therefore said to constitute a covalent bond and the orbital $\psi_A + \psi_B$ is called a bonding orbital. Occupation of the higher lying orbital $\psi_A - \psi_B$ by two electrons, on the other hand, produces a repulsive force between A and B and this orbital, therefore, is referred to as antibonding.

These results can readily be generalized for heteronuclear molecules. One only has to remember that in this case $|\lambda| \neq 1$ and that $E_A \neq E_B$. Since, depending on whether $|\lambda| < 1$ or $|\lambda| > 1$, the electrons in the orbital $\psi = \psi_A + \lambda\psi_B$ are more concentrated on atom A or atom B, respectively, it follows that λ is a measure for the heteropolarity of the bond formed between A and B. Unfortunately, λ cannot as a rule be calculated from first principles so that one has to rely on empirical data to determine the ionicity of the bonding. One of the most convenient means of estimating bond ionicities is the electronegativity scale and we will make extensive use of this means in Sections 3 and 5.

In a homonuclear molecule, corresponding bonding and antibonding orbitals both correlate with the same atomic energy level $E_A = E_B$. In a heteronuclear molecule we find that if $E_A < E_B$ then the bonding level correlates with E_A while the antibonding level correlates with E_B. As a consequence the energy difference between two corresponding bonding and antibonding levels is normally higher in a heteronuclear molecule than in the isoelectronic homonuclear molecule, and we will see that this is of importance for the size of the energy gaps in semiconductors.

The reason for discussing in some detail the molecular orbital description of diatomic molecules is that, in applying it to solids, we can as a first approximation form linear combinations with the atomic orbitals $\psi_{1A}, \psi_{2A}, \ldots$ of a central atom A and with those $\psi_B, \psi_C, \ldots$ of its nearest neighbours B, C, ..., and neglect all interactions between the orbitals $\psi_1 = \psi_{1A} + \lambda_1\psi_B$, $\psi_2 = \psi_{2A} + \lambda_2\psi_C, \ldots$ (see, for example, Hund;[32] Krebs and Schottky[6]). The results found above are then still valid since under these circumstances the electrons in the solid occur in pairs, each pair being completely isolated from the other. This approximation corresponds to a complete localization of the bonds.

So far we have assumed that on forming a molecule the s and p orbitals lead to different molecular orbitals. This assumption is not necessarily justified in solids. Thus, the tetrahedral neighbourhood of the carbon atoms in diamond, for example, suggests that on forming diamond one can regard each atom as being first excited into a (virtual) valence state in which the s and p orbitals are hybridized so that four equivalent sp^3 orbitals result. Using these orbitals to form linear combinations $\psi_1 = \psi_{1A} + \lambda_1\psi_B$, $\psi_2 = \psi_{2A} + \lambda_2\psi_C$, ... ($|\lambda_1| = |\lambda_2| = \ldots = 1$) and neglecting interactions between orbitals ψ_1, ψ_2, etc., one is led to the following energy scheme for diamond (this description of diamond has been discussed in more detail by Krebs and Schottky[6]): The electrons of the K shell of each carbon atom are internally paired and the K shell, therefore, is completely filled. The s and p orbitals of the L shell are hybridized and they produce two energy levels, one bonding, one antibonding. Because each atom has four sp^3 orbitals, each of these two levels is $4N$-fold degenerate, N being the number of atoms in the

crystal. The eight electrons involved in the four bonds extending from each atom just suffice to fill the bonding levels completely and the antibonding levels remain empty. If now we allow for interaction between the orbitals $\psi_1, \psi_2, \ldots$ both bonding and antibonding levels will broaden into bands. However, since at the observed interatomic distance these interactions are smaller than those between corresponding atomic orbitals ψ_{1A} and ψ_B, ψ_{2A} and ψ_C, etc., the band widths are smaller than the energy difference between the original bonding and antibonding levels. The final result, therefore, is a completely filled valence band which is separated from the next higher completely empty conduction band by a finite energy gap. Thus, we can not only see how in diamond the energy bands correlate with the atomic energy levels, but also that the completion of the octets through the formation of electron pair bonds, as postulated by the valence bond treatment, corresponds in the band model to a semiconductor band structure.

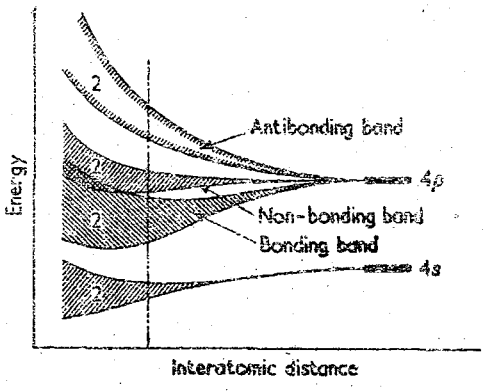


Figure 1. Band structure of selenium (schematic). The numbers indicate how many states per atom each band contains

Let us next look at hexagonal selenium. From its short range order—each selenium atom is joined to its two neighbours by two bonds which are approximately at right angles to each other—we deduce that the atomic orbitals give rise to the bands shown in Figure 1. No sp^3 hybridization is observed here, the 4*s* electrons of each atom being paired internally. However, we meet a non-bonding band which arises from that *p* orbital of each atom which is perpendicular to the two bonds formed by the atom. Since the interaction between two of these *p* orbitals on neighbouring atoms is small we deduce that the non-bonding band is narrow and that on the energy scale it is not shifted away appreciably from the position of the original atomic *p* levels. The band sequence indicated in Figure 1 and the separation between the filled non-bonding and the empty antibonding bands are, therefore, warranted (see Gaspar;[93] also Reitz[94]) and we thus again arrive at a semiconductor band structure.

As a third example we consider the compound semiconductor magnesium stannide (Mg_2Sn). In this substance, the tin atoms are surrounded by eight magnesium atoms sitting on the corners of a cube and the magnesium atoms are tetrahedrally co-ordinated by four tin atoms. These co-ordination configurations

can be interpreted as arising from a resonance among two valence structures which are both characterized by sp^3 hybrids on the tin and magnesium atoms. In the first valence structure, the four bonds formed by each tin atom point towards four tetrahedrally disposed corners of the co-ordination cube of each tin atom, and the two bonds going out from each magnesium atom point towards two corners of the co-ordination tetrahedron of each magnesium atom. In the second equivalent valence structure the bonds are directed towards the other corners of the co-ordination polyhedra. These valence structures are compatible with the band splitting outlined in Figure 2. Since the anion, tin, is more electronegative than the cation, magnesium, the sp^3 levels of the tin atoms lie lower than those of the magnesium atoms and the bonding band, therefore, correlates with the tin levels while the antibonding band correlates with the magnesium levels. This example shows why, for the occurrence of semiconducting properties in compounds, the bond picture only requires filled subshells on the anions: *the valence band arises from the*

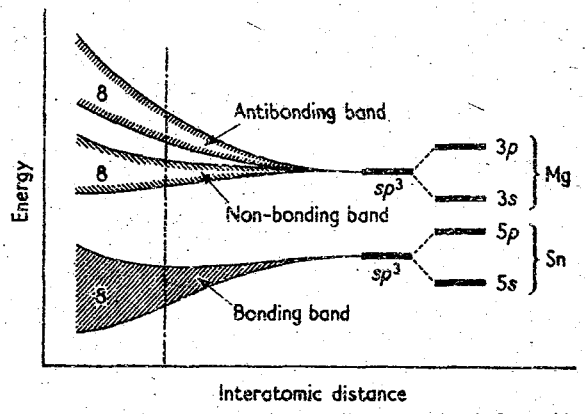


Figure 2. *Band structure of magnesium stannide (schematic). The numbers indicate how many states per formula unit each band contains*

anion orbitals and if they are filled the valence band will also be filled. As in selenium, we find a non-bonding band which here correlates with the two empty magnesium orbitals and therefore remains unoccupied.

It should be emphasized that in establishing qualitative band structures for diamond, selenium, and magnesium stannide we have made use only of the short range order of these materials. Since the above treatment is readily applicable to other semiconductors as well, it follows that *the main qualitative features of the band structures of many if not all semiconductors are determined by the short range order* and this is the reason why it is at all possible to define a specific semiconducting bond.

The molecular orbital treatment in the form used here is a very crude approximation of the actual situation, and in determining the exact band structure of a solid (band shapes, state densities, energy gaps, etc.) the knowledge of the full crystal symmetry is indispensable. If, however, one is interested only in predicting semiconductivity, the above treatment as well as the bond treatment are entirely sufficient and their application to the discussion of the occurrence of semiconductivity in solids is certainly more economical than full band calculations.

The band structures derived above show semiconductor characteristics only if the bands arising from the atomic orbitals do not overlap. An overlap between the bonding and antibonding bands in diamond for instance would result in metallic properties. The bonds would no longer be saturated; they would lose their directional properties and as a consequence the tetrahedral co-ordination of the atoms would change in favour of a short range order of the type found in non-directionally bonded phases (see also sub-section 3.2). Such a change is, indeed, observed in the Group IVB elements on going from germanium (diamond structure) to lead (close packed cubic structure). Tin, lying between germanium and lead, marks the border between directionally and non-directionally bonded phases: at low temperatures the diamond structure is stable but at high temperatures the higher co-ordinated structure of metallic (white) tin is favoured. It is of interest to show how the molecular orbital treatment accounts for these changes. As the principal quantum number n of the valence shell increases from $n = 2$ in diamond to $n = 6$ in lead, increasing parts of the valence orbitals fall within the

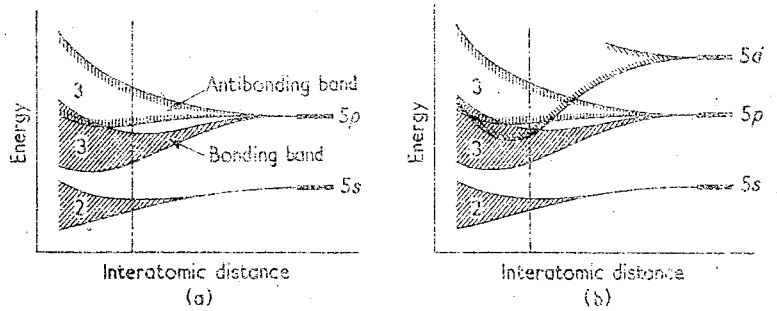


Figure 3. *(a) Band structure of amorphous antimony (schematic); (b) Band structure of crystalline antimony (schematic)*

atomic cores. The overlap of the valence orbitals of neighbouring atoms, therefore, decreases and so does the energy difference between bonding and antibonding levels. At the same time the interaction between different bonds going out from the same atom becomes more and more important so that the widening of the levels into bands, which results from this interaction, becomes more pronounced the larger the value of n. Both of these effects act to reduce the gap between bonding and antibonding bands with increasing n until finally at $n \geq 5$ the bands overlap, producing the changes discussed above.

Since most semiconductors are characterized by completed *octets* on the anions, we deduce that the separation on the energy scale of the atomic s and p orbitals from any higher orbitals is normally large enough to prevent the bands corresponding to these higher orbitals from overlapping with the valence band. There is, however, evidence of such an overlap in some solids containing Group VB to VIIB atoms only, and according to the above discussion this overlap should be reflected in the short range order of these materials. Antimony may serve as an example. In its structure, double layers are formed in which each atom acquires three nearest neighbours. Figure 3(a), therefore, is an adequate illustration of the splitting into bands of the s and p orbitals. However, the distance between

neighbouring double layers is considerably shorter than would be expected if these layers were held together by van der Waals' bonds only, so that in addition to its three nearest neighbours, each atom has three more neighbours at an only slightly longer distance. According to Mooser and Pearson[35,36] this co-ordination configuration is brought about by an appreciable interaction between the *d* orbitals of the antimony atoms in neighbouring double layers, the corresponding band being wide enough to overlap the bonding *p* band. Taking the *d* levels into account we therefore obtain the band structure sketched in Figure 3 (b) which explains the semi-metallic properties of crystalline antimony.† Nevertheless, the semiconductor band structure of Figure 3(a) is realized in nature: amorphous antimony consists of small flakes of the double layers met in the crystalline modification. These flakes are randomly arranged with respect to one another so that no interaction between the *d* orbitals occurs. In deriving the band structure of amorphous antimony one can therefore neglect the *d* levels, being led at once to the situation outlined in Figure 3(a). Amorphous antimony is, indeed, a semiconductor. We should add that in the discussion of selenium given earlier we have neglected the interaction between neighbouring selenium chains. The short interchain distances suggest that this neglect is not warranted and Figure 1 should therefore be completed by taking into account the *d* bands. Many of the rather anomalous properties of hexagonal selenium can then be interpreted in terms of a small band overlap.[35,36] As it stands, Figure 1 represents the band structure of amorphous and liquid selenium.

Finally some comment should be made with respect to the band structure of magnesium stannide. The short range order in this compound leaves little doubt about the correctness of the band scheme outlined in Figure 2. However, a tetrahedral co-ordination of the tin atoms in conjunction with a digonal co-ordination of the magnesium atoms would give rise to a similar band splitting. Thus it follows that the relationship between the short range order and the splitting of the atomic levels into bands is not necessarily unequivocal. However, this ambiguity is at least partially removed by the following postulate (Pauling[7]): If, in some atoms of a solid, empty orbitals are found (e.g. the empty sp^3 orbitals which in magnesium stannide constitute the non-bonding band) then *pivotal resonance* of the bonds can occur and the co-ordination numbers of the atoms in the solid can exceed their valencies. Unfortunately it seems impossible to prove this postulate within the

Table 2

Metallic properties	Semiconducting properties
Uninhibited resonance completely delocalized bonds	*No resonance* completely localized bonds *Synchronous and pivotal resonance* partly delocalized bonds

† The degeneracy of the atomic *d* levels is actually removed in the crystalline field so that, like the *p* levels, the *d* levels give rise to a series of different bands. For our qualitative discussion it is, however, sufficient to consider only the lowest of these bands.

framework of a one-electron treatment (see also Krebs and Schottky[6]) although it is well supported by a wealth of empirical data.

In this connection we should mention that a criterion for semiconductivity can be given in terms of the resonance forms which occur in solids (see Choquard, Janner and Ascher[37]). We have already seen that *uninhibited* resonance produces metallic properties and that *pivotal* resonance such as is observed in magnesium stannide does not destroy semiconducting properties. A third form of resonance, namely *synchronous* resonance (Pauling[7]), is also permitted in semiconductors. These findings are summarized in Table 2.

3. CRYSTAL CHEMISTRY OF SEMICONDUCTORS

If now we proceed to study the crystal chemistry of semiconductors we do so for the following reasons. First, there exists a large family of semiconductors whose structures are closely related to one another and it is possible to develop a classification of these structures which further illuminates the role in semiconductors of short range order and stoichiometric composition. Secondly, since the short range order gives reliable information on the valence structures of solids only if the bonds are directional, it is of importance to investigate the dependence of the directional properties of the bonds upon the chemical composition of valence compounds.

3.1. Structures Based on Close-packed Arrays of Anions

The semiconducting structures which we are to classify here are all characterized by close-packed anion sub-lattices, the cations being located in the interstices of these anion sub-lattices. Let us therefore first look into the geometrical properties of close-packed structures. One can think of any close-packed three dimensional array of equal spheres as being built up of close-packed plane layers. Let A be the set of normals to such a layer drawn through the centres of its spheres. There are two different ways in which a second identical layer can be brought into contact with the first one such that each sphere of the second layer touches three spheres of the first and vice versa. Let B and C be the corresponding sets of normals. If to arrive at a three dimensional close-packing more and more layers are stacked on top of each other, one can characterize the stacking by indicating on which of the sets of normals, A, B, or C, the spheres of each layer lie. Thus the sequences ABCABC... and ABABAB... correspond respectively to cubic and hexagonal close-packing. Next, if the centre of each sphere is joined to the centres of its nearest neighbours by lines, polyhedra are formed, the smallest among them being tetrahedra and octahedra. There are always two tetrahedra and one octahedron per close packed site and these fill space completely. Allowing one cation to be accommodated in the centre of each of these polyhedra we therefore have the following restricting condition for our structures:

$$\sum_i c_i \leq 3$$

where c_i is the number per anion (per close-packed site) of cations of the ith kind. In Table 3 a list of the known structures of this type is given and for each structure the fractions of filled tetrahedral and octahedral holes are indicated. As is seen from the Table, not all possible degrees of filling occur. The main reason for this

is that most of the solids and certainly all semiconductors crystallizing in these structures are normal valence compounds adhering to the octet rule (1) which if $b_a = b_c = 0$ can be written in the form

$$\sum_i c_i v_i = 8 - N \qquad \ldots(1a)$$

where c_i is defined as above, v_i is the valency of the cations of kind i, and N is the Group number of the anions.

The filling patterns of the anion sub-lattices are further restricted by the particular arrangement in these sub-lattices of the tetrahedra and octahedra (see e.g.

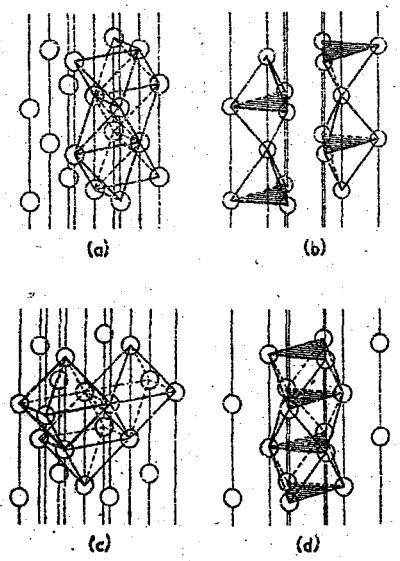


Figure 4. *The arrangement of the tetrahedra: (a) In close-packed cubic and (b) in close-packed hexagonal arrays. The arrangement of the octahedra: (c) In close-packed cubic and (d) in close-packed hexagonal arrays*

Wells[38]). Thus in the cubic close-packing, neighbouring tetrahedra share edges only (Figure 4(a)). Their centres, therefore, are far enough apart to allow all the tetrahedral interstices to accommodate a cation, the resulting structure being the antifluorite type. In the hexagonal close-packed structure, however, the tetrahedra appear in pairs, each tetrahedron sharing a face with its partner (Figure 4(b)). This close proximity of neighbouring tetrahedra prevents the filling of more than half of them so that no hexagonal analogue of the antifluorite structure exists. Concerning the octahedra the situation is similar. In the cubic packing neighbouring octahedra share edges, whereas in the hexagonal packing they share faces (Figure 4(c), (d)). However, because of the larger size of the octahedral holes complete filling is possible in each case, although in the hexagonal packing the cations must have low effective charges if all octahedra are to be filled. Complete

Table 3. Structures with Close-packed Anion Sub-lattices†

Kind and fraction of holes occupied by the cations		Sequence of close-packed anion layers			
Tetrahedral	Octahedral	ABAB...	ABCABC...	ABAC...	Other
$\frac{1}{8}$	0		SnI_4 ($D1_1$)		
$\frac{1}{6}$	0	Al_2Br_6			
$\frac{1}{4}$	0		HgI_2 (C13) SiS_2 (C42)		
$\frac{1}{3}$	0	β-Ga_2S_2 (Superlattice of B4)	α-Ga_2S_3 (B3)		
$\frac{2}{5}$	0	$ZnAl_2S_4$ (H.T.) (B4)	α-Ag_2HgI_4 (B3) $CdAl_2S_4$ (E3) $CdIn_2Se_4$		
$\frac{1}{2}$	0	ZnS (B4) (wurtzite)	ZnS (B3) (zinc blende) CuTe	SiC III (B5)	SiC II (B6) SiC I (B7)
		Cu_3AsS_4 ($H2_5$)	$AgInSe_2$ ($E1_1$) Cu_2FeSnS_4 ($H2_6$)		
$\frac{3}{5}$	0		Mg_3As_2 ($D5_3$) Zn_3P_2 ($D5_9$) ($D5_5$)		
1	0		Mg_2Sn (C1) LiMgSb ($C1_b$) Li_3AlN_2 ($E9_d$) Li_5GeP_3 (C1)		
0	$\frac{1}{6}$	α-WCl_6			
0	$\frac{1}{3}$	BiI_3 ($D0_5$)	$CrCl_3$ ($D0_4$)		
0	$\frac{1}{2}$	CdI_2 (C6)	$CdCl_2$ (C19) TiO_2 (C5) (anatase)	$HgBr_2$ (C24) $HgCl_2$ (C25) CdI_2 (C27)	
0	$\frac{2}{3}$	α-Al_2O_3 ($D5_1$) $LiSbO_3$			Bi_2Te_3 (C33)
0	1	NiAs (B8) MnP (B31)	NaCl (B1) HgS (B9) GeS (B16) SnS (B29)		
$\frac{1}{8}$	$\frac{1}{2}$	Mg_2SiO_4 ($H1_2$) (olivine)	$MgAl_2O_4$ ($H1_1$) (spinel)		
$\frac{1}{2}$	$\frac{1}{8}$		Co_9S_8 ($D8_9$)		
$\frac{1}{2}$	$\frac{1}{2}$		Cu_4Te_3 (38)		
$\frac{1}{2}$	$\frac{1}{2}$	Mg_3Bi_2 ($D5_2$)			
$\frac{1}{2}$	1		Cu_2Sb (C38)		
(2/3)			Cu_2Te		
1	1		Li_3Bi ($D0_3$) Li_2MgSn ($D0_{3b}$)		

† Compounds with deformed close-packed anion arrays are also included. Moreover, in all the listed structures the anions occupy only close-packed sites and they fill them completely.

filling of the octahedral interstices in cubic and hexagonal close-packed sub-lattices leads to the rocksalt (B1)† and the nickel arsenide (B8) structure types respectively. The first is found in both covalently and ionically bonded phases, the second only in covalently bonded ones (the largest electronegativity difference found in nickel arsenide phases is $\Delta x = 1$ corresponding to a relatively small ionic character of the bonds).

Of the phases crystallizing in the structures listed in Table 3 only those obeying equation (1*a*) are of interest to us. It is possible, therefore, to partition them according to the numbers c_i and valencies v_i of their cations and in this way to arrive at the classification outlined in Table 4. A similar classification has been derived by Goodman[18] from a consideration of the possible substitutions in the diamond type structure.

It can be shown[39,40] that phases arising from a filling of the tetrahedral holes in a close-packed anion sub-lattice and satisfying equation (1*a*) automatically fulfil the conditions for semiconductivity given in sub-section 2.1. Table 4, therefore, is very useful in a search for new semiconducting materials. Phases formed by filling the octahedral holes, on the other hand, are not all semiconductors even if their compositions are such as seemingly to satisfy equation (1*a*). The reason for

† Strukturbericht type.

Table 4

v_i	Σv_i	$\Sigma c_i v_i = 8-N$			
		1	2	3	4
1	1	AgI (B3, B4) NaCl (B1)			
1,1	2		Li_2Se (C1) [Cu_2Te]		
2		HgI_2 (C13) CdI_2, PbI_2 (C6) $CdCl_2$ (C19) $HgBr_2$ (C24) $HgCl_2$ (C25) CdI_2 (C27)	ZnS (B3, B4) CdO, NiO, PbS (B1) HgS (B9) GeS (B16) SnS (B29) [MnTe (B8)] [CuTe]		
1,1,1	3			Li_3Bi, [Cu_3Sb] ($D0_3$)	
1,2				LiMgSb ($C1_b$) [Cu_2Sb (C38)]	
3		$CrCl_3$ ($D0_4$) BiI_3 ($D0_5$)		InSb (B3) InN (B4) [CeAs (B1)] [NiAs (B8)] [MnP (B31)]	

Table 4—*continued*

v_i	Σv_i	$\Sigma c_i v_i = 8-N$			
		1	2	3	4
1,1,1,1	4				
1,1,2		α-Ag_2HgI_4 (B3)			Li_2MgSn ($D0_{3b}$)
2,2			$CdZnSe_2$ (B3) $PbSnS_2$ (B29)		Mg_2Sn (C1)
1,3			$AgInSe_2$ ($E1_1$) $AgBiS_2$ (B1)		
4		SnI_4, TiI_4 ($D1_1$)	SnS_2, PtS_2 (C6) TiO_2 (C5)		SiC (B5, B6, B7) [ZrC (B1)] [PtSn (B8)] [PtGe (B31)]
1,1,1,1,1,1	6				
1,1,1,1,2					
1,1,2,2					
2,2,2			Zn_2FeS_3 (~B3)	Mg_3As_2 ($D5_3$) Zn_3P_2 ($D5_9$) Mg_3Bi_2, $ZnMg_2Sb_2$ ($D5_2$)	
1,1,1,3			$Cu_3(Sb,As)S_3$ (~B3)	Li_3AlN_2 ($E9_d$)	
1,2,3					
3,3		Al_2Br_6	Ga_2S_3 (B3, B4) Bi_2Te_3 (C33) α-Al_2O_3 ($D5_1$) γ-Al_2O_3 ($H1_1$)	$GaInSb_2$ (B3)	
1,1,4					
2,4				$ZnGeP_2$ ($E1_1$)	
1,5			$LiSbO_3$		
6					
Higher derivatives					
2,3,3	8		$CdAl_2S_4$ (E3) $CdIn_2Se_4$ $MgAl_2O_4$ ($H1_1$)		
1,1,2,4	8		Cu_2FeSnS_4		
2,2,4	8		Mg_2SiO_4 ($H1_2$) Ni_2GeO_4 ($H1_1$)		
1,1,1,5	8		Cu_3AsS_4 ($H2_5$)		
1,1,1,1,1,4	9			Li_5GeP_3 (C1)	
1,1,1,2,3,4	12		$Cu_3Fe_2SnS_6$ (B4)		
1,3,3,3,3,3	16		$LiAl_5O_8$ ($H1_1$)		
	16		[Co_9S_8]		

Square brackets indicate that the representatives of the corresponding family of compounds are not necessarily semiconductors.

this is that octahedral co-ordination need not only result from the formation of saturated covalent, i.e. directional, bonds. Indeed, octahedrally co-ordinated atoms are very frequently met in metallic and ionic, i.e. non-directionally bonded, materials. It is of interest, therefore, to discuss here the main types of covalent bonds which lead to octahedral co-ordination. Thus the p orbitals of an atom can readily be used to form up to three electron pair bonds directed along the axes of a rectangular co-ordinate system. If these p bonds undergo pivotal resonance the central atom acquires an octahedral neighbourhood. p bonds are formed by the atoms listed in Table 5 which tend to keep their s electrons unshared. Since, according to Pauling,[7] at least one of any two atoms linked by a pivoting bond must contain empty orbitals in the valence shell and since in semiconductors the anions complete their outermost subshells, the empty orbital must occur on the cations. It therefore follows that, unless an electron transfer occurs, pivoting p bonds are only met in semiconductors containing the Group IIIB and IVB cations of Table 5. The best known representatives of such semiconductors are those of the lead sulphide family. To account for the octahedral co-ordination of Group VB cations which is nevertheless observed in many semiconductors, Mooser and Pearson[36]

Table 5

IIIB	IVB	VB	VIB	VIIB
				Cl
		As	Se	Br
In	Sn	Sb	Te	I
Tl	Pb	Bi		

suggested that in these cations the p and d orbitals of the valence shell are hybridized. Thus the p^3d^3 hybrid orbitals which extend towards the corners of a trigonal antiprism (distorted octahedron) fit very well into the octahedral co-ordination of, for example, the bismuth atoms in bismuth telluride.† When there are not enough valence electrons available to form six bonds such a hybridization provides the empty orbitals necessary for resonance. Since the tendency to keep the s electrons unshared is less pronounced in the lighter atoms of Table 5 one might also expect these lighter atoms to form octahedral sp^3d^2 bonds in certain compounds. Indeed, the cadmium iodide (C6) structure met in silicon ditelluride ($SiTe_2$), tin disulphide (SnS_2), and tin diselenide ($SnSe_2$) is readily understood if one assumes silicon and tin to form resonating sp^3d^2 bonds. Lead, on the other hand, in which the tendency to keep an unshared pair is well developed, does not form dichalcogenides. Nevertheless, there exists a lead compound which crystallizes in the cadmium iodide structure, namely lead diiodide (PbI_2). However, in this compound lead is divalent, forming two resonating p bonds.

Bonds involving d orbitals are of particular importance in compounds containing transition elements. Thus, it would seem that a transitional atom could readily

† In contrast to this, Drabble and Goodman[43] assume that sp^3d^3 bonds form in bismuth telluride. However, their bond scheme seems to account less satisfactorily for the observed interatomic distances.

form d^2sp^3 bonds when located in an octahedral hole of a close-packed anion array. However, short range order and even magnetic properties do not as a rule permit unequivocal determination of the valence states of these atoms in solids. The knowledge of how many *d* electrons and orbitals are involved in the bonding is, therefore, in general lacking. There is evidence that in the nickel arsenide phases, for example, at low axial ratios ($c/a \lesssim 1{\cdot}55$) cation–cation bonds form which run throughout the whole crystal, thus producing metallic properties (Pearson[42]). Nevertheless, there exists no really clear-cut criterion for predicting which of the nickel arsenide phases are metallic and which are semiconducting. Some progress in this direction has recently been made by Hulliger.[74] For a review on semiconductors containing transition elements, the reader is referred to the recent article by Morin.[43]

3.2. The Directional Character of Covalent Bonds and the Composition of Normal Valence Compounds

The classification discussed above shows how a great number of semiconducting structures can be derived by putting cations into the interstices of close-packed anion sub-lattices. It does not, however, state which atoms actually lend themselves to the formation of these structures. The empirical relationship between the directional character of the bonds in solids and some suitably chosen parameters of the component atoms[44] represents a step towards the solution of this problem. Such a relationship is all the more important since, as was shown in sub-section 2.1, the very concept of the semiconducting bond rests on the intimate connection between the short range order in semiconductors and the valence states of the constituents. This connection is limited to *directionally* bonded phases. The structures of *non-directionally* bonded phases are governed by geometrical principles[45] which state that such structures tend towards the best possible filling of space and the highest symmetry, and that the networks formed by the links between nearest neighbours tend towards the highest dimension. The short range order found in these structures is therefore no longer related to the valence states of the component atoms, and it becomes of interest to study the changes in short range order and structure which occur in a series of chemically similar solids on going from the directionally to the non-directionally bonded ones.

It was pointed out by Dehlinger[46] that the principal quantum number *n* of the outermost occupied shell of an atom measures the directional character of the bonds formed by this atom with atoms of the same kind: as *n* increases the bonds gradually lose their directional properties (see also sub-section 2.2). To extend this relationship to compounds we introduce a suitably averaged principal quantum number

$$\bar{n} = \frac{\sum_i a_i n_i}{a_i}$$

where n_i is the principal quantum number of the valence shell of the atom of the *i*th kind and a_i is the number, per formula unit, of atoms of this kind.

In compounds the directional character of the bonds does not depend on the value of $\bar{n}$ only. Indeed, it rapidly decreases regardless of $\bar{n}$ as the ionicity of the bonds increases. A convenient measure for the bond ionicity is the difference Δx

between the electronegativities of the anions and cations, and in order to be able to discuss compounds with more than two components we define

$$\Delta x = |\bar{x}_a - \bar{x}_c|$$

where $\bar{x}_a$ and $\bar{x}_c$ are the arithmetic means of the anion and cation elec ronegativities. (The electronegativities listed by Gordy and Thomas[47] are used throughout our discussion.)

If various chemically similar compounds are plotted in a $\bar{n}$ versus Δx diagram then according to the above discussion it is expected that covalent, i.e. directionally bonded phases, occur in a region of the diagram where $\bar{n}$ and/or Δx are small. Non-directionally bonded compounds, on the other hand, are expected to lie in

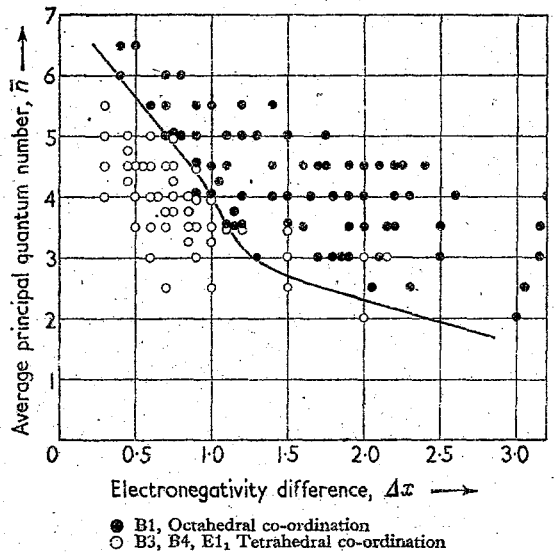


Figure 5. *Normal valence compounds of composition* AX *with tetrahedral and octahedral co-ordination*

the region of large $\bar{n}$ and/or Δx values. Such expectations are borne out to a rather pleasing extent for normal valence compounds of composition $A_iX_j(i,j = 1,2,3)$. Thus, if in Figure 5 we plot the AX compounds formed by non-transitional elements in their normal (highest) valence state which crystallize in the B1, B3, and B4 types of structures we see that a sharp separation occurs between the tetrahedrally and octahedrally co-ordinated structures. (The chalcopyrite phases of composition ABX_2 are also plotted in Figure 5 because their $E1_1$ type of structure is essentially a superlattice of the zinc blende structure.) A more complete account of the $\bar{n}, \Delta x$ classification of AX compounds is given elsewhere.[44] Here we should add only that the separation between different structures is further amplified if the electronic configuration of the component atoms is taken into account. This is seen in Figure 6 where we have plotted normal valence compounds containing A Group cations only. The $\bar{n}$ versus Δx diagram is now partitioned into different regions each containing one of the following structures: zinc blende (B3), wurtzite (B4),

rocksalt (B1), and caesium chloride (B2). It is noteworthy that the arrangement of the regions is such that at any value of $\bar{n}$ the structure with the higher Madelung constant always occurs at the higher Δx value.

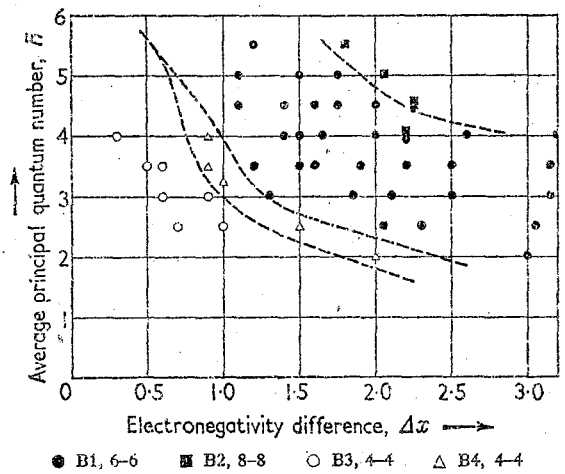


Figure 6. *Normal valence compounds of composition* AX *containing* A *Group cations only. The co-ordination numbers are indicated*

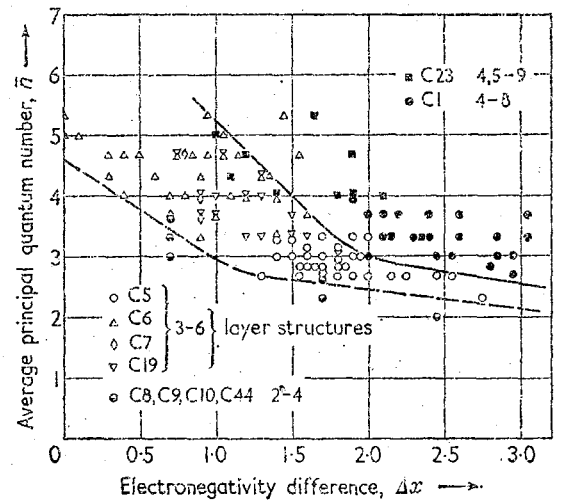


Figure 7. *The structures of normal valence compounds of composition* AX_2

Next, discussing *anion-rich* compounds of composition AX_2, AX_3, and A_2X_3 we find that, with the exception of the rather complex lead chloride (C23) and uranium trichloride (UCl_3) structures, all structures to the right of the solid lines in Figures 7–9 are based on *close-packed arrays of cations* whose interstices are

filled up to various degrees by anions (Table 6). Because the anions outnumber the cations the filling of the cation sub-lattices is high, giving rise to densely packed structures which come close to fulfilling the requirements of Laves'[45] geometrical

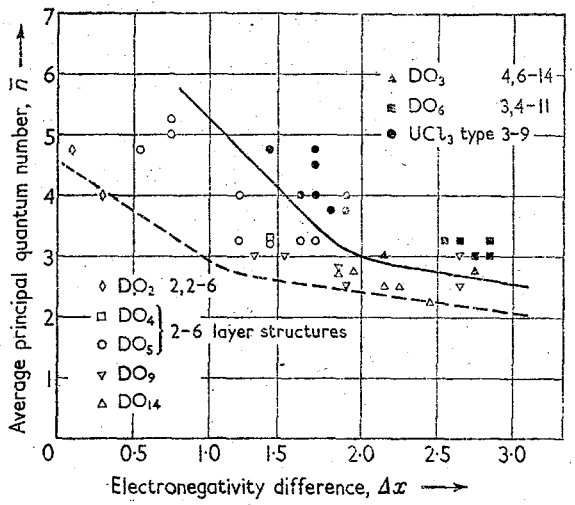


Figure 8. *The structures of normal valence compounds of composition* AX_3

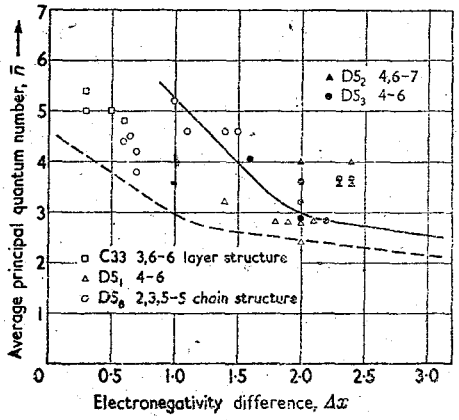


Figure 9. *The structures of normal valence compounds of composition* A_2X_3

principles. This is exactly what would be expected of the structures falling into the non-directionally bonded region (high $\bar{n}$ and/or Δx). Because the tetrahedral interstices are twice as numerous as the octahedral ones (sub-section 3.1) most of the anions occupy tetrahedral holes in the cation sub-lattices.

Structures in the directionally bonded region (i.e. lying to the left of the solid lines in Figures 7–9) are, on the other hand, characterized by *close-packed arrays*

Table 6

Phase type	AX_2		AX_3		A_2X_3	
Close-packing of cations	cubic	hex	cubic	hex	cubic	hex
Structure types	C1	—	$D0_3$	$D0_5$	$D5_3$	$D5_2$

of anions (Table 7). Because of the low cation to anion ratio the filling of the interstices in the anion sub-lattices is rather incomplete, so that the resulting structures are less densely packed and show lower co-ordination numbers than the non-directionally bonded ones. In many instances layer structures occur.

Table 7

Phase type	AX_2		AX_3		A_2X_3	
Close-packing of anions	cubic	hex	cubic	hex	hex	other
Structure types	C19	C6	$D0_4$	$D0_5$ $D0_{24}$	$D5_1$	C33

We can now appreciate why many semiconductors have structures based on close-packed anion arrays as discussed in sub-section 3.1: in anion-rich valence compounds close-packed anion sub-lattices form only if they are covalently, i.e. directionally, bonded and we have seen that the covalent nature of the bonds is an outstanding characteristic of semiconductors.

In anion-rich structures with close-packed anion sub-lattices the cations are usually located in the octahedral holes so that in the AX_2 compounds, for example, a 3-6 co-ordination results. The C18 type structure of mercury diiodide (HgI_2), however, is an exception, the cations occupying 1/4 of the tetrahedral holes of the anion sub-lattice. The resulting 2-4 co-ordination is normally found only in the very open framework structures of silicon dioxide (SiO_2) (C8, C9, C10) and germanium disulphide (GeS_2) (C44), and in the chain structure of silicon disulphide (SiS_2) (C42). As would be expected from our discussion these very open structures occur at the lowest values of $\bar{n}$ and/or Δx and accordingly they are separated from all the other AX_2 compounds in a $\bar{n}$ versus Δx plot (Figure 7).

In most of the *anion-rich* compounds the anions also form close-packed sub-lattices. This might be taken as an indication of the directional character of their bonds since in the *anion-rich* phases we have met close-packed anion sub-lattices only in directionally bonded structures. However, here the cations are more numerous than the anions and the degree of filling of the anion sub-lattice is therefore high. Indeed, it is found that the most common structures in which the cation-rich phases crystallize are the *antitypes* of the densely packed structures listed in Table 6. Many A_2X compounds for instance adopt the antifluorite (C1)

structure while the A_3X compounds favour either the anti-$D0_3$ structure or the $D0_{18}$ structure which is essentially the antitype of the $D0_6$ structure. Finally, in the A_3X_2 compounds we meet almost exclusively the anti-$D5_2$, the anti-$D5_3$, and the closely related $D5_9$ types. Because of their dense packing these structures are expected to be stable even when the bonding is non-directional and, as shown for example in Figure 10, the C1 phases are scattered over the whole $\bar{n}$ versus Δx diagram. Similar plots are obtained for the A_3X and A_3X_2 compounds but, because of the more stringent valence requirements, the scatter is somewhat less extended in these compounds.

In some ternary compounds of composition ABX which crystallize in an ordered fluorite-like structure, the anions do not occupy the close-packed (calcium) sites. Consideration of the bonding led to the prediction and subsequent discovery of the *metallic* properties of these compounds.[30] In contrast to this, *semiconducting*

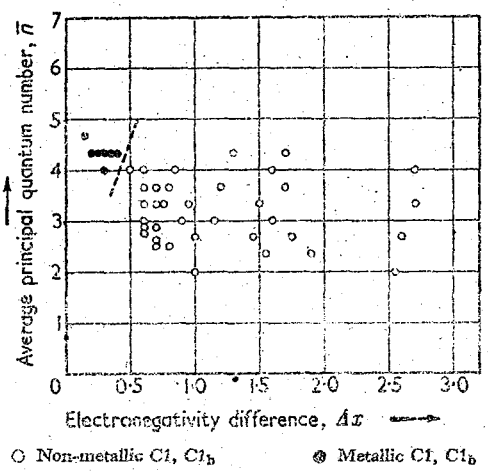


Figure 10. The antifluorite type phases

properties were predicted and found in the ABX compounds in which the anions are close-packed. It is interesting to note that the metallic ABX phases only occur at high $\bar{n}$ and low Δx values (Figure 10), i.e. in a region where the bonding is expected to be metallic.

4. THE SURFACE STRUCTURE OF SEMICONDUCTORS

The constitution of semiconductor surfaces is of considerable interest to solid state physicists because many electronic processes in semiconductors are controlled by the deviations from periodicity of the crystal potential near a surface. Thus we know that charge carrier traps and recombination centres exist in semiconductor surfaces and phenomenological theories have been developed to describe the influence of these traps and centres upon the properties of semiconductors. Nevertheless, our knowledge of the actual structure of semiconductor surfaces is rather limited. By means of low energy electron diffraction, information can be gained

about the spacings between surface atoms and about the structure of clean surfaces and its modification on exposure to various gases (cf. Farnsworth and Schlier[48,49]). Studies of crystal habit, microcleavage, and etch pits can give a wealth of information on surface energy, surface structure, effective charges associated with surface atoms and with atoms in the bulk, and on the effects of chemical agents on the structure of a clean surface (cf. Wolff and Broder,[23] Pearson and Wolff[50]). Such information is of considerable help in understanding the nature of the chemical bonding in semiconductors and for this reason it is appropriate to consider some of the experimental results which have been obtained so far.

In substances which crystallize in the diamond or zinc blende structures, whose interior atoms, therefore, are tetrahedrally co-ordinated, two kinds of surface atoms are found: (1) those lying on a 111 face (or any *hhl* face for which $h > l$); and (2) those lying on a 001 face. Surface atoms of the first kind have three neighbours in the bulk; those of the second kind only have two neighbours. All other surfaces are made up of atoms of these two kinds. The properties of a surface of these crystals then depend on the type of atoms which lie on this surface and on the electronic and chemical processes which occur to saturate the broken bonds. Thus it is found that in the elements silicon, germanium, and grey tin, the valence *s* and *p* orbitals of the atoms on a 111 face for instance are no longer hybridized. The surface atoms acquire an octahedral neighbourhood by moving into the bulk of the crystal by 1/4 of the length of the body diagonal. This can readily be understood if one assumes the surface atoms to form resonating *p* bonds. Since the *s* orbitals of the surface atoms contain only one electron each the surface atoms act as acceptors, giving rise to the P-type surface conduction which has been observed in silicon and germanium. In mercury selenide (HgSe), mercury telluride (HgTe), and, somewhat less pronounced, also in cadmium telluride (CdTe), similar deformations are found which in these compounds lead to rocksalt type surface layers. It is noteworthy that in Figure 5 all the zinc blende type compounds in which this rocksalt type surface structure occurs lie near the borderline between the tetrahedrally and octahedrally co-ordinated phases.

The clean surfaces of a semiconductor can also, and generally do, attain stability through chemical reactions with ambients so that further information about the constitution of such surfaces can be gained by subjecting them to the etching action of various gases, solutions, and liquid metals. The reactivity of a particular surface is found to depend on the valency of the ambient. Thus, monovalent atoms (halogens) prefer to become attached to surface atoms of type (1) while divalent atoms (oxygen, sulphur) show a preference for atoms of type (2). Moreover, size effects play a role, so that surface attack by large iodine atoms can differ appreciably from that by the smaller chlorine atoms.

The optical examinations of microcleavage patterns reported by Wolff and Broder[23] are of particular interest in connection with the subject of this article since they give information on the nature of the surface and bulk bonding in solids. The cleavage patterns of zinc blende type compounds for instance are influenced by the effective charges on the component atoms: because of these charges all zinc blende type compounds cleave on 011. However, if the charges are very small, cleavage on 111 is also found. The ratio of 111 to 011 cleavage, therefore, is a measure of the ionicity of the bonds and the results obtained so far indicate that in these compounds the effective charges on the atoms are less than unity.

The relative strengths of the bonds in ternary compounds can also be estimated from examination of micro-cleavage patterns. In chalcopyrite ($CuFeS_2$) and in enargite (Cu_3AsS_4) chains of the type cation_I–anion–cation_II are found to be considerably stronger than chains of the types cation_I–anion–cation_I and cation_II–anion–cation_II. In particular it is found that in chalcopyrite one of the bonds (probably that between Fe and S) is much stronger than the other (Cu–S).

We have indicated some of the ways in which information on the surface structure of semiconductors can be gained and how this information leads to a better understanding of the surface and bulk bonding in semiconductors. Moreover, our discussion would suggest that a rigorous application of the bond treatment to the description of surface phenomena might well increase our, at present rather limited, knowledge of the relationship between surface structure and surface states. Some of the problems met in the physics and chemistry of semiconductor surfaces are akin to those arising from the presence in semiconductors of physical defects such as vacancies and dislocations. It is expected, therefore, that in this field too, the bond approach will prove to be of interest.

5. ENERGY GAPS AND CHARGE CARRIER MOBILITIES AND THEIR DEPENDENCE ON BONDING

Following Welker's discovery of the semiconducting properties of the $A^{III}X^V$ compounds the chemical approach to semiconductivity as outlined in Section 2 rapidly provided us with the necessary means for predicting and classifying semiconducting materials. Little progress has, however, been made so far in predicting and calculating such fundamental parameters as band gaps and charge carrier mobilities of semiconductors. The reasons for this are manifold. First we note that present-day electron theories cannot cope with this problem in a general way. Indeed, rather than producing theoretical values for band gaps, accurate band calculations make use of the experimentally determined values to derive the proper energy–momentum relationship in a semiconductor. The theories dealing with charge carrier mobilities often cannot even give the right temperature dependence of the mobilities and the situation is worse as regards absolute mobility values. Moreover, it seems that in low mobility semiconductors the very concept of charge carrier velocity becomes meaningless.[51]

Under these circumstances, at best one might seek for empirical relationships between the semiconductor parameters and some suitably chosen parameters characterizing the component atoms and possibly the crystal structure. If these relationships are to be of any value, they should not only hold for the energy gaps and mobilities of known semiconductors, but they should allow the properties of new semiconductors to be deduced by extrapolation or interpolation. Even this empirical approach is, in the case of the charge carrier mobilities, severely limited because the experimentally determined mobilities do not necessarily represent intrinsic properties but may be affected by deviations from ideal composition and by crystal imperfections. The dependence on the date of measurement (!) of the electron and hole mobilities in germanium clearly reflects this unfortunate state of affairs (Shockley[52]). Moreover, while in a comparative study of energy gaps one can readily extrapolate to the values at absolute zero, such an extrapolation is not warranted in the case of the mobilities and normally one has to compare

mobilities at some arbitrary temperature (i.e. room temperature). Such a procedure obviously casts doubt upon the usefulness of the resulting relationship.

In spite of all these complications and uncertainties we now discuss some attempts at relating the values of energy gaps and mobilities to the chemical composition and structure of semiconductors. We do so because however dubious these attempts may be, they are, nevertheless, often very helpful to the experimentalist. Moreover, most of the recently established relationships are based on the bond treatment and thus afford further insight into the merits and shortcomings of this approach.

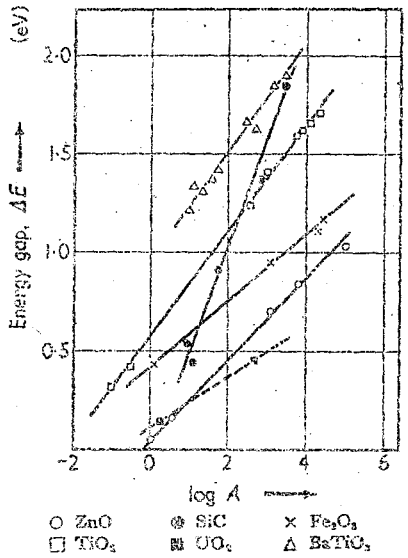


Figure 11. The Meyer–Neldel rule for various compound semiconductors (after Busch[55])

For completeness we should first of all mention some of the earlier established relationships. The temperature dependence of the electrical conductivity σ of many semiconductors in the extrinsic range is given by

$$\sigma = A \exp -(\Delta E/2kT)$$

According to Meyer and Neldel,[56] the energy gap ΔE in this formula is related to the constant A by a logarithmic law:

$$\log A = \alpha + \beta \Delta E$$

α and β being two constants depending on the material under investigation. Figure 11 shows the Meyer–Neldel rule for some compound semiconductors. Although attempts have been made to explain this rule theoretically,[55, 56] it seems that it is better considered as a purely empirical law. This attitude should be taken also towards the relationships between lattice constant a and energy gap ΔE as established for the Group IVB elements by Moss[57] (Figure 12) and towards the relationships between the ratio r_1/r_2 and ΔE as discussed by Miyauchi[58] (Figure 13).

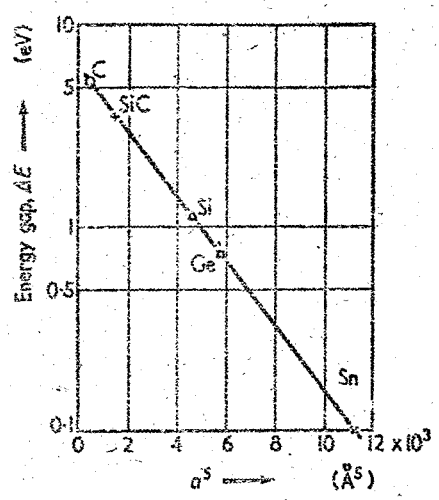


Figure 12. Empirical relationship between energy gap ΔE and lattice constant a of the Group IV elements (after Moss[67])

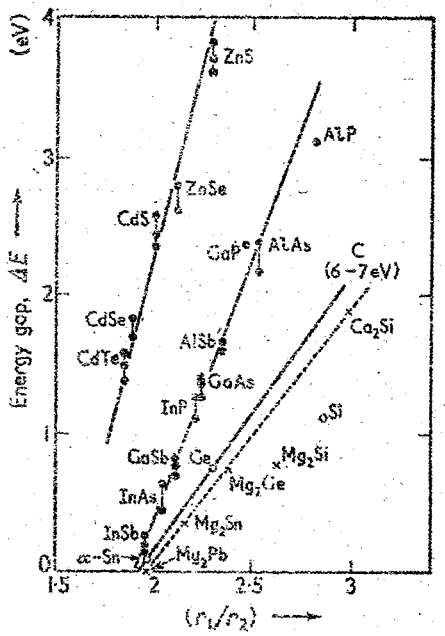


Figure 13. Empirical relationship between ΔE and the ratio r_1/r_2. r_1 is half the interatomic distance and r_2 is the average of the positive ionic radii (after Miyauchi[68])

Here r_1 is half the distance between neighbouring atoms in the crystal and r_2 is the average of the positive ionic radii of the constituents.

Welker's rules[4] on the other hand do at least afford a qualitative interpretation. These rules can be summarized as follows:

(1) In chemically and structurally similar semiconductors the one with the strongest bonds normally has the highest band gap and the highest electron mobility.

(2) Under otherwise equal conditions the semiconductor with the largest ionic component in the bonding normally has the highest band gap and the lowest electron mobility.

Theoretical interpretations of these rules have been given by Seraphin,[59] Heywang and Seraphin,[60] Gubanov,[61] and by Adavi.[62] Rather than reproducing the arguments of these authors we should point out here that the simple molecular orbital treatment given in sub-section 2.2 affords a qualitative interpretation of those of Welker's rules which relate to the size of the energy gaps. Thus, one readily recognizes that the splitting in a diamond type semiconductor—and Welker's rules are designed for these—between bonding and antibonding levels, and hence the gap between valence and conduction band, are largest when the bonding is strongest. As to the heteropolarity of the bonding we have pointed out in subsection 2.2 that it is expected to produce a larger splitting between bonding and antibonding levels. The molecular orbital treatment would also suggest that in the diamond-like structures of composition ABX_2 (chalcopyrites) in which A–X and B–X bonds are formed, the energy gap should be determined by the splitting between the bonding and antibonding levels corresponding to the weaker of the two bonds (see also Goodman,[18] Pearson and Wolff[50]). Finally, we note that if the valence orbitals in addition to bonding and antibonding bands also form a non-bonding band then complications in the relationships between bond strength and band gap can arise which might well account for the deviations from Welker's rules as observed, for example, in the lead sulphide family of semiconductors.

Attractive as Welker's rules appear at first sight there are, nevertheless, great difficulties in their application to the prediction of band gaps and charge carrier mobilities because both bond strength and ionic component of the bond are not readily measurable parameters. In applying Welker's rules various authors, therefore, have used different measures for these parameters. Starting from the relation between bond strength and bond length as established by Pauling,[63] Goodman[55] plots the square root of the energy gap ΔE of the Group IVB elements against $1/d^2$, where d is the interatomic distance (Figure 14). He then assumes that if the bonds in the $A^{III}X^{V}$ compounds were purely covalent the corresponding ΔE values would also obey the linear relationship given in Figure 14 so that any deviation, ΔE_{ion}, can be attributed to the ionicity of the bonds. Goodman is able to relate ΔE_{ion} to the charge carrier mobilities found in the $A^{III}X^{V}$ compounds. Unfortunately his treatment only covers the energy gaps and mobilities of a few zinc blende type semiconductors and a comparison between Figure 14 and Figure 12 would suggest that relationships between the energy gaps ΔE and structural parameters such as a and d are quite accidental.

The heat of formation has been used as a measure for the strength of the bonds in a semiconductor and a relationship between heat of formation and band gap has

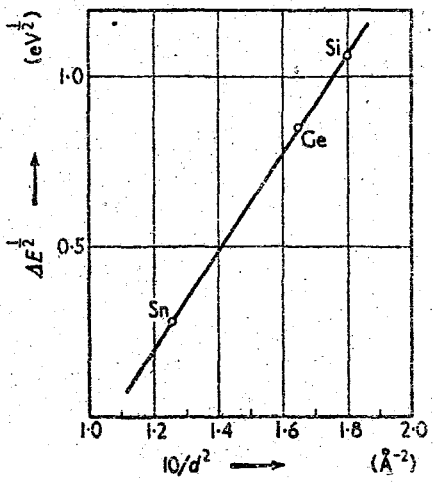


Figure 14. *Empirical relationship between ΔE and the interatomic distance d of the Group IV elements (after Goodman[64])*

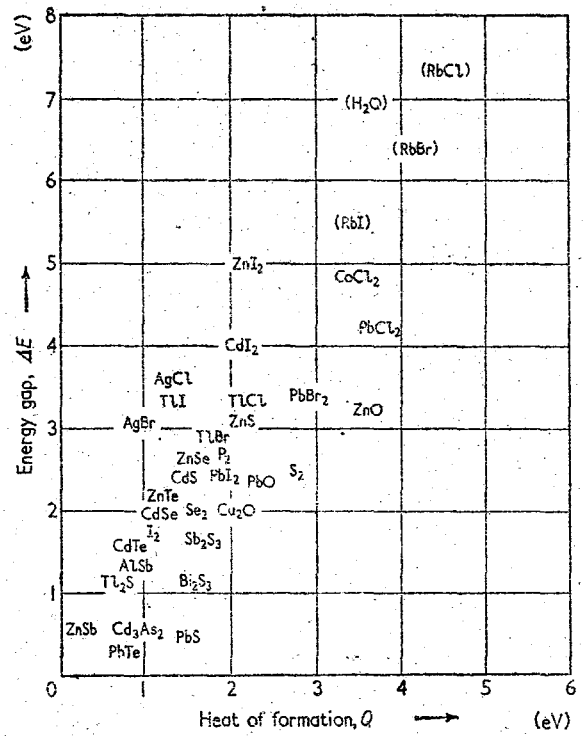


Figure 15. *Empirical relationship between ΔE and the heat of formation Q (after Ruppel, Rose, and Gerritsen[65])*

been discussed by Ruppel, Rose, and Gerritsen[65] (Figure 15). These authors also give a method for estimating band gaps of ionic compounds from the heats of hydration of the constituents (Figure 16). Zhuze,[66] on the other hand, related the

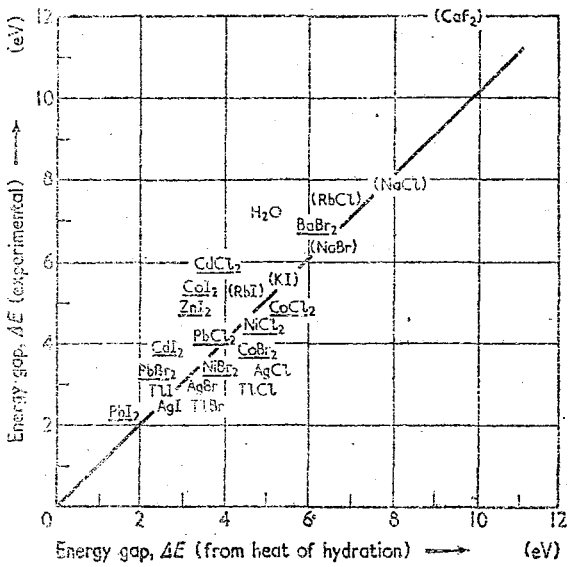


Figure 16. *Comparison between experimentally determined energy gaps and energy gaps as calculated from the heats of hydration (after Ruppel, Rose, and Gerritsen[65])*

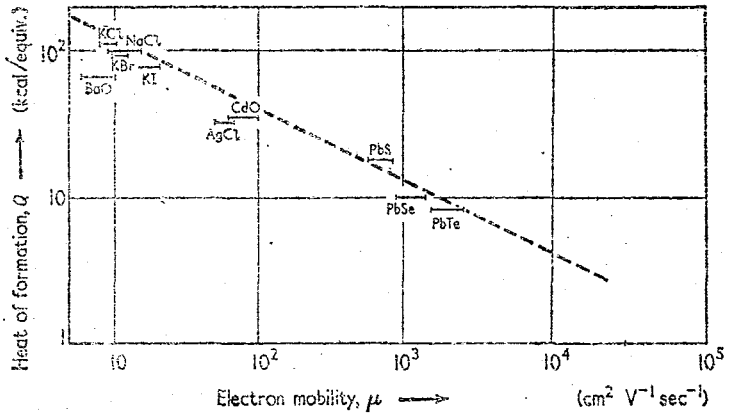


Figure 17. *Empirical relationship between the electron mobility μ and the heat of formation Q of some rocksalt type compounds (after Zhuze[66])*

heat of formation of a semiconductor to the electron mobility and in the case of some rocksalt and zinc blende type compounds his relationship seems quite satisfactory (Figures 17 and 18). Taking the difference Δx between the anion and cation electronegativities as a measure for the bond ionicity, Zhuze also shows the

influence of an ionic bond component upon the size of the electron mobility (Figure 19). Ormont[67] criticizes the use of the heat of formation as a measure of bond

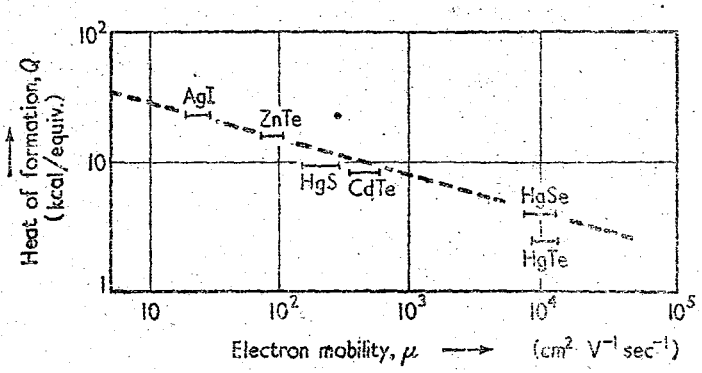


Figure 18. *Empirical relationship between μ and Q of some zinc blende type compounds (after Zhuze[66])*

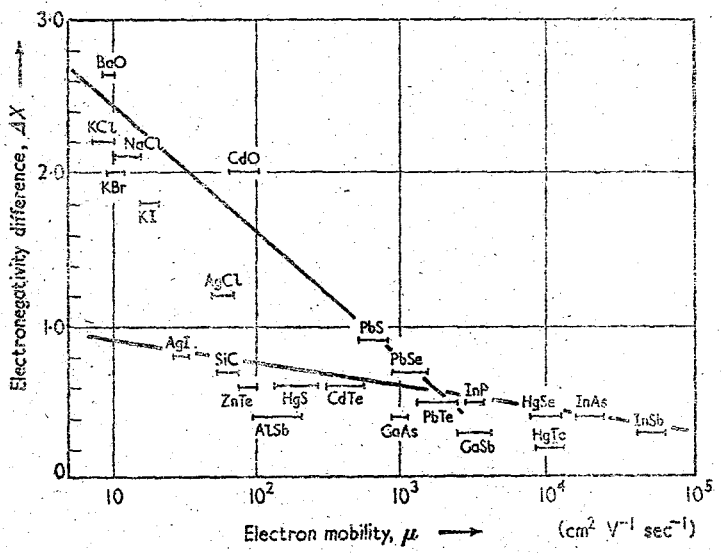


Figure 19. *Empirical relationship between μ and the electronegativity difference Δx of some rocksalt and zinc blende type compounds (after Zhuze[66])*

strength and suggests that dissociation energy Ω and surface energy ϵ_{hkl} are better suited for this purpose. He proposes relationships of the following form:

$$\Delta E = \left(\frac{N_a}{N_c}\right)^p (C-M-I)\epsilon_{hkl}$$

and

$$\Delta E = \left(\frac{N_a}{N_c}\right)^p (c-m-i)\Omega$$

where N_a and N_c are the numbers of valence electrons of the anion and cation respectively, p is a constant, and C, M, I, and c, m, i are terms gauging the relative sizes of the covalent, metallic, and ionic components of the bonding, respectively. C and c are both taken as constant, M and m are taken to be either proportional to, or an exponential of the sum of the atomic numbers, and finally I and i are assumed to be proportional to Δx. While Ormont's formulae reproduce the energy gaps of many semiconductors fairly accurately they reflect but little of the original simplicity of Welker's rules. It would therefore seem that not too much significance can be attached to the interpretation of the parameters C, M, I, c, m, and i.

The original work of Welker[4] as well as the interpretations of the Welker rules show that the influence of an ionic component in the bonding upon energy gap and mobility is mostly due to the difference of the depth of the potential wells of the anions and cations. Clearly this difference is not determined by the electronegativity difference alone as has usually been assumed. To explain the trends observed in the sizes of the energy gaps and mobilities of the $A^{III}X^{V}$ compounds Folberth[68] (see also Folberth and Welker[69]) therefore introduces a polarization which accounts for the asymmetry of the charge clouds associated with the bonds in these compounds. The way in which Folberth arrives at this polarization is, however, open to criticism and, indeed, his treatment seems to be restricted to *some* of the $A^{III}X^{V}$ compounds only.

The dependence of the energy gap on the principal quantum number n (sub-section 2.2) and the successful application of the parameters $\bar{n}$ and Δx in classifying the structures of normal valence compounds (sub-section 3.2) has resulted in yet another interesting relationship. In a recent paper Pearson[70] finds that for the iso-electronic series Si, AlP ($\bar{n} = 3$); Ge, GaAs, ZnSe, CuBr ($\bar{n} = 4$); and Sn, InSb, CdTe, AgI ($\bar{n} = 5$); empirical equations of the form

$$\Delta E - \Delta E_0 = c.\Delta x^b$$

hold (Figure 20), where ΔE_0 is the energy gap of the element of each series and c and b are parameters depending on $\bar{n}$. Pearson uses these equations to construct lines of equal energy gaps in a $\bar{n}$ versus Δx diagram. Realizing, however, that the effect of the polarization as discussed by Folberth is not fully accounted for by Δx alone he introduces a third parameter, namely the ratio r_c/r_a of the tetrahedral radii of the cations and anions. Projecting the lines of equal energy gaps on to the curved surface on which the above compounds lie in a $\bar{n} - \Delta x - r_c/r_a$ diagram (Figure 21) he finds that the band gaps of the remaining zinc blende type compounds lying on this surface are predicted within an average accuracy of 3 per cent. Compounds lying below the surface (i.e. with smaller r_c/r_a values) have larger gaps than would be expected from the lines of equal gaps, and the band gaps of compounds lying above the surface are smaller. This behaviour can be explained by discussing the dependence of the polarization on r_c/r_a. As compared to the relationships mentioned above, his treatment has the advantage of accounting for the energy gaps of *all* the zinc blende type semiconductors for which $\bar{n} \geq 3$.

Most of the relationships mentioned above are limited to series of structurally and chemically similar semiconductors. Attempts have been made by Goodman[18] and by Cornish[71] to systematize the experimentally established trends and thus to arrive at more general relationships. However, it is obvious from these attempts that, at present, any generalization can be gained only at the expense of accuracy.

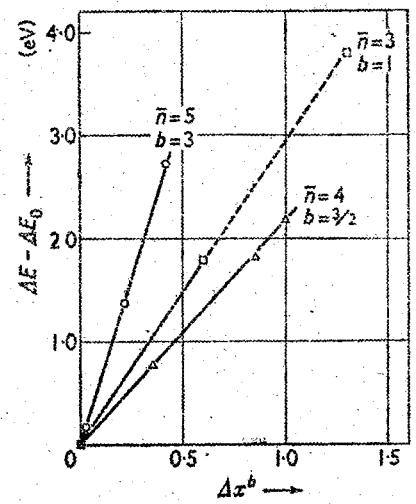


Figure 20. *Empirical relationship between ΔE and Δx in the iso-electronic series of the zinc blende type compounds (after Pearson[10])*

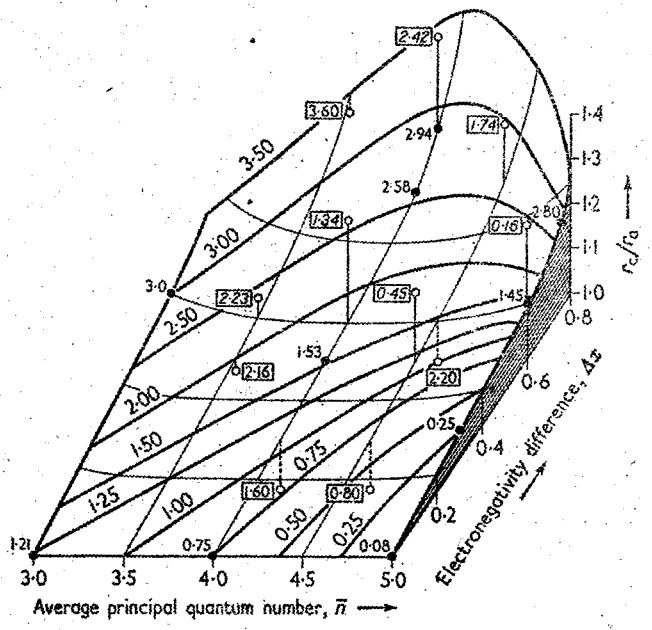


Figure 21. *The energy gaps of the zinc blende type compounds plotted against $\bar{n}$, Δx, and r_c/r_a (after Pearson[10])*

The present discussion of the brave but only partially successful attempts at relating energy gaps and mobilities to some atomic and structural parameters as well as to the character of the bonds, should not be concluded without mention being made of the serious criticism voiced by Herman.[73] According to him the differences in energy gaps and mobilities in semiconductors of the same structure can only be explained in terms of changes in the bonding if all these semiconductors have the same type of band structure and in particular if the band maxima and minima all occur in the same location in momentum space. Unfortunately, there is strong evidence that this is not even the case in the relatively simple zinc blende type semiconductors.

6. CONCLUDING REMARKS

In conclusion it can be said that systematic studies of the chemical bonding in semiconductors have led to the rapid discovery of a large number of new semiconducting materials. This discovery, which is of considerable practical interest, was made possible by the conclusion reached from the bond approach that all multi-component semiconductors are normal valence compounds. This could not have been reached from the band calculations which are available at present since these calculations tend to put emphasis on the symmetry of a crystal rather than on its chemical composition. Moreover, since it is readily possible to introduce chemical information into a bond discussion, such a discussion is well adapted for simultaneous surveys and comparative studies of many semiconductors. In spite of this the experimentally observed trends in the energy gaps and charge carrier mobilities of series of chemically and structurally related semiconductors have, up to now, defied any rigorous interpretation and the future of the bond approach will depend on whether or not it can successfully solve this problem. In studies of the structure and chemistry (adsorption, catalytic action) of semiconducting surfaces considerations of chemical bonding will continue to play an essential role.

REFERENCES

1. M. H. Cohen and V. Heine. *Phil. Mag.* **7**, 395 (1958)
2. E. Friederich. *Z. Phys.* **31**, 813 (1925); **34**, 637 (1925)
3. W. Meyer. *Z. Elektrochem.* **50**, 274 (1944)
4. H. Welker. *Z. Naturf.* **7a**, 744 (1952)
5. A. F. Joffé. *Bull. Acad. Sci. U.R.S.S.* **15**, 477 (1951)
6. H. Krebs and W. Schottky. *Halbleiterprobleme* I, p. 25 (Friedr. Vieweg & Sohn, Braunschweig, 1954)
7. L. Pauling. *Proc. roy. Soc.* **A196**, 343 (1949)
8. E. Mooser and W. B. Pearson. *Phys. Rev.* **101**, 1608 (1956)
9. E. Mooser and W. B. Pearson. *J. Electron.* **1**, 629 (1956)
10. A. H. Wilson. *Proc. roy. Soc.* **A133**, 458 (1931)
11. C. H. L. Goodman and R. W. Douglas. *Physica, 's Grav.* **20**, 1107 (1954)
12. V. M. Glazov, M. S. Mirgalovskaya, and L. A. Petrakova. *Bull. Acad. Sci. U.R.S.S.* **10**, 68 (1957)
13. O. G. Folberth and H. Pfister. *Semiconductors and Phosphors*, p. 474 (Interscience, New York, 1958)
14. I. G. Austin, C. H. L. Goodman, and A. E. S. Pengelly. *J. electrochem. Soc.* **103**, 609 (1956)
15. C. H. L. Goodman. *Nature, Lond.* **179**, 828 (1957)
16. D. N. Nasledov. *Semiconductors and Phosphors*, p. 641 (Interscience, New York, 1958)
17. J. Appel. *Z. Naturf.* **9a**, 265 (1954)

18. C. H. L. Goodman. *J. Phys. Chem. Solids* **6**, 305 (1958)
19. H. Krebs. *Acta cryst., Camb.* **9**, 95 (1956)
20. I. M. Tsidil'kovskii. *C. R. Acad. Sci. U.R.S.S.* **102**, 737 (1955)
21. G. F. J. Garlick, J. M. Hough, and R. A. Fatehalley. *Proc. phys. Soc. Lond.* **72**, 925 (1958)
22. G. A. Jeffrey, G. S. Parry, and R. L. Mozzi. *J. Chem. Phys.* **25**, 1024 (1956)
23. G. A. Wolff and J. D. Broder. *Acta cryst., Camb.* **12**, 313 (1959)
24. L. Pauling. *The Nature of the Chemical Bond*, 2nd Ed. (Cornell University Press, New York, 1945)
25. G. Busch. *Suppl. Nuovo Cimento* Ser. X, **7**, 696 (1958)
26. S. Geller and J. H. Wernick. *Acta cryst., Camb.* **12**, 46 (1959)
27. H. Winston. *Phys. Rev.* **94**, 328 (1954)
28. P. Fielding, G. Fischer, and E. Mooser. *J. Phys. Chem. Solids* **8**, 434 (1959)
29. W. E. Spicer and A. H. Sommer. *J. Phys. Chem. Solids* **8**, 437 (1959)
30. P. Junod, E. Mooser, and H. Schade. *Helv. phys. acta* **29**, 193 (1956)
31. C. A. Coulson. *Valence* (Clarendon Press, Oxford, 1952)
32. F. Hund. *Z. Phys.* **73**, 1 (1931)
33. R. Gaspar. *Acta physiol. hung.* **7**, 289 (1957)
34. J. R. Reitz. *Phys. Rev.* **105**, 1233 (1957)
35. E. Mooser and W. B. Pearson. *Canad. J. Phys.* **34**, 1369 (1956)
36. E. Mooser and W. B. Pearson. *J. Phys. Chem. Solids* **7**, 65 (1958)
37. Ph. Choquard, A. Janner, and E. Ascher. *J. Phys. Chem. Solids* **8**, 366 (1959)
38. A. F. Wells. *Solid State Physics* **7**, p. 425 (Academic, New York, 1958)
39. E. Mooser and W. B. Pearson. *J. Chem. Phys.* **26**, 893 (1957)
40. G. Fischer and W. B. Pearson. *Proceedings of the Symposium on the Role of Solid State Phenomena in Electric Circuits*, p. 133 (Polytechnic Institute of Brooklyn, Brooklyn, New York, 1957)
41. J. R. Drabble and C. H. L. Goodman. *J. Phys. Chem. Solids* **5**, 142 (1958)
42. W. B. Pearson. *Canad. J. Phys.* **35**, 886 (1957)
43. F. J. Morin. *Semiconductors*, p. 600 (Ed. N. B. Hannay) (Reinhold, New York, 1959)
44. E. Mooser and W. B. Pearson. *Acta cryst., Camb.* **12**, 1015 (1959)
45. F. Laves. *Theory of Alloy Phases*, p. 124 (American Society for Metals, Cleveland, Ohio, 1956)
46. U. Dehlinger. *Theoretische Metallkunde* (Springer, Berlin, 1955)
47. W. Gordy and W. J. O. Thomas. *J. Chem. Phys.* **24**, 439 (1956)
48. R. E. Schlier and H. E. Farnsworth. *Semiconductor Surface Physics*, p. 3 (Univ. of Penna. Press, 1956)
49. H. E. Farnsworth and R. E. Schlier. *Bull. Amer. phys. Soc.* Ser. II, **3**, 30 (1958)
50. W. B. Pearson and G. A. Wolff. *Disc. Farad. Soc.* In press
51. A. F. Joffé. *J. Phys. Chem. Solids* **8**, 6 (1959)
52. W. Shockley. *Electrons and Holes in Semiconductors* (Van Nostrand, New York, 1950)
53. W. Meyer and H. Neldel. *Z. tech. Phys.* **18**, 588 (1937)
54. G. Busch. *Z. angew. Math. Phys.* **1**, 3 (1950)
55. G. Busch. *Helv. phys. acta* **19**, 189 (1946)
56. J. H. Gisolf. *Ann. Phys.* **1**, 3 (1947)
57. T. S. Moss. *Photoconductivity in the Elements* (Butterworths, London, 1952)
58. T. Miyauchi. *J. phys. Soc. Japan* **12**, 308 (1957)
59. B. Seraphin. *Z. Naturf.* **9a**, 450 (1954)
60. W. Heywang and B. Seraphin. *Z. Naturf.* **11a**, 425 (1956)
61. A. J. Gubanov. *J. tech. Phys., Moscow* **26**, 2170 (1956)
62. J. Adavi. *Phys. Rev.* **105**, 789 (1957)
63. L. Pauling. *J. Phys. Chem.* **58**, 662 (1954)
64. C. H. L. Goodman. *J. Electron.* **1**, 115 (1955)
65. W. Ruppel, A. Rose, and H. J. Gerritsen. *Helv. phys. acta* **30**, 238 (1957)
66. V. P. Zhuze. *J. tech. Phys., Moscow* **25**, 2079 (1955)
67. B. F. Ormont. *C. R. Acad. Sci. U.R.S.S.* **124**, 129 (1959)
68. O. G. Folberth. *Z. Naturf.* **13a**, 856 (1958)
69. O. G. Folberth and H. Welker. *J. Phys. Chem. Solids* **8**, 14 (1959)
70. W. B. Pearson. *Canad. J. Chem.* **37**, 1191 (1959)
71. A. J. Cornish. *J. electrochem. Soc.* **106**, 685 (1959)
72. F. Herman. *J. Phys. Chem. Solids* **8**, 20 (1959)
73. J. P. Suchet. *J. Phys. Chem. Solids* **12**, 74 (1959)
74. F. Hulliger. *Helv. phys. acta* **32**, 615 (1959)

THERMAL CONDUCTIVITY OF SEMICONDUCTORS

J. APPEL, Dr. rer. nat.

Osram Studiengesellschaft, Augsburg, Germany

At present

General Atomic, Division of General Dynamics Corporation, John Jay Hopkins Laboratory for Pure and Applied Science, San Diego, Cal., U.S.A.

MS. received July 1959

THERMAL CONDUCTIVITY OF SEMICONDUCTORS

1. INTRODUCTION

Thermal conduction belongs to the equalization processes. It arranges the equalization of energy, therefore thermal conduction means the same as energy transport which occurs when a temperature difference is applied between two points of a crystal. In this paper we want to get inside the physical elemental processes concerned with heat conduction in semiconductors and will formulate the corresponding mathematical principles for the entirety of these processes.

In a homogeneous isotropic crystal the heat current flows in the direction of the temperature gradient which should be small:

$$\mathbf{w} = -\chi \operatorname{grad} T \qquad \ldots (1)$$

The proportionality factor χ is called thermal conductivity. Its dimensions are given by the international system of measuring (W cm^{-1} deg.K^{-1}).

The phenomenological theory of equalization processes and, therefore, also that of thermal conduction belongs to the intricate thermodynamics of irreversible processes. Therefore, in thermal conduction as well as in all transport processes, the time derivate of entropy produced within the system by irreversible processes is of great importance. Onsager[1] was the first to recognize this fact. Some ideas of his theory resulting in symmetry relations between transport coefficients, named after him, will be dealt with in Section 2. Furthermore, we will examine the influence of crystal symmetry on the conductivity tensor. In Section 3 the first step is taken for treating the problem in the framework of statistical mechanics. By disregarding all interactions the eigenvalues of those Hamilton operators are determined which are obtained for the motion of electrons and lattice particles, i.e. atoms, ions, nuclei, etc. After introducing interactions the stationary non-equilibrium states are examined. The transport coefficients can be formulated by means of known distribution functions of the perturbed state. This is stated in Sections 4–8 for various interactions. According to Krumhansl,[2] the transport processes may be subdivided into two groups. With the first one, low excitation energies are of importance: electrons, phonons, photons. With the second one, higher excitation energies are necessary: free electron–hole pairs and excited electron states, called excitons. Finally, in Section 9 a practical application of thermoelectric effects is pointed out wherein thermal conduction is of some importance.

2. PHENOMENOLOGICAL THEORY

2.1. On the Theory of Onsager

The irreversible transport phenomena may be represented to a first approximation by phenomenological equations of the general form

$$J_i = \sum_{k=1}^{n} L_{ik} X_k \quad (i = 1, 2, \ldots, n) \qquad \ldots (2)$$

These linear equations define in a suitable manner the generalized 'forces' X_k and the conjugated generalized 'currents' J_i. The L_{ik} are the phenomenological transport coefficients. When regarding L_{ik} as components of a second-order tensor, then its diagonal elements according to

$$J_i = L_{ii} X_i \quad (i = 1, 2, \ldots, n) \qquad \ldots (3)$$

are the usual transport coefficients. These are, for example, the diagonal elements of conductivity tensors. They specify *one* current component caused by *one* force component. When two or more transport processes of the same or of a different type take place *simultaneously*, cross-phenomena occur and non-diagonal elements of equation (2) are not equal to zero. In any case, with a proper choice of currents and flows, the symmetry relations

$$L_{ik} = L_{ki} \quad (i, k = 1, 2, \ldots, n) \qquad \ldots (4)$$

are satisfied by the L_{ik} values. The L_{ik} values form a symmetric tensor which is, according to equation (5), positively definite. The symmetry relations (4) express the interrelations between transport coefficients on simultaneous occurrences of several irreversible transport processes.

By the proper choice[3] of currents and forces the definition of these conjugated variables is to be understood in such a way that the 'internal' entropy production rate can be written in the simple form

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \sum_{i=1}^{n} \dot{a}_i \frac{\partial S}{\partial a_i} = \sum_{i=1}^{n} J_i X_i > 0 \qquad \ldots (5)$$

The time variation of a thermodynamic state variable a_i is the current or flow J_i, the conjugated force X_i of which is $\partial S/\partial a_i$. A general proof of Onsager's theorem (4) can be found in a book by de Groot.[3] It is given by the theory of fluctuation phenomena, having microscopic reversibility of elemental processes as its fundamental assumption. This assumption is justified[4] in that the fundamental equations of mechanics are invariant with regard to time inversion. Therefore, in thermal equilibrium each elemental process takes place as many times as the reverse. In quantum theory this means that the transition probabilities are reversible. This results quite generally from the hermiticity of Hamiltonians.

A fundamental difficulty is to understand from the time reversible equations of motion, with which one is concerned in mechanics and quantum mechanics, the irreversibility of macroscopic transport processes, e.g. thermal conduction. These difficulties have been examined thoroughly in recent years by several authors.[5]

2.2. One of Onsager's Reciprocal Relations

By way of example a short derivation of the relations between two thermoelectrical effects shall be given, the Seebeck effect on the one hand, and the Peltier effect on the other. The latter is of practical importance for thermoelectric cooling (see Section 9). The time derivative of entropy for a two-conductor circuit, which has in one of its arms either an electric voltmeter or a voltage source, has been calculated by de Groot. As shown in Figure 1, the soldered junctions 1 and 2 and

every point of both conductors is in contact with a heat reservoir of large heat capacity. With the equation of Gibbs

$$T\,\mathrm{d}S = \mathrm{d}U - \rho\,\mathrm{d}q \qquad \ldots(6)$$

where ρ = electrical potential,
q = electric charge,
U = inner energy,

the author finds for the entropy production of the total system

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\mathrm{d}U}{\mathrm{d}t}\frac{\Delta T}{T^2} - \frac{\mathrm{d}q}{\mathrm{d}t}\frac{\Delta\rho}{T} \qquad \ldots(7)$$

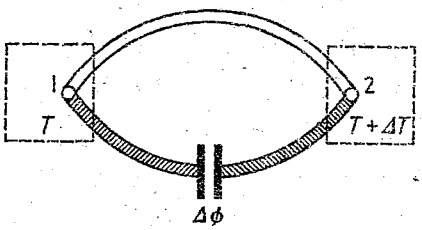


Figure 1. *Seebeck effect: Temperature difference ΔT generates a thermoelectric current and if the two-conductor circuit is interrupted, the thermoelectric voltage $\Delta\phi$ occurs. Peltier effect: A battery of voltage $\Delta\phi$ generates a direct current leading to a temperature difference ΔT between the two soldered junctions (for definition of signs see Lautz[69])*

This expression, involving products of the currents $\mathrm{d}U/\mathrm{d}t$, $\mathrm{d}q/\mathrm{d}t$ and the conjugated forces, is a specific case of equation (5). Accordingly, applying equation (2),

$$J_1 = -L_{11}\frac{\Delta\rho}{T} - L_{12}\frac{\Delta T}{T^2} \qquad \ldots(8a)$$

$$J_2 = -L_{21}\frac{\Delta\rho}{T} - L_{22}\frac{\Delta T}{T^2} \qquad \ldots(8b)$$

In order to understand the physical statement of the Onsager relation $L_{12} = L_{21}$ two cases shall be examined. The first is

$$\left.\begin{aligned}\Delta T &\neq 0\\ J_1 &= 0\end{aligned}\right\}\ \frac{\Delta\rho}{\Delta T} = -\frac{L_{12}}{L_{11}}T \qquad \ldots(9)$$

$\Delta\rho/\Delta T$ represents the differential thermoelectric power Q of the system. Equation (9) describes the Seebeck effect. The second case is

$$\left.\begin{aligned}\Delta T &= 0\\ J_1 &\neq 0\end{aligned}\right\}\ \frac{J_2}{J_1} = \frac{L_{21}}{L_{11}} = \Pi \qquad \ldots(10)$$

Π represents the Peltier coefficient; it measures the heat produced or destroyed on passage of current at a soldered junction per unit charge. An Onsager relation between the Seebeck and Peltier effect is therefore given by the equation

$$L_{21} = L_{12}, \qquad Q = \frac{\Delta\rho}{\Delta T} = -\frac{\Pi}{T} \qquad \ldots(11)$$

This is the first equation of Thomson. It shows that both effects are energetic reversals of one and the same process.

On the foundation of statistical mechanics Sondheimer[6] proved exactly the relation (11) for the coupled system electrons-lattice.

2.3. Anisotropic Crystals

For thermal conduction in anisotropic crystals, equation (1) must be generalized:

$$w_i = -\sum_{k=1}^{3} \chi_{ik} \frac{\partial T}{\partial x_k} \quad (i = 1, 2, 3) \qquad \ldots(12)$$

The matrix (χ_{ik}) of this equation system is called a tensor. Below we want to examine what relations result between the nine tensor components with regard to

(1) Onsager's theorem;
(2) The crystal symmetry.

(1) According to equation (12) the thermal conduction may be regarded formally as a superposition of three irreversible processes, each being assigned to a coordinate axes direction. This seems to be a specific case of the equation system (2) with $L_{ik} = \chi_{ik}$. In fact, an examination by Casimir[7] reveals that this is not exactly the case; rather the Onsager relations lead to the expressions

$$\frac{1}{2}\sum_{k=1}^{3}\left(\frac{\partial \chi_{ik}}{\partial x_k} - \frac{\partial \chi_{ki}}{\partial x_k}\right) = 0 \quad (i = 1, 2, 3) \qquad \ldots(13)$$

The physical statement of equation (13) is that antisymmetric components of the thermal conductivity tensor do not add to the divergence of heat flow, div w, and therefore cannot be observed. Therefore, it is permitted to equate the antisymmetrical components to zero. Then applies

$$\chi_{ik} = \chi_{ki} \quad (i, k = 1, 2, 3) \qquad \ldots(14)$$

This equation also results from equation (13) provided it is agreed that each component χ_{ik} disappears in vacuum and the value of χ_{ik} in the crystal is independent of the shape of the specimen. When inner or outer surfaces are important for the transport mechanism, equation (14) is not valid any more because the assumption of microscopic reversibility for scattering processes on surfaces is *not* warranted.

To deduce equation (13) Casimir starts from the fact that heat flow components w_i in anisotropic crystals are *not* time derivatives of thermodynamic state variables a_i as is the case for the heat current $|\mathbf{w}|$ in an isotropic conductor. Besides, the heat current in an anisotropic crystal cannot be defined uniquely, because only the divergence of the heat flow

$$\operatorname{div} \mathbf{w} = \frac{\partial w_1}{\partial x_1} + \frac{\partial w_2}{\partial x_2} + \frac{\partial w_3}{\partial x_3} \qquad \ldots(I)$$

is an observable quantity. For this reason the heat flow w remains unchanged when adding to the χ_{ik} values a term p_{ik}, so long as the corresponding divergence

$$\sum_{i,k} \frac{\partial}{\partial x_i}\left(p_{ik}\frac{\partial T}{\partial x_k}\right) = 0 \qquad \ldots \text{(II)}$$

This equation is always satisfied if

$$p_{ik} = -p_{ki} \quad \text{and} \quad \sum_i \frac{\partial p_{ik}}{\partial x_i} = 0 \qquad \ldots \text{(III)}$$

The addition of an antisymmetric tensor to the χ_{ik} has no physical meaning, provided equation (II) can be applied. Accordingly, the Onsager relation must not necessarily yield $\chi_{ik} = \chi_{ki}$. Casimir's paper explains more closely how to find the exact symmetry relations (13) when starting from local entropy production caused by temperature fluctuations in a suitable definition of generalized forces and currents.

(2) According to equation (14) the thermal conductivity tensor is symmetric. The Onsager relations reduce the number of independent components from nine to six. Also, generally, these six components are not independent of each other. Because of the crystal symmetry,[9] linear relations between them may exist. The crystal symmetry is described by non-trivial symmetry operations of the lattice. These are transformations arranging the crystal because of its symmetry into an identical position to a fixed co-ordinate system.

The symmetry of each perfect crystal is characterized by a group of symmetry operations of the lattice. These are motions bringing a crystal or an unbounded space lattice to an identical position with regard to a fixed co-ordinate system. The entirety of symmetry operations is called space group R of the crystal in question. The elements of the space groups, the symmetry operations, are apart from the primitive translations, proper and improper rotations (the latter include the ordinary reflections as well as the inversion), and operations due to screw axes and glide planes. The two last symmetry elements, which correspond to non-primitive translations followed by a proper or improper rotation, are omitted in the so-called simple or symmorphic space groups, so that these consist of a translation group T and a crystal point group P as sub-groups. The translations permute with the elements of the space group, therefore forming an invariant sub-group. This applies also to non-symmorphic space groups, as, for example, those being assigned to the diamond lattice.

The operations to be considered for determining the independent tensor components are those which leave invariant any point of the crystal. These operations form the 'site group' L.[8] By means of group theory this is defined as follows:

If R is the finite space group of a finite crystal and T the group of inherent primitive translations with one-dimensional representations $\exp(2\pi i\mathbf{kt})$, k therein being a vector in the reciprocal lattice (Section 3) and t a translation of T, then only the factor group F from R to T is of major interest. This one is isomorphic to one of the 32 crystal point groups. Generally, the site group is a sub-group of the factor group. For $\mathbf{k} = 0$ the factor group and site group are identical. In this case, being of interest to us, three-dimensional matrices correspond to the site group (= point group of R), these matrices being defined by *homogeneous* co-ordinate transformations.

Out of the symmetry operations of the lattice only those which leave invariant at least one point of the crystal are of interest here. The total of these operations forms the so-called site group.[9] The corresponding motions are represented by orthogonal transformations

$$x_i' = \sum_{k=1}^{3} a_{ik} x_k \quad (i = 1, 2, 3) \qquad \ldots (15)$$

With equation (15) the component w_i of the polar heat current vector turns to

$$w_i' = \sum_{k=1}^{3} \chi_{ik}' \frac{\partial T}{\partial x_k'} \quad (i = 1, 2, 3) \qquad \ldots (16)$$

Since equation (15) describes a site transformation, $w_i = w_i'$ and $\chi_{ik} = \chi_{ik}'$. With this in mind and with the help of the known transformation law for tensor components, the linear relations searched for are easily found. In Table 1, following Meissner

Table 1

Crystal classes (point groups)	Tensor components
C_1, C_1^i	$\chi_{11}, \chi_{22}, \chi_{33}, \chi_{12}, \chi_{13}, \chi_{23}$
C_2, C_2^h, C_s	$\chi_{11}, \chi_{22}, \chi_{33}, \chi_{12}$
C_3, C_3^i C_4, C_4^h, S_4 C_6, C_6^h C_3^h	$\chi_{11}, \chi_{22} = \chi_{11}, \chi_{33}, \chi_{12}$
D_2, D_2^h, C_2^v	$\chi_{11}, \chi_{22}, \chi_{33}$
D_3, D_3^d, C_3^v D_4, D_4^h, C_4^v, D_2^d D_6, D_6^h, C_6^v D_3^h	$\chi_{11}, \chi_{22} = \chi_{11}, \chi_{33}$
O, O^h, T, T^d, T^h	$\chi_{11}, \chi_{22} = \chi_{11}, \chi_{33} = \chi_{11}$

and Kohler,[10] the tensor components different from zero are summarized for each crystal point group. The point groups are designated according to Schoenfliess.

There is still another method which permits us to define independent tensor components in a simple manner. This method is based upon group theory and therefore has the advantage of leading in a simple and distinct manner to the determination of the independent tensor components, even with tensors of a higher order, which occur, for example, on examination of thermal conduction in a magnetic field. For the sake of simplicity the mathematical method is explained on second-order tensors which combine, as in the case of thermal conduction, two polar vectors.

A prerequisite for the following is the representation theory of rotation groups and of crystal point groups.[11] The latter are:

(1) Proper rotation groups;
(2) Improper rotation groups (= rotary reflection groups);
(3) Improper rotation groups not containing the inversion itself.

The matrices of the irreducible representations of an interesting point group are well known. Müller,[12] for example, has given them explicitly for cubic and hexagonal crystal classes. Now it is easy to see that the transformation matrices $(t_{lm,ik})$, being defined in the following manner

$$\chi'_{lm} = \sum_{i,k=1}^{3} t_{lm,ik}\chi_{ik} \quad (l,m = 1,2,3) \qquad \ldots(17)$$

and connecting the components of thermal conductivity tensors in equivalent co-ordinate systems, must be elements of a many-dimensional representation Γ of the crystal point group (= crystal class). Therefore, Γ can be reduced into irreducible representations of the point group, the dimensions of which (for cubic crystal systems) are at most three. It results therefrom that certain linear combinations of one, two, or three tensor components in all symmetry operations of a point group mutually transform. These linear combinations are the centre of interest as shown by Juretschke.[13] They transform according to a representation Γ'. This is a completely reduced set of matrices resulting from the reduction of the many-dimensional representation Γ. Symbolically this is written

$$\Gamma' = A^{-1}\Gamma A = \sum_i n_i\Gamma_i \qquad \ldots(18)$$

In symmetry operations the values of the linear combinations are not allowed to vary, neither are those of the individual tensor components, because the transformation matrices of the individual components and of the linear combinations emerge from each other by similarity transformations. *Two* kinds of linear combinations are to be distinguished:

(1) Those of which the value is different from zero. These are the actual invariants of the tensor. Their number is equal to the number of independent tensor components.
(2) Those of which the value is equal to zero. They furnish linear relations between the components. These are necessary in order to determine *which* tensor components are different from each other for a given crystal class.

The number of independent tensor components is easily found by taking into account that the linear combinations (1) transform according to the one-dimensional representations of the point group. How many one-dimensional representations are contained in the reducible representation Γ results from the well-known counting-off formula of Wigner

$$n_i = \frac{1}{h}\sum_{\mathrm{G}} h(\mathrm{C})\kappa(\mathrm{C})\kappa_i(\mathrm{C}) \qquad \ldots(19)$$

n_i states how many times the representation Γ_i is contained in the reducible representation Γ, h = number of group elements, $h(\mathrm{C})$ = number of elements in class C, $\kappa(\mathrm{C})$ = character of C in the representation Γ, $\kappa_i(\mathrm{C})$ = character of C in

the representation Γ'. A paper of Müller[12] states by means of a crystallographic decomposition table for all the 32 crystal classes how often irreducible representations of the same dimension are contained in the reducible representations of tensors of arbitrary order. For the number of one-dimensional representations within the three different above-mentioned kinds of crystal classes special formulae are given. For calculation of characters $\kappa(C)$ reference is made to Juretschke,[13] who also shows how to determine the linear combinations (1) and (2).

For this purpose Γ is decomposed into the irreducible representations $\mathbf{D}^{(l)}$ of the rotation group. In case the point group contains rotations followed by inversions, the place of the rotation group is taken by the rotary reflection group. It is the direct product of the rotation group and the group consisting of the unit E and the inversion I:

$$E = \begin{pmatrix} 1 & 0 & 0 & . & . \\ 0 & 1 & 0 & . & . \\ 0 & 0 & 1 & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix} \qquad I = \begin{pmatrix} -1 & 0 & 0 & . & . \\ 0 & -1 & 0 & . & . \\ 0 & 0 & -1 & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix} \qquad \ldots (IV)$$

The rotatory deflection group therefore has two irreducible representations $\mathbf{D}^{(l+)}$ and $\mathbf{D}^{(l-)}$. In the representation $\mathbf{D}^{(l+)}$ the matrix $\mathbf{D}^{(l)}(\phi)$ also corresponds to rotatory deflection $\bar{\phi}$, in the representation $\mathbf{D}^{(l-)}$ the matrix $-\mathbf{D}^{(l)}(\phi)$ is assigned to the rotatory deflection. The characters are

$$\begin{aligned} \kappa_{l+}(\phi) &= \kappa_l(\phi); \qquad \kappa_{l-}(\phi) = \kappa_l(\phi) \\ \kappa_{l+}(\bar{\phi}) &= \kappa_l(\phi); \qquad \kappa_{l-}(\bar{\phi}) = -\kappa_l(\phi) \end{aligned} \qquad \ldots (V)$$

The irreducible decomposition of the representation of a second-order tensor in irreducible representations of the rotatory deflection group results in

$$\begin{aligned} \kappa(\Gamma) &= \kappa_{3\times 3}(\phi) = \kappa_1(\phi) + \kappa_3(\phi) + \kappa_5(\phi) \\ \kappa(\Gamma_s) &= \kappa_{(3\times 3)_s}(\phi) = \kappa_1(\phi) + \kappa_5(\phi) \end{aligned} \qquad \ldots (VI)$$

On the right-hand side, the characters κ_{l+} are to be replaced by characters κ_{l-} provided Γ and Γ_s are representations of polar second-order tensors.[13] With axial tensors $\kappa_l = \kappa_{l+}$; $\kappa_{3\times 3}$ is the character of the product representation of two three-dimensional rotation groups. s indicates the symmetric direct product; the character of its representation has been calculated by Tisza.[14]

Table 2

Representation of a sub-space	Linear combinations transforming according to representation $\mathbf{D}^{(l)}$
$\mathbf{D}^{(0)}$	$\frac{1}{3}(\chi_{11}+\chi_{22}+\chi_{33})$
$\mathbf{D}^{(1)}$	$\frac{1}{\sqrt{2}}(\chi_{23}-\chi_{32}), \frac{1}{\sqrt{2}}(\chi_{13}-\chi_{31}), \frac{1}{\sqrt{2}}(\chi_{12}-\chi_{21})$
$\mathbf{D}^{(2)}$	$\frac{1}{\sqrt{6}}(2\chi_{33}-\chi_{22}-\chi_{11}), \frac{1}{\sqrt{2}}(\chi_{22}-\chi_{11})$ $\frac{1}{\sqrt{2}}(\chi_{13}+\chi_{31}), \frac{1}{\sqrt{2}}(\chi_{12}+\chi_{21}), \frac{1}{\sqrt{2}}(\chi_{23}+\chi_{32})$

A result of the decomposition shows the number of sub-spaces and their dimension. This dimension is equal to the number of linear combinations belonging to the sub-space. Juretschke has shown how to find all the linear combinations starting from the transformation properties given by the representation matrices belonging to one sub-space: $\mathbf{D}^{(0\pm)}$ transforms like a scalar and a pseudo-scalar, respectively; $\mathbf{D}^{(1\pm)}$ like an axial and a polar vector, respectively; etc. The requirement according to which linear combinations belonging to different sub-spaces are mutually orthogonal, resulting from the fact that linear combinations of a sub-space transform among themselves, helps to determine these combinations. The result for the thermal conductivity is summarized in Table 2.

The number of linear combinations in Table 2 belonging to equations (1) or (2) can be stated using equation (19). The last step is to determine which of the linear combinations belong to (1) and which to (2). For this purpose by means of the transformation matrices of a sub-space it is determined whether or not a linear combination with all symmetry operations is invariant.

3. STATIONARY STATES AND THEIR OCCUPATION

Prior to formulating in detail the theory of thermal conduction, the stationary states of the system consisting of electrons and lattice particles must be known. Therefore, first of all the eigenvalues of the Hamiltonian of the total system must be found, on the presumption of the usual approximations, especially by neglecting the 'dynamical' interactions between both partial systems, electrons and lattice particles. Further, it is supposed that interactions between the quasi-particles of the partial systems (electrons; lattice oscillators or phonons) may be neglected. An eigenvalue of the Hamiltonian may then be represented by the sum of the eigenvalues of one-particle Hamiltonians (electrons, oscillators; first quantization) or by two sums of the products of quasi-particle energies and corresponding occupation numbers (electrons, phonons; second quantization).

3.1. Adiabatic Approximation

An entire system is regarded as consisting of a fundamental region (unit volume) with N lattice particles of a monatomic lattice and with n electrons. The Schrödinger equation for the eigenfunctions $\Psi(\mathbf{R},\mathbf{r})$, with $\mathbf{R}$ as collective symbol for the position vectors $\mathbf{R}_j (j = 1,2,\ldots,N)$ of the lattice particles and $\mathbf{r}$ as collective symbol for the position vectors $\mathbf{r}_i (i = 1,2,\ldots,n)$ of the electrons, reads

$$\mathscr{H}\Psi = i\hbar\frac{\partial\Psi}{\partial t} \qquad \ldots (20)$$

The Hamiltonian $\mathscr{H}$ of the entire system in a non-polar crystal comprises the kinetic energy of electrons, their coulomb interaction with the lattice particles, their mutual Coulomb interaction, and the Hamiltonian of the lattice $\mathscr{H}_{\text{latt}}$:

$$\mathscr{H} = \sum_i \frac{p_i^2}{2m} + \sum_{i,j} v(\mathbf{r}_i - \mathbf{R}_j) + \tfrac{1}{2}\sum_{i,k}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_k|} + \mathscr{H}_{\text{latt}} \qquad \ldots (21)$$

The Hamiltonian of the lattice consists of kinetic and potential energy of the lattice particles and their interactions when in equilibrium positions:

$$\mathscr{H}_{\text{latt}} = \sum_j \frac{P_i^2}{2M} + \mathscr{H}_{\text{latt}}(\mathbf{R}-\mathbf{R}^{(0)}) + \mathscr{H}_{\text{latt}}(\mathbf{R}^{(0)}) \qquad \ldots (22)$$

In polar crystals the Hamiltonian $\mathscr{H}$ of the entire system is to be supplemented by various polarization potentials and more particularly by an interaction potential which is not lattice periodic.[15]

To $\mathscr{H}_{\text{latt}}$ is added a polarization (a) consisting of dipole moments corresponding to the displacement of ions. In addition, on displacement the shell electrons are polarized (b). To the electron–lattice interaction potential a lattice periodic part is to be added which takes into account the polarization of electron shells by free electrons (c). Finally, a non-lattice periodic part is added to the interaction potential describing the electrostatic interaction between free electrons and inert polarization potentials (a, b).

For the adiabatic approximation it is characteristic that the eigenfunction of the entire system is given by the product

$$\Psi(\mathbf{R}, \mathbf{r}) = \kappa(\mathbf{R})\Psi(\mathbf{R}, \mathbf{r}) \exp\left(-\frac{i}{\hbar}\epsilon t\right) \qquad \ldots (23)$$

This total eigenfunction is easy to interpret. Apart from the time dependent exponential function the first factor describes the motion of the lattice particles, the second one that of the electrons, if the position of the lattice particles is fixed by $\mathbf{R}$. Therefore, with the total wave function (23) it is supposed that the electrons adiabatically follow the thermal motion of the lattice particles. This is implied by the small mass ratio m/M. ϵ is the total energy in the adiabatic approximation. With (23) an equation is obtained for the eigenfunctions Ψ of the electron ensemble with fixed nuclear positions, and a second equation for the eigenfunction κ of the lattice in which the eigenvalues of the electron equation appear as potential energy. The expression (23) is not the exact solution of (20), but represents the solution of an essentially simplified equation. The neglections caused by the adiabatic approximation have been examined thoroughly by several authors,[16] especially by Haug and Sauermann. It results, according to the theoretical point of view, that up to now there is no exact justification for the adiabatic approximation, though it is the foundation for successful calculations on stationary states and transport phenomena, etc.

The main result of the paper by Haug and Sauermann is that the Born–Oppenheimer principle is based upon a wrong classification of the operator of kinetic energy of the lattice particles. The kinetic energy of the lattice particles is of the order of magnitude $(m/M)^{1/4}$ compared with the kinetic energy of the electrons (order of magnitude 1) and not, as supposed by Born and Oppenheimer, in the order of magnitude $(m/M)^{1/2}$. For this reason the adiabatic approximation does not remain consistent in itself, when the *neglected* terms involved in this approximation are regarded as perturbation.

It is important that the two Schrödinger equations obtained by the adiabatic approximation for the electron ensemble and the lattice decompose into two systems of one-particle Schrödinger equations with simple supplemented suppositions.

3.2. Electrons

Considering crystal electrons being dominant for the conduction mechanism (free or valence electrons) the neglect of exchange and of deflection or correlation energy[18] leads to a Hartree representation of the wave function

$$\psi = \prod_{n,\mathbf{k}} \psi_{n\mathbf{k}}(\mathbf{r}) \qquad \ldots (24)$$

The one-electron eigenfunctions $\psi_{n\mathbf{k}}$ are solutions of the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\Delta+V(\mathbf{r})\right]\psi_{n\mathbf{k}}(\mathbf{r})=\epsilon_n(\mathbf{k})\psi_{n\mathbf{k}}(\mathbf{r}) \qquad \ldots(25)$$

For the stationary energy states $\epsilon_n(\mathbf{k})$ of the electrons which are defined by the two quantum numbers n and $\mathbf{k}$ disregarding the spin, it is agreed that:

(1) In a lattice with the primitive translations $(\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3)$ the known periodicity conditions yield

$$\psi_{n\mathbf{k}}(\mathbf{r}+G\mathbf{t})=\psi_{n\mathbf{k}}(\mathbf{r}) \qquad \ldots(26)$$

G represents a large number ($G^3=N$) and

$$\mathbf{t}=n_1\mathbf{t}_1+n_2\mathbf{t}_2+n_3\mathbf{t}_3 \qquad \ldots(27)$$

is a translation vector (n_i = integral number).

(2) $\mathbf{k}$ is a reduced wave vector:

$$\mathbf{k}=\frac{2\pi}{G}(m_1\mathbf{b}_1+m_2\mathbf{b}_2+m_3\mathbf{b}_3), \quad -\frac{G}{2}<m_i\leqslant+\frac{G}{2} \quad (i=1,2,3) \qquad \ldots(28)$$

The inverse vectors $\mathbf{b}_j$ define the reciprocal lattice:

$$\mathbf{t}_i\mathbf{b}_j=\delta_{ij} \qquad \ldots(29)$$

The G^3 wave vectors $\mathbf{k}$, each of which stands perpendicularly on a band of parallel planes in the co-ordinate space, end in the interior and on the surface of a polyhedron. This polyhedron, which is a consequence of the translation symmetry of the crystal, is called the first Brillouin zone.

In equation (25) $V(\mathbf{r})$ is still to be defined. This is the periodic potential of the lattice, if the lattice particles are thought to be in equilibrium positions, $\mathbf{R}_j=\mathbf{R}_j^{(0)}$ (static approximation). The 'average' Coulomb interaction between one electron, picked out of the ensemble of free electrons, and the remaining electrons, the charge of which is uniformly distributed over the whole crystal, is taken into consideration by means of a constant supplementary potential, so that

$$V(\mathbf{r})=\sum_j v(\mathbf{r}-\mathbf{R}_j^{(0)})+\text{const} \qquad \ldots(30)$$

The potential $V(\mathbf{r})$ takes on the symmetry of the lattice, so is invariant under symmetry operations. A sub-group of the space group which is always an invariant sub-group too, forms the group $\mathbf{T}$ of the translations. Thus, the lattice potential yields the periodicity relation

$$V(\mathbf{r}+\mathbf{t})=V(\mathbf{r}) \qquad \ldots(\text{VII})$$

The fact that $V(\mathbf{r})$ is invariant with respect to the primitive translations $\mathbf{t}$ permits as single solution of equation (25) the Bloch functions

$$\psi_{n\mathbf{k}}(\mathbf{r})=\exp(i\mathbf{k}\mathbf{r})u_{n\mathbf{k}}(\mathbf{r}) \qquad \ldots(31)$$

where $u_{n\mathbf{k}}(\mathbf{r})$ is a lattice periodic function.

The analytical form of $u_{n\mathbf{k}}$ is given by the wave vector $\mathbf{k}$ and the crystal symmetry. $u_{n\mathbf{k}}$ has the complete symmetry of the lattice at $\mathbf{k}=0$, the function being invariant in symmorphic lattices with respect to all symmetry operations of the crystal point group. Near the nuclei the functions $u_{n\mathbf{k}}$ pass over into atomic wave functions.

A fundamental problem of solid state physics[18] is the calculation of the stationary energy states of electrons belonging to equation (31). Some general results for monatomic lattices due to Hund[19] are:

(1) To each wave vector $\mathbf{k}$ belongs an indefinite number of eigenvalues. These are restricted to certain energy ranges ($n = 1,2,\ldots$) which may be separated by forbidden zones. Every atom—with multi-atomic lattices every unit cell—furnishes one state per band and two if the Pauli principle is taken into account.

(2) The allowed energy ranges decompose the reciprocal lattice into separated zones of equal volume which are to be reduced to the fundamental zone (= first Brillouin zone) according to the definition of $\mathbf{k}$. At distinct points and certain lines the energy surfaces in $\mathbf{k}$-space of different zones may coincide[20] (degeneration).

There are various reasons for the degeneracy of energy surfaces. On neglection of spin–orbit interaction energy, surfaces will coincide always at such points in the $\mathbf{k}$-space at which the group of the corresponding $\mathbf{k}$-vector has 'small representations' of higher dimension than one. A typical example is the coincidence of valence bands of semiconductors crystallizing in diamond or zinc blende lattice at $\mathbf{k} = 0$ (germanium, silicon, III–V compounds). The corresponding representation is three-dimensional. When taking into account the spin–orbit interaction it decomposes into a two-dimensional and a one-dimensional representation. Beside the spatial crystal symmetry in non-symmorphic lattices, being such whose symmetry operations contain operations given by screw axes and/or glide planes, a coincidence of energy surfaces may be caused by the fact that the Hamiltonian of the electrons is real and, therefore, invariant with regard to time inversions.

(3) In each reduced Brillouin zone there exist several energy extrema. In the vicinity of a non-degenerated extreme value at $\mathbf{k} = \mathbf{k}_0$ the energy $\epsilon_n(\mathbf{k})$ may be represented to a first approximation by a quadratic function

$$\epsilon_n(\mathbf{k}) = \epsilon_n(\mathbf{k}_0) + \sum_{i,j=1}^{3} c_{ij}^{(n)}(K_i - K_{0i})(K_j - K_{0j}) \qquad \ldots (32)$$

The type of coefficients $c_{ij}^{(n)}$ which may occur is fixed[21] here and also in the case of degenerate eigenvalues by the crystal symmetry and $\mathbf{k}_0$. The quantitative values of $c_{ij}^{(n)}$ may be defined empirically. If the energy extrema are not located at $\mathbf{k} = 0$, as in the case of the conduction bands of germanium, silicon, and the Group III–V compounds, then *a priori* calculation of the band structure is necessary in order to determine $\mathbf{k}_0$.

(4) In simple cases, for example, when in cubic lattices an energy extremum occurs at $\mathbf{k} = 0$ and is not degenerate, equation (32) may be rewritten in the form

$$\epsilon_n(\mathbf{k}) = \epsilon_n(0) + \frac{\hbar^2}{2m^*}|\mathbf{k}|^2 \qquad \ldots (33)$$

The electron moves like a free particle with the effective mass m^*. This isotropic approximation is often applied in transport problems. Of course, the physical insight obtained by applying this approximation is limited by the qualitative nature of the results.

The occupation probability or average occupation number without interactions is governed by the Fermi–Dirac statistics:

$$\mathfrak{f}_0 = \frac{1}{\exp[(\epsilon - \xi)/kT] + 1} \qquad \ldots (34)$$

The number of states in the volume element of the **k**-space is relative to the unit volume of the ordinary co-ordinate space

$$\frac{2}{(2\pi)^3}\,\mathbf{f}_0\,\mathrm{d}K_1\,\mathrm{d}K_2\,\mathrm{d}K_3 \qquad \ldots(35)$$

With extrinsic semiconductors it is permitted[23] to relate the energy to a band edge as the zero point of the energy scale, i.e. $\epsilon(0)=0$, ($\mathbf{k}_0=0$). In intrinsic semiconductors it is expedient to choose another zero-point because the position of the band edges is temperature-dependent and this has a direct influence on the heat current density.

3.3. Lattice Vibrations, Phonons

The Schrödinger equation for the thermal motion of the lattice particles yields:

$$[\mathscr{H}_{\mathrm{latt}}-\mathscr{H}_{\mathrm{latt}}(\mathbf{R}^{(0)})]\,\phi(\mathbf{R})=\epsilon_{\mathrm{latt}}\,\phi(\mathbf{R}) \qquad \ldots(36)$$

By introducing suitable variables the equation (36) may be separated into a system of eigenequations for normal modes. For this purpose the potential energy V of the total system defined in equation (22) is expanded in terms of displacements $\delta\mathbf{R}_j=\mathbf{R}_j-\mathbf{R}_j^{(0)}$ and terms of a higher than second order are disregarded:

$$V=V_2=\tfrac{1}{2}\sum_{j,k}\sum_{\mu,\nu=1}^{3}A_{jk}^{\mu\nu}\,\delta\mathbf{R}_j^{(\mu)}\,\delta\mathbf{R}_k^{(\nu)} \qquad \ldots(37)$$

The coupling parameter $A_{jk}^{\mu\nu}$ forms a second-order tensor with regard to the indices $(\mu\nu)$. On decomposition of the displacements in terms of progressing plane waves with propagation vector $\mathbf{q}$ and polarization s by the normal co-ordinates, $\xi(\mathbf{q},s)$ defined below is

$$\phi(\mathbf{R})=\prod_{\mathbf{q},s}\phi[\xi(\mathbf{q},s)] \qquad \ldots(38)$$

$\phi(\xi)$ is an eigenfunction of the Schrödinger equation of the harmonic oscillator:

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}\xi^2}+\frac{2M}{\hbar^2}(\epsilon-M\omega^2\xi^2)\,\phi=0 \qquad \ldots(39)$$

The energy eigenvalues are

$$\epsilon_n(\mathbf{q},s)=(n+\tfrac{1}{2})\hbar\omega(\mathbf{q},s) \qquad \ldots(40)$$

The orthonormal set of oscillator eigenfunctions is given by

$$\phi_n(\xi)=\left(\frac{\alpha}{\pi^{1/2}2^n n!}\right)^{1/2}\exp(-\tfrac{1}{2}\alpha^2\xi^2)\,H_n(\alpha\xi) \qquad \ldots(41)$$

where

$$\alpha^2=\frac{M\omega}{\hbar}$$

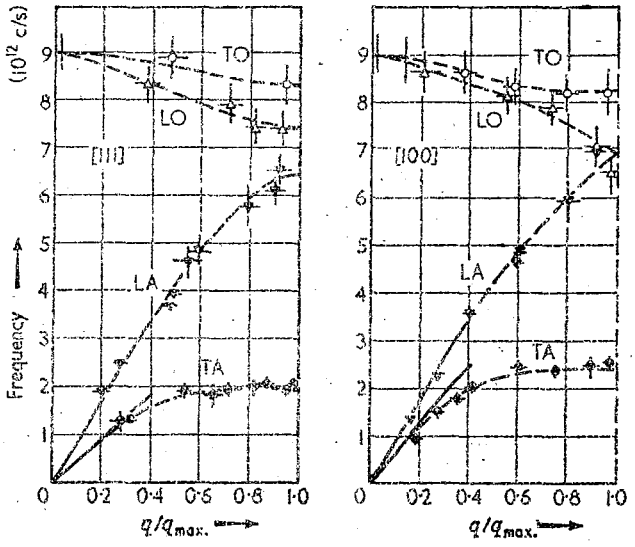
and $H_n=n$th Hermite polynominal. The unit of wave-excitation of a quantized lattice vibration is called a lattice quantum or *phonon* by analogy with the light

quantum or photon in electrodynamics. The propagation vectors **q** have the same geometrical properties as the **k**-vectors:

$$\mathbf{q} = \frac{2\pi}{G}(p_1\mathbf{b}_1+p_2\mathbf{b}_2+p_3\mathbf{b}_3), \quad -\frac{G}{2} < p_i \leqslant +\frac{G}{2}(i = 1,2,3) \quad \ldots(42)$$

The lattice dynamics yield three characteristic frequencies $\omega(\mathbf{q},s)$ to each of the G^3 propagation vectors and the three inherent polarization directions $\mathbf{n}(\mathbf{q},s)$, mutually orthogonal. Generally the co-ordinate axes of $\mathbf{n}(\mathbf{q},s=1,2,3)$ are situated arbitrarily with respect to the propagation direction. In order to keep the polarization of longitudinal waves steady at $\mathbf{q}=0$ one writes

$$\mathbf{n}(\mathbf{q},s) = -\mathbf{n}(-\mathbf{q},s) \quad \ldots(43)$$



(By courtesy of *Physics and Chemistry of Solids*)

Figure 2. *The dispersion curves for the* (100) *and* (111) *directions of germanium* (*according to Brockhouse and Iyengar*[26])

Only in certain distinct directions, which often are identical with main axes and are given by the crystal symmetry, may pure longitudinal and transverse waves exist.[25] For germanium, the dispersion curves $\omega(\mathbf{q},s)=\omega(-\mathbf{q},s)$ determined experimentally by Brockhouse[25] are represented in Figure 2 for two of three main axes of cubic systems. By this means it is possible to define the phase and group velocities

$$v_p = \frac{\omega(\mathbf{q},s)}{q} \quad \text{and} \quad v_g = \frac{\partial\omega(\mathbf{q},s)}{\partial q} \quad \ldots(44)$$

respectively. The latter is important for the energy transport and it is to be observed from Figure 2 that the group velocities of the optical lattice waves (LO and TO) are small compared with those for acoustic lattice waves (LA and TA),

especially with large wavelengths. The strong dispersion of the degenerate transverse acoustic branch in the (100) direction is remarkable. According to recent examinations by Ghose[26] et al.—the authors measured the dispersion in the (111) direction too—this dispersion is not quite so distinct.

In a diatomic lattice, e.g. the zinc blende lattice, there are three acoustical and three optical phonons for each q-vector lying in a non-distinct direction of the q-space. If n atoms are present in the unit cell then lattice dynamics furnish three acoustical and $3(n-1)$ optical branches of normal vibrations. The dispersion for the latter is always small. Furthermore, at low temperatures the number of high frequency optical modes is relatively small.

The average occupation number $N_0(\mathbf{q}, s)$, i.e. the average number of phonons of the energy $\hbar\omega(\mathbf{q}, s)$ is governed by the Bose–Einstein statistics.[27] By disregarding the phonon–phonon interaction this yields

$$N_0(\mathbf{q}, s) = \frac{1}{\exp\left[\hbar\omega(\mathbf{q}, s)/kT\right] - 1} \qquad \ldots (45)$$

Thus the average number of phonons in the volume element of the q-space relative to the volume unit of the ordinary co-ordinate space is

$$\frac{1}{(2\pi)^3} N_0 \, dq_1 \, dq_2 \, dq_3 \qquad \ldots (46)$$

With equations (40) and (46) we know the stationary oscillator states and their average occupation numbers which result from the first quantization of the classical oscillator equation of motion. The introduction of suitable real normal co-ordinates $\xi(\mathbf{q}, s)$ was the starting point for this representation of energy states of the lattice. These are defined according to Wilson[28] in the following manner:

$$\delta\mathbf{R}_j = \frac{1}{\sqrt{G^3}} \sum_{\mathbf{q}, s} \mathfrak{n}(\mathbf{q}, s) \left[\xi(\mathbf{q}, s) \cos \mathbf{q}\mathbf{R}_j - \frac{p(\mathbf{q}, s)}{M\omega(\mathbf{q}, s)} \sin \mathbf{q}\mathbf{R}_j \right] \qquad \ldots (47)$$

where the time dependence of the real normal co-ordinates is given by

$$\xi(\mathbf{q}, s) = \xi_0 \cos\left[\omega(\mathbf{q}, s) + \beta\right] \quad \text{and} \quad p(\mathbf{q}, s) = M\dot{\xi}(\mathbf{q}, s)$$

The term $\mathbf{q} = 0$ is always omitted because in the acoustical branch it corresponds to a translation of the whole lattice. The total energy ϵ_{latt} is, with regard to equation (47)

$$\epsilon_{\text{latt}} = \sum_{\mathbf{q}, s} \epsilon(\mathbf{q}, s) = \sum_{\mathbf{q}, s} \left[\frac{p^2(\mathbf{q}, s)}{2M} + M\omega^2(\mathbf{q}, s)\, \xi^2(\mathbf{q}, s) \right] \qquad \ldots (48)$$

The first quantization at once leads to equation (39). By means of solutions $\phi_n(\xi)$ the known matrix elements of the harmonic oscillator[29] are found.

4. THERMAL CONDUCTION IN DIELECTRICS

It is a well-known fact that—starting from the harmonic approximation—a dielectric would not allow a transport phenomenon which might be defined as thermal conduction. As shown in Section 3 this approximation describes the

motion of the lattice particles by plane waves. Therefore there is no interaction between these waves and they may propagate in the crystal without attenuation. If one regards the harmonic approximation as a limiting case one might say that the mean free path of phonons and therefore the thermal conductivity is infinitely large.[30] For this reason no thermal equilibrium can be established because, once a random distribution of energy ϵ_{latt} upon the individual oscillators exists, then this distribution is maintained at all times.

A finite thermal conductivity resulting from energy exchange of different plane waves cannot be obtained before taking into consideration the *anharmonic* coupling between the thermal motion of lattice particles. This is caused by terms of higher than second order in the expansion of the potential energy of the lattice particles in terms of displacements. The coupling enables the existence of stationary non-equilibrium states, leading to a finite thermal conductivity. The first preliminary statement for calculating the thermal conductivity within the framework of elastic continuum theory has been made by Debye.[30] He reflects that in an amorphous solid the thermal lattice waves are scattered on the constitutional density variations, as the light is in a cloudy medium. Spatial density variations in a crystal are caused by the anharmonicities in the thermal motion of the lattice particles. Debye calculates a respective mean free path and finds the well-known T^{-1} law. Then he shows that for calculation of thermal conductivity the formula[31]

$$\chi = \tfrac{1}{3} u c_v l \qquad \ldots (49)$$

may be applied. This formula is easily derivable and is known from gas kinetics. c_v (W sec cm^{-3} deg.K^{-1}) is the specific heat per volume unit, u is an average sound velocity.

The lattice theory of thermal conductivity goes back to Peierls.[32] The interaction of phonons is examined assuming that at no time phase relations between different oscillators, viz. 'single phonon' states, exist[33] and by applying Dirac's theory of time-proportional transition probabilities. The formulation of the transport problem then results from the condition that the distribution functions of the phonons are stationary under the influence of a temperature gradient and the phonon–phonon interaction. This leads to a system of three coupled Boltzmann equations for the perturbed distribution functions $N(\mathbf{q}, s = 1, 2, 3)$ assuming that at all times the state of the unperturbed system can be characterized by average occupation numbers obtained in the harmonic approximation.

The statistical fundamental equation

$$\frac{df}{dt} = 0 \qquad \ldots (VIII)$$

represents in a stationary non-equilibrium state the total time derivative of a distribution function f depending on the generalized co-ordinates and velocities of particles of a mechanical system. The fundamental equation is often spoken of as *Liouville's theorem*, since Liouville was the first who formulated (VIII) within the framework of classical statistical mechanics. For solving the fundamental equation in the kinetic gas theory the well-known 'Stoszahlenansatz' is made for the number of scattering processes per volume and time unit. Thus certain correlations between position and velocity of particles are disregarded. The 'Stoszahlenansatz' represents an average value on a large number of scattering processes, and therefore a statistical element is brought into the calculation as shown by Becker[5] in detail, leading to the *distinction* of a *time direction* (H-theorem).

In quantum theory the classical fundamental equation (VIII) is replaced by the quantum mechanical analogue of Liouville's theorem. This is the 'motion of equation' for the density matrix of an ensemble of quantum mechanical systems (R. C. Tolman. *The Principles of Statistical Mechanics* (Oxford University Press, London, 1950)). When starting from a system with the Hamiltonian

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1 \qquad \ldots (IX)$$

at which the solutions $u_n(q)$ of the unperturbed operator $\mathcal{H}_0$ are known as a set of orthonormal eigenfunctions to the eigenvalues $\epsilon_n^{(0)}$

$$\mathcal{H}_0 u_n(q) = \epsilon_n^{(0)} u_n(q) \qquad \ldots (X)$$

then the state of the perturbed system ($\mathcal{H}_1$ = small perturbation) may be represented by an expansion in terms of eigenfunctions of the unperturbed system

$$\psi(q,t) = \sum_k c_k(t)\, u_k(q) \exp\left(-\frac{2\pi i}{\hbar}\epsilon_k^{(0)} t\right) \qquad \ldots (XI)$$

The density matrix of the ensemble then consists of the components

$$\rho_{nm} = \overline{c_m^*(t)\, c_n(t)} \exp\left[-\frac{2\pi i}{\hbar}\left(\epsilon_n^{(0)} - \epsilon_m^{(0)}\right) t\right] \qquad \ldots (XII)$$

The average value, denoted by the double bar, is to be taken on all systems of the ensemble (e.g. on all electrons within the fundamental region of a crystal). The time derivative of the matrix components is given by

$$\frac{\partial \rho_{nm}}{\partial t} = -\frac{2\pi}{i\hbar}\sum_k (\rho_{nk}\mathcal{H}_{km} - \mathcal{H}_{nk}\rho_{km}), \qquad \mathcal{H}_{nk} = \int u_n^* \mathcal{H} u_k \, dq \qquad \ldots (XIII)$$

when applying the perturbation theory of Dirac. This equation corresponds to the classical fundamental equation (VIII). If at all times only diagonal elements in the density matrix are different from zero, then at no time do phase relations exist between the pure states of the system. Therefore the diagonal elements as mean occupation numbers of the pure states are equivalent to the distribution function if a proper normalization is applied. The assumption that at all times non-diagonal elements disappear is in accordance with the 'Stoszahlenansatz' of the kinetic gas theory. In both cases one arrives at the Boltzmann equation. Various authors have tried to prove the supposition of molecular disorder or the disregard of phase relations. Kümmel[5] shows that the supposition of molecular disorder is comprehensible by means of the irreversible statistics. This argument appears to state the essential point. Greenwood in the case of electric transport phenomena avoids the repeated neglection of phase relations by the assumption that the scattering centres for the electrons are randomly distributed. According to van Hove[6] certain characteristic properties of the perturbation may justify the neglect of phase relations as well. Adams[5] criticizes the papers of Greenwood, Luttinger, and Kohn,[5] and van Hove by proving that just the phase relations (= non-diagonal elements of the density matrix), which correspond to a certain kind of position velocity coherences of particles, are required for the validity of the equation (XIII) at all times. Therefore, according to Adams, just these coherences determine the time direction in which the irreversible process proceeds. Although the non-diagonal elements of the density matrix are of importance for the existence of a thermodynamic irreversible process and for the validity of the transport equation, these terms generally do not fit into the transport equation itself or into the expression for a current.

If the three functions $N(\mathbf{q}, s)$ are known, then the density of heat flow is given by

$$\mathbf{w} = \frac{1}{V}\sum_{\mathbf{q},s} N(\mathbf{q},s)\,\hbar\omega(\mathbf{q},s)\,\mathbf{u}_g(\mathbf{q},s) \qquad \ldots (50)$$

V is the periodicity volume with $N = G^3$ atoms. On calculation of the functions $N(\mathbf{q}, s)$ phonon scattering processes are taken into account which are induced by

the third-order terms in the development of the potential energy in terms of displacements. Generally, the fourth-order terms play only a small part in heat conduction, especially at low temperatures, $T \ll \theta_D$ (= Debye temperature). According to Pomeranchuk[34] they may be important for the phonon interaction only at very high temperatures, $T \gg \theta_D$.

The solution of transport problems becomes particularly simple and easy to survey if it is possible to characterize the time derivative of a distribution function due to scattering processes by a pure relaxation or collision time which is independent of external 'forces'. The convenient definition of a relaxation time τ in case of phonon–phonon interaction would be given by

$$\left.\frac{\partial N(\mathbf{q},s)}{\partial t}\right|_{\text{ph-ph}} = \frac{N(\mathbf{q},s)-N_0(\mathbf{q},s)}{\tau(\mathbf{q},s)} \qquad \ldots (51)$$

In general such a preliminary statement is not possible because the time derivation of $N(\mathbf{q},s)$ depends implicitly in a complicated manner from the occupation numbers of many other states. Principally too, in general, it is not possible to define a pure relaxation time $\tau(\mathbf{q},s)$. A necessary presumption would be that the influence of the externally applied temperature gradient in the three perturbation functions $N(\mathbf{q},s)-N_0(\mathbf{q},s)$ could be expressed by a common factor which depends on $|\operatorname{grad} T|$ only. In general this is not the case as is shown by the simple formula (68). Only at low temperatures and for low frequency lattice waves is the definition of $\tau(\mathbf{q},s)$ according to equation (51) as time of operation permitted, which states how long a single wave excited beyond the thermal equilibrium needs to adapt itself again to the equilibrium. In this case to a good approximation it may be supposed that all other lattice waves are in thermal equilibrium.

4.1. Interaction of Phonons

The mathematical formulation of the phonon–phonon interaction becomes very clear if we replace the real normal co-ordinates (47) by complex normal co-ordinates according to Becker and Leibfried[35]

$$b = \frac{1}{\sqrt{(2\hbar)}}\left[\sqrt{(M\omega)}\xi - \mathrm{i}\frac{p}{\sqrt{(M\omega)}}\right], \quad b^* = \frac{1}{\sqrt{(2\hbar)}}\left[\sqrt{(M\omega)}\,\xi + \mathrm{i}\frac{p}{\sqrt{(M\omega)}}\right] \ldots (52)$$

$$bb^* - b^*b = 1, \qquad \epsilon = \hbar\omega(bb^* + \tfrac{1}{2}) \qquad \ldots (53)$$

With this notation the displacement vector according to equation (47) is given by

$$\delta\mathbf{R}_j = \frac{1}{\sqrt{G^3}} \sum_{\mathbf{q},s} \sqrt{\left(\frac{\hbar}{2M\omega(\mathbf{q},s)}\right)}\, \mathbf{n}(\mathbf{q},s) \times$$
$$\times \left[b(\mathbf{q},s)\exp(-\mathrm{i}\mathbf{q}\mathbf{R}_j) + b^*(\mathbf{q},s)\exp(+\mathrm{i}\mathbf{q}\mathbf{R}_j)\right] \qquad \ldots (54)$$

Equation (54) represents a decomposition of the displacement into *running* waves,[37] whereas in equation (47) the displacement is written in terms of *progressing* plane waves. Now, on the basis of quantum theory for further calculation there are two slightly different possibilities both leading, of course, to the same result. On the one hand the phonon interaction may be calculated in the framework

of ordinary quantum mechanics. When doing this, Dirac's perturbation theory of time proportional transition probabilities leads to the matrix elements

$$\langle n|\, b(\mathbf{q},s)\,|n+1\rangle = \langle n+1|\, b^*(\mathbf{q},s)\,|n\rangle = (n+\tfrac{1}{2})^{1/2} \qquad \ldots(55)$$

Or, on the other hand, in the framework of second quantization of the Schrödinger equation (39) we define creation and annihilation operators for phonons which act on the state vectors of the unperturbed system when performing time dependent perturbation calculation.

We join in the second formulation the distinct representation of the mathematical problem by Leibfried and Schlömann.[36] In order to obtain the conservation law for the wave vectors $\mathbf{q}$ in a simple fashion, the authors have introduced co-ordinates which are slightly different from those defined in equation (52). Since equation (53) applies for the new co-ordinates too, we will not introduce a new notation. The new variables are defined in such a way that the displacement vector can be rewritten in the simple form:

$$\delta\mathbf{R}_j = \frac{1}{\sqrt{G^3}} \sum_{\mathbf{q},s} \sqrt{\left(\frac{\hbar}{2M\omega(\mathbf{q},s)}\right)}\, \mathfrak{n}(\mathbf{q},s) \exp(i\mathbf{q}\mathbf{R}_j)[b(\mathbf{q},s) - b^*(-\mathbf{q},s)] \qquad \ldots(56)$$

An important step towards quantum theoretical treatment of the problem consists in that the complex conjugate amplitudes $b(\mathbf{q},s)$, $b^*(\mathbf{q},s)$ are assigned to the Hermitian conjugate operators $b_{\mathbf{q},s}$, $b^+_{\mathbf{q},s}$. For these, according to equation (53), the commutator relation applies:

$$[b_{\mathbf{q},s}, b^+_{\mathbf{q}',s'}] = \delta_{\mathbf{q}\mathbf{q}'}\,\delta_{ss'} \qquad \ldots(57)$$

In the present notation the Hamiltonian of the unperturbed problem and the associated eigenvalues are diagonal:

$$\mathscr{H}_{\text{latt}} - \mathscr{H}_{\text{latt}}(\mathbf{R}^{(0)}) = \mathscr{H}^{(0)}_{\text{latt}} = \sum_{\mathbf{q},s} \hbar\omega(\mathbf{q},s)(b^+_{\mathbf{q},s} b_{\mathbf{q},s} + \tfrac{1}{2}) \qquad \ldots(58)$$

$$\epsilon_{\text{latt}} = \epsilon[\ldots, n(\mathbf{q},s), \ldots] = \sum_{\mathbf{q},s} \hbar\omega(\mathbf{q},s)[n(\mathbf{q},s) + \tfrac{1}{2}] \qquad \ldots(59)$$

The eigenvalues (59) are the *phonon states* of the total system. It is $n = b^+ b$. By the above it is evident that the operator $b^+_{\mathbf{q},s}$ represents the creation operator for a phonon $\hbar\omega(\mathbf{q},s)$ and $b_{\mathbf{q},s}$ represents the annihilation operator of this phonon. The operators b, b^+ can be related to a set of state vectors $\psi(\ldots, n(\mathbf{q},s), \ldots)$ which depend on the occupation numbers $n(\mathbf{q},s) = 0, 1, 2, \ldots$ and are orthonormal. This yields

$$\begin{aligned} b^+_{\mathbf{q},s}\psi[\ldots, n(\mathbf{q},s), \ldots] &= \sqrt{[n(\mathbf{q},s)+1]}\,\psi[\ldots, n(\mathbf{q},s)+1, \ldots] \\ b_{\mathbf{q},s}\psi[\ldots, n(\mathbf{q},s), \ldots] &= \sqrt{[n(\mathbf{q},s)]}\,\psi[\ldots, n(\mathbf{q},s)-1, \ldots] \end{aligned} \qquad \ldots(60)$$

The Hamiltonian of the perturbed problems is

$$\begin{aligned} \mathscr{H}^{(0)}_{\text{latt}} + V_3 &= \mathscr{H}^{(0)}_{\text{latt}} + \frac{1}{3!} \sum_{j,k,l} \sum_{\lambda,\mu,\nu=1}^{3} B^{\lambda\mu\nu}_{jkl}\, \delta\mathbf{R}^{(\lambda)}_j\, \delta\mathbf{R}^{(\mu)}_k\, \delta\mathbf{R}^{(\nu)}_l \\ &= \mathscr{H}^{(0)}_{\text{latt}} + \frac{1}{3!} \sum_{\substack{\mathbf{q},\mathbf{q}',\mathbf{q}'' \\ s,s',s''}} B(\mathbf{q},s;\mathbf{q}',s';\mathbf{q}'',s'')[(b_{\mathbf{q},s} - b^+_{-\mathbf{q},s})(b_{\mathbf{q}',s'} - b^+_{-\mathbf{q}',s'})(b_{\mathbf{q}'',s''} - b^+_{-\mathbf{q}'',s''})] \end{aligned} \qquad \ldots(61)$$

The coefficients

$$B(\mathbf{q},s;\mathbf{q}',s';\mathbf{q}'',s'') = \frac{1}{G^{3/2}}\sum_{j,k,l}\sum_{\lambda,\mu,\nu=1}^{3} B_{jkl}^{\lambda\mu\nu}\left(\frac{\hbar}{2M}\right)^{3/2}\frac{\mathbf{n}_\lambda(\mathbf{q},s)\,\mathbf{n}_\mu(\mathbf{q}',s')\,\mathbf{n}_\nu(\mathbf{q}'',s'')}{[\omega(\mathbf{q},s)\omega(\mathbf{q}',s')\omega(\mathbf{q}'',s'')]^{1/2}}\times$$
$$\times\exp\left[i(\mathbf{q}\mathbf{R}_j+\mathbf{q}'\mathbf{R}_k+\mathbf{q}''\mathbf{R}_l)\right] \qquad \ldots(62)$$

are invariant with regard to the translations $\mathfrak{t}$ of equation (27). From this results

$$B(\mathbf{q},s;\mathbf{q}',s';\mathbf{q}'',s'') = 0 \qquad \ldots(63)$$

if not

$$\mathbf{q}+\mathbf{q}'+\mathbf{q}'' = 2\pi\mathbf{b} \qquad \ldots(64)$$

From the translation invariance of V_3 already results the conservation law for the wave vectors. It is obvious that V_3 induces three-phonon scattering processes and V_4 four-phonon scattering processes, etc. The transition probability W_{if} for three-phonon processes connecting

$$\text{an initial state:} \quad \psi_i = \psi[\ldots,n_i(\mathbf{q},s),\ldots]$$
$$\text{with a final state:} \quad \psi_f = \psi[\ldots,n_f(\mathbf{q},s),\ldots]$$

is to be calculated by means of Dirac's perturbation calculation. The time derivatives of the distribution functions $N(\mathbf{q},s)$ due to phonon–phonon interaction are given by

$$\left.\frac{\partial N(\mathbf{q},s)}{\partial t}\right|_{\text{ph-ph}} = \sum_f \frac{dW_{if}}{dt} = \sum_f \frac{2\pi}{\hbar}\left|\langle i|V_3|f\rangle\right|^2\delta(\epsilon_f-\epsilon_i) \qquad \ldots(65)$$

$\delta(x)$ is the usual integral definition of Dirac's δ-function. After multiplying out the bracketed terms in equation (61) altogether eight operator products in V_3 will occur. Out of the corresponding transitions owing to energy conservation law, six are allowed and two are forbidden (the energy conservation law for these two processes reads: $\hbar(\omega+\omega'+\omega'')=0$, that is $\omega'=\omega(\mathbf{q}',s')$, etc.). The six allowed transitions in which the sum of phonon quantum numbers $n+n'+n''$ changes by $\Delta n = \pm 1$ may be arranged with the energy and the momentum conservation law in the following manner

$$\hbar\omega = \hbar\omega'+\hbar\omega''; \quad \hbar\omega+\hbar\omega' = \hbar\omega''; \quad \hbar\omega+\hbar\omega'' = \hbar\omega'$$
$$\mathbf{k} = \mathbf{k}'+\mathbf{k}''+2\pi\mathbf{b}; \quad \mathbf{k}+\mathbf{k}' = \mathbf{k}''+2\pi\mathbf{b}; \quad \mathbf{k}+\mathbf{k}'' = \mathbf{k}'+2\pi\mathbf{b} \qquad \ldots(66)$$

In a collision process a phonon splits into two other phonons or vice versa. However, it is to be distinguished mainly between normal collision (N-collision), $\mathbf{b}=0$, and Umklapp collision (U-collision), $\mathbf{b}\neq 0$. With regard to Umklapp collisions which do not conserve momentum, the importance of which for the lattice heat conductivity has been recognized for the first time by Peierls, the vector $\mathbf{q}''=\mathbf{q}+\mathbf{q}'+2\pi\mathbf{b}$ for example, may form an obtuse angle with the vector $\mathbf{q}+\mathbf{q}'$. Therefore, roughly, in U-collisions, contrary to N-collisions, the direction may be changed in which the energy is transferred. It results in a finite thermal resistance. As Leibfried shows graphically at least two of three wave vectors from phonons

taking part in a scattering process must be located near the boundary of the Brillouin zone (sub-section 3.3). Therefore, at low temperatures when only low frequency lattice waves with small wave vectors $\mathbf{q}$ are excited the number of U-collisions will become relatively small.

To avoid subsequent misunderstanding the expression U-collisions always means three-phonon collisions for which $\mathbf{b} \neq 0$. In the theory of *electrical* transport phenomena a different kind of Umklapp collisions may play a part for the wave vector $\mathbf{k}$ of a plane electron wave. In an electron-phonon scattering process the conservation law yields

$$\mathbf{k}' = \mathbf{k} \pm \mathbf{q} + 2\pi\mathbf{b} \qquad \ldots (XIV)$$

Both kinds of U-processes are of fundamental importance in so far as only by taking into account *at least one* of these processes in an exact theory, do they result in finite transport coefficients such as a finite lattice heat conductivity or a finite electrical conductivity. This is apparent from regarding the time derivative of the 'total wave number'

$$\mathbf{G} = \sum_{\mathbf{k}} f(\mathbf{k})\,\mathbf{k} + \sum_{\mathbf{q},s} N(\mathbf{q},s)\,\mathbf{q} \qquad \ldots (XV)$$

After a simple intermediate calculation[33,88] the Boltzmann equations of the stationary state

$$\left.\frac{\partial N(\mathbf{q},s)}{\partial t}\right|_{\text{ph-ph}} + \left.\frac{\partial N(\mathbf{q},s)}{\partial t}\right|_{\text{temp}} = 0 \qquad \ldots (67a)$$

may be written in extensive fashion. For this purpose the six allowed groups of three-phonon processes are combined in a suitable manner and the summation on the $\mathbf{q}$-vectors is to be replaced by an integration across the $\mathbf{q}$-space. We finally get

$$\begin{aligned}\frac{\partial N}{\partial t}\left(\frac{\partial \omega}{\partial \mathbf{q}}\operatorname{grad} T\right) = \frac{2\pi}{\hbar^2}\int \mathrm{d}^3\mathbf{q}' \Bigg\{ & 2\sum_{s',s''} |B(\mathbf{q},s;\mathbf{q}',s';s'')|^2 \times \\ & \times \delta(\omega+\omega'-\omega'')[(N+1)(N'+1)N'' - NN'(N''+1)] + \\ & + \sum_{s',s''} |B(\mathbf{q},s;\mathbf{q}',s';s''')|^2 \delta(\omega-\omega'-\omega'') \times \\ & \times [(N+1)N'N'' - N(N'+1)(N''+1)] \Bigg\} \qquad \ldots (67b)\end{aligned}$$

Equation (67) represents a system of three coupled integral equations for the stationary distribution functions $N(\mathbf{q},s)$ in the perturbed state, provided, of course, that all information about the phonon system is contained in the $N(\mathbf{q},s)$. By the notation

$$N(\mathbf{q},s) = N_0(\mathbf{q},s)\{1 + [N_0(\mathbf{q},s)+1]\,g(\mathbf{q},s)\} \qquad \ldots (68)$$

the calculation of the perturbation functions $g(\mathbf{q},s)$ is the most difficult task from the mathematical point of view in the theory of thermal conductivity. Even with simple assumptions about the lattice frequency spectrum this problem could not, up to now, be solved in general form.

4.2. Unbounded Perfect Crystal

According to the above definitions of the perturbation functions $g(\mathbf{q}, s)$ and to equation (50), the heat current density and, consequently, the components of the thermal conductivity tensor depend linearly on the perturbation functions. Although up to now it was not possible to succeed in calculating these functions in a consistent manner under general assumptions, important results were obtained by Peierls[32] and later on by Herpin,[39] Herring,[40] Klemens,[41] Leibfried and Schlömann,[36] and Pomeranchuk,[34] which allow a qualitative survey on the temperature dependence of thermal conductivity. Peierls established and discussed the equations (67) for the first time. He finds the following results:

(1) The formula

$$g(\mathbf{q}, s) = \text{const}(\mathbf{q}\,\text{grad}\,T) \qquad \ldots (69)$$

represents a non-trivial solution of the *homogeneous* integral equations (67*b*)—the left-hand side of equation (67*b*) is equal to zero—provided U-collisions are disregarded. However, an energy flow which is different from zero will occur, although there is no temperature difference between any two points. Now, according to the 'Alternativsatz' of Fredholm[42] on inhomogeneous integral equations, the formula (69) is no solution of the inhomogeneous equations (67*b*). Therefore, a consequence of this general statement is that normal processes alone are not sufficient for adjustment of a stationary non-equilibrium state. This consequence from the physical point of view is easy to interpret if the 'total wave number' of the phonons

$$\mathbf{G}_{\mathbf{q}} = \sum_{\mathbf{q},s} N(\mathbf{q}, s)\mathbf{q} \qquad \ldots (70)$$

is examined. Its time derivative must vanish in the stationary state. While in normal processes the first term of the time derivative does not change, i.e. $(\partial \mathbf{G}_{\mathbf{q}}/\partial t)_{\text{ph-ph}} = 0$, the second term $(\partial \mathbf{G}_{\mathbf{q}}/\partial t)_{\text{temp}} \neq 0$ because the lattice waves preferably travel in the opposite direction to the temperature gradient. However, when U-collisions are taken into account we have $(\partial \mathbf{G}_{\mathbf{q}}/\partial t)_{\text{ph-ph}} \neq 0$.

While N processes alone give no rise to thermal resistance, they result in shifting the unperturbed distribution $N_0(\mathbf{q})$ in momentum space. Klemens[41] recognized that the stationary distribution for N processes is given by

$$N_{\lambda}(\mathbf{q}, s) = \frac{1}{\exp\{[\hbar\omega + (\lambda\mathbf{q})]/kT\} - 1} \qquad \ldots (\text{XVI})$$

where λ is an arbitary vector. Since N processes change the mean occupation numbers of different single phonon states with the tendency to produce a quasi-equilibrium distribution (XVI), they are effective in determining the final steady state distributions $N(\mathbf{q}, s)$ due to N processes and U processes.

(2) At low temperatures U-collisions seldom occur. Thus, the right-hand side of equation (69) is to be supplemented by a correction term only:

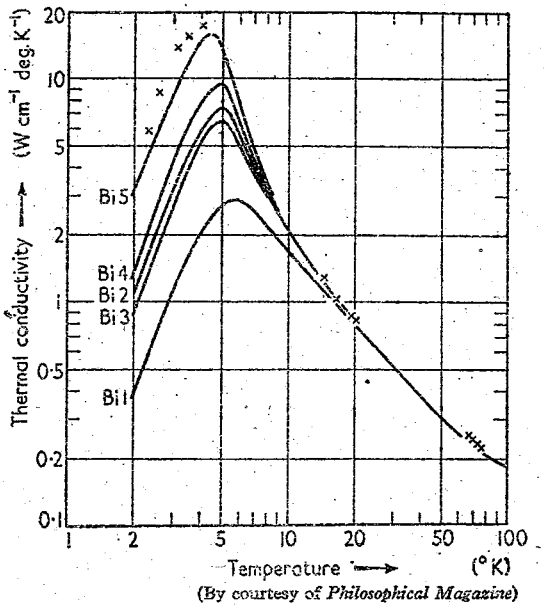
$$g(\mathbf{q}, s) = \text{const}(\mathbf{q}\,\text{grad}\,T) + \hbar(\mathbf{q}, s) \qquad \ldots (71)$$

Principally, with this notation the constant is determined in the same manner as the solution of the Bloch integral equation at low temperatures. In both cases the

theorem of Fredholm is utilized. The result for the constant and, consequently, for the thermal conductivity is given by:

$$\chi \propto T^{v} \exp\left(-\tfrac{1}{2}\frac{\theta}{T}\right), \qquad \frac{T}{\theta} \ll 1 \qquad \ldots(72)$$

The characteristic temperature θ in general is equal to the 'mean' Debye temperature θ_D. For v Peierls has not reported any value. According to Klemens[41] $v = +2$. An exponential dependence of $\chi(T)$ for example has been observed by Berman, Simon, and Ziman[43] on diamond ($\theta_D \simeq 2{,}065°$ K) and by White and Woods[44] on polycrystalline bismuth crystals (Figure 3, $\theta_D \simeq 120°$ K).



(By courtesy of *Philosophical Magazine*)

Figure 3. *The thermal conductivity of bismuth crystals of different diameters (according to White and Woods[44]). The crosses show measuring points according to Shalyt[44]*

(3) At high temperatures $N+1 \simeq N \simeq kT/\hbar\omega$. This simplification allows us to deduce by comparing the dimensions of both sides of equation (67) that $g \propto T^{-3}$ and therefore

$$\chi \propto T^{-1}, \qquad \frac{T}{\theta} \gg 1 \qquad \ldots(73)$$

The coupling parameter $B(\mathbf{q},s;\mathbf{q}',s';\mathbf{q}'',s'')$ is independent of temperature. When regarding the fourth-order terms in the expansion of potential energy in terms of displacements only, we obtain

$$\chi \propto T^{-2}, \qquad \frac{T}{\theta} \gg 1 \qquad \ldots(74)$$

Supposing that a mean free path may be defined, Pomeranchuk[34] obtains $\chi \propto T^{-3/2}$. The measured T-dependence of the heat conductivity at high temperatures varies

between T^{-1} and T^{-2}. An example is the thermal conductivity of germanium measured by Erdmann[45] and shown in Figure 4.

The results of Peierls are summarized above; they were obtained without the concept of a relaxation time. A very crude estimate of the constants of proportionality is possible if a relaxation time $\tau(\mathbf{q},s)$ or a mean free path $\mathbf{l}=\tau\mathbf{u}_g$, $|\mathbf{u}_g|=u_g$, is introduced and the isotropic thermal conductivity is calculated,[33,38,41] for example, for cubic crystals according to the relation

$$\chi=\sum_{\mathbf{q},s} c_v(\mathbf{q},s)\,u_g^2(\mathbf{q},s)\,\tau(\mathbf{q},s)\cos^2\theta \qquad \ldots(75)$$

In this notation θ is the angle between $\mathbf{q}$ and grad T and $c_v=\sum_{\mathbf{q},s} c_v(\mathbf{q},s)$ is the specific heat per unit volume.

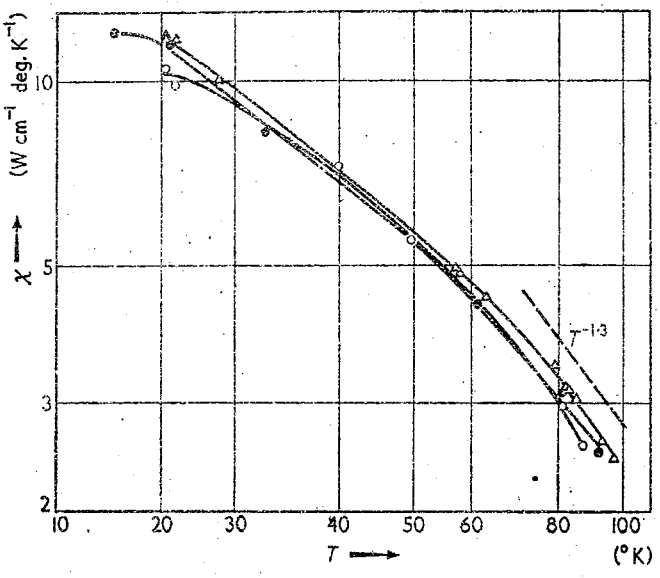


Figure 4. *The thermal conductivity of pure germanium single crystals (according to Erdmann[45])*

At low temperatures preferably the low frequency phonons are excited. The introduction of a relaxation time $\tau(\mathbf{q},s)$ for these phonons is justified[40] in that they interact in most cases with those phonons for which $\mathbf{q}$ is large and therefore the 'relaxation time' is very small. Thus, it is allowed when calculating τ from equation (67) to put $N'=N_0'$ etc. Then it yields

$$\begin{aligned}\frac{1}{\tau(\mathbf{q},s)} &= -\lim_{N\to N_0}\frac{\partial N/\partial t|_{\text{ph-ph}}}{N-N_0}\\ &= -\frac{2\pi}{\hbar^2}\int \mathrm{d}^3\mathbf{q}'\Big[2\sum_{s',s''}|B|^2\delta(\omega+\omega'-\omega'')\frac{(N_0'+1)N_0''}{N_0}+\\ &\quad+\sum_{s',s'''}|B|^2\delta(\omega-\omega'-\omega''')\frac{N_0'N_0'''}{N_0}\Big] \qquad \ldots(76)\end{aligned}$$

When presuming $N_0 \simeq kT/\hbar\omega$ and

$$\lim_{|\mathbf{q}|\to 0} \frac{|B(\mathbf{q},s;\mathbf{q}',s';\mathbf{q}'',s'')|^2}{\omega\omega'\omega''} = \text{const} \qquad \ldots(77)$$

Herring[40] has calculated the relaxation times for longitudinal and transverse phonons in elastic isotropic and anisotropic crystals by a geometrical reflection on the conservation surfaces assigned to a $\mathbf{q}$-vector by means of the energy and momentum conservation law for normal processes. He arrives at the general relation

$$\frac{1}{\tau(\lambda\mathbf{q},\lambda T)} \propto \lambda^5 \qquad \ldots(78)$$

With elastic isotropy this is reduced to

$$\left.\begin{aligned} \frac{1}{\tau_l(\mathbf{q})} &\propto q^4 T \\ \frac{1}{\tau_t(\mathbf{q})} &\propto qT^4 \end{aligned}\right\} \quad \begin{aligned} q &\leqslant q_c \\ \hbar\omega(q_c) &\simeq 0{\cdot}15\,kT \end{aligned} \qquad \ldots(79)$$

As is easily observed the longitudinal phonons contribute a divergent part to the thermal conductivity. Since the disregarded U-collisions for low frequency phonons are improbable at low temperatures, allowing for elastic anisotropy is a reasonable course for disposing of these difficulties, as could be shown by Herring. For cubic and hexagonal crystal classes

$$\left.\begin{aligned} \frac{1}{\tau_l(q)} &\propto q^2 T^3 \\ \frac{1}{\tau_t(q)} &\propto qT^4 \end{aligned}\right\} \quad q \leqslant q_c \qquad \ldots(80)$$

These equations are reliably valid for small elastic anisotropy occurring, for example, in germanium. In this case the anisotropy constant is

$$\frac{2c_{44}}{c_{11}-c_{12}} = 1{\cdot}65 \qquad \ldots(81)$$

The elastic constants c_{ij} are defined according to Voigt.[46] In equation (80) the constants of proportionality may be direction-dependent. The knowledge of the relaxation time $\tau_l(q)$ is important for calculating the electrical transport phenomena of semiconductors if the deviation of the phonons from thermal equilibrium is taken into account. For the phonon–electron interaction just those phonons come into question of which the relaxation time is $\tau_l(q)$. The latter enters into that part of differential thermoelectric power which is induced by the deviation of the phonons from thermal equilibrium and which can be much larger than the electronic drift part. The mean free path calculated from the measured thermoelectric power of N-type germanium is represented in Figure 5 as a function of temperature.[47]

Finally, in this paragraph we would like to point out that a promising method of calculating the thermal conductivity may be given by the formulation of the three coupled integral equations (67) as variation principle. Schlömann and Leibfried[69] lately applied this method to the problem of thermal conductivity of insulators, being introduced successfully by Kohler[68] into the theory of the electrical transport phenomena of metals. In this way it would be possible, as in the case of electrical conductivity,[68] to take into account in a distinct manner

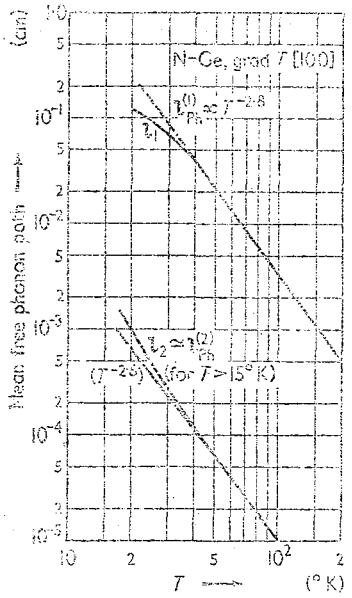


Figure 5. The temperature dependence of the mean free phonon path. $\lambda_{ph}^{(1)}$ results from the measured values of the differential thermoelectric power. $\lambda_{ph}^{(2)}$ results from the measured thermal conductivity (after Appel[67])

additional scattering processes to which the phonons in a crystal with imperfections are subjected and which will be discussed in the next section.

4.3. Crystal with Imperfections

In real crystals the phonons are additionally scattered by:

(1) Crystal boundary surfaces;
(2) Grain boundary surfaces;
(3) Point imperfections;
(4) Dislocations.

These static lattice imperfections have a common factor in that the corresponding transition probabilities of elastic scattering processes are independent of temperature. The heat resistances associated with lattice imperfections, contrary

to the electrical residual resistance, is temperature-dependent only because the mean occupation numbers $N_0(\mathfrak{q}, s)$ are substantially dependent upon the energy $\hbar\omega(\mathfrak{q}, s)$ of a phonon.

A detailed theoretical examination on the influence of static lattice imperfections upon the thermal conductivity has been accomplished by Klemens[50] within the framework of time-dependent perturbation theory. The perturbation Hamiltonian has the form

$$\mathscr{H}' = \sum_{\substack{\mathfrak{q},\mathfrak{q}' \\ s,s'}} C(\mathfrak{q}, s; \mathfrak{q}', s')[(b_{\mathfrak{q}s} - b^{+}_{-\mathfrak{q}s})(b_{\mathfrak{q}'s'} - b^{+}_{-\mathfrak{q}'s'})] \qquad \ldots (82)$$

The first step is to determine the function $C(\mathfrak{q}, s; \mathfrak{q}', s')$, being characteristic of each lattice imperfection. In cases (3) and (4) the elastic constants of interatomic linkages in small imperfect lattice domains and/or the mass of lattice particles are changed with regard to the ideal crystal. By the second step, transition probabilities and corresponding relaxation times are calculated. The relaxation time concept is allowed at low temperatures for low frequency phonons when a single scattering mechanism obviously dominates. Klemens obtained the following results:

(1) Crystal boundary surfaces:

$$\frac{1}{\tau_b} \propto \frac{1}{L}, \qquad \chi_b \simeq \frac{104\pi k^4}{h^3 u^2} LT^3 \qquad \ldots (83)$$

where L = smallest crystal dimension.

(2) Grain boundary surfaces:

$$\frac{1}{\tau_g} \propto N_g, \qquad \chi_g \simeq \frac{104\pi k^4 \alpha^2 T^3}{h^3 u^2} \frac{}{N_g} \qquad \ldots (84)$$

where N_g = number of boundaries per centimetre,
α = average angle of tilt.

(3) Point imperfections:

$$\frac{1}{\tau_p} \propto S^2 N_p q^4, \qquad \chi_p \simeq \frac{0{\cdot}3hu^2}{4\pi^2 a^3} \frac{1}{S^2 N_p T} \qquad \ldots (85)$$

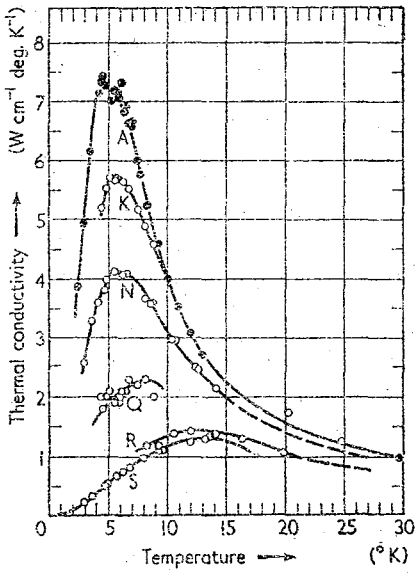
where N_p = concentration of imperfections,
S = scattering parameter with order of magnitude 1, depending on the type of imperfections,
a^3 = volume of the unit cell.

(4) Single dislocations:

$$\frac{1}{\tau_d} \propto N_d q, \qquad \chi_d \simeq \frac{28bk^3}{u\gamma^2 h^2} \frac{T^2}{N_d b^2} \qquad \ldots (86)$$

where N_d = number of dislocations per square centimetre,
b = magnitude of the Burgers vector,
γ = the Grüneisen constant.[41]

Already in 1938 the T^3 law had been established by Casimir,[51] presuming that the thermal lattice waves on the crystal surfaces are diffusely scattered. In the meantime it has been proved by several authors that the T^3 law is in accordance with experiment (Figure 3). The further results (2, 3, 4) have been discussed in detail by Klemens[41] in comparison with experience. Additionally, we should like to refer to a recent paper by Slack[52] (Figure 6) who examined in detail the thermal conductivity of potassium chloride single crystals at low temperatures. Atomic lattice imperfections consisting of a Ca^{2+} ion and a K^+ vacancy have been produced by adding calcium chloride. Oxide semiconductors such as, for example, zinc oxide, cadmium oxide, and titanium oxide are further substances suitable for an



(By courtesy of *Physical Review*)

Figure 6. *The thermal conductivity of potassium chloride single crystals doped with calcium chloride (according to Slack[52])*

examination of the influence of atomic imperfections, especially of vacancies, on thermal conductivity.

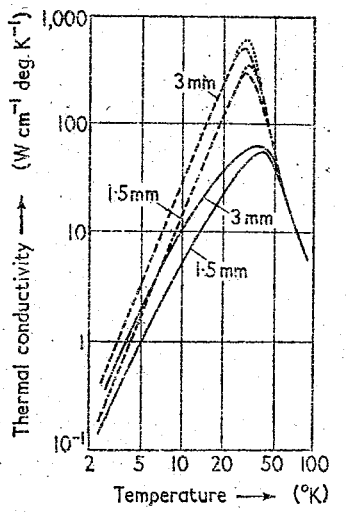
Generally, in a real crystal, several scattering processes will superpose at low temperatures. It is observed quite generally that the definition of a mixed relaxation time according to

$$\frac{1}{\tau(\mathbf{q},s)} = \frac{1}{\tau_1(\mathbf{q},s)} + \frac{1}{\tau_2(\mathbf{q},s)} + \dots \qquad \dots (87)$$

may still represent a sufficiently good approximation. But the definition of a mean relaxation time

$$\frac{1}{\bar{\tau}} = \frac{1}{\bar{\tau}_1} + \frac{1}{\bar{\tau}_2} + \dots \qquad \dots (88)$$

and thus of the thermal conductivity according to equation (49) is merely a poor approximation, if there is a substantial difference in the q-dependencies of the individual relaxation times. This applies to the direction dependence as well as to the $\mathbf{q}$-dependence of the relaxation times. The difference between the true thermal resistance and that calculated with the simple formula (88) has its largest values if two of the resistance terms are equal. However, as in the case of electronic thermal conductivity of metals, an exact calculation of the thermal conductivity of insulators, regarding the superposition of different scattering processes, should be possible if one succeeds in formulating the general transport problem as variation



(By courtesy of *Advances in Physics*)

Figure 7. *The thermal conductivity of diamond single crystals, diameters* 0·15 *and* 0·3 *cm (after Berman*[53]*) (solid line); theoretical results according to Klemens*[41]*: assuming the simple addition formula* (88) *for reciprocal* mean *relaxation times (broken line), using the addition formula* (87) *for reciprocal relaxation times (dotted line)*

principle and in solving this at least with simplifying suppositions (e.g. elastic isotropy).

In order to show that the total heat resistance calculated with equation (88), $W=\sum_i W_i$ in comparison with experiment always yields too small values, the thermal conductivity of two diamond single crystals measured by Berman[53] is represented in Figure 7. However, the difference $W_{\text{exp}} \gg W$ in the vicinity of the extreme values according to Berman et al.[54] is due to a great extent to impurities. Another discrepancy between theory and experiment consists in the fact that in several cases, e.g. on potassium chloride,[55] tellurium,[56] and germanium and silicon,[57] no exponential temperature behaviour of the thermal conductivity has been observed on pure crystals at low temperatures. According to Berman, Forster, and Ziman,[58] this difference is due to different isotopes constituting the

crystals and which, according to Klemens,[58] cause an additional heat resistance. However, Geballe and Hull[60] stated with regard to germanium that the scattering of phonons by natural isotopes cannot explain the lack of an exponential temperature dependence of χ. Because of the strong dispersion of the transversal acoustic lattice vibrations ($\theta_t < \theta_D$) the number of U-collisions does not diminish until temperatures $T \ll \theta_t$.

5. ELECTRONIC THERMAL CONDUCTION

We suppose that the total thermal conductivity of a semiconductor consists additively of two parts, the electronic conductivity χ_e and the phonon conductivity χ_{ph} and accordingly, we define

$$\chi^{(0)} = \chi_e(N = N_0) + \chi_{ph}(\mathfrak{f} = \mathfrak{f}_0) \qquad \ldots (89)$$

It has to be investigated how well the so-defined conductivity $\chi^{(0)}$ approximates the true conductivity χ of the total system consisting of electrons and phonons. The zero index indicates that on calculation of the electronic thermal conductivity about the distribution functions of electrons and phonons the two substantial presumptions have been made:

$$(1)\ N = N_0 \quad \text{and} \quad (2)\ \mathfrak{f} - \mathfrak{f}_0 \quad \text{is independent of } N - N_0.$$

Similarly, on calculation of the phonon conductivity $\chi_{ph}(\mathfrak{f} = \mathfrak{f}_0)$ it is assumed that:

$$(3)\ \mathfrak{f} = \mathfrak{f}_0 \quad \text{and} \quad (4)\ N - N_0 \quad \text{is independent of } \mathfrak{f} - \mathfrak{f}_0.$$

The total thermal conductivity calculated with presumptions (1) to (4) has been defined by us as $\chi^{(0)}$ in order to eliminate the usual obscurity. The additional terms in thermal conductivity and in energy transport occurring if the above presumptions are abandoned will together be called superposition effects. In semiconductors these are so small that $\chi^{(0)}$ represents the true total conductivity χ with a maximum inaccuracy of 1 per cent. ($\chi_{ph}(\mathfrak{f} = \mathfrak{f}_0)$ is *not* the insulator lattice conductivity.) For this reason it seems quite logical to examine the energy transport of phonons and electrons first of all separately. In the following the electronic thermal conduction is discussed.

5.1. Extrinsic Semiconductor

For an N-type or P-type semiconductor the thermal conductivity ($N = N_0$) is given by the Wiedemann–Franz–Lorenz law

$$\chi_e = L\sigma T = A2\left(\frac{k}{e}\right)^2 \sigma T \qquad \ldots (90)$$

$$\text{electrons:} \quad \sigma = \begin{cases} \sigma_e \\ \\ \sigma_p \end{cases} \quad L = \begin{cases} L_e \\ \\ L_p \end{cases} \quad n = \begin{cases} n \\ \\ p \end{cases} \quad \mu = \begin{cases} \mu_e \\ \\ \mu_p \end{cases}$$

$$\text{holes:}$$

L is the Lorenz number, $\sigma = ne\mu$ is the electric conductivity. This law applies with assumptions satisfied at higher temperatures:

(1) The deviation of the phonon distribution function from the thermal equilibrium is neglected;

(2) Relaxation times can be defined for the charge and energy transport and are equal, i.e. it implies

$$\left.\frac{\partial \mathbf{f}}{\partial t}\right|_{\text{scatt.}} = -\frac{\mathbf{f}-\mathbf{f}_0}{\tau(\mathbf{k})} \qquad \ldots(91)$$

independent of the external forces, being an electrical field and/or a temperature gradient.

In isotropic semiconductors L is defined[61] by the energy dependence of the collision time $\tau(\epsilon)$ and the value of the parameter $\eta = \zeta/kT$ which includes the

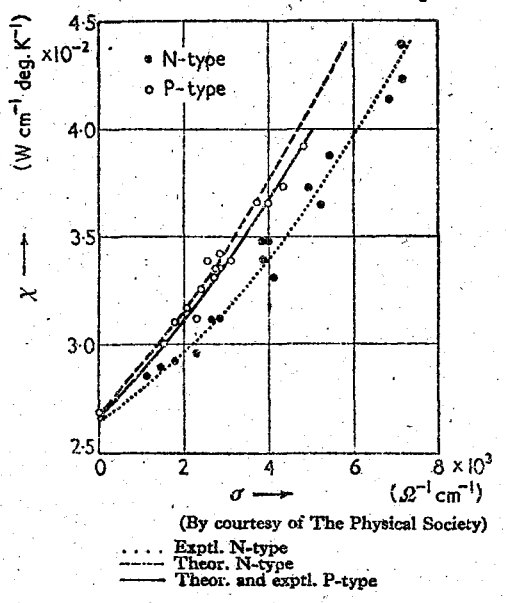


Figure 8. *The heat conductivity of bismuth telluride plotted against the electrical conductivity (after Goldsmid[62])*

Fermi energy. With the zero point of the energy scale fixed as in Section 3, L is given by

$$L = \frac{1}{e^2T^2}\left[\frac{J_3}{J_1} - \left(\frac{J_2}{J_1}\right)^2\right] \qquad \ldots(92)$$

with

$$J_n = \frac{16\pi(2m^*)^{1/2}}{3h^3}\int_0^\infty \tau(\epsilon)\frac{\partial \mathbf{f}_0(\eta)}{\partial \epsilon}\epsilon^{n+1/2}\,\mathrm{d}\epsilon \qquad \ldots(93)$$

If the electrons are exclusively scattered by the phonons, $\tau(\epsilon) \propto \epsilon^{-1/2}$, in non-degenerated semiconductors, $\eta \to -\infty$, then $A = 1$. Then, with $\sigma \propto T^{-3/2}$ the electronic thermal conductivity is proportional to $T^{-1/2}$.

In Figure 8, the thermal conductivity of N-type and P-type Bi_2Te_3 crystals measured by Goldsmid[62] is plotted against the electrical conductivity. The point of intersection on the ordinate resembles the phonon conductivity, which is equal to the insulator lattice conductivity because $\sigma = 0$. If this is also approximately true with $\sigma \neq 0$, then the deviation of the measuring points from a straight line should be explicable by the dependence of the Lorenz number on the scattering mechanism and the degree of degeneration. Actually, Goldsmid is in close agreement with theory for the P-type material if the values for ζ are taken from the measured thermoelectric power, and if the energy dependence $\tau_p \propto \epsilon^{-2\cdot83}$ is accepted for the electron–phonon relaxation time according to the measured temperature behaviour of the mobility. In the case of N-type crystals an explanation of the experimental data is impossible in this sense. The difference is attributed to increased scattering of phonons by foreign halogen atoms with increasing electrical conductivity. By measurements of the thermal conductivity on N-type bismuth telluride which is doped with tellurium, lithium, and aluminium, Goldsmid proves that the scattering of phonons by electrons does not reduce the thermal conductivity measurably. In this case there is good agreement between theory and experiment. Thus the author shows that with electron concentrations of $n = 10^{18}$ to 10^{19} cm^{-3} and $T = 150°$K the scattering of phonons by electrons exerts no distinct influence upon thermal conductivity in the case of the bismuth telluride.

5.2. Intrinsic Semiconductor

In a semiconducting crystal within the intrinsic range the number of electrons in the conduction band is about equal to the number of holes in the valence band. In isotropic conductors both kinds of charge carriers travel under the influence of a temperature gradient in the direction $-\mathrm{grad}\,T$. By this diffusion process, the driving force of which, as in the case of an extrinsic semiconductor, is the difference in the mean thermal velocities of the charge carriers at the hot and the cold ends, energy is transferred. In gases the corresponding diffusion process is called the Sorret effect. This thermal drift motion of the charge carrier is superposed by another effect which is caused by the strong temperature dependence from equilibrium concentrations of electrons and holes. Consequently, when applying a temperature gradient, a concentration gradient for both kinds of charge carriers results. Thus, the concentration gradient is the driving force for an additional energy transfer. This is determined by the excitation energy of free electron–hole pairs, and it has been calculated by Price,[63] Davydov and Shmuskevitsch,[64] and Madelung.[65] Below, the outline of Price is taken into consideration.

In a non-degenerated intrinsic semiconductor the drift velocities v_n, v_p of electrons and holes resulting from a *simultaneous* temperature and concentration gradient are subjected to about the same mathematical interrelationship which governs the thermal diffusion in gases. Since the electronic thermal conductivity is defined in such a way that no electric current is allowed, it yields

$$nv_n = pv_p = N \qquad \ldots (94)$$

If one knows $v_n[\mathrm{grad}(\log n), \mathrm{grad}\,T, E]$ and v_p, it is possible to calculate from equation (94) the unknown electric field strength E and thus the flux N. The total

electronic heat-current density, provided scattering of electrons is by phonons only, is then given by the following formula:

$$\mathbf{w}_e = \mathbf{N}\left(\epsilon_g + \frac{4}{kT}\right) + \mathbf{w}_n + \mathbf{w}_p \qquad \ldots (95)$$

The two last terms correspond to the thermal drift motion of both carriers. The first term results from the temperature behaviour of the equilibrium concentration. The excitation energy of an electron–hole pair $\epsilon_g(T)$ is to be inserted into the brackets. Also the individual terms of $\mathbf{N}$ may be summarized in such a way, as pointed out by Madelung,[23] that $\epsilon_g(T)$ occurs. If the Lorenz number L is defined according to

$$\chi^{(0)} = \chi_{ph} + L_i \sigma_i T, \qquad \sigma_i = \sigma_n + \sigma_p \qquad \ldots (96)$$

then

$$L_i = \left(\frac{k}{e}\right)^2 \left[\frac{\sigma_n \sigma_p}{\sigma^2}\left(\frac{\epsilon_g(T)}{kT} + 4\right)^2 + 2\right] \qquad \ldots (97)$$

applies. At present it is difficult to perform a quantitative comparison with experiment since exact measurements of thermal conductivity above room temperature meet with a large number of experimental difficulties so that considerable errors of measurement might occur. With bismuth telluride which becomes intrinsic below room temperature, Goldsmid and Price[66] succeeded in reconciling theory and experiments with the value $L_i \simeq 25(k/e)^2$. Further experimental results on bismuth telluride by Kanai and Nii[67] as well as by Satterthwaite and Ure[68] may be explained too by equation (97). The thermal conductivity of indium antimonide has been measured by Busch and Schneider,[69] Stuckes,[70] and Weiss.[71] The ambipolar effect is much smaller here than with bismuth telluride because the mobility ratio μ_n/μ_p is very large. The thermal conductivity of germanium up to about 1,000° K has been measured by Joffé,[64] Kettel,[72] and Abeles.[73] A final decision as to whether or not the measured temperature dependence can be explained by the ambipolar diffusion is not yet possible.

6. SCATTERING OF PHONONS BY ELECTRONS, SUPERPOSITION EFFECTS

For analysing the total heat-current density of a semiconductor we started from equation (89). Here it has been an essential presumption for calculating $\chi^{(0)}$ that the energy transports and thus the thermal conductivities of both sub-systems, electrons and phonons, are independent of each other. This independence does not apply exactly and, therefore, we have to prove how the transport problem may be re-formulated in a consistent manner.

The usual presumption of Bloch's theory is made. Thus it is assumed that electrons interact with the longitudinal acoustical phonons only and that the total system electrons and longitudinal phonons can be described by the distribution functions $\mathbf{f}(\mathbf{k})$ and $N(\mathbf{q})$ (see Section 3). These functions are determined by two Boltzmann equations. For the sake of simplicity it is presumed that the electric

field and the temperature gradient are orientated in x_1-direction. Then it yields ($-|e|$ = elemental charge):

$$-\frac{\hbar}{m^*}\frac{\partial \mathbf{f}_0}{\partial \epsilon}K_1\left[eE+T\frac{\partial}{\partial x_1}\left(\frac{\zeta}{T}\right)+\frac{\epsilon}{T}\frac{\partial T}{\partial x_1}\right] = \left.\frac{\partial \mathbf{f}}{\partial t}\right|_{\text{el-ph}} \qquad \ldots (98)$$

$$u\frac{\partial N_0}{\partial t}\frac{\partial T}{\partial x_1} = \left.\frac{\partial N}{\partial t}\right|_{\substack{\text{ph-el}\\ \text{ph-ph}}}$$

The trial for small deviations from equilibrium reads:

$$\begin{aligned}\mathbf{f}(\mathbf{k}) &= \mathbf{f}_0(y)-\frac{\phi(\mathbf{k})}{kT}\frac{\partial \mathbf{f}_0}{\partial y}\\ N(\mathbf{q}) &= N_0(z)-\frac{\psi(\mathbf{q})}{kT}\frac{\partial N_0}{\partial z}\end{aligned} \qquad \ldots (99)$$

For abbreviation we have written $y = \epsilon(\mathbf{k})/kT$ and $z = \hbar\omega(\mathbf{q})/kT$. By this notation it is possible to split up the time derivatives of distribution functions due to scattering processes into two linear operators which depend only upon ϕ or ψ:[49]

$$\begin{aligned}\left.\frac{\partial \mathbf{f}}{\partial t}\right|_{\text{el-ph}} &= -L_1(\phi)-L_2(\psi)\\ \left.\frac{\partial N}{\partial t}\right|_{\substack{\text{ph-el}\\ \text{ph-ph}}} &= -L_3(\phi)-L_4(\psi)\end{aligned} \qquad \ldots (100)$$

Further scattering processes, to which the electrons or phonons alone are subjected, will be performed by positive additional terms to the operators L_1 and L_4. In non-degenerate isotropic semiconductors by means of the notation

$$\phi(\mathbf{k}) = K_1 c(K), \qquad \psi(\mathbf{q}) = q_1 b(q) \qquad \ldots (101)$$

the operators L_n are given by:[74]

$$\begin{aligned}L_1(\phi) &= +\frac{\alpha}{2\hbar l_e}K_1\int_0^{2K}\frac{q^3}{K^3}\mathbf{f}_0(K)c(K)\,\mathrm{d}q\\ L_2(\psi) &= -\frac{\alpha}{2\hbar l_e}K_1\int_0^{2K}\frac{q^3}{K^3}\mathbf{f}_0(K)b(q)\,\mathrm{d}q\\ L_3(\phi) &= +\frac{\alpha}{\hbar l_e}q_1\int_{q/2}^{\infty}\frac{K}{q}\mathbf{f}_0(K)c(K)\,\mathrm{d}K\\ L_4(\psi) &= -\frac{\alpha}{\hbar l_e}q_1\int_{q/2}^{\infty}\frac{K}{q}\mathbf{f}_0(K)b(q)\,\mathrm{d}K\end{aligned} \qquad \ldots (102)$$

l_e is the mean free path of the charge carriers for their interaction with phonons, $\alpha = \hbar^2/2m^*kT$.

With the above defined scattering operators the simultaneous solution of equations (98) is mainly a problem of analysis. It might be solved in various ways. A very logical approach consists in formulating the problem as Kohler's generalized variation principle. This allows the application of distinct solution methods also, when adding further scattering mechanism. The results found thus by Dorn[49] in first approximation for the thermal conductivity are exactly in accordance with those we have found by applying a simple algebraic iteration method.[74] Moreover, in this way we were able to calculate the *thermomagnetic* effects, such as, for example, the change of thermal conductivity in a magnetic field. This method, being applied in a similar manner by Parrot,[75] in principle consists of developing the perturbation functions $c(y)$ and $b(z)$ in terms of a coupling parameter λ:

$$\begin{aligned} c(y) &= c_0(y)+\sum_{i=1}^{\infty}\left(\frac{1}{\lambda}\right)^i c_i(y) \\ b(z) &= b_0(z)+\sum_{i=1}^{\infty}\left(\frac{1}{\lambda}\right)^i b_i(z) \end{aligned} \qquad \ldots(103)$$

That is

$$\lambda = 4{\cdot}48\times 10^{12}\left(\frac{m^*}{m}\right)^{3/2}\frac{T^{5/2}\mu_e}{u l_{ph}^{(1)} n} \qquad \ldots(104)$$

$l_{ph}^{(1)}$ is the mean free path $u\bar{\tau}_l$ (equation (80)) of those phonons which interact substantially with the charge carriers. The coupling parameter λ outlines the relative influence of phonon–electron scattering upon the perturbation $b(z)$. The smaller λ is, the stronger the insulator term $b_0(z)$ is reduced. $c_0(y)$ and $b_0(z)$ are the drift terms. The coefficients $c_i(y)$ and $b_i(z)$ have to be calculated in a self-consistent manner as functions of external forces. A necessary consequence of notation (103) is the splitting of the transport coefficients into two components, an electronic and a phonon contribution. Both are connected by λ. The total heat-current density w in x_1-direction is given by

$$\mathbf{w} = \mathbf{w}_e+\mathbf{w}_{ph} = \frac{\hbar^2}{6\pi^2 m^*}\int_0^{\infty} \mathbf{f}_0 c(y)\left(\frac{y}{\alpha}\right)^{5/2} dy+\frac{1}{6\pi^2 u^3}\left(\frac{kT}{\hbar}\right)^4\int_0^{\theta/T}\frac{N_0 b'(z) z^4}{1-\exp(-z)}\,dz \qquad \ldots(105)$$

That is

$$b'(z) = \begin{cases} b(z), & q \leqslant q_c \\ -\dfrac{\hbar u l_{ph}^{(2)}}{T}\dfrac{\partial T}{\partial x_1}, & q > q_c \end{cases} \qquad \ldots(106)$$

Thus the heat-current density of the phonons is decomposed into two components. $l_{ph}^{(2)}$ is the mean free path, entering into the insulator lattice conductivity (equation (49)).

The results for thermal conductivity may be summarized in the following manner:

$$\begin{aligned} \chi &= \chi_e(N)+\chi_{ph}(\mathbf{f}) \\ &= \chi^{(0)}+\delta\chi_e+\delta\chi_{ph} \end{aligned} \qquad \ldots(107)$$

We use the notation

$$\begin{aligned} \chi_e(N_0) &= A2\left(\frac{k}{e}\right)^2 \sigma T \\ \delta\chi_e &= uBnk\,\mathfrak{k}_{ph}^{(1)} \\ \chi_{ph}(\mathfrak{f}_0) &= \chi_{ph}^{(0)} + \Delta\chi_{ph}(\mathfrak{f}_0) \\ \delta\chi_{ph} &= C\Delta\chi_{ph}(\mathfrak{f}_0) \end{aligned} \qquad \ldots(108)$$

The constants A and B entering into the electronic conductivity are of the order of magnitude 1. With mixed scattering of the electrons by phonons and foreign ions we find for non-degenerated semiconductors

$$A = \tfrac{1}{2}\left[\frac{K_{10}}{K_6} - \left(\frac{K_8}{K_6}\right)^2\right], \qquad B = \frac{4}{3\sqrt{\pi}}\left[\frac{K_9}{K_6} - \left(\frac{K_7}{K_6}\right)^2\right] \qquad \ldots(109)$$

where

$$K_n = \frac{1}{\sqrt{2}} \int_0^\infty \frac{\exp(-y)}{y^2+\beta^2}\, y^{n/2}\, dy$$

$\beta^2/6$ is the mobility ratio for the scattering of the charge carriers by phonons and by foreign ions.

$\chi_{ph}^{(0)}$ is the insulator lattice conductivity being reduced by the scattering of phonons by electrons in proportion to the electron concentration, $\Delta\chi_{ph} < 0$. This reduction depends only to a small extent on the deviation $\mathfrak{f} - \mathfrak{f}_0$ as in the case of metals. For $\lambda \gg 1$

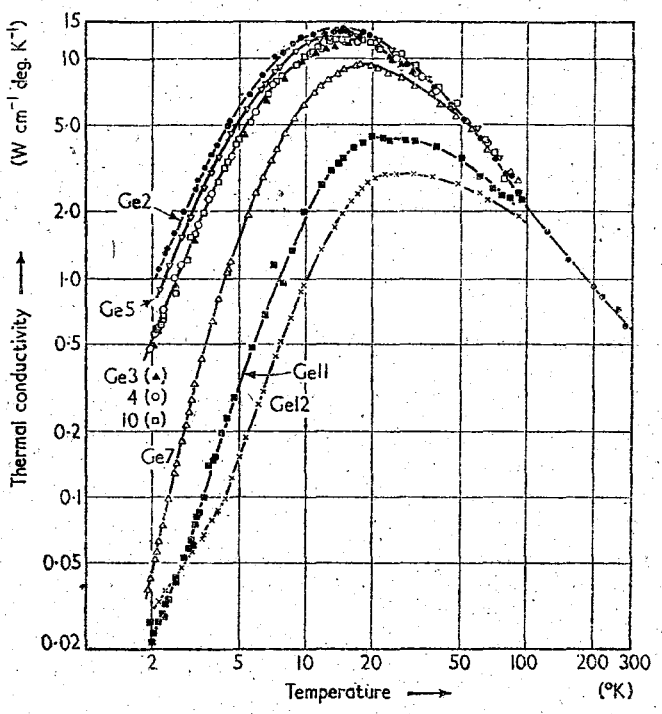
$$\chi_{ph}(\mathfrak{f}) - \chi_{ph}^{(0)} = (1+C)\Delta\chi_{ph} = -\frac{2}{9\pi}\frac{ne}{T\mu}(u\mathfrak{k}_{ph}^{(1)})^2 \qquad \ldots(110)$$

This result conforms exactly with that obtained by Dorn[49] by means of the generalized variation principle in first approximation. Because of $C \ll 1$, equation (110) is valid approximately even if the electrons are subjected to further scattering processes. On inserting data for N-type germanium into equation (110), we find

$$\chi_{ph}(\mathfrak{f}) - \chi_{ph}^{(0)} \simeq -2\times 10^{-10} nT^{-5} \qquad \ldots(111)$$

Accordingly, with $T = 20°$ K the concentration should be $n \simeq 10^{16}$ cm^{-3}, in order to reduce $\chi_{ph}^{(0)}$ by about 10 per cent. This corresponds qualitatively to experimental results of Carruthers et al.[76] They have measured the temperature behaviour of χ on germanium single crystals doped with different amounts of foreign atoms from Group III of the periodic system. The results are shown in Figure 9. The smaller thermal conductivity of the doped crystals compared with that of the pure crystal Ge2 is due to the scattering of phonons by holes. The authors explain the strong reduction of thermal conductivity of some samples below 20° K by the scattering of phonons by charge carriers in the impurity band. Thus it is assumed that phonon–electron or phonon–hole scattering occurs in the impurity band in a similar manner to the corresponding scattering processes of free charge carriers. This presumption appears to be justified because several authors, e.g. Fritsche and Lark-Horwitz,[77] as well as Finke and Lautz[78] observed a normal conductivity

behaviour in so far as a finite conductivity and a Hall effect were proved experimentally. From the theoretical point of view, phonon–electron scattering processes in the impurity band are possible as in the case of free charge carriers, provided the eigenfunctions of the charge carriers may be approximated sufficiently well by eigenfunctions of free particles, i.e. by the Bloch functions with $u_{n\mathbf{k}}(\mathbf{r}) \simeq$ constant, cited in Section 3, the wave vector $\mathbf{k}$ thus representing a good quantum number.[18]



(By courtesy of The Royal Society)

Figure 9. *Heat conductivity of germanium single crystals. Ge2 and Ge5 are pure N-type crystals, the other crystals are doped with elements from Group III of the Periodic System (after Carruthers et al.*[76])

The electronic thermal conductivity is of only little importance in non-degenerate semiconductors at low temperatures. With $A = 1$, it is

$$\frac{\chi_e(N_0)}{\chi_{ph}^{(0)}} = 1{\cdot}48 \times 10^{-8} \frac{\sigma T}{\chi_{ph}^{(0)}} \qquad \ldots (112)$$

if the conductivities are measured by international units. In the case of N-type germanium above 20° K the right-hand side of equation (112) is approximately $10^{-22}\, nT$. The correction term then is $\delta\chi_e \simeq 10^{-20}\, nT^{-3}$.

7. INFLUENCE OF MAGNETIC FIELD

As already stated by Onsager, the thermodynamic theory of the thermomagnetic effects in anisotropic crystals yields a general symmetry relation between the components of the thermal conductivity tensor:

$$\chi_{ik}(\mathbf{H}) = \chi_{ik}(-\mathbf{H}) \qquad \ldots(113)$$

A proof is found with Fieschi, de Groot, and Mazur.[79]

We take into consideration that the tensor components χ_{ik} are defined by the relation (12) and develop the χ_{ik} in terms of the components H_l of an external magnetic field $\mathbf{H}$:

$$\chi_{ik}(\mathbf{H}) = \chi_{ik}(0) + A_{ikl}H_l + A_{iklm}H_lH_m + \ldots \qquad \ldots(114)$$

The coefficients A_{ikl}, A_{iklm} are called *thermomagnetic* coefficients. It is easily observed with equation (12) that the A_{ikl}, transforming as components of a third-order axial tensor like the product of two axial vectors (9 components), take into account the Nernst effect. The A_{iklm}, transforming like components of a fourth-order polar tensor, such as the product of two symmetric tensors (36 components), are decisive for the change of thermal conductivity in weak magnetic fields ($\mu^2H^2 \ll 1, |\mathbf{H}| = \mathbf{H}(\text{Gauss})$). The number of independent coefficients A_{ikl}, A_{iklm} depends on the crystal class (point group) appropriate to the semiconductor. The number of independent tensor components can be determined by the group-theoretical method mentioned in Section 2. Furthermore, this method allows us to state which of the tensor components actually occurs. The results correspond exactly to those for the *galvanomagnetic* coefficients, because in both cases the components of two polar vectors—electric current and electric field strength, heat current, and temperature gradient—are connected with each other. Therefore, the results obtained by Kohler,[80] Okada,[81] and Juretschke[82] for the tensor representation of the galvanomagnetic coefficients may be utilized in obtaining the thermomagnetic coefficients.

With all thermomagnetic effects the *isothermic* and the *adiabatic* effects are to be distinguished. In the first case a heat current is allowed to flow perpendicularly to the primary temperature gradient and to the magnetic field. In the second case the surfaces in this direction are adiabatically isolated. The difference between isothermic and adiabatic effects is negligible[74] provided

(1) The electronic component χ_e of the thermal conductivity is small with regard to the lattice component; or
(2) The effects are investigated in strong magnetic fields.

Both apply as presumptions on non-degenerate extrinsic semiconductors. Thus, the isothermic heat conductivity of an isotropic conductor in a magnetic field[74] yields:

$$\chi_e(H) = A(H)\,2\left(\frac{k}{e}\right)^2 \sigma(0)\,T + B(H)\,u l_{\mathrm{ph}}^{(1)} n \qquad \ldots(115)$$

The functions $A(H)$ and $B(H)$ result from A and B by substituting

$$K_n(H) = \frac{1}{\sqrt{2}} \int_0^\infty \frac{(y^2+\beta^2)\exp(-y)}{(y^2+\beta^2)+\gamma^2 y^2}\, y^{n/2}\, dy \qquad \ldots(116)$$

instead of K_n. The magnetic field strength parameter is $\gamma = (3\sqrt{\pi}/4)10^{-8}\mu H$. Madelung[83] has calculated $A(H)$ for weak fields. One gets

$$A(H) = 1 - \left(\frac{4}{\pi} - \frac{1}{8}\right)\left(\frac{3\pi}{8}\mu H\right)^2, \qquad (\mu H)^2 \ll 1 \qquad \ldots (117)$$

provided that the only interactions of importance for the charge carriers are those with the phonons (mobility μ). With impurity scattering the coefficient of the squared term is increased by a factor 10. With increasing magnetic field strength the thermal conductivity decreases as monotonic function of H. In the limit $H \to \infty$ the perturbation $c(y)$ vanishes. Therefore, it results

$$\lim_{H \to \infty} \chi_e = 0 \qquad \ldots (118)$$

The measuring of thermal conductivity in strong magnetic fields, equation (118), would permit us to determine the phonon conductivity provided this would be independent of the perturbation $c(y) \propto \mathbf{f} - \mathbf{f}_0$. Since this is only approximately true, one obtains

$$\lim_{H \to \infty} \chi_{ph}(\mathbf{f}) = \chi_{ph}(\mathbf{f} = \mathbf{f}_0) \qquad \ldots (119)$$

The correction $\delta\chi_{ph}$, the maximum value of which amounts to about 1 per cent of χ_{ph} with semiconductors as with metals, vanishes in the limit $H \to \infty$.

8. FURTHER TRANSPORT PROCESSES

According to Krumhansl the various energy transport processes may be divided into two groups. By the first group low mean excitation energies are transferred, by the second group higher excitation energies come into play. In the following, out of both groups one energy transport process is discussed which might be of interest in special cases. Neither experimentally nor theoretically, however, have the two processes been examined so thoroughly that the present results can be definitively regarded as final. In this section we want to discuss the transport of thermal energy by radiation and by excitons.

8.1. **Radiation**

This transport process is of some interest at high temperatures. The corresponding thermal conductivity has been calculated by Czerny and Genzel.[84] By applying the Stefan–Boltzmann law for the total spontaneously emitted electromagnetic radiation, they find for the corresponding heat current of an isotropic dielectric:

$$w_{rad} = -\frac{4}{3}\frac{b n_m^2}{k_m}\frac{dT^4}{dx_1} \qquad \ldots (120)$$

The temperature gradient is orientated in the x_1-direction. $b = 5{\cdot}73 \times 10^{-12}$ W cm^{-2}deg.K^{-4} is the Stefan–Boltzmann constant. n_m and k_m are the values of the index of refraction and of the optical absorption constant (cm^{-1}) for that

wavelength which corresponds to the maximum of radiation energy of a black body with temperature T. In the stationary state one obtains

$$\frac{\mathrm{d}w_{\mathrm{rad}}(x_1)}{\mathrm{d}x_1} = 0 \rightarrow \frac{\mathrm{d}T^4}{\mathrm{d}x_1} = \mathrm{const} \qquad \ldots (121)$$

From this results a linear temperature gradient in T^4, if only the energy transport by radiation is taken into account. With equation (121) we can rewrite equation (120) in the final form

$$w_{\mathrm{rad}} = -\frac{16}{3}\frac{bn_m^2 T^3}{k_m}\frac{\mathrm{d}T}{\mathrm{d}x_1} \qquad \ldots (122)$$

It is permitted to add this energy transport caused by radiation to the one carried by electrons and phonons and thus define the total thermal conductivity for a linear temperature gradient, provided k_m is sufficiently high. In this case only neighbouring volume elements mutually interchange radiation. In small distances the difference between temperature gradients which are linear in T^4 and in T, respectively, becomes small.

According to Joffé, Devyatkova, and Moyges[85] the measured thermal conductivities of tellurium crystals may be explained quantitatively by taking into account the energy transport by radiation with the formula (122).

8.2. Excitons

As Haken[86] pointed out for excited electron states called excitons it is characteristic that:

(1) The excited electron state remains near the corresponding hole state;
(2) Excitation energy is transferred through the lattice, thus omitting a charge transport.

Of particular interest in connection with thermal conduction is point (2). It starts from the fundamental result according to which the excited states in a perfect lattice cannot be exactly located. In other words, excitons always have a finite travelling or diffusion velocity. An empirical verification for this fact in the case of alkali halides is given by experiments of Apker and Taft[87] who investigated the external photoelectric effect on crystals with F-centres. If such crystals are exposed to light radiation in the maximum of (KJ: $h\nu = 5{\cdot}66$ eV) absorption a distinct maximum of emissivity, electrons per irradiated light quantum, is observed. The authors explain this observation by assuming that the excitons, primarily created by irradiation, travel through the lattice and recombine on F-centres. Thus electrons in a defined energy range are emitted. Another convincing indication concerning the diffusion of excitons is given by Seitz.[88] He explains the creation of F-centres at very low temperatures in the following way. Excitons are formed by irradiation, they diffuse through the crystal until reaching a dislocation. Here they recombine, spending their excitation energy in building up F-centres.

A sufficiently high diffusion velocity together with an appropriate concentration of excitons should influence thermal conduction. It is reasonable that in certain semiconductors at higher temperatures the thermal energy of the lattice does not suffice for creating free charge carriers, though it might suffice for creation of

excitons if there exist such excited states. The excitons, under the influence of a temperature and concentration gradient, would travel through the crystal and thus transfer thermal translation energy as well as internal excitation energy. For the corresponding thermal conductivity Pikus[89] notes:

$$\chi_{\text{exc.}} = \frac{16\pi}{3}\left(\frac{k}{h}\right)^3 T^3 m_{\text{exc.}} l_{\text{exc.}} \left[\left(\frac{\Delta\epsilon_{\text{exc.}}}{kT}+2\right)^2+2\right]\exp\left(-\frac{\Delta\epsilon_{\text{exc.}}}{kT}\right) \quad \ldots (123)$$

where $m_{\text{exc.}}$ is the effective total mass of an exciton and $\Delta\epsilon_{\text{exc.}}$ is the internal excitation energy. The mean free path of excitons, $l_{\text{exc.}}$, has been calculated by several authors[83] in different ways. The investigations hitherto carried out have in common as an important presumption the fact that the excitons are elastically scattered by the phonons. Thus the possible change of internal excitation energy in scattering processes is neglected. This neglect is certainly allowed in non-polar crystals where the scattering of excitons by the acoustical phonons predominates. On the other hand, in polar crystals this neglect, in case of equal effective masses of electrons and holes, $m_n = m_p$, results in an indefinitely large free path, because the exciton appears to the polarization vibrations of the optical branch as a neutral particle for all wavelengths. Therefore, in polar crystals the hitherto known results for $l_{\text{exc.}}$ are applicable only if the effective masses of electrons and holes are substantially different.

Up to now only Deviatkova[90] in experiments on lead telluride observed a contribution to thermal conduction which might be due to excitons. Accepting this, the interpretation of the experiments yields $\Delta\epsilon_{\text{exc.}} = 0{\cdot}12$ eV. The band distance of lead telluride amounts to $0{\cdot}35$ eV. For the mean free path of the excitons with $(123; m_{\text{exc.}} \simeq 10^{-27}$ g) at 300° K a value of 5×10^{-6} cm is obtained.

9. THERMAL CONDUCTION AND THERMOELECTRIC COOLING

Out of the different thermoelectric effects, the Seebeck effect and the Peltier effect are particularly suited for practical applications. We want to demonstrate in the case of thermoelectrical cooling the importance of thermal conduction by taking advantage of the Peltier effect.

The theoretical examinations of thermoelectric cooling with regard to cooling efficiency or to the attainable temperature difference have been summarized by Joffé,[91] Justi,[92] and Lautz.[93] The simple presuppositions of the first theoretical paper of Altenkirch[94] on Peltier cooling with regard to the heat balance of the cold junction read, in the stationary state (see Figure 1):

$$\tfrac{1}{2}\,\text{Joule's heat} + \text{heat balance by thermal conduction} + \text{cooling efficiency} = \text{Peltier cooling}\,(= QTI)$$

The cooling efficiency, i.e. the quantity of heat taken from the heat reservoir at the cold junction per time unit, is proportional to the current $I[A]$ and the differential thermoelectric power Q on the right-hand side. The efficiency decreases with increasing irreversible heat-losses on the left-hand side. Thus, the significance of thermal conduction is evident. The detailed calculation shows that

Table 3. Electrical and Thermal Properties of the System Bi_2Te_3–Sb_2Te_3; Measuring Temperature: 300° K (after Birkholz[96])

Composition (atomic per cent)		Q (μV deg.K^{-1})	$\sigma_p \times 10^{-2}$ (Ω^{-1} cm^{-1})	$\chi \times 10^2$ (W cm^{-1} deg.K^{-1})	$\chi_{ph} \times 10^2$ (W cm^{-1} deg.K^{-1})	WFL-number $\times 10^8$ (V^2 deg.K^{-2})	R_H (cm^3 A^{-1} sec^{-1})	μ_p (cm^2 V^{-1} sec^{-1})	$p \times 10^{-19}$ (cm^{-3})	$z \times 10^3$ (deg.K^{-1})
Bi_2Te_3	Sb_2Te_3									
100	0	+240	5·25	2·0	1·77	12·8	+0·602	(267)	(1·23)	1·51
90	10	+227	5·64	1·82	1·66	16·3	–	–	–	1·59
80	20	+202	7·35	1·78	1·49	8·5	+0·598	372	1·25	1·68
60	40	+153	11·35	–	–	6·3	+0·340	327	2·17	–
50	50	+153	13·2	1·75	1·19	4·65	+0·273	302	2·73	1·77
40	60	+146	16·65	1·85	1·09	3·55	+0·202	285	3·65	1·91
30	70	+146	20·0	2·08	1·10	3·14	+0·144	245	5·11	2·05
20	80	+136	25·0	2·26	1·16	3·06	+0·102	216	7·30	2·04
10	90	+108	38·0	3·42	1·76	2·67	+0·084	271	8·80	1·30
0	100	+79	43·8	4·38	2·36	3·20	+0·084	312	8·80	0·74

the maximum cooling efficiency and the maximum temperature difference $(\Delta T)_{max}$ (Figure 1, cooling efficiency = 0) are given by the quantity

$$z = \frac{Q^2}{[\sqrt{(\chi_1/\sigma_1)}+\sqrt{(\chi_2/\sigma_2)}]^2}, \qquad Q = Q_1 - Q_2 \qquad \ldots (124)$$

Q_1 and Q_2 are the differential absolute thermoelectric powers having reversed signs for a P-type conductor and an N-type conductor.[95] This yields

$$(\Delta T)_{max} = \frac{z}{2} T_k$$

where T_k = temperature of the cold junction.

Therefore, semiconductors having low thermal conductivity, high electrical conductivity, and high thermoelectric power are of interest for thermoelectric cooling. The system Bi_2Te_3–Sb_2Te_3, which forms solid solutions, as shown in Table 3,[96] exemplifies how, by doping, an optimum might be obtained in the sense of a maximum z-value.

With regard to the thermal conductivity of such solid solutions only empirical rules can be given. The theoretical examination of the thermal conductivity of mixed crystals is a complicated matter for various reasons.

(1) At higher temperatures the phonon–phonon interaction in a perfect crystal cannot be described by a relaxation time or a free path. Therefore, even if in an imperfect crystal we could calculate a relaxation time for scattering of phonons by foreign atoms, it is not possible to define a mixed relaxation time according to equation (87).

(2) On substituting foreign atoms in a perfect crystal or a binary compound a description of the additional scattering of phonons by foreign atoms cannot be outlined by a free path, which is calculated by means of time dependent perturbation theory. The perturbation of the stationary states, especially the change of the lattice dispersion due to mass differences and to the change of linkages between the lattice particles is so strong that the application of a perturbation calculation is generally not permitted.

(3) The concentration of the free charge carriers in the mixed crystals is of the order of magnitude 10^{18} to 10^{19} cm^{-3}. Degeneration starts, and at low temperatures the reduction of the insulator lattice conductivity due to the scattering of phonons by electrons must be taken into account.

ACKNOWLEDGEMENTS

The interest of Professor Krautz in this work is gratefully acknowledged. I wish to thank my colleagues J. Erdmann and H. Schultz for several stimulating discussions.

REFERENCES

1. L. Onsager. *Phys. Rev.* **37**, 405 (1931); **38**, 2265 (1931)
2. J. A. Krumhansl. *J. Phys. Chem. Solids* **8**, 343 (1959)
3. S. R. de Groot. *Thermodynamics of Irreversible Processes* (North-Holland, Amsterdam, 1952)
4. J. Meixner. *Z. phys. Chem.* **B53**, 235 (1943)

5. R. Becker. *Theorie der Wärme* (Springer, Berlin, 1955); H. Kümmel. *Z. Phys.* **143**, 219 (1955); L. van Hove. *Physica, 's Grav.* **21**, 517 (1955); W. Kohn and J. M. Luttinger. *Phys. Rev.* **108**, 590 (1957); E. N. Adams. Private communication of April 1959
6. E. H. Sondheimer. *Proc. roy. Soc.* **234**, 391 (1956)
7. H. B. G. Casimir. *Rev. mod. Phys.* **17**, 343 (1945)
8. A. F. Koster. *Solid State Physics* 5 (Academic, New York, 1957)
9. H. Winston and R. S. Halford. *J. chem. Phys.* **17**, 607 (1949)
10. W. Meissner and M. Kohler. *Handb. d. Exp. Phys.* (Leipzig, 1935)
11. H. Bethe. *Ann. Phys.* **3**, 133 (1929)
12. K. Müller. *Abh. braunschw. wiss. Ges.* **9**, 115 (1957); *Diplom-Arbeit*, Braunschweig (1950)
13. H. J. Juretschke. *Notes for Physics* 8493 (Polytechnic Institute of Brooklyn, 1951)
14. L. Tisza. *Z. Phys.* **82**, 48 (1933)
15. H. Haken. *Halbleiterprobleme* II (Vieweg, Braunschweig, 1955); S. J. Pekar. *Fortschr. Phys.* **1**, 367 (1954)
16. M. Born and K. Huang. *Dynamical Theory of Crystal Lattices*, p. 166 (Clarendon, Oxford, 1954); A. Haug and A. Sauermann. *Z. Phys.* **153**, 269 (1958); B. Goodman. *Phys. Rev.* **110**, 888 (1958); A. V. Chester and A. Houghton. *Proc. phys. Soc. Lond.* **73**, 609 (1959)
17. H. Koppe. *Ergebn. exakt. Naturw.* **23**, 283 (1950)
18. J. C. Slater. *Handbuch der Physik* **19** (Springer, Berlin, 1956)
19. F. Hund. *Z. Phys.* **119**, 99 (1936)
20. L. P. Bouckaert, R. S. Smoluchowski, and E. Wigner. *Phys. Rev.* **50**, 58 (1936); C. Herring. *Phys. Rev.* **52**, 361 (1937)
21. E. O. Kane. *J. Phys. Chem. Solids* **8**, 38 (1958) (here one finds further literature); I. A. Firsov. *J. exp. theor. Phys.* **32**, 1350 (1957)
22. F. Herman. *Phys. Rev.* **93**, 1214 (1954)
23. O. Madelung. *Z. Naturf.* **13a**, 22 (1958)
24. A. J. E. Foreman and W. M. Lomer. *Proc. phys. Soc. Lond.* **B70**, 1143 (1957)
25. B. N. Brockhouse. *J. Phys. Chem. Solids* **8**, 400 (1959); B. N. Brockhouse and P. K. Iyengar. *Phys. Rev.* **111**, 747 (1958)
26. A. Ghose, H. Palevsky, P. J. Hughes, I. Pelah, and C. M. Eisenhauer. *Phys. Rev.* **113**, 49 (1959)
27. E. Schlömann. *Dissertation*, Göttingen (1954)
28. A. H. Wilson. *Theory of Metals* (Cambridge University Press, London, 1953)
29. L. I. Schiff. *Quantum Mechanics* (McGraw-Hill, New York, 1955)
30. P. Debye. *Vorträge über die kinetische Theorie der Materie* (Teubner, Leipzig, 1914)
31. J. O. Hirschfelder and C. F. Curtiss. *Molecular Theory of Gases and Liquids*, p. 10 (Chapman and Hall, London, 1954)
32. R. Peierls. *Ann. Phys., Lpz.* **3**, 1055 (1929)
33. R. Peierls. *Quantum Theory of Solids* (Clarendon, Oxford, 1956)
34. I. Pomeranchuk. *Phys. Rev.* **60**, 821 (1941); *J. Phys. (USSR)* **4**, 259 (1941); **6**, 237 (1942)
35. R. Becker and G. Leibfried. *Z. Phys.* **125**, 347 (1948); *Phys. Rev.* **60**, 34 (1946)
36. G. Leibfried and E. Schlömann. *Nachr. Akad. Wiss. Göttingen* **IIa**, 71 (1954)
37. G. Süssmann. *Z. Naturf.* **11a**, 1 (1956)
38. G. Leibfried. *Handbuch der Physik* **7** (Springer, Berlin, 1955)
39. A. Herpin. *Ann. Physique* **7**, 91 (1952)
40. C. Herring. *Phys. Rev.* **95**, 954 (1954)
41. P. G. Klemens. *Handbuch der Physik* **14** (Springer, Berlin, 1956); P. G. Klemens. *Proc. roy. Soc.* **A208**, 108 (1951)
42. G. Hamel. *Integralgleichungen* (Springer, Berlin, 1949)
43. R. Berman, F. E. Simon, and J. M. Ziman. *Proc. roy. Soc.* **A220**, 171 (1953)
44. G. K. White and S. B. Woods. *Phil. Mag.* 342 (1958); S. Shalyt. *J. Phys. (USSR)* **8**, 315 (1944)
45. J. Erdmann. *Dissertation*, Braunschweig (1957)
46. W. Voigt. *Lehrbuch der Kristallphysik* (Teubner, Leipzig, 1910)
47. J. Appel. *Z. Naturf.* **12a**, 410 (1957)
48. M. Kohler. *Z. Phys.* **124**, 772 (1948); **125**, 670 (1949)
49. D. Dorn. *Z. Naturf.* **12a**, 739 (1957)
50. P. Klemens. *Proc. roy. Soc.* **A68**, 1113 (1955)
51. H. B. G. Casimir. *Physica, 's Grav.* **5**, 495 (1938)
52. G. A. Slack. *Phys. Rev.* **105**, 831 (1957)
53. R. Berman. *Adv. Phys.* **2**, 103 (1953)
54. R. Berman, E. L. Forster, and J. M. Ziman. *Proc. roy. Soc.* **A231**, 130 (1955)

55. G. A. Slack. *Phys. Rev.* **105**, 827 (1957)
56. G. Fischer, G. K. White, and S. B. Woods. *Phys. Rev.* **106**, 480 (1957)
57. G. K. White and S. B. Woods. *Phys. Rev.* **103**, 569 (1956)
58. R. Berman, E. L. Forster, and J. M. Ziman. *Proc. roy. Soc.* **237**, 344 (1956)
59. P. G. Klemens. *J. Phys. Chem. Solids* **8**, 345 (1959)
60. T. H. Geballe and G. W. Hull. *Phys. Rev.* **110**, 585 (1958)
61. A. H. Wilson. *The Theory of Metals* (Cambridge University Press, London, 1953)
62. H. J. Goldsmid. *Proc. phys. Soc. Lond.* **69**, 203 (1956); **72**, 17 (1958)
63. P. J. Price. *Phil. Mag.* **46**, 1252 (1955)
64. A. F. Joffé et al. *Canad. J. Phys.* **34**, 1342 (1956)
65. O. Madelung. *Handbuch der Physik* **20** (Springer, Berlin, 1957)
66. P. J. Price. *Proc. phys. Soc. Lond.* **B69**, 8 (1956)
67. Y. Kanai and R. Nii. *J. Phys. Chem. Solids* **8**, 338 (1959)
68. C. B. Satterthwaite and R. W. Ure. *Phys. Rev.* **108**, 1164 (1957)
69. G. Busch and M. Schneider. *Physica, 's Grav.* **20**, 1084 (1954)
70. A. P. Stuckes. *Phys. Rev.* **107**, 427 (1957)
71. H. Weiss. *Halbleiter und Phosphore* (Vieweg, Braunschweig, 1958)
72. F. Kettel. *Tagung der Nordwestdeutschen Phys. Ges.* (Münster, 1954)
73. B. Abeles. *J. Phys. Chem. Solids* **8**, 340 (1959)
74. J. Appel. *Z. Naturf.* **13a**, 386 (1958)
75. J. E. Parrot. *Proc. phys. Soc. Lond.* **70**, 590 (1957)
76. J. A. Carruthers, T. H. Geballe, H. M. Rosenberg, and J. M. Ziman. *Proc. phys. Soc. Lond.* **238**, 502 (1957)
77. H. Fritsche and K. Lark-Horwitz. *Physica, 's Grav.* **20**, 834 (1954)
78. G. Finke and G. Lautz. *Z. Naturf.* **12a**, 223 (1957)
79. R. Fieschi, S. R. de Groot, and P. Mazur. *Physica, 's Grav.* **20**, 67 (1954)
80. M. Kohler. *Ann. Phys., Lpz.* **20**, 878 (1934); **20**, 981 (1934); *Z. Phys.* **95**, 365 (1935)
81. J. P. Jan. *Solid State Physics* 5 (Academic Press, New York, 1957)
82. H. J. Juretschke. *Acta cryst., Camb.* **8**, 716 (1955)
83. O. Madelung. *Z. Naturf.* **13a**, 386 (1958)
84. M. Czerny and L. Genzel. *Glastech. Ber.* **25**, 387 (1952)
85. A. Joffé et al. *J. Phys. Chem. Solids* **8**, 6 (1959)
86. H. Haken. *Fortschr. Phys.* **6**, 271 (1958)
87. L. Apker and E. Taft. *Phys. Rev.* **79**, 964 (1950)
88. F. Seitz. *Rev. mod. Phys.* **26**, 7 (1954)
89. G. E. Pikus. *J. Tech. Phys. (USSR)* **26**, 49 (1956)
90. E. D. Deviatkova. *J. Tech. Phys. (USSR)* **27**, 661 (1957)
91. A. Joffé. *Physik der Halbleiter* (Akademie, Berlin, 1958)
92. E. Justi. *Elektrothermische Kühlung und Heizung*, Arbeitsgemeinschaft f. Forschg. d. Landes Nordrhein-Westfalen, No. 70
93. G. Lautz. *Halbleiterprobleme* IV (Vieweg, Braunschweig, 1958)
94. E. Altenkirch. *Z. Phys.* **12**, 920 (1911); *Z. ges. Kälteind.* **19**, 1 (1912)
95. V. A. Johnson. *Progress in Semiconductors* 1 (Heywood, London; Wiley, New York, 1956)
96. U. Birkholz. *Z. Naturf.* **13a**, 780 (1958)

INDIUM ANTIMONIDE

T. S. MOSS, M.A., Ph.D.

Royal Aircraft Establishment, Farnborough, Hants, U.K.

MS. received August 1959

INDIUM ANTIMONIDE

1. INTRODUCTION AND GENERAL PROPERTIES

Although indium antimonide is a relative newcomer in the field of semiconductors, it has already been the subject of a great deal of intensive research work. Many of its properties have extreme values compared with other well-known semiconductors and already several effects have been observed for the first time (in some cases the only time) with this compound.

Both from the point of view of fundamental measurements and the application in a wide variety of devices, indium antimonide may well be considered the most interesting of all semiconductors.

Many of the interesting properties of indium antimonide stem from its low intrinsic energy gap and from the very low effective mass of the electrons. As a result of the former, the absorption edge lies in an interesting part of the infra-red spectrum, and the material is being used for radiation detectors and filters.

Because of the small effective mass of the electrons, magnetic fields produce large effects. For example, the Landau splitting ($Be\hbar/m^*$) is almost seventy times larger than for an effective mass of unity.† Coupled with the fact that in the neighbourhood of the absorption edge a relatively small energy separation corresponds to a readily resolvable wavelength interval. This means that a variety of magneto-optic effects are well defined in indium antimonide.

The small effective mass is mainly responsible for the very high mobility of indium antimonide—70,000 $cm^2V^{-1}sec^{-1}$ at room temperature—which is, for example, some 20 times higher than in germanium.

Indium antimonide crystallizes in the zinc blende lattice, with the same interatomic spacing as grey tin, namely 2·80 Å. The coefficient of linear expansion is $5{\cdot}5 \times 10^{-6}$ per degree Centigrade at room temperature. An interesting feature of the thermal expansion coefficient is that it drops to zero at 56° K and becomes negative for temperatures below this.[1]

The elastic constants have been studied by Potter[2,3] who used the data to estimate the degree of ionicity of the compound.

2. PREPARATION

Not least among the attractions of indium antimonide is the relative ease with which material of high purity may be prepared.

The normal method of preparation is by direct fusion in vacuo of stoichiometric quantities of commercially available high purity elements, i.e. 'Specpure' quality, followed by zone refining. The low melting point of indium antimonide (523° C) simplifies these processes—for example, radiant heating by nichrome wires or strips is perfectly satisfactory for a zone melter.

Thirty or forty passes through the zone refiner usually produce material which

† Rationalized M.K.S. units are generally used for equations, but popular practical units are used descriptively in the text.

is intrinsic at room temperature, i.e. the net extrinsic carrier concentration is $< 10^{16}$ cm^{-3}. It has been shown by Hulme and Mullin[4] that the most likely impurities then remaining are zinc and cadmium, both of which have segregation coefficients near to unity. These workers show that these impurities may be removed by volatilization, and that if zone refining is carried out after such treatment, material with $< 10^{14}$ cm^{-3} excess donors may be obtained.

The growing of single crystals of the pure material is quite straightforward.[5]

For production of doped N- or P-type material, it is customary to use Group VI and Group II elements respectively. These impurities have low activation energies. According to Edmond,[6] N-type material with at least 8×10^{18} cm^{-3} free carriers can be produced by use of tellurium; whereas P-type material of 2×10^{19} cm^{-3} can be made by doping with magnesium.

3. BAND STRUCTURE

Many of the interesting properties of indium antimonide can be explained from a detailed knowledge of the energy band structure, but conversely, this knowledge is largely gained from detailed experimental studies of electrical and optical properties, and it is not possible to discuss the latest band theory without using some of the most up-to-date experimental data which will only be given in detail in later sections of this review.

The earliest theoretical treatment designed to give a general explanation of the properties of the Group III–V compounds was that of Seraphin.[7] He considered a simple one-dimensional Kronig–Penney type model designed to represent the potential distribution through the crystal in a (111) direction. By perturbing the uniform potential wells representative of a Group IV element, so that adjacent wells were alternately deeper and shallower, he was able to make a representative study of the effect of increasing ionicity. The general trends were:

(1) Steady increase of energy gap over that of the Group IV element;
(2) Steady increase of the effective mass of holes;
(3) Marked decrease of the effective mass of electrons for a small degree of ionicity; after passing through a minimum, the mass increases again for large ionicities.

This simple theory thus explains the facts that the energy gap of indium antimonide is somewhat greater than that of grey tin (0·18 eV compared with 0·09 eV), while the hole mobility is lower than in grey tin. The occurrence of the low electron mass and high electron mobility is also seen to have a fundamental explanation.

Early speculations by Herman[8] were that both conduction and valence band extrema were at $\bar{k}000$, and that the energy surfaces there were spherical. There is now a considerable variety of mechanical, optical, and electrical data which supports this view (see Potter[3] or Tuzzolino,[9] Dumke,[10] and Frederickse and Hosler,[11] respectively).

As in germanium, there should be three valence bands; two degenerate at $\bar{k}000$ but with different hole masses, and a third separated from these by spin–orbit splitting. This splitting is estimated to be fairly large[12] so that this band will not usually have any effect on the properties.

A detailed calculation of the shape of the energy bands has been carried out by Kane[12] using the Kronig–Penney perturbation approach. It was assumed at the outset that the extrema lay at $k000$ and the values of the effective masses and the size of the energy gap were taken as known data. The most important result of this calculation is that the conduction band is shown to be non-parabolic and, in consequence, the effective mass of the electrons is not a single-valued parameter but is energy dependent. Figure 1, in which the energy bands are plotted against k^2, shows that the heavy hole band—which is the one containing the majority of holes—is parabolic, but the light hole band is not. No experimental evidence for

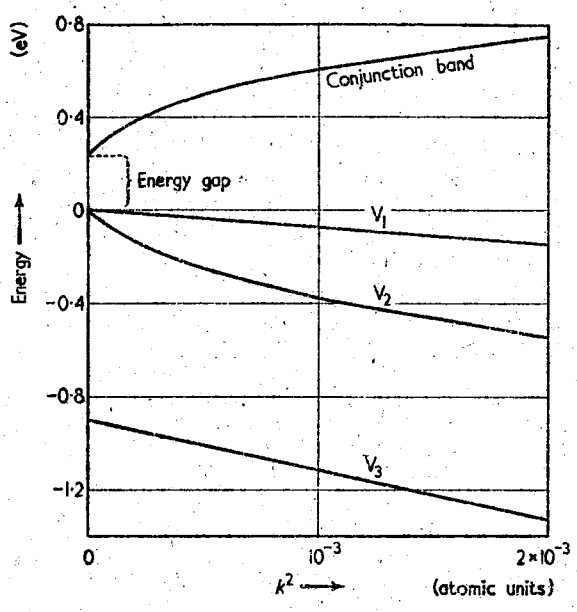


Figure 1. *Energy bands in indium antimonide at 0°K (from Kane[12])*

energy dependence of either hole mass has yet been obtained, however, and there is so far no experimental value for the spin–orbit splitting of the V_3 band.

Kane finds that the conduction band is given by

$$\epsilon'(\epsilon'-G)(\epsilon'+\Delta)-k^2P^2(\epsilon'+2\Delta/3) = 0 \qquad \ldots(1)$$

where G is the gap width at 0° K (0·23 eV), Δ is the spin–orbit splitting, estimated theoretically as 0·9 eV, and ϵ, the energy measured above the valence band, is given by

$$\epsilon = \epsilon' - \hbar^2k^2/2m_0 \qquad \ldots(2)$$

and P^2 is the matrix element given by

$$\frac{2P^2}{3}\left(\frac{2}{G}+\frac{1}{G+\Delta}\right) = \frac{\hbar^2}{m_0^*}+\frac{\hbar^2}{m_0} \qquad \ldots(3)$$

where m_0^* is the effective mass at $k = 0$ and m_0 is the free electron mass.

The term $\hbar^2k^2/2m_0$ is quite negligible compared with ϵ for all relevant energies, so that ϵ may be substituted for ϵ' in equation (1) giving

$$k^2P^2 = \frac{\epsilon(\epsilon-G)(\epsilon+\Delta)}{(\epsilon+2\Delta/3)} \qquad \ldots(4)$$

Now it has been shown by Stephen and Lidiard[13] that if the effective mass is determined from the Faraday effect under degenerate conditions (see sub-section 4.8) the effective mass is not given by the *curvature* of the ϵ–k curve, as is normally the case, but by

$$\frac{\hbar^2}{m^*} = \left(\frac{\mathrm{d}\epsilon}{k\,\mathrm{d}k}\right)_{\mathrm{F}} \qquad \ldots(5a)$$

where the subscript F means that the value is that at the Fermi level. Hence, in atomic units, where $\hbar = m_0 = 1$

$$m^* = \tfrac{1}{2}\left(\frac{\mathrm{d}k^2}{\mathrm{d}\epsilon}\right)_{\mathrm{F}} \qquad \ldots(5b)$$

so that by direct differentiation of equation (4) the theoretical value of the effective mass at any energy level may be obtained.

To a good approximation, it is found that

$$m^{*2} = \frac{A^2G^2(\Delta+G)^2}{4P^4(G+2\Delta/3)^2} - \frac{Ak^2(G+\Delta)}{P^2(G+2\Delta/3)} \qquad \ldots(6)$$

and from this equation we may derive an expression for the zero-k effective mass in terms of the mass found from any given Faraday experiment, namely

$$\frac{1}{m_0^*} = \frac{A}{m^{*2}G}[k^2+\sqrt{(k^4+G^2m^{*2})}] - 1 \quad \text{atomic units} \qquad \ldots(7)$$

where the parameter $A = 1$ for small energies and $A = 0{\cdot}97$ for energies $\simeq 0{\cdot}1$ eV above the bottom of the conduction band. This expression is accurate to better than ± 1 per cent for energies up to $0{\cdot}14$ eV above the conduction band (i.e. the energy range covered by the Faraday effect results given in sub-section 4.8) and is also independent of the spin–orbit splitting except for its slight influence on A. For example, changing Δ from $0{\cdot}5$ eV to $1{\cdot}4$ eV changes m_0^* by less than 1 per cent.

Values of m_0^* calculated by equation (7) from effective masses determined by the infra-red Faraday effect are given in sub-section 4.8. They give the average value $m_0^* = 0{\cdot}0143m_0$.

4. OPTICAL PROPERTIES

Measurements of absorption in indium antimonide have now been made over the wavelength range 2–200 microns and cover three regions of specific interest.

(1) Short wavelength absorption and the absorption edge.
(2) Free carrier absorption.
(3) Reststrahlen absorption.

In addition to absorption measurements, interesting work has been carried out on dispersion, magneto-optic effects, and recombination radiation.

4.1. Short Wavelength Absorption

Values of the absorption coefficient K, obtained by direct transmission measurements through very thin parallel slices of pure single crystal material, obtained by Moss et al.[14] are shown in Figure 2.

The main absorption edge is seen to lie near 7 microns. The wavelength of maximum rate of change of K, which has been shown by Moss[15] to be a reasonable criterion for locating the absorption edge, is 6·94 microns or 0·179 eV. The Figure also shows the absorption calculated by Kane[12] from his band theory on the basis of direct transitions. As the only adjustable parameter in the theory was the energy gap, the agreement in magnitude at short wavelengths is very good.

The measured values of K at 1·5 microns exceed 10^4 cm^{-1}, which is a much higher absorption level than that at which the onset of vertical transitions in

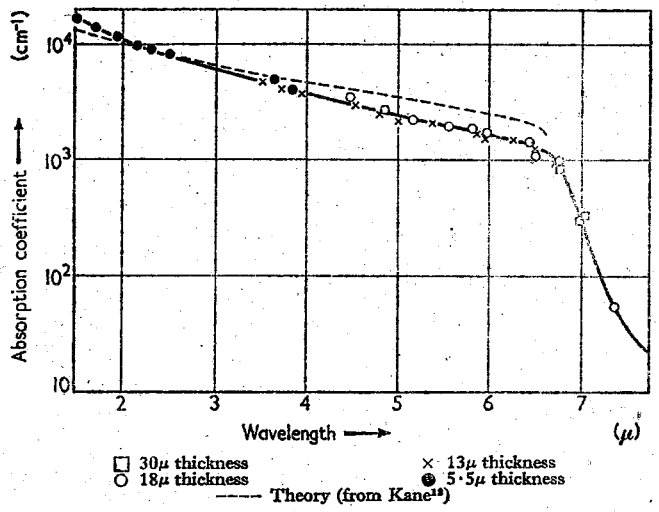


Figure 2. *Absorption in indium antimonide*

germanium and silicon shows a secondary absorption edge. This fact indicates that there is only the one absorption edge, and that the conduction band minimum is therefore vertically above the valence band maximum.

Interpretations to the contrary—in terms of non-vertical transitions involving one, two, or four acoustical mode phonons and valence band maxima away from $k000$—have been made by Roberts and Quarrington,[16] Blount et al.,[17] and Potter.[2] A later treatment of this subject[10] shows that the absorption edge can best be explained by a model with both extrema at $k000$, where, at energies slightly below those of direct transitions, there is absorption due to virtual transitions to intermediate states in the conduction band, followed by scattering by optical mode phonons of energy appropriate to the Reststrahlen wavelength. Dumke's results compare well with the experimental results of Roberts and Quarrington[16] for temperatures of 200°–300° K. Dumke emphasizes that for temperatures of 77° K and less, the *shape* of the absorption curve is independent of temperature, further cooling merely moving it to shorter wavelengths.

A very precise value for the energy gap has been obtained from a study of the oscillatory magneto-absorption effect by Zwerdling et al.[18] Using magnetic fields of 20,000–37,000 gauss and specimens ~ 7 microns thick, they were able to locate four absorption minima in the wavelength range 5–7 microns. When the photon energies of these minima are plotted against magnetic field, it is found that on extrapolation to zero field they all yield the same photon energy. This is the energy gap, its value being $0{\cdot}180 \pm 0{\cdot}002$ eV at 298° K, in good agreement with the value obtained from the maximum slope of the absorption curve.

A shift of the absorption edge in high magnetic fields was observed by Burstein, Picus, and Gebbie.[19] By using fields up to 58,000 gauss, they found that for a constant transmission level there was a shift of $2{\cdot}3 \times 10^{-7}$ eV gauss^{-1}. This is approximately only 60 per cent of the expected shift given by $\frac{1}{2}\hbar\omega_c$ (where the angular cyclotron resonance frequency is $\omega_c = Be/m^*$), possibly because of a variation in absorption magnitude due to change of electron populations produced by the field.[20]

4.2. Pressure and Temperature Dependence of Activation Energy

As there are great technical difficulties in providing window materials transparent at ~ 8 microns which will stand high pressures, measurements of the pressure dependence of the absorption edge have not so far been reported. However, data are available on the pressure dependence of the resistivity and Hall constant.[21–24] The average increase of energy gap with pressure is $d\epsilon/dp = 15 \times 10^{-6}$ eV kg^{-1} cm^{-2}.

From the elastic constants,[2,25] the average compressibility is

$$\chi = 2{\cdot}2 \times 10^{-6} \text{ kg}^{-1}\text{cm}^{-2}$$

Thus, the shift per unit strain is

$$\frac{d\epsilon}{dV/V} = -6{\cdot}8 \text{ eV}$$

and the equivalent temperature dependence is

$$\left(\frac{d\epsilon}{dT}\right)_d = -1{\cdot}1 \times 10^{-4} \text{ eV deg.C}^{-1}$$

The temperature dependence of the energy gap, which has been shown[16] to be linear with temperature above 100° K, has been measured in many ways. The results, which have been summarized by Moss,[15] give an average shift of $-3{\cdot}3 \times 10^{-4}$ eV deg.C^{-1}.

By comparison with the dilatation contribution this figure leads to a lattice broadening contribution of

$$\left(\frac{d\epsilon}{dT}\right)_b = -2{\cdot}2 \times 10^{-4} \text{ eV deg.C}^{-1}$$

4.3. Dependence of Absorption Edge on Impurity Concentration

Tanenbaum and Briggs[26] first observed that the position of the absorption edge in indium antimonide was strongly influenced by the amount of N-type impurity. No significant effect due to P-type impurity was found.

The cause of this shift is now known to be the progressive filling of the lower levels in the conduction bands by extrinsic electrons so that transitions can only take place to the upper, empty, levels.

The degree of filling for a given carrier concentration N can be calculated in terms of the momentum vector at the Fermi level (k_F) since

$$N = \frac{2}{(2\pi)^3}\frac{4\pi k_F^3}{3} \quad \text{giving} \quad k_F = (3\pi^2 N)^{1/3} \qquad \ldots (8)$$

The values of k_F can then be converted to equivalent Fermi levels by use of Figure 1. Some representative figures are given in Table 1.

Table 1. Extent to which Conduction Band is Filled for Various Carrier Concentrations

N (cm^{-3})	k_F (atomic units)	$\Delta\epsilon$ (eV)
10^{16}	0·0036	0·012
10^{17}	0·0077	0·046
10^{18}	0·0167	0·18
10^{19}	0·036	0·42

Thus for 10^{19} electrons cm^{-3}, the absorption edge should have moved to approximately 2 microns—as is observed to be the case from the data of Hrostowski et al.[27] in Figure 11(b).

4.4. Free Carrier Absorption

From the classical treatment of absorption by free carriers,[15] the real part of the dielectric constant is given by

$$2nk\omega = \frac{Ne^2/\tau m^* \epsilon_0}{\omega^2 + \tau^{-2}} \qquad \ldots (9)$$

This relation corresponds to the assumption of an energy independent collision time τ. Several workers[28, 29] have concluded from their analysis of electrical data that for indium antimonide τ is indeed energy independent at least up to room temperature. Thus, for this material, equation (9) should be well obeyed.

Expressing τ in terms of the mobility—again for the case of energy independent scattering—we have

$$\tau = \mu m^*/e$$

$$\text{giving} \quad nK = \frac{Ne^3}{c\epsilon_0 \mu m^{*2} \omega^2} \qquad \ldots (10)$$

for infra-red wavelengths where $\omega \gg 1/\tau$.

Thus for conditions where the absorption is not too intense, so that n does not vary significantly, the absorption constant should increase as the square of the wavelength. Such behaviour is confirmed by the experimental results of Figure

3(a). From the slope of the line, using the measured values of N and μ, and putting $n = 3{\cdot}9$, the calculated effective mass is $m^* = 0{\cdot}023m_0$. This is in very good agreement with the Faraday effect determination of effective mass in a slightly purer specimen listed in Table 2 (page 204).

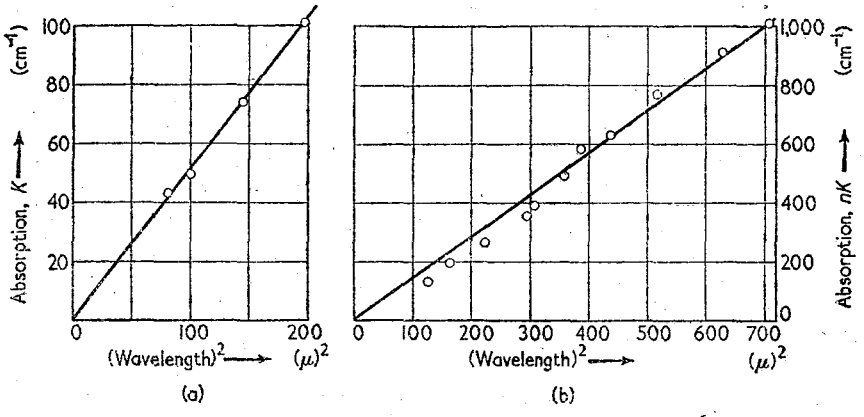


Figure 3. *Wavelength dependence of free carrier absorption in N-type indium antimonide:* (*a*) $N = 3{\cdot}5 \times 10^{17}$ cm^{-3}, $\mu = 18{,}000$ $cm^2V^{-1}sec^{-1}$; (*b*) $N = 6{\cdot}2 \times 10^{17}$ cm^{-3}, $\mu = 25{,}000$ $cm^2V^{-1}sec^{-1}$

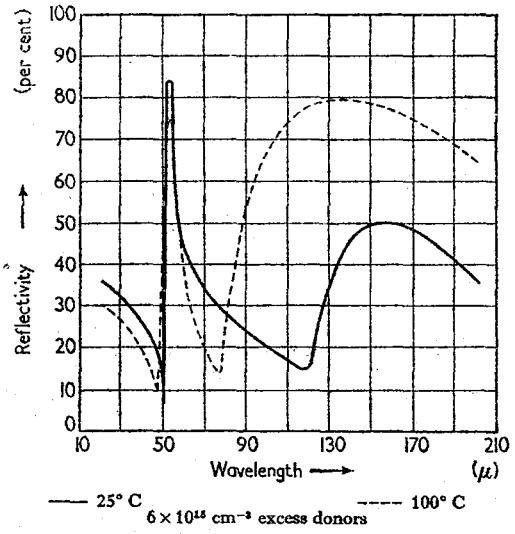


Figure 4. *Long wavelength reflectivity of indium antimonide (from Yoshinaga and Oetjen*[31])

For higher absorption levels it is necessary to allow for the variation in refractive index, by plotting nK against λ^2. As shown by Figure 3(b), which uses data obtained by Spitzer and Fan,[30] the expected proportionality is then obtained for absorption coefficients up to at least 1,000 cm^{-1}. Using the measured parameters

and equation (10) the effective mass for this specimen is found to be $m^* = 0{\cdot}030m_0$, which agrees very closely with the value calculated by Spitzer and Fan[30] from dispersion data on this same specimen, namely $m^* = 0{\cdot}029m_0$.

Yoshinaga and Oetjen[31] have measured the reflectivity of indium antimonide for wavelengths up to 200 microns. From their results, some of which are plotted in Figure 4, they have deduced the absorption coefficient and refractive index. The material used was intrinsic at room temperature and so the absorption was low at short wavelengths. Beyond 70 microns, K increased roughly as λ^2. From their analysis of the data in the 100 micron region, these workers concluded that the optical behaviour corresponded to mobilities and carrier concentrations within 2 per cent of the d.c. values.

4.5. Reststrahlen Absorption

The reflectivity of pure indium antimonide crystals in the region of the lattice absorption band has been measured.[31] At short wavelengths the room temperature reflectivity falls steadily with increasing wavelength to reach a minimum at 49·8 microns (Figure 4). There is then a rapid rise to a maximum reflectivity of 83 per cent at 54·6 microns. This value is presumably more accurate than the 78 per cent quoted earlier by Yoshinaga.[32]

These measured values can be used to derive the parameters of the equivalent classical oscillator using the equations given by Moss.[15] From the value of the maximum reflectivity and the wavelengths of maximum and minimum reflectivity, it is found that the resonance frequency is given by $\omega_0 \equiv 177{\cdot}6$ cm^{-1}, the bandwidth by $1/\tau \equiv 5{\cdot}7$ cm^{-1}, and the amplitude by $A \equiv 12 \times 10^4$ cm^{-2}.

From these values, the magnitude of the reflection minimum and the calculated reflectivity at 40 microns and 60 microns are all in good agreement with the experimental values. Thus it is concluded that a single classical oscillator represents the Reststrahlen band to a good degree of approximation.

These parameters represent a very narrow and intense absorption band, with a maximum absorption coefficient of $K = 1{\cdot}7 \times 10^4$ cm^{-1} at 56·4 microns. The net contribution to the low frequency dielectric constant by the band is given by A/ω_0^2 which has the value 3·7.

The effective value of ionic charge (e^*) may be obtained from this value, using the Szigetti relation

$$\Delta\epsilon = (n_1^2+2)^2 Ne^{*2}/9\mu\epsilon_0\omega_0^2 \qquad \ldots(11)$$

where the reduced mass $\mu = 10^{-25}$ kg, density of 'molecules' $N = 1{\cdot}5 \times 10^{28}$ m^{-3}, and the undispersed refractive index is $n_1 = 3{\cdot}96$.

Thus $e^* = 0{\cdot}5$ free electron charge, in good agreement with the charge deduced by Ehrenreich[33] from mobility data.

4.6. Refractive Index and Dispersion

The dispersion of indium antimonide has been measured accurately by an interference method[14] in the wavelength region 7–20 microns. In this work, measurements were made on specimens so thin that the fringe number could be determined unambiguously from the relation

$$2nd = N\lambda \qquad \ldots(12)$$

where N is the number of the transmission maximum, and d the specimen thickness. Measurements were then repeated on thicker specimens to increase the accuracy of the results.

The results obtained are shown in Figure 5(a). It will be seen that the index falls almost linearly with increasing wavelength between 10 and 20 microns. As there is no region where the index is wavelength independent, it is not possible to determine a specific value for the long wavelength undispersed refractive index directly from the measured data. It is necessary to resolve the dispersion into its four components.

(1) *Tail band absorption.* The main absorption band is arbitrarily treated as two separate bands, the 'tail' band stretching from 1·5 microns to the absorption edge and the 'short waveband' for $\lambda < 1{\cdot}5$ microns. This process is adopted simply because the absorption has been measured from 1·5 to 8 microns and these data can thus be used to compute the dispersion. It has been shown by the author[15] that the index at wavelength λ_a is related to the integrated absorption as follows:

$$n_a - 1 = \frac{1}{2\pi^2} \int_0^\infty \frac{K d\lambda}{1-\lambda^2/\lambda_a^2} \qquad \ldots (13)$$

The results of this integration show that most of the dispersion between 6 and 10 microns arises from this absorption region.

(2) *Short waveband absorption.* No data are available on the wavelength region below 1·5 microns, but as its contribution for $\lambda > 7$ microns is small, it can be calculated with sufficient accuracy by treating the absorption as that of a simple classical oscillator. The contribution to the dielectric constant by this band is found to be

$$\Delta\epsilon = 3{\cdot}5/\lambda^2$$

with λ in microns.

(3) *Free carrier dispersion.* This is given by equation (10) with $N = 1{\cdot}6 \times 10^{22}$ m^{-3} and $m^*/m_0 = 0{\cdot}0143$, giving

$$\Delta\epsilon = 9{\cdot}5\lambda^2 \times 10^{-4}$$

with λ in microns.

(4) *Reststrahlen band.* From the parameters calculated earlier for the Reststrahlen band we have

$$\Delta\epsilon = -4(\lambda_0^2/\lambda^2 - 1)$$

where $\lambda_0 = 56{\cdot}4$ microns. This term is somewhat larger than the free carrier contribution for intrinsic material.

These four contributions are plotted in Figure 5(b) and their sum is seen to agree quite well with the measured data in Figure 5(a).

As there is no wavelength region where the refractive index becomes constant, it is difficult to specify a unique value for the 'long wavelength refractive index'. Study of the individual contributions, however, shows that if there were no free

carrier or Reststrahlen dispersion, the index would have a long wavelength value equal to that measured at 9·8 microns, i.e.

$$n_1 \equiv n_{9\cdot8} = 3\cdot96$$

The dispersion due to free carriers has been studied by Spitzer and Fan,[30] who point out that the theoretical dispersion equation, i.e.

$$n^2 = \epsilon - \frac{Ne^2/m^*\epsilon_0}{\omega^2 + 1/\tau^2} \qquad \ldots (14)$$

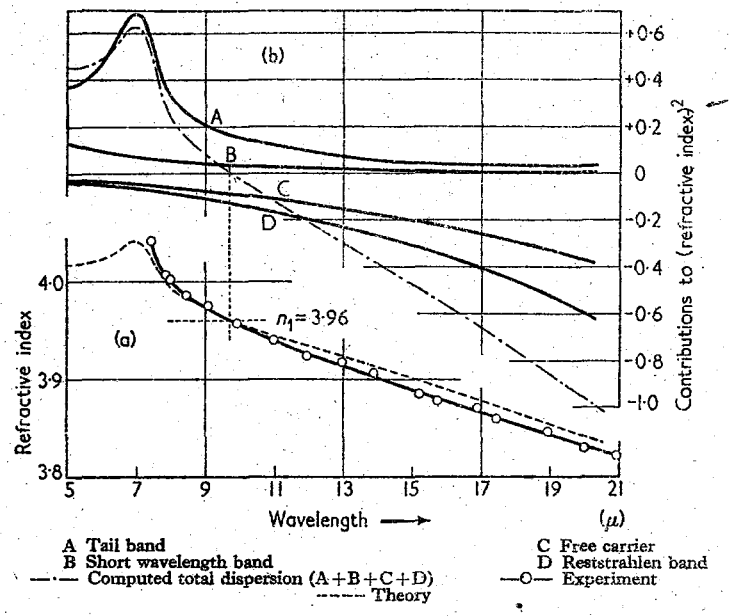


Figure 5. *Measurement and analysis of refractive index of pure indium antimonide*

is independent of the collision time τ for the infra-red region where $\omega \gg 1/\tau$. Thus, even if τ is energy dependent so that the free carrier absorption no longer follows a simple λ^2 law, the dispersion should show strict proportionality to λ^2.

Such behaviour was verified experimentally for both P- and N-type material. For the former, although K was virtually independent of λ, the dispersion followed a good λ^2 law from which an effective hole mass $m_h = 0\cdot2m_0$ was obtained. For electrons, as discussed in Section 3, the effective mass is energy dependent due to the non-parabolic nature of the conduction band. In these conditions, the effective mass found from equation (14) is that given by equation (5*a*). Measurements were made on material containing various electron concentrations. For the two most impure samples, the masses found were

$$N = 2\cdot8\times10^{18}\ \text{cm}^{-3}, \qquad m_e^* = 0\cdot040m_0$$

$$N = 4\cdot0\times10^{18}\ \text{cm}^{-3}, \qquad m_e^* = 0\cdot041m_0$$

Use of equations (7) and (8) for the values of k appropriate to these concentrations shows that the masses should be $0 \cdot 042\, m_0$ and $0 \cdot 48\, m_0$, respectively. The former value is in good agreement with experiment while the latter indicates that at the highest energy levels the mass is not increasing as much as expected, and is perhaps approaching a constant, saturation, value.

4.7. Recombination Radiation and Emissivity

Both recombination radiation and emission from heated material have been measured for indium antimonide, and it has been shown that the spectral distributions in the two cases are closely related.[15]

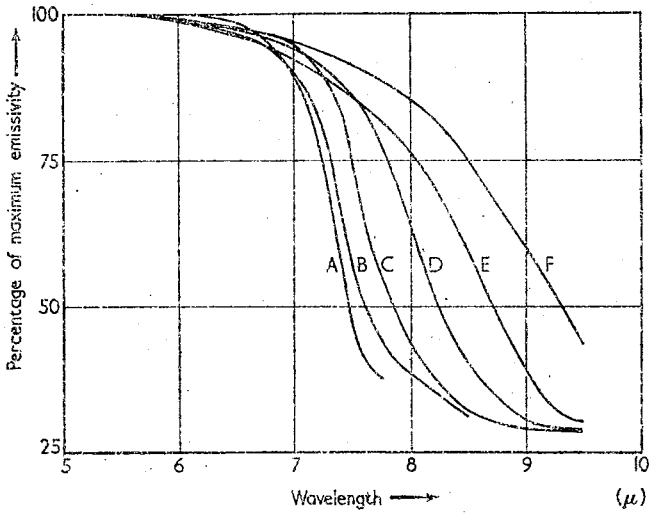


Figure 6. *Measurements of the emissivity of single crystals of indium antimonide by Moss and Hawkins*

From the absolute intensity of the recombination radiation it is estimated that about 20 per cent of the recombinations are radiative. The percentage of radiative recombinations can also be estimated from a comparison of observed lifetimes with the theoretical radiation lifetime calculated from absorption data. This data gives $\tau_R = 0 \cdot 79\ \mu$sec, while even for good quality crystals the longest observed lifetimes are $\sim 4 \times 10^{-8}$ sec. Hence the percentage of radiant recombinations appears to be only about 5 per cent. Landsberg[34] has suggested that this discrepancy could be due to the fact that the short wavelength light, used to generate excess carriers in the recombination radiation experiment, produced more than one electron–hole pair per quantum, and such behaviour has now been observed.[35]

Measurements of the emissivity of single crystals of indium antimonide by Moss and Hawkins[36] are shown in Figure 6. As expected, the emissivity falls rapidly in the neighbourhood of the absorption edge. From the graph the shift of the 'emission edge' is found to be $-2 \cdot 9 \times 10^{-4}$ eV deg.C^{-1}.

4.8. Magneto-optic Effects

The underlying cause of the various magneto-optic effects is the coalescing of the continuum of allowed levels in the conduction and valence bands into relatively narrow sub-bands which are separated by the Landau splitting of $\hbar\omega_c$, where ω_c is the angular cyclotron resonance frequency given by

$$\omega_c = \frac{Be}{m^*}$$

In indium antimonide, this splitting can be particularly large for electrons, because of their very low mass.

These levels are involved in the three main magneto-optic phenomena in the following ways.

(1) *Oscillatory magneto-absorption.* If transmission is measured near the main absorption edge then an oscillatory spectrum will be obtained as transitions occur between the various valence and conduction band levels. Making the simplifying assumption that the splitting in the valence band is negligible due to the relatively large carrier mass, then transitions will occur from a unique level at the top of the valence band to the various levels in the conduction band, giving absorption peaks at the appropriate photon energies. It is necessary to use good spectral resolution and high magnetic fields ($\sim$ 30,000 gauss, so that $\omega_c \gg 1/\tau$) in order to get well resolved spectra.

The effect was first observed by Burstein and Picus[37] and has since been studied extensively by Lax[38] and his co-workers. If the photon energies for various absorption maxima are plotted against magnetic field they fall approximately on straight lines and may all be extrapolated backwards to zero field to give the same value of photon energy. This is the true activation energy of the semiconductor, the value obtained being 0·180 eV. In view of the energy dependence of effective electron mass there is some non-linearity in the energy–magnetic field relation and hence some uncertainty in the extrapolation. However, in spite of this, the method yields the most precise figure yet obtained for the energy gap.

(2) *Cyclotron resonance.* In cyclotron resonance experiments, absorption occurs due to transitions between the Landau levels within a given band. As these levels are relatively close together, the absorption frequency is very low unless the magnetic field is very high. Even for electrons in indium antimonide, with their very low mass, it is necessary to use magnetic fields $\sim 10^5$ gauss to produce resonance at convenient infra-red wavelengths. Such measurements have been performed by Burstein, Picus, and Gebbie[19] and Keyes et al.[39] The latter workers, who used pulsed magnetic fields up to 3×10^5 gauss, were able to observe the increase in effective mass with increasing magnetic field.

Cyclotron resonance measurements have also been carried out at radio frequencies.[40] At these low frequencies, the main difficulty is to ensure that the collision time is sufficiently long to obtain a well-defined resonance, i.e. $\omega_c\tau \gg 1$. On cooling with liquid helium, the above workers observed resonance and so obtained effective masses $m_e = 0{\cdot}013m_0$ and $m_h = 0{\cdot}18m_0$.

(3) *Faraday effect.* The free carrier Faraday effect is essentially a measure of the dispersion caused by cyclotron resonance absorption.[15,41]

From classical electrodynamic theory it is found that the rotation per unit length θ is given by

$$\theta = \frac{BNe^3}{2nc\epsilon_0 m^{*2}\omega^2} \qquad \ldots (15)$$

where the angular infra-red frequency ω is such that $\omega\tau \gg 1$ and $\omega \gg \omega_c$. These inequalities are easy to satisfy by using moderate magnetic fields, convenient infra-red wavelengths, and room temperature, and for this reason the Faraday effect method is the most convenient and flexible of the accurate methods of measuring effective mass. An important feature of equation (15) is that it is independent of the relaxation time, so that the type of scattering occurring is immaterial.

A more rigorous treatment of the Faraday effect[13] shows that, for any degenerate semiconductor with spherical energy bands, an equation identical with equation (15) is obtained where the effective mass is given in terms of the slope of the ϵ–k curve at the Fermi level by equation (5*a*).

Table 2

	Electrons cm^{-3}	m_e^*/m_0	k_F
Faraday	$6{\cdot}4\times10^{17}$	0·029	0·0140
	$2{\cdot}0\times10^{17}$	0·022	0·0096
	$2{\cdot}9\times10^{16}$	0·0178	0·0050
	$1{\cdot}1\times10^{16}$	0·0165	0·0037
	$2{\cdot}6\times10^{15}$	0·0131	0·0023
Dispersion	$3{\cdot}5\times10^{17}$	0·023	0·0115
	$6{\cdot}5\times10^{17}$	0·029	0·0141
	$1{\cdot}2\times10^{18}$	0·032	0·0175
	$2{\cdot}8\times10^{18}$	0·040	0·0232

By measuring the Faraday rotation in specimens with various electron concentrations, Smith et al.[42] have made a detailed study of the energy dependence of the mass and hence of the deviation of the ϵ–k curve from the simple parabolic form. The results are summarized in Table 2, together with some dispersion results by Spitzer and Fan[30] (see sub-section 4.6).

Analysis of the results using equation (7) gives a mean value for the zero-k mass of $0{\cdot}0143m_0$.

Large Faraday rotations are readily obtainable with indium antimonide. For an absorption loss of e:1, the theoretical rotation is simply μB radians. Thus for 10^4 gauss, relatively pure indium antimonide gives a rotation of 7 radians, or just over 360°.

5. PHOTOEFFECTS

Radiation of wavelengths below that of the absorption edge is absorbed by indium antimonide with the generation of hole–electron pairs which may be made to give useful signals in three distinct ways.

(1) Photoelectromagnetic effect (PEM).
(2) Photoconductivity (PC).
(3) Photovoltage or photocurrent at a P–N junction.

It transpires that for indium antimonide all these three modes of operation give the same order of sensitivity as infra-red detectors, and it is therefore of interest to consider the theory underlying all three.

5.1. Photoelectromagnetic Effect

It is convenient to treat this effect prior to simple photoconductivity because the same equations—with the magnetic field made zero—may then be used.

In the PEM effect we consider a slab of material illuminated on its upper surface (of area $A = wX$) as in Figure 7(a), the irradiation being such that q quanta sec^{-1} are absorbed per unit area of surface. Unit quantum efficiency is assumed. Photoelectrons and photoholes generated near the surface diffuse downwards. The

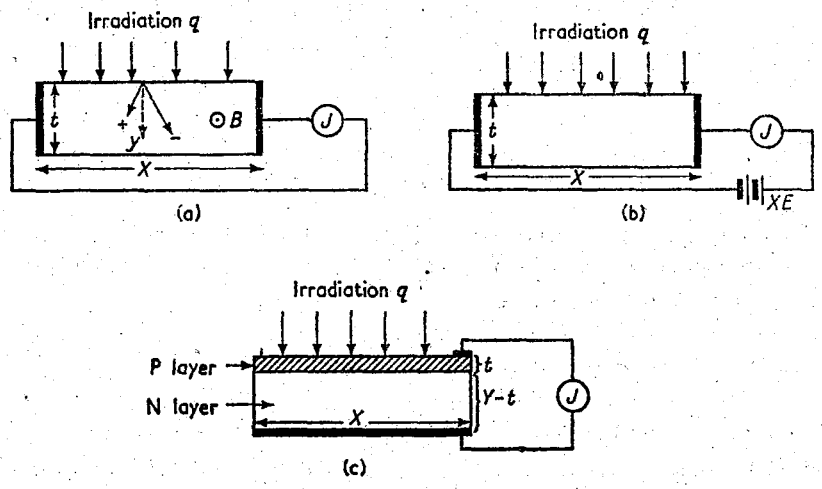


Figure 7. *Schematic arrangements of photodetectors: (a) PEM; (b) PC; (c) P–N junction. All detectors have width w, and q = quanta $cm^{-2}sec^{-1}$*

magnetic field, at right angles to the direction of irradiation and the concentration gradient, deflects the two types of carrier in opposite directions causing a current to flow between the end electrodes.

Following the treatment of Moss,[15] we assume that the radiation is strongly absorbed so that the carriers may be considered to be generated *at* the surface, and obtain for the downward diffusion current density due to holes

$$j_y^+ = \frac{eq\sinh(t-y)/f+\alpha\cosh(t-y)/f}{(1+\alpha^2)\sinh t/f+2\alpha\cosh t/f} \qquad \ldots (16)$$

where the surface recombination velocity s is taken to be the same on each surface, and $\alpha=\tau s/f$. The effective diffusion length is given by

$$f^2 = L^2\Big/\left[1+B^2\mu_p\mu_n\left(\frac{bp+n}{bn+p}\right)\right] \qquad \ldots (17)$$

and the ambipolar diffusion length is given by

$$L^2 = D_n\tau\left(\frac{n+p}{bn+p}\right) \qquad \ldots (18)$$

and $b=\mu_n/\mu_p$.

From equation (16) the PEM short circuit current per unit width of specimen is

$$J_{sc} = \int_0^t j_y^+ \mathrm{d}y = \frac{eqfB(\mu_p+\mu_n)}{\alpha+\coth t/2f} \qquad \ldots (19)$$

If the surface recombination velocity is small so that $\alpha \ll 1$ and the specimen is several diffusion lengths thick, the denominator approaches unity and the PEM current becomes

$$\Delta J = wJ_{sc} = eqwBf(\mu_p+\mu_n) \qquad \ldots (20)$$

This expression can be considered for three conditions of magnetic fields.

(1) Small fields where $B^2\mu_p\mu_n \ll 1$. Then

$$J_1 = eqwB(\mu_p+\mu_n)\left[\frac{(n+p)\tau D_n}{bn+p}\right]^{1/2} \qquad \ldots (21)$$

(2) For the particular case of $B^2\mu_p\mu_n = 1$, which represents a conveniently high permanent magnetic field of about 12,000 gauss, the expression is independent of n and p, namely:

$$J_2 = wBeq(\mu_n+\mu_p)\left(\frac{D_n\tau}{b+1}\right)^{1/2} \simeq wBeq\mu_n(D_p\tau)^{1/2} \qquad \ldots (22)$$

since $b \gg 1$ in indium antimonide.

(3) For very high magnetic fields (only moderately high fields if the material is very P-type) the current saturates at a value

$$J_3 = weq(\mu_n+\mu_p)\left[\frac{D_n\tau(n+p)}{\mu_n\mu_p(bp+n)}\right]^{1/2} \qquad \ldots (23)$$

The predominant noise of a PEM detector should be the Johnson noise in the specimen resistance, i.e.

$$\Delta J^2 = \frac{4kT\Delta f}{R} \qquad \ldots (24)$$

By expressing R in terms of the carrier concentrations and mobilities, the optimum ratio of n to p to give the maximum signal–noise ratio may be found.

For case (1), i.e. small fields, $p/n \simeq 3b$, giving the minimum detectable signal in quanta per second, $Q = qwX$, to be

$$Q_1 = \left(\frac{2\cdot 3}{B\mu_n}\right)(A\Delta f)^{1/2}\left(\frac{tn_i}{\tau}\right)^{1/2} \qquad \ldots (25)$$

For case (2), maximum resistivity material is required, i.e. $p = bn$. With $b \sim 100$ this gives

$$Q_2 \simeq (A\Delta f)^{1/2}\left(\frac{tn_i}{\tau}\right)^{1/2} \qquad \ldots (26)$$

For the high field case (3), the optimum material has $p+n = (b^{1/2}-b^{-1/2})n_i$ giving

$$Q_3 \simeq 0{\cdot}63(A\Delta f)^{1/2}\left(\frac{tn_i}{\tau}\right)^{1/2} \qquad \ldots (27)$$

5.2. Photoconductivity

The photocurrent ΔJ for the conditions of Figure 7(b) is given in terms of equation (16) with $f = L$ as $B = 0$, namely

$$\begin{aligned}\Delta J &= -\tau w E(\mu_p+\mu_n)[j_y^+]_0^t \\ &= \frac{eq\tau w E(\mu_p+\mu_n)}{(1+\alpha\coth t/2L)} \qquad \ldots (28) \\ &\simeq eq\tau w E(\mu_p+\mu_n)\end{aligned}$$

for well-etched specimens which are more than a few diffusion lengths thick.

In terms of the heat dissipation per unit area of specimen h, and the specimen resistance R, the applied field is

$$E^2 = \frac{Rhw}{X}$$

Assuming that Johnson noise is the predominant noise source (as observed by Suits et al.[43] and Oliver[44]) then signal = noise occurs when

$$\Delta J^2 = \frac{4kT\Delta f}{R}$$

which gives a photon signal

$$Q = Xwq = \left(\frac{4kT\Delta f X w}{h}\right)^{1/2}\left(\frac{t}{\tau}\right)\left(\frac{bn+p}{b+1}\right) \qquad \ldots (29)$$

This expression has its optimum value when $p = bn$, giving a minimum detectable signal

$$Q_{\min} = \frac{2b^{1/2}}{b+1}\left(\frac{tn_i}{\tau}\right)\left(\frac{4kT\Delta f A}{h}\right)^{1/2} \qquad \ldots (30)$$

5.3. P–N Junction Detector

Consider a junction consisting of a thin P-layer exposed to the radiation as in Figure 7(c). For these conditions it has been shown[15] that the current density across the junction is given by

$$\frac{j}{e} = Z\left[\exp\left(\frac{eV}{kT}-1\right)\right]-q' \qquad \ldots (31)$$

The parameter Z is given by

$$Z = \frac{p_0 L_h}{\tau_h} + \frac{n_0 L_e}{\tau_e}\left(\frac{\alpha \cosh t/L_e + \sinh t/L_e}{\alpha \sinh t/L_e + \cosh t/L_e}\right) \quad \ldots (32)$$

and

$$q' = \frac{q}{(\alpha \sinh t/L_e + \cosh t/L_e)}$$

where p_0 and n_0 are minority carrier concentrations, the subscript e applies to electrons in the P-region, and $\alpha = \tau_e s/L_e$. It is relatively simple to make the P-layer so thin that $t \ll L_e$ giving

$$\cosh \frac{t}{L_e} \simeq 1 \quad \text{and} \quad \sinh \frac{t}{L_e} \simeq \frac{t}{L_e}$$

and it is also fairly easy to reduce the surface recombination velocity to the level where

$$\frac{\alpha t}{L_e} \ll 1 \quad \ldots (33)$$

so that $q' = q$ and

$$Z = \frac{p_0 L_h}{\tau_h} + n_0\left(\frac{s+t}{\tau_e}\right) \quad \ldots (34)$$

Now the short circuit current density is simply

$$j_{sc} = -eq$$

and, from equation (31), the resistance of the junction is

$$R = \frac{kT}{e^2 wXZ} \quad \ldots (35)$$

Assuming as before that Johnson noise sets the limit of detection we have for signal = noise

$$J_{sc}^2 = \frac{4kT\Delta f}{R}$$

or

$$Q = 2n_i(A\Delta f)^{1/2}\left[\left(\frac{D_h}{n^2\tau_h}\right)^{1/2} + \frac{s}{p} + \frac{t}{p\tau_e}\right]^{1/2} \quad \ldots (36)$$

as $p_0 = n_i^2/n$ and $n_0 = n_i^2/p$ where n and p are majority carrier concentrations.

For good sensitivity of the P–N detector it is clearly necessary for both n and p to be large. At large concentrations, however, recombination rates are high and in fact the lifetime will tend to become inversely proportional to the concentration. Thus $p\tau_e$ would be expected to be roughly constant, and such behaviour has now been verified by Hilsum,[45] who obtains $p\tau_e \sim 10^{-9}$ cm^{-3}sec. The term s/p can be made negligible in comparison with this by good etching if $p \gg 10^{17}$. To make the remaining term small, n should be as large as possible (since n^2 will increase more rapidly than τ_h falls) and for $n > 10^{18}$ cm^{-3} this term will be negligible. Thus

$$Q \simeq 2n_i\left(\frac{A\Delta ft}{p\tau_e}\right)^{1/2} \quad \ldots (37)$$

with $p\tau_e \sim 10^9$ cm^{-3}sec.

For a typical detector with $A = 10^{-2}$ cm^2, $\Delta f = 1$ c/s, $t = 10^{-3}$ cm, at room temperature, equation (37) gives Q(P–N) $= 3 \times 10^9$ quanta sec^{-1}.

For the photoconductor the minimum detectable signal will not be quite as low as indicated by equation (30) because of the dependence of τ on p as discussed above. From the results of Hilsum[45] the minimum value of equation (29) will occur near $p/n_i = 3 \cdot 5$, giving Q(PC) $= 2 \cdot 6 \times 10^9$ quanta sec^{-1}.

For the PEM detector, it is clearly necessary to use as high a magnetic field as possible, and equation (26) represents a convenient practical limit. The optimum p–n ratio for this equation is the same as for equation (29) giving Q(PEM) $= 3 \cdot 2 \times 10^9$ quanta sec^{-1}.

All these values are thus very similar, corresponding to approximately 10^{-10} W of 6 micron radiation. So far there are no published data on room temperature P–N junction cells but results published for the other two modes show that sensitivities quite close to the above figures are obtainable (see Figure 12).

For operation with cooling, the PC and P–N detectors, having direct dependence on n_i, will be superior to the PEM cells. Results on cooled P–N detectors have been given by Mitchell et al.,[46] Avery et al.,[47] and Lasser et al.[48] For a 1 mm^2 detector at liquid air temperatures the former workers found signal = noise for 3×10^{-11} W of 2 microns radiation—equivalent to only 10^{-11} W for 6 microns radiation. Other results are plotted in Figure 13 and discussed in sub-section 7.1. As the lifetime increases on cooling[49] this will give an additional improvement to the PC detector.[50]

In all the above theory, the quantum efficiency has been assumed to be unity. This has been shown to be true for wavelengths above 2·5 microns by Tauc,[51] the quantum yield exceeding unity for higher energy photons. For wavelengths > 6 microns the sensitivity of detectors begins to fall because of falling absorption. For the PC detector the sensitivity has fallen to half its maximum value (on a power, not photon, basis) at $\lambda_{1/2} = 7 \cdot 5$ microns. For the PEM detector, the $\lambda_{1/2}$ value is somewhat smaller.

6. ELECTRICAL PROPERTIES

The electrical properties of indium antimonide are of immediate interest because of the very high electron mobility and the high ratio of electron-to-hole mobility.

In view of this high mobility ratio, the conductivity of all pure specimens is dominated by the electron conductivity and it has been difficult to obtain accurate values for the hole mobility. The best values for this mobility seem to be those of Hilsum and Barrie.[29] They obtain from an analysis of Hall conductivity and magneto-resistance data the room temperature (18° C) values $\mu_n = 78{,}000$ and $\mu_p = 750$ cm^{2}V^{-1}sec^{-1} for their purest material, with $b = 104$ for materials with up to 2×10^{15} cm^{-3} excess acceptors. These workers also find $n_i = 1 \cdot 62 \times 10^{16}$ cm^{-3} at this temperature.

The conductivity is given by

$$\sigma = e(bn + p)\mu \qquad \ldots (38)$$

This expression is readily shown to have its minimum value when $p = bn$ if we ignore the dependence of mobility on impurity content, which means that maximum resistance material is far from intrinsic. The magnitudes are $\sigma_{min} = 40$ Ω^{-1}cm^{-1} and $\sigma_i = 200$ Ω^{-1}cm^{-1}.

For materials with various hole concentrations, Hilsum and Barrie[29] find the mobility values given in Table 3.

Table 3. Mobilities in Indium Antimonide

p (cm^{-3})	μ_n(cm^{2}V^{-1}sec^{-1})	μ_p(cm^{2}V^{-1}sec^{-1})	b
$< 10^{14}$†	78,000	750	104
2×10^{15}†	75,000	720	104
$5{\cdot}2 \times 10^{16}$	41,000	660	62
7×10^{16}	37,000	620	60
$1{\cdot}07 \times 10^{17}$	28,500	550	52
$1{\cdot}86 \times 10^{17}$	24,000	520	46

† Excess hole concentration.

Conductivity and Hall constant of various specimens have been measured over a wide temperature range by Hrostowski et al.[52] The results for their purest specimens are shown in Figure 8, from which it can be seen that the P-type

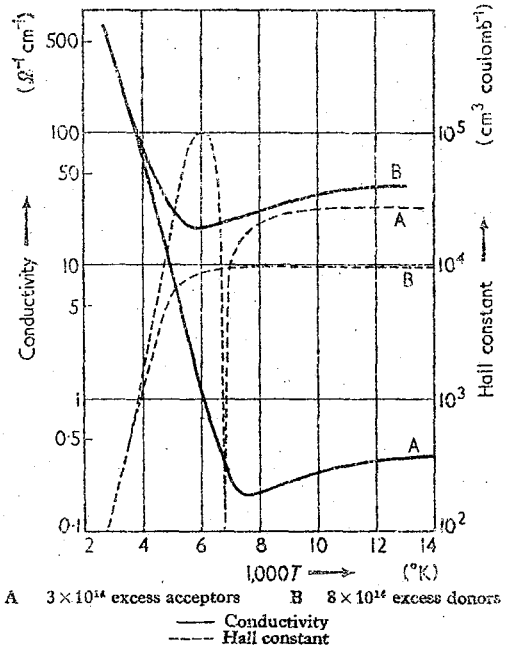


Figure 8. *Temperature dependence of conductivity and Hall constant (from Hrostowski et al.*[52])

specimen is essentially intrinsic for temperatures as low as 145° K, where the Hall reversal occurs. At 78° K the mobility of specimen B was 375,000 cm^{2}V^{-1}sec^{-1}.

For the temperature range 200°–600° K, these workers find that the carrier concentration is given by

$$np = n_i^2 = 3{\cdot}6\times10^{29}T^3\exp(-0{\cdot}26/kT) \qquad \ldots(39)$$

giving an activation energy of 0·26 eV at 0° K.

Although the electron mobility is so high in indium antimonide it has been difficult on theoretical grounds to understand why it was not very much higher. A calculation[21] on the basis of deformation potential scattering (the predominant scattering mechanism in germanium for example) using the pressure data

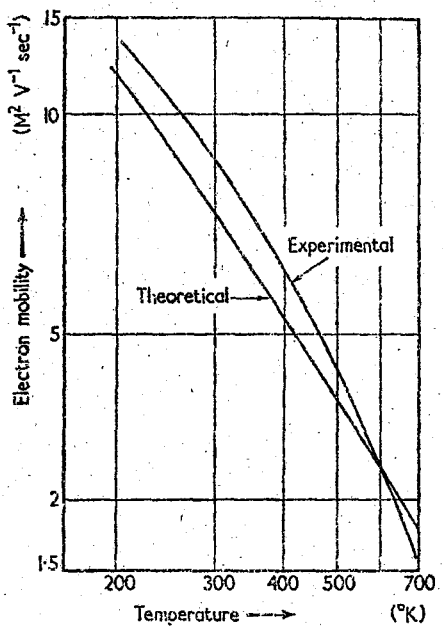


Figure 9. *Comparison of experimental mobility with theory for combined screened-polar and electron–hole scattering (from Ehrenreich[59])*

quoted in sub-section 4.2 gave an expected room temperature mobility of $\sim 10^7\,\mathrm{cm^2V^{-1}sec^{-1}}$.

It has been shown by Ehrenreich[33] that the predominant scattering mechanism is polar scattering by optical modes resulting from the significant degree of ionicity in this compound. This worker uses an effective charge (e_c^*) which is given in terms of that used in equation (11) by the relation

$$e_c^* = \frac{(n_1^2+2)}{3n_1^2}e^* = 0{\cdot}375e^* \qquad \ldots(40)$$

Equation (11) gives $e^* = 0{\cdot}5e$, so that e_c^* should be $0{\cdot}19e$. Ehrenreich finds the best fit to the mobility data is obtained for $0{\cdot}18e$, so that the agreement is very good.

A recent and very comprehensive treatment of transport phenomena in indium antimonide by Ehrenreich[53] includes the influence of the non-parabolic nature of the conduction band, the screening effect of the intrinsic carriers, and electron–hole scattering. This analysis gives $e_c^* = 0 \cdot 20e$, again in very close agreement with the charge obtained from Reststrahlen data. Alternatively if the Reststrahlen value of effective charge is taken as correct, Ehrenreich[53] obtains good agreement between theory and experiment for the temperature variation of both conductivity and thermoelectric power without the use of any adjustable parameters, as shown by the curves of Figures 9 and 10. In the case of thermoelectric power the analysis is

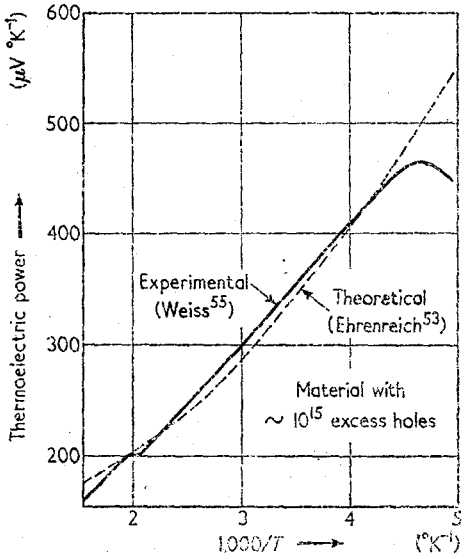


Figure 10. *Comparison of experimental thermoelectric power with theory for combined screened-polar and electron–hole scattering*

particularly important in fitting the experimental data by use of the correct cyclotron resonance value of effective mass, whereas previous simple analyses of thermoelectric data[54, 55] have indicated much higher masses.

Ehrenreich[53] also analyses the effect of scattering on the Hall constant, and finds that if this is written as

$$R = -\frac{A}{ne} \qquad \ldots (41)$$

then the scattering parameter A lies between $1 \cdot 02$ and $1 \cdot 10$ for all temperatures between 200° K and 700° K for intrinsic material. The value is thus always well within the limits of unity (as used for degenerate, energy independent, conditions) and $3\pi/8$ (as appropriate for simple lattice scattering for parabolic bands). As the minimum value $A = 1 \cdot 02$ occurs at 300° K, the use of $R = -1/ne$ for room temperature data is a very good approximation.

7. APPLICATIONS

The very high mobility of electrons in indium antimonide and the small intrinsic energy gap lend themselves to making a variety of solid state devices. These may be considered under three headings.

(1) Optical and photoelectric devices.
(2) Magneto-resistance devices.
(3) Hall effect devices.

So far no indium antimonide transistors have been made, because at room temperature the energy gap is too small for formation of effective P–N junctions. At liquid air temperature, however, it is known that good P–N junctions can be made, and it may yet turn out to be worthwhile making transistors to use at this temperature, because under such conditions electron mobilities at least a hundred times higher than in germanium (at room temperature) are attainable. An Esaki diode with a response time $< 10^{-9}$ sec has been reported.[69]

7.1. Optical and Photoelectric Devices

Indium antimonide is useful as an infra-red filter. For intrinsic material the absorption edge lies at ~ 7 microns, and for a slice of such material 0·1 mm thick the attenuation would be 10^4:1 at 6·7 microns and vastly greater at all shorter wavelengths down to the ultra-violet. The transmission of such an indium antimonide filter, with and without blooming, is shown in Figure 11(a). This filter is well suited for work in the 8–13 microns atmospheric 'window'.

The position of the absorption edge may be moved progressively to shorter wavelengths by making the material more N-type, as shown by the results of

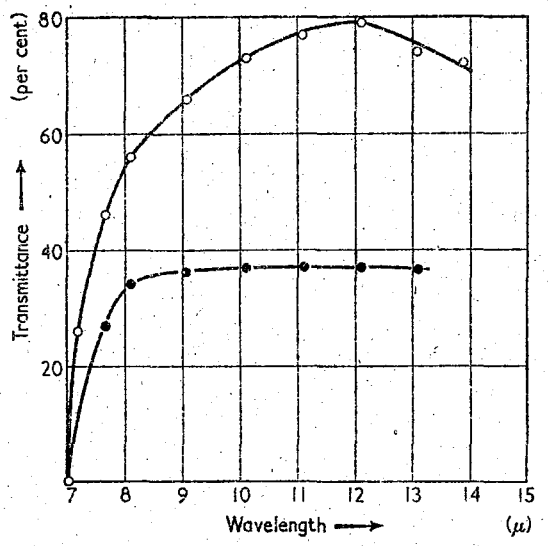
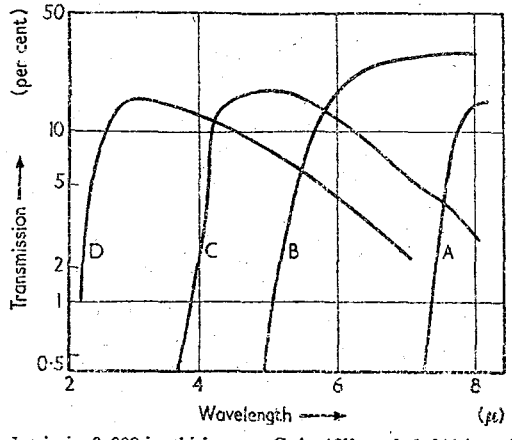


Figure 11(a). *Spectral transmission of indium antimonide*

Hrostowski et al.[27] in Figure 11(b). Thus a relatively sharp absorption edge can be obtained at any desired wavelength between $\simeq 2{\cdot}5$ and $7{\cdot}5$ microns, with positive rejection of short wavelength radiation in all cases.



A Intrinsic, 0·009 in. thick
B 7×10^{17} cm^{-3}, 0·010 in. thick
C 4×10^{18} cm^{-3}, 0·011 in. thick
D $\sim10^{19}$ cm^{-3}, 0·003 in. thick

Figure 11(b). *Spectral transmission of N-type indium antimonide (from Hrostowski et al.[27])*

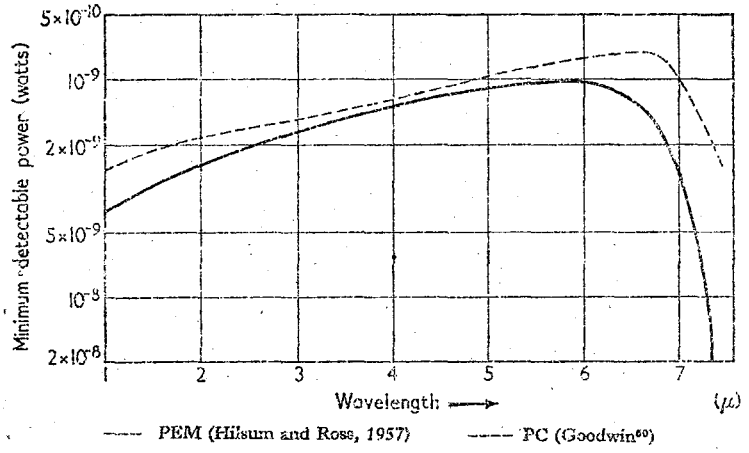
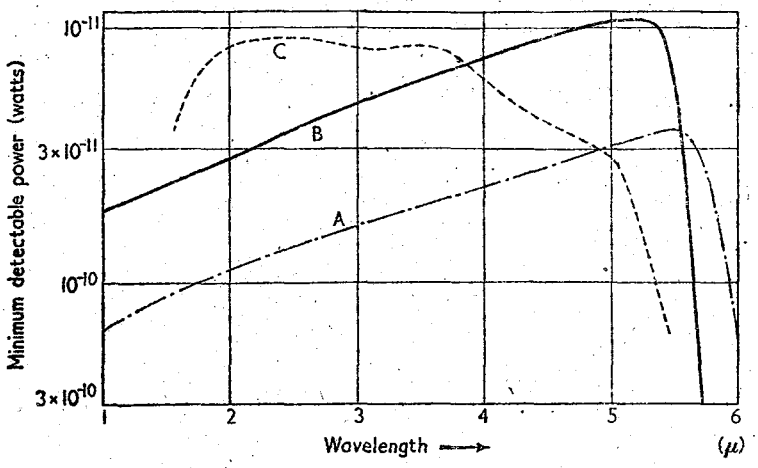


Figure 12. *Spectral sensitivities of indium antimonide detectors at room temperature*

The sensitivity of infra-red photodetectors of this material has already been discussed in Section 5. Absolute spectral sensitivity curves for uncooled PEM and PC detectors shown in Figure 12 indicate that the detection limits of the two types (both $\simeq 1{\cdot}5$ mm^2 in area) are nearly the same at short wavelengths, with some advantage to the PC detector in the neighbourhood of the absorption edge.

Figure 13 shows the absolute sensitivities of various detectors cooled with

liquid air. For wavelengths < 4 microns the surface P–N junction is superior, from 4 to 5·6 microns the edge junction is best, and for the longest wavelengths



A Photoconductor (Goodwin[59]) B P–N junction (edge irradiation) (Avery et al.[47])
C P–N junction (surface irradiation) (Lasser et al.[48])

Figure 13. *Spectral sensitivities of liquid-air cooled indium antimonide detectors*

there is an advantage in using the photoconductor. The areas of these three detectors are: A, 1·4 mm²; B, 0·14 mm²; and C, 4 mm².

7.2. Magneto-resistance Effect

The increase in resistance of a semiconductor under the influence of a magnetic induction B is

$$\frac{\Delta\omega}{\omega} = 0{\cdot}38\mu^2 B^2 \qquad \ldots (42)$$

Thus to obtain large effects, μ should be as large as possible. For N-type indium antimonide with $B = 1$ weber m^{-2} (equivalent to 10,000 gauss), the resistance increase thus reaches 18:1, which contrasts with a mere 6 per cent increase for N-type germanium, for example.

In general, the whole of this resistance increase is not obtained because of limitations imposed by specimen geometry, the restriction being worst for long thin specimens as shown by results of Welker and Weiss[56] in Figure 14. Clearly the Corbino disk should be used when the maximum possible resistance change is required. With such a configuration, increases of resistance of 60:1 have been achieved with fields of 20,000 gauss at room temperature. If it is necessary to use a long thin bar, then the resistance increase can be enhanced by plating metallic rings around the bar so that it becomes effectively several short wide segments in series.[56]

At low temperatures and magnetic fields of 160 kgauss, Figure 15 shows that the resistance increases more than 10,000 times over the zero field value.[57]

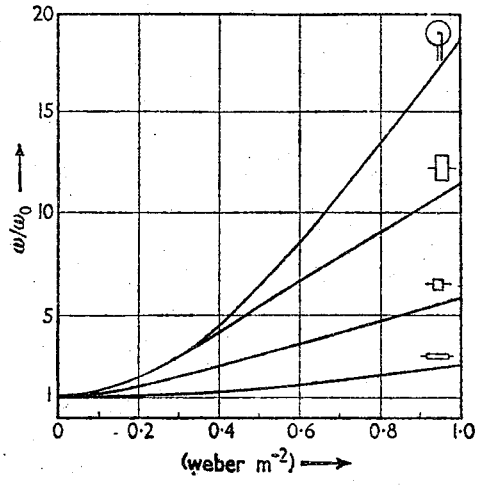


Figure 14. *Magnetic resistance effects in various indium antimonide specimens (from Welker and Weiss[56])*

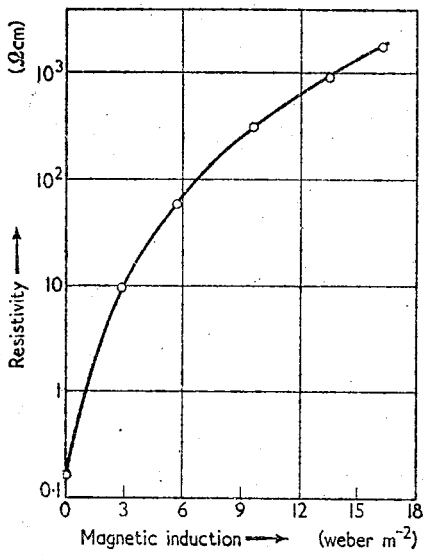


Figure 15. *Magneto-resistance of indium antimonide (from Haslett and Love[57]). Indium antimonide at 4° K. $2{\cdot}8 \times 10^{14}$ excess donors. Electric field 0·27 $V cm^{-1}$*

Various devices using the magneto-resistance effect are as follows.

(1) *Variable resistor.* The resistance can be varied by moving the indium antimonide relative to a permanent magnetic field, or the resistance can be altered by a controlling current in an auxiliary coil. The latter device may be particularly useful for servo systems and for a series control element for high current low voltage systems. Such a variable resistor has no sliding contact.

(2) *Magnetically operated relay.* Change in the resistance of the indium antimonide as it moves into a magnetic field can be made to operate a relay. No direct physical contact is required to operate such a switch, and it can handle considerable power.

(3) *Displacement gauge.* Using indium antimonide elements moving in a non-uniform magnetic field, Ross and Saker[58] obtained a deflection of 5 mm on a robust galvanometer for a 1 micron movement of the stylus, i.e. a linear magnification of 5,000 times. With suitable amplification such a device can be used as a surface roughness meter.

(4) *Microphone or gramophone pick-up.* By moving an indium antimonide probe in a non-uniform magnetic field, resistance modulation will be obtained. By using high currents the output power available from such a pick-up or microphone is potentially much greater than in conventional types, and it is conceivable that a pick-up could be made to drive a loudspeaker directly.

7.3. Hall Effect

For a rectangular specimen, the Hall field E_H is related to the applied field E_x by the expression

$$E_H = \mu_n B E_x \qquad \ldots (43)$$

for an N-type specimen.

The voltage output of a Hall generator, therefore, increases with mobility, so that the advantages of indium antimonide for such devices are obvious. If such generators are considered as power sources rather than voltage sources, then it is readily shown that the efficiency of the device, i.e. the ratio of the output power to the input power, is proportional to $(\mu_n B)^2$, with a maximum possible efficiency of 17 per cent imposed by geometrical considerations.[56]

For applications where stability with temperature is important, it is advisable to avoid intrinsic material for which, by definition, the carrier concentrations vary rapidly. With moderate doping the temperature variation will be greatly reduced for ordinary working temperatures. If the use of intrinsic material is essential, various methods of temperature compensation are described by Saker et al.[59] Some of the practical problems and uses of Hall generators have been discussed by Kuhrt,[60] Hartel,[61] and Weiss.[55] Various applications are listed below.

(1) *Flux meters and compass.* The most direct application of the Hall effect is in measurement of magnetic fields. Already such flux meters using germanium are available commercially. With use of indium antimonide with its greater output power, robust direct reading meters can be used, and smaller fields can be measured. As a natural extension of this application, indium antimonide can be used to measure the earth's field, and so act as an electric compass.[62] With the use of mumetal rods to concentrate the field, the attainable sensitivity is such that $\sim 10^{-8}$ W of output power is produced per degree change around the null position. This is sufficient change to give a significant deflection on a sensitive meter.

(2) *Magnetometer and susceptibility bridge.* Using a d.c. amplifier, the sensitivity of the flux meter can be greatly increased. A transistorized magnetometer described by Hilsum and Thompson[63] can detect 10^{-4} oersted. For observation times of 1 sec, Ross and Saker[62] estimate the limit of field measurement to be as low as 10^{-6} oersted if mumetal concentrators are used.

A particularly novel application of the high sensitivity of indium antimonide to magnetic fields is the use of an indium antimonide detector in a susceptibility bridge.[64]

(3) *Clip-on direct current meter.* Here the Hall effect element is inserted in a small air gap in a magnetic circuit of high permeability material, which can be constructed in the form of the conventional clip-on a.c. meter. Measurement of direct currents down to ~ 1 mA is possible in this way.

(4) *Multiplier as computer element.* The Hall effect multiplier may be used in analogue computers when the parameters can be represented as current and magnetic field respectively. As shown by Saker et al.[59] the Hall voltage dependence on magnetic field is linear provided the exciting *current* is maintained constant, although the exciting voltage varies.

(5) *Modulator.* By applying audio frequencies to the magnetic field while high frequency current is being passed through the specimen, a modulated carrier will be generated at the Hall electrodes.[65] Using indium antimonide, a high percentage modulation depth should be obtainable. In the same way the device can be used as a d.c. to a.c. convertor by driving the magnetic field at any convenient frequency.

(6) *A.C. power meters.* If the current to be measured is passed through the specimen and the voltage used to provide the magnetic field, then the Hall voltage gives a measure of the mean power irrespective of waveform or phase angles. Of particular interest is the fact that such a device will operate equally well at high frequencies. For example, the specimen may be mounted directly in a coaxial line or waveguide where it is automatically subject to the appropriate magnetic field. Barlow[66] has shown that good performance is obtainable at 300 Mc/s with germanium elements.

(7) *Electric motor torque measurement.* If a Hall element is inserted in the armature of an electric motor so that it experiences the same current and magnetic field as a typical armature conductor, then the Hall voltage produced will be directly proportional to the force on the conductor. Thus, by observing the Hall voltage on an oscilloscope, the instantaneous torque of the motor can be measured during actual operation.[67]

(8) *Hall effect amplifier.* As a magnetic field can, in theory, be established by the use of negligible power, it is possible to make an amplifier by feeding the input power into the magnetic field and taking power out of the Hall electrodes. With use of indium antimonide such an amplifier is a practical possibility, and Ross and Thompson[68] have achieved a power gain of 5 for both d.c. and low frequency a.c. conditions.

(9) *Hall effect displacement gauge.* Ross and Saker[62] estimate that using an attainable field gradient of 10 gauss micron^{-1}, the sensitivity of a displacement gauge should be adequate to measure movements of only 1 Å.

(10) *Hall gyrator.* By shunting a Hall effect element with suitable resistances a non-reciprocal four terminal network is obtained. With indium antimonide the

forward losses are particularly low. Ross and Saker[62] obtained < 8 dB. With care, a reverse loss of 100 dB is attainable.

Although few of these devices have yet reached the stage of commercial production, the wide variety of possible devices which may be made from indium antimonide indicates that this material will be of considerable importance in the future.

ACKNOWLEDGEMENT

Acknowledgement is due to the Chief Scientist, Ministry of Supply, for permission to publish this paper.

REFERENCES

1. D. F. Gibbons. *Phys. Rev.* **112**, 136 (1958)
2. R. F. Potter. *Phys. Rev.* **103**, 47, 861 (1956)
3. R. F. Potter. *J. Phys. Chem. Solids* **3**, 223 (1957)
4. K. F. Hulme and J. B. Mullin. *J. Electron. Control* **3**, 160 (1957). See also A. J. Strauss. *J. appl. Phys.* **30**, 559 (1959)
5. W. D. Lawson and S. Neilsen. *Preparation of Single Crystals* (Butterworth, London, 1958)
6. J. T. Edmond. *Proc. phys. Soc. Lond.* **73**, 622 (1959)
7. B. Seraphin. *Z. Naturf.* **9a**, 450 (1954)
8. F. Herman. *J. Electron. Control* **1**, 115 (1955)
9. A. J. Tuzzolino. *Phys. Rev.* **108**, 1411 (1957); *Phys. Rev.* **109**, 1980 (1958)
10. W. P. Dumke. *Phys. Rev.* **108**, 1419 (1957)
11. H. P. R. Frederickse and W. R. Hosler. *Phys. Rev.* **108**, 1136, 1146 (1957)
12. E. O. Kane. *J. Phys. Chem. Solids* **1**, 249 (1957)
13. M. J. Stephen and A. B. Lidiard. *J. Phys. Chem. Solids* **9**, 43 (1959)
14. T. S. Moss, S. D. Smith, and T. D. H. Hawkins. *Proc. phys. Soc. Lond.* **B70**, 776 (1957)
15. T. S. Moss. *Optical Properties of Semiconductors* (Butterworth, London; Academic, New York, 1959)
16. V. Roberts and J. E. Quarrington. *J. Electron. Control* **1**, 152 (1955)
17. E. Blount, J. Callaway, M. Cohen, W. P. Dumke, and J. Phillips. *Phys. Rev.* **101**, 563 (1956)
18. S. Zwerdling, B. Lax, and L. M. Roth. *Phys. Rev.* **108**, 1402 (1957)
19. E. Burstein, G. S. Picus, and H. A. Gebbie. *Phys. Rev.* **103**, 825 (1956)
20. P. T. Landsberg. *Proc. phys. Soc. Lond.* **71**, 69 (1958)
21. R. W. Keyes. *Phys. Rev.* **99**, 490 (1955)
22. D. Long. *Phys. Rev.* **99**, 388 (1955)
23. D. Long and P. H. Miller. *Phys. Rev.* **98**, 1192 (1955)
24. R. F. Potter. *Phys. Rev.* **108**, 652 (1957)
25. L. H. de Vaux and F. A. Pizzarello. *Phys. Rev.* **102**, 85 (1956)
26. M. Tanenbaum and H. B. Briggs. *Phys. Rev.* **91**, 1561 (1953)
27. H. J. Hrostowski, G. H. Wheatley, and W. F. Flood. *Phys. Rev.* **95**, 1683 (1954)
28. S. W. Kurnick and R. N. Zitter. *J. appl. Phys.* **27**, 278 (1956)
29. C. Hilsum and R. Barrie. *Proc. phys. Soc. Lond.* **71**, 676 (1958)
30. W. G. Spitzer and H. Y. Fan. *Phys. Rev.* **106**, 882 (1957)
31. H. Yoshinaga and R. A. Oetjen. *Phys. Rev.* **101**, 526 (1956)
32. H. Yoshinaga. *Phys. Rev.* **100**, 753 (1955)
33. H. Ehrenreich. *J. Phys. Chem. Solids* **2**, 131 (1957)
34. P. T. Landsberg. *Proc. phys. Soc. Lond.* **B70**, 1175 (1957)
35. J. Tauc and A. Abraham. *Czech. J. Phys.* **9**, 95 (1959)
36. T. S. Moss and T. D. H. Hawkins. *Proc. phys. Soc. Lond.* **72**, 270 (1958)
37. E. Burstein and G. S. Picus. *Phys. Rev.* **105**, 1123 (1957)
38. B. Lax. *Rev. mod. Phys.* **30**, 122 (1958)
39. R. J. Keyes, S. Zwerdling, S. Foner, H. H. Kolm, and B. Lax. *Phys. Rev.* **104**, 1804 (1956)
40. G. Dresselhaus, A. F. Kip, C. Kittel, and G. Wagoner. *Phys. Rev.* **98**, 556 (1955)
41. T. S. Moss, S. D. Smith, and K. W. Taylor. *J. Phys. Chem. Solids* **8**, 323 (1959)
42. S. D. Smith, T. S. Moss, and K. W. Taylor. *J. Phys. Chem. Solids* **11**, 131 (1959)
43. G. H. Suits, W. D. Schmitz, and R. W. Terhune. *J. appl. Phys.* **27**, 1385 (1956)

44. D. J. Oliver. *Proc. phys. Soc. Lond.* **B70**, 331 (1957)
45. C. Hilsum. *Proc. phys. Soc. Lond.* **74**, 81 (1959)
46. G. R. Mitchell, A. E. Goldberg, and S. W. Kurnick. *Phys. Rev.* **97**, 239 (1955)
47. D. G. Avery, D. W. Goodwin, and A. E. Rennie. *J. sci. Instrum.* **34**, 394 (1957)
48. M. E. Lasser, P. Cholet, and E. C. Wurst. *J. opt. Soc. Amer.* **48**, 468 (1958)
49. G. K. Wertheim. *Phys. Rev.* **104**, 662 (1956)
50. D. W. Goodwin. *J. sci. Instrum.* **34**, 367 (1957)
51. J. Tauc. *J. Phys. Chem. Solids* **8**, 219 (1959)
52. H. J. Hrostowski, F. J. Morin, T. H. Geballe, and G. H. Wheatley. *Phys. Rev.* **100**, 1672 (1955)
53. H. Ehrenreich. *J. Phys. Chem. Solids* **9**, 129 (1959)
54. R. P. Chasmar and R. Stratton. *Phys. Rev.* **102**, 1686 (1956)
55. H. Weiss. *Z. Naturf.* **11a**, 131, 684 (1956)
56. H. Welker and H. Weiss. *Solid State Physics* **3**, 1 (1956)
57. J. C. Haslett and W. F. Love. *J. Phys. Chem. Solids* **8**, 518 (1959)
58. I. M. Ross and E. W. Saker. *Nature, Lond.* **178**, 1196 (1956)
59. E. W. Saker, F. A. Cunnel, and J. T. Edmond. *Brit. J. appl. Phys.* **6**, 217 (1955)
60. F. Kuhrt. *Siemens-Z.* **28**, 299 (1954)
61. W. Hartel. *Siemens-Z.* **28**, 376 (1954)
62. I. M. Ross and E. W. Saker. *J Electron. Control* **1**, 223 (1955)
63. C. Hilsum and N. A. C. Thompson. To be published
64. C. Hilsum and A. C. Rose-Innes. *Nature, Lond.* **182**, 1082 (1958)
65. W. P. Mason, W. H. Hewitt, and R. F. Wick. *J. appl. Phys.* **24**, 166 (1953)
66. H. E. M. Barlow. *Proc. Instn elect. Engrs* **102B**, 179, 199 (1955)
67. F. Kuhrt and E. Brauenstreuther. *Siemens-Z.* **28**, 370 (1954)
68. I. M. Ross and N. A. C. Thompson. *Nature, Lond.* **175**, 518 (1955)
69. R. L. Batdorf, G. C. Dacey, R. L. Wallace, and D. J. Walsh. *J. appl. Phys.* **31**, 613 (1960)

MAGNETO-OPTICAL PHENOMENA IN SEMICONDUCTORS

B. LAX

and

S. ZWERDLING

M.I.T. Lincoln Laboratory, Lexington, Mass., U.S.A.

MS. received October 1959

The work reported in this paper was performed by Lincoln Laboratory, a centre for research operated by the Massachusetts Institute of Technology with the joint support of the U.S. Army, Navy, and Air Force.

MAGNETO-OPTICAL PHENOMENA IN SEMICONDUCTORS

1. INTRODUCTION

The investigation of semiconductors at microwave frequencies by cyclotron resonance [1,2] revealed a number of interesting aspects of the electronic behaviour of holes and electrons in a magnetic field. The quantum-mechanical interpretation of the results which was made in terms of transitions between quantized magnetic levels suggested a number of other phenomena which could be most readily observed at infra-red frequencies. Several of these have now been observed by absorption of such radiation in specimens of appropriate thickness. We have chosen to designate this class of phenomena therefore by the term *magneto-absorption*, although some of these can also be readily examined by studying the dependence of the dispersive properties of the media upon the frequency of the electromagnetic waves. Although we shall not discuss it in this exposition, the first phenomenon to be studied at infra-red frequencies was *cyclotron resonance* which is associated with transitions of carriers between magnetic levels in the same band. The next to be studied were the *oscillatory magneto-absorption* of the *direct transition* in semiconductors and the *magneto-absorption* of the *indirect transition* in germanium. These involve transitions of electrons from the magnetic levels of the valence band to those of the conduction band. Another closely related effect is the *magneto-absorption* of the *forbidden transitions* such as those that occur between the split valence bands in germanium, gallium arsenide, etc., which have been analysed theoretically and are yet to be observed experimentally. *The Zeeman effect of impurity levels* in the presence of a magnetic field presents an interesting study of the transitions of electrons or holes from the ground state to the excited states. In addition, transitions from the ground state to the Landau levels in the conduction and valence bands, respectively, give rise to an *impurity oscillatory magneto-absorption* similar to that of the direct transition between bands. The *Zeeman effect of excitons*, involving transitions of electrons from the valence band to bound magnetic levels just below the conduction band, has been observed in germanium for both direct and indirect excitons. *The Faraday effect* which depends on the differential dispersion of two counter-rotating electromagnetic waves has been successfully applied to investigate the shape of the conduction band in energy–momentum space. *The Faraday effect of bound valence electrons* provides information about the oscillator frequency associated with the binding of these electrons. *The Faraday effect of impurity levels* as well as that of other phenomena has been considered theoretically. *Magneto-plasma phenomena*, depending on the dispersive properties of free carriers in semiconductors, have also been considered theoretically. Among these are the *magneto-plasma reflection* and the *rotatory dispersive effect* (Kerr magneto-optic effect) which is closely related to the Faraday effect. Finally, *de Haas–van Alphen-type oscillations* of free carrier absorption in degenerate semiconductors at low temperatures, analogous to those observed in

metals, should be observable by infra-red spectrometric techniques in a magnetic field.

2. MAGNETO-ABSORPTION

2.1. Direct Transition

In the presence of a magnetic field, two simple parabolic bands whose energy–momentum relations at zero field are given by

$$\epsilon_1 = -\hbar^2 k^2/2m_1 \quad \text{and} \quad \epsilon_2 = \epsilon_g + \hbar^2 k^2/2m_2$$

where m_1 is the effective mass of a hole, m_2 that of the electron, and ϵ_g the energy gap, give rise to one-dimensional energy bands as follows:

$$\left.\begin{aligned} \epsilon_{1n} &= -(n_1+1/2)\hbar\omega_{c_1} - \hbar^2 k_H^2/2m_1 \\ \epsilon_{2n} &= (n_2+1/2)\hbar\omega_{c_2} + \hbar^2 k_H^2/2m_2 + \epsilon_g \end{aligned}\right\} \quad \ldots (1)$$

where $\omega_{c_{1,2}} = eH/m_{1,2}c$ is the cyclotron frequency and H the magnetic field intensity. We have for the present neglected any effect due to the spin of the electrons and holes. The expressions of equation (1) are obtained by solving an effective mass Schroedinger equation in which the momentum operator includes the magnetic vector potential. These are harmonic oscillator-like levels in the transverse co-ordinates whose separation in energy depends upon the cyclotron frequency and the effective mass of each band, respectively. In order to calculate the absorption coefficient associated with the transitions of electrons from the valence band to the conduction band, it is necessary to evaluate a dipole matrix involving wave functions which are products of Bloch functions at the band edges, evaluated over the unit cell, and Landau functions which are solutions of the effective mass wave equation and are evaluated over the entire crystal. In evaluating the matrix over the Bloch functions, it is assumed that the valence and conduction bands involved are of different symmetry and therefore permit a first-order transition. This type of calculation for the absorption coefficient at zero field, $\alpha(0)$, has been given by Hall, Bardeen, and Blatt[3] as

$$\alpha(0) = A(\hbar\omega - \epsilon_g)^{1/2} \quad \ldots (2)$$

where

$$A = \frac{2e^2}{\eta m_0^2 \omega c}\left(\frac{2\mu}{\hbar^2}\right)^{3/2} |\mathbf{p}_{12}.\mathbf{e}|^2$$

where η is the index of refraction, the reduced effective mass $\mu = m_1 m_2/(m_1+m_2)$, $\mathbf{p}_{12}$ is the momentum matrix, and $\mathbf{e}$ the unit polarization vector of the electric field. In an analogous manner, the absorption coefficient can be evaluated in the presence of a magnetic field to give[4]

$$\left.\begin{aligned} \alpha(H) &= \frac{A\hbar\omega_c}{2}\sum_n (\hbar\omega - \epsilon_n)^{-1/2} \\ \epsilon_n &= \epsilon_g + (n+\tfrac{1}{2})\hbar\omega_c + (g_2 M_2 - g_1 M_1)\beta H \end{aligned}\right\} \quad \ldots (3)$$

where $\omega_c = eH/\mu c$, g_1 and g_2 the spectroscopic splitting factors, and the magnetic quantum numbers associated with total angular momenta of hole and electron spins are given by M_1 and M_2, respectively, and $\beta = e\hbar/2m_0 c$ is the gyromagnetic

ratio. The selection rules for these transitions, which now involve an envelope function corresponding to the solution of the effective mass equation with a magnetic field, require that for the allowed direct transition, $\Delta n = 0$ for two simple parabolic bands. In addition, for the electric field parallel to the magnetic field, $\mathbf{E} \| \mathbf{H}$, $(M_1 - M_2) = \Delta M = 0$, and for the electric vector perpendicular to the magnetic field, $\mathbf{E} \perp \mathbf{H}$, $\Delta M = \pm 1$. These additional selection rules apply when there is spin–orbit coupling which permits magnetic transitions due to an electric

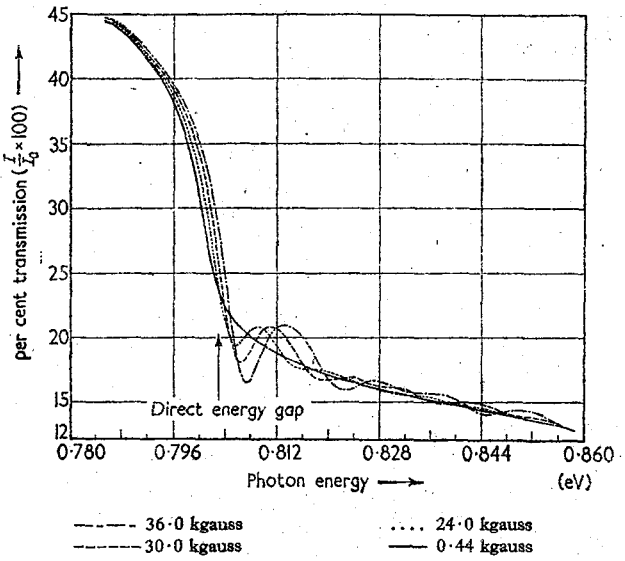


Figure 1. *The oscillatory magneto-absorption spectrum for the direct transition in germanium at room temperature*[7]

field. If we include a Lorentzian term τ to account for losses, the expression under the summation sign of equation (3), with $\epsilon_n = \hbar\omega_n$, takes the form

$$\hbar^{-1/2}(\omega - \omega_n)^{-1/2} \rightarrow \hbar^{-1/2}\left[\frac{x + \sqrt{(x^2+1)}}{2\tau^{-1}(x^2+1)}\right]^{1/2} \qquad \ldots (4)$$

where $x = (\omega - \omega_n)\tau$. An equivalent but more complicated expression has been obtained by Burstein and co-workers.[4] These theoretical results indicate an absorption which has an oscillatory character as a function of photon energy for a given magnetic field. Such a phenomenon was observed experimentally independently by the group at Lincoln Laboratory in germanium, and by the N.R.L. group in indium antimonide.[5] An interesting feature of the magneto-absorption, as shown in Figure 1, is that the absorption edge is shifted to higher energies and changes shape in a magnetic field. The initial experiments on the effect of a magnetic field on the spectral absorption were made by observing the shift of the absorption edge.[6] The change of the energy gap was interpreted as the shift in energy along the isotransmission lines between the parallel portions of the curves shown in Figure 2. This apparent shift of the gap did not vary linearly, but

quadratically with the magnetic field. Such a quadratic shift is predicted from a series expansion of equation (4) in H. This establishes that, to the first order, there is no shift of the absorption edge. Consequently, the proper interpretation of the change in the energy gap is best made in terms of the motion of the transmission minima as a function of magnetic field. These minima occur at photon energies above the absorption edge, and correspond to transitions between the quantized magnetic levels of the two bands.

The oscillations are most effectively presented as a ratio of the transmitted intensity in the presence of the magnetic field to that at zero field versus photon energy, as shown in Figure 2. These curves, taken at room temperature using prism dispersion, show a characteristic anisotropy when the magnetic field $\mathbf{B}$ is along different crystallographic directions, which is due to the properties of the

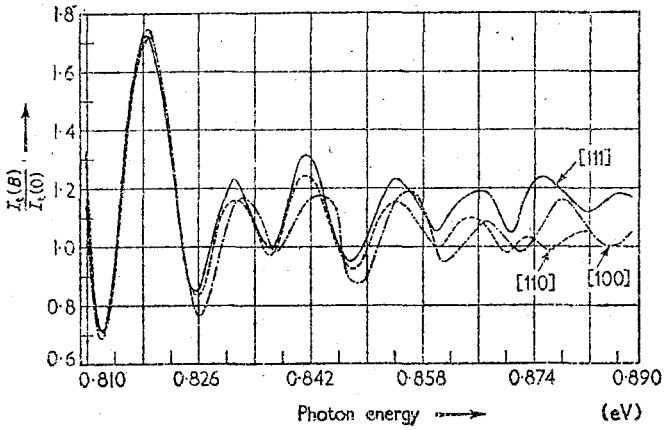


Figure 2. *Anisotropy of the oscillatory magneto-absorption of the direct transition in germanium at room temperature.* $\mathbf{B} = 35 \cdot 7$ *kgauss, along the directions indicated*

valence band in germanium. Such spectra for the direct transition magneto-absorption in germanium were studied with prism dispersion at room temperature, 77° K, and 4° K.[7] The characteristics of the spectra remained essentially the same with decreasing temperature, except that the positions of the transmission minima occurred at increasing energies. In order to obtain quantitative results from these data, the positions of the minima were plotted as a function of the magnetic field as shown in Figure 3. The straight lines drawn through the experimental points converged at zero field to a value which represents an accurate determination of the direct energy gap† in germanium and give a value at room temperature $\epsilon_g = 0 \cdot 803 \pm 0 \cdot 001$ eV. The linear behaviour with magnetic field is consistent with the theoretical results of equation (3). Another quantitative result obtained from this plot is the determination of the effective mass of the electron. This has been done either by using the slope of the first line, or taking the difference between the lines 5 and 7 at a fixed magnetic field. Either quantity can be interpreted in terms of the cyclotron frequency corresponding to a reduced effective mass μ of

† For more recent work see section added in proof, page 268.

the heavy hole and the electron at $k=0$. Since the effective mass of the heavy hole m_1 is known from cyclotron resonance, it was a simple matter to deduce the value of the electron mass to give a result $m_2=\mu m_1/(m_1-\mu)=(0\cdot036\pm0\cdot002)m_0$. This value which represents the first experimental determination of this electron mass[7] is in very good agreement with the theoretical estimate $0\cdot034\,m_0$ by Dresselhaus, Kip, and Kittel[1] and $0\cdot037\,m_0$ by Dumke.[8] It is noteworthy that by this magneto-absorption technique, an experimental measurement of the effective mass of a higher conduction band minimum was made, an achievement which could not be accomplished by cyclotron resonance.

The experiments which were done by the N.R.L. group[4] at room temperature

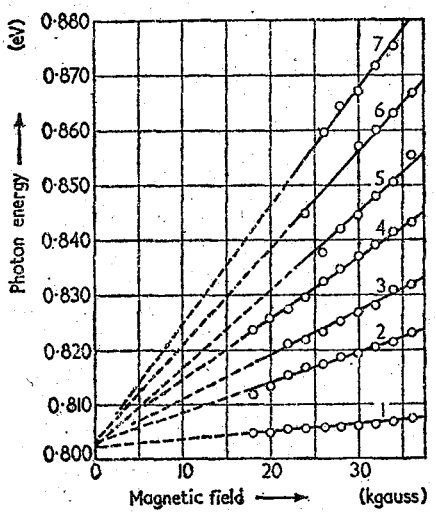


Figure 3. *Energy values of transmission minima versus magnetic field for successive transitions of electrons between Landau levels of valence and conduction bands in germanium. Convergence of lines yields energy gap value $\epsilon_g=0\cdot803\pm0\cdot001$ eV at room temperature*[5]

differed from those described above in that a longitudinal magnetic field, parallel to the direction of propagation was used, in conjunction with circularly polarized radiation. Two of the spectra obtained for $\mathbf{H}\|[100]$ at 46·6 kgauss are shown in Figure 4. The difference between these spectra is associated with unresolved fine structure, which has been observed under high resolution at low temperature by the Lincoln group.

The analysis thus described for two simple parabolic bands has been extended to take into account the complex form of the energy bands in germanium. It was first necessary to solve a 4×4 matrix wave equation for the magnetic levels of the valence band. This was done by Luttinger and Kohn and later modified by Luttinger.[9] Using the results of this theory and the parameters obtained from cyclotron resonance, the energy levels have been calculated,[10] yielding four sets

of levels which are indicated in Figure 5. The diagram shows two sets of levels which are closely spaced, characteristic of heavy holes, and another two, which are spaced farther apart, corresponding to light holes. For quantum numbers zero and

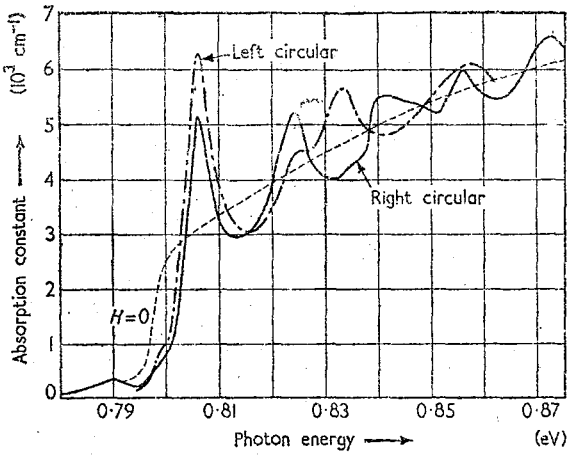


Figure 4. *Comparison of circularly polarized magneto-absorption spectra of the direct transition in germanium (Abstracted from published curves of Burstein, Picus, Wallis, and Blatt*[12])

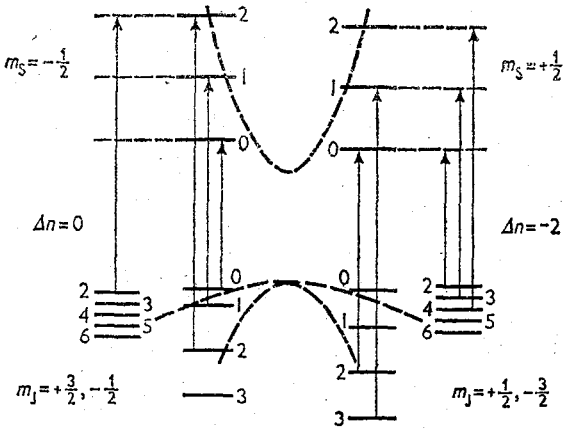


Figure 5. *Schematic diagram showing the magnetic levels for the valence and $k=0$ conduction bands in germanium.*[10] *The transitions shown are those allowed for* $\mathbf{E}\|\mathbf{H}$

one, there is no real distinction between light and heavy holes. In a similar manner, two sets of levels for electrons in the conduction band have been plotted to take into account the effect of spin. The diagram shows the appropriate transitions in which the selection rule for the orbital magnetic quantum numbers are $\Delta n = 0$ and $\Delta n = -2$. This rule is a consequence of the first-order solution for the magnetic

levels of the four-fold degenerate valence bands in which each level designated is actually a linear combination of levels n and $n-2$. In addition to the orbital selection rule, the experiments which have been carried out by Zwerdling, Lax, Roth, and Button[11] using linearly polarized radiation, require that for $\mathbf{E}\|\mathbf{H}$, $\Delta M = 0$ and $\mathbf{E} \perp \mathbf{H}$, $\Delta M = \pm 1$, and for those by the N.R.L. group[12] that $\Delta M = +1$ for right-handed and $\Delta M = -1$ for left-handed circularly polarized radiation, respectively. The grouping and the values of the total angular momentum numbers for the valence bands are ascribed to the spin–orbit splitting of these bands in which the four-fold degenerate $p_{3/2}$ states were split to form a set of bands 0·28 eV above the two-fold degenerate $p_{1/2}$ states. The results of this theoretical spectrum have been compared with the more recent experiments of ZLRB,[11] and are shown in Figure 6. These data were obtained at liquid helium temperatures with high resolution provided by diffraction grating dispersion. The solid lines correspond

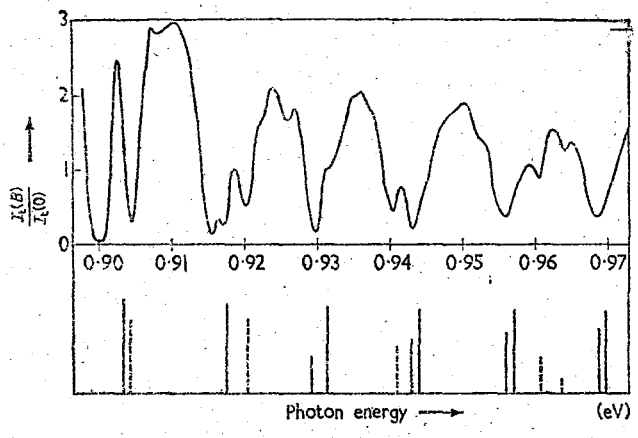


Figure 6. *Comparison of theoretical spectrum with experimental data for germanium at* 4·2°*K with high resolution grating spectrometer and a field of* 38·9 *kgauss* ($\mathbf{E}\|\mathbf{H}$)

to heavy hole transitions and the dashed lines to light hole transitions, except for the two prominent lines which were shown to be due to exciton formation.[13] From these experiments, a determination of the electron effective mass associated with the conduction band at $k = 0$, gave $m_s = (0{\cdot}037 \pm 0{\cdot}001)m_0$. When corrected for higher order terms in the momentum, the energy levels are given by the expression

$$\epsilon_n = (27{\cdot}1 \pm 0{\cdot}6)(n+\tfrac{1}{2})\hbar\omega_{c_0} - (0{\cdot}33 \pm 0{\cdot}1)(n+\tfrac{1}{2})^2(\hbar\omega_{c_0})^2 \mp 1{\cdot}2\beta H \quad \ldots (5)$$

where ω_{c_0} is the cyclotron angular frequency of the free electron mass m_0. The second term, which is quadratic in the quantum number n, is due to the decreasing curvature of the conduction band with increased energy. In addition, from a comparison of the data for the two polarized spectra with $\mathbf{E}\|\mathbf{H}$ and $\mathbf{E} \perp \mathbf{H}$, it was possible to measure the splitting of the magnetic levels in the conduction band due to spin–orbit effects. This gave an anomalous g value, $g = -2{\cdot}5$ as obtained from the last term in equation (5) where the ($\mp$) signs refer to spin up and spin down respectively, the separation being a measure of the spin splitting. This effect,

which is more striking in indium antimonide, is described in somewhat more detail subsequently.

The oscillatory magneto-absorption spectra of the direct transition in indium antimonide,[7] indium arsenide,[7] and gallium antimonide[14] have also been observed by the Lincoln group. The experimental results for indium antimonide are shown in Figure 7, in which the first four transmission minima observed have been plotted as a function of magnetic field, and when extrapolated to zero field, give a value of the energy gap $\epsilon_g = 0{\cdot}180 \pm 0{\cdot}001$ eV at room temperature. Similar data corresponding to the lowest three lines were also obtained by the N.R.L. group.[6] In order to interpret the data, a dashed line shown in Figure 6 was drawn corresponding to the predicted behaviour of the first minimum when spin–orbit effects

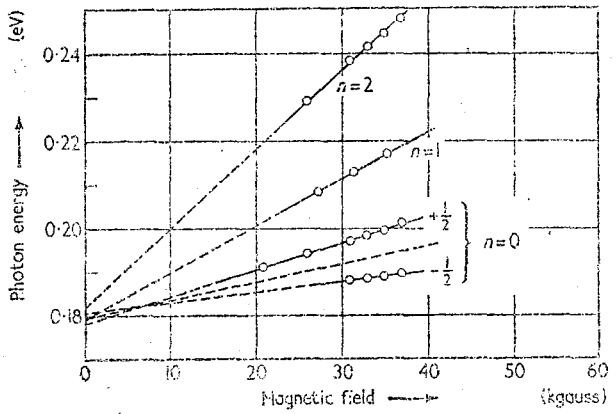


Figure 7. *Energy values of transmission minima versus magnetic field for electron transitions between Landau levels of valence and conduction bands in indium antimonide.*[6] *The first level ($n = 0$) is split by spin–orbit interaction yielding an anomalous g value for the conduction electron of $g \simeq 54$. The energy gap at zero field and room temperature obtained from the convergence of the lines is $\epsilon_g = 0{\cdot}180 \pm 0{\cdot}002$ eV (Zwerdling, Lax, and Roth*[7])

are ignored, and the electron effective mass taken from cyclotron resonance ($m_2 = 0{\cdot}013\ m_0$) and the hole effective mass ($m_1 \simeq 0{\cdot}18\ m_0$) are used. This theoretical line bisects the first two observed lines, and therefore it is assumed that these corresponded to the $\Delta n = 0$ transition which was split due to spin–orbit effects. From the experimental results, the calculated effective g factor of the electron corresponding to the energy splitting is equal to 54. This anomalous g value has been accounted for by Roth[10] who developed a theory which takes into account the effect of the split valence bands. The expression for the g value is then given by

$$g = 2\left[1 + \left(1 - \frac{m_0}{m_2}\right)\frac{\Delta}{3\epsilon_g + 2\Delta}\right] \qquad \ldots (6)$$

where m_2 is the electron effective mass at the conduction band minimum, ϵ_g is the value of the gap obtained from the magneto-absorption measurements, and the

estimated spin–orbit splitting from atomic data is $\Delta \simeq 0 \cdot 9$ eV. When these values are substituted into the above equation, the theoretical value of $g \simeq -56$ is obtained in good agreement with experiment. Similar data with lower resolution were obtained at room temperature with a prism spectrometer for an 18 micron thick polycrystalline sample of indium arsenide.[7] Two transmission minima were observed at 38·9 kgauss, but this time no fine structure was resolved corresponding to a spin–orbit effect of the lowest Landau state as in indium antimonide. Nevertheless, the preliminary data gave an energy gap $\epsilon_g = 0 \cdot 360 \pm 0 \cdot 002$ eV and an effective mass value for the electron $m_2 \simeq 0 \cdot 03\ m_0$. Another intermetallic semiconductor which has been studied, but this time with the high resolution grating

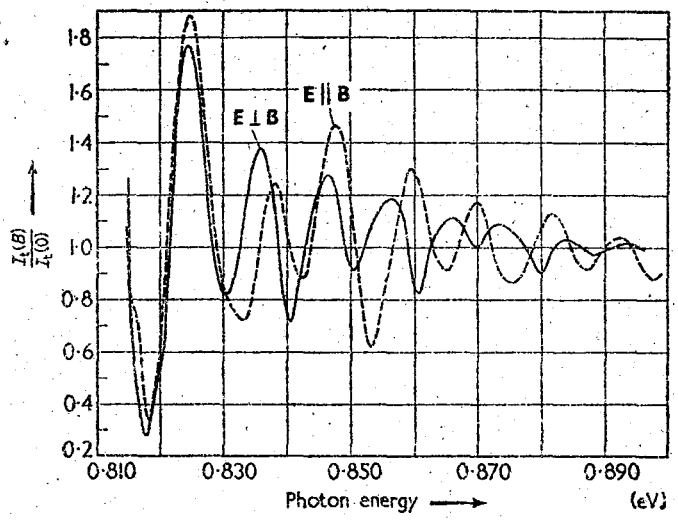


Figure 8. *The oscillatory magneto-absorption spectrum of the direct transition for parallel and perpendicular polarization in gallium antimonide at 1·5° K. The plot shows the ratio of the transmission of infra-red radiation at 38·9 kgauss to that at zero field as a function of photon energy. The magnetic field is along a* [111] *crystal axis*[14]

magneto-spectrometer and at liquid helium temperature was gallium antimonide.[14] The magneto-absorption data obtained for a single-crystal sample 4 microns thick closely resembled the oscillatory spectrum for germanium at room temperature. The occurrence of this type of spectrum, together with the large absorption coefficient, establishes that the energy gap corresponds to a direct transition. Although the fine structure was not resolved, probably due to the high impurity concentration ($\sim 10^{17}$ cm^{-3}), evidence for its existence may be inferred from the indication of a possible exciton line for the polarized spectrum with $\mathbf{E} \| \mathbf{H}$ as shown in Figure 8. The difference between this spectrum and the one for $\mathbf{E} \perp \mathbf{H}$ also indicates the existence of fine structure. The analysis of the data gave an energy gap for gallium antimonide at 4·2° K, $\epsilon_g = 0 \cdot 813 \pm 0 \cdot 001$ eV and an electron effective mass, $m_2 = (0 \cdot 047 \pm 0 \cdot 003)\ m_0$.

2.2. Indirect Transition

The absorption edge associated with the normal energy gap in germanium and silicon involves the transitions of electrons from the top of the valence band at $k = 0$ to the bottom of the conduction band, which is displaced from the centre of the Brillouin zone. For germanium, the conduction band minima are at the edge of the zone along the [111] directions, and in silicon approximately 5/6 of the distance in momentum space along the [100] direction. It has been shown theoretically[3] and experimentally[15] that for such a transition to take place, it is necessary that the electron emit or absorb a phonon to conserve momentum. In order to calculate the absorption coefficient for this process, it is necessary to make a second-order perturbation calculation in which matrices are evaluated for transitions from the initial states to intermediate states and phonon matrices for transitions from the intermediate states to the final states in the conduction band. The calculations at zero field made by Hall, Blatt, and Bardeen[3] were modified by Macfarlane and Roberts[15] to take into account the phonon population as a function of temperature, yielding an expression for the absorption coefficient

$$\alpha(0)_{\pm} = C_{\pm}[\hbar\omega - \epsilon_g \mp k\theta]^2 \qquad \ldots (7)$$

with

$$C_{\pm} = \pm\frac{D}{\omega}[\exp \pm \theta/T - 1]$$

where T is the absolute temperature and D involves fundamental constants: density of states associated with the valence and conduction bands, a factor for multiple minima, and a product of the phonon matrices and the direct transition matrices. The plus or minus sign refers respectively to the absorption or emission of a phonon whose energy is $k\theta$. When a similar calculation is made including the presence of a magnetic field, the expression is modified and takes the form

$$\alpha(H) = 2C_{\pm}(\hbar^2\omega_{c_1}\omega_{c_2})\sum_{n_1 n_2} S(\hbar\omega - \epsilon_{n_1 n_2}) \qquad \ldots (8)$$

where

$$\epsilon_{n_1 n_2} = \epsilon_g + (n_1 + \tfrac{1}{2})\hbar\omega_{c_1} + (n_2 + \tfrac{1}{2})\hbar\omega_{c_2} \pm k\theta + (M_2\bar{g}_2 - M_1 g_1)\beta H$$

and $S(\hbar\omega - \epsilon_{n_1 n_2})$ is a step function. The above result applies in general since ω_{c_1} and ω_{c_2} in equation (8) correspond to the appropriate cyclotron angular frequencies for the two bands, thus taking into account the anisotropy of the phenomenon. Due to the phonon transitions, no selection rules are imposed on the Landau orbital quantum numbers n_1 and n_2. However, the selection rules for the spin angular momenta are still retained with $\Delta M = 0$ for $\mathbf{E} \| \mathbf{H}$ and $\Delta M = \pm 1$ for $\mathbf{E} \perp \mathbf{H}$. The symbol $\bar{g}_2$ denotes the value of the appropriate tensor component of the g factor to take into account the anisotropy of the electron spin which has been postulated by Roth on theoretical grounds. The magneto-absorption spectrum predicted by equation (8) should therefore consist of a series of steps or absorption edges with increasing photon energy above that of the indirect gap. In an actual crystal, relaxation effects occur and the absorption edges have a finite slope, since the function $S(x) \to \arctan x$ where $x = (\omega - \omega_{n_1 n_2})\tau$ and $\epsilon_{n_1 n_2} = \hbar\omega_{n_1 n_2}$. The energy values of the transitions between quantized levels correspond to the centres or inflection points of the steps in the transmission spectrum.

Such a spectrum has been observed in germanium by the Lincoln group[11] as shown in Figure 9 for the transmission through a 6 mm sample at 1·5° K in magnetic fields up to 38·9 kgauss. The plot is a portion of the zero field transmission spectrum of the sample showing the absorption edge of the indirect exciton line

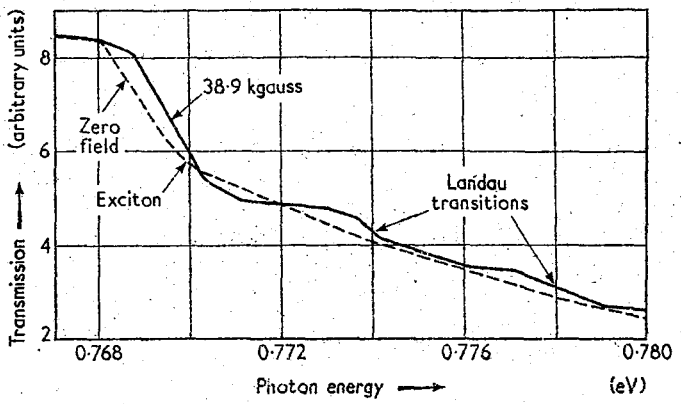


Figure 9. *The magneto-absorption spectrum of the indirect transition in germanium at 1·5°K showing the exciton absorption edge at zero field and the development of the 'stair case' absorption edges for the Landau transitions at higher energies at 38·9 kgauss*[11]

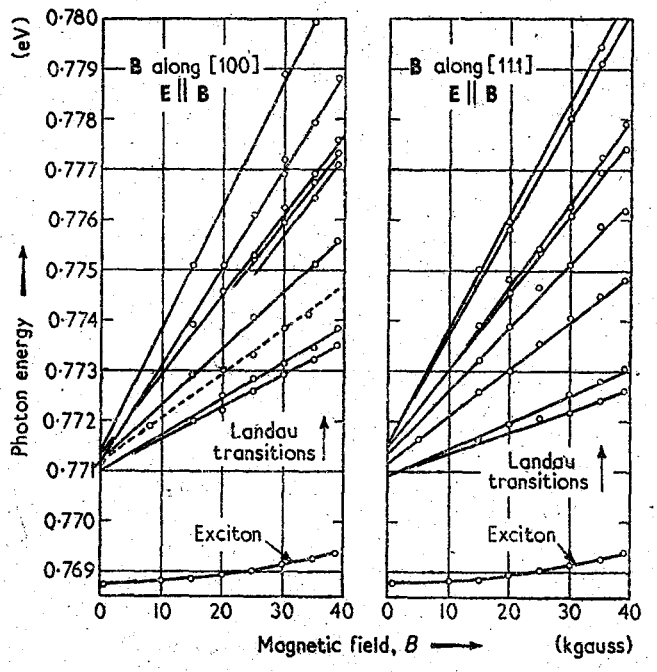


Figure 10. *Position of indirect transitions in germanium at 1·5°K as a function of magnetic field showing the quadratic Zeeman effect of the exciton and linear convergence of the Landau levels*[11]

as reported by Macfarlane, McLean, Quarrington, and Roberts (MMQR).[16] With the magnetic field, not only does the exciton edge shift to higher energy, but a series of weaker absorption edges corresponding to the interband Landau transitions are developed. The individual lines have been analysed in detail and the energy values have been plotted as a function of magnetic field as shown in Figure 10. The results for two orientations of the magnetic field with $\mathbf{E}\|\mathbf{B}$ are shown. The positions of the various edge centres fall on straight lines as a function of magnetic field in accordance with the theory, and at zero field extrapolate to a value of $0{\cdot}7713 \pm 0{\cdot}0004$ eV. This corresponds to the indirect energy gap at $1{\cdot}5°$ K plus the energy of the emitted phonon. The latter has been evaluated by MMQR from their studies of two observed indirect exciton lines. By taking the difference between one formed by emission of a phonon and a second formed by absorption of a phonon, they obtain a value for the longitudinal acoustical phonon which is $0{\cdot}0276 \pm 0{\cdot}0005$ eV. The value of this phonon energy has also been determined by Haynes[17] from recombination emission studies and by Brockhouse and Iyengar[18] from neutron diffraction. Consequently, the energy gap for the indirect transition becomes $\epsilon_g = 0{\cdot}744 \pm 0{\cdot}001$ eV including the cumulative error for the phonon energy determination and the magneto-absorption measurements.

2.3. The Forbidden Transition

A third type of magneto-absorption spectrum which may be observed is that of the direct transition which is not allowed at $k = 0$, since the symmetry of the two bands is the same. Such transitions can occur between the valence bands in germanium, gallium arsenide, and other similar semiconductors in which the valence band has been split due to spin–orbit coupling. They are also possible in a direct transition between the valence and conduction bands of some semiconductors, and it is believed that the observations of Gross[19] and Nikitine[20] in cuprous oxide are of this type. In the absence of a magnetic field, a theory for these forbidden transitions has been developed by Kahn,[21] in which he showed that the absorption coefficient has a spectral dependence of the form

$$\alpha(0) = B(\hbar\omega - \epsilon_g)^{3/2}$$

$$B = \frac{2e^2\hbar^2}{3\eta m^2 c\omega}|\mathbf{M}_{12}|^2\left(\frac{2\mu}{\hbar^2}\right)^{5/2}$$

$$\mathbf{M}_{12} = \frac{1}{m}\sum_j \left[\frac{\mathbf{p}_{1j}(\mathbf{e}\cdot\mathbf{p}_{j2})}{\epsilon_1 - \epsilon_j} - \frac{(\mathbf{e}\cdot\mathbf{p}_{1j})\mathbf{p}_{j2}}{\epsilon_j - \epsilon_2}\right] \qquad \ldots (9)$$

In the presence of a magnetic field, an analysis similar to that at zero field leads to an expression for the absorption coefficient which has two components, depending on the orientation of the polarized radiation relative to the magnetic field. When the electric vector is parallel to the magnetic field, the expression becomes

$$\alpha_\parallel(H) = \frac{3\hbar\omega_c}{2}B\sum_n [\hbar\omega - \epsilon_n]^{1/2} \qquad \ldots (10)$$

where

$$\epsilon_n = \epsilon_g + (n+\tfrac{1}{2})\hbar\omega_c + (M_2 g_2 - M_1 g_1)\beta H$$

The selection rules for this transition are $\Delta n = 0$ and also $\Delta M = 0$. The resultant absorption curve will have a shape which appears as a superposition of a series of absorption edges. When the electric vector is perpendicular to the magnetic field, the absorption coefficient has the form

$$\alpha_{\perp}(H) = \tfrac{3}{2}\left(\frac{\hbar\omega_c}{2}\right)^2 B \sum_n (n+1)[(\hbar\omega - \epsilon_{n_1})^{-1/2} + (\hbar\omega - \epsilon_{n_2})^{-1/2}] \quad \ldots (11)$$

where

$$\epsilon_{n_1,_2} = \epsilon_g + (n+\tfrac{1}{2})\hbar\omega_c + (M_2 g_2 - M_1 g_1)\beta H + \hbar\omega_{c_1,_2}$$

The selection rules for this transition are $\Delta n = \pm 1$ and also $\Delta M = \pm 1$. This result predicts that for the forbidden transition, absorption maxima occur, and if the spin effects are neglected, there are two series of maxima in which the intensities increase with n, and the maxima are separated by $\hbar\omega_c$ corresponding to the magnetic level spacing for the reduced effective mass. The reduced effective mass for

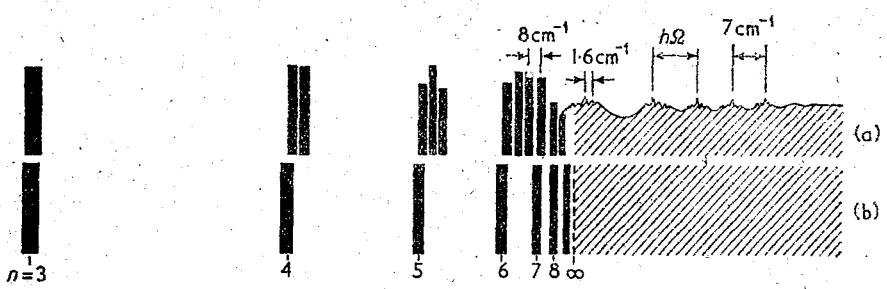


Figure 11. *The exciton and magneto-absorption spectrum in cuprous oxide.*[23] *(a) Discrete lines show the Zeeman splitting of the exciton levels and the oscillatory magneto-absorption in the 'continuum' for a magnetic field of 29 kgauss; (b) Exciton spectrum at zero field. (The symbol $h\Omega$ is equivalent to $\hbar\omega_c$ used in the text)*

two bands whose curvatures are in the opposite sense in energy–momentum space is given by the usual expression

$$\frac{1}{\mu} = \frac{1}{m_1} + \frac{1}{m_2}$$

However, if the band curvatures are in the same sense, as for the split valence bands in a semiconductor, then the reduced effective mass is obtained from the *difference* of the reciprocal effective masses of the two bands. Due to the presence of the extra terms, one associated with the conduction band, the other with the valence band, the maxima for circularly polarized radiation will be shifted in the two series relative to one another by $\hbar(\omega_{c_1} - \omega_{c_2})$. The significance of this is that from the separation of the maxima and the shift of the two series, one can determine the effective masses of both holes and electrons. If spin–orbit coupling is included, then the last term splits each series into two components, thus yielding four lines for each value of the quantum number n.

Although the forbidden transition in germanium has not been observed because the experimental conditions have not yet been favourable, Elliott[22] suggests that

from symmetry considerations the transitions observed by Gross in cuprous oxide must be of this type. Since Gross, Zaharcenja, and Pavinskij[23] have reported an oscillatory magneto-absorption effect above the ionization energy of the excitons as shown in Figure 11, there is reason to believe that it is a forbidden direct transition. The spectrum is rather complicated and not as well defined as those in other semiconductors where the masses are smaller and the intensities of the magneto-absorption maxima are greater. The spectrum shows fine structure with a spacing of $1 \cdot 6$ cm^{-1} which is superimposed on a larger oscillation whose spacing between absorption maxima is not uniform and occurs with spacings of 9, 8, 7, and 7 cm^{-1} at a magnetic field of 29 kgauss. The latter spacing does not correspond to the reduced effective mass of $0 \cdot 25\ m_0$ deduced from the exciton absorption.

3. ZEEMAN EFFECT OF EXCITONS

3.1. Cuprous Oxide

The Zeeman effect of excitons has been observed by Gross and Zaharcenja[24] in cuprous oxide and cadmium sulphide.† A more complex Zeeman effect of the direct and indirect excitons in germanium has been observed by Zwerdling, Lax, Roth, and Button.[11, 25] On the theoretical side, Samojlovic and Korenblit[26] have considered the linear Zeeman effect of excitons. Further work including transition probabilities of excitons in a magnetic field were treated by Elliott and Loudon.[27] The theory of the complex excitons in germanium was given by Roth and Kleiner[25] in terms of spin Hamiltonians. In its simplest form for spherical bands, an effective mass Schroedinger equation for a hole and electron in a magnetic field bound together by a Coulomb force takes the form

$$\left[\frac{1}{2m_1}\left(\mathbf{p}_1-\frac{e\mathbf{A}_1}{c}\right)^2+\frac{1}{2m_2}\left(\mathbf{p}_2+\frac{e\mathbf{A}_2}{c}\right)^2-\frac{e^2}{\kappa(\mathbf{r}_1-\mathbf{r}_2)}\right]\Psi=\epsilon\Psi \qquad \ldots(12)$$

where m_1 and m_2 are the effective masses of the hole and electron, respectively, and the magnetic vector potential $\mathbf{A}=\frac{1}{2}\mathbf{H}\times\mathbf{r}$, κ is the dielectric constant, and ϵ is the energy of the exciton state relative to the conduction band minimum. This equation is usually rewritten in a centre-of-mass co-ordinate in the form

$$\left[\frac{-\hbar^2}{2\mu}\nabla^2+\frac{ie\hbar}{2c}\left(\frac{m_2-m_1}{m_1m_2}\right)\mathbf{H}.\mathbf{r}\times\nabla+\frac{e^2}{8\mu c^2}(\mathbf{H}\times\mathbf{r})^2-\frac{e^2}{\kappa r}\right]F(\mathbf{r})=\epsilon F(\mathbf{r}) \ldots(13)$$

where

$$\Psi(\mathbf{r}_1,\mathbf{r}_2)=F(\mathbf{r})\exp\left(-\frac{ie}{2\hbar c}\mathbf{H}\times\mathbf{R}.\mathbf{r}\right)$$

$$\mathbf{r}=\mathbf{r}_1-\mathbf{r}_2 \quad \text{and} \quad \mathbf{R}=\frac{(m_1\mathbf{r}_1+m_2\mathbf{r}_2)}{(m_1+m_2)}$$

and μ is the reduced effective mass. The above equation is the usual type encountered in the study of the Zeeman effect where the second term is responsible for the linear effect. The third term is the diamagnetic term which is responsible for the quadratic Zeeman effect. The linear term is the important one for low magnetic fields, whereas the quadratic term becomes the dominant one at high magnetic fields. In cuprous oxide, the absence of the linear Zeeman effect led

† See section added in proof, page 268.

Gross[28] to the conclusion that he was dealing with a positronium-like structure in which $m_1 \simeq m_2$. From his spectrum, which formed a hydrogen-like series, and from the measured index of refraction $\eta = 2{\cdot}5$ for cuprous oxide, he deduced a reduced effective mass $\mu = 0{\cdot}25$ which, from the absence of the linear Zeeman effect, gives effective masses of $m_1 = m_2 = 0{\cdot}5\, m_0$. He also observed a large quadratic Zeeman effect, which for the simple case is given by

$$\Delta\epsilon = \frac{n^2}{8}(n^2-1)\,\kappa^2 \frac{\hbar^4\omega_c^2}{\mu e^4} \quad \text{for } M = 0,\ l = 0 \qquad \ldots (14)$$

where κ is the dielectric constant, n is the total quantum number, and ω_c is the cyclotron frequency associated with the reduced effective mass μ. Indeed, Gross

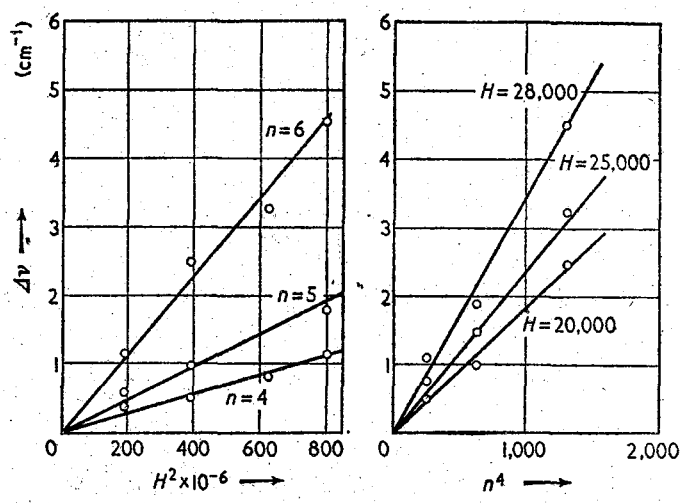


Figure 12. *The quadratic Zeeman spectrum of excitons in cuprous oxide showing the energy shift versus H^2 for the excited states and also as a function of n^4 for different values of the magnetic field demonstrating the validity of equation* (14)[24]

and Zaharcenja[28] showed that for these quadratic Zeeman effects, the shift was proportional to n^4 and H^2 for high quantum numbers, in accordance with the theory. Some of these results are shown in Figure 12. They have also observed the Zeeman effect of the excitons in cadmium sulphide, but for the magnetic fields available, only the linear effect was observed. The exciton structure in cadmium sulphide was also studied by this group in collaboration with Grillot and Bancie-Grillot.[29] This time an emission spectrum was obtained at 4° K in a magnetic field of 28,000 oersteds.

3.2. Direct Transition Exciton in Germanium

In studying the magneto-absorption spectrum of the direct transition in germanium, Zwerdling, Roth, and Lax[13] reported the presence of two intense lines at high fields which could not be accounted for in terms of Landau transitions and

which were identified as exciton lines. The spectrum was then studied as a function of magnetic field, and it was noted that at zero magnetic field, the two prominent lines were retained as shown in Figure 13. Subsequent observations of this exciton were also made by MMQR.[30] Further studies of these exciton lines were made by the Lincoln group at 1·5° K, 4·2° K, and 77° K. The essential features of the exciton structure and its magnetic behaviour varied little for these different temperatures with the exception that the absolute energy of exciton formation did shift as a function of temperature due to the corresponding change in the direct energy gap. Several interesting features are to be noted. The lowest exciton line had a width at half intensity at zero field of approximately 10^{-3} eV, corresponding

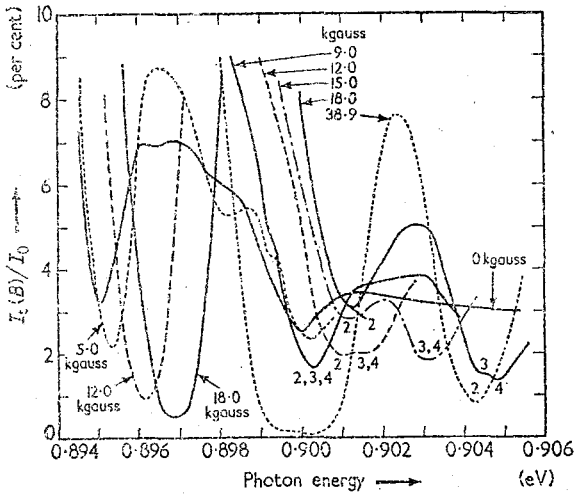


Figure 13. *The Zeeman effect of the direct exciton in germanium at 1·5° K for different values of magnetic field. First exciton line exhibits the onset of two-fold splitting at 38·9 kgauss*[11]

to a lifetime $\tau \simeq 10^{-12}$ sec, and increased in width with magnetic field approximately by a factor of 3, as shown in Figure 13. This is not surprising in view of the fact that in considering the theory of this exciton, it is necessary to write an 8×8 matrix Hamiltonian associated with the four-fold degeneracy of the valence band and the two-fold degeneracy of the conduction band. Due to selection rules and the orientation of the linearly polarized electric vector relative to the magnetic field, the multiplicity of lines is reduced to six or less. Nevertheless, the excessive line width produced overlapping, so that the fine structure associated with this exciton was not resolved, although some indication of a splitting at 1·5° K does appear. In order to resolve this structure, it will probably be necessary to use higher magnetic fields.

In examining the behaviour of the exciton as a function of magnetic field up to approximately 10 kgauss, the lines marked 1 and 3, which are the ground and first excited state respectively (Figure 14), show a quadratic Zeeman effect up to 12 and 4 kgauss, respectively, in accordance with the theory indicated

above. In addition, at higher fields they show a linear behaviour owing to the fact that in the Hamiltonian shown in equation (13), the diamagnetic term becomes considerably larger than the Coulomb term. Under these circumstances, the exciton levels can be considered as conduction band Landau levels which are perturbed by a Coulomb potential and hence have a linear dependence on the magnetic field. A theory using a variational calculation for the ground state for spherical energy bands was developed by Yafet, Keyes, and Adams[31] and gave a value of the energy separation between the ground state and the first Landau state of $4{\cdot}5\times10^{-3}$ eV. This compares favourably with the experimental value of approximately 4×10^{-3} eV obtained at different temperatures where this phenomenon was observed. If the Landau levels are extrapolated to zero field, a binding

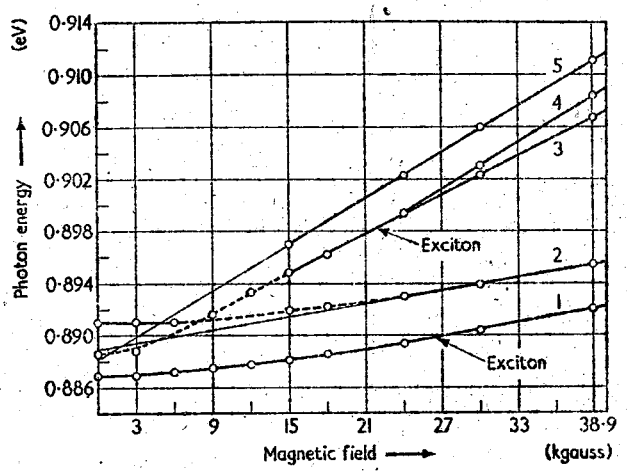


Figure 14. *Position of direct exciton absorption lines and Landau transitions in germanium as a function of magnetic field at 77°K.*[11] *The Landau lines converge to a value of energy gap* $\epsilon_g = 0{\cdot}8886\pm0{\cdot}0004$ *eV. The exciton binding energy, therefore, is* $\epsilon_{ex} = 0{\cdot}0^{\prime}18\pm0{\cdot}0004$ *eV*

energy for the exciton of $\epsilon = (1{\cdot}8\pm0{\cdot}4)\times10^{-3}$ eV at 77° K is obtained.† This is in good agreement with the theoretical value of the exciton binding energy obtained from

$$\epsilon_{ex} \simeq \frac{13{\cdot}6\mu}{\kappa^2 n^2 m_0} \qquad \ldots(15)$$

where μ is the reduced effective mass evaluated using the $k=0$ conduction band electron mass obtained from the magneto-absorption data and an appropriate effective mass for the holes which takes into account the character of the degenerate valence bands, giving a value of $\epsilon_{ex} = 1{\cdot}5\times10^{-3}$ eV.

An anomalous feature which was present in this spectrum is the existence of a line which appears at an energy larger than the direct gap and is quite prominent at zero field. This line has now been accounted for by MMQR[32] who showed that it

† See section added in proof, page 268.

is due to strain in the sample, which was cemented to glass prior to cooling, and was absent when a free mounted sample was used. The strain causes a splitting of the degeneracy of the valence bands, resulting in these two lines. A theoretical analysis of this phenomenon has been given by Kleiner and Roth[33] who have shown that the splitting by strain in two crystal directions can be used to measure the deformation potential.

3.3. Indirect Transition Exciton in Germanium

The Zeeman effect of the indirect transition exciton in germanium has been measured by Zwerdling, Lax, Roth, and Button[11,14] in fields up to 38·9 kgauss. The existence of this exciton absorption was reported by MMQR[16] in their study of the fine structure of the indirect transition absorption edge in germanium.

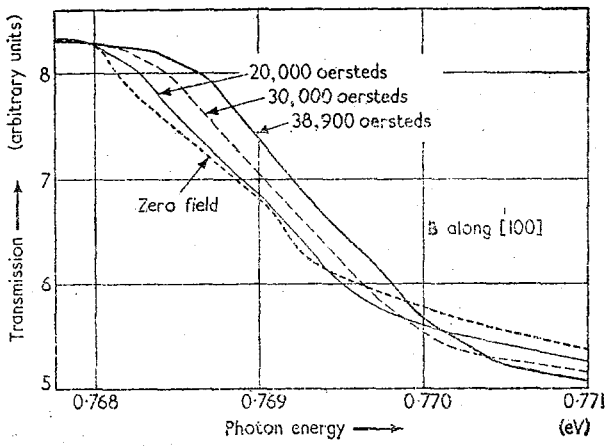


Figure 15. *The transmission trace of the indirect exciton at* 1·5° *K for several values of magnetic field as a function of photon energy. Exciton corresponds to transition with emission of longitudinal acoustical phonon. Traces show structure and non-linear shift of exciton absorption edge with magnetic field*[11]

Zwerdling conducted precise experiments under very high resolution, and in addition observed a fine structure at zero field which indicated the presence of two components of the exciton line as shown in Figures 15 and 16. The experimental binding energies at zero field for these two components were found to be $2{\cdot}1 \times 10^{-3}$ eV and $3{\cdot}2 \times 10^{-3}$ eV. These results are in good agreement with theoretical values of $2{\cdot}5 \times 10^{-3}$ and $3{\cdot}3 \times 10^{-3}$ eV.[11,14] The presence of the two components is primarily due to the lifting of the degeneracy by the terms in the matrix Hamiltonian arising from the ellipsoidal character of the electron energy bands. With increasing magnetic field, the exciton absorption edge shifts to higher energy quadratically and, furthermore, develops additional fine structure which is best analysed by taking the derivative of the transmission curve as shown in Figure 16. This fine structure has been studied by the Lincoln group,[25] involving a detailed examination of the lines as a function of magnetic field, as shown in Figure 17. Roth and Kleiner developed a theory in terms of a spin Hamiltonian, similar to

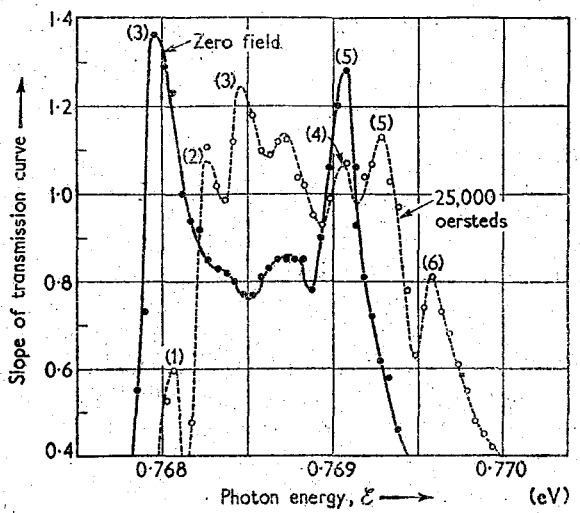


Figure 16. *Derivative curves of the exciton absorption for the indirect transition in germanium for* $\mathbf{H} \| [110], \mathbf{E} \| \mathbf{H}$ *at* $1 \cdot 5°K$. *The two lines corresponding to the zero field splitting of the exciton move to higher energy with magnetic field and develop structure as shown above*[25]

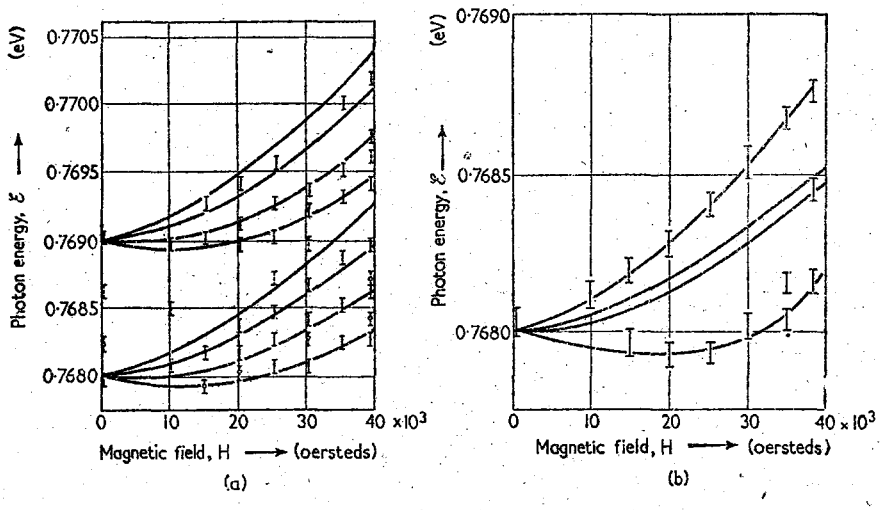


Figure 17. (*a*) *The Zeeman spectrum of indirect exciton in germanium for* $\mathbf{H} \| [100], \mathbf{E} \| \mathbf{H}$ *at* $1 \cdot 5°K$. *The 'bars' represent experimental points. The solid lines are the theoretical curves*[25] *obtained from the spin Hamiltonian of equation* (16) *where* $\mathcal{J} = 3/2$, $g_1 = 1 \cdot 6$, $g_2 = 1 \cdot 6$, $C = 0 \cdot 53$, *and* $D = 10 \cdot 0$; (*b*) *The Zeeman spectrum for* $\mathbf{H} \| [111], \mathbf{E} \| \mathbf{H}$. *The theoretical curves using the above parameters demonstrate agreement with experimental data* (*Button and Roth, unpublished*)

that used in paramagnetic resonance, in which the energies of the system can be obtained from

$$\mathscr{H} \simeq D\left(J_z^2-\frac{J^2}{3}\right)-\mathbf{g}_1.\beta\mathbf{J}.\mathbf{H}+\mathbf{g}_2.\beta\mathbf{S}.\mathbf{H}+CH^2 \qquad \ldots(16)$$

where the first term represents the zero field splitting and J is the total angular momentum of the hole having the value 3/2, the second is the linear Zeeman effect due to holes, the third is the linear Zeeman effect term due to electrons with a spin $S=\frac{1}{2}$, and the last is an isotropic quadratic term which becomes significant above 10,000 oersteds. Button fitted the theory to the experimental data and determined the best values for these parameters with the magnetic field along the [110] direction and $\mathbf{E}\|\mathbf{H}$, as shown in Figure 17(a). The same parameter values also gave very good agreement with the experimental data for the magnetic field along the [111] direction, as shown in Figure 17(b). Although the results did not give a highly accurate value of the g factor of the electrons, nevertheless, they did suggest that it was anisotropic. A theoretical study by Roth[34] indicated that the g factor should be less than 2 and highly anisotropic. It is apparent that to determine these parameters with sufficient accuracy, it will be necessary to perform these experiments at much higher fields and with thicker samples than those used.

The g value of $1\cdot6\pm0\cdot2$ in the [100] direction determined by the Zeeman effect of the indirect exciton in germanium was the first measurement of this quantity. Recently, the spin resonance experiments of Feher, Wilson, and Gere[35] in germanium, which are inherently more accurate, confirmed the above result. Consequently, from the theory of Roth[34] which showed that two components of the electron g factor are related to the properties of the bands by

$$g_{\|}-2=-\frac{\delta}{\Delta\epsilon_{13'}}\left(\frac{m_0}{m_t}-1\right) \qquad \ldots(17)$$

and by

$$g_{\perp}-2=-\frac{\delta}{\Delta\epsilon_{13'}}\left(\frac{m_0}{m_l}-1\right)+\Delta g'_{\perp} \qquad \ldots(18)$$

Roth and Lax[36] were able to deduce further information about the bands. The quantities $g_{\|}$ and $g_{\perp}$ are the components of g factor parallel and perpendicular to the principal axis of the energy ellipsoid of the electrons; $m_t=0\cdot082\ m_0$ and $m_l=1\cdot68\ m_0$ are the electron mass parameters from cyclotron resonance;[1] δ is the spin–orbit splitting of the $L_{3'}$ valence band at the [111] edge of the Brillouin zone; and $\Delta\epsilon_{13'}$ is the energy gap between the $L_{3'}$ and the L_1 band edges, where L_1 is the [111] conduction band minimum and $\Delta g'_{\perp}$ is a small term whose sign is ambiguous but whose value can be estimated. The parameter $\delta\simeq2\Delta/3$, where $\Delta\simeq0\cdot3$ eV, is the spin–orbit splitting of the valence band at $k=0$. Using the experimental data of Philipp and Taft,[37] taken at room temperature, it was deduced that $\Delta\epsilon_{13'}$ corresponded to the absorption edge found at $\simeq2\cdot0$ eV. This gave a value of $g_{\perp}=2\cdot04\pm0\cdot04$ and $g_{\|}=0\cdot9$. From the spin resonance, Roth and Lax computed $g_{\|}=0\cdot63\pm0\cdot08$ and a value of $\Delta\epsilon_{13'}\simeq1\cdot6$ eV corresponding to the gap at 4° K. The discrepancy in the gap value can possibly be due to a temperature effect, although the sign of the change is opposite to that of the usual direct or indirect

gap. The other possibility is that the estimate of δ is inaccurate. Nevertheless, the identification of the L_1 and $L_{3'}$ transition appears credible.

4. MAGNETO-ABSORPTION OF IMPURITIES

4.1. Zeeman Effect of Impurities

The impurity levels of donors and acceptors in germanium and silicon have been studied both theoretically and experimentally. The theory for the donors has been more extensively developed and has been most successfully applied to silicon. The theoretical work for acceptors is rather involved and is still in its preliminary stages. Experiments in N-type germanium have been carried out in the far infra-red by Boyle[38] and also by Fan and Fisher;[39] the latter have also made measurements in P-type germanium at these wavelengths.[40] Complete Zeeman spectra of both N- and P-type silicon have been obtained by Zwerdling, Button, and Lax.[41]

The theory for the donors can be formulated using the effective mass Schroedinger equation which takes the form

$$\left(\frac{\pi_x^2+\pi_y^2}{2m_t}+\frac{\pi_z^2}{2m_l}-\frac{e^2}{\kappa r}\right)\Psi=\epsilon\Psi \qquad \ldots (19)$$

where

$$\pi_j=\frac{\hbar}{i}\left(\frac{\partial}{\partial j}-\frac{eA_j}{c}\right)$$

$$\mathbf{A}=\tfrac{1}{2}\mathbf{H}\times\mathbf{r} \quad \text{and} \quad j=x,y,z$$

When this is written in explicit form, one obtains expressions which contain in addition to the zero field Hamiltonian, terms linear and also quadratic in the magnetic field. At low fields, the quadratic terms can be neglected, and the linear terms in H may be treated as a perturbation upon the hydrogenic solutions of the zero field wave equation. For germanium and silicon, such a treatment has been developed by Lax, Puff, and Kleiner[42] and also by Haering[43] in which the splitting of the excited $p_\pm$ states was considered. They have shown that this splitting for each ellipsoid is given by $\Delta\epsilon=\hbar\omega_t\cos\Phi$, where $\omega_t=(eH)/(m_t c)$, m_t is the transverse effective mass, and Φ is the angle between the magnetic field and the principal axis of the ellipsoidal energy surface. The anisotropy of the linear Zeeman effect using the above result has been plotted for both silicon and germanium as shown in Figure 18(a) and 18(b). In silicon, when the magnetic field is along the [001] axis, the $p_\pm$ states will show a maximum splitting for the two ellipsoids along this axis, but for the four ellipsoids whose principal axes are transverse to this direction, there will be no splitting. As the magnetic field is rotated in the (110) plane so as to be parallel to the [111] axis, all six ellipsoids will be equivalent and will contribute equally to the splitting of the $p_\pm$ states. With the field along the [110] axis, the two ellipsoids perpendicular to the field will show no splitting, and the other four will contribute to split the $p_\pm$ states as shown in Figure 18(a). This plot is very helpful in interpreting the data obtained in the Zeeman spectrum of donors in silicon. Furthermore, it can be shown that if linearly polarized radiation is used, only those ellipsoids can participate in transitions from the donor ground state to the $p_\pm$ states for which the electric vector has a component perpendicular to the principal axis. In a similar way, one can account for the spectrum in germanium

associated with the four ellipsoids, which are located along the [111] directions. Consequently, when the magnetic field is along the [100] directions, all the ellipsoids are equivalent and the linear Zeeman spectrum for the $p_{\pm}$ states consists of

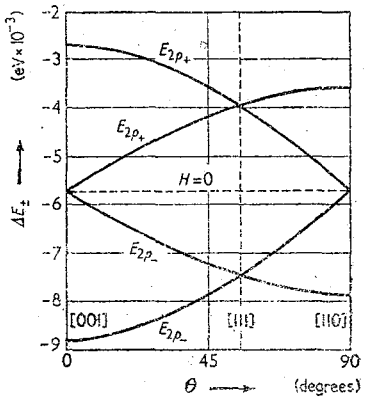


Figure 18(a). *Anisotropy of the linear Zeeman splitting of the $2p_{\pm}$ states in silicon in a* (110) *plane as a function of angle relative to a* [001] *direction*[4] *for a field of* 100 *kgauss*

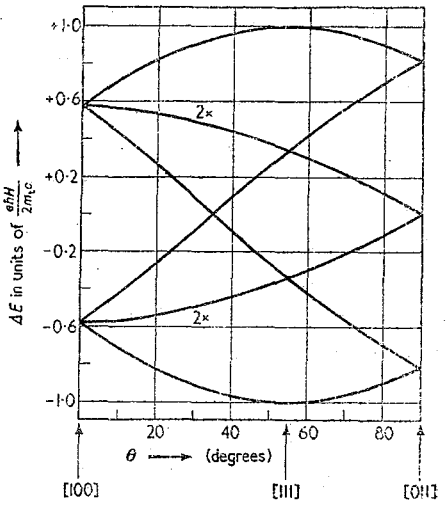


Figure 18(b). *Anisotropy of the linear Zeeman splitting of the $2p_{\pm}$ states in germanium in a* (110) *plane as a function of angle relative to a* [100] *direction.*[43] *(The symbol 2× indicates a two-fold degeneracy)*

two lines. With **H** along the [111] direction, the ellipsoid whose principal axis coincides with the magnetic field direction will have a large Zeeman splitting as shown in Figure 18(b), but the other three will show a smaller splitting. Finally,

with **H** along the [011] direction, two ellipsoids which are perpendicular to this direction will show no splitting but two which make a relatively small angle ($\simeq 35°$) will produce a splitting. Hence three absorption lines will occur for this direction of the field.

To interpret the Zeeman spectrum obtained experimentally for both silicon and germanium with the d.c. magnetic fields currently available, it is necessary to consider the Schroedinger equation, including the quadratic terms in the magnetic field. These magnetic terms can be treated as a perturbation. Such an approach is quite effective for silicon where the Zeeman effect has been studied up to 38·9 kgauss. In germanium, however, even above 10 kgauss, a perturbation treatment is not appropriate for the excited states, since the quadratic or diamagnetic term becomes more dominant than the Coulomb term. In this event, one can treat the problem as corresponding to a nearly free electron with the Coulomb term being a small perturbation on the Landau levels. The perturbation treatment can still be applied effectively to the ground state of N-type germanium even at these fields. In doing so, the zero field solutions are assumed to be of the variational form $\Psi \sim \exp(-\alpha r)$, where α is evaluated in the usual way. The energy shift is then calculated by treating the magnetic terms (including the quadratic terms) as a perturbation, thereby obtaining an expression for the quadratic Zeeman effect given by

$$\Delta\epsilon = \frac{\kappa^2 \hbar^2 \omega_c^{*2}}{16 m_t e^4} [(3+\rho)+(1-\rho)\cos 2\theta] \qquad \ldots (20)$$

where

$$\rho = m_t/m_l, \quad \omega_c^* = \frac{eH}{m^* c}, \quad \text{and} \quad \frac{1}{m^*} = \frac{2}{3m_t} + \frac{1}{3m_l}$$

and θ is the angle between **H** and the [001] axis in the $(1\bar{1}0)$ plane. From this expression, the anisotropy of the quadratic Zeeman effect, including all four ellipsoids, can be readily obtained.[44] Such an anisotropy diagram is very similar to that obtained for the excited states if one ignores the small Coulomb perturbation on the Landau levels and calculates the Zeeman pattern using the effective masses for electrons obtained from cyclotron resonance. The linear Zeeman effect gives information only about the transverse effective mass, whereas equation (20) shows that the quadratic Zeeman effect includes terms both for the transverse and longitudinal masses and hence in effect permits the determination of both parameters.

Calculations of the Zeeman effect for the excited states of N-type silicon[45] were made by using perturbation theory to higher order terms and assuming zero field wave functions of the form

$$\exp\{-\sqrt{[A^2(x^2+y^2)+B^2z^2]}\}$$

In one approximation $A = B$, and in another A and B were determined by the variational solution of the zero field effective mass equation obtained by Kohn and Luttinger[46] and also by Kleiner.[47] The solution for the former approximation has been carried out by Brown[45] for the $2p$ states.

The Zeeman splitting of the $2p_\pm$ state in N-type phosphorus-doped germanium was obtained by Boyle[38] with the magnetic field along the [100] direction and is

shown in Figure 19. The two-fold splitting is in accordance with the theory and in approximate agreement with the result calculated from $\Delta\epsilon = \hbar\omega_t \cos\Phi$ when the transverse effective mass of the electron obtained from cyclotron resonance[1] is used. Similar results for N-type arsenic-doped germanium were obtained by Fan and Fisher[39] for fields from 6,000 to 16,000 gauss. Using the value of the Zeeman splitting at the lowest field where the theory is applicable, they obtained a value for the transverse mass, $m_t = (0{\cdot}077 \pm 0{\cdot}005)m_0$, slightly lower than the cyclotron resonance value of $0{\cdot}082\, m_0$, but nevertheless in reasonable agreement with theory.

The Zeeman spectrum for silicon has been more extensively investigated and was measured for the magnetic field along the three principal directions in the

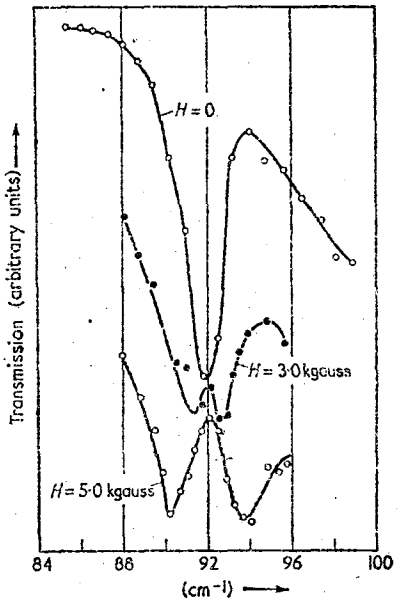


Figure 19. *The Zeeman splitting of a $2p_\pm$ phosphorus impurity level in germanium*[38]

(110) plane by Zwerdling, Button, and Lax.[41] These experiments were performed at 4·2° K using magnetic fields up to 38·9 kgauss with linearly polarized radiation oriented both parallel and perpendicular to the magnetic field. A spectrometer utilizing double pass potassium bromide prism dispersion in the wavelength region 15–22 microns was used. The donor impurity, bismuth, was chosen for the N-type material and the acceptor impurity, aluminium, for the P-type material as they are both appropriate for study in this spectral region. A plot of the initial data obtained for the Zeeman spectrum of the excited donor levels with the magnetic field along a [100] axis and, for the polarized electric vector of the radiation both parallel and perpendicular to the field, is shown in Figure 20. With the electric vector perpendicular to the magnetic field, the Zeeman splitting of the $2p_\pm$ and the $3p_\pm$ states is quite apparent, arising from the two ellipsoids with major axes parallel

to the field and perpendicular to the electric vector. There is also a 'central' component in these absorptions, contributed in this case by the other four ellipsoids, since their major axes are not parallel to the electric vector, but are orthogonal to the field. However, when the electric vector is parallel to the field, only the central component can occur in the Zeeman spectrum, since the electric vector is now orthogonal to the major axes of the latter four ellipsoids, but is parallel to those of the former two which, although producing a splitting, do not contribute to the absorption since their momentum matrix elements vanish. For this particular case, only the quadratic Zeeman effect is apparent and becomes observable above 20 kgauss. Furthermore, the quadratic Zeeman effect of the p_0 states may thus be

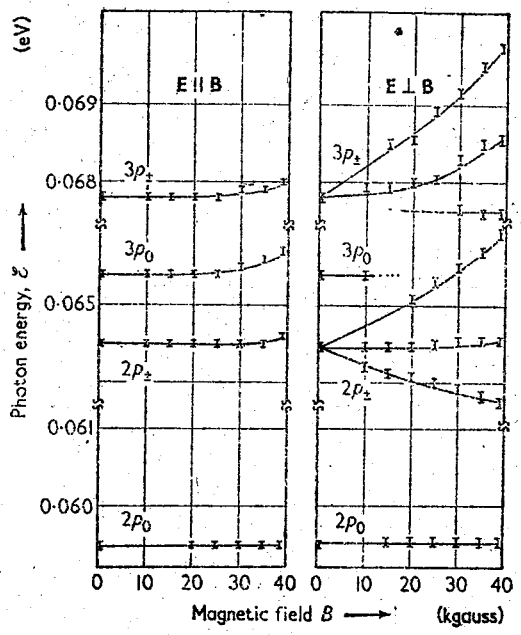


Figure 20. *The Zeeman pattern of bismuth donor levels in silicon for parallel and perpendicular polarized radiation.*[41] B∥[100]

studied without interference from overlapping by split components of the $p_\pm$ states, as is evident in Figure 20 for the $3p_0$ and $2p_+$ absorptions.

A comparison of initial experimental data for B∥[110] with the perturbation theory calculated for the $2p_\pm$ states is shown in Figure 21. The calculation included the quadratic terms in the magnetic field and matrix elements for higher states up to the $3p$ levels, and the transverse effective mass of the electron ($m_t = 0{\cdot}19\ m_0$) as obtained from cyclotron resonance[1] was used. The agreement is surprisingly good for the central component and the p_- state, but the theoretical curve falls below the experimental data for the p_+ state. This is probably due to the fact that this state shifts sharply upward towards the ionization limit with increasing field and consequently the coupling with higher excited states such as the $4p$ and above should be included in the theory, which would give better agreement. Similar

detailed data for the Zeeman spectrum of aluminium-doped P-type silicon have also been observed.[41] The energies of the excited states at zero field are in good agreement with the data obtained by Hrostowski and Kaiser[48] and in approximate agreement with the theoretical results of Schechter and Kohn.[49] From first-order theory, it is expected that the lowest four states would be four-fold, two-fold, and two-fold degenerate, respectively. This is borne out by the experimental data where the magnetic field removes these degeneracies. A quantitative comparison between theory and experiment awaits the results of a theoretical calculation in which the Zeeman spectrum is to be represented by a spin Hamiltonian analogous to that for the indirect exciton. The coefficients for the linear and quadratic terms are evaluated in a manner similar to that used for calculating the Zeeman spectrum

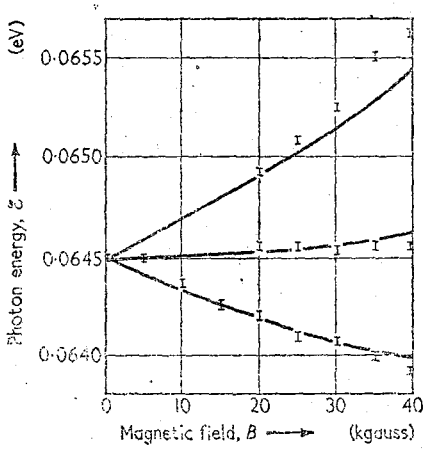


Figure 21. *Comparison of theory and experiment of the Zeeman effect of the $2p_{\pm}$ bismuth impurity levels.* **B**‖[110]. *The solid lines are obtained from the first-order and second-order perturbation theory including terms for 3p states. The 'bars' are experimental data (Brown and Button, unpublished)*

of the donor states. The zero field wave functions obtained by Schechter[49] are used in the perturbation theory with the magnetic terms in the four-fold matrix Hamiltonian for the excited states as the perturbing terms.

4.2. Oscillatory Magneto-absorption of Impurities

In a manner analogous to the oscillatory magneto-absorption observed for the direct transition across a forbidden energy gap, it is possible to observe a similar phenomenon of a transition from an impurity level to a magnetic level beyond the ionization limit. The usual circumstance is that transitions occur from the occupied ground state of an impurity to a Landau level in the conduction or valence bands, for donor or acceptor impurities respectively. It is probable in this case that the Landau level is perturbed by the Coulomb potential of the impurity. The study of this type of transition has the advantage over that of the interband transition in that it involves the properties of only one of the bands at a time, whereas the latter

involves both conduction and valence bands simultaneously. The only drawback to the use of the impurity magneto-absorption method of energy band studies is that the ground state at high magnetic fields also has a quadratic Zeeman effect and that, where it occurs, the degeneracy of bands is removed, splitting the ground state or effectively broadening it. Hence, in interpreting the oscillatory spectrum obtained, it would be necessary to include these effects quantitatively. The spacing of absorption maxima would not necessarily correspond to the separations of Landau levels as encountered in cyclotron resonance.

The first report of an impurity magneto-absorption measurement was made by Boyle and Brailsford[50] who studied the infra-red absorption of a 0·01 cm thick sample of N-type indium antimonide containing 2×10^{14} carriers cm^{-3} at liquid helium temperatures at wavelengths from 70 to 120 microns. The spectrum in the presence of a magnetic field exhibited two transmission minima at 18 and 19 kgauss. The data were interpreted as transitions between the ground state and bound Landau levels represented by wave functions of the form $\Phi(n,l)\exp -(z^2/4a^2)$, where $\Phi(n,l)$ is the solution of the wave equation in cylindrical co-ordinates[51] for a free carrier in a magnetic field. The exponential term in z accounts for the Coulomb attraction and n and l are quantum numbers associated with the transverse co-ordinates. Using a variational solution, the transitions were identified as those from level $n = 0$, $l = 0$ to $n = 1$, $l = 0$, 2 with a separation calculated theoretically to be $0{\cdot}5\hbar\omega_c$. From the slope of the curve for the energy of the transitions as a function of magnetic field, the value of the effective mass was determined to be $0{\cdot}0146\ m_0$.

A theoretical treatment of the impurity magneto-absorption has been given by Wallis and Bowlden.[52] They have made specific calculations of the absorption for indium antimonide at high fields and also made a qualitative comparison of the data of Boyle and Brailsford with their theoretical results. The discrepancy between the two was attributed to the absence of complete freeze-out of carriers at the magnetic fields used, as had been observed by Keyes and Sladek[53] at higher magnetic fields of 50–60 kgauss. However, there is a difference in the physical interpretations in that Wallis and Bowlden make calculations of transitions from a bound ground state to free Landau states, whereas Boyle and Brailsford consider their results as being transitions between bound states only. Elliott and Loudon[27] comment on this matter by suggesting that, in general, one can neglect the effect of the coulomb perturbation at high magnetic fields without serious errors, i.e. for eigenvalues. However, in order to calculate line intensities, it is necessary to include this effect.

The transitions to the higher Landau levels in the conduction band, showing the oscillatory character of the impurity magneto-absorption spectrum, were demonstrated by Boyle[28] in N-type germanium where he obtained oscillations of the transmission at 73 microns as a function of field, well above the ionization limit for the donor, arsenic. The amplitude of the oscillations apparently increased with magnetic field. A more useful representation depicted in Figure 22 shows the oscillations as a function of energy at a fixed magnetic field of 21·8 kgauss, similar in character to that observed for the direct transition in germanium.[5] The intensity decreased with increasing energy corresponding to transitions from the ground state to Landau levels of higher n values. Boyle plotted the energy of the transmission minima for two different values of magnetic field versus the Landau

quantum number n, as shown in Figure 23. The intersection of the two lines at $n = -1/2$ corresponds to the value of the ionization energy.† The ionization energy

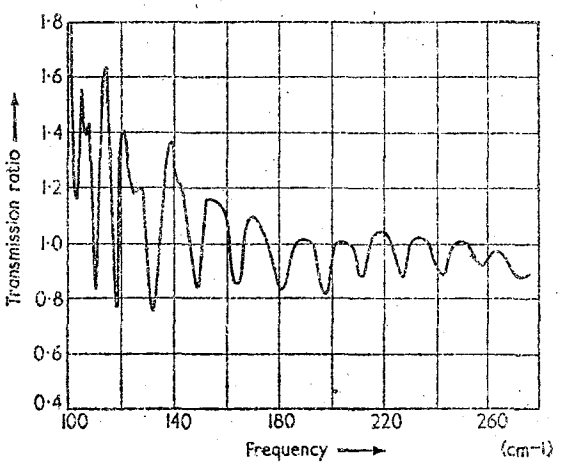


Figure 22. *The oscillatory magneto-absorption of arsenic impurities in germanium as a function of frequency.*[38] *The plot shows the transmission ratio with and without the field for* $\mathbf{B} = 21 \cdot 8$ *kgauss*

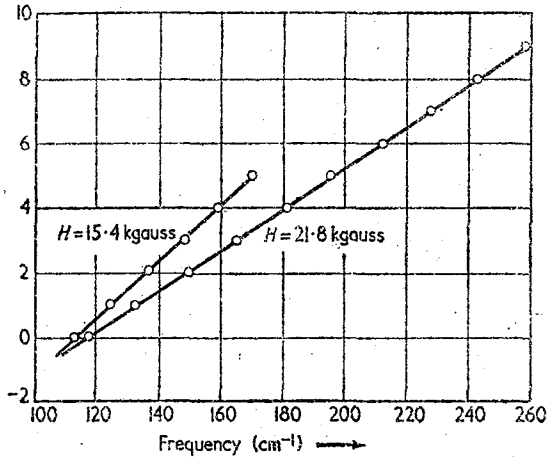


Figure 23. *Position of absorption maxima as a function of magnetic quantum number n for the oscillatory magneto-absorption of arsenic in germanium. Intersection at* $n = -1/2$ *corresponds to ionization* $\epsilon_I = 0 \cdot 0134$ *eV* (108 cm^{-1})

for arsenic in germanium is thus found to be 108 cm^{-1} or 0·0134 eV. From the slopes of the lines he also obtains an electron effective mass value for the zone edge in the [100] direction of 0·13 m_0, in good agreement with cyclotron resonance

† See section added in proof, page 268

results.[1] Similar experiments were also performed by Fan and Fisher[39] for both N- and P-type impurities in germanium. For the P-type material, the oscillations showed a pattern having two series of transmission minima corresponding to the light and heavy holes.

5. FARADAY EFFECT

5.1. Indium Antimonide

Experiment utilizing the Faraday effect in semiconductors is fundamentally a technique which measures the rotation of a plane polarized electromagnetic wave propagating parallel to a magnetic field in the material. The rotation is a consequence of the differential dispersion of the two circularly polarized components of the plane polarized wave, in that the left- and right-hand circularly polarized waves interact differently with holes and electrons in the presence of a field. Faraday rotation was observed in germanium by Rau and Caspari[54] at microwave frequencies. However, it was pointed out by Mitchell[55] that similar experiments on free carriers at infra-red frequencies could also be performed and used to determine effective masses of such carriers. The first experiments at infra-red frequencies were performed by von Kimmel[56] who made measurements in silicon, gallium arsenide, indium phosphide, and gallium phosphide in the region from 0·5 to 1·5 microns. However, no quantitative correlation between the experimental data and the fundamental band properties of the semiconductors was attempted. Experiments yielding quantitative results and interpretation have now been successfully performed on indium antimonide† by Moss, Smith, and Taylor[57] and also by Brown and Lax.[58] The theory of the phenomenon has been considered by Stephen and Lidiard.[59]

For free carriers, the theory of the Faraday rotation can be derived from a simple classical analysis of the Drude–Zener type. An expression for the effective conductivity σ of electrons for circularly polarized waves may be derived, giving

$$\sigma_{\pm} = \frac{ne^2\tau}{m_2[1+j(\omega \mp \omega_{c_2})\tau]} \qquad \ldots (21)$$

where n is the electron density in number per unit volume, ω_{c_2} is the cyclotron frequency of the electron with effective mass m_2, and the minus and plus sign indicates the two senses of circular polarization. Using Maxwell's equations, it can be shown that the phase constant β is given by

$$\beta_{\pm} \simeq \frac{\omega\sqrt{\kappa}}{c}\left(1 \pm \frac{ne^2\omega_{c_2}}{2\kappa\omega^3 m_2}\right) \qquad \ldots (22)$$

The last expression in the equation assumes that the magnetic field is far below resonance and the collision time is sufficiently long so that losses due to collision may be neglected. The Faraday rotation per unit length, θ, is defined as

$$\theta = \frac{\beta_+ - \beta_-}{2} \qquad \ldots (23)$$

† See section added in proof, page 268.

If the results of equation (22) are substituted into this expression, then

$$\theta = \frac{ne^2}{2m_2 c\omega^2\sqrt{\kappa}}\frac{\omega_{c_2}}{} = \frac{ne^3 H}{2m_2^2 c^2\omega^2\sqrt{\kappa}} = \frac{ne^3 H\lambda^2}{8\pi^2 m_2^2 c^4\sqrt{\kappa}} \qquad \ldots (24)$$

where $\omega_{c_2} \ll \omega$ and $\omega_{c_2}\tau \gg 1$. This result shows that the Faraday rotation well below cyclotron resonance is a linear function of magnetic field, inversely proportional to the square of the effective mass, and directly proportional to the square of the wavelength, λ. The importance of this result is that it indicates that one can make determinations of the effective mass of a carrier at infra-red wavelengths where cyclotron resonance is not possible, even though $\omega\tau > 1$, because of the unavailability of sufficiently high magnetic fields. Furthermore, attempts to utilize cyclotron resonance at lower frequencies in the microwave region where adequate magnetic fields are available have thus far been mainly unsuccessful

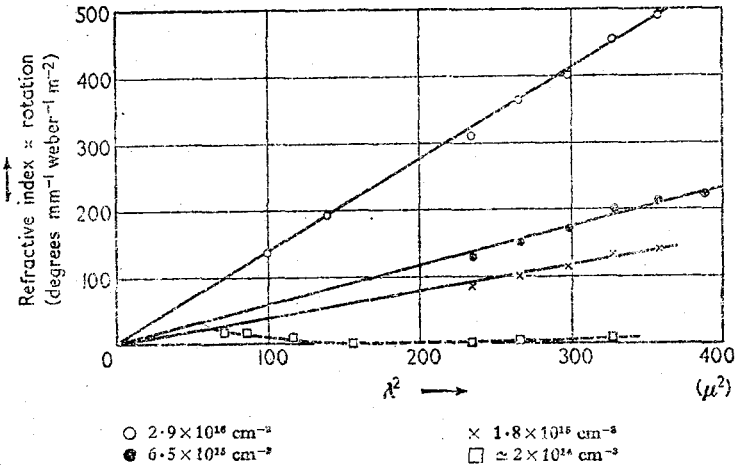


Figure 24. *Faraday rotation in N-type indium antimonide as a function of λ^2 showing a linear behaviour for $\mathbf{B} = 4$ kgauss*[57]

because then $\omega\tau < 1$ for nearly all semiconducting materials at present available except germanium and silicon. An effective mass value for electrons in indium antimonide has been obtained from experiments designed to utilize the relationship given by equation (24). In the wavelength region where the rotation is a linear function of magnetic field and also follows the λ^2 dependence, the rotation is determined experimentally for a particular λ and H. The electron concentration is obtained independently from a Hall effect measurement, and by using the dielectric constant which is calculated at these wavelengths from reflectivity measurements, i.e. $\kappa = 16$, the effective mass may be calculated.

The experimental results of Moss, Smith, and Taylor (MST)[57] were obtained both at room temperature and at 77° K. Their data plotted as a function of λ^2 for a magnetic field of 4 kgauss is given in Figure 24, and showed a linear behaviour in accordance with the simple relation of equation (24). By varying the impurity concentration in indium antimonide, MST were able to evaluate the effective mass

as a function of increasing energy in the conduction band, since at 77°K the measurements yield the values at the Fermi level, which shifts to higher energies with increased concentration. Their results showed that the energy surface was non-parabolic, consistent with results of cyclotron resonance at high magnetic fields[60] and free carrier absorption for varying impurity concentrations.[61] However, in interpreting such Faraday rotation data for more general energy surfaces, Stephen and Lidiard[59] extended the calculation by solving the Boltzmann transport equation and showed that for a degenerate semiconductor the usual effective mass relationship is replaced by the expression

$$\frac{1}{m_2} = \frac{1}{\hbar^2}\left(\frac{\mathrm{d}\epsilon}{k\,\mathrm{d}k}\right)_{\epsilon_F} \qquad \ldots (25)$$

This shows that the Faraday effect measures a reciprocal mass related to the slope of the energy–momentum curve rather than the curvature, which is the

Table 1. Effective Masses of Electrons in Indium Antimonide from Faraday Rotation[57]

$N(\mathrm{cm}^{-3})$	m_2/m_0†	Fermi energy (eV)	m_2/m_0‡
$6{\cdot}4\times10^{17}$	0·0288	0·135	0·0265
$2{\cdot}9\times10^{16}$	0·0178	0·027	0·0158
$7{\cdot}3\times10^{15}$	0·0137	0·010	0·0141
$2{\cdot}6\times10^{15}$	0·0131	0·005	0·0136
0	—	0	0·0130

† See reference 57. ‡ See reference 62.

second derivative. For a simple parabolic band, it can be shown that the two are equivalent. Keeping this in mind, the effective mass results show a dependence on impurity concentration, which was interpreted as shown in Table 1 and compared with the theoretical estimates of Kane[62] obtained from perturbation theory. The correspondence between the theory and experiment is quite satisfactory, with the low mass value at low concentrations in good agreement with that obtained by cyclotron resonance at microwaves.

An explanation of the deviation of the Faraday rotation from the λ^2 dependence for decreased wavelengths has been suggested by Stephen and Lidiard. The reasoning given by them was only qualitative and based on the assumption that as the wavelength approaches a value corresponding to that of the energy gap, there would be absorption due to the associated oscillator frequency. However, an interpretation of such experimental results can be made by a phenomenological treatment.† At decreasing wavelengths, the bound electrons in the valence band are disturbed by the electromagnetic radiation and behave as simple harmonic oscillators with a frequency ω_0. If we include the effect of the magnetic field, it

† See section added in proof, page 268.

can be shown, in a manner similar to that for the free carriers, that the effective dielectric constant due to the bound electrons is

$$\kappa_{\pm} = \frac{Ne^2\omega}{m_0[\omega_0^2-\omega(\omega\pm\omega_{c_0})]} \qquad \ldots (26)$$

Consequently, the total Faraday rotation is given by†

$$\theta = \frac{e^3H}{2c^2\omega^2\sqrt{\kappa}}\left[\frac{n}{m_2^2} - \frac{N\omega^2}{m_0^2(\omega_0^2-\omega^2)^2}\right] \qquad \ldots (27)$$

where N is the total number of electrons in the valence band (which for indium antimonide is of the order of 10^{22}) and m_0 is taken as the free space electron mass.

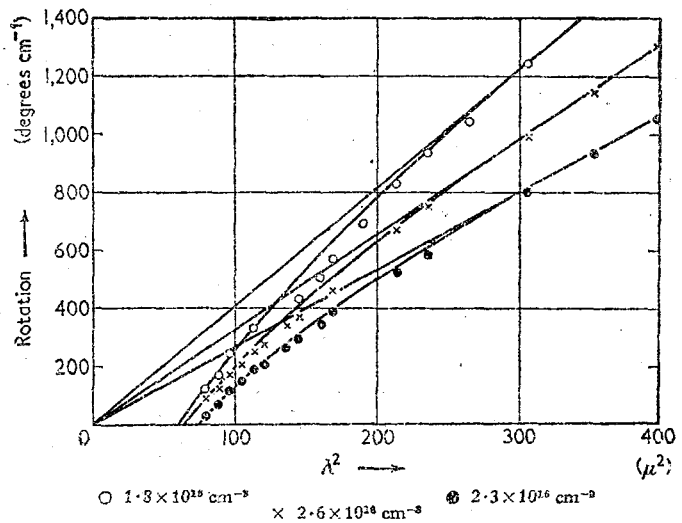


Figure 25. Faraday rotation in N-type indium antimonide at $\mathbf{B} = 15$ *kgauss field as a function of* λ^2 *showing non-linear behaviour at lower wavelengths due to bound valence electrons*[86]

Using this expression and taking $\theta = 0$, then from the results of Figure 25, a value of the oscillatory frequency of the bound electrons is found corresponding to a wavelength $\lambda_0 \simeq 2$ microns, or to an energy which is considerably higher than that of the energy gap. The important consequence of this treatment is that it does show the possibility of negative Faraday rotation due to the valence electrons. However, the theory appears somewhat oversimplified and in its present form does not account quantitatively for the value of the effective binding of the valence electrons since it thus gives an energy of $\simeq 0{\cdot}6$ eV.

5.2. Germanium and Silicon

The Faraday effect has also been observed by Walton and Moss[83] in N-type germanium. The theory for germanium in the linear magnetic field region has

† See section added in proof, page 268.

been considered by Lidiard and Stephen.[64] This work is very interesting in that it is the first measurement of the effect in a material whose energy surfaces are anisotropic. Although the linear Zeeman effect implicitly reflects the anisotropy of the ellipsoidal conduction band surfaces, it does not reveal information about their parameters. The Faraday effect, of course, can be used to study more complicated energy bands by using sufficiently high fields and including higher order terms in the theory. For germanium this is not possible with the fields at present available. Nevertheless, for the more general case of conduction electrons in silicon with the six ellipsoidal surfaces along the cube axes and the magnetic field along the [111] axis, and also for germanium with the four ellipsoidal surfaces along the cube diagonals and the magnetic field along the [100] axis, the Faraday rotation derived from the results of Lax and Roth[65] is given by

$$\theta = \frac{ne^2(K+2)b}{2\omega c m_t \sqrt{\kappa}[3K-b^2(K+1)]} \qquad \ldots (28)$$

where $K = m_l/m_t$ and $b = \omega_t/\omega$. Similar but more complicated expressions have been developed for the magnetic field along other principal crystallographic directions for silicon and germanium. By assuming that $b \ll 1$, then from equation (28)

$$\theta = \frac{ne^3 K(K+2)H}{6\omega^2 c^2 m_l^2 \sqrt{\kappa}} \qquad \ldots (29)$$

The above result is that obtained by Lidiard and Stephen[64] and corresponds to the isotropic linear Faraday effect for a cubic crystal. It is obtained as the first term in the expansion of equation (28) in a power series of b or H. Using the expression of equation (29), the results of Hall effect measurements to determine the carrier concentration, and the value of the dielectric constant in the wavelength region 6–13 microns, i.e. $\kappa = 16$, an effective mass $m^* = [3/K(K+2)]^{1/2} m_l$ was evaluated from the observed Faraday rotation by Walton and Moss. They found that $m^* = (0{\cdot}135 \pm 0{\cdot}004) m_0$, in good agreement with the equivalent value from cyclotron resonance data which gave $m^* = 0{\cdot}134\, m_0$.

In a manner analogous to the treatment for electrons in germanium and silicon, one can obtain an expression for the linear Faraday effect in P-type material in which the energy surfaces are given by the relation

$$\epsilon = \frac{\hbar^2}{2m_0}\{Ak^2 \pm \sqrt{[B^2k^4 + C^2(k_x^2 k_y^2 + k_x^2 k_z^2 + k_y^2 k_z^2)]}\} \qquad \ldots (30)$$

Expressions for the conductivity tensor for these surfaces have been developed by Lax and Mavroides[66] which take into account the Boltzmann distribution of the holes. Using these results, the linear Faraday rotation angle for holes is given by the equation

$$\theta = \frac{e^3 H}{2\omega^2 c^2 \sqrt{\kappa}}\left(\frac{n_H}{m_H^2} + \frac{n_L}{m_L^2}\right) \qquad \ldots (31)$$

where

$$m_{H,L} = \frac{m_0}{(A \mp B')}(1 + 0{\cdot}025\,\gamma_{\mp} + 0{\cdot}00786\,\gamma_{\mp}^2 + \ldots)$$

$B' = (B^2 + \frac{1}{6}C^2)^{1/2}$, and $\gamma_{\mp}$ is a parameter which is evaluated in reference 66. The ratio of the light and heavy hole populations is given by $n_L/n_H = (m_{dL}/m_{dH})^{3/2}$ where m_{dL} and m_{dH} are the effective masses corresponding to the density-of-states values. For silicon, $n_L/n_H = 0{\cdot}136$, $m_H = 0{\cdot}52\ m_0$, $m_L = 0{\cdot}164\ m_0$, and for germanium, $n_L/n_H = 0{\cdot}040$, $m_H = 0{\cdot}35\ m_0$, $m_L = 0{\cdot}044\ m_0$. Thus for silicon, the ratio of $\theta_L/\theta_H = 1{\cdot}37$ indicates that the light holes contribute more to the Faraday rotation than the heavy holes. Similarly for germanium, the ratio $\theta_L/\theta_H = 2{\cdot}55$ shows that the light holes dominate even more. However, if the magnetic field becomes large, the analysis becomes more involved, particularly for the light hole in germanium, and as suggested by equation (28) the ratio $\theta_L/\theta_H \simeq (\omega_L\omega)/(\omega^2 - \omega_L^2)$, where $\omega_L = eH/m_L c$. Thus for large fields, but for values still below cyclotron resonance where the Faraday rotation experiments are appropriate, the contribution of the light hole becomes more important than that of the heavy hole and the departure from linear dependence on the field can be used to measure its mass.

5.3. Interband Transitions

It has been suggested that the Faraday rotation can be used to study cyclotron resonance directly. However, if the required fields are available, the absorption techniques would be more sensible. Nevertheless, there are situations where the Faraday rotation procedure could be used advantageously compared to the absorption technique. It is possible to study all the magneto-absorption phenomena previously described by means of Faraday rotation. Associated with the magneto-absorption of the direct transition there is also a dispersive component which can be shown to have the form

$$\beta_{\pm} \simeq \left[\frac{(x_{\pm}^2+1)^{1/2} - x_{\pm}}{x_{\pm}^2+1}\right]^{1/2} \qquad \ldots (32)$$

where $\hbar x_{\pm} = \{\hbar\omega - [\epsilon_g + (n+1/2)\hbar\omega_c \pm g\beta H]\}\tau$, ω_c is the cyclotron frequency for the reduced mass of the electron and hole, and g is an effective spectroscopic factor which takes into account the anomalous values for both the hole and the electron due to spin–orbit coupling, as stated earlier. The significance of this result is that the Faraday rotation associated with the direct transition, where $\Delta n = 0$, distinguishes between positive and negative circular polarization only, due to the spin effects and the presence of spin–orbit coupling. Hence no rotation will occur if the latter is absent and in general if the g value is small the rotation may be difficult to observe. The oscillatory phenomenon will then occur with a splitting about each Landau transition. If the magnetic field is sufficiently high to resolve the splitting into two lines, each will be associated with a frequency-dependent anomalous dispersion curve, one for right-hand and the other for left-hand circularly polarized radiation. As the spectrum is scanned from shorter to longer wavelengths than both these lines respectively, using linearly polarized radiation with $\mathbf{E} \perp \mathbf{B}$, the Faraday rotation will change from one sense to the other twice, but in opposite sequence. In an analogous manner, one can also study the Faraday rotation of the forbidden direct transition, where an oscillatory spectrum occurs only when the electric vector of the radiation is perpendicular to the magnetic field. For this polarization, transitions are allowed only for $\Delta n = \pm 1$, so that a Faraday rotation spectrum would contain contributions associated with the transition between

Landau levels in addition to those involving the effect of the spins. Consequently, such a spectrum would have more structure than that of the allowed direct transitions. Perhaps one of the most interesting applications of the Faraday rotation technique to magneto-absorption phenomena would be that which is associated with the indirect transition. In this case, the absorption coefficient has the form $\alpha_{\pm} \sim \arctan x_{\pm}$, where $x_{\pm}$ is proportional to the expressions for the energy given by equation (8) and similar to that for the direct transition except that it involves phonon energies and two quantum numbers n_1 and n_2 associated with the valence and conduction bands, respectively. Since there are no selection rules for these quantum numbers, the Faraday rotation will not be affected by the transition associated with the Landau levels, but will distinguish between the splitting due to the spin effects. Consequently, the Faraday rotation should be a useful method for determining the g values of the electron in germanium since the g values of the holes are known from the oscillatory magneto-absorption of the direct transition. The intriguing aspect of this experiment is that whereas the magneto-absorption in this instance produces a step-function in the spectrum as shown in Figure 9, the Faraday effect would result in a spectrum which would contain corresponding maxima of rotation, since $\beta_{\pm} \sim \log(x_{\pm}^2+1)^{-1}$. In a similar manner, the Faraday rotation associated with the Zeeman effect of the indirect exciton would be determined by spin effects arising from spin–orbit coupling, and would again result in rotation maxima rather than the 'edges' of absorption spectra. For sufficiently high fields, this would provide a convenient study of the fine structure of the Zeeman spectrum for the indirect exciton, and allow the determination of the g factors associated with the holes and electrons.

5.4. Impurity Levels

Another experiment that would be of interest is the study of the Faraday effect of the transitions from the ground state to the excited states of impurities in semiconductors. The theory of such a phenomenon has been considered for the classical atom by Rosenfeld[67] some time ago. After modifying and simplifying his results to fit the present problem, it can be shown that the differential dispersion or phase constant of interest becomes

$$\Delta\beta_{\pm} = \frac{Ne^2}{2m^* c\sqrt{\kappa}} \sum_n \frac{f_n \omega_n}{(\omega_n \pm \omega_c^*)^2 - \omega^2} \qquad \ldots (33)$$

where ω_c^* is the cyclotron resonance frequency for carriers in a spherical band, ω_n is the oscillator frequency, and f_n is the oscillator strength. However, for conduction electrons as in germanium or silicon at low magnetic fields where the linear theory holds, ω_c^* should be replaced by $\omega_t \cos\phi$, where ϕ is the angle the magnetic field makes with the principal axis of the ellipsoid, and ω_t is the cyclotron frequency corresponding to the transverse electron effective mass. The Faraday rotation in this case would only occur for the magnetic field splitting of the $p_{\pm}$ states, since these are the ones in which the sense of polarization is differentiated. It is also apparent that the expression of equation (28) would apply with respect to the oscillatory absorption spectrum for the transition between the ground state and bound Landau levels above the ionization limit.

6. MAGNETO-PLASMA EFFECTS

6.1. Magneto-plasma Reflection

Spitzer and Fan[61] have investigated the optical properties of a number of semiconductors by studying the reflection of infra-red radiation at frequencies above and below the plasma frequency of the free carriers. The experiments permitted them to determine the effective masses of these carriers. This had been particularly successful for electrons in indium antimonide. However, in order to obtain the mass values, it was necessary to determine the electron concentration by Hall measurements and the dielectric constant by reflection coefficient at short wavelengths or with low carrier concentration. The accuracy of the results depends on the latter measurements rather than on the optical wavelength determination. The experiments can be made more versatile and independent of electrical measurements by the use of a magnetic field. The mass can be measured directly, and the carrier density calculated from the optical data. The fundamental phenomenon is based upon the relationship between the index of refraction and the parameters of the electrons or holes as given by

$$\eta_{\parallel}^2(H) = \eta^2(0) = \kappa\left[1-\frac{\omega_p^2}{\omega^2}\right] \qquad \ldots (34a)$$

$$\eta_{\perp}^2(H) = \kappa\left[1-\frac{(\omega^2-\omega_p^2)\omega_p^2}{(\omega^2-\omega_p^2-\omega_c^{*2})\omega^2}\right] \qquad \ldots (34b)$$

$$\eta_{\pm}^2(H) = \kappa\left[1-\frac{\omega_p^2}{\omega(\omega\mp\omega_c^*)}\right] \qquad \ldots (34c)$$

where $\omega_p^2 = ne^2/m^*\kappa$, $\eta_{\parallel}(H) = \eta(0)$ is the index of refraction for linearly polarized radiation propagating transverse to the magnetic field with the electric vector parallel to the magnetic field. For an isotropic carrier, $\eta_{\parallel}(H)$ is identical to the index at zero field. $\eta_{\perp}(H)$ is the index for propagation in a direction transverse to the magnetic field, but with the electric vector linearly polarized perpendicular to H. $\eta_{\pm}(H)$ is the index for right- and left-hand circularly polarized radiation propagating in the direction of the applied field. For a semi-infinite semiconducting slab, the reflection coefficient R for perpendicular incidence is given by

$$R = \frac{(\eta-1)^2}{(\eta+1)^2} \qquad \ldots (35)$$

The expressions in equations (34) neglect losses, since it can be shown that at these wavelengths $\omega\tau \gg 1$ and the medium is essentially dispersive. In this case, the refractive index at short wavelengths ($\omega^2 \gg \omega_p^2$) is essentially independent of the plasma, so that $\eta^2 = \kappa$, resulting in a reflection coefficient $R \simeq 36$ per cent for indium antimonide, but as the wavelength increases, $R \to 0$ as $\omega^2 \to \omega_p^2(1-\kappa^{-1})^{-1}$ and $\eta \to 1$. For a relatively small further increase in wavelength, R goes rapidly to unity when $\omega = \omega_p$ and $\eta = 0$, corresponding to the cut-off condition for propagation in the medium. Thus from the wavelength of zero (or minimum) reflectivity, it is possible to estimate m^* from equation (34*a*) if the electron density and κ are known. However, when the magnetic field is applied, the critical conditions

for which $R = 0$ and $R = 1$ are altered. The critical conditions corresponding to cut-off with $R = 1$ are the same for both the transverse propagation of equation (34*b*) and longitudinal propagation of equation (34*c*), namely

$$\omega \simeq \omega_p \pm \frac{\omega_c^*}{2} + \frac{\omega_c^{*2}}{8\omega_p} \qquad \ldots (36)$$

where $\omega_p \gg \omega_c^*$. Similarly, for the condition of zero reflection with a magnetic field, one obtains expressions with a slight correction

$$\omega_\pm \simeq \omega_p \left(1+\frac{\delta}{2}\right) \pm \frac{\omega_c^*}{2} + \tfrac{1}{8}\frac{\omega_c^{*2}}{\omega_p}\left(1-\frac{\delta}{2}\right) \qquad \ldots (37a)$$

$$\omega_\perp \simeq \omega_p \left(1+\frac{\delta}{4}\right) \pm \frac{\omega_c^*}{2} + \frac{\omega_c^{*2}}{4\omega_p}\left(1-\frac{\delta}{4}\right) \qquad \ldots (37b)$$

where $\delta = 1/\kappa$. The important distinction between the zero field and the magneto-plasma reflection is that where there was one minimum for the former, there are

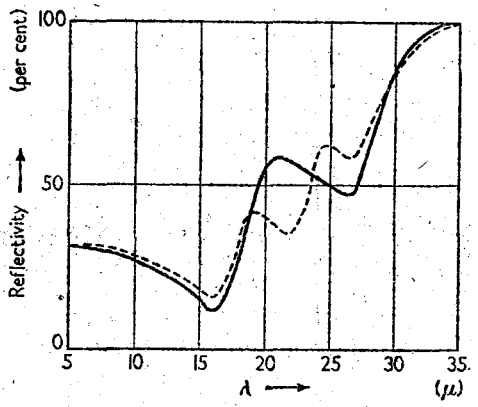


Figure 26. *Sketch of theoretical magneto-plasma reflectivity of indium antimonide with* $\sim 10^{18}$ *electrons* cm^{-3} *and a field of* 80 *kgauss. The solid line represents the behaviour of the reflectivity for unpolarized radiation in the case of longitudinal propagation. The dashed line represents the reflectivity for unpolarized radiation propagating transverse to the field. For the case of transverse propagation and linear polarization perpendicular to the field (not shown), the 'central' component of the dashed curve is absent and the reflectivity minima approach zero. The separation between the minima of the solid curve corresponds to the cyclotron frequency of the electron,* ω_{c_2}

now two for the latter, which are separated by ω_c^*, the cyclotron frequency. This then permits a direct measurement of the effective mass for linear polarized radiation when the small correction of a few per cent for the dielectric constant is made. The curves for such an experiment would have the approximate appearance shown in Figure 26. The solid lines show the results expected for linear polarization for

either transverse or longitudinal propagation. For unpolarized radiation, the latter condition would still give two minima, since the electric vector would always be perpendicular to the magnetic field and could be resolved into two equal circularly polarized components of opposite ense of rotation. For transverse propagation, unpolarized radiation can be resolved into two linear components parallel and perpendicular to the field. The latter will produce the splitting into two minima, whereas the former will retain the 'central' minimum as shown in Figure 26 by the dotted curve. Consequently, the magnetic data permit the effective mass to be obtained from ω_c^*, and by using the value of κ obtained from the value of R at short wavelengths, the electron density can be readily calculated from the reflection minimum at zero field.†

6.2. Rotary Dispersion of a Magneto-plasma (Kerr Magneto-optic Effect)

The magneto-plasma effects which are observable by simple reflection can also be studied by the rotation of linearly polarized radiation which propagates parallel to the magnetic field and is reflected from the surface of a semiconductor near the plasma frequency. The phenomenon, which has been encountered in other solids, is known as the *Kerr magneto-optic effect*. In order to observe it in the usual materials, it is necessary that there exist in the medium a dissipative mechanism in addition to dispersive properties, so that the wave damping factor $\alpha \neq 0$. In the case of the magneto-plasma, the electric vectors of the two counter-rotating components of the linearly polarized incident wave are rotated upon reflection at the interface through angles given by

$$\tan\theta_{\pm} = \frac{2\alpha_{\pm}\beta_0}{\beta_0^2-\beta_{\pm}^2-\alpha_{\pm}^2} \simeq \frac{2\beta_0}{\alpha_{\pm}} \qquad \ldots (38)$$

The approximation holds beyond 'cut-off' when $\alpha_{\pm} > \beta_0$ in a purely dispersive medium. The actual rotation of the plane of polarization (or the major axis of the resultant elliptically polarized reflected wave) is obtained from the differential effect, $\theta = (\theta_+ - \theta_-)/2$. In the region beyond 'cut-off' of either or both components, there is a Kerr rotation even though the medium is not dissipative, but only dispersive. The medium is only of interest when the 'cut-off' frequency is reached for each circularly polarized component, i.e. when the phase constants β_+ and β_- vanish, respectively, and only the reactive damping factors α_+ and α_- exist. The medium is perfectly reflecting for each component after 'cut-off' since no loss occurs. Since the penetration depth is different for the two counter-rotating electric vectors, they emerge after reflection with $\theta_+ \neq \theta_-$, giving rise to the rotatory dispersion by the magneto-plasma. The appropriate expressions for the damping factors above their 'cut-off' frequencies are

$$\alpha_{\pm}^2 = \beta_0^2\kappa\left[\frac{\omega_p^2}{\omega(\omega \mp \omega_c)} - 1\right] \qquad \ldots (39)$$

where $\beta_0 = \omega/c$, $\omega_p^2 = ne^2/m^*\kappa$. As in the case of magneto-plasma reflection, the critical frequencies at 'cut-off' are given by the expressions of equation (36), for which $\alpha_+ = \beta_+ = 0$ and also $\alpha_- = \beta_- = 0$, as is evident from equation (39). As the

† See section added in proof, page 268.

frequency decreases and reaches 'cut-off' for the positive circular component at

$$\omega_+ \simeq \omega_p + \frac{\omega_c^*}{2} + \frac{\omega_c^{*2}}{8\omega_p}$$

θ_+ rises sharply from zero to 90° as $\alpha_+ \to \beta_0$. The latter occurs at

$$\omega \simeq \omega_p\left(1-\frac{\delta}{2}\right) + \frac{\omega_c^*}{2} + \frac{\omega_c^{*2}}{8\omega_p}\left(1+\frac{\delta}{2}\right)$$

Thereafter, the total rotation θ continues to increase from 45° more slowly until 'cut-off' is reached for the negative circular component at

$$\omega_- \simeq \omega_p - \frac{\omega_c^{*2}}{2} + \frac{\omega_c^{*2}}{8\omega_p}$$

whereupon θ_- rises sharply from zero and reaches 90° as $\alpha_- \to \beta_0$. At $\omega = \omega_-$, θ decreases sharply, then more slowly after

$$\omega \simeq \omega_p\left(1-\frac{\delta}{2}\right) - \frac{\omega_c^*}{2} + \frac{\omega_c^{*2}}{8\omega_p}\left(1+\frac{\delta}{2}\right)$$

As ω continues to decrease from these values where $\omega \gg \omega_c^*$, α_- continues to increase monotonically, so that $\theta_- \to 180°$, but as $\omega \simeq \omega_c^*$, α_+ suddenly increases strongly so that $\theta_+ \simeq 180°$. Just beyond $\omega = \omega_c^*$, there is a sign reversal and $\alpha_+ = 0$. The medium thereafter becomes propagating for the positive circular component ($\beta_+ \neq 0$), accompanied by an increasing partial reflection. At the resonance frequency $\omega = \omega_c^*$, the Kerr rotation θ suddenly increases from nearly zero to nearly 90°, and continues to approach this value asymptotically.

For $\omega \gg \omega_c^*$, the frequency separation between $\theta = 0$ and θ_{max} is $\Delta\omega \simeq \omega_c^*$. The angle θ_{max} occurs just prior to the sharp drop at 'cut-off' for the negative circular component. Thus it is possible to determine the effective mass of the carrier in a manner analogous to that for magneto-plasma reflection. The experiment will also permit additional data to be obtained from the ellipticity of the emerging wave, since between $\theta = 0$ and θ_{max} the positive component is nearly fully reflected and the negative component is partially reflected. Thus, the reflection also provides information about the relative amplitudes of these components.

7. EXPERIMENTAL TECHNIQUES

7.1. Magneto-absorption

The observation of the magneto-absorption phenomena described above involves a variety of experimental techniques of spectroscopy and cryogenics which merit description in moderate detail. The magnetospectrophotometric system developed by Zwerdling[11] and co-workers for experimental measurements on the oscillatory effects and the Zeeman effect of excitons and impurity levels consisted of a double-pass infra-red spectrophotometer and a 12 in. electromagnet which supported a liquid helium optical cryostat. The magnet was equipped with specially designed pole pieces of Hiperco alloy tapered down to a pole face

diameter of $\frac{7}{8}$ in. and having a $\frac{5}{16}$ in. air gap. At the maximum rated magnet power of 4 kW, this system provided a magnetic field of 38·9 kgauss. The spectrophotometer employed a rapid interchange multiple source unit containing tungsten filament and globar sources and interchangeable monochromator units utilizing prism or grating dispersion covering the spectral region from 0·2 to 50 microns. The monochromatic radiation from the exit slit transmitted by the sample which was inside the cryostat and between the pole faces was measured by an infra-red detector suitable for the wavelength region being scanned at various constant magnetic field intensities. The initial experiments on the oscillatory absorption in germanium were performed using dense flint prism dispersion with a chromatic resolving power ~ 1,000 corresponding to a spectral slit width equivalent to 10^{-3} eV. This resolution, although adequate at room temperature, was insufficient to observe the fine structure at low temperatures where the lines become narrower. Consequently, it was necessary to utilize diffraction grating dispersion which increased the resolving power to values between 10,000 and 20,000, or an equivalent spectral width ~ 10^{-4} eV. Typical low temperature results for germanium obtained with the diffraction grating are shown in Figure 6. In order to conduct these experiments at temperatures down to that of liquid helium, it was necessary to construct a cryostat capable of maintaining the sample at low temperature in the air gap of the magnet while allowing the radiation to pass through it from the monochromator to the detector. An all-metal cryostat was built containing a liquid helium compartment surrounded by a liquid nitrogen compartment for radiation shielding, both inside an outer wall having two polished salt windows transparent in the wavelength range of interest. The sample was mounted in a copper holder attached to the helium compartment and cooled by thermal conduction. It was possible to lower the temperature of the liquid helium bath to 1·2° K by high speed mechanical pumping. For the direct transition experiments, the samples were prepared by cementing flat slabs of semiconducting material to suitable transparent substrates and grinding and polishing to final thicknesses from 4 to 15 microns. For the indirect transitions, samples with thicknesses ranging from 1 to 6 mm were prepared having optically flat polished faces. Linearly polarized radiation was produced either by means of infra-red transmitting type HR Polaroid in the wavelength region 1 to 2 microns, or at longer wavelengths by means of a silver chloride sheet pile polarizer set to the Brewster angle. By rotating these polarizers, the electric vector was aligned either parallel or perpendicular to the magnetic field vector. For the experiments with grating dispersion at wavelengths between 1·1 microns and 2 microns, the higher orders of the grating were removed by an anti-reflection coated silicon transmission filter.

The magneto-absorption experiments with germanium at the Naval Research Laboratory were conducted using a Bitter-type air core solenoid capable of achieving magnetic field intensities up to 60 kgauss. The optical configuration made it possible to propagate the radiation parallel to the field so that experiments with right- and left-hand circularly polarized radiation were possible. The use of circularly polarized radiation has the feature of separating the spectra for transitions with spin selection rules $\Delta M = +1$ and $\Delta M = -1$, whereas with linear polarization, it is only possible to separate spectra with $\Delta M = 0$ and $\Delta M = \pm 1$. The circularly polarized radiation was generated from linearly polarized radiation by the use of a Fresnel rhomb made of sodium chloride.

7.2. Zeeman Effect

The experiments on the Zeeman effect of impurity levels in germanium and silicon have been done, respectively, in two different spectral regions, one in the far infra-red from 50 to 200 microns, and the other from 15 to 22 microns. Spectrophotometric techniques in the far infra-red require the use of diffraction grating dispersion, an enclosure for the optical path from source to detector which is either evacuated or flushed with dry nitrogen to eliminate water vapour absorption, liquid helium sample temperatures, intense gas discharge sources such as the mercury arc lamp, and a very sensitive radiation detector. It is still difficult to generate adequate radiation intensity in this region, and the best practical sources above 100 microns are gas discharge tubes which yield a higher output than the more familiar black body radiators. Although absorption by atmospheric water vapour is a problem requiring the solution indicated, its known absorption lines in this region provide a convenient means of spectrometer calibration. At these wavelengths, the Golay pneumatic infra-red detector may be used successfully, although a helium temperature carbon resistor bolometer mounted just behind the sample has been employed by Boyle[38] and others.[68] The latter arrangement has the advantage that the sample is exposed to room temperature background radiation from only one side instead of two, and furthermore, that the effective noise temperature of the detecting element is very low. Most far infra-red spectrometers have a high radiation gathering power, or a low f number, requiring the use of large aperture optical components. As with all diffraction grating spectrometers, it is necessary to isolate the desired spectral order. This is usually done for wavelengths $\lambda > 50$ microns with an appropriate combination of transmission filters (such as fused silica), reststrahlen reflection filters of various salts, and specular reflection of the desired wavelengths from special gratings or roughened mirrors which scatter higher orders. Scattered radiation of the much more intense near infra-red, visible, and ultra-violet wavelength regions are usually removed by transmission through sooted quartz or special black polyethylene.

Since the energy required for transitions from the ground state of the usual monovalent impurities in germanium to the excited states or the adjacent band is very small, of the order of 10^{-2} eV, the depopulation of the ground state and line broadening effects by thermal excitation can be effectively avoided only by maintaining the sample at liquid helium temperature (at 4° K, $kT \simeq 3 \times 10^{-4}$ eV). Furthermore, undesired photon-induced excitations due to absorption of undispersed radiation from the $\simeq 300°$ K background with an emission maximum close to 9 microns must also be minimized. The latter problem may be solved by using helium temperature windows having suitable transmission properties, mounted internally in the optical path at the sample. For example, fused silica may be used, since it is opaque from $\simeq 4$ to 50 microns, but becomes transparent at longer wavelengths.

These problems are much less severe for the study of monovalent impurities in silicon, since the ground states are considerably farther removed from the excited states or the adjacent band extremum than for germanium. For example, the donor, bismuth, and the acceptor, aluminium, studied by Zwerdling, Button, and Lax[41] have ionization energies close to 7×10^{-2} eV, so that the Zeeman effect of their excitation spectra were measured in the potassium bromide prism dispersion region 15–22 microns. For this region, standard infra-red components were used,

including a double pass potassium bromide prism monochromator, a globar source, thermocouple detector, and front surface reflecting optics in an $f/4$ system. The samples mounted in the helium optical cryostat between the electromagnet pole pieces had thicknesses ranging from 1 to 6 mm, depending upon the impurity concentration and the absorption intensity of the transition studied. The optimum impurity density range was found to be 10^{15}–10^{16} cm^{-3}, being neither too small to require excessive thicknesses nor too large, which may result in impurity banding and line broadening effects.[59] The use of host silicon material essentially free of impurities other than the one being studied (particularly oxygen) was essential for successful spectral measurements. Since polarized radiation for correlating the Zeeman effect of the excitation spectrum with the theory required a Brewster angle reflection polarizer, the available source intensity was reduced by somewhat more than half. Consequently, attention to the details of sample selection and preparation, optical alignment, and the use of a special high sensitivity thermocouple detector were all critically required in order to avoid using excessive spectrometer slit widths, accompanied by the loss of spectral resolution and detail. In the experiments of ZBL,[41] the mean calculated resolving power of $3 \cdot 5 \times 10^{-4}$ eV for the measured polarized spectra was comparable to the line widths found, so that the inherent spectral line widths for the samples used were not thus determined. Nevertheless, the photon energies of the various absorption maxima could be determined to an uncertainty which is an order of magnitude smaller than this resolving power.

The magneto-absorption experiments of Gross and co-workers,[39] including the Zeeman effect of excitons and transitions between Landau levels in cuprous oxide, were performed from room temperature down to 1·3° K in magnetic fields up to $\simeq$ 30 kgauss. The measurements were made photographically with prism and diffraction grating spectrophotometers having linear dispersion as large as 1·5 Å mm^{-1}. For cuprous oxide, the spectra were obtained by transmission through crystal plates of thicknesses ranging from 1·7 mm down to about 20 microns, the latter prepared by a careful polishing procedure. Linearly polarized radiation was employed for measuring the Zeeman effect of the exciton lines, yielding the σ spectrum ($\mathbf{E} \perp \mathbf{H}$) and the π spectrum ($\mathbf{E} \| \mathbf{H}$). Lowering the sample temperature to 1·3° K resulted in appreciable line narrowing, not only for the first component of the exciton series, but also for the higher components. Consequently, the existence of the quadratic Zeeman effect was observable at a field of 29,000 oersteds for the transitions with higher quantum numbers, as well as absorption maxima at energies greater than the exciton series limit.

7.3. Faraday Effect

The experimental techniques required for studying the Faraday effect in semiconductors at infra-red frequencies are in many ways similar to those for magneto-absorption. The principal difference is that one measures the intensity of plane polarized radiation transmitted by the sample, not to determine absorption, but to determine the angle through which the polarization plane has been rotated by the sample in the magnetic field. In addition, it is necessary that the direction of radiation propagation be parallel to the magnetic field, so that the differential dispersion of the right- and left-hand circularly polarized components will produce the desired rotation of the plane polarized wave. The usual procedure is to obtain

monochromatic radiation of the desired wavelength from an infra-red monochromator having the appropriate prism or grating dispersing element. After passing this radiation through a linear polarizer, it is brought to a focus on the sample. The field induced rotation is determined by compensation with a measured equal rotation of a polarizing analyser mounted between the sample and the detector. In the experiments of Moss, Smith, and Taylor[57] on indium antimonide, transmission polarizers made of thin sheets of polyethylene set at the Brewster angle were used. The rotations by the sample were determined in the wavelength region 10–30 microns using magnetic fields up to 4 kgauss. Brown and Lax[58] also studied indium antimonide at room temperature, but used reflection polarizers consisting of germanium mirrors at the Brewster angle arranged so that there was no beam deviation as the unit was rotated. For particular wavelengths and magnetic field values in the potassium bromide and sodium chloride prism regions, the radiation intensity transmitted by the polarizer, sample, and analyser was recorded continuously during the rotation of the analyser from an initial orientation with the polarizer fixed. The rotation of the plane of polarization produced by the field was then determined from shift of the $\cos^2\theta$ curve along the θ co-ordinate, where θ is the angle between the polarization planes of polarizer and analyser. Magnetic fields up to 15 kgauss were obtained by using a 4 in. calibre Bitter-type air core solenoid. With low temperature cryostats, the Faraday rotation experiments provide more detailed and precise data since the carriers in a degenerate semiconductor reflect the properties at the energy defined by the Fermi surface. The low temperature Faraday rotation experiments of Moss and coworkers indicate the advantages which can be thus achieved.

8. SUMMARY AND DISCUSSION

The main purpose of the magneto-absorption experiments described in this paper has been to obtain quantitative information about the electronic band structure of semiconductors. The earlier cyclotron resonance experiments in germanium and silicon done at microwave frequencies[1,2] were eminently successful in elucidating the band properties of these semiconductors at the extrema of the valence of conduction bands. However, even in these materials, no information was obtained about additional extrema occurring farther into these bands. Furthermore, the microwave results gave quantitative information only to the extent of $\sim 10^{-3}$ eV beyond the respective extrema and did not probe deeper into these bands. Still another weakness of the microwave method was that cyclotron resonance could not succeed in semiconductors of even moderate impurity concentrations because the necessary condition that $\omega\tau > 1$ could not then be fulfilled, since τ would be too small at microwave frequencies. Magnetospectroscopy in semiconductors, which includes cyclotron resonance at infra-red frequencies and the other magneto-optical phenomena previously discussed, has overcome many of these difficulties. The first achievement was the measurement of the effective mass of the electron at the $k = 0$ conduction band minimum in germanium, as well as an accurate determination of the energy gap associated with the direct transition.† The so-called 'quantum effects' of the magnetic levels of the valence band predicted by Luttinger and Kohn[9] were experimentally verified. The anomalous g

† See section added in proof, page 268.

factors of the electrons due to spin–orbit coupling in indium antimonide and in germanium at $k = 0$ were discovered and quantitatively determined. The effective mass of the electron in gallium antimonide was measured for the first time with fair precision, the energy gap at liquid helium temperature determined, and the associated transition shown to be direct.

Perhaps the most dramatic successes of the magneto-absorption measurements were connected with the direct and indirect transitions in germanium. Firstly, the energy gaps at liquid helium temperature were measured with great accuracy to three significant figures, or about 0·1 per cent. Secondly, the direct transition exciton in germanium was discovered for the first time and a technique was developed which permitted a definitive determination of the binding energy of both the direct† and indirect transition excitons. The study of the latter was particularly fruitful, since the fine structure was observed as a function of magnetic field, and quantitative parameters of a spin Hamiltonian which describes the exciton spectrum were determined and the g factor of the electron in germanium with its anisotropy was finally uncovered.

The Zeeman effect of the excited states of both donor and acceptor impurities has now been observed in both germanium and silicon, but has not yet produced new fundamental information.† This work has served primarily to verify the theory developed, based on the model of the bands established by microwave cyclotron resonance. In bismuth-doped silicon, however, the Zeeman effect experiments of Zwerdling, Button, and Lax,[41] which were performed in great detail using polarized radiation and with the magnetic field alternately along the three principal crystallographic directions, have served to identify some of the absorption maxima of low intensity, previously observed by Hrostowski and Kaiser[48] at zero field, as due to transitions from the ground state to specific higher quantum number states. The new phenomenon which has evolved with the Zeeman effect is the oscillatory magneto-absorption of impurities,[52] corresponding to transitions from the ground state of the impurity to the bound Landau states in the valence or conduction band. Making use of such observations, Boyle[38] has been able to obtain an experimental value for the ionization energy of arsenic impurity in germanium, using a procedure analogous to that developed for determining the energy gap in oscillatory magneto-absorption experiments with germanium.[5]

Another capability of the type of experiments treated in this paper is the measurement of the curvature of bands as a function of energy. The decreasing curvature of the conduction band in indium antimonide was initially demonstrated by the infra-red cyclotron resonance experiments using pulsed magnetic fields.[60] This was subsequently verified by the experiments of Spitzer and Fan using free carrier absorption.[61] A similar determination of the conduction band curvature in indium antimonide was provided by the Faraday rotation experiments of Moss, Smith, and Taylor.[57] It is probable that the probing of the curvature of bands is most directly and accurately achieved by the oscillatory magneto-absorption. In germanium, changes of a few per cent in the curvature of the $k = 0$ conduction band minimum were determined with high accuracy, as shown by equation (5). Similarly, small changes were also observed by means of the impurity oscillatory effect as demonstrated by Figure 22.

Many of the phenomena which have been observed in germanium and silicon

† See section added in proof, page 268.

will probably be extended to other semiconducting materials when higher magnetic fields of the order of 100 kgauss become available, particularly when combined with very low temperatures and high resolution spectrometers. For instance, the oscillatory magneto-absorption of indium antimonide which was investigated at room temperature should provide new quantitative information, particularly about the valence band structure under more favourable conditions of field and temperature. Similarly, fine structure in gallium antimonide, which was evident but not resolved with the present apparatus, should become well defined at higher fields. Indium antimonide, gallium arsenide, gallium phosphide, and other compound semiconductors will yield to quantitative measurements from which the effective masses of holes and electrons will be deduced. The magneto-absorption of the forbidden transition between the valence bands in germanium, which has not yet been observed, should be detected at higher fields. The indirect transition magneto-absorption in silicon which was barely perceptible at 39 kgauss will certainly yield fruitful results at higher fields. This can be best appreciated by looking at equation (8), which shows that for this transition, the absorption coefficient increases as the square of the magnetic field.

In addition to the extension of already observed effects to new materials, new phenomena can be studied with the magneto-spectrometers now available. Although the experiments discussed have been studied by transmission techniques, they can be also observed by reflection. These may be more difficult, however, but where the absorption coefficient is high and the preparation of very thin samples impractical, they may be necessary. The magneto-plasma effects which are best studied by reflection experiments, should be observable and most fruitful at high magnetic fields. These, together with Faraday rotation experiments, should provide information about the band structure where the quadratic effects at high fields give rise to anisotropy. The high fields, particularly in larger volumes, will be helpful in extending the investigations using recombination radiation techniques in semiconductors. The experiments of Haynes[17] in germanium and silicon, involving the indirect transition, produce strong emission lines whereas the corresponding transitions in absorption produce 'edges'. Consequently, at liquid helium temperatures and magnetic fields of 60 kgauss or more, the emission lines should exhibit resolvable structure. If the Zeeman effect of excitons can then be observed in emission, greater resolution of the fine structure should permit the determination of the tensor components of the electron g factor from the anisotropy of the effect.

Extension of the Zeeman experiments to shallow impurities in silicon would be desirable. In silicon, the Zeeman effects have thus far been observed only with the relatively deeper levels of aluminium and bismuth impurities. With shallower impurities, such as boron for an acceptor and antimony for a donor, it should be possible to obtain more intense absorption with the same concentration, or narrower lines with lower concentration than before. This is so because the ground state wave function is less localized and overlaps those of the excited states more, thus increasing the matrix element and hence the transition probability. For the same reason, the intensity of the oscillatory magneto-absorption of the impurities will also increase and permit quantitative measurements of the ionization energies to be made. These measurements will have to be performed with a caesium bromide prism in the wavelength region of 25 to 36 microns.

Another experiment involving the oscillations of the absorption, analogous to the de Haas–van Alphen effect in metals, can be made in semiconductors. These oscillations arise from the quantization of the magnetic levels in a degenerate semiconductor. The free carrier absorption at moderate wavelengths, e.g. in the potassium bromide (10–20 microns) region, could exhibit this effect at low temperature. The oscillations would be analogous to those found for indium antimonide and indium arsenide by Frederickse[70] and also Sladek[71] in their magnetoresistance and Hall effect measurements. For semiconductors, the de Haas–van Alphen type oscillations of free carriers will occur for simple bands when the scattering is energy dependent. The theory is essentially that developed by Argyres[72] for the magnetoresistance in high fields, except that instead of a d.c. electric field, the expressions for the conductivity should contain the frequency of the electromagnetic field. Actually, such a phenomenon has already been observed in bismuth by Boyle[73] in which oscillations as a function of magnetic field at $\simeq 20$ microns and $4{\cdot}2^\circ$ K were observed in thin, 50 micron, samples.

These are some of the accomplishments to date in a field of research which is barely three years old. This is an impressive beginning which portends a bright future for magnetospectroscopy, not only in semiconductors, but in other solids as well. The combination of high resolution spectroscopy, low temperatures, and high magnetic fields should provide a tool for fundamental research which will be as revolutionary as the microwave techniques which have been exploited for solids during the last dozen years.

ADDED IN PROOF

Since this paper was written, a number of important developments have occurred which are closely related to the topics discussed. In particular, further work on the Zeeman effect of excitons in cadmium sulphide by Wheeler and Dimmock[74] and also by Hopfield and Thomas[75] has produced new results on the band structure of this compound and on the g values from the complex fine structure of the Zeeman effect. The latter authors have also described a new magneto-optical phenomenon associated with the momentum of the photon in which they observed marked changes in the absorption spectrum of cadmium sulphide with reversal of the magnetic field direction. In their experiments, the radiation was linearly polarized parallel to the c axis and normally incident on the sample, and the magnetic field was perpendicular to the c axis and the direction of propagation.

New developments dealing with the excitons in germanium have been twofold. First, a proposal for observing a third exciton formed by a direct transition from the $L_{3'}$ or [111] valence band to an energy just below the [111] conduction band was advanced by Lax.[76] The binding energy of $\simeq 0{\cdot}0048$ eV was estimated for this exciton, and the possibility of determining the effective masses and the components of the g factor of the $L_{3'}$ band from the Zeeman pattern and the magneto-reflection spectrum was described. The second development involves the interpretation of the exciton binding energy at $k = 0$. According to the experimental results of Edwards and Lazazzera[77] the linear portion of a plot of the energy of absorption maxima versus field obtained from the magneto-absorption spectrum extrapolates to a convergence point at $\mathbf{B} = 0$ which falls below the energy of the zero field exciton line in an unstrained sample. These authors

interpret this result as evidence for direct transitions to the higher bound exciton levels which are magnetically shifted, rather than to free conduction band Landau levels, and invoke the theoretical work of Elliott and Loudon[27] and Howard and Hasegawa.[78] The latter suggest that the magneto-absorption spectrum of transitions from an *impurity* ground state in germanium as observed by Boyle[38] are transitions to *bound* excited states rather than free Landau states. The implication of these theories is that the experimental values of the binding energy of the donor impurity in germanium, as well as the direct transition energy gap obtained from a convergence plot, must be corrected in an appropriate manner. These considerations apparently do not apply to the indirect exciton, since the phonon-assisted transitions to higher states in the conduction band have no selection rules on the quantum number n. This conclusion is indicated by the data of Figure 10, in which the Landau transitions converge well above the exciton line in an unstrained sample.

In studying the Zeeman effect of P-type impurities in silicon, Zwerdling, Button, Lax, and Roth[76] discovered the existence of impurity levels within the continuum of the $p_{3/2}$ valence bands, just above the split-off $p_{1/2}$ band. The spin–orbit splitting of the valence bands in silicon was deduced from the spectrum to be $\Delta = 0{\cdot}0442 \pm 0{\cdot}0004$ eV and the hole effective mass of the $p_{1/2}$ band to be $0{\cdot}25\ m_0$. The Zeeman spectrum of these 'internal' levels, although complex, is consistent with this mass value. In addition, fine structure due to spin effects should provide quantitative information about g values. Another interesting aspect was that transitions to quantized magnetic levels were observed extending to energies well above the series limit of the internal levels as determined by fitting the zero field data to a modified Rydberg formula. The energy of these transitions showed linear behaviour at high fields, and the straight lines, extrapolated to zero field, coalesced to an energy well below the series limit. Since it was also possible to follow these transitions to comparatively lower fields where they showed quadratic behaviour, and appeared to coalesce *just below* the series limit, these observations are in accord with the theory of Howard and Hasegawa.[78]

The magneto-plasma reflection effect treated theoretically in this paper has now been observed experimentally and is described by Lax and Wright[80] for N-type indium antimonide and mercury selenide with donor concentrations $\sim 10^{18}$ cm^{-3}. The effective masses were found to be $0{\cdot}041m_0$ and $0{\cdot}045m_0$, respectively. A dielectric constant of $\kappa = 14{\cdot}9$ was deduced for mercury selenide from the optical measurements. The experimental points were found to fit theoretical curves when losses due to scattering were included.

New results on the Faraday rotation in indium antimonide have been reported by Smith, Moss, and Taylor.[81] The authors analysed in greater detail the energy dependence of the electron mass in this material. In addition, Moss and Walton[82] have determined the effective mass of the electrons in gallium arsenide and indium phosphide with carrier concentrations of the order of 5×10^{16} cm^{-3} and obtained effective mass values of the order of $0{\cdot}072m_0$ and $0{\cdot}073m_0$, respectively, by the use of the infra-red Faraday effect. On the theoretical side, a quantum-mechanical analysis of Faraday rotation at photon energies just below the gap has been made by Lax. The classical analysis in the text indicated that Faraday rotation can result from a bound electron in the presence of a magnetic field, but quantitative agreement with experimental results was not satisfactory. Quantum mechanically,

however, the rotation can be analysed by considering the existence of virtual transitions at photon energies below that of the gap, which give rise to dispersion. At zero magnetic field, the phase constant is given approximately by

$$\beta_0 + A\sqrt{(\epsilon_g - \hbar\omega)}$$

where A is the constant defined in equation (2) and $\beta_0 = \omega\epsilon\mu_0$ is the phase constant associated with the dielectric properties of the medium. When the magnetic field is introduced, one obtains the result

$$\beta_\pm \simeq \beta_0 + \frac{A}{2}\sum_n \frac{\hbar\omega_c}{\sqrt{(\epsilon_n - \hbar\omega \pm \frac{1}{2}g\beta H)}}$$

where $\epsilon_n = \epsilon_g + (n+\frac{1}{2})\hbar\omega_c$ is the energy associated with the magnetic quantum levels of the electron and the heavy hole bands, ω_c is the cyclotron angular frequency for the reduced effective mass of the electron and the heavy hole, and the selection rule $\Delta n = 0$ applies. The radical contains the spin g factor of the electron at the bottom of the band, and β is the Bohr magneton. The g factor of heavy hole is comparatively small compared to that of the electron. The Faraday rotation due to this effect alone is then

$$\theta = \frac{A\hbar\omega_c}{8} g\beta H \sum_n \left(\epsilon_n - \hbar\omega\right)^{-3/2}$$

This last result shows the proper singularity near the energy gap and gives rise to an increasing component of Faraday rotation as the photon energy approaches the gap value. The important consequence of this quantum-mechanical treatment is that this rotation component is due primarily to electron spin, in the presence of spin–orbit coupling, and to virtual transitions between the heavy hole and the electron. The virtual transitions between the light hole and the electron do not contribute, since the spins in the two bands are reversed and no Faraday rotation results to a first approximation. The other interesting consequence is that if the spin is ignored, there is no Faraday rotation due to the diamagnetic terms alone, contrary to the result predicted by the classical theory. Lastly, the experiment indicates that the interband Faraday rotation is opposite to that of the free electrons, which provides confirmation that the sign of the g factor for the electron in indium antimonide is negative, as predicted theoretically by equation (6).

ACKNOWLEDGEMENTS

We should like to acknowledge the contributions of our colleagues and co-workers at the Lincoln Laboratory who have contributed directly and indirectly to much of the research that has been described in this article. We wish to thank Laura M. Roth for several valuable discussions and suggestions concerning the theoretical aspects of the magneto-absorption work presented in this manuscript. We are grateful to Kenneth J. Button for important contributions on the Zeeman effect of excitons and impurity levels which are to be published. We have drawn upon some of the unpublished theoretical work on the impurity Zeeman effect developed by Walter H. Kleiner and Richard N. Brown. The experimental data in Figure 25

originate from the unpublished thesis of R. N. Brown. The theoretical result for the Faraday rotation effect in P-type semiconductors was obtained by John G. Mavroides. We have benefited greatly from discussions with G. B. Wright concerning magneto-plasma phenomena in semiconductors. We also owe a debt of gratitude to J. P. Theriault for his capable assistance throughout the course of the experimental work of the Lincoln group herein described, both published and unpublished. Finally, we should like to thank Mrs. P. M. Dougherty for her help during the preparation of the manuscript.

REFERENCES

1. G. Dresselhaus, A. F. Kip, and C. Kittel. *Phys. Rev.* **98**, 368 (1955); R. N. Dexter, H. J. Zeiger, and B. Lax. *Phys. Rev.* **104**, 637 (1956)
2. For reviews of cyclotron resonance, see B. Lax. *Rev. mod. Phys.* **30**, 122 (1958); also B. Lax and J. Mavroides. *Solid State Physics* (Academic, New York) (to be published)
3. L. H. Hall, J. Bardeen, and F. J. Blatt. *Phys. Rev.* **95**, 146, 559 (1954); *Proc. of Conf. on Photoconductivity, Atlantic City* 1956 (Wiley, New York)
4. R. J. Elliott, T. P. McLean, and G. G. Macfarlane. *Proc. phys. Soc. Lond.* **72**, 553 (1958); B. Lax, L. M. Roth, and S. Zwerdling. *J. Phys. Chem. Solids* **8**, 311 (1959); E. Burstein, G. S. Picus, R. F. Wallis, and F. J. Blatt. *J. Phys. Chem. Solids* **8**, 305 (1959); *Phys. Rev.* **113**, 15 (1959)
5. S. Zwerdling and B. Lax. *Phys. Rev.* **106**, 51 (1957); E. Burstein and G. S. Picus. *Phys. Rev.* **105**, 1123 (1957)
6. E. Burstein, G. S. Picus, H. A. Gebbie, and F. J. Blatt. *Phys. Rev.* **103**, 826 (1956); S. Zwerdling, R. J. Keyes, S. Foner, H. H. Kolm, and B. Lax. *Phys. Rev.* **104**, 1805 (1956)
7. S. Zwerdling, B. Lax, and L. M. Roth. *Phys. Rev.* **108**, 1402 (1957)
8. W. P. Dumke. *Phys. Rev.* **105**, 139 (1957)
9. J. M. Luttinger and W. Kohn. *Phys. Rev.* **97**, 869 (1955); J. M. Luttinger. *Phys. Rev.* **102**, 1030 (1956)
10. L. M. Roth, B. Lax, and S. Zwerdling. *Phys. Rev.* **114**, 90 (1959)
11. S. Zwerdling, B. Lax, L. M. Roth, and K. J. Button. *Phys. Rev.* **114**, 80 (1959)
12. E. Burstein, G. S. Picus, R. F. Wallis, and F. J. Blatt. *Phys. Rev.* **113**, 15 (1959)
13. S. Zwerdling, L. M. Roth, and B. Lax. *Phys. Rev.* **109**, 2207 (1958)
14. S. Zwerdling, B. Lax, K. J. Button, and L. M. Roth. *J. Phys. Chem. Solids* **9**, 320 (1959)
15. G. G. Macfarlane and V. Roberts. *Phys. Rev.* **97**, 1714 (1955); *Phys. Rev.* **98**, 1865 (1955)
16. G. G. Macfarlane, T. P. McLean, I. Quarrington, and V. Roberts. *Phys. Rev.* **108**, 1377 (1957)
17. J. R. Haynes, M. Lax, and W. F. Flood. *J. Phys. Chem. Solids* **8**, 392 (1959)
18. B. N. Brockhouse and P. K. Iyengar. *Phys. Rev.* **108**, 894 (1957) *Phys. Rev.* **111**, 747 (1958); B. N. Brockhouse. *J. Phys. Chem. Solids* **8**, 400 (1959)
19. E. F. Gross. *Nuovo Cim. Suppl.* **3**, 672 (1956)
20. S. Nikitine. *J. Phys. Chem. Solids* **8**, 190 (1959)
21. A. H. Kahn. *Phys. Rev.* **97**, 1647 (1955)
22. R. J. Elliott. *Discussions—J. Phys. Chem. Solids* **8**, 193 (1959)
23. E. F. Gross, B. P. Zaharcenja, and P. Pavinskij. *J. tech. Phys., Moscow* **27**, 2177 (1957)
24. E. F. Gross and B. P. Zaharcenja. *J. Phys. Radium* **1**, 68 (1957)
25. K. J. Button, L. M. Roth, W. H. Kleiner, S. Zwerdling, and B. Lax. *Phys. Rev. Letters* **2**, 161 (1959)
26. A. G. Samojlovic and L. L. Korenblit. *C. R. Acad. Sci. U.R.S.S.* **101**, 55 (1955)
27. R. J. Elliott and R. Loudon. *J. Phys. Chem. Solids* **8**, 382 (1959)
28. E. F. Gross. *J. Phys. Chem. Solids* **8**, 172 (1959)
29. E. Grillot and M. Bancie-Grillot. *J. Phys. Chem. Solids* **8**, 187 (1959)
30. G. G. Macfarlane, T. P. McLean, I. Quarrington, and V. Roberts. *Proc. phys. Soc. Lond.* **71**, 863 (1958)
31. Y. Yafet, R. W. Keyes, and E. N. Adams. *J. Phys. Chem. Solids* **1**, 69 (1956)
32. G. G. Macfarlane, T. P. McLean, I. Quarrington, and V. Roberts. *Phys. Rev. Letters* **2**, 252 (1959)
33. W. H. Kleiner and L. M. Roth. *Phys. Rev. Letters* **2**, 334 (1959)

34. L. M. Roth. Unpublished work
35. G. Feher, D. K. Wilson, and E. A. Gere. *Phys. Rev. Letters* 3, 25 (1959)
36. L. M. Roth and B. Lax. *Phys. Rev. Letters* 3, 217 (1959)
37. H. R. Philipp and E. A. Taft. *Phys. Rev.* 113, 1002 (1959)
38. W. S. Boyle. *J. Phys. Chem. Solids* 8, 321 (1959)
39. H. Y. Fan and P. Fisher. *J. Phys. Chem. Solids* 8, 270 (1959)
40. P. Fisher and H. Y. Fan. *Phys. Rev. Letters* 2, 456 (1959)
41. S. Zwerdling, K. J. Button, and B. Lax. *Bull. Amer. phys. Soc.* Ser. II, 4, 145 (1959); *Phys. Rev.* 118, 975 (1960)
42. B. Lax, R. Pupe, W. H. Kleiner. *Bull. Amer. phys. Soc.* Ser. II, 3, 31 (1958)
43. R. E. Harring. *Canad. J. Phys.* 36, 1161 (1958)
44. B. Lax, L. M. Roth, and S. Zwerdling. *J. Phys. Chem. Solids* 8, 311 (1959)
45. W. H. Kleiner, R. N. Brown, and B. Lax. *Bull. Amer. phys. Soc.* Ser. II, 4, 144 (1959)
46. W. Kohn and J. M. Luttinger. *Phys. Rev.* 97, 883 (1954)
47. W. H. Kleiner. *Phys. Rev.* 97, 1722 (1954)
48. H. J. Hrostowski and R. H. Kaiser. *J. Phys. Chem. Solids* 4, 148 (1958)
49. W. Kohn and D. Schechter. *Phys. Rev.* 99, 1903 (1955); D. Schechter. *Ph.D. Thesis* Carnegie Inst. of Tech. (1958)
50. W. S. Boyle and A. D. Brailsford. *Phys. Rev.* 107, 903 (1957)
51. M. H. Johnson and B. A. Lippman. *Phys. Rev.* 76, 828 (1949)
52. R. F. Wallis and H. J. Bowlden. *J. Phys. Chem. Solids* 7, 78 (1958); 8, 318 (1959)
53. R. W. Keyes and R. J. Sladek. *J. Phys. Chem. Solids* 1, 143 (1956)
54. R. R. Rau and M. E. Caspari. *Phys. Rev.* 100, 632 (1955)
55. E. W. J. Mitchell. *Proc. phys. Soc. Lond.* B68, 973 (1955)
56. H. von Kimmel. *Z. Naturf.* 12A, 1016 (1957)
57. T. S. Moss, S. D. Smith, and K. W. Taylor. *J. Phys. Chem. Solids* 8, 323 (1959)
58. R. N. Brown and B. Lax. *Bull. Amer. phys. Soc.* Ser. II, 4, 133 (1959)
59. M. J. Stephen and A. B. Lidiard, *J. Phys. Chem. Solids* 9, 43 (1959)
60. R. J. Keyes, S Zwerdling, S. Foner, H. H. Kolm, and B. Lax. *Phys. Rev.* 104, 1804 (1956)
61. W. G. Spitzer and H. Y. Fan. *Phys. Rev.* 106, 882 (1957)
62. E. O. Kane. *J. Phys. Chem. Solids* 1, 249 (1957)
63. A. K. Walton and T. S. Moss. *J. appl. Phys.* 30, 951 (1959)
64. A. B. Lidiard and M. J. Stephen. *J. Phys. Chem. Solids*, to be published
65. B. Lax and L. M. Roth. *Phys. Rev.* 98, 548 (1955)
66. B. Lax and J. G. Mavroides. *Phys. Rev.* 100, 1650 (1955)
67. L. Rosenfeld. *Z. Phys.* 57, 835 (1930)
68. P. L. Richards and M. Tinkham. *Phys. Rev. Letters* 1, 318 (1958)
69. R. Newman. *Phys. Rev.* 103, 103 (1956)
70. H. P. R. Frederickse and W. R. Hosler. *Phys. Rev.* 108, 1136, 1146 (1957)
71. R. J. Sladek. *Phys. Rev.* 110, 817 (1958)
72. P. N. Argyres. *Phys. Rev.* 109, 1115 (1958)
73. W. S. Boyle and K. F. Rodgers. *Phys. Rev. Letters* 2, 338 (1959)
74. R. G. Wheeler and J. O. Dimmock. *Phys. Rev. Letters* 3, 372 (1959)
75. J. J. Hopfield and D. G. Thomas. *Phys. Rev. Letters* 4, 357 (1960)
76. B. Lax. *Phys. Rev. Letters* 4, 511 (1960)
77. D. F. Edwards and V. J. Lazazzera. *Bull. Amer. phys. Soc.* 5, 177 (1960)
78. R. E. Howard and H. Hasegawa. *Bull. Amer. phys. Soc.* 5, 178 (1960)
79. S. Zwerdling, K. J. Button, B. Lax, and L. M. Roth. *Phys. Rev. Letters* 4, 173 (1960)
80. B. Lax and G. B. Wright. *Phys. Rev. Letters* 4, 16 (1960)
81. S. D. Smith, T. S. Moss, and K. W. Taylor. *J. Phys. Chem. Solids* 11, 131 (1959)
82. T. S. Moss and A. K. Walton. *Proc. phys. Soc. Lond.* 74, 131 (1959), *Physica* 25, 1142 (1959)

THE BAND STRUCTURE AND ELECTRONIC PROPERTIES OF GRAPHITE CRYSTALS

R. R. HAERING, B.A, M.A., Ph.D.

Assistant Professor of Physics, McMaster University, Hamilton, Ontario, Canada

At present

IBM Research Laboratories, Yorktown Heights, N.Y., U.S.A.

and

S. MROZOWSKI, Ph.D.

Professor of Physics,

University of Buffalo, Buffalo, New York, U.S.A.

MS. received October 1959

THE BAND STRUCTURE AND ELECTRONIC PROPERTIES OF GRAPHITE CRYSTALS

1. INTRODUCTION

Although not as popular as germanium or silicon, graphite has lately drawn a considerable amount of attention. A mixed type crystal, being the end product of the process of aromatization, exhibits a number of unique electronic properties of interest to the theoretical as well as to the experimental physicist. The pioneering papers of Ganguli and Krishnan, of Wallace, and of Coulson, followed by an ever increasing stream of publications, have led to a general understanding of a number of fundamental properties of graphite. Today graphite is one of the best known, if not best understood, single crystals. As usual, a great deal of work remains to be done and, consequently, in addition to reviewing what is already understood, this paper will be devoted, in part, to those effects which have not yet been satisfactorily explained.

In an effort to present the whole subject in an orderly and logical manner, a unified point of view had to be adopted. As a result, some of the important contributions or ideas have undoubtedly not been stressed as much as they deserve to be, and perhaps some have not even been mentioned. The subject of graphite is, however, so broad today that a complete discussion is out of the question in a review article of this length.

2. BAND STRUCTURE OF GRAPHITE

The neighbouring carbon atoms in a plane hexagonal layer of graphite are separated by 1·42 Å. On the other hand, the distance between adjacent layers (regardless of how these layers are stacked) is 3·37 Å. Because of this large anisotropy, it is reasonable to base any theoretical model of the band structure of graphite on a discussion of a single hexagonal layer. Since the interlayer spacing is large, we expect that the interaction of neighbouring planes can then be treated in the tight binding approximation.[1] In the following section, we shall therefore consider the band structure of two-dimensional graphite.

2.1. Band Structure of Two-dimensional Graphite

The electronic configuration of an isolated carbon atom is $1s^2 2s^2 2p^2$. Since the 1s electrons are well localized, we shall consider them to be part of the 'ion core' and discuss only the remaining four valence electrons. When carbon atoms are arranged in a plane hexagonal layer, three of the four valence electrons form covalent bonds with valence electrons on neighbouring atoms. These three electrons are described by hybrid orbitals.[2] The normalized orthogonal hybrid orbitals for carbon atoms in a graphite layer are:

$$\psi_{2\pi} = \frac{1}{\sqrt{3}}[\psi_{2s} + \sqrt{(2)}\psi_{2p_x}]$$

$$\psi_{2\pi/3} = \frac{1}{\sqrt{3}}\left[\psi_{2s} - \frac{1}{\sqrt{2}}\psi_{2p_x} + \sqrt{(\tfrac{3}{2})}\psi_{2p_y}\right]$$

$$\psi_{4\pi/3} = \frac{1}{\sqrt{3}}\left[\psi_{2s} - \frac{1}{\sqrt{2}}\psi_{2p_x} - \sqrt{(\tfrac{3}{2})}\psi_{2p_y}\right]$$

The remaining electron is described by a $2p_z$ atomic orbital.

We next observe that the Hamiltonian of a graphite layer is invariant under a reflection in the plane containing the atoms. For this reason, we may label the eigenstates according to whether they are even or odd with respect to the above reflection. We shall denote even and odd states by σ and π respectively. We note that an approximate (tight binding) wave function made up of a linear combination

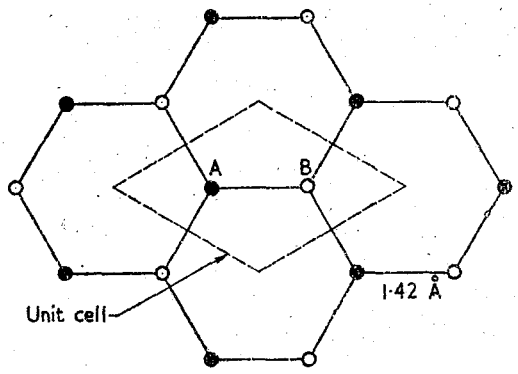


Figure 1. *A single layer of the graphite lattice showing the unit cell*

of hybrid orbitals on different atoms is of the σ type, while a linear combination of $2p_z$ orbitals yields a wave function of the π type. Since the classification of states into σ and π is a rigorous one, σ–π mixing can be excluded and the problems of solving for the σ and π bands are independent.

Figure 1 shows the unit cell of a graphite layer. Since this cell contains two atoms, and since there are three σ orbitals and one π orbital for each atom, we shall obtain eight bands which all merge into the $n = 2$ atomic level in the limit of infinite lattice spacing. Of these eight bands, six are of the σ type and two are of the π type. However, only four of these eight bands will normally be occupied, since the number of states in each band is $2N$ (N is the number of unit cells) and there are $8N$ electrons to be accommodated. We shall refer to the four normally occupied bands as the valence bands and to the four normally empty bands as the conduction bands.

We expect, on the basis of chemical binding theory, that the states of lowest energy correspond to the three occupied σ bands. The order of the remaining five bands should be π, π, σ in order of increasing energy. If the two π bands do not overlap, the Fermi energy, E_F, lies at the maximum energy of the π valence band.

Since the only energy region of interest in most physical problems is the vicinity of the Fermi energy, Wallace[3] was led to ignore the σ bands completely in his calculation of the band structure. The role of the σ bands has been considered by Lomer[4] and Corbato.[5] Lomer estimates that the occupied and unoccupied σ band edges lie $\simeq 1$ eV below the above E_F respectively, while Corbato obtains $\simeq 6$ eV by a very elaborate self-consistent field calculation. The experimental evidence on this point is indecisive. Taft and Apker[6] have studied the energy distribution of photoelectrons from polycrystalline graphite; the results are consistent with a density of states relatively small at the Fermi level and rising to a value several-fold higher 1 eV away. No new band seems to appear up to this depth. In the emission of the soft X-ray K_α line of graphite a broad band is observed (width $\simeq 20$ eV or wider) which shows some fine structure.[7] This band undoubtedly corresponds to the σ and π valence bands. The density of states at the absorption edge is very low and increases rapidly away from this edge. Since only one broad band is observed, the σ and π valence bands must at least partly overlap. If the estimate of Coulson and Taylor[8] for the total width of the π valence band is correct ($\simeq 5$ eV), a value for the depression of the upper edge of the σ band intermediate to those of Lomer and Corbato would seem to be indicated. A tentative decomposition of the fine structure of the K_α band into two components would be consistent with a value 3 to 4 eV for this depression.[7] In view of the above discussion, we shall confine our attention to the two π bands.

A number of calculations of the structure of the π bands have been reported in the literature (Wallace,[3] Coulson,[9] Coulson and Taylor,[8] Carter,[10] Lomer,[4] Corbato,[5] Slonczewski and Weiss[11]). These investigations all agree as to the nature of the π bands near the Fermi energy. It is found that the two π bands are degenerate in energy at the six Brillouin zone corners and that, in the vicinity of the corners, the energy is a linear function of κ. ($\boldsymbol{\varkappa}$ is the wave vector $\mathbf{k}$ measured from a zone corner; $\boldsymbol{\varkappa}=\mathbf{k}-\mathbf{k}_u$.) The treatments given by Carter[10] and by Slonczewski and Weiss[11] are by far the most satisfying. These authors show that the degeneracy is solely due to crystal symmetry. The behaviour of the energy in the vicinity of the degeneracy point is obtained by an application of $\boldsymbol{\varkappa}.\mathbf{p}$ perturbation theory. Diagonalizing the perturbation:

$$\mathscr{H}' = \frac{\hbar}{m}\boldsymbol{\varkappa}.\mathbf{p}+\frac{\hbar^2\kappa^2}{2m} \qquad \ldots(1)$$

where $\boldsymbol{\varkappa}=\mathbf{k}-\mathbf{k}_u$.

One finds to first order in κ:

$$E(\mathbf{k}_u+\boldsymbol{\varkappa}) \simeq E_u \pm \frac{\hbar p_0}{m}\kappa + O(\kappa^2) \qquad \ldots(2)$$

In equation (2), p_0 is a positive constant and E_u is the π band energy at a zone corner.

It is worth while to compare equation (2) with the original result obtained by Wallace.[3] Wallace based his calculation on tight binding wave functions ignoring overlap, and keeping only nearest-neighbour exchange integrals. His result is:

$$E(\mathbf{k}_u+\boldsymbol{\varkappa}) = E_u \pm \frac{\sqrt{3}}{2}\gamma_0 a\kappa + O(\kappa^2) \qquad \ldots(3)$$

where $a = 2{\cdot}46$ Å and

$$\gamma_0 = -\int \mathrm{d}\tau \psi^*_{2p_z}(\mathbf{r}-\mathbf{r}_A)(V-V_{at})\psi_{2p_z}(\mathbf{r}-\mathbf{r}_B) > 0$$

is the 'resonance' integral for π orbitals on nearest-neighbour atoms. In the expression for γ_0, V is the periodic one-electron potential and V_{at} is the potential of an isolated atom.

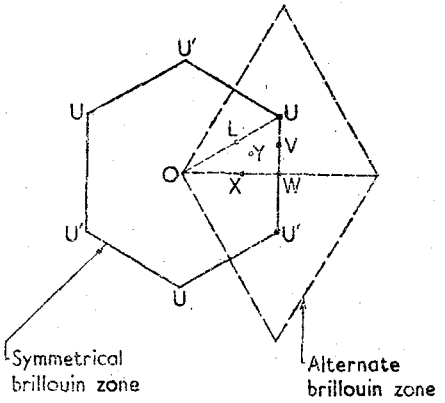


Figure 2. *Brillouin zone for a single layer of graphite*

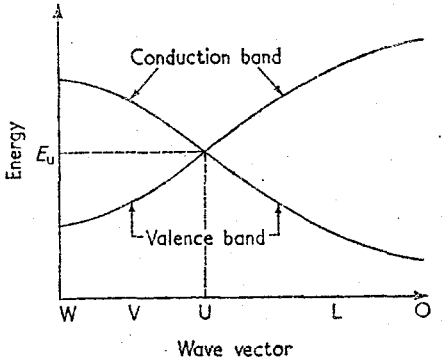


Figure 3. *Variation of π band energies with wave vector along OLUVW of Figure 2*

The value of the constant p_0 in equation (2) is not given by group theoretical considerations and is usually taken to be a parameter to be fixed by experiment. We see that equation (2) agrees with the earlier result of Wallace (cf. equation (3)) if we put

$$p_0 = \frac{\sqrt{3}\,m}{2\hbar}\gamma_0 a$$

Various attempts have been made to estimate p_0 (or γ_0) theoretically. Wallace[3] quotes an early estimate by Coulson ($\gamma_0 \simeq 0{\cdot}9$ eV). A later estimate by Coulson and Taylor[8] yields $\gamma_0 \simeq 3$ eV. Corbato[5] estimates $\gamma_0 \simeq 3{\cdot}8$ eV. Slonczewski and

Weiss,[11] finally, estimate $\gamma_0 \simeq 2{\cdot}4$ eV. We shall see later that experiment indicates a value of 2–3 eV for γ_0, in fair agreement with the latter three estimates.

Slonczewski and Weiss[11] have also pointed out that a negative value of p_0 would imply an overlap of the two π bands. However, if one believes that at least the sign of p_0 is given correctly by a tight binding estimate, such an overlap may be excluded. These authors also justified the neglect of spin–orbit coupling. They estimate that the spin–orbit interaction removes the π band degeneracy but that the resultant splitting is less than 10^{-4} eV and can thus be ignored even at very low temperatures.

Figure 2 shows the Brillouin zone of two-dimensional graphite. The notation is that of Slonczewski and Weiss.[11] Figure 3 shows the variation of energy along OLUVW of this zone. Equations (2) and (3) apply in the vicinity of U (or U'), where the constant energy contours are circles.

2.2. Band Structure of Three-dimensional Graphite

We have seen that the various investigations of the band structure of a layer of graphite all agree as to the nature of the energy contours for energies near the Fermi energy, E_F. The reason for this agreement is that the qualitative shape of the energy surfaces does not depend on the magnitude of the parameter γ_0, provided that we restrict ourselves to positive values of this constant. For any positive value of γ_0, the π bands form two touching cones. However, in the three-dimensional case, the energy bands depend on several parameters and even the qualitative nature of the bands is greatly affected by the relative magnitudes of these parameters. The only satisfactory procedure is then to obtain the most general possible set of energy surfaces in terms of a number of parameters. The values of these parameters must then be determined by comparison with experiment.

There is some doubt as to what the crystalline structure of graphite is. However, it seems fairly certain that the Bernal structure[12] accounts for most of a good single crystal of graphite, and theoretical calculations generally assume this structure. This assumption is certainly open to criticism, since it has been shown (Haering[13]) that the stacking order greatly affects the band structure. We shall return to this point later, but for the present we will assume Bernal stacking. This structure may be obtained from Figure 4 by repeating the layers shown in the manner 121212.... The unit cell of this structure contains four atoms, namely the atoms labelled A, B, A', B' in Figure 4. Since we are considering four orbitals per atom (three σ orbitals and one π orbital), the band energies will be obtained from the solution of a 16×16 secular determinant. As the Hamiltonian is no longer invariant under a reflection in a plane parallel to a graphite layer, we cannot expect that the σ and the π orbitals will not mix. That is, for an arbitrary value of the wave vector $\mathbf{k}$, the Hamiltonian will have matrix elements between σ and π states. The effect of this mixing has been investigated by Johnston[14] and has been found to result in an overlap of the π bands. However, we shall see later that a tight binding calculation taking into account next-nearest plane π overlaps (but ignoring σ–π mixing) yields a similar term in the energy spectrum. For this reason, we may ignore the effect of σ–π mixing provided that we keep next-nearest plane terms for π orbitals and provided that we do not attempt to estimate the magnitude of these terms using the tight binding method. This procedure is mathematically simpler, since it reduces the secular determinant from 16×16 to 4×4.

The Brillouin zone for three-dimensional Bernal graphite is shown in Figure 5. The first calculation of the three-dimensional band structure was made by Wallace.[3] In the vicinity of the lines HKH or H'K'H' of Figure 5, Wallace finds the following energy bands:

$$E(\mathbf{k}) = E_3 \pm \gamma_1 \cos\left(k_z \frac{c_0}{2}\right) \pm \left[\gamma_1^2 \cos^2\left(k_z \frac{c_0}{2}\right) + \tfrac{3}{4}\gamma_0^2 a^2 \kappa^2\right]^{1/2} \qquad \ldots (4)$$

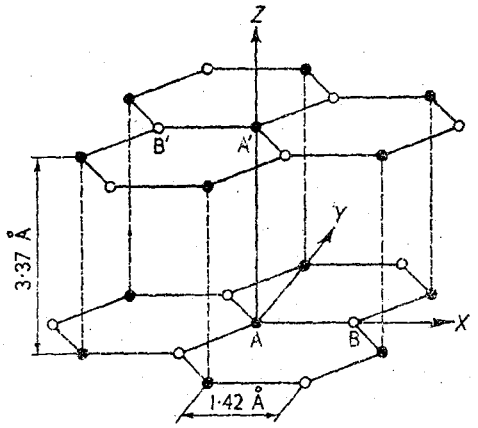


Figure 4. *The graphite crystal lattice*

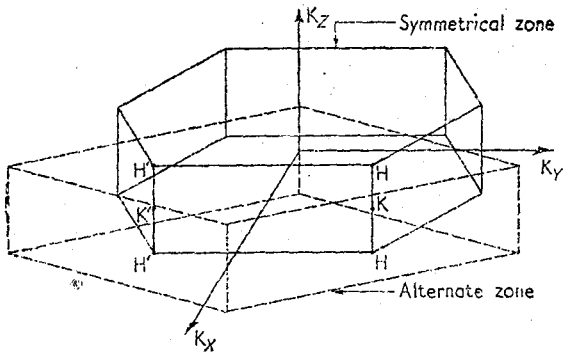


Figure 5. *Possible choices for the Brillouin zone of the Bernal structure*

In equation (4), κ has the same meaning as in equations (1) and (3) and $c_0 = 6{\cdot}74$ Å is the height of the unit cell. The parameter γ_1 is the resonance integral between A and A' atoms. Resonance integrals other than those between A–B and A–A' atoms are ignored in equation (4). Equation (4) also ignores the fact that carbon atoms at locations A and B in Figure 4 have different surroundings as far as their out-of-plane environment goes. This has been pointed out by Carter and Krumhansl[15] and we shall return to this point later.

It is convenient to adopt a more compact notation due to McClure.[16] Let us define the dimensionless quantities:

$$\sigma = \frac{\sqrt{3}}{2} a\kappa \qquad \Gamma = 2\cos\frac{\xi}{2}$$

$$\alpha = \tan^{-1}\left(\frac{\kappa x}{\kappa y}\right) \qquad S = \sigma \exp(i\alpha)$$

$$\xi = k_z c_0$$

then equation (4) becomes:

$$E(\sigma, \xi) = E_3 \pm \tfrac{1}{2}\gamma_1 \Gamma \pm [(\tfrac{1}{2}\gamma_1 \Gamma)^2 + (\gamma_0 \sigma)^2]^{1/2} \qquad \ldots (4a)$$

The variation of energy in the four bands described by equation (4*a*) is plotted in Figure 6 for the zone edge HKH. For any value of ξ, the four bands form a set of

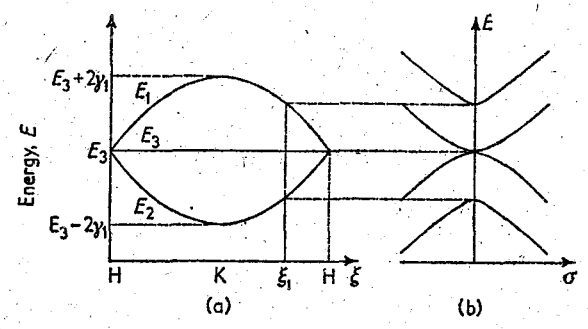


Figure 6. *Variation of energy with wave vector according to Wallace: (a) Variation with ξ along a line HKH (or H′ K′ H′) in Figure 5; (b) Hyperbolic variation with σ in a plane $\xi = \xi_1$ perpendicular to HKH*

hyperboloids of revolution. The outstanding feature of this band structure is the fact that two of the bands are degenerate anywhere along the zone edge and that these two bands are 'flat'; i.e. display no variation with ξ along an edge HKH. Later investigations have shown this 'flatness' to be due to the approximation employed. A variation of E_3 with ξ would result in 'vertical overlap' of the two bands. For the *same* value of ξ the two bands would still just touch; however, the conduction band energies for one value of ξ will overlap valence band energies for a different ξ value. Since the Fermi energy, E_F, coincides with E_3 in the Wallace model, such an overlap would result in free carriers (holes and electrons) and may be expected to have a profound effect on the transport properties, at least at low temperatures. Polder[17] and later Horton and Tauber[18] and Johnston[19] have pointed out that contributions from more distant neighbours than those considered by Wallace[8] can yield a band overlap of the type discussed above. In view of this sensitive dependence of the energy spectrum on distant neighbour interactions, it

is desirable to abandon completely the tight binding approach and to consider what features of the band structure are entirely due to the lattice symmetry. Such group theoretical studies of the band structure have been carried out by Carter,[10] Slonczewski,[20] and Slonczewski and Weiss.[11] Here, we shall outline the work of Slonczewski and Weiss.[11]

The basic idea of the Slonczewski–Weiss model (hereafter referred to as SW) may be stated quite simply. We first note that previous calculations indicate that energies near E_F occur only in the vicinity of the zone edges. We thus forgo a complete calculation of the band structure and confine our attention to the neighbourhood of the zone edges.

Consider for the moment a single layer of graphite. We have already seen that the two π states are degenerate at the zone corners. Let us denote the two basic functions corresponding to such a corner by u_1 and u_2. By considering separately two adjacent planes, we obtain in this way four layer functions of the π type, namely, $\hat{u}_1, \hat{u}_2$, and u_1, u_2. In order to introduce k_z dependence into the problem, we now make up Bloch sums of such layer functions. This procedure is completely analogous to making up a tight binding wave-function for a one-dimensional crystal. Moreover, Slonczewski and Weiss[11] showed that if one defines:

$$\left.\begin{aligned}
\psi_1 &= \frac{1}{\sqrt{2}}\left\{\sum_n \exp(in\xi)\, u_2(\mathbf{r}-n\mathbf{c}_0)+\sum_n \exp\left[i(n+\tfrac{1}{2})\,\xi\right]\hat{u}_1(\mathbf{r}-n\mathbf{c}_0)\right\}\\
\psi_2 &= \frac{1}{\sqrt{2}}\left\{\sum_n \exp(in\xi)\, u_2(\mathbf{r}-n\mathbf{c}_0)-\sum_n \exp\left[i(n+\tfrac{1}{2})\,\xi\right]\hat{u}_1(\mathbf{r}-n\mathbf{c}_0)\right\}\\
\psi_{31} &= \sum_n \exp\left[i(n+\tfrac{1}{2})\,\xi\right]\hat{u}_2(\mathbf{r}-n\mathbf{c}_0)\\
\psi_{32} &= \sum_n \exp(in\xi)\, u_1(\mathbf{r}-n\mathbf{c}_0)
\end{aligned}\right\} \quad \ldots (5)$$

the four functions ψ_1, ψ_2, ψ_{31}, ψ_{32} form the basis of an irreducible representation of the wave vector group corresponding to an arbitrary point on the zone edge. For such an arbitrary point, ψ_1 and ψ_2 are the basic functions for two one-dimensional representations (except at the points H and H′, where these two representations merge into a two-dimensional one). The two functions ψ_{31} and ψ_{32} form the basis of a two-dimensional representation anywhere along the edges HKH or H′K′H′.

Since $c_0 = 6{\cdot}74$ Å is the distance between similar layers, and since u_1 and u_2 are layer functions on the first layer in the cell while $\hat{u}_1$ and $\hat{u}_2$ are layer functions of the second layer, we see at once that ψ_{32} involves only layers 1, 3, 5, ... Similarly, ψ_{31} involves only layers 2, 4, 6, ... However, ψ_1 and ψ_2 involve all layers.

Equations (5) are more general than the corresponding tight binding expressions. Nevertheless, it is instructive to study the tight binding limit of these expressions in order to gain insight into their meaning. This may be done by comparing the transformation properties of the above basic functions with the transformation properties of atomic orbitals. Slonczewski[20] has shown that in the tight binding limit one obtains:

$$\left.\begin{aligned}
\psi_1 &\sim \frac{1}{\sqrt{2}}\left[\sum_{\mathrm{A}} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{A}})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{A}}) + \right. \\
&\qquad \left. + \exp(i\xi) \sum_{\mathrm{A}'} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{A}'})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{A}'})\right] \\
\psi_2 &\sim \frac{1}{\sqrt{2}}\left[\sum_{\mathrm{A}} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{A}})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{A}}) - \right. \\
&\qquad \left. - \exp(i\xi) \sum_{\mathrm{A}'} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{A}'})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{A}'})\right] \\
\psi_{31} &\sim \exp(i\xi) \sum_{\mathrm{B}'} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{B}'})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{B}'}) \\
\psi_{32} &\sim \sum_{\mathrm{B}} \exp(i\mathbf{k}.\mathbf{R}_{\mathrm{B}})\, p_z(\mathbf{r}-\mathbf{r}_{\mathrm{B}})
\end{aligned}\right\} \qquad \ldots(5a)$$

The variation of energy along an edge is now easily established from equations (5) or (5*a*). We keep only the lowest order terms, since the sums over layers may be expected to converge very rapidly due to the large layer separation. In the notation of McClure[16] one obtains:

$$\left.\begin{aligned}
E_1 &= \Delta + 2\gamma_1 \cos\frac{\xi}{2} \\
E_2 &= \Delta - 2\gamma_1 \cos\frac{\xi}{2} \\
E_3 &= 2\gamma_2 \cos^2\frac{\xi}{2}
\end{aligned}\right\} \qquad \ldots(6)$$

The quantities Δ, γ_1, and γ_2 are constants with the dimensions of energy. The notation has been chosen so as to agree as much as possible with that of Wallace.[3] However, it should be remembered that the form of equations (6) does not depend on the tight binding approximation. Nevertheless, we will later find it convenient to interpret the various constants in terms of tight binding in order to get some idea of their relative magnitudes.

The next step is to consider the variation of the energy in a plane k_z = constant. Here, we restrict ourselves to the vicinity of the zone edge and use $\boldsymbol{\kappa}.\mathbf{p}$ perturbation theory. Using the symmetry properties of our basic functions (5), we must establish the form of the momentum matrix. We will not do this here, but will simply write down the form of the Hamiltonian matrix. One finds:

$$H = \begin{pmatrix} E_1 & 0 & H_{13} & H_{13}^* \\ 0 & E_2 & H_{23} & -H_{23}^* \\ H_{13}^* & H_{23}^* & E_3 & H_{33} \\ H_{13} & -H_{23} & H_{23}^* & E_3 \end{pmatrix} \qquad \ldots(7)$$

In equation (7), the rows and columns are ψ_1, ψ_2, ψ_{31}, ψ_{32}, in that order, and:

$$\left.\begin{aligned} H_{13} &= \frac{1}{\sqrt{2}}\left(-\gamma_0+\gamma_4\cos\frac{\xi}{2}\right)S \\ H_{23} &= \frac{1}{\sqrt{2}}\left(+\gamma_0+\gamma_4\cos\frac{\xi}{2}\right)S \\ H_{33} &= \left(\gamma_3\cos\frac{\xi}{2}\right)S \end{aligned}\right\} \qquad \ldots (8)$$

Here γ_0, γ_3, and γ_4 are again constants and $S = \sigma\exp(i\alpha)$ (cf. equation (4*a*)).

The Hamiltonian matrix (7) is the result of the SW calculation. The expressions (6) and (8) for the matrix elements are obtained if only the lowest order terms are kept in the sums over layers. The four bands are obtained from the solution of the secular equation

$$|H-EI| = 0 \qquad \ldots (9)$$

where I is the unit matrix. The six parameters involved in equations (6) and (8) may be estimated in the tight binding limit. However, we cannot have much faith in such estimates because of our lack of knowledge of the crystal potential.

The nature of the energy surfaces may now be obtained from equation (9). If we ignore γ_3 and γ_4 for the moment, we find:

$$\left.\begin{aligned} E &= \tfrac{1}{2}(E_1+E_3) \pm [\tfrac{1}{4}(E_1-E_3)^2+(\gamma_0\sigma)^2]^{1/2} \\ E &= \tfrac{1}{2}(E_2+E_3) \pm [\tfrac{1}{4}(E_2-E_3)^2+(\gamma_0\sigma)^2]^{1/2} \end{aligned}\right\} \qquad \ldots (10)$$

For any value of ξ, we again have a set of four hyperboloids of revolution (cf. equation (4)). However, because of the variation of E_3 along the edges HKH or H′K′H′, one of the valence bands may overlap one of the conduction bands. This is illustrated in Figure 7. In the absence of overlap, the Fermi energy, E_F, would coincide with E_3(K). However, because of overlap, E_F is lowered (see Figure 7) with the result that there are two pockets containing free electrons and a central region containing holes. We note that if γ_2 were chosen negative, we could obtain a central region containing electrons and two 'wings' containing holes. Yet another possibility is shown in Figure 8, in which γ_2 was assumed positive and Δ was chosen greater than $2\gamma_1$. In this case, there is a gap between the valence and conduction bands. If γ_2 were chosen positive and $E_3(\mathrm{K}) = E_2(\mathrm{K})$, the valence bands would just touch the lower conduction band at K and the result would be a zero-gap semiconductor.

So far, we have ignored terms involving γ_3 and γ_4. γ_4 might easily be included but does not affect the arguments qualitatively (cf. equation (8)). On the other hand, the inclusion of γ_3 introduces extra structure. This was first found by Johnston.[19] The effect is most pronounced for energies in the vicinity of E_3, and results in trigonal warping and in additional degeneracies for $\sigma \neq 0$ (Slonczewski and Weiss,[11] McClure,[16] and Nozières[21]). (See, for example, Figure 5 of Nozières.[21])

We may summarize this section by saying that the conduction and valence bands of three-dimensional graphite may touch, overlap, or be separated. It is left to experiment to decide which of these possibilities actually occurs. In fitting the band structure to experiment, we may use the tight binding limit as a guide. This

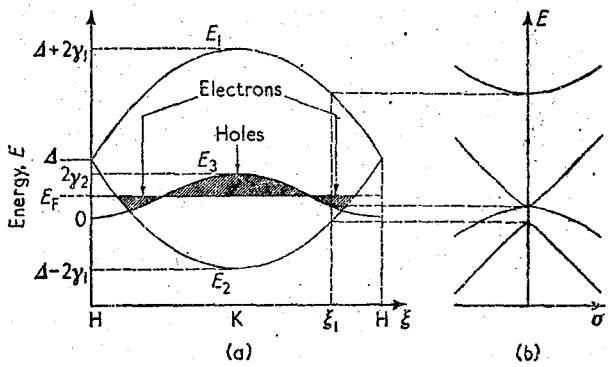


Figure 7. *Variation of energy with wave vector according to Slonczewski and Weiss: (a) Variation with ξ along a line HKH (or H′K′H′) in Figure 5; (b) Hyperbolic variation with σ in a plane ξ = ξ1 perpendicular to HKH. γ2 has been chosen positive*

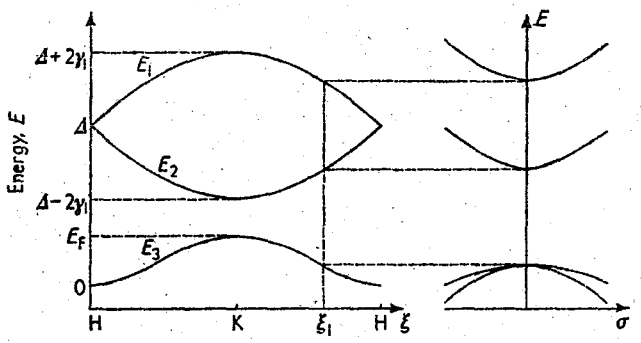


Figure 8. *Possible variation of energy with wave vector according to Slonczewski and Weiss. In this case there is a gap between the highest valence band and the lowest conduction band*

gives us some idea as to the relative magnitudes of the six parameters of the preceding band model. For instance, it is fairly safe to assume $\gamma_0 > \gamma_1 > \gamma_2$. This restriction alone greatly limits the number of possibilities. It means, for example, that the variation of E_3 along a zone edge is less than that of E_1 or E_2. We have used this in plotting Figures 6 and 7. One may also feel reasonably confident that $\Delta < \gamma_1$ (Carter and Krumhansl[15]). This condition alone is sufficient to cause the bands to overlap, at least for positive values of γ_2. However, one cannot estimate the relative magnitudes of γ_1, γ_3, and γ_4 with any confidence.

3. APPLICATIONS OF THE BAND MODEL

3.1. General Remarks

In the following sections we shall mainly outline the various attempts which have been made to correlate the previously mentioned model of the band structure with experiment. Since this model involves six parameters, no single experiment is a severe test of its validity. Our aim is, therefore, to find a consistent set of band parameters which allow us to interpret *all* available experiments. We shall find it convenient to distinguish between two completely different approaches to this problem. The first of these is due to Haering and Wallace,[22] hereafter called the small γ_1 model. The second approach is due to McClure,[16] Nozières,[21] and Boyle and Nozières[23], and will be referred to as the overlap model.

The small γ_1 model is primarily based on a calculation of the diamagnetic susceptibility. Its basic features are an unusually small value of γ_1 ($\gamma_1 \simeq 0{\cdot}005$ eV) and a certain number of excess electrons ($\sim 10^{-4}$ per atom) whose origin is unexplained. Since we may be confident that $\gamma_2 < \gamma_1$, the band overlap discussed in the last section is extremely small. Hence, the hyperbolic bands in Figure 7 approach their asymptotes very quickly, so that for an energy $E \gg E_3$ the bands are well described by the Wallace expression (cf. Figure 6). However, the small value of γ_1 implied by the susceptibility forces one to choose the Fermi energy $E_F \simeq 0{\cdot}06$ eV (measured from E_3 in Figure 6), in order that the de Haas–van Alphen effect may be explained. This choice of E_F implies the presence of about 10^{-4} free electrons per atom. At present, there is no plausible explanation for the origin of such electrons. The small γ_1 model has also been used to calculate the electrical conductivity, the Hall coefficient, the magnetoresistance (Haering and Wallace[22]), the orientation dependence of the de Haas–van Alphen effect and the electronic specific heat (Haering[24]). Attempts have also been made to explain the cyclotron resonance (Murphy[25]) and the high field galvanomagnetic effects (Storey[26]).

In the overlap model, the free carriers arise from the 'vertical' band overlap discussed previously. The basic feature of this model is the simultaneous presence of positive and negative carriers (cf. Figure 7). The presence of holes as well as electrons is confirmed by cyclotron resonance measurements (Galt et al.[27]) and also by the Hall effect data of Soule.[28,29] The six parameters of the overlap model are determined by first fitting the de Haas–van Alphen effect measurements. This procedure determines the ratio γ_0^2/γ_1 and also the values of γ_1, γ_2, and Δ in terms of an assumed value for γ_0. A lower limit for γ_0 ($\gamma_0 \geqslant 1{\cdot}17$ eV) is also obtained, as well as the Fermi energy E_F, which is treated as a seventh parameter. The de Haas–van Alphen effect is not sensitive to either γ_3 or γ_4. In principle, one would expect γ_3 to be determined by the cyclotron resonance data. However, the analysis of this data is complicated (Lax and Zeiger,[30] and Nozières[21]) and has so far only confirmed the value of γ_0^2/γ_1 obtained from the de Haas–van Alphen effect. Infra-red absorption studies (Boyle and Nozières[23]) yield an independent value of γ_1 ($\gamma_1 = 0{\cdot}14$ eV). Another independent value for γ_1 is obtained from studies of the energy loss of electrons in graphite. The energy spectrum consists of several fairly sharp lines with positions independent of the incident energy; the angular dependence of loss and of the intensity were also observed, although the agreement between different workers is not good in this case. Ichikawa[31] has given good

reasons to identify the 7·5 eV loss line with the collective excitation of electrons, and has estimated γ_1 to be $\simeq 0{\cdot}32$ eV on this basis. The reader is referred to the paper by Ichikawa[31] for references to the literature on the subject of electron energy losses in graphite. Finally, one may determine γ_0 from the diamagnetic susceptibility (McClure[32]). We shall defer a discussion of the internal consistency of these band parameter determinations until we have discussed them in more detail.

Since free carriers arise in a natural way in the overlap model, this approach is more satisfying than the small γ_1 model, in which excess carriers must be postulated. However, a surprising number of phenomena may be successfully explained in terms of either model.

In the following sections we shall discuss the various experiments which have contributed to our understanding of graphite. Whenever a certain experiment has been interpreted in terms of either of the above mentioned band models, we shall discuss both interpretations. Section 4 contains a review of investigations which have not yet been interpreted in terms of a band structure model. A summary of the results is given in Section 5. This latter section also contains some general remarks concerning those aspects of graphite about which we know very little at present.

3.2. Diamagnetic Susceptibility

It has been known for a long time that graphite possesses an anomalously high diamagnetic susceptibility in the direction perpendicular to the graphite planes. By analogy with other aromatic ring systems it was presumed that this susceptibility is due to π electrons, possibly representing some kind of generalized London-type diamagnetism. Ganguli and Krishnan[33] investigated the susceptibility of a single crystal over a wide temperature range and found a strong temperature dependence of a type characteristic of an electron gas with a low degeneracy temperature. Ganguli and Krishnan attempted to explain the relation found by considering the zone structure of graphite and the probable shape of the energy surfaces for π electrons. Remarkably good agreement between the theoretical curve and the experimental data was obtained. Some inconsistencies in their theory were removed by Mrozowski,[34] who also pointed out that a correction for London diamagnetism should be introduced before a comparison of the experimental data with the theory is made. Some later attempts to calculate the susceptibility were based on the Wallace band structure (Smoluchowski,[35] Eatherly,[36] and Hove[37]) although it was already clear at that time that Wallace's model was too simplified and that the π bands probably overlapped (Mrozowski[38]). All this work, as well as the more recent calculation of Mase[39] was based on the Landau–Peierls formula,[40] which as shown by Ganguli and Krishnan[33] fixes the low temperature limit and the high temperature behaviour of the susceptibility, leaving little room for testing the band structure as far as the shape of the χ vs $1/T$ curve is concerned (see below the Pacault–Marchand relation). Although these calculations give the correct temperature dependence of the susceptibility, they give its magnitude incorrectly (a factor of 100 too low). This is not surprising since the Landau–Peierls relation was never intended to apply to a case such as graphite where, near the Fermi surface, we have degenerate bands. A magnetic field will introduce off-diagonal matrix elements which couple not only the states in a given

band but which also couple the various bands. The possible importance of such band-to-band terms was pointed out by Adams.[41]

McClure[42] succeeded in calculating the susceptibility of two-dimensional graphite taking into account the above mentioned band-to-band transitions. His calculation yields the correct temperature dependence *and* magnitude for the high temperature susceptibility, provided that γ_0 (the only band parameter of the two-dimensional model) is chosen to be 2·6 eV.

It is worth while to reproduce McClure's approximate calculation of the diamagnetism, since it gives a good deal of insight into the origin of the large susceptibility. In the absence of a magnetic field, the density of states per unit volume predicted by equation (3) is:

$$N_0(E) = \frac{8|E-E_u|}{3\pi c_0 \gamma_0^2 a^2} \qquad \ldots (11)$$

In the presence of a magnetic field, the energy spectrum consists of a series of discrete levels spaced according to:

$$E-E_u = \pm\sqrt{(\tfrac{3}{2})}\gamma_0 a\left(\frac{eH}{\hbar c}\right)^{1/2}\sqrt{n} \qquad \ldots (12)$$

where $n = 0, 1, 2, \ldots$. The degeneracy of each level is

$$\frac{4}{\pi c_0}\left(\frac{eH}{\hbar c}\right)$$

per unit volume. The diamagnetism has its origin in the increase of the total energy associated with this rearrangement of the electrons. It is readily seen that at high temperatures most of this energy increase is due to those electrons which, in the presence of the magnetic field (applied perpendicularly to the graphite layer), occupy the $n = 0$ level. We estimate this increase to be (cf. Figure 9):

$$\Delta E \simeq -\int_{E_u-\delta}^{E_u+\delta} [E-E_u]\, N_0(E) f(E)\, \mathrm{d}E \qquad \ldots (13)$$

where

$$\delta = \frac{\sqrt{3}}{2}\gamma_0 a\left(\frac{eH}{\hbar c}\right)^{1/2}$$

and $f(E)$ is the Fermi–Dirac distribution function. The integral is readily evaluated when $kT \gg \delta$ and yields:

$$\Delta E \simeq \frac{3\gamma_0^2 a^2}{2\pi c_0}\left(\frac{eH}{\hbar c}\right)^2\left[\frac{\partial f}{\partial E}\right]_{E=E_u} \qquad \ldots (13a)$$

The susceptibility resulting from this estimate is:

$$\chi \simeq -\frac{1}{H}\frac{\partial(\Delta E)}{\partial H} = -\frac{3\gamma_0^2 a^2}{4\pi c_0}\left(\frac{e}{\hbar c}\right)^2 \frac{\operatorname{sech}^2(E_F-E_u/2kT)}{kT} \qquad \ldots (14)$$

where E_F is the Fermi energy ($\sim E_u$) and k is Boltzmann's constant. A more exact calculation shows equation (14) to be an over-estimate by a factor of about 3. Choosing $\gamma_0 = 2{\cdot}6$ eV, one obtains:

$$\chi \simeq -\frac{0{\cdot}010}{T}\,\mathrm{sech}^2\left(\frac{E_F - E_u}{2kT}\right)\ \text{e.m.u. gm}^{-1} \qquad \ldots(15)$$

in good agreement with the experimental data (Ganguli and Krishnan[33]).

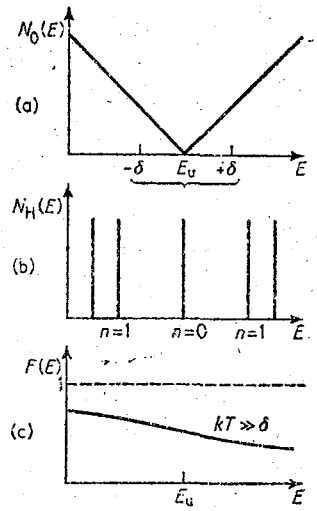


Figure 9. *Origin of the susceptibility of two-dimensional graphite: (a) Density of states in the absence of a magnetic field; (b) Density of states in the presence of a magnetic field; (c) Fermi–Dirac distribution function for* $kT \gg \delta$

The reason for the large susceptibility is apparent from Figure 9. The $n = 0$ level lies precisely at the zero of $N_0(E)$. When electrons 'condense' into this level, they condense from the relatively large energy range $-\delta \leqslant E - E_u \leqslant +\delta$. If band-to-band transitions were ignored, the $n = 0$ level would be replaced by two levels, the corresponding energy spectrum being:

$$E - E_u = \pm\sqrt{(\tfrac{3}{2})}\gamma_0 a\left(\frac{eH}{\hbar c}\right)^{1/2}\sqrt{(n+\tfrac{1}{2})} \qquad \ldots(16)$$

where $n = 0, 1, 2, \ldots$. It is readily verified that equation (16) leads to a negligibly small susceptibility, a fact which clearly demonstrates the importance of band-to-band terms.

It is still possible to treat the band-to-band terms when the three-dimensional model is considered. This was first done by Haering and Wallace[22] (hereafter HW), who based their calculation on equation (7) with $\Delta = \gamma_2 = \gamma_3 = \gamma_4 = 0$. The

energy levels in a magnetic field applied along the c_0 axis are then determined by the equation:

$$(2n+1)\left(\tfrac{3}{4}\gamma_0^2 a^2 \frac{eH}{\hbar c}\right) = E^2 \pm \left[4E^2\gamma_1^2\cos^2\frac{\xi}{2} + \left(\tfrac{3}{4}\gamma_0^2 a^2 \frac{eH}{\hbar c}\right)^2\right]^{1/2} \quad \ldots(17)$$

where $n = 0, 1, 2, \ldots$, and E is measured from E_3 in Figure 6. The degeneracy of each level is still given by $(4/\pi c_0)(eH/\hbar c)$ per unit volume. The chief feature of the level spectrum given by equation (17) is that there is still a sharp level corresponding to $n = 0$. All other levels get broadened by the introduction of k_z. HW then argue that the chief contribution to the susceptibility still comes from the $n = 0$ level. They estimate the susceptibility in a manner analogous to McClure's estimate, using, however, the density of states appropriate to the three-dimensional model, namely:

$$N_0(E) = \frac{16\gamma_1}{3\pi^2 c_0 \gamma_0^2 a^2} + \frac{8|E|}{3\pi c_0 \gamma_0^2 a^2} \quad \ldots(18)$$

Since the density of states is non-vanishing at $E = 0$, the electrons which condense into the $n = 0$ level come from an energy range very much smaller than 2δ. Therefore, proper advantage cannot be taken of the preferential occupation of the lower energy states, and hence the energy of the system will not increase greatly when a field is applied. HW showed that if one chooses

$$\gamma_1 \ll \frac{\sqrt{3}}{2}\gamma_0 a\left(\frac{eH}{\hbar c}\right)^{1/2} \quad \ldots(19)$$

the high temperature susceptibility predicted by the three-dimensional model agrees with equation (15). When equation (19) is not satisfied, the susceptibility is field dependent and approaches zero as the field approaches zero. Since the experimental values show no such field dependence, HW conclude that $\gamma_1 \lesssim 0{\cdot}005$ eV. Even with this value of γ_1, the susceptibility should decrease with magnetic field below about 800 oersteds. However, no experimental information is available in this range.

The above argument is the basis of the small γ_1 model. The calculation of the low temperature susceptibility is fairly involved but agrees with experiment provided the Fermi level at these temperatures is chosen to be at $0{\cdot}06$ eV. Although there is an appreciable shift of E_F with temperature, E_F approaches E_u only asymptotically at high temperatures (cf. Figure 1 of HW[22]). For this reason, the small γ_1 model would predict a slight initial increase of χ as π electrons are depleted by some mechanism such as bromination. This is due to the variation of the sech^2 factor in equation (15). Such an increase has been found by Hennig and McClelland.[43] However, a quantitative comparison with theory is difficult and the startling agreement of the small γ_1 model with the results of Hennig and McClelland is probably fortuitous. In the next section we shall see that the small γ_1 model is also capable of explaining the two periods of the de Haas–van Alphen effect.

Very recently, Uemura and Inoue[44] and McClure[32] have calculated the energy spectrum in the presence of a magnetic field using equation (17) in its more complete form, but still ignoring γ_3 and γ_4. They find that the levels are determined by the equation:

$$(2n+1)\left(\tfrac{3}{4}\gamma_0^2 a^2 \frac{eH}{\hbar c}\right) = (E-E_3)(E-E_3-\Delta)\pm$$
$$\pm\left[(E-E_3)^2(E_1-E_2)^2+\left(\tfrac{3}{4}\gamma_0^2 a^2 \frac{eH}{\hbar c}\right)^2\right]^{1/2} \quad \ldots(20)$$

This result is a suitable generalization of equation (17). The calculation of the susceptibility from equation (20) is complicated and must be done numerically. This has been done by McClure.[32] In order to limit the number of levels which contribute, the calculation is carried out with a field of 5×10^4 oersteds. Although the introduction of γ_1 reduces the susceptibility, McClure[32] finds that this reduction is not as severe as the argument of HW would indicate. The reason is that

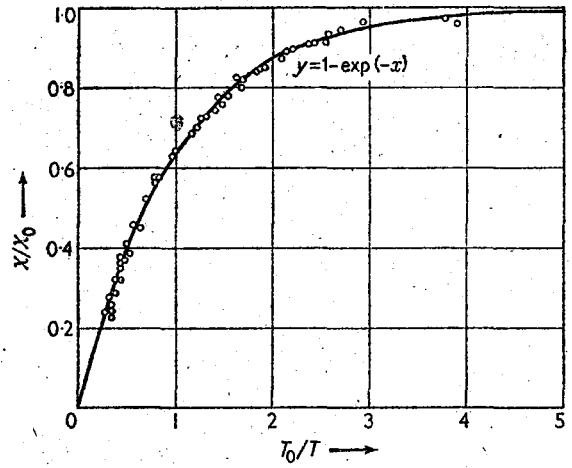


Figure 10. *Temperature variation of the diamagnetic susceptibility of graphite* $\chi_\perp$ *(corrected for core and London diamagnetism). Experimental data of Ganguli and Krishnan,*[33] *the theoretical curve according to equation* (21). $\chi_0 = 28{\cdot}95\times10^{-6}$; $T_0 = 350°K$

there is a small energy increase of *all* levels in the energy range $\pm 2\gamma_1$ and the sum of these contributions yields a term proportional to γ_0^2/γ_1 in the susceptibility. Preliminary results indicate that the following choice of parameters will fit ($\pm$ 10 per cent) both the susceptibility and the de Haas–van Alphen effect:

$$\gamma_0 \simeq 3{\cdot}0\text{ eV} \qquad \gamma_1 \simeq 0{\cdot}35\text{ eV}$$
$$\gamma_2 \simeq 0{\cdot}016\text{ eV} \qquad \Delta \simeq 0{\cdot}022\text{ eV}$$

It is not clear that one is justified in assuming that the susceptibility calculated in fields smaller than 5×10^4 oersteds would agree with McClure's value. This suspicion is based on equation (19). A calculation at lower fields would be highly desirable; however, since the amount of numerical computation increases greatly at lower fields, this may not be feasible.

In conclusion we mention the work of Pacault and Marchand[45] which is

essentially an extension of the paper by Ganguli and Krishnan.[53] Assuming the carriers to constitute a two-dimensional nearly free electron gas and using the Landau–Peierls relation, they derive an expression for the temperature dependence of the susceptibility of the form:

$$\chi = \chi_0[1-\exp(-T_0/T)] \qquad \ldots (21)$$

where T_0 is the degeneracy temperature of the gas. Fitting such a curve to the data of Ganguli and Krishnan, they find $T_0 = 350°$ K for graphite. The concentration of carriers is found to be 1×10^{-5} and their effective mass is about $m_0/300$. The first two of these quantities are in fairly good agreement with the results of a more refined analysis of the band model (see next section). Although the assumptions made in deriving the relation (21) are clearly too simplified, the formula is remarkable since it fits the experimental data (corrected for core and London diamagnetism) very well over the entire temperature range (see Figure 10). Furthermore, Pacault and Marchand show that curves of the type of equation (21) can be fitted equally well to data obtained for different polycrystalline carbons whether or not these are graphitized; thus the formula (21) seems to have very general applicability and might be used as a two-parameter empirical relation until such time as it is derived on the basis of a more refined model.

Very recently, Marchand[46] has reported the results of precision measurements of susceptibility in magnetic fields from 5,000 to 10,000 oersteds. Having found no variation with the field intensity, he concludes that if the electronic energy is proportional to κ^n, n must be equal to either 1 or 2.

3.3. de Haas–van Alphen Effect

The de Haas–van Alphen effect of graphite displays two periods of $2{\cdot}2 \times 10^{-5}$ and $1{\cdot}6 \times 10^{-5}$ oersted^{-1}. The oscillatory dependence of magnetic susceptibility was first observed by Shoenberg.[47] Berlincourt and Steele[48] investigated a single crystal of good quality and obtained results in complete agreement with those of Shoenberg. The theoretical description of the susceptibility oscillations at low temperatures is considerably simpler than the description of the constant part of the diamagnetism. The reason for this is that the oscillatory behaviour is a large quantum number effect while the main contribution to the constant part of the susceptibility comes from small quantum numbers. For this reason, one may use the semi-classical (WKB) approximation in discussing the de Haas–van Alphen effect (Lifshitz and Kosevich[49]). The periods and amplitudes may then be evaluated in terms of quantities which are determined directly by the shape of the Fermi surface. Figures 11 and 12 show the Fermi surfaces of the small γ_1 and overlap models. Figure 11 employs the double height zone convention (Wallace[3]).

Basically, the Fermi surface of the small γ_1 model is a cylinder of variable cross-section whose axis extends along a zone edge HKH (cf. Figure 5). The periods in the de Haas–van Alphen effect are determined by the maximum and minimum areas of cross section of the surface $E = E_F$ with a plane $k_z =$ constant. (We assume here that the field is applied along the c_0 axis.) Equation (4) yields with the double zone convention:

$$(\gamma_0 \sigma)^2 = E_F\left(E_F + 2\gamma_1 \cos\frac{\xi}{2}\right) \qquad \ldots (22)$$

Figure 11 is readily drawn from equation (22). Unless $E_F > 2\gamma_1$, the Fermi surface is spindle shaped and has no minimum cross-sectional area. We identify the two observed periods with S_{max} and S_{min} in Figure 11. Using

$$T = \frac{2\pi e}{\hbar c} S_{ext}^{-1} \quad (T = \text{period}) \qquad \ldots (23)$$

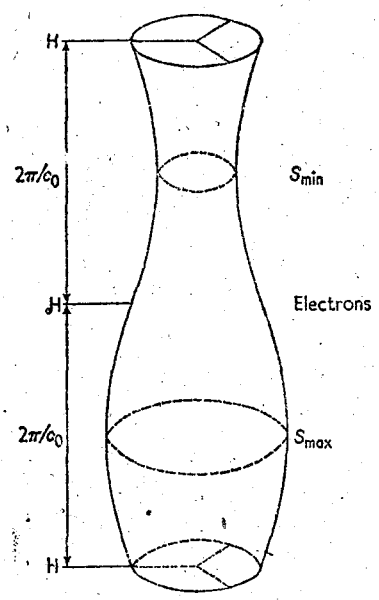


Figure 11. *The Fermi surface of the small γ_1 model. It is assumed that $E_F > 2\gamma_1$ and the double height zone convention is used. The drawing is not to scale*

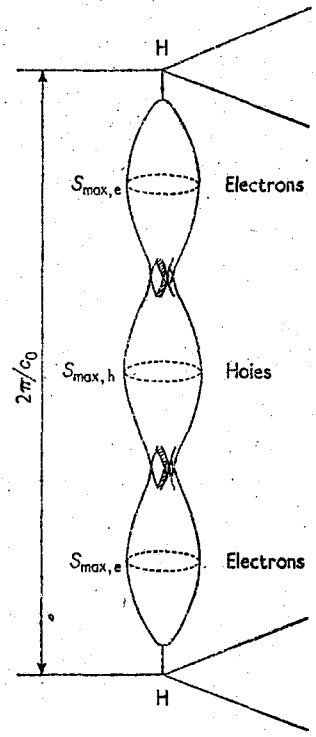


Figure 12. *The Fermi surface of the overlap model. γ_2 has been assumed to be positive. The drawing is not to scale*

one readily finds $E_F \simeq 0{\cdot}065$ eV; $E_F/\gamma_1 \simeq 14$. The amplitudes of the susceptibility oscillations may also be obtained from the work of Lifshitz and Kosevich.[49] Even if temperature damping is accounted for, the calculated amplitudes are about eight times larger than the observed amplitudes of $\simeq 2$ e.m.u. g^{-1}. It seems unlikely that collision damping could account for this discrepancy.

In the small γ_1 model, a complete description of the susceptibility, including its oscillations, fixes *all* band parameters. We find:

$\gamma_0 \simeq 2{\cdot}6$ eV — from high temperature susceptibility

$\gamma_1 \simeq 0{\cdot}005$ eV, E_F (at 0° K) $\simeq 0{\cdot}065$ eV — from de Haas–van Alphen effect

Any further correlation of this model with experiment must, therefore, be made without further adjustment of parameters.

In the overlap model, the two periods both arise from maximum areas of cross-section of the Fermi surface. However, one of these ($2{\cdot}2\times10^{-5}$ oersted^{-1}) is due to electrons, while the other ($1{\cdot}6\times10^{-5}$ oersted^{-1}) is due to holes. The 'feet' of the constant energy surface in Figure 12 are the result of γ_3. It is clear that the periods in the de Haas–van Alphen effect are not sensitive to γ_3, except perhaps when the field is applied perpendicularly to the c_0 axis. McClure[16] also estimates that the neglect of γ_4 only introduces an error of about 10 per cent. The overlap model, as applied to the de Haas–van Alphen effect, is thus a five-parameter model involving γ_0, γ_1, γ_2, Δ, and E_F. The relation analogous to equation (22) is:

$$(\gamma_0\sigma)^2 = (E_F-E_2)(E_F-E_2) \qquad \ldots (24)$$

Equation (24) follows from equation (10) when the double height zone convention is used.

The observed periods may be fitted with any $\gamma_0 > 1{\cdot}17$ eV and with either positive or negative γ_2. A positive value for γ_2 is favoured, since this choice of sign leads to excess electrons in agreement with Soule's Hall effect data (Soule[20, 50]).

In the previous section we have seen that the steady part of the susceptibility requires $\gamma_0 \sim 3$ eV. For this choice of γ_0, the remaining band parameters as determined by the de Haas–van Alphen effect are (McClure[16]):

$$\gamma_1 = 0{\cdot}377 \text{ eV} \qquad \gamma_2 = 0{\cdot}016 \text{ eV}$$

$$\Delta = 0{\cdot}008 \text{ eV} \qquad E_F = 0{\cdot}022 \text{ eV}$$

McClure's analysis also yields the carrier concentrations, the mass, and the mass anisotropy ratio of the carriers responsible for the susceptibility fluctuations. These are:

$$n_e = 2{\cdot}3\times10^{-5} \text{ per atom} \qquad n_p = 1{\cdot}9\times10^{-5} \text{ per atom}$$

$$(m_\perp)_e = 0{\cdot}036\, m_0 \qquad (m_\perp)_p = 0{\cdot}07\, m_0$$

$$\left(\frac{m_\parallel}{m_\perp}\right)_e = 130 \qquad \left(\frac{m_\parallel}{m_\perp}\right)_p = 130$$

An experimental measurement of these mass ratios involves a measurement of the amplitude of the susceptibility oscillations. In this way, Shoenberg[47] obtains (for $\gamma_2 > 0$):

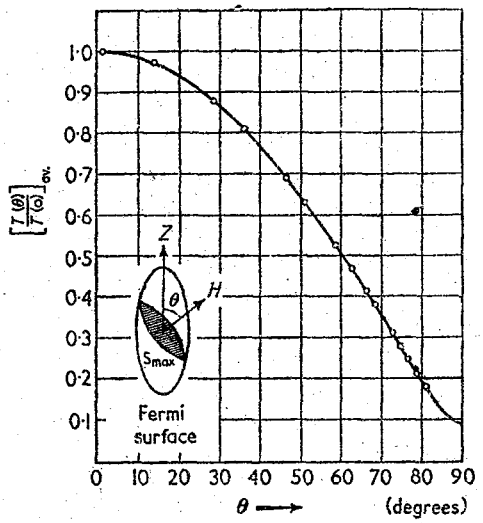
$$\left(\frac{m_\parallel}{m_\perp}\right)_e \simeq 350 \qquad \left(\frac{m_\parallel}{m_\perp}\right)_p \simeq 90\text{–}10{,}000$$

In view of the difficulty of the experimental analysis, this disagreement is not considered serious. It is interesting to note that a negative value for γ_2 would make the disagreement more serious.

McClure's analysis of the de Haas–van Alphen effect is based on Shoenberg's periods of $2{\cdot}20\times10^{-5}$ and $1{\cdot}65\times10^{-5}$ oersted^{-1}. McClure and Smith[51] have recently re-analysed Shoenberg's data. They find periods of $2{\cdot}209\times10^{-5}$ and $1{\cdot}631\times10^{-5}$ oersted^{-1} and mass ratios of 100 and 225 for electrons and holes

respectively. These mass ratios are in good agreement with those determined from the band structure analysis.

A study of the orientation dependence of the de Haas–van Alphen effect is of great interest. Although the small γ_1 and overlap models yield similar periods when the field is applied along the c_0 axis, they predict a somewhat different orientation dependence. Soule and McClure[52] have studied this problem. (Actually, the oscillations used were those in the magnetoresistance, not those in the susceptibility. However, since these have the same origin, we mention their results here.) Only the variation of the *average* period has so far been studied. This variation may be explained with an ellipsoidal energy surface of axis ratio 11·6 to 1. The results are shown in Figure 13.



O de Haas–van Alphen periods
Inset: Ellipsoid with axis ratio 11·6 to 1

Figure 13. *Dependence of the de Haas–van Alphen period on the angle between the magnetic field and the c_0 axis. $T = (2\pi e\hbar)/cS_{max}$. (After Soule and McClure)*

On the basis of the small γ_1 model, one expects

$$\left[\frac{T(\theta)}{T(0)}\right]_{av} = \cos\theta$$

which follows for a cylindrical energy surface. The difference between this prediction and Figure 13 is so small that with the scale used, the two curves would be distinct only for angles $\theta > 80°$. Unfortunately, there are no experimental points in this range of angles.

The individual variation of the two periods would be much more significant than the variation of their average. In this manner, one should be able to establish whether $\partial^2 E/\partial k_z^2$ is positive for both periods (cf. Figure 12) or positive for one period and negative for the other (cf. Figure 11).

3.4. Cyclotron Resonance

Cyclotron resonance effects in graphite have been found and investigated by Galt et al.[27] The magnetic field was applied along the c_0 axis and circularly polarized radiation was employed. The data displays the following general features.

(1) The variation of the power absorption coefficient with magnetic field is approximately symmetric about $H = 0$, suggesting the presence of electrons and holes in approximately equal numbers.
(2) The derivative of the power absorption coefficient shows a number of sharp peaks, suggesting the presence of minority carriers.

The experiment was performed at two different microwave frequencies. These measurements suggested a linear relation between the resonance frequency and the magnetic field.

The approximate symmetry of the power absorption curve does not necessarily imply the presence of both types of carriers. If the constant energy surfaces are extremely anisotropic in the plane perpendicular to the applied magnetic field, circularly polarized radiation will not discriminate between the two types of carriers. However, our discussion of the band structure gave us no reason to believe that the constant energy contours might show an eccentricity of the above type. It is true that some anisotropy exists in the overlap model because of the presence of γ_3 (cf. Figure 12). However, this warping is probably insufficient and, in any case, it does not affect the 'majority carriers', i.e. those carriers located near $S_{\max,e}$ and $S_{\max,h}$ of the Fermi surface. We must therefore conclude that both holes and electrons are present in graphite. This conclusion is a severe blow to the small γ_1 model. In spite of this difficulty, an attempt was made to explain the data of Galt et al. on the basis of this model (Murphy[25]). On the basis of the small γ_1 model, certain peaks in the absorption data may be explained as well as the linear relationship between resonance frequency and magnetic field. However, the symmetry of the power absorption curve does not follow on the basis of the small γ_1 model. A possible way out of the difficulty lies in a suggestion by Chambers.[53] Chambers suggests that (in bismuth) the dip in the power absorption is a magnetoresistance effect. As such, it would certainly be expected to be symmetrical in the magnetic field. Even if this explanation is adopted, most of the complicated detail of the experimental data remains unexplained.

Lax and Zeiger[30] have interpreted the cyclotron resonance data in terms of majority and minority electrons and holes. The shape of the experimental curves suggests that classical resonance conditions apply (mean free path $\gg$ skin depth). Lax and Zeiger show that the dominant peaks in the data resemble very closely the theoretical shape expected for a minority carrier in the presence of a majority carrier of the same type. They conclude that the peaks at -550 and -120 oersteds are due to majority and minority electrons; those at $+750$ and $+240$ oersteds are presumed to be due to majority and minority holes. The ratio of majority to minority carrier concentration is estimated to be about 5:1 and the majority to minority mass ratio is taken to be 3:1. The remaining peaks in the data are then attributed to harmonics of the fundamental majority and minority resonances. Such harmonics are the result of a warped energy surface. In this way, the peaks at $+120$ oersteds and $+91$ oersteds are attributed to second and

third harmonics of the minority holes, while those at -65 and -45 oersteds are interpreted as second and third harmonics of the minority electrons. The weaker peaks at -230, -150, and $+330$ oersteds are attributed to majority harmonics.

Although the above interpretation is not very closely linked to any band model we have discussed, it is possible to make certain statements regarding the symmetry of constant energy surfaces which are compatible with this explanation. This follows because of the existence of certain harmonic selection rules (Zeiger, Lax, and Dexter[54]). For instance, for an energy surface with a threefold axis of symmetry, only harmonics with frequency $\omega_c(1+3n)$ can occur (ω_c = cyclotron frequency; n = integer). The above interpretation, which involves second and third harmonics, therefore excludes an energy surface along HKH of Figure 5. For this reason it is not possible to reconcile this explanation with *any* of our models of the band structure. It is true that the harmonic selection rule does not apply to the outer 'feet' of the Fermi surface (cf. Figure 12), but it would be unreasonable to suppose that all carriers contributing to cyclotron resonance are located in these 'feet'. Furthermore, Nozières[21] has shown that carriers in these 'feet' cannot give rise to sharp resonances.

Nozières[21] has proposed an explanation of the cyclotron resonance data which is consistent with the overlap model. He attributes the broad variation of the absorption coefficient to majority holes and electrons. The shape of the derivative curve indicates that $\omega_c\tau \simeq 2$ for these carriers. The sharp peaks are all presumed to be due to minority carrier resonances and their harmonics. The sharpness of the peaks indicates that the relaxation time for these minority carriers is much longer than that for the majority carriers. In terms of the Fermi surface in Figure 12, these minority carriers are electrons and holes located *near but not in* the 'feet' of the surface. In this way, all observed lines may be accounted for in terms of a fundamental resonance for electrons and holes, and harmonics consistent with the harmonic selection rule mentioned earlier. We briefly summarize Nozières' conclusions.

(1) There is a continuous distribution of mass values over the Fermi surface ranging from 0 to $0{\cdot}054m_0$ for electrons and from $0{\cdot}054m_0$ to $0{\cdot}066m_0$ for holes.
(2) The mass values at $S_{\max,e}$ and $S_{\max,h}$ (cf. Figure 12) are $0{\cdot}031m_0$ for electrons and $0{\cdot}066m_0$ for holes. These values should be compared with the results of the de Haas–van Alphen effect studies.
(3) From the masses of the carriers, one finds $\gamma_0^2/\gamma_1 \simeq 25$ eV, which should again be compared with the de Haas–van Alphen result.
(4) The sign of γ_2 may be established as positive from the fact that the strongest minority resonance occurs on the electron side.

In view of the complexity of a quantitative analysis, the agreement of Nozières' over-simplified calculation with the de Haas–van Alphen results is considered satisfactory. Although it is the incorporation of γ_3 into the calculation which results in the occurrence of harmonic resonances, Nozières' analysis gives no indication of the magnitude of this parameter. At first sight, it would appear that there must be a considerable number of minority carriers present because of the pronounced nature of the minority peaks. Yet, with a reasonable value of γ_3 ($\simeq 0{\cdot}1$ eV), one obtains warping only in the immediate vicinity of the 'feet' of the

Fermi surface, and hence only very few minority carriers. Without further investigation, the agreement of cyclotron resonance results with the overlap model cannot, therefore, be considered quantitative.

3.5. Absorption and Reflection of Light

The reflectivity of various samples of graphite to light at normal incidence has been measured in the range of wavelengths 6,000 Å to 2,300 Å by Humphreys-Owen and Gilbert.[55] From the experimental data the refractive index and the extinction coefficient have been derived. All three quantities show a peak at around 2,600 Å (the refractive index at slightly longer and the extinction coefficient at slightly shorter wavelengths), thus indicating the existence of an anomalous dispersion region. The peak probably corresponds to transitions between the states of maximum density in the π valence and the π conduction bands. Such an interpretation seems to be in agreement with the work of Coulson and Taylor,[8] who estimated the energy difference between these states to be about 5 eV. The numerical agreement with the photoelectric work function (4·5 to 4·8 eV, Taft and Apker[6]) and the thermionic emission work function (4·39 eV, Braun and Busch[56]) is certainly fortuitous. Measurements of reflectivity were also performed by McCartney and Ergun[57] with light of a wavelength $\simeq$ 5,500 Å on a $(10\bar{1}2)$ face of a graphite crystal for different planes of polarization; from these data, the optical constants perpendicular and parallel to the graphite planes have been derived. The extinction coefficient is in good agreement with that of Humphreys-Owen and Gilbert,[55] but the index of refraction and the reflectivity do not agree. The disagreement in these values might be due to the preliminary nature of the results of the latter investigators (Humphreys-Owen and Gilbert).

Recently, high temperature emissivity and reflectivity measurements were performed by two groups of investigators (Null and Lozier,[58] Plunkett and Kingery[59]) leading to some clarification of disagreements existing in the literature.

Boyle and Nozières[23] have measured the infra-red absorption of graphite and have interpreted their results in terms of the band structure. Their measurements consist of a determination of the emissivity of a single crystal of graphite at 250° C and a determination of the reflection coefficient, R, at 4° K. According to Kirchhoff's law, the emissivity is proportional to $1-R$, and hence to the amount of incident energy absorbed in a reflection. Of the two experiments, the emissivity determination is considered more reliable, but the low power level of the emitted radiation limits the measurements to wavelengths shorter than 12 microns. The low temperature measurements of the reflection coefficient extend to 30 microns and agree qualitatively with the reflection coefficient inferred from the emissivity measurements in the range from 4 to 7 microns.

Figure 14 shows the results obtained. Since the work did not extend to wavelengths shorter than 1·5 microns only the beginning of the flattening out of the curve is noticeable. Humphreys-Owen and Gilbert[55] found a plateau extending from 0·6 to about 0·35 micron, that is, up to the point where the increase associated with the 0·26 micron peak starts to show up.

Although the low temperature measurements are, in principle, a more sensitive test of the band structure, they are difficult to interpret because the light is not at normal incidence. Boyle and Nozières[23] therefore restrict their quantitative analysis to the emissivity measurements. In this analysis, certain simplifying

approximations about the band structure are made. The parameters γ_2, γ_3, γ_4, and Δ are all equated to zero. In this approximation, the difference between the small γ_1 and overlap models vanishes, except for the position of the Fermi surface, and at 250° C even this difference is negligible. The band structure is, therefore, just the Wallace structure as shown in Figure 5. One expects that the neglect of γ_2, γ_3, γ_4, and Δ is justified if $\hbar\omega$, the energy of a quantum of radiation, is larger than any of these parameters. This will be the case in the above range of frequencies.

With these approximations, the absorption properties of graphite are determined by γ_1. Figure 14 shows the measured emissivity as well as the theoretical curve. The value of γ_1 has been adjusted so that the break in the theoretical curve comes

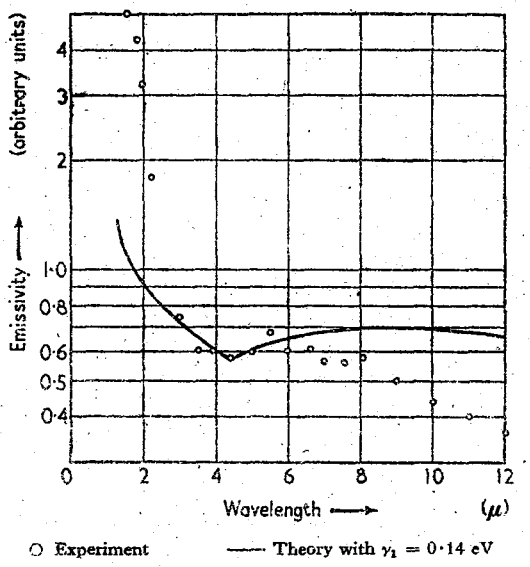


Figure 14. *The emissivity of a single crystal of graphite (After Nozières)*

at 4·5 microns. This yields $\gamma_1 = 0{\cdot}14$ eV. Since theory and experiment agree poorly both above and below this point, the value of γ_1 determined in this way seems unreliable. It is also difficult to see how at 250° C, where $kT \simeq 0{\cdot}05$ eV, one can have any sensitive criterion for a parameter as small as 0·14 eV. A more convincing criterion for γ_1 is obtained from the susceptibility measurements. For the overlap model, we have seen that these measurements require $\gamma_1 \simeq 0{\cdot}35$ eV.

3.6. Electronic Specific Heat

The specific heat of graphite has received a good deal of attention in the literature. We shall here concern ourselves only with the electronic contribution. The electrons contribute a term γT to the total specific heat. The constant γ is proportional to the density of electron states at the Fermi surface and is readily calculated when the band structure is known.

Measurements of the electronic specific heat have been made by Keesom and Pearlman[60] and by DeSorbo and Nichols.[61] The value of the constant γ derived from these measurements is $0{\cdot}6\times10^{-5}$ cal mole^{-1} deg.$^{-2}$. For further measurements of the (total) specific heat the reader is referred to the bibliography of the paper by DeSorbo and Nichols.

Calculations of the constant γ in terms of the band parameters have been carried out by Komatsu and Nagamiya,[62] Bowman and Krumhansl,[63] McClure,[16] and Haering.[24] Using the three-dimensional Wallace band structure (cf. Figure 5), one finds (Bowman and Krumhansl[63]):

$$\gamma = 6{\cdot}62\times10^{-5}\frac{\gamma_1}{\gamma_0^2}\left\{1+\pi\left|\frac{E_F}{2\gamma_1}\right|+\frac{3}{2}\left(\frac{E_F}{2\gamma_1}\right)^2\right\}\text{cal mole}^{-1}\text{ deg.}^{-2} \quad \ldots(25)$$

Here, E_F is the position of the Fermi level relative to E_3 and γ_0, γ_1, and E_F are to be expressed in electron volts.

In calculating γ on the basis of the overlap model, one should use equation (18) of McClure.[16] However, this density of states is so nearly equal to that implied by equation (25) with the same values of γ_0, γ_1, and E_F, that we may use equation (25) without introducing any serious error. Using $\gamma_0^2/\gamma_1 = 25$ eV, $\gamma_1 = 0{\cdot}35$ eV, $E_F = 0{\cdot}02$ eV, one obtains

$$\gamma \simeq 0{\cdot}3\times10^{-5}\text{ cal mole}^{-1}\text{ deg.}^{-2}$$

This is one-half of the observed value. A similar value is obtained on the basis of the small γ_1 model. In this case, equation (25) must be replaced (Haering[24]) by:

$$\gamma = 20{\cdot}9\times10^{-5}\frac{E_F}{\gamma_0^2}\text{ cal mole}^{-1}\text{ deg.}^{-2} \quad \ldots(26)$$

With $E_F = 0{\cdot}065$ eV and $\gamma_0 = 2{\cdot}6$ eV, this yields

$$\gamma \simeq 0{\cdot}2\times10^{-5}\text{ cal mole}^{-1}\text{ deg.}^{-2}$$

Both of these values are low. The higher experimental value might be due to presence of excess carriers (especially in the case of polycrystalline samples). However, in view of the difficulties of the experiments and the uncertainties of extrapolation procedure, the disagreement may not be real.

3.7. Non-oscillatory Hall Effect and Magnetoresistance

It was convenient to discuss the steady susceptibility and the de Haas–van Alphen effect separately. Our discussion of the Hall effect and the magnetoresistance will be similarly subdivided. We first discuss non-oscillatory effects. A correlation of the observed properties with the band structure model is complicated by the dependence of these properties on the scattering mechanisms. This remark does not apply to the high field oscillatory behaviour. These oscillations, which will be discussed in the next section, have their origin in the passage of quantized levels through the Fermi surface. They confirm the values of the band parameters deduced from the de Haas–van Alphen effect.

Of the non-oscillatory transport properties, only the Hall coefficient, R, is relatively insensitive to the relaxation time. In fact, under the assumption of a constant relaxation time, τ, over the Fermi surface, the zero field Hall coefficient

is independent of τ, and may be calculated when the band structure is known. This has been done by Johnston[64] and by Haering and Wallace.[22] The results of these investigations are qualitatively similar. The Hall coefficient is found to be negative ($\simeq -0{\cdot}7$ cm^3 C^{-1}) at low temperatures and approaches zero asymptotically at high temperatures.

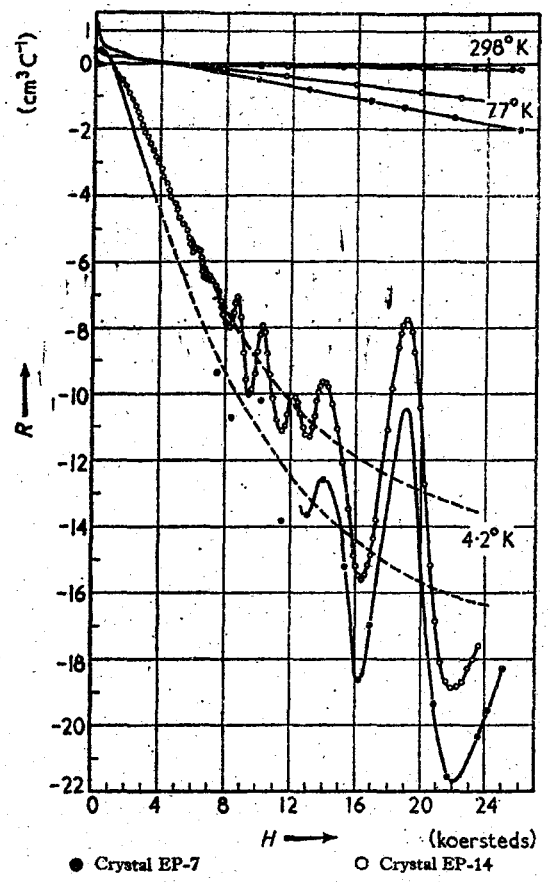


Figure 15. *Field dependence of the Hall coefficient of samples EP-7 and EP-14 at* 298°, 77°, and 4·2°*K* (*After Soule*)

The first measurements of the Hall constant for single crystals at low temperatures were performed by Kinchin.[65] The Hall constant was found to be strongly field dependent, so that a direct comparison of the results with the calculations of Johnston and of Haering and Wallace was not possible without extrapolation to zero field. The general form of the temperature and field dependence found by Kinchin for his single crystals was not substantiated by the more precise measurements of Berlincourt and Steele[48] and of Soule.[29] The latter two papers are in substantial agreement with each other, although individual differences exist between various samples. Figures 15 and 16 show the results obtained by Soule[29]

and illustrate this effect. Therefore, it seems possible that the results of Kinchin are simply a further manifestation of this variation from sample to sample. In extending the measurements to very low fields, Soule has shown that the Hall coefficient is strongly field dependent even at fields as low as 100 oersteds, which invalidates all attempts at extrapolation of the data to zero field. A theoretical analysis of the data requires therefore a curve fitting procedure which has to be done for each sample individually. This has been done by McClure[66] using Soule's samples EP-7 and EP-14. McClure's analysis is based on the magnetoconductivity tensor (McClure[67]) and involves the magnetoresistance as well as the Hall effect data. It is assumed that the magnetic field is applied along the c_0 axis (here called the z axis) and that the electric field is applied in the plane perpendicular

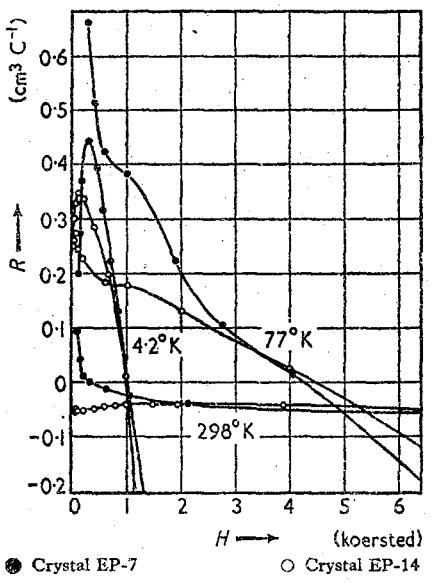


Figure 16. *Field dependence of the Hall coefficient in low fields (After Soule)*

to this axis. The energy surfaces are assumed to have rotational symmetry about the c_0 axis, an approximation which is reasonable for the majority electrons and holes, but which does not describe the minority carriers very well. In effect, we are thinking of the overlap model but ignoring the warping due to γ_3. Under these conditions, the magnetoconductivity tensor has only two independent elements σ_{xx} and σ_{xy}. These are related to the zero field, xy plane conductivity σ, and to the Hall coefficient R through:

$$\left.\begin{aligned} \sigma_{xx} &= \frac{\sigma}{1+(R\sigma H)^2} \\ \sigma_{xy} &= \frac{\sigma(R\sigma H)}{1+(R\sigma H)^2} \\ R &= \frac{\sigma_{xy}}{H(\sigma_{xx}^2+\sigma_{xy}^2)} \end{aligned}\right\} \quad \ldots (27)$$

In obtaining equations (27), it has been assumed that the Boltzmann equation is valid and that the relaxation time is constant over a path in k-space defined by the intersection of the Fermi surface with a plane k_z = constant. For large magnetic fields, where both the magnetoresistance (see Figure 17) and the Hall coefficient (see Figure 15) display oscillations at low temperatures, a Boltzmann treatment

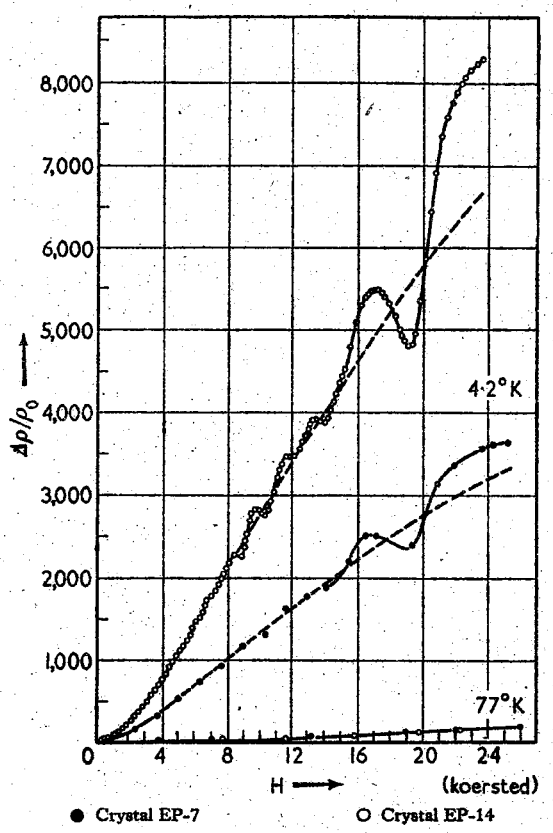


Figure 17. *Field dependence of the magnetoresistance of samples EP-7 and EP-14 at 298°, 77°, and 4·2° K (After Soule)*

cannot be trusted. Certainly, such a treatment will not reproduce the oscillations and it is doubtful whether one should even apply it to the 'base line'.

The experimental curves of σ_{xx} and σ_{xy}/H versus H are fitted to a linear combination of 'Lorentz factors':

$$\sigma_{xx} = \sum_n \frac{A_n}{1+\left(\frac{H}{H_n}\right)^2} \qquad \ldots (28)$$

A similar expression is valid for σ_{xy}/H. The best values of the coefficients A_n and the 'saturation fields' H_n are obtained by a least-squares fitting procedure. These

are listed in Table 1. A relation for σ_{xx} more general than equation (28) would be:

$$\sigma_{xx} = \sum_n \int \frac{\mathrm{d}sg_n(s)}{1+(sH)^2} \qquad \ldots (29)$$

where $g(s)$ is a distribution function and the sum on n extends over the various carriers present. When Fermi–Dirac statistics are applicable, $g(s)$ is sharply peaked so that we may replace the integral in equation (29) by the value of the integrand (apart from $g(s)$) at the peak. This yields equation (28). The Hall coefficient and magnetoresistance calculated on the basis of the entries in Table 1

Table 1. Parameters from Least-square Fit of Magnetoconductivity Tensor Elements (After McClure[66])

For $\sigma_{xx} \times 10^{-16}$

Sample	Temperature (°K)	A_1	H_1	A_2	H_2	A_3	H_3
EP-14	4·2	1430·0	267·0	5720·0	74·4	2·0	14000·0
EP-14	77·0	416·0	1230·0	19·2	6630·0	—	—
EP-14	298·0	181·0	7290·0	36·0	19800·0	—	—
EP-7	4·2	442·0	428·0	341·0	116·0	1·7	14000·0
EP-7	77·0	358·0	1190·0	60·9	3710·0	—	—
EP-7	300·0	181·0	7200·0	38·3	22000·0	—	—

For $\left(\frac{\sigma_{xy}}{H}\right) \times 10^{-11}$

EP-14	4·2	−1240·0	267·0	15010·0	74·4	—	—
EP-14	77·0	−87·4	1410·0	128·0	1140·0	15·0	180·0
EP-14	300·0	−2·22	6870·0	0·123	28300·0	−1·09	260·0
EP-7	4·2	−236·0	487·0	2220·0	145·0	—	—
EP-7	77·0	−210·0	1240·0	315·0	1000·0	500·0	50·0
EP-7	300·0	−2·28	6850·0	0·0863	32010·0	18·3	67·0

All quantities are in c.g.s. (Gaussian) units.

fit the data displayed in Figures 15–17 very well. The reader is referred to the papers by Soule[29] and by McClure[66] for diagrams and for a more complete discussion.

The results summarized in Table 1 may be interpreted in terms of majority and minority carriers. We note that the terms in columns A_1 and A_2 of Table 1 are present at all temperatures, while the term A_3 is required only at 4·2° K in σ_{xx} and only at 77° and 300° K in σ_{xy}/H. Furthermore, at a given temperature, the saturation field H_3 is always very different from H_1 or H_2, while the latter two are of the same order. It is therefore reasonable to assume that the terms in column A_3 of Table 1 represent minority carriers.

It is possible to infer densities and mobilities for the various carriers, through the

use of certain dispersion relations involving the elements of the magnetoconductivity tensor. Rewrite the conductivity σ_{xx} as:

$$\sigma_{xx} = \sigma_{xx,e} + \sigma_{xx,h} \qquad \ldots (30)$$

where $\sigma_{xx,e}$ and $\sigma_{xx,h}$ stand for the partial conductivities due to electrons and holes. It is possible to show (McClure[66]) that

$$(n+p)ec = \frac{2}{\pi}\int_0^\infty \mathrm{d}H\sigma_{xx}(H)$$

$$\sigma_{xx,h} - \sigma_{xx,e} = \frac{2}{\pi}\int_0^\infty \mathrm{d}H\frac{\sigma_{xy}}{H} \qquad \ldots (31)$$

where n and p are the electron and hole concentrations respectively. The integrals involved in these expressions may be evaluated numerically from the experimental data by extrapolating these results to infinity. McClure has applied equation (31) to the minority carriers *separately*; he has also applied equation (31) to the *total* contribution of the majority carriers, i.e. to the conductivities implied by the sum of the first two terms in equation (28). In this way, the results listed in Table 2 are obtained.

Using these results and the effective masses determined by the de Haas–van Alphen effect, one may also make estimates of the relaxation time of the majority

Table 2. The Properties of the Current Carriers in Graphite as Derived from Galvanomagnetic Data (After McClure[66])

Sample	EP-14			EP-7		
Temperature (°K)	4·2	77	298	4·2	77	300
Majority hole						
Density $\times 10^{-18}$ (cm^{-3})	2·88	2·19	7·04	2·08	2·39	7·22
Mobility $\times 10^{-4}$ ($cm^2V^{-1}sec^{-1}$)	104·0	7·33	1·01	64·5	6·53	0·99
Majority electron						
Density $\times 10^{-18}$ (cm^{-3})	2·92	2·24	7·04	2·14	2·46	7·36
Mobility $\times 10^{-4}$ ($cm^2V^{-1}sec^{-1}$)	83·9	6·38	1·13	59·1	4·86	1·09
Hole mobility/electron mobility	1·24	1·15	0·89	1·09	1·34	0·90
Theoretical density $\times 10^{-18}$ (cm^{-3})	2·4	3·6	13·4	2·4	3·6	13·4
Minority hole						
Density $\times 10^{-15}$ (cm^{-3})		3·3			8·7	0·57
Mobility $\times 10^{-4}$ ($cm^2V^{-1}sec^{-1}$)		5·7			20·0	15·0
Minority electron						
Density $\times 10^{-15}$ (cm^{-3})	200·0		0·50	200·0		
Mobility $\times 10^{-4}$ ($cm^2V^{-1}sec^{-1}$)	0·7		39·0	0·7		

The mobilities refer to motion in the layer plane. Theoretical density is the result of a numerical calculation based on the overlap model.

carriers. This is done through the use of the relation $\mu = e\tau/m^*$. At 4·2° K, the relaxation times are $3{\cdot}5 \times 10^{-11}$ sec and $1{\cdot}5 \times 10^{-11}$ sec for holes and electrons, respectively.

The above analysis yields carrier densities which agree within 50 per cent with those derived from the overlap model. This in itself can hardly be considered a success of this model, since we have seen that the density of states is sensitive only to the values of γ_0 and γ_1. However, the change of sign of the Hall coefficient in low fields is definite evidence for the presence of positive and negative carriers, although this does not establish that these carriers arise from band overlap. A change in sign of the Hall coefficient can also be obtained from the change in carrier concentration resulting when a magnetic field is applied (Uemura and Inoue[44]). However, these authors postulate acceptor levels in the conduction band. At present, it is not known whether such states are meaningful.

3.8. Oscillatory Hall Effect and Magnetoresistance

One expects that, at low temperatures, oscillations should appear in the galvanomagnetic properties just as they appeared in the susceptibility. These oscillations are the result of the fluctuations in the density of states which arise as the various quantized levels are swept out through the Fermi surface. The conditions that such oscillations appear are that the electron gas be highly degenerate ($kT/E_F \ll 1$) and that the levels be swept through the Fermi surface one at a time ($kT/\Delta E \lesssim 1$). Here ΔE is the energy difference between adjacent quantum levels corresponding to the same value of k_z. Just as the susceptibility oscillations were a large quantum number effect, so are the oscillations in the galvanomagnetic properties (except perhaps at very large fields). A semi-classical treatment is therefore again appropriate (Zilberman,[68] Lifshitz and Kosevich[49]).

Since the physical reason for the oscillations in the Hall effect and magnetoresistance is the same as the reason for the susceptibility fluctuations, we expect that the periods observed in any of these fluctuations will be identical. Berlincourt and Steele[48] were the first to observe these oscillations in graphite. They found a common period of $2{\cdot}15 \times 10^{-5}$ oersted^{-1}. In the de Haas–van Alphen effect an additional period of $1{\cdot}61 \times 10^{-5}$ oersted^{-1} was observed. More recently, Soule[50] has obtained similar results and has established that a (weaker) period of $1{\cdot}58 \times 10^{-5}$ oersted^{-1} is also present in the Hall effect and magnetoresistance.

Although the periods in all of the above mentioned effects are identical, there exist certain phase differences between them. Berlincourt and Steele[48] find a phase difference of $\pi/2$ between Hall effect and de Haas–van Alphen effect fluctuations. They also report a phase difference of $\pi/4$ between the oscillations in the magnetoresistance and the Hall effect; however, Soule[50] reports a value of π for this phase lag. The reason for this difference is not understood at present.

It is clear that an analysis of the galvanomagnetic fluctuations cannot establish the nature of the carriers. We assume that the carriers are electrons and holes as indicated by the cyclotron resonance and Hall effect measurements. It is, however, possible to obtain additional information about effective masses and carrier densities from such an analysis. The values obtained are (for sample EP-14):

$$m_e = 0{\cdot}030 m_0 \qquad m_p = 0{\cdot}060 m_0$$

$$n = 2{\cdot}8 \times 10^{18}\ \text{cm}^{-3} \qquad p = 2{\cdot}2 \times 10^{18}\ \text{cm}^{-3}$$

The mass values compare favourably with those obtained from cyclotron resonance and from the de Haas–van Alphen effect: The carrier densities should be compared to the majority hole and electron densities for EP-14 at 4·2° K (see Table 2).

4. RELATED STUDIES OF GRAPHITE

In this chapter we shall review a number of other investigations of the electronic properties of graphite. We list these in a separate section because the effects in question have not yet been sufficiently analysed in terms of a band model, either due to the mathematical difficulties of the problems or due to the complexity of the experimental findings or because the investigations are still at a preliminary stage.

4.1. Thermoelectric Effect and Electrical Conductivity

No studies of the thermoelectric effect were performed on single crystals of graphite. Since graphite is strongly anisotropic, it was to be expected that the effect would be different in the two directions along and across the graphite planes. Ubbelohde and Orr[69] have cut out samples from blocks of material in which the alignment of crystallites, although not perfect, was not far from that for a single crystal. The thermoelectric power of such graphite against copper was found (presumably at room temperature) to be about $+30\ \mu V\,deg.C^{-1}$ perpendicular to the plane as compared with about $+11\ \mu V\,deg.C^{-1}$ along the plane. Two samples cut out in these two directions were combined in a thermo-junction, which showed a current varying linearly with the temperature of the junction between 0° and 700° C. It is believed that the effect is due to the directional anisotropy of the work function, so that the conduction band is separated by an energy gap from the valence band in the c direction resulting in a junction which behaves like a semiconductor–metal contact.

It is difficult to find good single crystals of graphite of a reasonably large size; it is also difficult to ensure good contact between electrodes and the crystal. These are probably the main reasons for discrepancies in the values and in temperature dependence reported for the electrical conductivity ρ_a along the graphitic planes and ρ_c in the direction perpendicular to the planes. In measuring ρ_a one has to ensure that good contact of the electrodes with all graphitic planes is obtained. Since this is difficult, the experimental values will always tend to be too high. Since ρ_c is much higher than ρ_a, there are many reasons why the measured ρ_c might be too low: Rothstein[70] drew attention to the fact that the presence of one spiral dislocation can cut down the resistance by a large factor. Although such dislocations are rare in graphite crystals, one has to remember that anything that disturbs the perfect alignment of planes, like mosaic structure, will lead to a zigzag flow and to a decrease in ρ_c. On the other hand, a layer of impurities or a crack parallel to the planes will cause a considerable increase in ρ_c.

There is general agreement that at room temperature $\rho_a \simeq 5 \times 10^{-5}\ \Omega cm$. As far as ρ_c is concerned, the values reported vary from 1 Ωcm (Krishnan and Ganguli,[71] Dutta[72]) to $5 \times 10^{-3}\ \Omega cm$ (Washburn,[73] Primak and Fuchs[74]). Irregularities in the curves of Dutta[72] suggest some contact troubles in his experiments; on the other hand Primak and Fuchs[74] could directly estimate the ratio ρ_c/ρ_a by a study of the potential distribution on the surface of the crystals. It is for this reason that

the lower value of ρ_c seems to be the more reliable one. As to the temperature dependence, with increase of temperature ρ_c increases at first from 0° K up to a maximum located below room temperature[74] and subsequently decreases (measurements were made up to 450° C[72]). ρ_a increases steadily with increase of temperature, rapidly at first and then above room temperature in a linear fashion. Figure 18 gives the recent results of Soule[23] for his samples EP-7 and EP-14 and shows the general character of the variation of ρ_a in the low temperature range as originally found by Reynolds, Hemstreet, and Leinhardt.[75]

The shape of the curve in Figure 18 may be understood qualitatively in terms of the overlap model. If we refer to Table 2, we first note that the mobilities decrease with increasing temperature roughly as $T^{-1\cdot2}$. On the other hand, the carrier densities are roughly constant up to 77° K but increase by a factor of 2–3

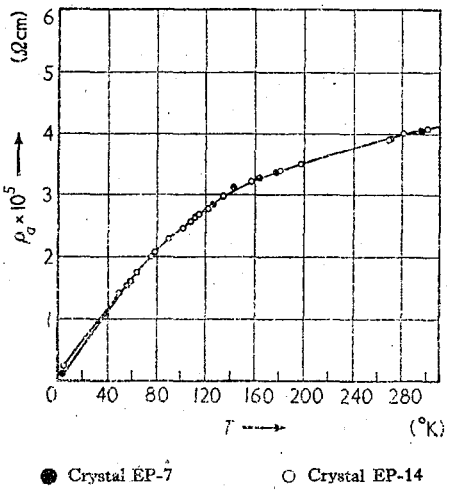


Figure 18. *Temperature dependence of the resistivity of graphite (After Soule)*

between 77° K and 298° K. Hence, up to $\simeq$ 120° K, the increase in the resistivities is mainly due to the decrease in the mobilities. Above this temperature, the increase in carrier concentration reduces the rate of increase of the resistivity, hence the knee in the curve. This implies an average degeneracy temperature of $\simeq$ 120° K for the carriers. An independent estimate can be obtained from the total band overlap ($2\gamma_2$). The average Fermi energy for the carriers should be just one-half of this overlap. Choosing $\gamma_2 = 0{\cdot}016$ eV, the value obtained earlier, this yields an average degeneracy temperature of $\simeq$ 190° K.

No attempts have yet been made to obtain further correlation of the overlap model with the electrical conductivities. Such an investigation might lead to information about the relevant scattering mechanisms, but cannot be expected to yield further information about the band structure.

Wallace, in his original paper,[3] has attempted to make some estimates of ρ_a and ρ_c. Recently, Haering and Wallace[22] have calculated the ratio ρ_c/ρ_a on the basis of the small γ_1 model. Assuming the same relaxation time for ρ_c as for ρ_a they find

$\rho_c/\rho_a \simeq 10^5$. This result cannot be trusted since the above assumption concerning the relaxation times cannot be justified. In general, it must be expected that the relaxation time is an arbitrary function of wave vector.

The resistivity ρ_c is much lower than the resistivity of crystals of multi-ring aromatic compounds as measured in the direction of the molecular stacks (the resistivity in the other direction is even higher). This can be readily understood from the point of view of the charge transfer effect between aromatic planes. As the number of rings in a molecule increases, the charge transfer interaction contributes more and more to the attractive force between layer planes, thus leading to a decrease in the interplanar distance. Increased charge transfer means a higher conductivity; unfortunately, not even an order of magnitude estimate of the resulting effect on ρ_c can be obtained in this manner.

4.2. **Paramagnetic Resonance Absorption**

The question of the nature of paramagnetic spin centres in chars, carbons, and graphites is still in a somewhat confused state in spite of some recent important advances in the field. As far as crystals of graphite are concerned, however, the situation has been lately greatly clarified. The paramagnetic resonance in graphite has two sources: (1) localized spin centres at the boundaries of crystals and at all kinds of defects; and (2) non-localized spins of free carriers. The spin resonance in graphite was discovered by Castle[76] and independently by Hennig and Smaller.[77] In highly purified Ceylon graphite powder, Castle found a line corresponding to $g = 2{\cdot}008$ with a width of about 20 oersteds and in view of the low spin concentration of 10^{-4}–10^{-5} *per atom* ascribed the absorption to free carriers. Hennig, Smaller, and Yasaitis[78] performed experiments with doped graphite and came to the conclusion that the spin centres formed by doping are different from those observed in regular graphite. This result would seem to support Castle's original view. Later, Hennig and Smaller[79] observed that exposure of samples to air and subsequent heat treatment strongly affected the intensity of the absorption. This important observation implies that the spin centres are localized. In view of the high spin concentration observed (10^{-3}–10^{-4}), Hennig and Smaller[79] concluded that the spin centres must be defects at internal surfaces or mosaic boundaries. In all these experiments graphite samples were finely ground in air before the measurements were performed. It is well known that when graphite is ground, the powder acquires a great chemical activity and absorbs oxygen abundantly; the oxygen being bound at the free carbon valences created along the broken surfaces. This oxygen can be driven out (presumably in the form of some complexes) only by heat treatment to above 800° C.

In order to see what happens to the broken carbon bonds when oxygen is not present, Andrew, Akamatu, and Mrozowski[80] have performed some preliminary experiments on grinding polycrystalline graphite in a neutral atmosphere. These experiments were further extended by Mrozowski and Andrew.[81] It was found that grinding creates a large number of spin centres ($g = 2{\cdot}0075$ and line width ≥ 15 oersteds) the concentration of which increases with the duration of grinding. When the grinding is performed in a helium atmosphere, subsequent admission of air kills about a third of the resonance irreversibly; but no matter under what conditions the grinding is performed (in vacuum, in helium, or in air) large numbers of spin centres will still be present and a strong resonance absorption will still be

observed. However, the resonance gradually weakens as the sample is heat treated to higher and higher temperature (in vacuum or in a neutral atmosphere) and disappears completely after heat treatment to 1,000° C whether or not the sample was exposed to air after grinding. This shows that the carbon atoms on the boundary of the aromatic lattice change from their regular state in which they possess a free valence to a different state in which there is no free spin available. This transition requires an activation energy, which is supplied in heat treatment. In reality there is a whole spectrum of activation energies, since the peripheral states formed in grinding probably have a variety of bond strain energies to start with. This new (saturated) state of the peripheral atoms would explain the absence of chemical absorption by powders which have been heat treated. Atoms in this state probably constitute some kind of electronic trap for π electrons. This is suggested by the work of Loebner[82] who has shown that grinding of polycrystalline graphite in air and subsequent removal of oxygen by heat treatment to 1,000° C changes the thermoelectric power of graphite, indicating the creation of an excess of holes in the π band. Coulson[83] explains the nature of this new state by proposing that single carbon atoms on the periphery of the crystals change under the influence of heat treatment from the configuration $2s2p^3$ with four valencies (one unsaturated σ bond) to the basic configuration $2s^2 2p^2$ with two valencies (saturated bonds). Such peripheral atoms could trap π electrons into their $2p_z$ orbitals. On edges where two carbon atoms belonging to a single aromatic ring are exposed, Coulson[83] expects a change into a state with a preponderant acetylene type triple bonding, again leading to a disappearance of the spin resonance. One can visualize, however, situations for some internal defects where a jump into an inactive state is not possible. Therefore, one can expect that the resonance due to localized spin centres will not be entirely removed by heat treatment (unless the graphite crystal used is quite free of internal defects).

Recently, Wagoner[84] has observed the spin resonance in a single crystal of graphite. The resonance is much weaker than the one observed for ground samples. The most striking feature of Wagoner's results is the large g-value anisotropy. At room temperature, g varies from $2{\cdot}0026 \pm 0{\cdot}0002$ to $2{\cdot}0495 \pm 0{\cdot}0002$ as the magnetic field is oriented from perpendicular to parallel to the c_0 axis. The g-value anisotropy increases with decreasing temperature: $g_\parallel$ becomes $2{\cdot}127$ at 77° K while $g_\perp$ remains constant. The line width is only a few oersteds. Wagoner establishes the resonance as being due to free carriers by showing that the observed line has the Dysonian shape expected for mobile spin centres and by demonstrating that the number of spin centres agrees roughly with the number of free carriers expected on the basis of the band structure. It is further shown that the temperature dependence of the number of spin centres agrees with band model predictions. At present, it is not known whether both electrons and holes contribute to the spin resonance. The large g shift must be due to spin–orbit coupling effects. Such effects were shown by Slonczewski[20] to have a small effect on the band structure. In a discussion of the spin resonance they will, however, be of prime importance, and it may be expected that a detailed analysis of Wagoner's results will yield further information about the band structure.

Abragam, Landesman, and Winter[85] have found an Overhauser effect in graphite. They observed the resonance absorption of carbon-13 nuclei and found that the nuclei exhibit a long relaxation time t_1 (2 – 10 min). When the sample was

irradiated with microwaves polarizing the electronic spins for a time long compared to t_1, the absorption of nuclei polarized by the interaction with electronic spins could be observed after the microwave field was turned off. An increase in signal by a factor of 300 in comparison with a non-irradiated sample was found at room temperature. The factor decreases very rapidly with decreasing temperature. One suspects that the electronic spins responsible for the polarization of the nuclei in these experiments are those connected with the free carriers; unfortunately, the electron spin resonance was not directly observed by these investigators. Abragam, Landesman, and Winter find that the Overhauser effect in neutron irradiated graphites shows a different behaviour, more in line with the behaviour of chars. It is likely that in this case the electronic spins in question are those localized at the defects.

4.3. Rhombohedral Graphite

Lipson and Stokes[86] established that graphite crystals sometimes show a rhombohedral structure which differs from the Bernal structure in that the stacking sequence is 123123.... A number of other investigators have confirmed this result.[87–89] The fraction of the rhombohedral form present may vary from 0 to 30 per cent. Hoerni[87] has found one purely rhombohedral crystal. When the rhombohedral form is present, groups of layers which are stacked this way alternate with layers which follow the more usual Bernal stacking, thus forming a mosaic. Well crystallized graphites which contain only a small percentage of the rhombohedral structure may have this percentage increased up to 20 per cent by mechanical grinding (Bacon[90]). Boehm and Hofmann[88] have found that the rhombohedral form is formed by slight pressure, and that this form disappears on heat treatment at 2,000°–3,000° C.

Since the essential difference between the above mentioned crystalline forms is the stacking order, it is reasonable to expect that a comparison of the electronic properties of these two forms would lead to useful information concerning the band parameter γ_1. This led Haering[13] to investigate the band structure of rhombohedral graphite. The band structure which emerges from this calculation is surprisingly complex. In the Wallace approximation, the conduction and valence bands still touch as they do for Bernal graphite. However, the loci of these touching points are no longer the lines HKH and H'K'H' (cf. Figure 5). Instead, these loci lie on cylinders whose axes are the lines HKH (or H'K'H') and whose radii depend on γ_1. It seems certain that a study of the electronic properties of such a crystal would contribute greatly to our understanding of graphite.

4.4. Graphite Compounds, Lattice Vacancies, and Radiation Damage Effects

Although the study of graphite compounds is not directly related to the study of a single crystal of elemental graphite, it is felt that a brief outline of our knowledge of such compounds is in order. The reason for this is the fact that numerous checks on the validity of the band structure model involve changing the carrier concentration by doping techniques. If we are to interpret such doping experiments, we must understand how the added impurities enter the graphite lattice, how they distribute themselves, and what their effect is on the electronic energy spectrum. A good deal of work along these lines has been done by Hennig and his

associates. For a summary of this work and for references, the reader is referred to Hennig.[91] Graphite compounds are classified according to whether the foreign atoms enter the lattice substitutionally or interstitially. Boron atoms are thought to enter substitutionally. By analogy with the semiconductors silicon and germanium, it is believed that atoms of nitrogen, phosphorus, antimony, and aluminium also enter graphite in this manner. A lattice vacancy may also be thought of as a substitutional compound. Vacancies and other lattice defects can apparently be introduced even into a single crystal by quenching from high temperatures.

Our knowledge of interstitial compounds is somewhat more complete than our knowledge of substitutional compounds. Atoms, molecules, and ions lodge themselves between the carbon layers. Such interstitial groups may act either as donors or as acceptors. In certain cases the atoms may also form covalent bonds with adjacent carbon atoms. The interpretation of the electronic properties of interstitial compounds is not simple and presents many serious problems. From the point of view of doping experiments, substitutional impurities are more attractive, since these probably distort the graphite lattice to a lesser extent and do not provide a 'bridge' for electron transfer between adjacent layers.

Radiation damage affects the electronic properties in two ways. Firstly, the number of free carriers is changed because of the formation of vacancies and possibly also of interstitial atoms and, secondly, the number of scattering centres is changed. The observed Hall coefficient, susceptibility, and thermoelectric power have been roughly correlated with the two-dimensional band structure and it has been found that qualitative agreement is possible (see a review by Hove[92] and a paper by Kinchin[93]). In view of the complexity of such damage effects, this is all that can be hoped for at the present time. Strong irradiation breaks up the graphite structure and changes a graphite crystal into a polycrystalline specimen. To some extent this is the opposite of the graphitization process by heat treatment. Therefore, the problems of radiation damage and its thermal annealing have much in common with the problems of polycrystalline graphites and carbons, and analogies can easily be noticed.[94]

5. SUMMARY OF THE BAND STRUCTURE

The graphite lattice is made up of hexagonal layers, usually stacked according to a 1212... stacking scheme. Each carbon atom has four valence electrons, three of which form tight covalent σ bonds with the three nearest-neighbour atoms located in the same plane. The remaining electron is in a π state and is more loosely bound. Since the interlayer distance (3·37 Å) is large compared to the interatomic spacing of atoms in a given plane (1·42 Å), a first approximation is obtained by considering a single layer of carbon atoms. Such a calculation shows that the three occupied and the three unoccupied σ bands are separated by a gap of approximately 5 eV and that this gap encloses the Fermi level in the π bands. The occupied and unoccupied π bands touch at the corners of the hexagonal Brillouin zone and the Fermi 'surface' consists of arcs of circles near the six zone corners.

The band structure of three-dimensional graphite is expected to be similar in the gross features to that of two-dimensional graphite, but may differ in detail near degeneracy points such as those of the π bands. The effect of σ–π mixing should now be included but this is expected to be small. The best approach is to study the

band structure in the vicinity of the zone edges by using group theory. Such an analysis yields the most general allowable energy surfaces in terms of a number of parameters. These parameters must be determined by experiment.

The small γ_1 model is an attempt to determine these parameters by experiment. From the calculation of the constant part of the susceptibility it is concluded that $\gamma_0 \simeq 2{\cdot}6$ eV and that $\gamma_1 \lesssim 0{\cdot}005$ eV. The two periods in the de Haas–van Alphen effect (or in the galvanomagnetic phenomena) may then be fitted only if a number of excess electrons is postulated ($\sim 10^{-4}$ per atom). The model also yields approximately the correct orientation dependence of the susceptibility fluctuations and roughly the correct specific heat. None of the above experiments determine whether the carriers are electrons or holes or both. Experiments able to settle this point are those dealing with the cyclotron resonance and Hall effect. Both of these indicate that electrons and holes are present simultaneously even at low temperatures. The small γ_1 model can account for the presence of holes only at high temperatures.

The overlap model is based on a small vertical overlap of the π bands, mainly the result of the parameter γ_3. Other band parameters are γ_0, γ_1, γ_3, γ_4, and Δ. γ_0 is the same parameter as enters into the two-dimensional band structure, and is determined by the susceptibility to be $\simeq 3$ eV, in good agreement with theoretical estimates. γ_1 represents the main interlayer interaction. The ratio γ_0^2/γ_1 is determined by the de Haas–van Alphen effect, by cyclotron resonance, and by the electronic specific heat to be $\simeq 25$ eV. With $\gamma_0 \simeq 3$ eV, this yields $\gamma_1 \simeq 0{\cdot}36$ eV. Independent estimates of γ_1 can be made from electron energy loss studies ($\gamma_1 = 0{\cdot}32$ eV) and from infra-red absorption data ($\gamma_1 = 0{\cdot}14$ eV). Experiments with rhombohedral graphite would provide a sensitive criterion for γ_1. γ_2 mainly determines the total band overlap and hence the number of carriers present. The analysis of the de Haas–van Alphen effect yields $\gamma_2 \simeq 0{\cdot}016$ eV. A rough check is obtained from the 'knee' in the resistivity curve. γ_3 is the chief reason for the trigonal warping of the energy surfaces. In principle, it should be determined by the cyclotron resonance data, but this has not yet been done. The number of minority carriers contributing to cyclotron resonance is quite large. It is not known whether a reasonable value of γ_3 ($\gamma_3 \simeq 0{\cdot}1$ eV) can yield this number. γ_4 has so far always been neglected and nothing is known about its magnitude. Δ reflects the non-identical surroundings of the A and B sites. From the de Haas–van Alphen effect, a value of $\simeq 0{\cdot}02$ eV is obtained.

Galvanomagnetic measurements yield convincing evidence about the gross features of the band structure, such as the simultaneous presence of majority and minority electrons and holes. A very rough indirect check on the values of the band parameters is also obtained, in so far as it is possible to estimate carrier concentrations from this data. It may also be expected that an analysis of the spin resonance data will yield information concerning the spin–orbit interaction and perhaps σ–π mixing.

The overlap model allows us to understand a large number of phenomena in a qualitative way. This in itself is a great success. It is also possible to obtain a great deal of quantitative agreement with this model. In view of the simplifying assumptions which have been made in the various calculations, an overall agreement of $\pm$ 10–20 per cent must be considered very satisfactory. However, there are two features which are somewhat out of line at the moment. The first concerns

the number of minority carriers observed in cyclotron resonance. The second is the value of γ_1 obtained from an analysis of the infra-red absorption data. In addition, there is the worry about the field dependence of the susceptibility at low fields. But by far the most serious gap in our understanding of graphite is the lack of an adequate model for describing impurity states. Let us consider this briefly.

In the semiconductors germanium and silicon, we have a fairly satisfactory theory of impurity states provided that these localized states are shallow; i.e. provided that they have small ionization energies and large Bohr orbits. Consider now the hypothetical experiment of making the gap between the conduction and valence bands of such a semiconductor smaller and smaller. What will happen to such an impurity state? Apart from the fact that we must now describe the state by considering both the valence and conduction bands, the ionization energy of such a state will become smaller and its Bohr orbit will increase. Finally, if the bands touch, we may expect the state to be completely non-localized. It seems therefore that impurities might alter the density of states near such a degeneracy point. Before doping experiments can be properly understood, we must know the answer to questions such as the one discussed here.

REFERENCES

1. See, for example, J. R. Reitz. *Solid State Physics* **1**, 1 (Academic, New York, 1955)
2. See, for example, C. A. Coulson. *Valence* (Oxford University Press, 1953)
3. P. R. Wallace. *Phys. Rev.* **71**, 622 (1947)
4. W. M. Lomer. *Proc. roy. Soc.* (L) **A227**, 330 (1955)
5. F. J. Corbato. *Quarterly Progress Report*, Solid State and Molecular Theory Group, MIT, No. 21; also Ph.D. Thesis, M.I.T. (1956); *Proceedings of the Third Conference on Carbon* p. 173 (Pergamon, London, 1959)
6. E. Taft and L. Apker. *Phys. Rev.* **99**, 1831 (1955)
7. See the article by D. H. Tombulian in *Handbuch der Physik* Vol. 30 (Springer, Berlin, 1957) pp. 274–6 for a review of X-ray data and for references to original papers
8. C. A. Coulson and H. Taylor. *Proc. phys. Soc. Lond.* **A65**, 815 (1952)
9. C. A. Coulson. *Nature, Lond.* **159**, 265 (1947)
10. J. L. Carter. Ph.D. Thesis, Cornell University (1953)
11. J. C. Slonczewski and P. R. Weiss. *Phys. Rev.* **109**, 272 (1958)
12. J. D. Bernal. *Proc. roy. Soc.* **A106**, 749 (1924)
13. R. R. Haering. *Canad. J. Phys.* **36**, 352 (1958)
14. D. F. Johnston. *Proc. roy. Soc.* **A237**, 48 (1956)
15. J. L. Carter and J. A. Krumhansl. *J. chem. Phys.* **21**, 2238 (1953)
16. J. W. McClure. *Phys. Rev.* **103**, 612 (1957)
17. D. Polder. Unpublished, see reference 38
18. G. H. Horton and G. E. Tauber. Unpublished
19. D. F. Johnston. *Proc. roy. Soc.* **A227**, 349 (1955)
20. J. C. Slonczewski. Ph.D. Thesis, Rutgers University (1955)
21. P. Nozières. *Phys. Rev.* **109**, 1510 (1958); *Proceedings of the Third Conference on Carbon* p. 197 (Pergamon, London, 1959)
22. R. R. Haering and P. R. Wallace. *J. Phys. Chem. Solids* **3**, 253 (1957)
23. W. S. Boyle and P. Nozières. *Phys. Rev.* **111**, 782 (1958)
24. R. R. Haering (1959). Unpublished
25. J. Murphy. M.Sc. Thesis, McGill University (1958)
26. S. Storey. Ph.D. Thesis, McGill University (1959)
27. J. K. Galt, W. A. Yager, and H. W. Dail, Jr. *Phys. Rev.* **103**, 1586 (1956); J. K. Galt, W. A. Yager, and F. R. Merritt. *Proceedings of the Third Conference on Carbon* p. 193 (Pergamon, London, 1959)
28. D. E. Soule. *Bull. Amer. phys. Soc.* **1**, 255 (1956)
29. D. E. Soule. *Phys. Rev.* **112**, 698 (1958)
30. B. Lax and H. J. Zeiger. *Phys. Rev.* **105**, 1466 (1957)
31. Y. H. Ichikawa. *Phys. Rev.* **109**, 653 (1958)
32. J. W. McClure. To be published in *Phys. Rev.*

33. N. GANGULI and K. S. KRISHNAN. *Proc. roy. Soc.* **A177**, 168 (1941)
34. S. MROZOWSKI. *Phys. Rev.* **85**, 609 (1952) and 'Errata' **86**, 1056 (1952)
35. R. SMOLUCHOWSKI. *Rev. mod. Phys.* **25**, 178 (1953)
36. W. P. EATHERLY (1953). See discussion following the paper by Smoluchowski[35]
37. J. E. HOVE. *Phys. Rev.* **100**, 645 (1955)
38. S. MROZOWSKI. *J. chem. Phys.* **21**, 492 (1953)
39. S. J. MASE. *J. phys. Soc. Japan* **13**, 563 (1958)
40. R. E. PEIERLS. *Z. Phys.* **80**, 763 (1933)
41. E. N. ADAMS. *Phys. Rev.* **89**, 633 (1953)
42. J. W. MCCLURE. *Phys. Rev.* **104**, 666 (1956)
43. G. R. HENNIG and J. D. MCCLELLAND. *J. chem. Phys.* **23**, 1431 (1955)
44. Y. UEMURA and M. INOUE. *J. phys. Soc. Japan* **13**, 382 (1958)
45. A. PACAULT and A. MARCHAND. *C. R. Acad. Sci., Paris* **241**, 489 (1955); *Proceedings of the Third Conference on Carbon* p. 37 (Pergamon, London, 1959); A. MARCHAND. *Ann. Chim.* (13) **2**, 469 (1957)
46. A. MARCHAND. *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 61, p. 26. To be published in the *Proceedings of the Fourth Conference on Carbon*
47. D. SHOENBERG. *Phil. Trans.* **245**, 1 (1952)
48. T. G. BERLINCOURT and M. G. STEELE. *Phys. Rev.* **98**, 956 (1955)
49. I. M. LIFSHITZ and A. M. KOSEVICH. *Soviet Phys. JETP* **2**, 636 (1956)
50. D. E. SOULE. *Phys. Rev.* **112**, 708 (1958)
51. J. W. MCCLURE and L. B. SMITH. TM 381, National Carbon Research Laboratories, Cleveland, Ohio (1959)
52. D. E. SOULE and J. W. MCCLURE. TM 356, National Carbon Research Laboratories, Cleveland, Ohio (1958)
53. R. G. CHAMBERS. *Phil. Mag.* **8**, 459 (1956)
54. H. J. ZEIGER, B. LAX, and D. L. DEXTER. *Phys. Rev.* **105**, 495 (1957)
55. S. P. F. HUMPHREYS-OWEN and L. A. GILBERT. *Industrial Carbon and Graphite, London Conference* p. 37 (Society of Chemical Industry, London, 1958)
56. A. BRAUN and G. BUSCH. *Helv. phys. acta* **20**, 33 (1947)
57. J. T. MCCARTNEY and S. ERGUN. *Proceedings of the Third Conference on Carbon* p. 223 (Pergamon, London, 1959)
58. M. R. NULL and W. W. LOZIER. *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 85, p. 34. To be published in the *Proceedings of the Fourth Conference on Carbon*
59. J. D. PLUNKETT and W. D. KINGERY. *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 86, p. 34. To be published in the *Proceedings of the Fourth Conference on Carbon*
60. P. H. KRESOM and N. PEARLMAN. *Phys. Rev.* **99**, 1119 (1955)
61. W. DESORBO and G. E. NICHOLS. *J. Phys. Chem. Solids* **6**, 352 (1958)
62. K. KOMATSU and T. NAGAMIYA. *J. phys. Soc. Japan* **6**, 438 (1958)
63. J. C. BOWMAN and J. A. KRUMHANSL. *J. Phys. Chem. Solids* **6**, 367 (1958)
64. D. F. JOHNSTON. *Proc. roy. Soc.* **A227**, 359 (1955)
65. G. H. KINCHIN. *Proc. roy. Soc.* **A217**, 9 (1953)
66. J. W. MCCLURE. *Phys. Rev.* **112**, 715 (1958)
67. J. W. MCCLURE. *Phys. Rev.* **101**, 1642 (1956)
68. G. E. ZILBERMAN. *Soviet Phys. JETP* **2**, 650 (1956)
69. A. R. UBBELOHDE and J. ORR. *Nature, Lond.* **179**, 193 (1957)
70. J. ROTHSTEIN. *Phys. Rev.* **95**, 370 (1954)
71. K. S. KRISHNAN and N. GANGULI. *Nature, Lond.* **144**, 667 (1939)
72. A. K. DUTTA. *Phys. Rev.* **90**, 187 (1953)
73. G. E. WASHBURN. *Ann. Physik* **4**, 236 (1915)
74. W. PRIMAK and L. H. FUCHS. *Phys. Rev.* **95**, 22 (1954)
75. J. M. REYNOLDS, H. W. HEMSTREET, and T. E. LEINHARDT. *Phys. Rev.* **91**, 1152 (1953)
76. J. CASTLE. *Phys. Rev.* **92**, 1063 (1953)
77. G. R. HENNIG and B. SMALLER. Argonne National Laboratory *Progress Report* ANL-5101 (1953)
78. G. R. HENNIG, B. SMALLER, and E. L. YASAITIS. *Phys. Rev.* **95**, 1088 (1954)
79. G. R. HENNIG and B. SMALLER. *Proceedings of the First and Second Conferences on Carbon* (University of Buffalo, New York, 1956)
80. J. F. ANDREW, H. AKAMATU, and S. MROZOWSKI. *Bull. Amer. phys. Soc.* **3**, 136 (1958)
81. S. MROZOWSKI and J. F. ANDREW. *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 4, p. 8. To be published in the *Proceedings of the Fourth Conference on Carbon*
82. E. LOEBNER. Ph.D. Thesis, University of Buffalo (1955)

83. C. A. COULSON. *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 5, p. 8. To be published in the *Proceedings of the Fourth Conference on Carbon*
84. G. WAGONER, TM 384, National Carbon Research Laboratories, Cleveland, Ohio (1959); also *Abstracts of the Fourth Conference on Carbon*, University of Buffalo, 1959, paper 24, p. 15. To be published in the *Proceedings of the Fourth Conference on Carbon*
85. A. ABRAGAM, A. LANDESMAN, and J. M. WINTER. *C. R. Acad. Sci., Paris* **247**, 1852 (1958)
86. H. LIPSON and A. R. STOKES. *Proc. roy. Soc.* **A181**, 101 (1942)
87. J. HOERNI. *Helv. phys. acta* **23**, 587 (1950)
88. H. P. BOEHM and U. HOFMANN. *Z. anorg. Chem.* **278**, 58 (1955)
89. A. HEROLD. *Bull. Soc. Chim. Fr.* 999 (1955)
90. G. E. BACON. *Acta Cryst., Camb.* **3**, 320 (1950)
91. G. R. HENNIG. *Proceedings of the First and Second Conferences on Carbon* p. 103 (University of Buffalo, New York, 1956) and a later review to be published in the *Proceedings of the Fourth Conference on Carbon*
92. J. E. HOVE. *Proceedings of the First and Second Conferences on Carbon* p. 125 (University of Buffalo, New York, 1956)
93. G. H. KINCHIN. *J. Nucl. Energy* **1**, 124 (1954)